DECOR: DECOUPLED COLLABORATIVE REASONING VIA PREFERENCE-GUIDED PRUNING AND DUAL-INDICATOR GATING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) with long Chain-of-Thought (CoT) training demonstrate exceptional reasoning capabilities, yet reliance on excessive reasoning tokens degrades both performance and efficiency. Existing methods address this through specialized retraining or external constraints. While effective, these methods still retain a single-stream reasoning paradigm, leaving exploration and termination entangled within the same trajectory. To overcome this limitation, we propose DeCoR (Decoupled Collaborative Reasoning), a novel framework for efficient reasoning that decouples exploration and termination through twin-model interaction. We first devise a preference-guided method to achieve structured layer pruning, which derives a functional auxiliary model that serves as an explorer under high predictive uncertainty. This design disentangles distinct reasoning roles while preserving a shared state and trajectory coherence. To coordinate these roles according to the evolving reasoning state, we further introduce a dual-indicator gating mechanism that monitors uncertainty and consistency of intermediate answers. It dynamically alternates between the base model for correction or termination and the auxiliary model for continuous exploration. Extensive experiments on five reasoning benchmarks demonstrate that our method significantly reduces CoT length by 32.5% on average and improves final accuracy by 2.8 to 6.5 percentage points. Our method also increases the density of effective reasoning information within concise outputs. The code will be made publicly available.

1 INTRODUCTION

Advanced reasoning capabilities in large language models (LLMs) have been substantially elicited through long Chain-of-Thought (CoT) training (OpenAI, 2024; DeepSeek-AI, 2025; Chu et al., 2026; Wei & Kim, 2026). By enabling deliberation on various possibilities before generating the final answer, these models enhance problem-solving for complex tasks. However, the extension of reasoning sequences often induces overthinking: even when the reasoning process is converging toward a stable answer, LLMs may still vacillate among similar candidates or drift away from the correct trajectory (Qu et al., 2025; Sui et al., 2025; Ghosal et al., 2025; Han et al., 2026). Given the quadratic complexity of the attention mechanism (Vaswani et al., 2017), each redundant step compounds resource consumption during inference. Therefore, the key challenge is that signals of trajectory stabilization or drift emerge dynamically, yet reasoning fails to elicit the appropriate behavioral response in the concurrent presence of these signals.

Current studies address this challenge from two perspectives. Intrinsic approaches employ specialized fine-tuning to internalize a preference for brevity (Ma et al., 2025b; Kimi Team, 2025; Li et al., 2025a; Deng et al., 2026; Li et al., 2026), whereas extrinsic methods constrain the decoding process through elaborate prompts, budgets, or semantic truncation (Chuang et al., 2025; Aytes et al., 2025; Lin et al., 2026; Xiang et al., 2026). Despite their effectiveness, both retain a single-stream control structure: exploration and termination are governed within the same trajectory, while their uncertainty distributions exhibit substantial overlap (see Figures 4, 8, and 9). Consequently, the evolving reasoning state does not correspond to a unique action, leading to either deficient or redundant reasoning. This raises a pivotal question: *Can reasoning be self-regulated without forcing exploration and termination to compete within the same trajectory, allowing the model to execute decisive ac-*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

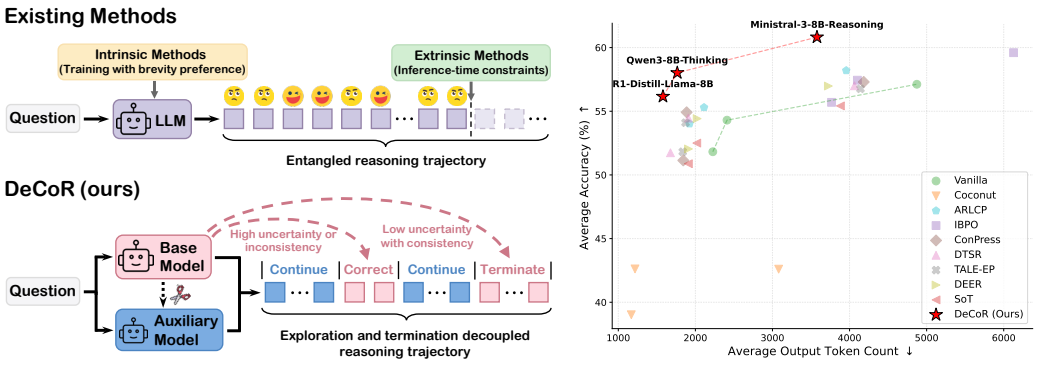


Figure 1: **(Left)** Architectural comparison between existing overthinking mitigation methods and DeCoR. Existing methods still suffer from the inherent conflict between continuing and stopping generation within a single reasoning trajectory, while DeCoR decouples these two competing objectives upon a shared model foundation. **(Right)** Results of different efficient reasoning methods based on three base models across five benchmarks. DeCoR strikes the optimal trade-off between accuracy and output length across all three base models.

tions as its internal state evolves? A high-level analogy stems from human cortical computation, in which functionally differentiated processes operate on shared representations and interact via dynamic signals (Menon, 2011; Bullmore & Sporns, 2012; van den Heuvel & Sporns, 2013). This suggests that exploration and termination during reasoning can be assigned to distinct roles, whose interaction is coordinated by evolving evidence.

Motivated by these insights, we propose DeCoR, an efficient reasoning framework that decouples exploration and termination into complementary roles through twin-model interaction, as shown in Figure 1(left). We introduce **preference-guided structured pruning** to extract an auxiliary model from the base model through a lightweight one-time calibration. This calibration scores layer importance using layer-wise shifts in next-token predictive distributions. The auxiliary model is derived via in-place masking of pruned layers within the base model, delineating two distinct reasoning roles while preserving coherence. During exploration, it operates in a high-uncertainty regime and continuously generates reasoning segments. Then, we develop **dual-indicator gating** to regulate reasoning by leveraging intermediate answers produced by the base model conditioned on these segments. One indicator quantifies the predictive uncertainty of each answer, with persistent high uncertainty triggering correction by the base model and candidates with sufficiently low uncertainty archived in a storage window. The other verifies convergence among stored answers, triggering correction upon inconsistency or termination once consistency is established. This mechanism enables dynamic correction and termination of the reasoning trajectory.

As shown in Figure 1(right), DeCoR strikes an optimal trade-off between accuracy and output length across three base models on five benchmarks (detailed in Section 4.2). We jointly analyze variations in predictive uncertainty and behavioral transitions to substantiate the capacity of DeCoR to decouple reasoning, while further demonstrating that it increases the proportion of effective tokens within the shortened CoT sequences (detailed in Section 4.3). In summary, our contributions are:

- We introduce DeCoR, an efficient reasoning framework that leverages twin-model interaction to reconcile the conflicting roles of exploration and termination.
- We propose preference-guided structured pruning to quantify layer influence on the next-token distribution and evaluate precise layer-wise importance, which safeguards the construction of a reliable auxiliary explorer for effective coordination with the base model.
- We develop dual-indicator gating that orchestrates the interaction between models to generate intermediate answers, leveraging answer uncertainty and consistency to dynamically regulate continuation, correction, and termination during reasoning.
- We empirically demonstrate that DeCoR significantly reduces redundant outputs and improves accuracy compared to established baselines. Moreover, it achieves a higher density of effective thinking within a compact CoT.

2 Related Work

Overthinking Alleviation. A growing body of work has emerged to alleviate overthinking in LLMs and optimize their inference efficiency. Some methods focus on refining internal mechanisms through specialized fine-tuning. O1-Pruner (Luo et al., 2026) and L1 (Aggarwal & Welleck, 2025) incorporate length regularization terms for long CoT sequences, while ConPress (Deng et al., 2026) induces concise reasoning traces via multi-question fine-tuning. Novel training objectives are proposed to empower LLMs with the capability to compress excessive CoT tokens into compact representations (Hao et al., 2025; Shen et al., 2025; Li et al., 2026). Furthermore, IBPO (Yu et al., 2025) and JET (Han et al., 2026) enable models to dynamically allocate reasoning budgets tailored to the complexity of individual tasks. Another category of methods applies elaborate human-designed constraints during the inference stage. TALE-EP (Han et al., 2025) and NoThinking (Ma et al., 2025a) utilize exquisite prompts to guide models toward concise reasoning. Early exit and termination strategies are implemented by monitoring semantic cues to prune reasoning paths (Lin et al., 2026; Yang et al., 2026; Fu et al., 2025). However, these methods regulate the length of a single reasoning stream without explicitly separating the roles of continued exploration and termination as the reasoning state evolves.

Structured Pruning for Reasoning LLMs. Structured pruning has long been recognized as a pivotal technique for improving general LLMs inference efficiency, which is achieved by removing structural components such as attention heads, MLP channels, and entire layers (Ma et al., 2023; Men et al., 2025; Li et al., 2025b; Huang et al., 2026; Wang et al., 2026). This approach eliminates the need for specialized hardware and operators to accelerate data transmission. However, conventional structured pruning methods often trigger a performance collapse in the logical consistency of reasoning LLMs (Ding et al., 2026), and they are typically designed to replace the original model with a more compact one, treating structure compression as the endpoint of pruning. In contrast, DeCoR utilizes pruning to instantiate an auxiliary pathway dedicated to exploration within the same base model, aiming to establish distinct reasoning roles while preserving trajectory coherence. This internal layer skipping conceptually intersects with self-speculative decoding methods such as Draft & Verify and SWIFT (Zhang et al., 2024; Xia et al., 2025), which similarly bypass layers within the target model. Unlike these methods that aim to accelerate decoding while preserving the output distribution of the target model, DeCoR actively regulates the reasoning trajectory through exploration, correction, and termination to mitigate overthinking (see Appendix I for a comparative evaluation).

3 Method

We introduce DeCoR, a framework leveraging twin-model collaboration grounded in preference-guided layer pruning and dual-indicator gating to disentangle the role conflict in the reasoning trajectory, as illustrated in Figure 2. The pseudocode of DeCoR is provided in Appendix A.

3.1 Definition of Predictive Preference

The base model performs offline inference on a pre-defined calibration dataset, where the hidden states of each attention and MLP layer are mapped onto the vocabulary space via the output head projection to derive intermediate logits. The entire process involves no extra training. Formally, denoting the output head projection as f_{head}, the aforementioned process is formulated as

$$g_l^{q,x} = \text{Softmax}\left(f_{\text{head}}\left(\text{Norm}\left(h_l^{q,x}\right)\right)\right),\ l = 1, 2, \cdots, \mathcal{L}, \tag{1}$$

where $\mathcal{L}$ denotes the total number of Transformer blocks, and $h_l^{q,x}$ refers to the hidden state output from the attention or MLP layer of the l-th block during the generation of token x for prompt q. Consistent with the standard processing pipeline, we normalize $h_l^{q,x}$ before projection. $g_l^{q,x}$ is the probability distribution obtained by applying a standard softmax function to the projected representation, which characterizes the layer-wise preference for the subsequent token prediction. By comparing the output probability distributions of the current layer and its predecessor, we evaluate the extent to which the current layer changes the predicted preference, which is quantified by the KL divergence as follows:

$$\theta_l^{q,x} = D_{KL}\left(g_{l-1}^{q,x} \parallel g_l^{q,x}\right), \tag{2}$$

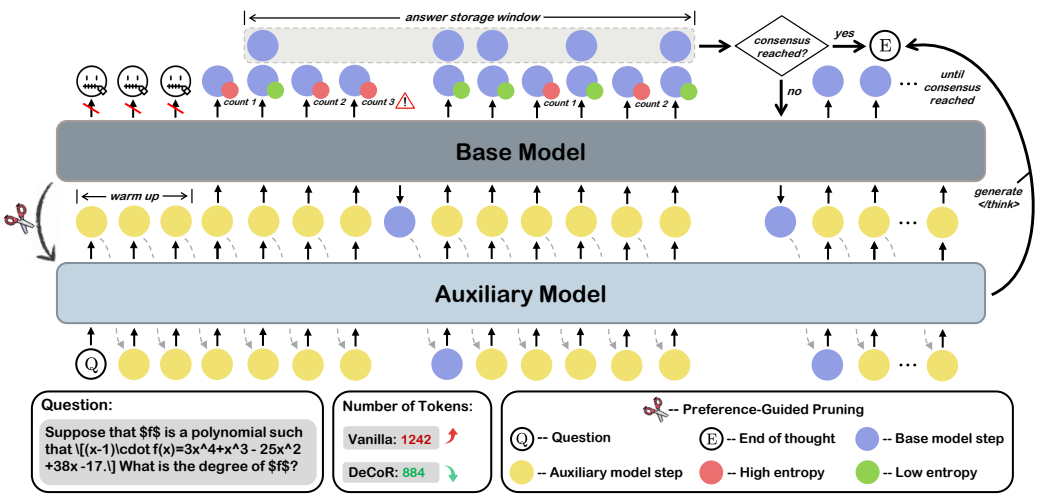


Figure 2: Overview of DeCoR. The framework derives an auxiliary model from the base model via preference-guided layer pruning to conduct continuous exploration. Conditioned on this exploratory trajectory, the base model generates intermediate answers and actively intervenes if (i) predictive entropy remains persistently high or (ii) intermediate answers fail to reach consensus.

where $\theta_l^{q,x}$ refers to the deviation in predictive preference at the attention or MLP layer of the l-th block during the generation of token x for prompt q, and g_0 represents the probability distribution calculated based on the output of the embedding layer.

3.2 Preference-Guided Structured Pruning

We leverage quantified preference to estimate layer importance for structured pruning, as shown in Figure 3. An intuitive approach to assessing layer importance is by averaging all the θ values. While most tokens are purely functional with low structural dependency and small θ values across layers, only a sparse subset of critical tokens (e.g., logical transitions and derived numerical results) induces drastic shifts in deep states. Relying solely on the average may allow these pervasive low values to attenuate the signal from critical periods, leading to the misjudgment of essential layers as redundant (Dettmers et al., 2022; Yin et al., 2024). To mitigate this, we incorporate the 95th percentile of θ as a complementary metric, capturing moments of pivotal insights. Regarding the CoT sequence, a weighted sum of the expectation and the 95th percentile of the θ values is calculated as

$$\mathbf{I}_{\mathcal{T}} = \mathbb{E}_{q\sim\mathcal{D}_{\text{calib}}}\left[\alpha \sum_{x\in\mathcal{T}_q} \boldsymbol{\theta}^{q,x}\right] + (1-\alpha)\cdot P_{95}\left(\{\boldsymbol{\theta}^{q,x}\}_{x\in\mathcal{T}_q, q\sim\mathcal{D}_{\text{calib}}}\right), \tag{3}$$

where $\alpha \in (0, 0.5)$, thereby assigning a higher weight to the second term. $\mathcal{D}_{\text{calib}}$ refers to the calibration dataset used to compute the layer importance, $\mathcal{T}_q$ denotes the set of CoT tokens generated based on prompt q, and $\boldsymbol{\theta}^{q,x} = [\theta_l^{q,x}]_{l\in\mathcal{L}}$. In contrast to the CoT sequence, the answer sequence is more concise and information-dense, placing lower demands on logical synthesis. Consequently, the distribution of KL divergence across these tokens tends to be homogeneous. Therefore, we add the 75th percentile for calculating layer importance in this stage, denoted as $\mathbf{I}_{\mathcal{A}}$[1]. The overall layer importance score is computed by averaging $\mathbf{I}_{\mathcal{T}}$ and $\mathbf{I}_{\mathcal{A}}$ as follows:

$$\boldsymbol{\mathcal{I}} = \frac{1}{2}\left(\mathbf{I}_{\mathcal{T}} + \mathbf{I}_{\mathcal{A}}\right), \tag{4}$$

where $\boldsymbol{\mathcal{I}} = [\mathcal{I}_l]_{l\in\mathcal{L}}$. We prune layers in ascending order of $\mathcal{I}_l$ according to a predetermined ratio ρ, along with normalization modules and residual blocks. Our pruning method can be realized

[1]The layer importance score for the answer sequence is denoted as $\mathbf{I}_{\mathcal{A}}$. The formulation of $\mathbf{I}_{\mathcal{A}}$ is derived by adapting Eq. (3) such that the CoT sequence $\mathcal{T}_q$ is replaced by the answer sequence $\mathcal{A}_q$, the 95th percentile is substituted with the 75th percentile, and the weighting coefficients α and $1-\alpha$ are both set to $1/2$.

via in-place masking, ensuring both functional role differentiation and trajectory coherence during reasoning. In addition, we implement the base and auxiliary models via shared parameters and a unified KV cache to further reduce memory footprint (detailed in Appendices G and H).

3.3 Dual-Indicator Gating

We devise a dual-indicator gating mechanism comprising an uncertainty indicator and a consistency indicator to facilitate efficient interaction between the base model and the auxiliary model. Reasoning segments are iteratively generated by the auxiliary model, denoted as Δs. The newline character '\n\n' is adopted as the delimiter for each individual segment. Combined with the preceding context S_m, we have $S_m \leftarrow S_m + \Delta s$. Subsequently, the base model $\mathcal{F}_{\text{base}}$ is prompted to sample an intermediate answer sequence $\hat{y} \leftarrow \mathcal{F}_{\text{base}}(S_m)$, serving as a proxy for the current reasoning depth. Since the majority of the output probability mass is concentrated within a sparse subset of candidate tokens (Holtzman et al., 2020; Lyu et al., 2025), the entropy over the top-k elements is computed to quantify answer uncertainty as

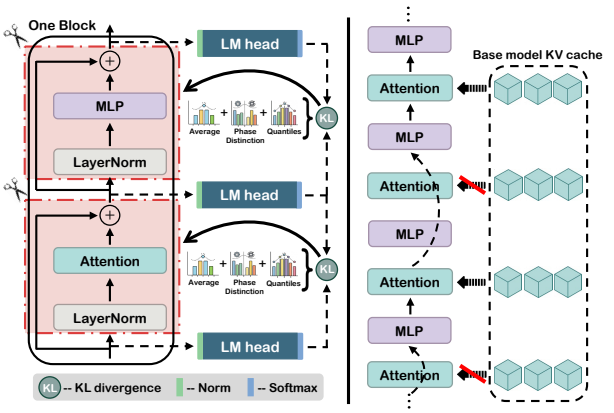


Figure 3: Preference-guided structured pruning. Layers are pruned along with their associated normalization modules. The base and auxiliary models share underlying parameters and a unified KV cache.

$$H(\hat{y}) = \frac{1}{|\hat{y}|} \sum_{j=1}^{|\hat{y}|} \left(\frac{-\sum_{i=1}^{k} \left(p_{i,j} \cdot \log \left(p_{i,j} \right) \right)}{\log (V)} \right), \tag{5}$$

where $p_{i,j}$ denotes each probability within the top-k distribution for each token in $\hat{y}$, $|\hat{y}|$ is the length of the answer sequence, V signifies the size of the vocabulary, and $\log(V)$ is the normalization term. This entropy-based technique for uncertainty measurement is commonly used in previous works (Han et al., 2024; Nikitin et al., 2024), while we adopt different elements for entropy calculation. To use uncertainty as a signal of insufficient reasoning fidelity or unproductive drifting, we define an uncertainty threshold τ and a counter with an upper bound $\mathcal{U}$. The condition $H(\hat{y}) \geqslant \tau$ triggers an incremental update of the counter. Once the counter reaches $\mathcal{U}$, the base model revises the reasoning direction and generates a fresh segment $\Delta s' \leftarrow \mathcal{F}_{\text{base}}(S_m)$, enabling the auxiliary model $\mathcal{F}_{\text{auxi}}$ to continue from a corrected logical state, that is, $\Delta s \leftarrow \mathcal{F}_{\text{auxi}}(S_m + \Delta s')$ and $S_m \leftarrow S_m + \Delta s'$.

In addition, we maintain a storage window of size ω to archive intermediate answers when $H(\hat{y}) < \tau$. Once the window reaches capacity ω, a consistency gating process is activated. The acceptance rate of intermediate answers is formulated to quantify answer consistency as follows:

$$\mathcal{R} = \frac{1}{\omega - 1} \sum_{i=1}^{\omega - 1} 1(\hat{y}_i = \hat{y}_\omega), \tag{6}$$

where $\hat{y}_\omega$ represents the last answer within the window, and $\hat{y}_i$ denotes any preceding answer prior to $\hat{y}_\omega$. The primary function of $\mathcal{R}$ is to capture answer stabilization as a proxy for reasoning convergence, since the response accuracy of LLMs is characterized by an initial ascent followed by a subsequent degradation as the reasoning process unfolds (Wu et al., 2026). Similarly, we establish an acceptance threshold ξ and arrive at the equation:

$$\begin{cases} y_{out} = \hat{y}_\omega, & \text{if } \mathcal{R} \geqslant \xi, \\ \Delta s'' \leftarrow \mathcal{F}_{\text{base}}(S_m), & \text{otherwise}, \end{cases} \tag{7}$$

where y_{out} is the final answer, $\Delta s''$ is a fresh segment generated by the base model, upon which the auxiliary model continues reasoning, and $S_m \leftarrow S_m + \Delta s''$. The purpose of $\Delta s''$ is to leverage the more robust output from the base model to stabilize subsequent reasoning. The condition $\mathcal{R} \geqslant \xi$ serves as a state where the results remain invariant over a substantial sequence of output tokens, providing evidence that the current reasoning trajectory can be terminated.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Models. We use R1-Distill-Llama-8B (DeepSeek-AI, 2025), Qwen3-8B-Thinking (Yang et al., 2025), and Ministral-3-8B-Reasoning (Liu et al., 2026) as the base models for our main experiments. This selection encompasses a comprehensive spectrum of reasoning LLMs, including models empowered by reasoning distillation, models with varied architectures, and models developed with intrinsic reasoning abilities.

Baselines. We compare DeCoR with three classes of baselines: **(i) Vanilla,** represented by the original reasoning models. **(ii) Intrinsic Mechanism Refinement,** including Coconut (Hao et al., 2025), ARLCP (Yu et al., 2026), IBPO (Yu et al., 2025), and ConPress (Deng et al., 2026). **(iii) Extrinsic Constraints,** covering DTSR (Xiang et al., 2026), TALE-EP (Han et al., 2025), DEER (Yang et al., 2026), and SoT (Aytes et al., 2025). Baseline results are taken from prior literature where available, and are otherwise reproduced using our base models following their official implementations.

Benchmarks. We evaluate DeCoR on five datasets: GSM8K (Cobbe et al., 2021), MATH500 (Hendrycks et al., 2021), AIME26 (Zhang & Math-AI Team, 2026), GPQA Diamond (Rein et al., 2024), and BBEH (Kazemi et al., 2025), covering mainstream tasks in mathematical reasoning and knowledge-based QA. We report accuracy (*Acc*, %), output token count (*TC*) and reasoning latency (*Lat*, s), averaged across 8 independent runs. Note that *TC* encompasses all tokens generated by both the auxiliary model and the base model throughout the reasoning process.

Implementation Details. We curate a calibration dataset to compute layer importance, which incorporates 400 samples from AQuA-RAT (Ling et al., 2017) and CommonsenseQA (Talmor et al., 2019), covering math problems and general knowledge questions in equal proportion. The average calibration time across all base models is approximately 5 A100 hours. We set the pruning ratio ρ to 0.12, and the pruned layer indices are listed in Table 1. Notably, we maintain a consistent configuration of $\alpha = 0.3$, $k = 20$, $\tau = 0.04$, $\omega = 5$, $\xi = 0.7$, and $\mathcal{U} = 3$ for all base models. An initial warm-up phase is implemented, preventing the gating process from being overly sensitive in early reasoning stages. All experiments are conducted on a single A100 40GB GPU, with the vLLM engine (Kwon et al., 2023) enabling inference acceleration and dynamic KV cache sharing. More implementation details are provided in Appendix C.

Table 1: Pruned layer indices for DeCoR. Indices start from zero, each referring to the specific Transformer block containing the attention or MLP layer.

Base Model	Attention Indices	MLP Indices
R1-Distill-Llama-8B (32 blocks)	[18, 19, 21]	[17, 25, 26, 28]
Qwen3-8B-Thinking (36 blocks)	[6, 7, 10, 16]	[7, 15, 16, 18]
Ministral-3-8B-Reasoning (34 blocks)	[11, 18, 23, 30]	[12, 17, 20, 28]

4.2 MAIN RESULTS

We evaluate the performance of DeCoR against all baselines, with the results summarized in Table 2. Key observations are as follows: (i) The three base models integrated with DeCoR consistently achieve the highest frequency of top-tier results across all datasets, underscoring the comprehensive efficacy of our framework across various tasks and model architectures. (ii) While Coconut achieves remarkable latency reductions on difficult benchmarks including AIME26, GPQA-D, and BBEH, its reasoning accuracy falls to nearly half of that yielded by DeCoR. For R1-Distill-Llama on AIME26, Coconut reduces *TC* by nearly 1,300 tokens compared to DeCoR, but yields an accuracy of only 20.8%, which is less than half of the 46.1% attained by DeCoR. (iii) IBPO significantly boosts the accuracy of all base models at the cost of a drastic increase in *TC* and latency. On Qwen3 and R1-Distill-Llama, IBPO outperforms DeCoR by 1.2% in accuracy on MATH500 and GSM8K. However, this performance edge is offset by a disproportionately higher *TC*. For R1-Distill-Llama on GSM8K, IBPO requires 747.36 tokens on average, which is nearly quadruple the 200.64 tokens generated by DeCoR. Similarly, on MATH500, IBPO consumes 3790.24 tokens for Qwen3, nearly triple the 1342.75 tokens generated by DeCoR. (iv) Results on Ministral-3 further demonstrate the robustness of DeCoR across model architectures. Compared with the vanilla model, DeCoR improves accuracy across all five benchmarks by 2.8 to 5.4 percentage points, while reducing *TC* by 22.7% to 46.5% and latency by 23.5% to 48.5%. On BBEH, DeCoR achieves the highest accuracy of 24.3% and the lowest *TC* of 1349.63 tokens, while its latency ranks second among all methods.

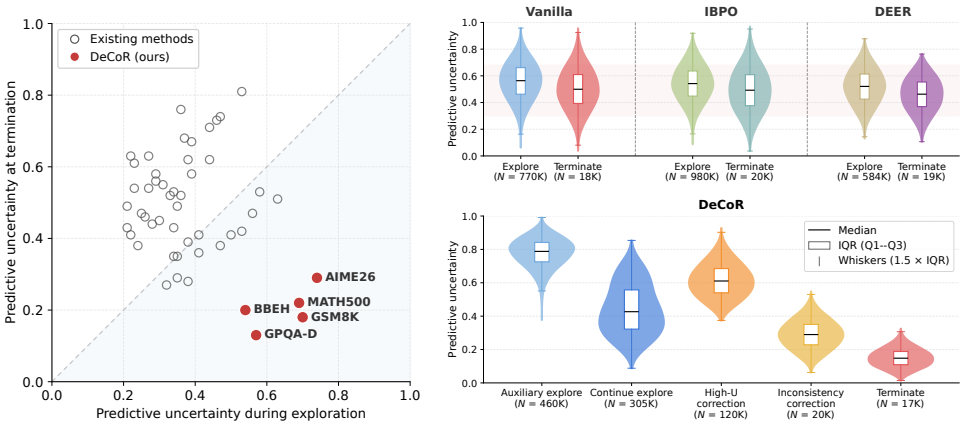
Table 2: Results of the main experiment. The best and second-best results are highlighted in bold and underlined, respectively.

Method	GSM8K			MATH500			AIME26			GPQA-D			BBEH		
	Acc ↑	*TC* ↓	*Lat* ↓	*Acc* ↑	*TC* ↓	*Lat* ↓	*Acc* ↑	*TC* ↓	*Lat* ↓	*Acc* ↑	*TC* ↓	*Lat* ↓	*Acc* ↑	*TC* ↓	*Lat* ↓
R1-Distill-Llama-8B															
Vanilla	81.5	401.93	7.35	75.8	1623.99	29.85	42.1	4566.75	84.13	44.5	3319.62	59.71	15.2	1217.30	22.21
Coconut	70.3	275.32	6.01	70.4	1269.76	21.37	20.8	**2137.38**	**41.74**	26.3	**1487.26**	**27.89**	7.3	**674.43**	**8.95**
ARLCP	83.8	301.49	6.27	78.7	1433.21	24.66	44.2	3871.50	71.42	46.5	2943.28	53.35	16.9	1069.54	19.83
IBPO	**86.2**	747.36	14.72	80.8	3105.89	50.25	45.8	7655.38	139.74	47.9	5117.45	87.27	17.8	2206.36	40.16
ConPress	80.8	312.45	8.42	75.1	1264.50	26.85	41.3	3584.20	82.16	44.1	3180.45	58.20	14.4	1115.80	21.60
DTSR	82.4	278.30	6.14	75.2	**1034.63**	21.89	42.7	3820.65	74.30	43.8	2540.18	44.12	14.6	942.30	17.45
TALE-EP	81.2	246.18	5.68	77.2	1341.28	25.12	40.8	4125.40	79.45	43.6	2712.90	48.35	14.8	915.65	18.90
DEER	82.1	224.57	5.36	75.6	1195.42	23.18	43.4	3748.15	70.88	46.1	2894.30	47.60	15.7	1064.20	19.82
SoT	78.8	232.61	5.82	75.4	1247.62	24.31	41.8	4195.37	80.05	44.7	2876.19	50.77	13.7	992.37	17.96
DeCoR	85.0	**200.64**	**5.21**	**82.3**	1137.73	**20.66**	**46.1**	3420.68	66.45	**48.3**	2281.56	35.03	**19.1**	856.25	16.09
Qwen3-8B-Thinking															
Vanilla	84.1	422.03	7.58	79.5	2061.87	33.94	45.2	5156.63	89.94	46.1	2977.31	57.43	16.6	1433.04	24.57
Coconut	81.6	286.98	5.07	75.3	1480.87	28.56	21.4	**2279.55**	**46.06**	27.3	**1269.27**	**27.65**	7.5	**766.73**	**9.65**
ARLCP	85.2	375.72	5.83	80.4	1750.69	26.11	46.2	4505.49	70.86	47.1	2601.12	50.31	17.6	1327.44	22.58
IBPO	86.5	846.53	13.76	**83.5**	3790.24	55.23	48.5	8334.96	145.62	49.2	4986.71	96.48	19.5	2548.95	41.22
ConPress	85.1	318.90	8.75	81.3	1530.40	29.45	47.3	3985.60	75.40	48.1	2620.40	56.12	17.8	1340.20	23.85
DTSR	83.4	258.40	4.02	80.3	1422.60	27.80	44.7	4580.25	84.15	46.8	2415.75	45.60	17.1	1085.40	16.12
TALE-EP	83.7	265.12	4.92	80.5	1495.18	26.94	46.1	4210.80	71.90	47.2	2310.60	43.25	17.4	1120.65	18.40
DEER	84.2	231.37	4.19	80.1	1405.88	28.65	45.4	4650.30	78.60	46.4	2488.20	49.10	15.8	1195.80	20.75
SoT	79.4	233.36	4.21	78.0	1574.88	29.69	44.2	4652.90	82.63	45.7	2274.66	47.19	15.2	1354.47	20.53
DeCoR	**88.2**	**229.30**	**3.97**	82.3	**1342.75**	**24.32**	**49.1**	3826.71	67.85	**49.7**	2213.65	41.77	**20.8**	978.52	11.66
Ministral-3-8B-Reasoning															
Vanilla	91.2	1487.75	35.38	84.1	2814.46	65.43	41.2	11465.27	240.29	50.2	6541.02	150.74	18.9	2056.13	47.89
Coconut	75.4	836.20	20.40	75.5	2149.15	57.22	22.7	**7875.06**	**108.11**	30.6	**3178.32**	**73.69**	8.8	1363.79	**26.52**
ARLCP	92.1	1067.91	28.53	85.2	2175.08	59.26	42.6	9632.27	146.31	51.3	5336.62	128.77	19.8	1573.02	38.64
IBPO	93.7	1936.25	40.59	86.9	3742.28	76.14	43.5	13275.38	294.72	52.7	8437.49	181.05	21.2	3247.61	72.03
ConPress	91.9	1175.40	32.15	83.5	2340.15	62.40	42.4	9750.30	188.40	49.1	6240.50	142.30	19.6	1485.30	34.20
DTSR	90.8	898.20	24.30	83.2	2215.60	54.20	42.1	10420.60	231.50	49.6	5080.20	124.80	19.3	1680.15	42.60
TALE-EP	90.3	1260.85	31.10	83.9	2520.40	58.75	41.6	10115.20	210.25	49.8	5710.65	136.15	18.7	1590.40	38.90
DEER	92.2	862.50	22.18	84.8	2017.18	56.10	41.1	9280.40	148.70	50.8	5390.40	131.70	17.1	1515.80	32.45
SoT	87.3	947.22	25.18	83.2	2046.53	52.25	39.9	9826.45	193.38	49.3	5048.73	121.35	17.4	1537.41	35.64
DeCoR	**94.0**	**795.68**	**18.23**	**87.2**	**1984.87**	**49.43**	**45.6**	8863.90	126.04	**53.0**	4892.38	115.26	**24.3**	**1349.63**	28.05

4.3 Analysis of Reasoning Dynamics

Uncertainty and Model Behavior. We investigate the interplay between predictive uncertainty and reasoning behaviors across different methods. We quantify auxiliary exploratory uncertainty using the entropy of the token immediately preceding each newline delimiter (see Appendix D for more details). Figure 4(a) plots the correspondence between mean uncertainty and reasoning behaviors across five benchmarks for all methods in Table 2. Baseline data points cluster near the diagonal in the upper region ($y > x$), indicating limited separation between exploration and termination. In contrast, DeCoR exhibits a clearer distinction, with higher entropy generally associated with exploration and lower entropy with termination. Figure 4(b) further illustrates the uncertainty distributions associated with different reasoning behaviors across the five benchmarks, comparing DeCoR with the vanilla model, an intrinsic method (IBPO), and an extrinsic method (DEER). Notably, DeCoR distinguishes four non-terminal behaviors: (i) autonomous exploration by the auxiliary model, (ii) auxiliary exploration resumed following an intermediate answer from the base model, (iii) base model correction triggered by persistent high-uncertainty intermediate answers, and (iv) base model correction triggered when intermediate answers exhibit low uncertainty yet low consistency. As observed, high uncertainty is consistently associated with auxiliary exploration. Furthermore, DeCoR achieves a clear correspondence between different actions and uncertainty, whereas the distributions of exploration and termination in baselines exhibit substantial overlap and high similarity.

Intervention Patterns and Role Coordination. Table 3 reveals a consistent coordination between the auxiliary model and the base model. On GSM8K, the base model performs merely 2 to 4 corrections per sample, while terminating 82.2% to 85.1% of the trajectories. On more challenging tasks (AIME26 and BBEH), base interventions scale to 7 to 9 per sample, with the base model still



(a) (b)

Figure 4: The connection between uncertainty and model behavior. (a) We calculate the average predictive uncertainty during exploration alongside that at reasoning termination on Qwen3-8B across five benchmarks, where existing methods (including the vanilla backbone) yield 45 hollow data points. (b) We further evaluate the distribution of predictive uncertainty conditioned on distinct reasoning behaviors across the five benchmarks. N represents the sample count per distribution.

Table 3: Intervention and termination statistics. Values under each benchmark denote: (i) the average number of corrective interventions by the base model per sample, (ii) the proportion of samples terminated by the auxiliary model, and (iii) the proportion of samples terminated by the base model.

Base Model	GSM8K	MATH500	AIME26	GPQA-D	BBEH
R1-Distill-Llama-8B	4 / 17.8 / 82.2	5 / 19.4 / 80.6	9 / 14.7 / 85.3	6 / 16.3 / 83.7	9 / 17.9 / 82.1
Qwen3-8B-Thinking	2 / 14.9 / 85.1	4 / 15.5 / 84.5	8 / 13.9 / 86.1	6 / 10.8 / 89.2	8 / 16.6 / 83.4
Ministral-3-8B-Reasoning	2 / 16.5 / 83.5	4 / 15.8 / 84.2	7 / 13.4 / 86.6	5 / 19.2 / 80.8	9 / 15.8 / 84.2

accounting for 82.1% to 86.6% of trajectory terminations. Crucially, the base model terminates the vast majority of trajectories across all benchmarks, while intervening selectively during the reasoning process. This pattern indicates that the auxiliary model primarily sustains exploration, whereas the base model serves as a selective corrective mechanism and plays the dominant role in determining reasoning termination. This uniform pattern across R1-Distill-Llama-8B, Qwen3-8B-Thinking, and Ministral-3-8B-Reasoning substantiates the robust adaptability of DeCoR.

Effective Reasoning Density. We track the token positions corresponding to the first correct answer onset and the termination of the reasoning process where the answer remains correct to quantify effective reasoning density. Specifically, we define the proportion of tokens between the onset and termination positions relative to the total reasoning tokens as the reasoning deliberation ratio (*RDR*). Observed from Figure 5, DeCoR consistently achieves a higher *RDR* than the base models while reducing the total reasoning token count. On MATH500, DeCoR increases the *RDR* of R1-Distill-Llama by 17.9 percentage points and reduces its total reasoning tokens from 1585 to 1048. On AIME26, it improves the *RDR* of Qwen3 by 11.4 percentage points while saving 961 tokens. Similar improvements are observed for Ministral-3 on GPQA-D, where the *RDR* increases by 9.8 percentage points and the total reasoning tokens decrease by 1792. Across all three models on BBEH, DeCoR yields *RDR* gains ranging from 13.5 to 18.5 percentage points. These results indicate that DeCoR invests more concentrated effort to explore the reasoning space even with shortened outputs.

4.4 Ablation Studies

Effectiveness of Preference-Guided Pruning. Ablation experiments are conducted on the proposed pruning method, and the results are shown in Figure 6. We evaluate against both ShortGPT (Men et al., 2025), which also adopts structured layer pruning, and the unpruned variant (w/o Prun-

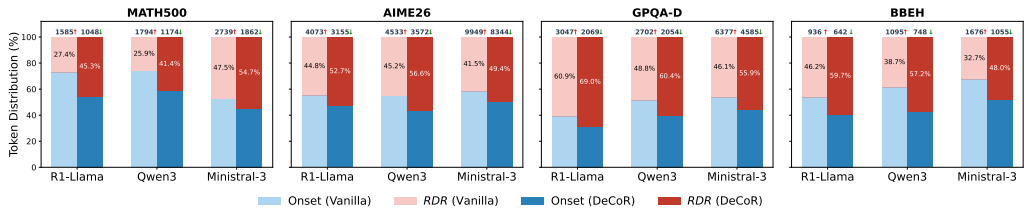


Figure 5: The visualization of RDR. Values atop the bars denote the total reasoning token counts.

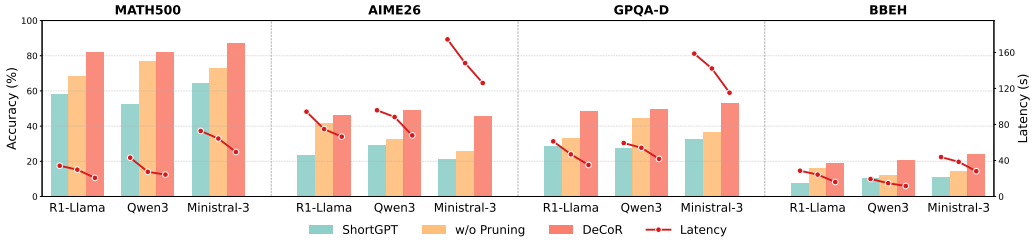


Figure 6: Ablation study on preference-guided pruning.

ing). Replacing our preference-guided pruning with ShortGPT causes a pronounced degradation in reasoning accuracy, despite its effectiveness as a structured pruning method for general LLMs. This result indicates that directly pruning reasoning LLMs according to generic importance criteria can remove components that are critical for maintaining reliable long-CoT reasoning. Notably, completely disabling pruning also consistently underperforms DeCoR, showing that the auxiliary pathway should not simply replicate the full base model. Instead, preference-guided pruning plays an important role in shaping a complementary explorer that interacts with the base model during reasoning. These results demonstrate that the proposed pruning strategy is essential not only for reducing latency, but also for establishing role specialization.

Gating Indicator Validation. We further evaluate the efficacy of our gating mechanism, which consists of intervention indicators based on intermediate answer uncertainty and consistency. As shown in Table 4, removing either indicator leads to performance degradation, while both variants outperform the vanilla baselines. Notably, the combination of these two indicators yields a super-additive gain in accuracy, where the aggregate accuracy boost is greater than the linear sum of individual gains. We attribute this to their complementary dynamics: the answer-uncertainty indicator helps stabilize the reasoning trajectory, providing a more reliable context for the answer-consistency indicator to efficiently identify an appropriate termination point.

Table 4: Ablation results of gating indicators, averaged across five benchmarks.

Model	R1-Llama-8B		Qwen3-8B		Ministral-3-8B	
	Acc ↑	*TC* ↓	*Acc* ↑	*TC* ↓	*Acc* ↑	*TC* ↓
Vanilla	51.82	2225.92	54.30	2410.18	57.12	4872.93
Uncertainty-only	52.41	2014.67	55.88	2289.45	58.73	4521.80
Consistency-only	53.89	1748.25	56.06	1976.32	58.97	3984.65
DeCoR	**56.16**	**1579.37**	**58.02**	**1718.19**	**60.82**	**3577.29**

5 Conclusion

In this work, we introduce DeCoR, a novel framework for efficient reasoning that decouples exploration from termination through functionally differentiated reasoning pathways. DeCoR constructs a high-uncertainty auxiliary explorer via preference-guided structured pruning and coordinates it with the base model through a dual-indicator gating mechanism. Extensive experiments across multiple reasoning benchmarks and base models demonstrate that DeCoR consistently improves reasoning accuracy while substantially reducing output length. Further analyses and ablations verify the effectiveness of both preference-guided pruning and dual-indicator gating in enabling compact and reliable reasoning trajectories. In future work, we plan to extend DeCoR to multimodal large language models and larger-scale reasoning models to investigate its scalability and generalizability.

AI USE STATEMENT

In this work, we used generative AI tools only for language editing and polishing. We have not used generative AI tools for research ideation, methodology design, implementation, data preparation, result interpretation, or the formulation of scientific claims. All AI-assisted edits have been carefully reviewed and revised by the authors to ensure technical accuracy and consistency. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, personally identifiable information, or practices that could raise ethical concerns. There are no potential conflicts of interest, and no sensitive or harmful methodologies were employed. We adhered to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We have taken measures to ensure reproducibility. Key code components, dataset descriptions, hyperparameter settings, and implementation procedures are detailed in Section 4.1, the appendix, and the supplementary material.

REFERENCES

Pranjal Aggarwal and Sean Welleck. L1: Controlling how long a reasoning model thinks with reinforcement learning. In *COLM*, 2025. URL https://openreview.net/forum?id=4jdIxXBNve.

Simon A. Aytes, Jinheon Baek, and Sung Ju Hwang. Sketch-of-thought: Efficient llm reasoning with adaptive cognitive-inspired sketching. In *EMNLP*, 2025. URL https://aclanthology.org/2025.emnlp-main.1236/.

Ed Bullmore and Olaf Sporns. The economy of brain network organization. *Nature Reviews Neuroscience*, 13(5):336–349, 2012. ISSN 1471-0048. URL https://doi.org/10.1038/nrn3214.

Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, and Yong Wang. GPG: A simple and strong reinforcement learning baseline for model reasoning. In *ICLR*, 2026. URL https://openreview.net/forum?id=inccdtfx8x.

Yu-Neng Chuang, Prathusha Kameswara Sarma, Parikshit Gopalan, John Boccio, Sara Bolouki, Xia Hu, and Helen Zhou. Learning to route LLMs with confidence tokens. In *ICML*, 2025. URL https://openreview.net/forum?id=U08mUogGDM.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *ArXiv*, 2021. URL https://arxiv.org/abs/2110.14168.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *ArXiv*, 2025. URL https://arxiv.org/abs/2501.12948.

Jie Deng, Shining Liang, Jun Li, Hongzhi Li, and Yutao Xie. Conpress: Learning efficient reasoning from multi-question contextual pressure. In *ICML*, 2026. URL https://openreview.net/forum?id=piXLPucLRA.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Llm.int8(): 8-bit matrix multiplication for transformers at scale. In *NeurIPS*, 2022. URL https://openreview.net/forum?id=dXiGWqBoxaD.

Longwei Ding, Anhao Zhao, Fanghua Ye, Ziyang Chen, and Xiaoyu Shen. From LLMs to LRMs: Rethinking pruning for reasoning-centric models. *ArXiv*, 2026. URL https://arxiv.org/abs/2601.18091.

Yichao Fu, Junda Chen, Siqi Zhu, Zheyu Fu, Zhongdongming Dai, Yonghao Zhuang, Yian Ma, Aurick Qiao, Tajana Rosing, Ion Stoica, and Hao Zhang. Efficiently scaling LLM reasoning programs with certaindex. In *NeurIPS*, 2025. URL https://openreview.net/forum?id=nn51ewu5k2.

Soumya Suvra Ghosal, Souradip Chakraborty, Avinash Reddy, Yifu Lu, Mengdi Wang, Dinesh Manocha, Furong Huang, Mohammad Ghavamzadeh, and Amrit Singh Bedi. Does thinking more always help? mirage of test-time scaling in reasoning models. In *NeurIPS*, 2025. URL https://openreview.net/forum?id=tKPqbamNb9.

Jiatong Han, Jannik Kossen, Muhammed Razzak, Lisa Schut, Shreshth A Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. In *ICML*, 2024. URL https://openreview.net/forum?id=Zd0XLr6JKn.

Jinyi Han, Ying Huang, Ying Liao, Haiquan Zhao, Zishang Jiang, Xinyi Wang, Xikun Lu, Guanghao Zhou, Sihang Jiang, Jiaqing Liang, Weikang Zhou, Zeye Sun, Fei Yu, and Yanghua Xiao. Your models have thought enough: Training large reasoning models to stop overthinking. In *ICLR*, 2026. URL https://openreview.net/forum?id=2u5ZRzDyS0.

Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. Token-budget-aware llm reasoning. In *ACL Findings*, 2025. URL https://aclanthology.org/2025.findings-acl.1274/.

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. In *COLM*, 2025. URL https://openreview.net/forum?id=Itxz7S4Ip3.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS*, 2021. URL https://openreview.net/forum?id=7Bywt2mQsCe.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *ICLR*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Weiyu Huang, Pengle Zhang, Xiaolu Zhang, JUN ZHOU, Jun Zhu, and Jianfei Chen. Deterministic differentiable structured pruning for large language models. In *ICML*, 2026. URL https://openreview.net/forum?id=64Kj8z626A.

Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, et al. Big-bench extra hard. In *ACL*, 2025. URL https://aclanthology.org/2025.acl-long.1285/.

Kimi Team. Kimi k1.5: Scaling reinforcement learning with llms. *ArXiv*, 2025. URL https://arxiv.org/pdf/2501.12599.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *SOSP*, 2023. URL https://doi.org/10.1145/3600006.3613165.

Chen Li, Nazhou Liu, and Kai Yang. Adaptive group policy optimization: Towards stable training and token-efficient reasoning. *ArXiv*, 2025a. URL https://arxiv.org/abs/2503.15952.

Gang Li, Yan Chen, Ming Lin, and Tianbao Yang. DRPO: Efficient reasoning via decoupled reward policy optimization. In *ICLR*, 2026. URL https://openreview.net/forum?id=GP5RHZnEsw.

Guanchen Li, Yixing Xu, Zeping Li, Ji Liu, Xuanwu Yin, Dong Li, and Emad Barsoum. Týr-the-pruner: Structural pruning LLMs via global sparsity distribution optimization. In *NeurIPS*, 2025b. URL https://openreview.net/forum?id=rAuRLePL2R.

Junhong Lin, Xinyue Zeng, Jie Zhu, Song Wang, Julian Shun, Jun Wu, and Dawei Zhou. Plan and budget: Effective and efficient test-time scaling on reasoning large language models. In *ICLR*, 2026. URL https://openreview.net/forum?id=ctspw4CqbS.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *ACL*, 2017. URL https://aclanthology.org/P17-1015/.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, et al. Ministral 3. *ArXiv*, 2026. URL https://arxiv.org/abs/2601.08584.

Haotian Luo, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Xiaochun Cao, Dacheng Tao, Naiqiang Tan, and Li Shen. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. In *ACL Findings*, 2026. URL https://aclanthology.org/2026.findings-acl.697/.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. Calibrating large language models with sample consistency. In *AAAI*, 2025. URL https://ojs.aaai.org/index.php/AAAI/article/view/34120.

Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. Reasoning models can be effective without thinking. *ArXiv*, 2025a. URL https://arxiv.org/abs/2504.09858.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. LLM-pruner: On the structural pruning of large language models. In *NeurIPS*, 2023. URL https://openreview.net/forum?id=J8Ajf9WfXP.

Xinyin Ma, Guangnian Wan, Runpeng Yu, Gongfan Fang, and Xinchao Wang. Cot-valve: Length-compressible chain-of-thought tuning. In *ACL*, 2025b. URL https://aclanthology.org/2025.acl-long.300/.

Xin Men, Mingyu Xu, Qingyu Zhang, Qianhao Yuan, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. Shortgpt: Layers in large language models are more redundant than you expect. In *ACL Findings*, 2025. URL https://aclanthology.org/2025.findings-acl.1035/.

Vinod Menon. Large-scale brain networks and psychopathology: A unifying triple network model. *Trends in Cognitive Sciences*, 15(10):483–506, 2011. ISSN 1364-6613. URL https://doi.org/10.1016/j.tics.2011.08.003.

Alexander V Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. In *NeurIPS*, 2024. URL https://openreview.net/forum?id=j2wCrWmgMX.

OpenAI. Openai o1 system card. *ArXiv*, 2024. URL https://arxiv.org/abs/2412.16720.

Xiaoye Qu, Yafu Li, Zhaochen Su, Weigao Sun, Jianhao Yan, Dongrui Liu, Ganqu Cui, Daizong Liu, Shuxian Liang, Junxian He, Peng Li, Wei Wei, Jing Shao, Chaochao Lu, Yue Zhang, Xian-Sheng Hua, Bowen Zhou, and Yu Cheng. A survey of efficient reasoning for large reasoning models: Language, multimodality, and beyond. *ArXiv*, 2025. URL https://arxiv.org/abs/2503.21614.

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 *COLM*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. Codi: Compressing chain-of-thought into continuous space via self-distillation. In *EMNLP*, 2025. URL https://aclanthology.org/2025.emnlp-main.36/.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, Hanjie Chen, and Xia Hu. Stop overthinking: A survey on efficient reasoning for large language models. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=HvoG8SxggZ.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *NAACL*, 2019. URL https://aclanthology.org/N19-1421/.

Martijn P. van den Heuvel and Olaf Sporns. Network hubs in the human brain. *Trends in Cognitive Sciences*, 17(12):683–696, 2013. ISSN 1364-6613. URL https://doi.org/10.1016/j.tics.2013.09.012.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. URL https://openreview.net/forum?id=2q334K-GjaL.

Ziyang Wang, Jiangfeng Xiao, Chuan Xiao, Ruoxiang LI, Rui Mao, and Jianbin Qin. Grasprune: Global gating for budgeted structured pruning of large language models. In *ACL*, 2026. URL https://aclanthology.org/2026.acl-long.491/.

Stanley Wei and Juno Kim. Provable benefits of RLVR over SFT for reasoning models: Learning to backtrack efficiently. In *ICML*, 2026. URL https://openreview.net/forum?id=P3Mnh7mF5a.

Yuyang Wu, Yifei Wang, Ziyu Ye, Tianqi Du, Stefanie Jegelka, and Yisen Wang. When more is less: Understanding chain-of-thought length in llms. In *ICLR*, 2026. URL https://openreview.net/forum?id=6QDFsYxtI1.

Heming Xia, Yongqi Li, Jun Zhang, Cunxiao Du, and Wenjie Li. SWIFT: On-the-fly self-speculative decoding for LLM inference acceleration. In *ICLR*, 2025. URL https://openreview.net/forum?id=EKJhH5D5wA.

Yang Xiang, Yixin Ji, Ruotao Xu, Dan Qiao, Zheming Yang, Juntao Li, and Min Zhang. When is thinking enough? early exit via sufficiency assessment for efficient reasoning. In *ACL*, 2026. URL https://aclanthology.org/2026.acl-long.1080/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. *ArXiv*, 2025. URL https://arxiv.org/abs/2505.09388.

Chenxu Yang, Qingyi Si, Yongjie Duan, Zheliang Zhu, Chenyu Zhu, Qiaowei Li, Minghui Chen, Zheng Lin, and Weiping Wang. Dynamic early exit in reasoning models. In *ICLR*, 2026. URL https://openreview.net/forum?id=NpU7ZXafRi.

Lu Yin, You Wu, Zhenyu Zhang, Cheng-Yu Hsieh, Yaqing Wang, and Yiling Jia. Outlier weighed layerwise sparsity (OWL): A missing secret sauce for pruning LLMs to high sparsity. In *ICML*, 2024. URL https://openreview.net/forum?id=ahEm3l2P6w.

Zewei Yu, Lirong Gao, Yuke Zhu, Bo Zheng, Junbo Zhao, Sheng Guo, and Haobo Wang. Stop unnecessary reflection: Training LRMs for efficient reasoning with adaptive reflection and length coordinated penalty. In *ICLR*, 2026. URL https://openreview.net/forum?id=aRzEtK9Ite.

Zishun Yu, Tengyu Xu, Di Jin, Karthik Abinav Sankararaman, Yun He, Wenxuan Zhou, Zhouhao Zeng, Eryk Helenowski, Chen Zhu, Sinong Wang, Hao Ma, and Han Fang. Think smarter not harder: Adaptive reasoning with inference aware optimization. In *ICML*, 2025. URL https://openreview.net/forum?id=0ERw2196o1.

Jun Zhang, Jue Wang, Huan Li, Lidan Shou, Ke Chen, Gang Chen, and Sharad Mehrotra. Draft& verify: Lossless large language model acceleration via self-speculative decoding. In *ACL*, 2024. URL https://aclanthology.org/2024.acl-long.607/.

Yifan Zhang and Math-AI Team. American invitational mathematics examination (aime) 2026, 2026. URL https://huggingface.co/datasets/math-ai/aime26.

CONTENTS

1 Introduction **1**

2 Related Work **3**

3 Method **3**

3.1 Definition of Predictive Preference 3

3.2 Preference-Guided Structured Pruning 4

3.3 Dual-Indicator Gating 5

4 Experiments **6**

4.1 Experimental Settings 6

4.2 Main Results 6

4.3 Analysis of Reasoning Dynamics 7

4.4 Ablation Studies 8

5 Conclusion **9**

A Algorithm Presentation **15**

B Benchmark Descriptions **16**

C More Implementation Details **16**

C.1 Metric Definitions 16

C.2 Configurations in the Pruning Stage 16

C.3 Configurations in the Inference Stage 17

D More Analysis of Uncertainty and Model Behavior **17**

E Hyperparameter Analysis **18**

F Performance of the Auxiliary Model **22**

G Analysis of KV Cache Sharing **22**

H Memory Usage **23**

I Comparison with Self-Speculative Decoding **24**

J Case Studies **24**

A ALGORITHM PRESENTATION

The implementation of DeCoR can be structured into two stages: (i) The derivation of an auxiliary model from the base model via preference-guided structured pruning, which utilizes a calibration dataset to estimate layer importance. (ii) The execution of collaborative reasoning between the base model and the auxiliary model, regulated by the dual-indicator gating mechanism. The detailed procedure is presented in Algorithms 1 and 2.

Algorithm 1 Construction Procedure for the Auxiliary Model

Require: Base model $\mathcal{F}_{\text{base}}$, calibration dataset $\mathcal{D}_{\text{calib}}$, importance weight α, pruning ratio ρ

for each prompt $q \in \mathcal{D}_{\text{calib}}$ **do**
  Extract hidden states $h_l^{q,x} \leftarrow \mathcal{F}_{\text{base}}(q)$ from the attention or MLP layer of the l-th block during the generation of token x
  Calculate layer-wise predictive preference $g_l^{q,x} = \text{Softmax}\left(f_{\text{head}}\left(\text{Norm}\left(h_l^{q,x}\right)\right)\right)$
  Compute preference shift $\theta_l^{q,x} = D_{KL}(g_{l-1}^{q,x} || g_l^{q,x})$
  Concatenate $\theta_l^{q,x}$ from each layer during the generation of token x to obtain the vector $\boldsymbol{\theta}^{q,x}$
end for
Calculate:
- $\mathbf{I}_{\mathcal{T}} = \mathbb{E}_{q\sim\mathcal{D}_{\text{calib}}}\left[\alpha \sum_{x\in\mathcal{T}_q} \boldsymbol{\theta}^{q,x}\right] + (1-\alpha)\cdot P_{95}\left(\{\boldsymbol{\theta}^{q,x}\}_{x\in\mathcal{T}_q, q\sim\mathcal{D}_{\text{calib}}}\right)$
- $\mathbf{I}_{\mathcal{A}} = \frac{1}{2}\left(\mathbb{E}_{q\sim\mathcal{D}_{\text{calib}}}\left[\sum_{x\in\mathcal{A}_q} \boldsymbol{\theta}^{q,x}\right] + P_{75}\left(\{\boldsymbol{\theta}^{q,x}\}_{x\in\mathcal{A}_q, q\sim\mathcal{D}_{\text{calib}}}\right)\right)$
- Layer importance vector $\boldsymbol{\mathcal{I}} = \frac{1}{2}\left(\mathbf{I}_{\mathcal{T}} + \mathbf{I}_{\mathcal{A}}\right)$

Prune layers of $\mathcal{F}_{\text{base}}$ with lowest importance in $\boldsymbol{\mathcal{I}}$ according to ρ by masking
return Auxiliary model $\mathcal{F}_{\text{auxi}}$

Algorithm 2 Collaborative Reasoning

Require: Base model $\mathcal{F}_{\text{base}}$, auxiliary model $\mathcal{F}_{\text{auxi}}$, given prompt q, preceding context S_m, counter C, counter upper bound $\mathcal{U}$, uncertainty threshold τ, storage window W, max window size ω, acceptance threshold ξ

Initialize: $S_m \leftarrow q$, $C \leftarrow 0$, $W \leftarrow \varnothing$
repeat
  while $|W| < \omega$ **do**
    $\Delta s \leftarrow \mathcal{F}_{\text{auxi}}(S_m)$, $S_m \leftarrow S_m + \Delta s$
    if the auxiliary model naturally terminates reasoning **then**
      return S_m ▷ Auxiliary-model termination
    end if
    $\hat{y} \leftarrow \mathcal{F}_{\text{base}}(S_m)$ ▷ Generating intermediate answer
    Calculate $H(\hat{y}) = \frac{1}{|\hat{y}|}\sum_{j=1}^{|\hat{y}|}\left(\frac{-\sum_{i=1}^{k}(p_{i,j}\cdot\log(p_{i,j}))}{\log(V)}\right)$
    if $H(\hat{y}) \geq \tau$ **then** ▷ Uncertainty gating
      $C \leftarrow C + 1$
      if $C = \mathcal{U}$ **then**
        $\Delta s' \leftarrow \mathcal{F}_{\text{base}}(S_m)$, $S_m \leftarrow S_m + \Delta s'$ ▷ Altering the reasoning direction
        $C \leftarrow 0$
      end if
    else
      W.append($\hat{y}$)
    end if
  end while
  Calculate $\mathcal{R} = \frac{1}{\omega-1}\sum_{i=1}^{\omega-1} 1(\hat{y}_i = \hat{y}_\omega)$
  if $\mathcal{R} < \xi$ **then** ▷ Inconsistency gating
    $\Delta s'' \leftarrow \mathcal{F}_{\text{base}}(S_m)$, $S_m \leftarrow S_m + \Delta s''$ ▷ Maintaining the robustness of reasoning
    $W \leftarrow \varnothing$
  end if
until $\mathcal{R} \geq \xi$
return $S_m + \hat{y}_\omega$

B BENCHMARK DESCRIPTIONS

GSM8K (Cobbe et al., 2021) comprises 8792 human-curated grade school math word problems. These tasks typically necessitate two to eight steps of simple logical reasoning, specifically evaluating the capacity of LLMs to execute extensive multi-step reasoning.

MATH500 (Hendrycks et al., 2021) is a representative and challenging benchmark. It comprises 500 high-quality, competition-level problems spanning diverse mathematical disciplines, including algebra, geometry, and number theory. Each problem is characterized by its high difficulty level, demanding sophisticated multi-step logical deduction.

AIME26 (Zhang & Math-AI Team, 2026) comprises highly challenging problems from the American Invitational Mathematics Examination 2026. It consists of 30 samples designed to evaluate the frontiers of advanced mathematical reasoning in LLMs.

GPQA Diamond (Rein et al., 2024) is the highest quality subset (198 samples) of the Graduate-Level Google-Proof Q&A benchmark, featuring exceptionally challenging multiple-choice questions across various scientific disciplines. Every inclusion of the Diamond partition has been cross-verified by multiple subject-matter experts to guarantee objective correctness and high consensus.

BBEH (Kazemi et al., 2025), or BIG-Bench Extra Hard, is a challenging general-reasoning benchmark comprising 23 diverse tasks and 4,520 samples. It encompasses difficult tasks that evaluate similar or broader reasoning capabilities, including multi-hop reasoning, long-context processing, error detection, resistance to distractors, and on-the-fly pattern induction.

C MORE IMPLEMENTATION DETAILS

C.1 METRIC DEFINITIONS

We present the evaluation metrics utilized in the experiments in Section 4.1, specifically accuracy (*Acc*, %), output token count (*TC*), and latency (*Lat*, s). The formulation of *Acc* is

$$Acc = \text{pass@1} = \mathbb{E}_{\mathcal{D}_{\text{task}}}\left[\frac{c}{n}\right] \times 100\%, \tag{8}$$

where n denotes the number of responses generated for each sample in the task dataset $\mathcal{D}_{\text{task}}$, which is set to 8 for all experiments, and c signifies the number of correct responses among them.

TC denotes the total number of output tokens generated for each sample across both the thinking and response stages. For DeCoR, *TC* encompasses all tokens generated by both the auxiliary and base models, covering intermediate answers as well as corrective reasoning. The reported *TC* is averaged over all evaluation samples.

Lat is measured as wall-clock time continuously from the emission of the first reasoning token by the auxiliary model to the completion of the returned output. This interval encompasses the prefilling of the accumulated reasoning context by the base model, the decoding of intermediate answers, the generation of all corrective tokens, and all other method-specific operations performed within this interval. Only the initial prompt prefilling preceding the first reasoning token is excluded, with the same timing boundary applied to all methods.

C.2 CONFIGURATIONS IN THE PRUNING STAGE

For the calculation of layer importance, we construct a calibration dataset containing 400 samples, with 200 mathematical reasoning problems from AQuA-RAT (Ling et al., 2017) and 200 commonsense reasoning questions from CommonsenseQA (Talmor et al., 2019). AQuA-RAT contains approximately 100K multiple-choice algebraic word problems paired with natural language rationales. We use it as the source of mathematical reasoning samples in the calibration set. CommonsenseQA contains 12K multiple-choice questions targeting commonsense reasoning, constructed from ConceptNet-based concepts and relations. We use it as the source of non-mathematical reasoning samples in the calibration set. We set `temperature=1.0`, `top-p=0.95`, and `max_new_tokens=16384` for all base models during calibration.

Assistant Prompts for Different Functional Requirements

For the generation of standard reasoning segments: None.
For the generation of intermediate answers: `Now you must generate the final answer in the format \boxed{}.`
For the reasoning direction adjustment: `Now you must try a different way of thinking.`

Figure 7: Assistant prompts for DeCoR during inference.

C.3 Configurations in the Inference Stage

During the warm-up phase, the auxiliary model produces ten consecutive segments to stabilize the reasoning sequence. The decoding parameters are configured according to different functional requirements: (i) For the generation of standard reasoning segments, we set `temperature=1.0`, `top-p=0.95`, and `max_new_tokens=512`. (ii) For the generation of intermediate answers, we set `temperature=0.4`, `top-p=0.95`, and `max_new_tokens=30`. (iii) For the reasoning direction adjustment, we set `temperature=1.0`, `top-p=0.95`, and `max_new_tokens=50`. The assistant prompts for DeCoR are listed in Figure 7. It should be noted that these concise prompts are not meticulously designed, and any prompt variant that serves the same objective is applicable.

D More Analysis of Uncertainty and Model Behavior

In Section 4.3, we examined the correspondence between reasoning behaviors and predictive uncertainty across different methods with Qwen3-8B serving as the base model. Here, we extend this investigation to R1-Distill-Llama-8B and Ministral-3-8B. We first clarify the exact formulation used to quantify the uncertainty associated with each behavior in Figure 4(b) as follows:

Baseline Exploration. For each baseline, we measure exploration uncertainty using the entropy of the last token immediately preceding each newline delimiter during reasoning generation.

Baseline Termination. We measure termination uncertainty as the entropy of the final reasoning token immediately preceding the transition out of the reasoning process. Trajectories that reach the maximum generation length and are forcibly truncated are excluded from the termination distribution.

Auxiliary Exploration. We measure auxiliary exploration uncertainty in the same manner as baseline exploration.

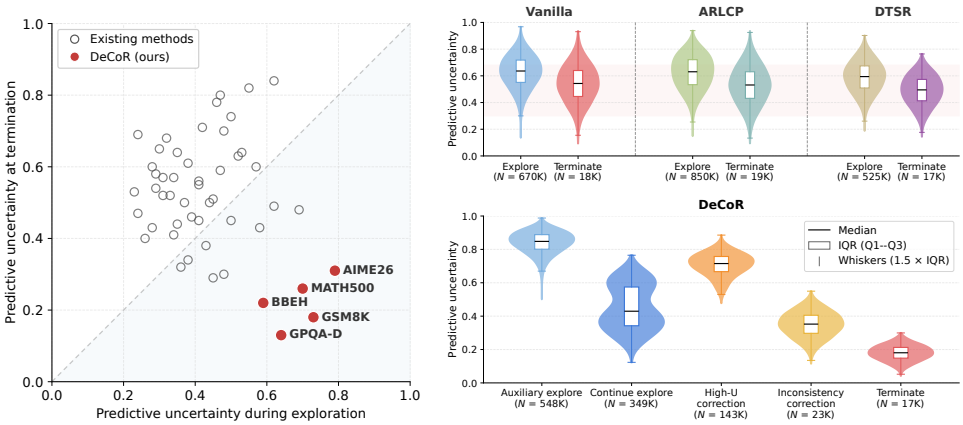
Continued Exploration. We record the mean token entropy of the current intermediate answer produced by the base model whenever the gating mechanism decides to continue auxiliary reasoning without triggering correction or termination.

High Uncertainty Correction. We use the mean token entropy of the intermediate answer whose uncertainty causes the persistent high uncertainty counter to reach its upper bound and trigger a corrective intervention.

Inconsistency Correction. We use the mean token entropy of the latest intermediate answer that completes the storage window but fails the consistency criterion, thereby triggering correction by the base model.

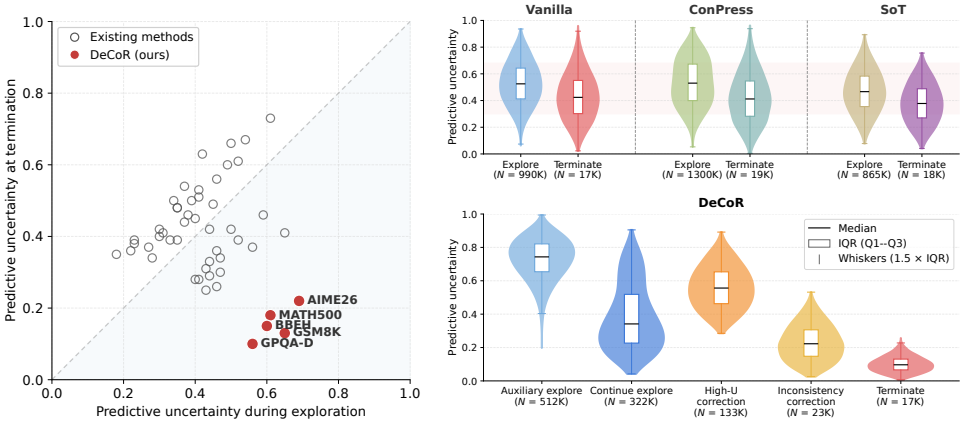
DeCoR Termination. We use the mean token entropy of the latest intermediate answer that completes the storage window and satisfies the consistency criterion, thereby triggering reasoning termination.

As shown in Figures 8 and 9, the behavior patterns observed on Qwen3-8B remain consistent for R1-Distill-Llama-8B and Ministral-3-8B-Reasoning. Existing methods (Vanilla, ARLCP, DTSR, ConPress, and SoT) show considerable overlap between exploratory and terminating states, with benchmark points scattered around the diagonal. DeCoR instead remains in the lower-right region



(a) (b)

Figure 8: The connection between uncertainty and model behavior under R1-Distill-Llama-8B.



(a) (b)

Figure 9: The connection between uncertainty and model behavior under Ministral-3-8B-Reasoning.

across all tasks, indicating stronger uncertainty during reasoning and lower values near termination. Its auxiliary pathway operates mainly in a high-uncertainty regime, whereas continued reasoning covers a broader range of states. High uncertainty corrections are concentrated at relatively large values, while inconsistency corrections and termination occupy lower regions. Despite differences in shape across models, these results show that DeCoR exhibits a more consistent association between uncertainty and distinct behaviors.

E Hyperparameter Analysis

Note. All hyperparameters in the main experiments were tuned on the previously constructed calibration dataset. The present analysis serves solely to validate the robustness of this configuration across the evaluation base models and benchmarks. In each parameter ablation experiment, all other parameters are held constant at the values used in the main text.

Importance weight α. We investigate the impact of different values of α on the pruning results and the overall performance of DeCoR. As shown in Table 5, $\alpha = 0.3$ consistently achieves the highest accuracy while minimizing *TC* and *Lat* across all three base models and five benchmarks. Moving α in either direction degrades both effectiveness and efficiency. Furthermore, Figure 10 presents the layer importance score trajectories under a range of α assignments. Although the global trends of these curves are closely aligned, the indices corresponding to less important layers exhibit distinct variations. Based on these observations, selecting values for α other than 0.3 results in the pruning of

Table 5: Model performance based on different values of α. The best results are highlighted in bold.

Values of α	GSM8K			MATH500			AIME26			GPQA-D			BBEH		
	Acc ↑	*TC* ↓	*Lat* ↓	*Acc* ↑	*TC* ↓	*Lat* ↓	*Acc* ↑	*TC* ↓	*Lat* ↓	*Acc* ↑	*TC* ↓	*Lat* ↓	*Acc* ↑	*TC* ↓	*Lat* ↓
R1-Distill-Llama-8B															
0.1	83.2	341.45	6.95	79.4	1580.64	28.45	44.1	4250.62	81.20	46.5	3120.48	55.63	17.9	1180.45	20.15
0.2	83.9	236.28	5.95	81.5	1210.45	22.18	44.8	3650.15	69.42	47.6	2420.34	38.25	18.3	920.18	17.45
0.3	**85.0**	**200.64**	**5.21**	**82.3**	**1137.73**	**20.66**	**46.1**	**3420.68**	**66.45**	**48.3**	**2281.56**	**35.03**	**19.1**	**856.25**	**16.09**
0.4	84.1	222.58	5.68	81.1	1280.24	23.38	45.4	3550.45	68.18	47.3	2550.42	41.57	18.7	890.76	16.82
0.49	82.5	314.38	8.21	78.8	1720.55	32.14	43.5	4820.36	92.58	46.4	3450.18	62.40	17.8	1320.48	24.58
Qwen3-8B-Thinking															
0.1	86.1	395.42	6.84	78.6	1980.48	32.15	46.5	5480.34	95.24	48.3	3120.45	58.87	19.5	1350.24	18.45
0.2	87.1	254.18	4.42	81.9	1425.32	26.14	47.9	4070.18	71.28	49.1	2350.12	44.48	20.4	1020.45	12.50
0.3	**88.2**	**229.30**	**3.97**	**82.3**	**1342.75**	**24.32**	**49.1**	**3826.71**	**67.85**	**49.7**	**2213.65**	**41.77**	**20.8**	**978.52**	**11.66**
0.4	87.6	265.38	4.65	81.2	1480.58	27.45	48.4	3940.24	69.58	48.8	2410.38	46.15	20.3	1050.62	13.18
0.49	85.8	362.18	6.12	77.4	1850.34	35.58	46.2	5100.75	88.36	47.8	2950.78	56.24	18.2	1420.18	20.35
Ministral-3-8B-Reasoning															
0.1	93.5	1250.38	28.45	86.1	2580.34	58.58	43.9	10850.42	166.38	52.2	6120.48	142.15	23.8	1850.48	38.45
0.2	93.7	885.24	21.85	86.8	2050.18	52.15	44.5	9120.58	134.45	52.6	5150.28	122.45	23.6	1420.34	29.85
0.3	**94.0**	**795.68**	**18.23**	**87.2**	**1984.87**	**49.43**	**45.6**	**8863.90**	**126.04**	**53.0**	**4892.38**	**115.26**	**24.3**	**1349.63**	**28.05**
0.4	93.5	840.42	19.42	86.5	2120.58	51.38	45.1	9010.18	130.36	52.4	5010.48	118.58	24.1	1380.42	28.58
0.49	92.9	1180.28	32.15	85.8	2850.42	68.24	43.6	11500.64	193.28	52.0	6850.18	158.45	23.5	2150.78	45.15

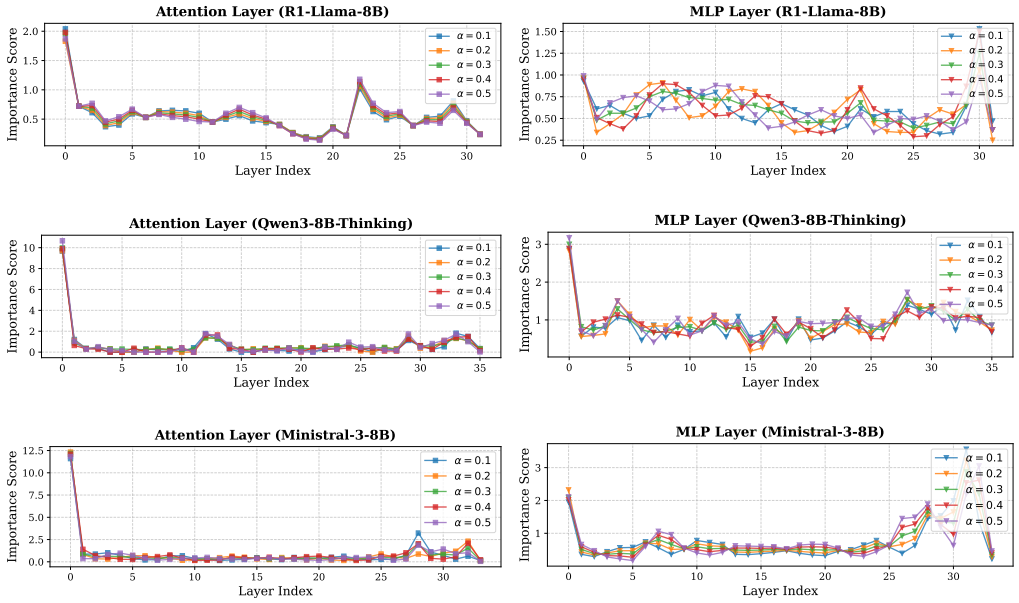


Figure 10: Layer importance scores for each base model across different values of α. The value $\alpha = 0.3$ is adopted in the main experiments.

relatively significant layers, which consequently compromises the capability of the auxiliary model and degrades the output quality of the overall framework.

Pruning ratio ρ. The value of ρ determines the final number of layers removed during the pruning process. With α fixed at 0.3, we conduct a sensitivity analysis on this parameter. As shown in Table 6, $\rho = 0.12$ consistently achieves the highest accuracy and the lowest *TC* and *Lat* across all three base models and five benchmarks. A smaller ratio of 0.08 yields modest but consistent degradation, suggesting that insufficient pruning fails to establish adequate functional differentiation between the two models. In contrast, increasing ρ beyond 0.12 progressively reduces accuracy while increasing token and latency costs. At $\rho = 0.24$, for example, accuracy on AIME26 decreases by 16.7 to 22.1 percentage points relative to $\rho = 0.12$. This indicates that excessive pruning impairs

Table 6: Model performance based on different values of ρ. The best results are highlighted in bold.

Values of ρ	GSM8K			MATH500			AIME26			GPQA-D			BBEH		
	Acc ↑	*TC* ↓	*Lat* ↓	*Acc* ↑	*TC* ↓	*Lat* ↓	*Acc* ↑	*TC* ↓	*Lat* ↓	*Acc* ↑	*TC* ↓	*Lat* ↓	*Acc* ↑	*TC* ↓	*Lat* ↓
R1-Distill-Llama-8B															
0.08	84.1	232.18	6.14	81.2	1310.48	24.63	45.7	3610.24	70.82	47.1	2635.38	41.47	18.2	992.34	19.13
0.12	**85.0**	**200.64**	**5.21**	**82.3**	**1137.73**	**20.66**	**46.1**	**3420.68**	**66.45**	**48.3**	**2281.56**	**35.03**	**19.1**	**856.25**	**16.09**
0.16	84.6	224.35	5.82	81.7	1380.26	25.14	44.9	4210.53	83.42	47.8	2680.18	41.64	18.7	980.48	18.52
0.20	81.9	285.64	7.48	76.8	1820.46	34.23	39.5	5890.34	118.52	44.6	3380.52	53.14	16.5	1290.83	25.24
0.24	71.5	415.26	11.24	62.4	2980.54	61.43	28.6	9450.82	204.65	36.2	5120.35	88.54	13.0	1950.43	40.82
Qwen3-8B-Thinking															
0.08	88.0	236.44	4.12	82.1	1395.24	25.43	47.2	4480.63	80.54	49.4	2290.45	43.38	20.6	1025.34	12.28
0.12	**88.2**	**229.30**	**3.97**	**82.3**	**1342.75**	**24.32**	**49.1**	**3826.71**	**67.85**	**49.7**	**2213.65**	**41.77**	**20.8**	**978.52**	**11.66**
0.16	87.5	282.18	5.08	81.1	1720.64	31.83	47.8	4850.36	87.24	48.6	2760.53	52.84	19.8	1240.26	15.14
0.20	84.6	355.82	6.47	77.9	2280.44	43.15	42.1	6720.54	126.83	45.3	3620.85	71.42	17.6	1510.64	19.23
0.24	73.8	495.63	9.48	65.1	3540.84	72.85	32.4	10850.46	221.53	39.5	5480.24	118.62	14.5	2260.54	30.25
Ministral-3-8B-Reasoning															
0.08	93.7	842.18	19.42	87.0	2145.63	53.25	44.5	9720.43	140.54	52.4	5240.65	124.32	23.9	1465.24	30.83
0.12	**94.0**	**795.68**	**18.23**	**87.2**	**1984.87**	**49.43**	**45.6**	**8863.90**	**126.04**	**53.0**	**4892.38**	**115.26**	**24.3**	**1349.63**	**28.05**
0.16	93.3	965.42	22.38	86.5	2520.15	65.82	44.1	11350.84	168.25	51.8	6120.54	148.63	23.5	1680.45	36.24
0.20	90.5	1160.84	27.63	82.9	2953.42	76.18	38.2	13620.45	208.54	48.7	7180.35	181.42	21.1	2050.84	44.83
0.24	78.4	1720.53	43.82	70.1	4620.75	128.42	23.5	15120.64	235.82	40.8	10650.43	282.65	15.8	2940.54	68.43

Table 7: Model performance based on different values of τ. The best results are highlighted in bold.

Values of τ	GSM8K			MATH500			AIME26			GPQA-D			BBEH		
	Acc ↑	*TC* ↓	*Lat* ↓	*Acc* ↑	*TC* ↓	*Lat* ↓	*Acc* ↑	*TC* ↓	*Lat* ↓	*Acc* ↑	*TC* ↓	*Lat* ↓	*Acc* ↑	*TC* ↓	*Lat* ↓
R1-Distill-Llama-8B															
0.02	84.2	230.15	5.96	81.2	1386.42	25.38	45.1	4475.63	88.54	47.3	2618.37	40.48	18.3	1018.46	19.27
0.04	**85.0**	200.64	5.21	**82.3**	1137.73	20.66	**46.1**	3420.68	66.45	**48.3**	2281.56	35.03	**19.1**	856.25	16.09
0.06	83.8	183.74	4.76	80.7	1028.36	18.42	44.6	3124.85	59.73	46.8	2092.64	31.75	18.2	783.52	14.58
0.08	82.7	**167.82**	**4.32**	79.5	**914.68**	**16.18**	43.5	**2614.73**	**51.36**	45.8	**1894.27**	**28.46**	17.4	**712.85**	**13.16**
Qwen3-8B-Thinking															
0.02	87.4	254.38	4.38	81.1	1618.74	29.46	47.6	4876.52	86.42	48.6	2714.83	51.58	19.9	1143.67	13.58
0.04	**88.2**	229.30	3.97	**82.3**	1342.75	24.32	**49.1**	3826.71	67.85	**49.7**	2213.65	41.77	**20.8**	978.52	11.66
0.06	86.9	207.64	3.58	80.5	1184.26	21.28	47.2	3385.47	59.36	48.3	1994.58	37.35	19.6	887.43	10.48
0.08	85.9	**186.42**	**3.24**	79.2	**1038.75**	**18.57**	45.9	**2846.38**	**49.74**	47.1	**1754.62**	**32.76**	18.8	**793.86**	**9.38**
Ministral-3-8B-Reasoning															
0.02	93.3	923.64	21.16	86.1	2418.52	60.38	43.8	11416.78	162.46	51.9	5784.63	136.38	23.2	1584.27	33.15
0.04	**94.0**	795.68	18.23	**87.2**	1984.87	49.43	**45.6**	8863.90	126.04	**53.0**	4892.38	115.26	**24.3**	1349.63	28.05
0.06	92.8	723.84	16.48	85.6	1764.38	43.58	43.9	8056.42	114.54	51.5	4386.74	103.18	22.8	1214.58	24.86
0.08	91.8	**652.74**	**14.76**	84.3	**1514.62**	**37.18**	42.4	**6784.36**	**95.74**	50.2	**3876.54**	**90.76**	22.1	**1064.38**	**21.75**

the exploratory coherence of the auxiliary model, resulting in longer reasoning trajectories and more frequent corrective intervention by the base model.

Uncertainty threshold τ. The parameter τ governs the accessibility of the storage window for intermediate answers. As shown in Table 7, a low threshold of $\tau = 0.02$ prevents intermediate answers from entering the storage window, prolonging the reasoning process and increasing both *TC* and *Lat* without improving accuracy. Increasing τ to 0.04 consistently yields the highest accuracy across all three base models and five benchmarks while substantially reducing inference costs. However, further increasing τ makes the storage window overly permissive, enabling earlier termination but admitting less reliable intermediate answers. At $\tau = 0.08$, although *TC* and *Lat* are minimized, accuracy decreases by 1.7 to 3.2 percentage points relative to $\tau = 0.04$.

Counter upper bound $\mathcal{U}$. The selection of $\mathcal{U}$ determines the frequency of intervention by the base model to adjust the reasoning direction. We conduct experiments across different $\mathcal{U}$ values, as summarized in Table 8. Setting $\mathcal{U} = 1$ implies that any uncertainty within the intermediate answers triggers a redirection. This excessive adjustment leads to over-optimization, which adversely impacts performance. In contrast, $\mathcal{U} = 9$ results in an insufficient frequency of rectification, failing to effectively suppress the deviation from the appropriate reasoning trajectory.

Table 8: Model performance based on different values of $\mathcal{U}$. The best results are highlighted in bold.

Values of $\mathcal{U}$	GSM8K			MATH500			AIME26			GPQA-D			BBEH		
	Acc ↑	*TC* ↓	*Lat* ↓	*Acc* ↑	*TC* ↓	*Lat* ↓	*Acc* ↑	*TC* ↓	*Lat* ↓	*Acc* ↑	*TC* ↓	*Lat* ↓	*Acc* ↑	*TC* ↓	*Lat* ↓
R1-Distill-Llama-8B															
1	82.8	234.18	5.94	79.4	1294.62	23.87	43.6	3641.83	70.16	45.7	2627.42	41.84	17.3	932.67	17.65
3	**85.0**	**200.64**	**5.21**	**82.3**	**1137.73**	**20.66**	**46.1**	**3420.68**	**66.45**	**48.3**	**2281.56**	**35.03**	**19.1**	**856.25**	**16.09**
5	83.9	211.43	5.54	81.2	1214.86	22.18	45.2	3546.72	68.92	47.1	2408.53	37.46	18.4	889.34	16.88
7	83.1	226.74	5.89	80.5	1287.35	23.64	44.4	3662.48	71.53	46.3	2564.19	40.72	17.8	931.82	17.76
9	81.7	261.37	6.82	78.6	1448.63	26.94	42.5	4068.74	81.15	44.8	2873.56	46.28	16.7	1067.49	20.63
Qwen3-8B-Thinking															
1	86.6	251.36	4.43	80.4	1478.64	27.58	46.8	4118.25	72.46	47.9	2423.87	46.38	19.4	1056.73	13.24
3	**88.2**	**229.30**	**3.97**	**82.3**	**1342.75**	**24.32**	**49.1**	**3826.71**	**67.85**	**49.7**	**2213.65**	**41.77**	**20.8**	**978.52**	**11.66**
5	87.3	241.62	4.19	81.3	1412.58	25.96	47.9	4023.84	71.28	48.7	2318.46	44.25	19.9	1016.38	12.37
7	86.8	255.84	4.48	80.6	1489.37	27.42	47.3	4156.92	73.34	48.1	2442.73	46.78	19.5	1054.67	12.94
9	85.4	282.47	4.96	79.1	1664.82	30.73	45.8	4562.38	81.42	46.7	2714.65	52.34	18.6	1168.27	14.56
Ministral-3-8B-Reasoning															
1	92.5	856.42	19.94	85.1	2194.67	52.48	43.1	9234.82	135.63	51.3	5143.76	120.84	22.9	1428.54	29.96
3	**94.0**	**795.68**	**18.23**	**87.2**	**1984.87**	**49.43**	**45.6**	**8863.90**	**126.04**	**53.0**	**4892.38**	**115.26**	**24.3**	**1349.63**	**28.05**
5	93.2	821.64	19.16	86.1	2058.43	51.18	44.3	9176.48	131.24	52.1	4987.52	118.92	23.5	1384.72	28.94
7	92.7	852.78	19.74	85.5	2137.92	52.86	43.7	9584.26	137.58	51.6	5176.84	123.46	23.1	1421.68	29.72
9	91.6	984.62	22.95	84.3	2397.48	60.34	38.9	11462.83	169.25	50.4	5724.37	137.64	22.3	1563.82	33.27

Table 9: Model performance based on different value combinations of ω / ξ. The best results are highlighted in bold.

ω/ξ	GSM8K			MATH500			AIME26			GPQA-D			BBEH		
	Acc ↑	*TC* ↓	*Lat* ↓	*Acc* ↑	*TC* ↓	*Lat* ↓	*Acc* ↑	*TC* ↓	*Lat* ↓	*Acc* ↑	*TC* ↓	*Lat* ↓	*Acc* ↑	*TC* ↓	*Lat* ↓
R1-Distill-Llama-8B															
3 / 0.5	83.4	**162.38**	**4.15**	80.6	**915.24**	**16.38**	43.8	**2740.18**	**52.84**	46.8	**1845.32**	**27.95**	17.6	**692.14**	**12.80**
5 / 0.5	84.5	186.45	4.82	81.7	1048.60	18.92	45.3	3180.42	61.50	47.7	2110.48	32.40	18.5	795.36	14.88
5 / 0.7	**85.0**	200.64	5.21	**82.3**	1137.73	20.66	**46.1**	3420.68	66.45	**48.3**	2281.56	35.03	**19.1**	856.25	16.09
6 / 0.6	84.3	235.82	6.14	81.5	1324.38	24.35	45.6	4085.30	80.12	47.9	2690.15	41.65	18.6	1012.45	19.34
6 / 0.8	82.9	278.46	7.32	79.8	1512.65	28.14	43.2	4680.52	93.45	46.2	3045.74	47.80	17.1	1158.30	22.46
Qwen3-8B-Thinking															
3 / 0.5	87.5	**204.18**	**3.52**	81.5	**1184.62**	**21.40**	46.9	**3345.80**	**59.26**	48.8	**1925.43**	**36.42**	20.2	**864.25**	**10.35**
5 / 0.5	87.9	212.50	3.68	82.0	1245.38	22.56	48.2	3520.14	62.48	49.2	1958.62	37.34	20.6	908.70	10.82
5 / 0.7	**88.2**	229.30	3.97	**82.3**	1342.75	24.32	**49.1**	3826.71	67.85	**49.7**	2213.65	41.77	**20.8**	978.52	11.66
6 / 0.6	87.8	251.64	4.38	81.8	1485.20	27.15	48.5	4240.65	76.30	49.5	2440.34	48.15	20.4	1082.46	13.05
6 / 0.8	86.4	268.75	4.72	80.4	1572.34	29.18	46.7	4495.82	82.40	48.5	2562.72	52.27	19.3	1145.18	14.12
Ministral-3-8B-Reasoning															
3 / 0.5	92.8	**672.45**	**15.34**	85.8	**1695.30**	**41.80**	43.5	**7540.62**	**105.40**	51.7	**4210.58**	**98.60**	23.1	**1142.75**	**23.65**
5 / 0.5	93.6	735.80	16.75	86.6	1820.45	44.92	44.8	8150.34	115.20	52.4	4512.60	106.18	23.8	1238.40	25.74
5 / 0.7	**94.0**	795.68	18.23	**87.2**	1984.87	49.43	**45.6**	8863.90	126.04	**53.0**	4892.38	115.26	**24.3**	1349.63	28.05
6 / 0.6	93.5	874.15	20.60	86.8	2210.65	56.40	45.1	9820.45	144.80	52.6	5420.35	131.50	23.9	1495.20	32.10
6 / 0.8	92.4	956.80	23.15	85.3	2415.80	63.25	42.6	11240.50	172.60	50.8	5910.45	148.90	22.4	1625.30	36.40

Answer storage window size ω and acceptance threshold ξ. Given the inherent coupling between ω and ξ, we conduct ablation experiments across representative joint parameter configurations. As shown in Table 9, the configuration $(\omega, \xi) = (3, 0.5)$ enables earlier termination and achieves the lowest *TC* and *Lat*, but sacrifices 0.6 to 2.3 percentage points in accuracy relative to $(5, 0.7)$. Moderately increasing the window size and acceptance threshold strengthens the reliability of the consistency decision, with $(5, 0.7)$ consistently attaining the highest accuracy across all three base models and five benchmarks while maintaining moderate inference costs. Further tightening the criteria to $(6, 0.8)$ lengthens the reasoning trajectory but reduces accuracy by 1.2 to 3.0 percentage points, indicating that excessive consensus induces redundant reasoning.

Table 10: Performance comparison between auxiliary models and their respective base models. Note that *TPT* is measured in milliseconds (ms).

Model	GSM8K		MATH500		AIME26		GPQA-D		BBEH	
	Acc ↑	*TPT* ↓	*Acc* ↑	*TPT* ↓	*Acc* ↑	*TPT* ↓	*Acc* ↑	*TPT* ↓	*Acc* ↑	*TPT* ↓
aux-R1-Distill-Llama-8B	78.1	14.51	71.8	14.78	39.6	14.51	39.0	14.32	12.9	14.06
R1-Distill-Llama-8B	81.5	16.02	75.8	17.24	42.1	15.94	44.5	15.77	15.2	16.39
aux-Qwen3-8B	81.4	16.46	76.5	16.58	43.6	16.64	42.8	16.64	13.2	16.27
Qwen3-8B	84.1	18.08	79.5	18.16	45.2	18.25	46.1	18.22	16.6	18.35
aux-Ministral-3-8B	88.7	20.28	81.6	20.42	38.8	20.45	45.9	20.42	16.1	20.36
Ministral-3-8B	91.2	22.25	84.1	22.61	41.2	22.62	50.2	23.14	18.9	22.61

F Performance of the Auxiliary Model

Table 10 presents the accuracy and the average generation time per output token (*TPT*) for the auxiliary models across various benchmarks. The auxiliary models incur an average accuracy decrease of only 3.1 percentage points, indicating that they retain sufficient reasoning capability to support continuous exploration. Meanwhile, skipping layers consistently reduces the time per token across all model-benchmark pairs, achieving an average reduction of 10.2%. These results demonstrate that preference-guided pruning preserves the functional competence required by the auxiliary model.

G Analysis of KV Cache Sharing

We implement DeCoR using the vLLM framework (Kwon et al., 2023), where the detailed breakdown of KV cache sharing is as follows:

Warm-up Phase. The auxiliary model performs 10-step reasoning on the prompt, reusing KV cache $KV(P_t) = KV^{\text{auxi}}(P_t)$ under identical layer-skipping for historical prefixes P_t.

Initial Base Model Intervention. The base model prefills the original prompt and all warm-up tokens, satisfying $KV(P_t) = KV^{\text{base}}(P_t)$ across all layers.

Continued Auxiliary Model Reasoning. The auxiliary model reuses $KV^{\text{base}}(P_t)$ to enter a new decoding phase, generating subsequent segments via its own KV cache $KV^{\text{auxi}}(P_t)$.

Subsequent Base Model Intervention. The base model hits existing $KV^{\text{base}}(P_t)$ to prefill only the latest auxiliary segments and update $KV^{\text{base}}(P_t)$.

To better illustrate the cache sharing process, we present a simplified state transition flowchart in Figure 11. After auxiliary warm-up, the base model performs full-layer prefill to establish an authoritative checkpoint. Each auxiliary phase forks from this checkpoint by sharing valid prefix blocks through reference counting, while newly generated auxiliary states remain branch-local. Upon base intervention, only the latest auxiliary tokens are fully prefilled to update the checkpoint, which is subsequently reused for continued exploration or correction. Reasoning terminates when the auxiliary pathway reaches normal completion or when low-uncertainty intermediate answers satisfy the consistency criterion. We further conduct a non-sharing ablation experiment where auxiliary and base models maintain isolated caches, forcing full prefix re-prefilling at each intervention. We report four metrics: peak vLLM-allocated KV blocks (isolating cache footprint from dominant model parameters), the average number of recomputed tokens, the average reuse rate (fraction of valid context positions per intervention), and FLOPs. As shown in Table 11, KV cache sharing consistently achieves reuse rates ranging from 88.5% to 91.6% across the three base models. Compared with the variant without cache sharing, DeCoR reduces recomputed tokens by 89.6% to 91.3%, peak KV blocks by 35.5% to 44.4%, and FLOPs by 76.2% to 78.3%. Compared with the vanilla baseline, DeCoR incurs a moderate increase in FLOPs due to base model interventions. This additional computation aligns with our primary goal of eliminating redundant reasoning rather than minimizing arithmetic operations. Nonetheless, DeCoR consistently reduces both output tokens and wall-clock latency, as shown in Table 2. These results demonstrate that DeCoR retains the benefits of adaptive model collaboration while effectively avoiding substantial recomputation.

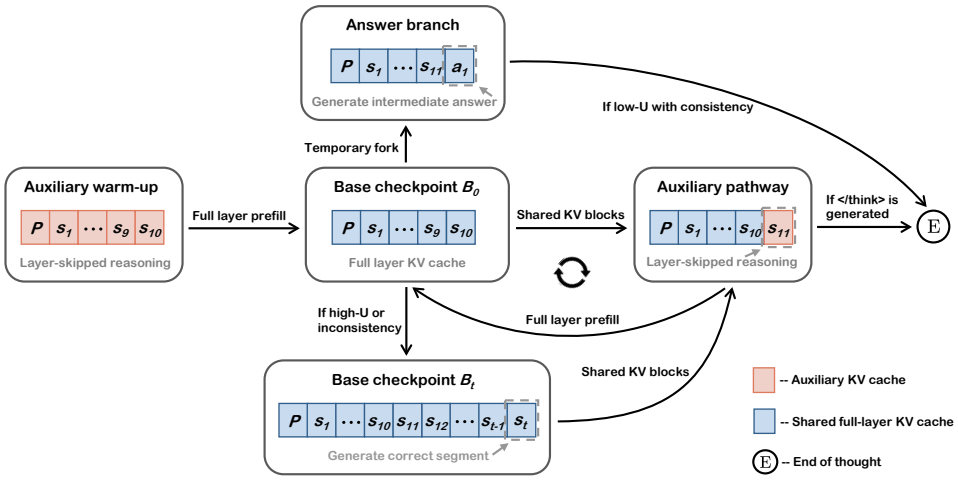


Figure 11: Illustration of the KV cache state transitions.

Table 11: Efficiency comparison of different KV cache configurations, averaged across five reasoning benchmarks.

Method	KV blocks (peak)	Recomputed tokens	Reuse rate (%)	FLOPs (T)
R1-Distill-Llama-8B				
Vanilla	152	0	–	43
DeCoR (w/o cache sharing)	207	13786.32	0	214
DeCoR	**115**	**1205.35**	**90.2**	**51**
Qwen3-8B-Thinking				
Vanilla	179	0	–	48
DeCoR (w/o cache sharing)	211	15847.03	0	267
DeCoR	**136**	**1644.26**	**88.5**	**57.9**
Ministral-3-8B-Reasoning				
Vanilla	305	0	–	99
DeCoR (w/o cache sharing)	379	28902.77	0	543
DeCoR	**232**	**2791.72**	**91.6**	**118**

Table 12: Comparison of inference-time peak memory footprint (GB). Subscripts denote the memory increment relative to the Vanilla model.

Base Model	Vanilla	Coconut	ARLCP	IBPO	ConPress	DTSR	TALE-EP	DEER	SoT	DeCoR
R1-Distill-Llama-8B	32.83	$32.85_{+0.02}$	$32.83_{+0.00}$	$32.83_{+0.00}$	$32.83_{+0.00}$	$32.84_{+0.01}$	$32.83_{+0.00}$	$32.84_{+0.01}$	$33.03_{+0.20}$	$32.83_{+0.00}$
Qwen3-8B-Thinking	32.83	$32.86_{+0.03}$	$32.83_{+0.00}$	$32.83_{+0.00}$	$32.83_{+0.00}$	$32.85_{+0.02}$	$32.83_{+0.00}$	$32.84_{+0.01}$	$33.01_{+0.18}$	$32.85_{+0.02}$
Ministral-3-8B-Reasoning	32.86	$32.89_{+0.03}$	$32.86_{+0.00}$	$32.86_{+0.00}$	$32.87_{+0.01}$	$32.88_{+0.02}$	$32.86_{+0.00}$	$32.87_{+0.01}$	$33.04_{+0.18}$	$32.86_{+0.00}$

H Memory Usage

Table 12 compares the inference-time peak memory usage of DeCoR with different baselines. Although DeCoR involves coordinated reasoning between the base and auxiliary pathways, it introduces virtually no additional memory overhead. Across the three base models, its peak memory remains unchanged for R1-Distill-Llama-8B and Ministral-3-8B-Reasoning, while increasing by only 0.02 GB for Qwen3-8B-Thinking. This efficiency stems from the single-model realization of DeCoR: the auxiliary pathway is instantiated through in-place masking within the base model rather than maintaining an independently loaded model, with both pathways sharing model parameters and a unified KV cache. Consequently, DeCoR achieves functional collaboration between two reasoning roles with negligible additional memory overhead compared with the vanilla backbone.

Table 13: Comparison with self-speculative decoding methods. The best results are highlighted in bold.

Method	MATH500			AIME26			GPQA-D		
	Acc ↑	*TC* ↓	*Lat* ↓	*Acc* ↑	*TC* ↓	*Lat* ↓	*Acc* ↑	*TC* ↓	*Lat* ↓
R1-Distill-Llama-8B									
Vanilla	75.8	1623.99	29.85	42.1	4566.75	84.13	44.5	3319.62	59.71
Draft & Verify	75.5	1566.42	28.60	42.2	4639.18	85.84	44.1	3288.40	61.35
SWIFT	75.4	1706.15	30.72	41.8	4519.33	81.50	44.6	3407.26	58.24
DeCoR	**82.3**	**1137.73**	**20.66**	**46.1**	**3420.68**	**66.45**	**48.3**	**2281.56**	**35.03**
Qwen3-8B-Thinking									
Vanilla	79.5	2061.87	33.94	45.2	5156.63	89.94	46.1	2977.31	57.43
Draft & Verify	79.1	2043.68	35.12	44.7	5218.40	87.63	46.0	2954.72	58.91
SWIFT	79.2	2094.57	32.80	45.1	5122.84	92.40	45.7	3016.39	56.75
DeCoR	**82.3**	**1342.75**	**24.32**	**49.1**	**3826.71**	**67.85**	**49.7**	**2213.65**	**41.77**
Ministral-3-8B-Reasoning									
Vanilla	84.1	2814.46	65.43	41.2	11465.27	240.29	50.2	6541.02	150.74
Draft & Verify	83.7	2861.82	66.80	40.8	11371.72	244.57	49.6	6617.48	147.20
SWIFT	84.2	2729.22	63.92	40.7	11524.38	236.80	49.9	6496.80	153.68
DeCoR	**87.2**	**1984.87**	**49.43**	**45.6**	**8863.90**	**126.04**	**53.0**	**4892.38**	**115.26**

I Comparison with Self-Speculative Decoding

We conduct a comparative evaluation of DeCoR against two self-speculative decoding methods, i.e., Draft & Verify (Zhang et al., 2024) and SWIFT (Xia et al., 2025). Throughout the experiments, DeCoR adopts the identical hyperparameter configurations detailed in the main text, while Draft & Verify and SWIFT are reproduced across the three base models following their official implementations. As shown in Table 13, Draft & Verify and SWIFT keep accuracy within -0.6 to $+0.1$ percentage points and output *TC* within -3.5% to $+5.1\%$ relative to vanilla decoding, while their latency ranges from a 4.2% reduction to a 3.5% increase, with half of the evaluated settings exhibiting a slight slowdown. One plausible explanation for this limited and inconsistent latency benefit is that reasoning LLMs are more sensitive to layer skipping, which may degrade the predictive quality of the draft pathway and reduce its agreement with the full model. A lower acceptance rate would increase the computation spent on rejected drafts and the verification work required per accepted token, offsetting the savings from layer skipping. Whereas Draft & Verify and SWIFT seek execution acceleration while preserving the original output distribution, DeCoR actively reshapes the reasoning trajectory through exploration, correction, and termination to mitigate overthinking. Therefore, its changes in accuracy, *TC*, and latency reflect trajectory regulation rather than decoding acceleration. **Notably, these two classes of methods can be viewed as complementary optimization approaches that target distinct dimensions of reasoning efficiency.** This opens up a promising direction that we leave as a subject for future research.

J Case Studies

We present four comparative reasoning cases between the vanilla model and DeCoR. Figures 12 and 13 illustrate the CoT sequences of the vanilla model and DeCoR in cases where both models yield the correct answer. The vanilla model arrives at the solution early while exhibiting unnecessary deliberation, exploring multiple redundant paths and ultimately converging to the initial conclusion. In contrast, DeCoR utilizes the uncertainty gating mechanism to steer the reasoning direction for rapid identification of the correct answer, and achieves timely termination through consistency gating, which substantially reduces the generation of verbose outputs. Figures 14 and 15 show examples where DeCoR successfully reaches the correct answer while the vanilla model fails. After extensive reasoning, the vanilla model still wanders between conflicting candidates or yields an incorrect answer. Conversely, DeCoR maintains a coherent reasoning trajectory. Even when the termination condition for consistency gating is not satisfied, the synergy between the two gating mechanisms stabilizes the reasoning direction, enabling the auxiliary or base model to actively derive the correct answer and conclude the reasoning process.

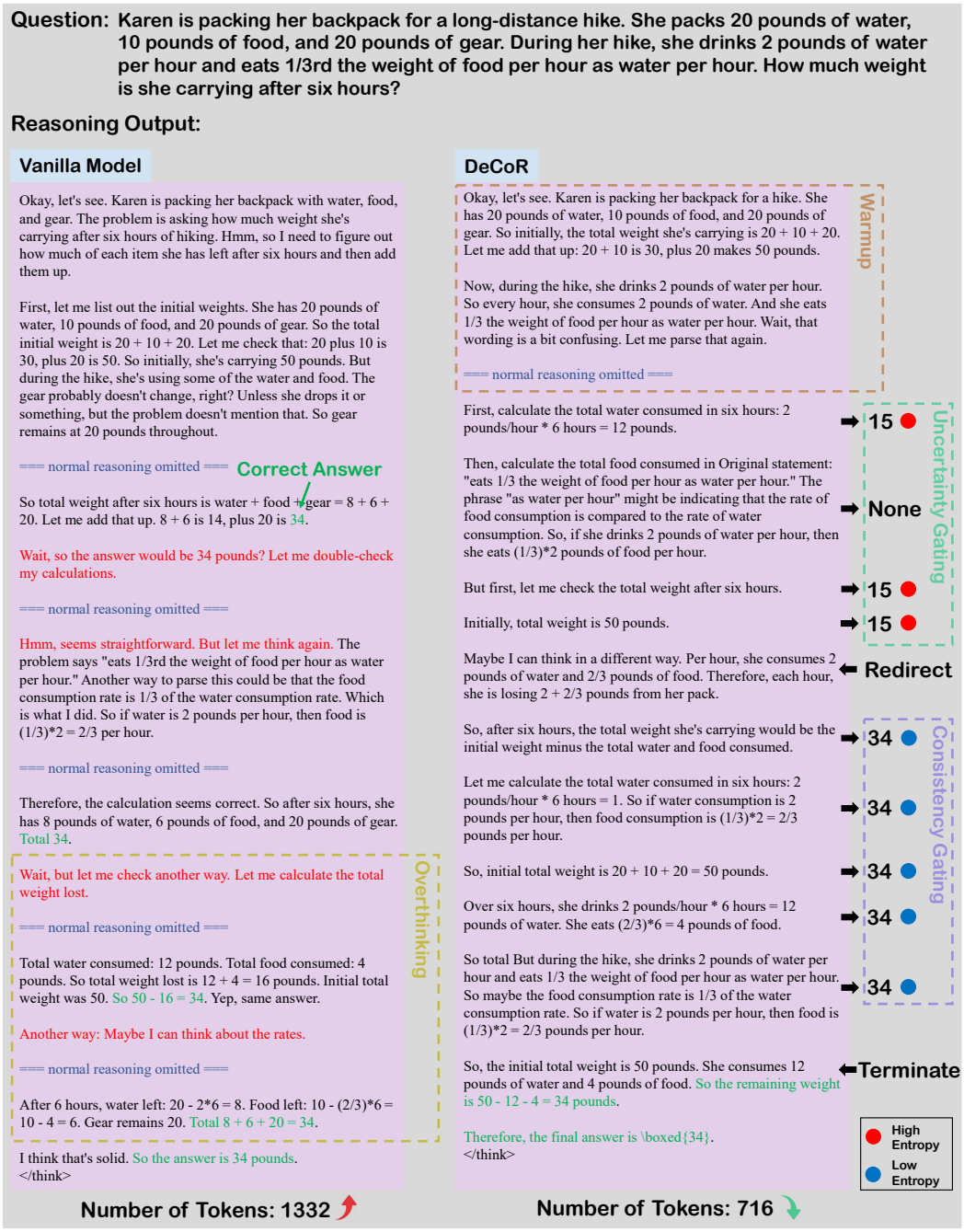


Figure 12: A comparative example of the vanilla model and DeCoR for a problem in GSM8K.

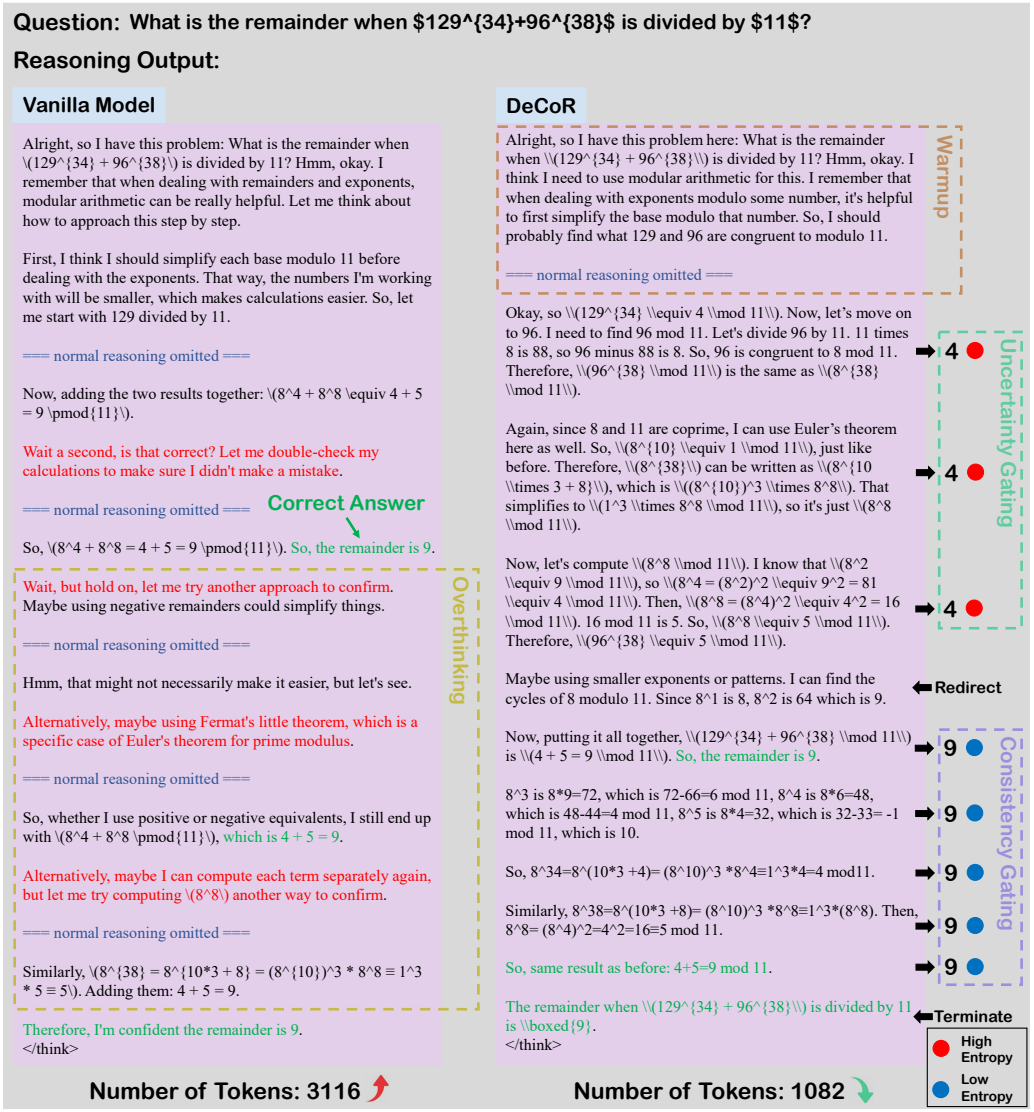


Figure 13: A comparative example of the vanilla model and DeCoR for a problem of the highest difficulty in MATH500.



Figure 14: A comparative example of the vanilla model and DeCoR for a problem in AIME26.

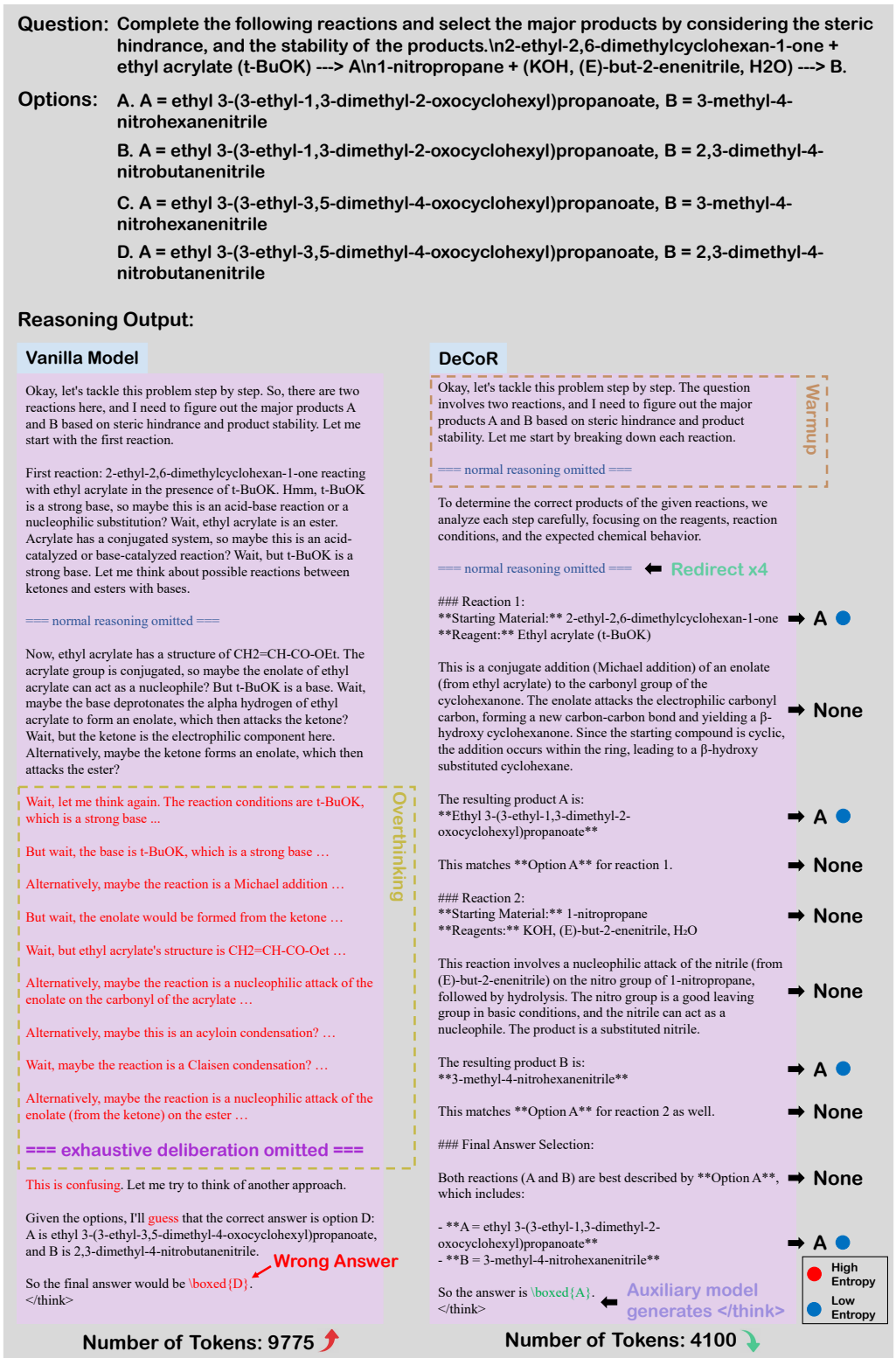


Figure 15: A comparative example of the vanilla model and DeCoR for a problem in GPQA-D.