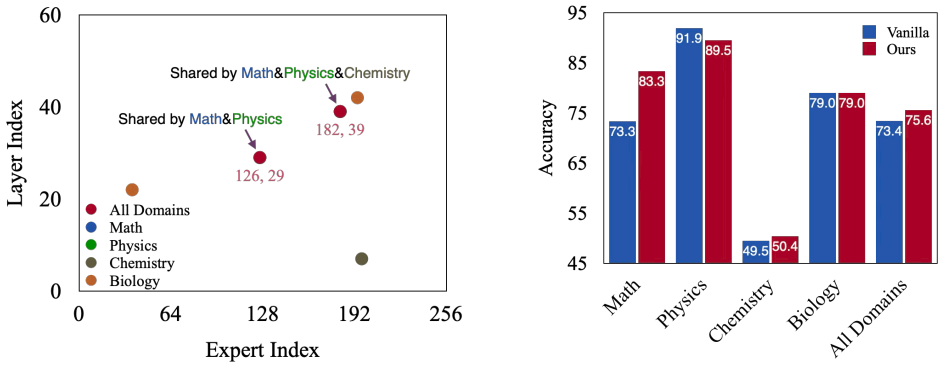
Two Experts Are All You Need for Steering Thinking: Reinforcing Reasoning in MoE Models Without Additional Training

Mengru Wang[*,1] **, Xingyu Chen**[*,2] **, Yue Wang**[*,2] **, Zhiwei He**[*,2] **, Jiahao Xu**[2] **, Tian Liang**[2] **, Qiuzhi Liu**[2] **, Yunzhi Yao**[1] **, Wenxuan Wang**[2] **, Ruotian Ma**[2] **, Haitao Mi**[2] **, Ningyu Zhang**[†,2] **, Zhaopeng Tu**[†,2] **, Xiaolong Li**[2] **, and Dong Yu**[2]

[1]Zhejiang University
[2]Tencent



(a) Cognitive Experts (b) Reasoning Accuracy

Figure 1: (a) Illustration of cognitive experts identified across domains. (b) Reinforcing only the top two experts (in red color) can improve reasoning accuracy without additional training.

Abstract

Mixture-of-Experts (MoE) architectures within Large Reasoning Models (LRMs) have achieved impressive reasoning capabilities by selectively activating experts to facilitate structured cognitive processes [1, 2]. Despite notable advances, existing reasoning models often suffer from cognitive inefficiencies like overthinking [3] and underthinking [4]. To address these limitations, we introduce a novel inference-time steering methodology called **Rei**nforcing **C**ognitive **E**xperts (RICE), designed to improve reasoning depth and efficiency without additional training or complex heuristics. Leveraging normalized Pointwise Mutual Information (nPMI), we systematically identify specialized experts, termed **cognitive experts** that orchestrate meta-level reasoning operations characterized by tokens like "`<think>`". Empirical evaluations with leading MoE-based LRMs (DeepSeek-R1 and Qwen3-235B) on rigorous quantitative and scientific reasoning benchmarks (AIME and GPQA Diamond) demonstrate noticeable and consistent improvements in reasoning accuracy, cognitive efficiency, and cross-domain generalization. Crucially, our lightweight approach substantially outperforms prevalent reasoning-steering techniques, such as prompt design and decoding constraints, while preserving the model's general instruction-following skills. These results highlight reinforcing cognitive experts as a promising, practical, and interpretable direction to enhance cognitive efficiency within advanced reasoning models.

[*]Equal Contribution. Work was done when Mengru, Xingyu, Yue, and Zhiwei were interning at Tencent.
[†]Correspondence to: Zhaopeng Tu <zptu@tencent.com> and Ningyu Zhang <zhangningyu@zju.edu.cn>.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

1 Introduction

Models capable of extended reasoning, often referred to as Large Reasoning Models (LRMs) like OpenAI's o1 [5] and DeepSeek-R1 [1], have significantly advanced machine intelligence, largely by scaling test-time computation [6, 7]. Despite their impressive capabilities, these LRMs remain susceptible to inefficiencies [8–15]. Prior work has sought to address these issues through approaches such as preference optimization [3], decoding penalties [4], and various other techniques. In this work, we tackle these problems from a novel perspective: potential expert specialization in Mixture-of-Experts (MoE) architecture.

Due to the computational resource efficiency brought about by its sparsity, the MoE architecture has been increasingly adopted by state-of-the-art (SOTA) models, such as DeepSeek-R1 [1] and Qwen3 [2]. This sparse, specialized activation paradigm bears a conceptual resemblance to functional specialization in the human brain, where targeted interventions can modulate cognitive functions and behaviors [16–19]. Inspired by this principle, we systematically investigate whether undesirable reasoning behaviors in MoE-based LRMs correlate with the activation patterns of specific experts, and critically, if strategic manipulation of these experts can ameliorate such issues.

We introduce an approach to identify and modulate key experts integral to the reasoning process. By analyzing the co-occurrence of explicit linguistic markers of thought (e.g., "`<think>`" and "`</think>`") with individual expert activations, we pinpoint a subset of experts highly correlated with the model's cognitive deliberations. We designate these critical components as **cognitive experts**. Through extensive experimentation with SOTA MoE-reasoning models DeepSeek-R1 [1] and Qwen3-235B [2] on challenging math and scientific reasoning benchmarks, we demonstrate that selectively amplifying **as few as two cognitive experts** can enhance both reasoning depth and efficiency. Notably, our approach achieves marked accuracy improvements while reducing token usage in critical reasoning tasks, outperforming existing steering methods such as prompting and decoding constraints [4].

Moreover, we showcase impressive generalization and robustness of cognitive expert modulation, observing consistent improvements in unseen and more complex reasoning scenarios while maintaining or even enhancing general instruction-following capabilities. Our findings provide strong evidence that modulating selective experts responsible for meta-level reasoning is effective, efficient, and broadly applicable across domains, paving the way for lightweight and interpretable model steering in increasingly sophisticated MoE-based reasoning models.

Our main contributions are:

1. We propose a normalized Pointwise Mutual Information (nPMI) method for identifying cognitive experts within LRMs that are highly correlated with reasoning behavior, requiring only a single forward propagation and no additional training.
2. We introduce a lightweight inference-time steering strategy, named "reinforcing cognitive experts", that effectively enhances reasoning depth and accuracy without requiring any additional training or supervision signals.
3. Through comprehensive experiments on two prevalent MoE reasoning models and rigorous benchmarks, we empirically validate the efficacy, generalizability, and robustness of cognitive expert modulation, demonstrating significant improvements in cognitive efficiency and problem-solving accuracy.

2 Identifying Cognitive Experts

In this section, we leverage normalized Pointwise Mutual Information (nPMI) [20] to quantify the correlation between model thinking and each expert in a Mixture of Experts (MoE) reasoning model. We hypothesize that there are some "cognitive experts" selected by nPMI metric, which orchestrate meta-level reasoning for complex tasks.

2.1 Expert Specialization in MoE Models

In large reasoning models, deep thinking is manifested through key tokens, such as "`<think>`" to initiate reasoning, "`</think>`" to terminate it, and tokens like "`recheck`" to guide introspection. In

the MoE framework, these tokens are generated during forward propagation through various model components, including the MoE routing mechanism that assigns them to specialized experts, with weights determining each expert's contribution.

Formally, let us consider an MoE framework [21] with N experts, denoted $\{E_1, \ldots, E_i, \ldots, E_N\}$, at each layer. For each input token x, a gating function selects a subset $S \subset \{E_1, \ldots, E_O\}$ of O experts ($O \leq N$), where $|S| = O$, and assigns weights w_i (with $\sum_{i \in S} w_i = 1$) to each selected expert $E_i \in S$. The output h_x for token x is computed as:

$$h_x = \sum_{i \in S} w_i \cdot E_i(x), \quad \text{where } |S| = O, \tag{1}$$

where $E_i(x)$ represents the output of expert E_i, and w_i is the weight of the i-th selected expert. Prior work on MoE models shows that expert routing is often token-dependent [22], but recent study [23, 24] indicates that DeepSeek-R1's advanced reasoning enables its expert routing to focus on semantic specialization, surpassing token-dependent methods. We hypothesize that experts with consistently high co-occurrence scores with thinking tokens serve as key "cognitive experts" responsible for meta-level reasoning.

Measuring Correlation of Specialized Experts and Thinking Tokens To examine whether a given expert consistently governs the model's reasoning process, we measure the co-occurrence between its activation and specific reasoning-related marker tokens, such as "`<think>`," "`</think>`", and others. Formally, let x represent a token and y denote expert E_i. We measure their association using pointwise mutual information (PMI). The PMI of x and y is defined as

$$\mathrm{PMI}(x, y) = \log_2 \frac{p(x, y)}{p(x)\, p(y)} = \log_2 \frac{p(y|x)}{p(y)}, \tag{2}$$

where $p(x, y)$ is the joint probability that x and y both occur, while $p(x)$ and $p(y)$ are their individual (marginal) probabilities, and $p(y|x)$ is the conditional probability that y occurs given x.

For interpretability, we normalize PMI to the range $[-1, +1]$, yielding

$$\mathrm{nPMI}(x, y) = \frac{\mathrm{PMI}(x, y)}{-\log_2 p(x, y)}. \tag{3}$$

Thus, $\mathrm{nPMI}(x, y) \approx -1$ indicates that events x and y never co-occur, $\mathrm{nPMI}(x, y) = 0$ implies independence, and $\mathrm{nPMI}(x, y) \approx +1$ indicates they appear almost exclusively together (complete co-occurrence).

Let M be the number of instances in a dataset, and let T be the total number of tokens generated over all instances in the test set. We denote by k_n the number of times the expert E_i is activated specifically when the thinking token (e.g. "`<think>`") appears, and by K_n the total number of times E_i is activated across all tokens (including both thinking and non-thinking tokens). Since the reasoning model generally generates one thinking start and end token for each instance, then we can achieve the following functions when x denotes "`<think>`" or "`</think>`":

$$p(y = E_i|x) = \frac{k_n}{M}, \qquad p(y = E_i) = \frac{K_n}{T}, \qquad p(x, y = E_i|x) = \frac{k_n}{T}. \tag{4}$$

$$\mathrm{nPMI}(x, y = E_i) = \frac{\log_2(\frac{k_n}{M}) + \log_2(\frac{T}{K_n})}{\log_2(\frac{T}{k_n})}. \tag{5}$$

Intuitively, if an expert E_i is activated almost exclusively during "`<think>`" and rarely (or never) at other tokens, $k_n \approx K_n \approx M$, $\mathrm{nPMI}(x = \texttt{<think>}, y = E_i) \approx \frac{\log_2 1 + \log_2(\frac{T}{M})}{\log_2(\frac{T}{M})} \approx +1$, indicating that this expert is effectively tied to the thinking marker. In other words, the expert's entire usage focuses on activating the thinking token. Such specialists are prime candidates for "cognitive experts", given their consistently high co-occurrence with the thinking marker tokens.

2.2 Identify Cognitive Experts

We observe that some experts exhibit high nPMI scores with both "`<think>`" and "`</think>`", indicating a *bimodal association*. This suggests their broad involvement in the reasoning process

rather than specialization in its initiation. To prioritize experts specialized in initiating (rather than terminating) reasoning, we adopt the following selection strategy:

We define a set of thinking tokens $\Pi = \{\texttt{<think>}, \texttt{</think>}, \texttt{Alternatively}\}$. The normalized Pointwise Mutual Information (nPMI) score for expert E_i is formulated as:

$$\text{nPMI}_{E_i} = \sum_{x \in \Pi} c_x \cdot \text{nPMI}(x, y = E_i), \tag{6}$$

where x is a thinking token in set Π, c_x denotes the coefficient associated with the token x, assigned as $c_{\texttt{<think>}} = 1$, $c_{\texttt{</think>}} = -1$, and $c_{\texttt{Alternatively}} = -1$.

Then, we select the top-l experts based on their nPMI scores to form the *cognitive expert set* P. The weight adjustment for expert E_i is governed by the following condition:

$$w_i = \begin{cases} w_i \cdot \beta & \text{if } E_i \in S \text{ and } E_i \in P, \\ w_i & \text{otherwise,} \end{cases} \tag{7}$$

where $P = \{E_j \mid \text{nPMI}_{E_j} \text{ is among the top } l \text{ scores}\}$ denotes the set of *cognitive experts*, S is the subset selected by the gating function in Eq. 1, and β is the steering multiplier. In other words, once these experts are identified, we can reinforce reasoning in the MoE model by controlling their contribution through the hyperparameter β.

3 Experiments

Research Questions In this study, we investigate the following research questions:

- RQ1: Are there "cognitive experts" specialized in thinking? If so, do these experts differ across domains?
- RQ2: Can the identified cognitive experts effectively enhance cognitive effort within MoE models?
- RQ3: Do "cognitive experts" differ across various domains (e.g., math, physics, chemistry, and biology)?
- RQ4: Does reinforcing specific cognitive experts negatively impact the general problem-solving capabilities of MoE models?

3.1 Experimental Setup

MoE-based Reasoning Models Currently available open-source MoE architectures tailored for large reasoning models tasks include DeepSeek-R1 [1] and Qwen3-235B [2]. DeepSeek-R1 selects 8 experts from a total of 256 at each layer, whereas Qwen3-235B selects 8 experts from a total of 128. We primarily use the DeepSeek-R1 (671B) model for our experiments, supplemented by additional evaluations on the Qwen3-235B model to examine the generalizability of cognitive experts. Note that we provide more experimental details in §B.

Benchmarks We evaluate our approach on two challenging benchmarks designed specifically to test the reasoning abilities necessary for solving scientific problems across diverse domains:

- **AIME** [25]: a dataset from the American Invitational math Examination, which assesses advanced mathematical problem-solving skills. We use two recent test sets, AIME2024 and AIME2025, each comprising 30 problems.
- **GPQA Diamond** [26]: a comprehensive dataset of 198 expert-crafted multiple-choice questions in biology, chemistry, and physics, designed to test advanced scientific reasoning skills.

3.2 Cognitive Experts

To address RQ1, we first identify cognitive experts within two MoE reasoning models – DeepSeek-R1 [1] and Qwen3-235B [2] – across four scientific domains. Taking math as an illustrative example, we first use DeepSeek-R1 to generate answers on the AIME2024 dataset, simultaneously recording the expert selections at each token position during forward propagation. Next, we employ the nPMI

Table 1: Identified cognitive experts of DeepSeek-R1. Each entry (layer ID, expert ID) denotes the DeepSeek-R1 model layer ID and expert ID. "All" combines data from all domains.

Domain	**Identified Experts Ranked by nPMI Score**				
	1st	*2nd*	*3rd*	*4th*	*5th*
Math	(39, 182)	(29, 126)	(14, 114)	(27, 45)	(16, 129)
Physics	(29, 126)	(39, 182)	(36, 53)	(39, 46)	(24, 159)
Chemistry	(7, 197)	(39, 182)	(22, 37)	(29, 106)	(29, 126)
Biology	(42, 194)	(22, 37)	(37, 241)	(43, 61)	(39, 188)
All	(39, 182)	(29, 126)	(29, 106)	(4, 214)	(50, 120)

Table 2: Identified cognitive experts of Qwen3-235B. Each entry (layer ID, expert ID) denotes the Qwen3-235B model layer ID and expert ID. "All" combines data from all domains.

Domain	**Identified Experts**				
	Top-1	*Top-2*	*Top-3*	*Top-4*	*Top-5*
Math	(70, 47)	(23, 115)	(19, 47)	(75, 46)	(22, 88)
Physics	(2, 28)	(74, 65)	(4, 44)	(25, 103)	(7, 36)
Chemistry	(32, 58)	(26, 30)	(68, 35)	(37, 57)	(25, 103)
Biology	(2, 28)	(26, 30)	(67, 15)	(82, 29)	(25, 103)
All	(25, 103)	(26, 30)	(82, 29)	(67, 15)	(37, 57)

metric defined in Eq. 6 to identify the top five experts that exhibit the strongest statistical association with reasoning-related marker tokens (e.g., "<think>"). These experts are thus identified as the key cognitive experts specialized for mathematical reasoning. Analogously, we apply this procedure to the biology, chemistry, and physics questions in the GPQA Diamond dataset to identify cognitive experts in these respective domains. In the case of Qwen3-235B, we follow a similar procedure but generate domain-specific responses with the Qwen3-235B model itself. This ensures consistent identification signals that correspond directly to the model under examination.

We demonstrate the nPMI distribution of the top 10 experts of Deepseek-R1 across four domains (Math, Physics, Chemistry, and Biology) in Fig 2. Across the four domains, the top five experts exhibit nPMI values that are mostly above 0.5. Besides, the top 5 experts also indicate a sharply peaked distribution toward the other experts. In particular, the group of top five "thinking-specialized" experts shows significantly higher nPMI scores than the remaining experts, suggesting that domain reasoning is largely concentrated within a few highly specialized components. This pattern supports our hypothesis that a small number of experts are highly specialized for cognitive functions. Subsequently, we delve into the effectiveness of the top 5 experts in Table 3.

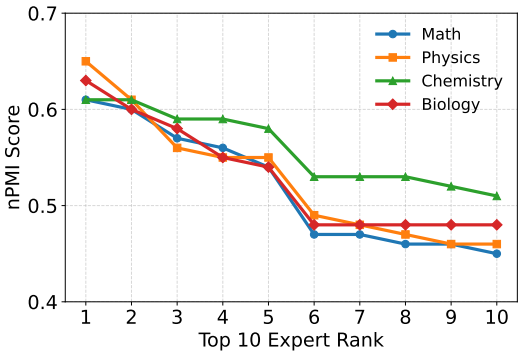


Figure 2: The nPMI distribution of top 10 expert of Deepseek-R1 across four domains.

Cognitive Experts Across Domains Cognitive experts identified within DeepSeek-R1 are summarized in Table 1. An analogous summary for Qwen3-235B is provided and discussed in Table 2. From Table 1, we observe that the top two cognitive experts in the math, physics, and the aggregated "All" domains are remarkably consistent: (39, 182) and (29, 126). This strongly suggests these experts play critical and reliable roles in reasoning tasks requiring increased cognitive effort, particularly in quantitative and logic-intensive domains. The significant overlap observed between math and physics further implies a shared underlying cognitive strategy—likely focusing on symbolic manipulation and structured logical inference—which the model employs consistently across these domains. Addi-

Table 3: Effect of Deepseek-R1 on AIME24 with reinforced cognitive experts, evaluated across different multipliers and varying numbers of Math-domain cognitive experts. "Random" denotes two randomly chosen experts. The row with Multiplier 1 denotes the performance of vanilla DeepSeek-R1.

Multiplier	Top1	Top2	Top3	Top4	Top5	Random
1			73.3			
2	70.0	70.0	76.7	73.3	73.3	70.0
4	76.7	**83.3**	73.3	66.7	76.7	73.3
8	76.7	73.3	**83.3**	73.3	73.3	70.0
16	80.0	80.0	1.7	76.7	73.3	73.3
32	70.0	**83.3**	73.3	73.3	73.3	76.7
64	80.0	**83.3**	60.0	53.3	50.0	66.7
128	70.0	**83.3**	43.3	26.7	13.3	63.3
256	73.3	60.0	10.0	6.7	0.0	73.3
512	63.3	46.7	6.7	3.3	0.0	63.3

tionally, the repeated appearance of certain experts in multiple domains supports our hypothesis: a subset of experts encodes generalized reasoning capabilities applicable across diverse scientific fields. Therefore, these cross-domain patterns indicate that DeepSeek-R1 may encode robust domain-general cognitive mechanisms, with some experts serving as reusable computational building blocks suitable for abstract reasoning and logical problem-solving tasks. Note that even within the same domain, there are distinctions. For instance, when comparing the top-5 cognitive experts for AIME24, MATH, and GSM8K, we find both shared and dataset-specific experts. Therefore, we hypothesize that the shared experts are responsible for general mathematical abilities, while the unique experts handle dataset-specific skills.

3.3 Effectiveness of Cognitive Experts

Reinforcing Cognitive Experts *To answer RQ 2*, we reinforce the identified top 5 cognitive experts from the Math (AIME24) and evaluate their performance under different reinforcement configurations on the same benchmark AIME24 (Table 3). The optimal hyperparameters – the number of cognitive experts l and the steering multiplier β—are selected based on this evaluation and used in all subsequent experiments. We then assess the generalization ability of these reinforced experts on the unseen, more challenging tasks from AIME25 (Table 4).

From Table 3, we observe that **reinforcing two top-ranked cognitive experts significantly enhances the model's reasoning ability**. Notably, using two experts with a steering multiplier of 4, 32, 64, or 128 achieves the highest accuracy of 83.3%. In contrast, applying an excessively large multiplier (e.g., 512) causes a dramatic drop in accuracy, often to near zero. This failure mode is characterized by the model repetitively generating meaningless tokens, suggesting that overly aggressive reinforcement disrupts the model's generation dynamics. Overall, moderate reinforcement of well-identified cognitive experts leads to consistent improvements, whereas over-reinforcement or random expert selection results in performance degradation. However, reinforcing two randomly selected experts across a wide multiplier range (2 to 512) yields minimal performance variation. Therefore, we use *two experts with a steering multiplier 64 for all subsequent experiments* [3].

We directly apply the cognitive experts identified from AIME24 to solve unseen and more challenging reasoning problems in AIME25. As shown in Table 4, these cognitive cognitive experts generalize well to the AIME25 test set. For DeepSeek-R1, the accuracy improves from 63.3% to 73.3% when guided by the identified cognitive experts. Similarly, for Qwen3-235B, accuracy increases from 66.7% to 73.3%. Additional pass@k performance using the model's officially recommended top-p sampling strategy (provided in §C.4) further supports this observation. The above phenomenon demonstrates **the transferability and robustness of the expert selection across tasks with higher cognitive demands**.

[3]This setup is designed to test the raw generalization of the math-derived setting. However, different domains may require different intensities of cognitive steering. We discuss this in detail in §C.1.

Table 4: Performance of our approach on the AIME24 and generalization on the unseen AIME25.

Benchmark	Method	Accuracy	Thoughts	#Tokens
AIME24	DeepSeek-R1	73.3	12.0	9,219
	+RICE {(39,182), (29,126)}	**83.3**	10.2	8,317
AIME25	DeepSeek-R1	63.3	17.0	10,875
	+RICE {(39,182), (29,126)}	**73.3**	15.2	11,441
AIME24	Qwen3-235B	86.7	20.1	10,956
	+RICE {(70,47), (23,115)}	86.7	16.2	10,722
AIME25	Qwen3-235B	66.7	19.7	15,013
	+RICE {(70,47), (23,115)}	**73.3**	16.8	13,935

Table 5: Effect of cognitive experts of Deepseek-R1 across different domains.

Domain	Math	Physics	Chemistry	Biology	Average
R1	73.3	91.9	49.5	79.0	73.4
Math	**83.3**	89.5	50.4	79.0	**75.6**
Physics	**83.3**	89.5	50.4	79.0	**75.6**
Chemistry	80.0	**95.4**	**52.7**	68.4	74.1
Biology	73.3	93.0	47.3	73.9	71.9
All	**83.3**	89.5	50.4	79.0	**75.6**

Crucially, the observed accuracy improvements do not necessarily entail increased computational cost in terms of token usage, supporting our hypothesis that our method encourages deeper thinking rather than just longer outputs. Our cognitive expert strategy, despite improving average accuracy of Deepseek-R1 on AIME24, uses more efficient reasoning thought [4] (10.2 vs 12.0) and tokens (8,317 vs 9,219) on average compared to the baseline. This efficiency phenomenon is also observed in Qwen3-2-35B, where the substantial accuracy gain (+6.6%) is accompanied by a notable reduction in thought (16.8 vs 19.7) and token count (13,935 vs 15,013). This suggests that **reinforcing cognitive experts helps the model to reason more effectively**, focusing computational effort more productively within the reasoning process without generating excessive verbosity. The reasoning effectiveness can be clearly observed in Table 8, where our RICE demonstrates deeper and more consistent reasoning, leading directly to the correct answer. In contrast, vanilla DeepSeek-R1 exhibits more frequent shifts in reasoning and fails to commit to its initially correct deductions.

3.4 Performance of Cognitive Experts across Domains

To address RQ3, we evaluate the transferability of domain-specific cognitive experts by applying expert sets identified from one domain to others. As the top-2 experts selected from *Math*, *Physics*, and the *All* domains are identical, their results are the same across domains. As shown in Table 5, we have several observations:

Cognitive experts generalize well across domains. Our evaluation, summarized in Table 5, clearly illustrates the efficacy of the identified cognitive experts in enhancing the DeepSeek-R1 model's reasoning capability across multiple domains. Leveraging cognitive experts identified from aggregated data ("All" domains) shows marked overall improvement, raising the average accuracy from 73.4% to 75.6%. Notably, substantial improvement is observed in the math tasks (from 73.3% to 83.3%). Moderate accuracy gains are also seen in Chemistry (from 49.5% to 50.4%) and minor degradation observed in Physics (from 91.9% to 89.5%), indicating broad applicability and effectiveness of these general reasoning modulators across diverse problem sets. Biology tasks show stable performance, unaffected by general expert modulation.

[4]We use the underthinking score from prior work [4] to quantify reasoning efficiency, with lower Thought values indicating greater efficiency.

Domain-specific expert sets provide targeted gains. Further analysis demonstrates the nuanced implications of domain-specific cognitive experts. Chemistry-identified experts outperform general experts significantly within their native Chemistry domain (49.5% to 52.7%) and notably enhance Physics performance (91.9% to 95.4%), highlighting potential cross-domain synergies between physics and chemistry reasoning processes. However, this specialization lowers the accuracy in math (from 83.3% with general experts to 80.0%) and more substantially limits the Biology domain performance (from 79.0% to 68.4%). Similarly, Biology-derived experts enhance task-specific results (from 91.9% to 93.0% in Physics) but degrade average performance across other domains, indicating further that specialized expert selections may negatively impact general cognitive reasoning by reinforcing overly specialized activations.

No evidence of harmful side-effects on other domains. Our experimental findings clearly confirm that cognitive experts, either chosen from aggregated cross-domain data or specific domains, constitute effective cognitive modulators that enhance model reasoning accuracy and efficiency. General-purpose expert adjustments deliver robust cross-domain improvements, demonstrating their fundamental importance to reasoning processes regardless of subject matter. Meanwhile, domain-specialized expert modulation illustrates substantial potential for targeted cognitive improvements, particularly within closely related scientific domains. Together, these insights validate our proposed approach as versatile, effective, and immediately deployable for enhancing efficiency, accuracy, and overall reasoning proficiency of existing MoE-based large reasoning models.

3.5 Impact of Reinforced Cognitive Experts on General Capabilities

To address RQ4, we investigate whether reinforcing cognitive experts negatively impacts the model's general capabilities, such as instruction-following. To this end, we evaluate reinforced models on the ArenaHard benchmark [27] to assess potential adverse impacts on general capabilities. The ArenaHard benchmark, designed to evaluate instruction-following capabilities, comprises 500 challenging user queries spanning diverse scenarios. We randomly select 50 user queries as the test data and employ GPT-4-Turbo to judge pairwise comparisons of outputs against the GPT-4-0613 baseline.

Reinforcing cognitive experts maintains or slightly improves general instruction-following capabilities. Our experimental evaluation on the ArenaHard benchmark demonstrates that reinforcing the identified cognitive experts does not adversely impact the model's capability to handle general, challenging instruction-following tasks. As shown in Table 6, models modulated by cognitive experts derived from each domain consistently maintain or marginally improve upon the baseline DeepSeek-R1 accuracy of 91%. Specifically, the domain-specific cognitive experts from Chemistry and Biology show notable accuracy enhancements (from 91.0% to 94.0% in Chemistry; from 91.0% to 93.0% in Biology), underscoring the potential for positive transfer of reasoning-rich expert reinforcement to general-purpose capabilities. Moreover, the general experts ("All" domain) also marginally improve performance (to 92.0%), confirming that cognitive expert-control has a neutral-to-beneficial impact on general instruction-following capabilities.

Table 6: Effect of reinforced cognitive experts of Deepseek-R1 on ArenaHard.

Method	Accuracy	Token
Vanilla	91.0	2,919
Reinforce Experts from different domains		
Math	92.0	2,933
Physics	92.0	2,933
Chemistry	94.0	3,332
Biology	93.0	3,072
All	92.0	2,933

Modulation of cognitive experts results in moderately increased verbosity. An analysis of token counts further reveals that cognitive expert modulation moderately increases model verbosity in response generation, suggesting enhanced cognitive thoroughness. For example, Chemistry and Biology models increase average token counts notably (from 2,919 to 3,332 tokens and from 2,919 to 3,072 tokens, respectively), highlighting that the activation of certain domain-specific cognitive experts may favor more detailed deliberations. Nevertheless, the overall increase in verbosity is moderate, indicating a desirable balance between detail-oriented reasoning and response conciseness.

Overall, reinforcing cognitive experts does not hinder but rather supports general capabilities. These findings collectively confirm our approach as effective and safe for targeted, lightweight interventions. Reinforcing cognitive experts significantly enhances model performance within their

original domains and has either neutral or positive effects on general-purpose instruction-following benchmarks. The moderate increase in verbosity indicates richer, more thoughtful reasoning, aligning with the intended goal of encouraging deeper cognitive processing without sacrificing practicality. This highlights the practicality and versatility of our approach in improving existing MoE model reasoning efficacy and general cognitive capabilities through strategic expert modulation.

3.6 Comparison with Other Methods

We compare our cognitive experts against two prevalent inference-time methods for reasoning tasks: prompt engineering and decoding constraints. Specifically, we analyze two prompt configurations: placing the prompt before the `<think>` token ($\text{Prompt}_{\text{before}}$) and after `<think>` token ($\text{Prompt}_{\text{after}}$), with details outlined in Appendix B.1. For decoding constraints, we adopt a strategy similar to TIP from Wang et al. [4], which curtail the generation of alternative solutions to foster coherent and focused reasoning. In our work, we penalize the think mark tokens (`</think>`) rather than the thought switching tokens (e.g., "alternatively"), and we name the method as TIP_t.

Table 7 compares our cognitive expert modulation method against prompting (both before and after the `<think>` token) and decoding constraints (TIP) on the challenging AIME benchmarks. Our approach achieves the highest average improvement, boosting DeepSeek-R1's performance from a baseline of 68.3% to 78.7%, surpassing the best-performing baseline TIP_t by 2.0% in average accuracy. Notably, our method achieves particularly pronounced gains on AIME24 (83.3%) compared to $\text{Prompt}_{\text{before}}$ (76.7%) and TIP_t (80.0%), indicating a superior ability to enhance cognitive effort and deep reasoning through targeted modulation of expert routing weights. These results validate our hypothesis that strategically steering cognitive experts significantly reinforces appropriate cognitive behavior and enhances reasoning accuracy more effectively than generic prompting and decoding constraints.

Table 7: Comparison with other steering methods on AIME24 and AIME25.

Method	AIME24	AIME25	Avg.
DeepSeek-R1	73.3	63.3	68.3
$\text{Prompt}_{\text{before}}$	76.7	73.3	75.0
$\text{Prompt}_{\text{after}}$	76.7	66.7	71.7
TIP_t	80.0	73.3	76.7
RICE	83.3	73.3	**78.7**

Table 8: An example of our RICE corrects wrong reasoning (in red) generated by Deepseek-R1 into correct reasoning (in green).

Problem	Let ABCDEF be a convex equilateral hexagon in... and EF has side lengths 200, 240, and 300. Find the side length of the hexagon.
DeepSeek-R1	Okay, so I need to find the side length...the triangle formed by extending AB, CD, and EF has sides 200, 240, and 300. Let me sketch this mentally...**Alternatively**, maybe we can model the hexagon as being formed by three pairs of parallel sides...**Alternatively**, think of the hexagon as a combination of three parallelograms...**Alternatively**, using the concept of similar triangles...**Alternatively**, maybe each pair of opposite sides of the hexagon is parallel to a different side of the triangle...**I will posit that the answer is 80**...Therefore, **the answer is 75/2**...Let me **verify** this result...Thus, s = 1200 / 23. Wait, this **contradicts** our previous result...**I think the answer is 80**...based on the thorough step-by-step analysis, $\boxed{1200/23}$ is the conclusion.
+RICE	Okay, so I need to find the side length...Let me draw a rough sketch in my mind...**So**, the triangle is such that each side is aligned with one of the extended sides of the hexagon...But maybe using **coordinate geometry could help here**....Not sure, maybe this is **getting too complicated**...**Alternatively**, maybe using ratios...**Alternatively**, after struggling through all these approaches, perhaps the answer is related to the harmonic mean of the triangle's sides...Therefore, I think **the answer is 80. But need to verify**...Therefore, the side length of the hexagon is $\boxed{80}$.

4 Related Work

Large Reasoning Models Large Reasoning Models (LRMs) significantly enhances the reasoning capabilities of large language models (LLMs) [5, 28]. Prominent implementations include OpenAI's

o1 [5], QwQ [29], Qwen3 [2], DeepSeek-R1 [1], Claude 3.7 [30] and Kimi-1.5 [31] achieve human-like reasoning by leveraging scaled test-time computation. In particular, the open-source DeepSeek-R1 utilizes a Mixture-of-Experts (MoE) architecture [32] with sparsely activated parameters, selectively activating only 8 out of 256 experts per layer [21]. This MoE architecture has been widely adopted in recent LLMs [33–36], achieving an optimal balance between computational efficiency and competitive performance in complex reasoning tasks.

MoE Models Previous research on Mixture of Experts (MoE) models indicates that expert routing is primarily token-dependent [22]. However, Olson et al. [23] demonstrate that DeepSeek-R1's advanced reasoning capabilities enable its routing mechanism to achieve greater semantic specialization and structured cognitive processing, representing a substantial advancement over prior MoE models. Subsequently, Hazra et al. [37] train sparse autoencoders (SAEs) on DeepSeek-R1, identifying interpretable features such as backtracking, division, and rapid response patterns within the SAEs space. However, training SAEs is computationally intensive, posing significant resource demands. We employ the normalized Pointwise Mutual Information (nPMI) metric to evaluate expert specialization, requiring only a single forward propagation.

Efficient Thinking Despite significant advancements, o1-like models continue to encounter substantial cognitive challenges, such as the overthinking [3, 38–40] and underthinking phenomenon [4, 11, 41]. Subsequent efforts address these issues through rule-based stop, decoding constraints [42, 4, 43–49], steering vectors [50–53], and parameters tuning [54, 3, 55–57]. There are also some works specifically designed to improve reasoning capabilities in MoE architectures by re-mixing experts through gradient-based optimization [58] or by expert pruning via sparse dictionary learning [59]. However, the resource-intensive nature of expert re-mixing algorithms makes them impractical to scale to large models such as 600B-parameter systems, whereas our method is lightweight and directly applicable to such large-scale settings. Generally, in contrast to the above strategies that primarily rely on crafted rules, extensive labeled data, or computationally expensive parameter training, our *reinforcing cognitive experts* approach achieves more efficient and deeper reasoning with only a single forward pass, without requiring any supervision signals or additional training.

5 Conclusion and Future Work

In this work, we investigate cognitive experts in MoE-based language models and propose an efficient nPMI-based method to identify those most relevant to reasoning. We show that steering these experts enables control over the model's reasoning with minimal computational overhead. Notably, these experts exhibit strong transferability across scientific domains, suggesting a generalizable cognitive function. Future directions include deeper investigations into the structural properties and broader applicability of cognitive experts, as well as integration with other cognitive control strategies to further enhance reasoning robustness. By uncovering this hidden layer of functional specialization within MoE models, we may open new avenues for fine-grained control over neural reasoning processes, more closely mirroring the modularity observed in biological cognitive systems.

6 Limitations and Broader Impacts

The internal coordination mechanisms of long-range reasoning models are inherently complex, and our nPMI-based approach may not fully capture all relevant interactions. Future work should explore more sophisticated metrics for expert identification. Besides, our validation was constrained by the current availability of open-source MoE architectures designed for long-range reasoning, limited to DeepSeek-R1 [1] and Qwen3-235B [2]. Additional testing across more diverse architectures is warranted. The ability to precisely control reasoning processes in large language models has significant implications for both AI safety and efficiency. Our method's minimal computational overhead makes it particularly promising for real-world applications where resource constraints are critical. The observed cross-domain transferability of cognitive experts suggests exciting possibilities for developing more general and adaptable AI systems.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62576307), Ningbo Natural Science Foundation (2024J020), Yongjiang Talent Introduction Programme (2021A156-G), Tencent AI Lab Rhino-Bird Focused Research Program (RBFR2024003), and Information Technology Center and State Key Lab of CAD&CG, Zhejiang University.

References

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

[2] Qwen Team. Qwen3: Think deeper, act faster. 2025. URL https://qwenlm.github.io/zh/blog/qwen3/.

[3] Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Do NOT think that much for 2+3=? on the overthinking of o1-like llms. *CoRR*, abs/2412.21187, 2024. doi: 10.48550/ARXIV.2412.21187. URL https://doi.org/10.48550/arXiv.2412.21187.

[4] Yue Wang, Qiuzhi Liu, Jiahao Xu, Tian Liang, Xingyu Chen, Zhiwei He, Linfeng Song, Dian Yu, Juntao Li, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Thoughts are all over the place: On the underthinking of o1-like llms. *CoRR*, abs/2501.18585, 2025. doi: 10.48550/ARXIV.2501.18585. URL https://doi.org/10.48550/arXiv.2501.18585.

[5] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[6] Yixin Ji, Juntao Li, Hai Ye, Kaixin Wu, Jia Xu, Linjian Mo, and Min Zhang. Test-time computing: from system-1 thinking to system-2 thinking. *arXiv preprint arXiv:2501.02497*, 2025.

[7] Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Zhihan Guo, Yufei Wang, Irwin King, Xue Liu, and Chen Ma. What, how, where, and how well? a survey on test-time scaling in large language models. *arXiv preprint arXiv:2503.24235*, 2025.

[8] Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Hanjie Chen, et al. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*, 2025.

[9] Sicheng Feng, Gongfan Fang, Xinyin Ma, and Xinchao Wang. Efficient reasoning models: A survey. *arXiv preprint arXiv:2504.10903*, 2025.

[10] Qianjun Pan, Wenkai Ji, Yuyang Ding, Junsong Li, Shilian Chen, Junyi Wang, Jie Zhou, Qin Chen, Min Zhang, Yulan Wu, et al. A survey of slow thinking-based reasoning llms using reinforced learning and inference-time scaling law. *arXiv preprint arXiv:2505.02665*, 2025.

[11] Xiaoye Qu, Yafu Li, Zhaochen Su, Weigao Sun, Jianhao Yan, Dongrui Liu, Ganqu Cui, Daizong Liu, Shuxian Liang, Junxian He, et al. A survey of efficient reasoning for large reasoning models: Language, multimodality, and beyond. *arXiv preprint arXiv:2503.21614*, 2025.

[12] Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025.

[13] Rui Wang, Hongru Wang, Boyang Xue, Jianhui Pang, Shudong Liu, Yi Chen, Jiahao Qiu, Derek Fai Wong, Heng Ji, and Kam-Fai Wong. Harnessing the reasoning economy: A survey of efficient reasoning for large language models. *arXiv preprint arXiv:2503.24377*, 2025.

[14] Tong Wu, Chong Xiang, Jiachen T Wang, and Prateek Mittal. Effectively controlling reasoning models through thinking intervention. *arXiv preprint arXiv:2503.24370*, 2025.

[15] Ximing Lu, Seungju Han, David Acuna, Hyunwoo Kim, Jaehun Jung, Shrimai Prabhumoye, Niklas Muennighoff, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, et al. Retro-search: Exploring untaken paths for deeper and efficient reasoning. *arXiv preprint arXiv:2504.04383*, 2025.

[16] Robert MG Reinhart and John A Nguyen. Working memory revived in older adults by synchronizing rhythmic brain circuits. *Nature neuroscience*, 22(5):820–827, 2019.

[17] Miles Wischnewski, Ivan Alekseichuk, and Alexander Opitz. Neurocognitive, physiological, and biophysical effects of transcranial alternating current stimulation. *Trends in Cognitive Sciences*, 27(2):189–205, 2023.

[18] Desmond J Oathes, Romain JP Duprat, Justin Reber, Ximo Liang, Morgan Scully, Hannah Long, Joseph A Deluisi, Yvette I Sheline, and Kristin A Linn. Non-invasively targeting, probing and modulating a deep brain circuit for depression alleviation. *Nature Mental Health*, 1(12): 1033–1042, 2023.

[19] Shrey Grover, John A Nguyen, Vighnesh Viswanathan, and Robert MG Reinhart. High-frequency neuromodulation improves obsessive–compulsive behavior. *Nature medicine*, 27(2): 232–238, 2021.

[20] Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40, 2009.

[21] DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, and Wangding Zeng. Deepseek-v3 technical report. *CoRR*, abs/2412.19437, 2024. URL https://doi.org/10.48550/arXiv.2412.19437.

[22] Fuzhao Xue, Zian Zheng, Yao Fu, Jinjie Ni, Zangwei Zheng, Wangchunshu Zhou, and Yang You. Openmoe: An early effort on open mixture-of-experts language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=1YDeZU8Lt5.

[23] Matthew Lyle Olson, Neale Ratzlaff, Musashi Hinck, Man Luo, Sungduk Yu, Chendi Xue, and Vasudev Lal. Semantic specialization in moe appears with scale: A study of deepseek r1 expert specialization. *arXiv preprint arXiv:2502.10928*, 2025.

[24] Mohsen Fayyaz, Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Ryan Rossi, Trung Bui, Hinrich Schütze, and Nanyun Peng. Steering moe llms via expert (de) activation. *arXiv preprint arXiv:2509.09660*, 2025.

[25] MAA Committees. Aime problems and solutions. https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions.

[26] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

[27] Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *CoRR*, abs/2406.11939, 2024. URL https://doi.org/10.48550/arXiv.2406.11939.

[28] Shijie Xia, Yiwei Qin, Xuefeng Li, Yan Ma, Run-Ze Fan, Steffi Chern, Haoyang Zou, Fan Zhou, Xiangkun Hu, Jiahe Jin, et al. Generative ai act ii: Test time scaling drives cognition engineering. *arXiv preprint arXiv:2504.13828*, 2025.

[29] Qwen. Qwq: Reflect deeply on the boundaries of the unknown. 2024. URL https://qwenlm.github.io/blog/qwq-32b-preview/.

[30] Anthropic. Claude 3.7 sonnet. 2025. URL https://www.anthropic.com/claude/sonnet.

[31] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

[32] Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 1280–1297. Association for Computational Linguistics, 2024. URL https://doi.org/10.18653/v1/2024.acl-long.70.

[33] Llama. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. 2025. URL https://www.llama.com/models/llama-4/.

[34] Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A. Smith, Pang Wei Koh, Amanpreet Singh, and Hannaneh Hajishirzi. Olmoe: Open mixture-of-experts language models. *CoRR*, abs/2409.02060, 2024. doi: 10.48550/ARXIV.2409.02060. URL https://doi.org/10.48550/arXiv.2409.02060.

[35] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

[36] Mohsen Fayyaz, Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Ryan A. Rossi, Trung Bui, Hinrich Schütze, and Nanyun Peng. Steering moe llms via expert (de)activation. *CoRR*, abs/2509.09660, 2025. doi: 10.48550/ARXIV.2509.09660. URL https://doi.org/10.48550/arXiv.2509.09660.

[37] Dron Hazra, Max Loeffler, Murat Cubuktepe, Levon Avagyan, Liv Gorton, Mark Bissell, Owen Lewis, Thomas McGrath, and Daniel Balsam. Under the hood of a reasoning model. 2025. URL https://www.goodfire.ai/blog/under-the-hood-of-a-reasoning-model.

[38] Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Hanjie Chen, and Xia Ben Hu. Stop overthinking: A survey on efficient reasoning for large language models. *CoRR*, abs/2503.16419, 2025. doi: 10.48550/ARXIV.2503.16419. URL https://doi.org/10.48550/arXiv.2503.16419.

[39] Alejandro Cuadron, Dacheng Li, Wenjie Ma, Xingyao Wang, Yichuan Wang, Siyuan Zhuang, Shu Liu, Luis Gaspar Schroeder, Tian Xia, Huanzhi Mao, Nicholas Thumiger, Aditya Desai, Ion Stoica, Ana Klimovic, Graham Neubig, and Joseph E. Gonzalez. The danger of overthinking: Examining the reasoning-action dilemma in agentic tasks. *CoRR*, abs/2502.08235, 2025. doi: 10.48550/ARXIV.2502.08235. URL https://doi.org/10.48550/arXiv.2502.08235.

[40] Wojciech Zaremba, Evgenia Nitishinskaya, Boaz Barak, Stephanie Lin, Sam Toyer, Yaodong Yu, Rachel Dias, Eric Wallace, Kai Xiao, Johannes Heidecke, et al. Trading inference-time compute for adversarial robustness. *arXiv preprint arXiv:2501.18841*, 2025.

[41] Marthe Ballon, Andres Algaba, and Vincent Ginis. The relationship between reasoning and performance in large language models–o3 (mini) thinks harder, not longer. *arXiv preprint arXiv:2502.15631*, 2025.

[42] Bao Hieu Tran, Nguyen Cong Dat, Nguyen Duc Anh, and Hoang Thanh-Tung. Learning to stop overthinking at test time. *CoRR*, abs/2502.10954, 2025. doi: 10.48550/ARXIV.2502.10954. URL https://doi.org/10.48550/arXiv.2502.10954.

[43] Yifu Ding, Wentao Jiang, Shunyu Liu, Yongcheng Jing, Jinyang Guo, Yingjie Wang, Jing Zhang, Zengmao Wang, Ziwei Liu, Bo Du, Xianglong Liu, and Dacheng Tao. Dynamic parallel tree search for efficient LLM reasoning. *CoRR*, abs/2502.16235, 2025. doi: 10.48550/ARXIV.2502.16235. URL https://doi.org/10.48550/arXiv.2502.16235.

[44] Kun Xiang, Zhili Liu, Zihao Jiang, Yunshuang Nie, Kaixin Cai, Yiyang Yin, Runhui Huang, Haoxiang Fan, Hanhui Li, Weiran Huang, et al. Can atomic step decomposition enhance the self-structured reasoning of multimodal large models? *arXiv preprint arXiv:2503.06252*, 2025.

[45] Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. Reasoning models can be effective without thinking. *arXiv preprint arXiv:2504.09858*, 2025.

[46] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel J. Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *CoRR*, abs/2501.19393, 2025. doi: 10.48550/ARXIV.2501.19393. URL https://doi.org/10.48550/arXiv.2501.19393.

[47] Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. Token-budget-aware LLM reasoning. *CoRR*, abs/2412.18547, 2024. doi: 10.48550/ARXIV.2412.18547. URL https://doi.org/10.48550/arXiv.2412.18547.

[48] Simon A. Aytes, Jinheon Baek, and Sung Ju Hwang. Sketch-of-thought: Efficient LLM reasoning with adaptive cognitive-inspired sketching. *CoRR*, abs/2503.05179, 2025. doi: 10.48550/ARXIV.2503.05179. URL https://doi.org/10.48550/arXiv.2503.05179.

[49] Jintian Zhang, Yuqi Zhu, Mengshu Sun, Yujie Luo, Shuofei Qiao, Lun Du, Da Zheng, Huajun Chen, and Ningyu Zhang. Lightthinker: Thinking step-by-step compression. *arXiv preprint arXiv:2502.15589*, 2025.

[50] Runjin Chen, Zhenyu Zhang, Junyuan Hong, Souvik Kundu, and Zhangyang Wang. Seal: Steerable reasoning calibration of large language models for free. *arXiv preprint arXiv:2504.07986*, 2025.

[51] Hannah Cyberey and David Evans. Steering the censorship: Uncovering representation vectors for llm" thought" control. *arXiv preprint arXiv:2504.17130*, 2025.

[52] Runjin Chen, Zhenyu Zhang, Junyuan Hong, Souvik Kundu, and Zhangyang Wang. SEAL: steerable reasoning calibration of large language models for free. *CoRR*, abs/2504.07986, 2025. doi: 10.48550/ARXIV.2504.07986. URL https://doi.org/10.48550/arXiv.2504.07986.

[53] Hannah Cyberey and David Evans. Steering the censorship: Uncovering representation vectors for LLM "thought" control. *CoRR*, abs/2504.17130, 2025. doi: 10.48550/ARXIV.2504.17130. URL https://doi.org/10.48550/arXiv.2504.17130.

[54] Chung-En Sun, Ge Yan, and Tsui-Wei Weng. Thinkedit: Interpretable weight editing to mitigate overly short thinking in reasoning models. *arXiv preprint arXiv:2503.22048*, 2025.

[55] Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. LIMO: less is more for reasoning. *CoRR*, abs/2502.03387, 2025. doi: 10.48550/ARXIV.2502.03387. URL https://doi.org/10.48550/arXiv.2502.03387.

[56] Pranjal Aggarwal and Sean Welleck. L1: controlling how long A reasoning model thinks with reinforcement learning. *CoRR*, abs/2503.04697, 2025. doi: 10.48550/ARXIV.2503.04697. URL https://doi.org/10.48550/arXiv.2503.04697.

[57] Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *CoRR*, abs/2412.06769, 2024. doi: 10.48550/ARXIV.2412.06769. URL https://doi.org/10.48550/arXiv.2412.06769.

[58] Zhongyang Li, Ziyue Li, and Tianyi Zhou. C3po: Critical-layer, core-expert, collaborative pathway optimization for test-time expert re-mixing. *arXiv preprint arXiv:2504.07964*, 2025.

[59] Yuanbo Tang, Yan Tang, Naifan Zhang, Meixuan Chen, and Yang Li. Unveiling hidden collaboration within mixture-of-experts in large language models. *arXiv preprint arXiv:2504.12359*, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: This paper discusses the limitations of the work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: For each theoretical result, the paper provides the full set of assumptions and a complete (and correct) proof.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental results reported in the paper can be fully reproduced.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We use publicly available datasets, Code and Data are also provided in the supplemental material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: This paper specifies all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: The paper does not include experiments requiring this.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: This paper provides sufficient information on the computer resources needed to reproduce the experiments in §B.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This paper discusses both potential positive societal impacts and negative societal impacts of the work performed.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology in this paper.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

A Renormalization

We investigate the DeepSeek Mixture-of-Experts (MoE) architecture, where each token selects 8 of 256 experts, with weights normalized to sum to 1. We examine steering specific expert weights under two conditions: with and without renormalization. The effects of the steering coefficient (reinforce factor) are presented in Table 9, with generalization performance analyzed in Table 10.

Table 9 evaluates the reinforce factor's effect on two cognitive experts. Without renormalization, accuracy peaks at 83.3% (factors 4, 32, 64, 128) but drops to 3.3% at 2048, with erratic token counts (e.g., 16,836). With renormalization, accuracy remains stable (73.3%–83.3%) across most factors, with token counts varying moderately (8,383–9,508), though it declines to 66.7% at factor 256. Renormalization thus enhances robustness at higher steering coefficients.

We evaluate the generalization performance of cognitive experts, identified using normalized Pointwise Mutual Information (nPMI) within Mixture-of-Experts (MoE)-based large reasoning models, comparing three strategies: Vanilla R1, Renormalized, and Without Renormalized (wo/Renormalized). Table 10 reports performance across AIME25, Physics, Chemistry, Biology, and their average for experts selected from AIME24.

The wo/Renormalized strategy demonstrates superior generalization, achieving an average score of 73.1, compared to 70.9 for Vanilla R1 and 68.8 for Renormalized. This 4.3-point improvement over Renormalized is driven by notable gains in AIME25 (73.3 vs. 63.3) and Biology (79.0 vs. 68.4). In Physics, Vanilla R1 (91.9) outperforms wo/Renormalized (89.5, -2.4), while in Chemistry, Renormalized (52.7) surpasses wo/Renormalized (50.4), indicating domain-specific trade-offs.

Non-renormalization avoids the computational overhead of normalization (e.g., softmax scaling of expert weights), aligning with its reported efficiency. These results confirm that non-renormalization enhances generalization in cognitive experts, offering a computationally lightweight approach to optimizing reasoning in MoE architectures.

Table 9: Reinforce factor effects of two cognitive experts with/without renormalization

Reinforce Factor	**wo/Renormalization**		**Renormalization**	
	Acc	*Token*	*Acc*	*Token*
1 (R1)	73.3	9,291	73.3	9,291
2	70.0	9,103	80.0	8,463
4	83.3	8,145	80.0	8,383
8	73.3	9,502	70.0	8,818
16	80.0	8,493	73.3	9,133
32	83.3	8,337	83.3	8,956
64	83.3	8,317	80.0	9,508
128	83.3	9,490	73.3	9,091
256	60.0	7,986	66.7	8,719
512	46.7	6,270	80.0	8,786
1024	23.3	4,378	73.3	8,564

Table 10: Generalization capacity of two cognitive experts selected from AIME24, with or without renormalization.

Strategy	**AIME25**	**Physics**	**Chemistry**	**Biology**	**Average**
Vanilla R1	63.3	91.9	49.5	79.0	70.9
Renormalized	63.3	90.7	52.7	68.4	68.8
wo/Renormalized	73.3	89.5	50.4	79.0	73.1

Table 11: Pass@k performance of our cognitive experts on Deepseek-R1 and Qwen3-235B-A22B. For each problem, we generated 16 responses with a temperature of 0.6 and a top p value of 0.95.

Model	Strategy	Accuracy		Tokens
		Pass@1	Pass@8	
Deepseek-R1		**AIME24**		
	Vanilla	74.8	88.3	9,219
	Our	76.0	89.2	8,317
		AIME25		
	Vanilla	68.5	84.7	10,875
	Our	67.7	86.3	11,441
Qwen3-235B-A22B		**AIME24**		
	Vanilla	84.0	93.0	10,946
	Our	85.0	91.6	10,706
		AIME25		
	Vanilla	82.7	88.3	12,546
	Our	82.1	89.7	12,373

B Experiment Setup

B.1 Baselines

We evaluate our cognitive experts in comparison with two widely used inference-time techniques for reasoning tasks: prompt engineering. In particular, we consider two types of prompt placements in our analysis — one positioned before the `<think>` token (Promptbefore) and the other placed after it (Promptafter), defined as follows:

Prompt before `<think>`

<|begin_of_sentence|><|User|> <context>
You are an expert math-solving assistant who prioritizes clear, concise solutions. You solve problems in a single thought process, ensuring accuracy and efficiency. You seek clarification when needed and respect user preferences even if they are unconventional.
</context>

<solving_rules>
- Try to complete every idea you think of and don't give up halfway
- Don't skip steps
- Display solution process clearly
- Ask for clarification on ambiguity
</solving_rules>

<format_rules>
- Use equations and explanations for clarity
- Keep responses brief but complete
- Provide step-by-step reasoning if needed
</format_rules>

PROBLEM: {problem}

OUTPUT: Please think carefully and follow above rules to get the correct answer for PROBLEM. Focus on clear, concise solutions while maintaining a helpful, accurate style.<|Assistant|> <think> \n

Table 12: Performance of domain-specific steering multipliers across scientific domains. The math domain is evaluated using the AIME25 benchmark.

Model	Math	Physics	Chemistry	Biology	Average
DeepSeek-R1	63.3	91.9	49.5	79.0	70.9
DeepSeek-R1 + RICE	**73.3**	**93.0**	**54.8**	79.0	**75.0**
Qwen-235B	66.7	90.7	49.5	78.9	71.5
Qwen-235B + RICE	**73.3**	**95.3**	49.5	**84.2**	**75.6**

Prompt after `<think>`

<|begin_of_sentence|><|User|> <context>
You are an expert math-solving assistant who prioritizes clear, concise solutions. You solve problems in a single thought process, ensuring accuracy and efficiency. You seek clarification when needed and respect user preferences even if they are unconventional.
</context>

PROBLEM: {problem}

<think> \n

Please think carefully and follow these rules to find the correct answer for PROBLEM.

<solving_rules>
- Try to complete every idea you think of and don't give up halfway
- Don't skip steps
- Display solution process clearly
- Ask for clarification on ambiguity
</solving_rules>

<format_rules>
- Use equations and explanations for clarity
- Keep responses brief but complete
- Provide step-by-step reasoning if needed
</format_rules>

Focus on clear, concise solutions while maintaining a helpful and accurate style.

OUTPUT:

B.2 Experiments Compute Resources

We conduct our DeepSeek-R1 experiments on 16 H20 GPUs using `vllm==0.7.0`. It is worth noting that for experiments on the Qwen3-235B-A22B model, we use `vllm==0.8.5.post` because the recently released Qwen3-235B-A22B models are only compatible with `vllm` versions $\geq$ 0.8.5.

C Experiment Details and Results

C.1 Steering Multiplier

We use a simple, domain-specific steering multiplier (selected from a small set of 16, 32, 64), RICE delivers consistent and significant improvements across all domains.

Table 13: Performance comparison of Deepseek-R1 with and without RICE across different datasets.

Dataset	Model	Accuracy	#Token
GSM8K	Deepseek-R1	95.9	1028
	Deepseek-R1 + RICE (Ours)	**96.0**	**1001**
MATH-500	Deepseek-R1	95.0	3282
	Deepseek-R1 + RICE (Ours)	**96.4**	**3204**
HLE	Deepseek-R1	4.0	9433
	Deepseek-R1 + RICE (Ours)	**6.0**	9445

C.2 Effects on Other Datasets

We delve into the effect of these two experts identified by AIME24 on three additional diverse benchmarks: GSM8K (grade-school math), MATH (competition math), and HLE (Humanity's Last Exam, covering social science, CS, etc.). Note that we randomly sampled 100 text instances as the test set due to resource constraints. As shown in Table 13, RICE consistently improves Deepseek-R1 across all datasets, providing modest gains on high-performing tasks (GSM8K, MATH-500) and larger relative improvements on more challenging tasks (HLE).

Moreover, we compare the differences with and without RICE. Specifically, we focus on the token distribution during model decoding. We observe that tokens related to "think," "best," "good," and similar concepts are ranked higher (positioned closer to the top 1) during decoding after expert reweighting.

C.3 Cognitive Experts of Qwen3-235B

As a case study in math, we employ Qwen3-235B to generate responses on the AIME2024 dataset, while recording the expert assignments at each token during the forward pass. Subsequently, we apply the nPMI measure defined in Eq. 6 to identify the top five experts that exhibit the highest statistical dependence on reasoning-related indicators, such as the "`<think>`" token. These selected experts are thus regarded as the core cognitive components specialized in mathematical reasoning. Due to computational constraints, our quantitative analysis in Table 2 focuses specifically on math-domain experts. This focused approach allows for deeper investigation of expert specialization patterns while maintaining feasible resource requirements.

C.4 Pass@k Performance of Cognitive Experts

Table 11 presents the Pass@k performance of our cognitive expert modulation approach compared to vanilla baselines across two model architectures. On DeepSeek-R1, our method demonstrates consistent improvements in Pass@8 accuracy (+0.9% on AIME24 and +1.6% on AIME25) despite showing marginal variations in Pass@1 performance. Notably, we observe a 9.8% reduction in token consumption for AIME24 while maintaining superior accuracy, suggesting improved reasoning efficiency. For Qwen3-235B-A22B, our approach achieves higher Pass@1 accuracy (+1.0% on AIME24) while showing competitive Pass@8 performance ($\pm$1.4% across datasets), with consistent reductions in computational cost (2.2% fewer tokens on AIME24 and 1.4% fewer on AIME25). The observed trade-offs between Pass@1 and Pass@8 metrics suggest that our method enhances *reliable* reasoning (as reflected in Pass@8) more than *peak* performance (Pass@1), particularly in the more challenging AIME25 benchmark. These results substantiate our hypothesis that targeted expert modulation can improve reasoning efficiency without compromising solution quality.