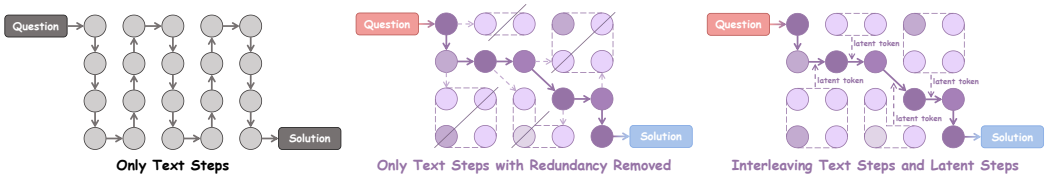
A*-THOUGHT-V2: EFFICIENT LATENT REASONING VIA GEOMETRIC DYNAMICS OF LLM

Anonymous authors
Paper under double-blind review

ABSTRACT

Chain-of-Thought (CoT) improves the reasoning ability of Large Language Models (LLMs) but incurs substantial computation and context costs. Existing methods either lose intermediate information through hard pruning or lack a principled criterion for continuous compression. We present **A*-Thought-V2**, a geometric dynamics of LLM guided framework that models CoT as a hidden-state trajectory and replaces hard deletion with an explicit–implicit interleaved latent architecture. After projecting question, step, and solution representations into a 3D PCA space, it measures alignment between each local transition and global question-to-solution direction. Aligned steps remain explicit text, whereas deviating steps are compressed into continuous latent tokens. Directional angles capture both local semantics and reasoning dynamics: small angles indicate direct execution and answer formation, while large angles more frequently involve checking, correction, and branch exploration; their temporal variation reveals exploration, convergence, and refinement stages. To train this architecture, we introduce stepwise embedding forcing, which pools each redundant step into a single latent embedding, and label forcing, which supervises that latent token with a soft multi-modal vocabulary distribution instead of a hard one-hot label. Experiments on Qwen3.5-9B and Qwen3.6-27B across six in-domain and out-of-domain benchmarks show that A*-Thought-V2 improves average accuracy by up to 2.6% while reducing response length by up to half, increasing Accuracy per Computation Unit by 2.29×, and reducing preprocessing and training time by 94.6% and up to 80.3%, respectively. Representation analyses suggest that latent states form a compact region distinct from textual states, while higher entropy at latent-token positions reflects broader soft targets that encourage richer step-level feature learning.



(a) Chain-of-Thought (b) A*-Thought (c) A*-Thought-V2

Figure 1: Comparison of reasoning paradigms. A*-Thought applies hard pruning, retaining selected steps and discarding others. A*-Thought-V2 encodes redundant steps as latent representations, forming an information-dense explicit–implicit interleaved sequence that preserves reasoning progression.

1 INTRODUCTION

Chain-of-Thought (CoT) enhances LLM reasoning but incurs computational and context costs (Muennighoff et al., 2025; Chen et al., 2026; Liu et al., 2026a). Existing compression methods adopt hard compression, retaining selected steps while discarding others (Xu et al., 2025; Jin et al., 2025; Xia et al., 2025), which may remove useful reasoning information. Latent reasoning instead encodes compressed steps into dense representations (Wu et al., 2025; Liu et al., 2026b; Zeng et al., 2026), but lacks an effective mechanism for selecting explicit steps and preserving the remaining information.

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.

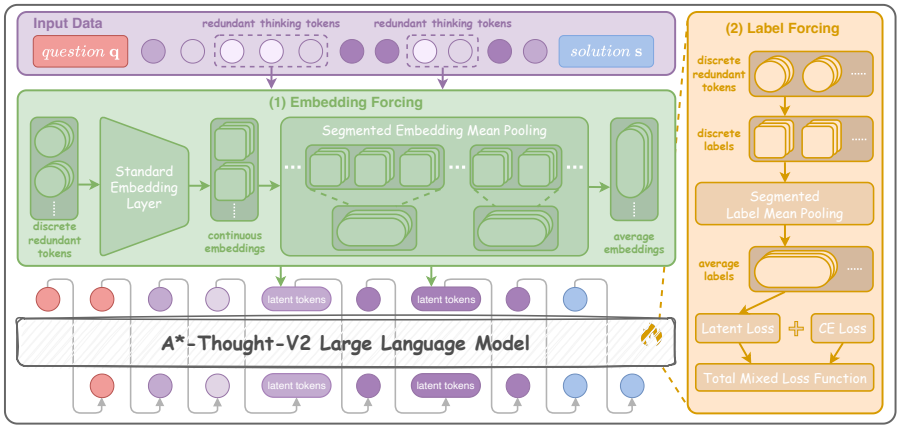


Figure 2: Framework of A*-Thought-V2. By heuristically constructing an explicit-implicit interleaved sequence, it compresses discrete redundant thinking tokens into continuous latent tokens via *Embedding Forcing* and supervises them via *Label Forcing*, integrating the latent loss with the standard cross-entropy loss.

In this paper, we introduce A*-Thought-V2, a dynamics-guided framework for information-preserving CoT compression with an explicit–implicit interleaved architecture. Unlike A*-Thought (Xu et al., 2025), which directly discards redundant steps, A*-Thought-V2 models reasoning as a trajectory in the LLM hidden-state space and encodes redundant steps in the latent space. We project the question, intermediate steps, and solution into a three-dimensional PCA space and measure the alignment between each local transition and the global question-to-solution direction. Our analysis identifies six 30° directional-angle intervals with semantic tendencies ranging from direct execution to correction and branch reconsideration. The temporal angle variation further characterizes three qualitative reasoning stages: exploration, convergence, and refinement. These geometric dynamics guide controllable explicit–latent step allocation and compression.

To learn the resulting interleaved sequence, we develop an SFT paradigm with *Embedding Forcing* and *Label Forcing*, as illustrated in Figure 2. Embedding Forcing converts compressed spans into segment-pooled latent embeddings, while Label Forcing supervises the corresponding latent positions with soft vocabulary distributions. Joint optimization of the latent and standard text losses encourages compact latent tokens to preserve the semantics of compressed reasoning.

The main contributions of A*-Thought-V2 are summarized as follows:

- We provide a geometric dynamics analysis of CoT trajectories that identifies six directional-angle intervals with distinct semantic tendencies and three reasoning stages, thereby guiding controllable and efficient data selection and compression.
- We develop an explicit–implicit interleaved latent architecture with step-wise embedding forcing and label forcing, replacing hard pruning with continuous latent representations supervised by step-level soft vocabulary distributions.
- Extensive experiments across two model scales and six benchmarks show that A*-Thought-V2 improves average accuracy by up to 2.6% and ACU by 2.29× over the baseline, while reducing compression time by 94.6% over A*-Thought. Ablation and representation analyses further validate the proposed latent mechanisms.

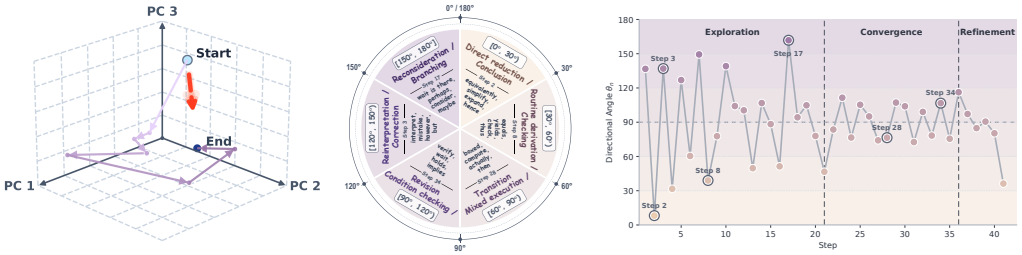
2 Preliminaries

For a T-step CoT, most existing compression methods adopt hard compression, preserving steps with $\text{flag}^{(n)} = 1$ while directly discarding those with $\text{flag}^{(n)} = 0$. A*-Thought Xu et al. (2025) searches for this binary flag sequence in a two-dimensional tree structure. In contrast, A*-Thought-V2 encodes steps with $\text{flag}^{(n)} = 0$ into latent representations to retain more reasoning information at

higher density, and models the reasoning process $\mathbf{q} \to \mathbf{t}^{(1)} \to \cdots \to \mathbf{t}^{(T)} \to \mathbf{s}$ as a trajectory in the PCA-projected hidden-state space, as illustrated in Figure 3(a). Using a frozen Qwen3.5-0.8B extractor by default, we mean-pool final-layer token hidden states within the question, each reasoning step, and the solution. We then apply PCA jointly to these vectors for each trajectory, obtaining the 3D representations $\mathbf{h}_q$, $\mathbf{h}_{t^{(n)}}$, and $\mathbf{h}_s$, respectively. We define the global question-to-solution direction as $\mathbf{z}_0 = \mathbf{h}_s - \mathbf{h}_q$, and the local transition directions as $\mathbf{z}_1 = \mathbf{h}_{t^{(1)}} - \mathbf{h}_q$ and $\mathbf{z}_n = \mathbf{h}_{t^{(n)}} - \mathbf{h}_{t^{(n-1)}}$ for $n = 2, \ldots, T$. We then quantify the directional alignment between each local transition and the global solution direction using their included angle:

$$\theta_n = \arccos\left(\frac{\mathbf{z}_n^\top \mathbf{z}_0}{\|\mathbf{z}_n\|\|\mathbf{z}_0\|}\right). \tag{1}$$

We analyze the PCA-projected CoT trajectory from both semantic and temporal perspectives. As shown in Figure 3(b), small-angle transitions are mainly associated with direct reduction and answer formation, whereas large-angle transitions more frequently involve checking, correction, reinterpretation, and branch reconsideration. Detailed interval proportions, semantic categories, and lexical markers are provided in Appendix B.1. Figure 3(c) further reveals three reasoning stages from the temporal variation of θ_n: exploration with strong oscillations, convergence with decreasing variation, and refinement before the final answer. The complete reasoning content corresponding to the marked steps is presented in Appendix B.2. Within each trajectory, 90° separates nonnegative from negative projected alignment, while 60° imposes a stricter alignment criterion, we therefore focus on these two thresholds in the main experiments.



(a) CoT Trajectory (b) Step Semantics (c) Reasoning Stages

Figure 3: PCA-projected CoT trajectory and its directional dynamics. (a) Three-dimensional representation trajectory. (b) Semantic tendencies across six angle intervals. (c) Exploration, convergence, and refinement stages reflected by step-wise angle variations.

3 METHODOLOGY

In this section, we introduce the construction and supervision of the explicit–latent architecture of A*-Thought-V2, followed by its autoregressive decoding procedure.

3.1 LATENT ARCHITECTURE

To enable LLMs to compress and learn high-dimensional latent thinking representations, we model latent tokens by embedding forcing and label forcing.

Embedding Forcing In a standard LLM, a sequence of discrete input tokens $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ is mapped by the embedding layer into continuous dense vector representations. At the token-level, the embedding of i-th token is:

$$\mathbf{e}_i = W_E(\mathbf{x}_i) \in \mathbb{R}^d \tag{2}$$

where $W_E \in \mathbb{R}^{V \times d}$ is a learnable embedding matrix, V is the vocabulary and d is the hidden dimension. Then $\mathbf{e} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n)$ proceeds to the subsequent transformer block, resulting in the corresponding hidden states $\mathbf{h} = (\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n)$.

In A*-Thought-V2, to achieve both less information loss and higher information compression density, we map redundant thinking steps $\{\mathbf{t}^{(n)} \in \mathbf{t} \mid \text{flag}^{(n)} = 0\}$ into the latent space via embedding

forcing, transforming the previously hard-pruned steps into continuous latent thought vectors. For a redundant thinking step $\mathbf{t}^{(n)}$ consisting of $l^{(n)}$ discrete tokens, let its corresponding variable-length standard embedding sequence be $\mathbf{e}^{(n)} = (\mathbf{e}_1, \ldots, \mathbf{e}_{l^{(n)}})$. To preserve the chronological logic and alleviate the information bottleneck, we map this step-level embedding sequence into a single latent token through segmented mean pooling:

$$\mathbf{c}^{(n)} = \frac{1}{l^{(n)}} \sum_{j=1}^{l^{(n)}} \mathbf{e}_j^{(n)} \in \mathbb{R}^d, \tag{3}$$

where $\mathbf{c}^{(n)}$ is the step-level latent embedding corresponding to the step $\mathbf{t}^{(n)}$. At the span-level, the step-level latent embeddings corresponding to a redundant span of m consecutive reasoning steps are concatenated to form a latent span:

$$\mathbf{c} = \left(\mathbf{c}^{(1)}; \mathbf{c}^{(2)}; \ldots; \mathbf{c}^{(m)}\right) \in \mathbb{R}^{m \times d}. \tag{4}$$

The interleaved reasoning trajectory is constructed by concatenating the original embeddings of retained steps and the latent embeddings of compressed steps in their original order:

$$\mathbf{e}_{\mathbf{t}'} = \left(\widetilde{\mathbf{e}}^{(1)}; \ldots; \widetilde{\mathbf{e}}^{(T)}\right), \qquad \widetilde{\mathbf{e}}^{(n)} = \begin{cases} \mathbf{e}^{(n)}, & \text{flag}^{(n)} = 1, \\ \mathbf{c}^{(n)}, & \text{flag}^{(n)} = 0, \end{cases} \tag{5}$$

where $\mathbf{e}^{(n)}$ denotes the original token-embedding sequence of step n, $\mathbf{c}^{(n)}$ is its single-token latent representation when compressed, and semicolons denote concatenation along the sequence dimension. During supervised training, the complete teacher-forced embedding sequence is formed by concatenating the question, interleaved reasoning trajectory, and solution representations:

$$\mathbf{e}_{input} = (\mathbf{e}_\mathbf{q}; \mathbf{e}_{\mathbf{t}'}; \mathbf{e}_\mathbf{s}), \tag{6}$$

where $\mathbf{e}_\mathbf{q}$ and $\mathbf{e}_\mathbf{s}$ denotes the standard embeddings of the question $\mathbf{q}$ and solution $\mathbf{s}$. Through this streamlined embedding forcing mechanism, A*-Thought-V2 effectively compresses lengthy redundant text into dense latent spatial representations. This allows the model to retain crucial heuristic reasoning information as a continuous semantic flow, while significantly reducing the overall context window occupancy.

Label Forcing Standard LLMs are trained through next-token prediction using deterministic one-hot targets. Let $\hat{p}_{i,v}$ denote the predicted probability of vocabulary item v at token position i. For a sequence of length N, the standard cross-entropy loss is

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{v=1}^{V} y_{i,v} \log \hat{p}_{i,v}, \tag{7}$$

where $y_{i,v} \in \{0, 1\}$ is the one-hot target distribution.

In A*-Thought-V2, each latent token is expected to simultaneously encode the semantic information of multiple original tokens, requiring a multi-modal target distribution. Deterministic hard labels are therefore ill-suited for this objective. To supervise the compressed representations generated by embedding forcing, we introduce a step-wise label forcing mechanism using continuous soft probability distributions instead of discrete one-hot vectors.

Corresponding to the step-level compression strategy described in embedding forcing, we map all target tokens of a redundant step into a single multi-modal distribution. For the redundant step $\mathbf{t}^{(n)}$ consisting of $l^{(n)}$ tokens, the target soft label $\mathbf{y}_{soft}^{(n)} \in \mathbb{R}^V$ is constructed by averaging the one-hot encoded ground truth vectors of all tokens within that entire step:

$$\mathbf{y}_{soft}^{(n)} = \frac{1}{l^{(n)}} \sum_{j=1}^{l^{(n)}} \mathbf{y}_j^{(n)}. \tag{8}$$

This multi-modal probability distribution acts as a segment-wise soft vocabulary target, forcing the latent token to simultaneously predict the macro-semantics of its corresponding reasoning step. The

cross-entropy loss for this specific latent token, denoted as $\mathcal{L}_{latent}^{(n)}$, is then calculated between the latent output distribution and the step-wise soft target:

$$\mathcal{L}_{latent}^{(n)} = -\sum_{v=1}^{V} \mathbf{y}_{soft,v}^{(n)} \log \hat{p}_{n,v}, \tag{9}$$

where $\hat{p}_{n,v}$ is the model's predicted probability for the v-th vocabulary item at the latent position corresponding to step $\mathbf{t}^{(n)}$.

Finally, the overall training objective integrates the standard hard-label loss for regular text tokens and the soft-label loss for latent tokens. To ensure the model adequately learns this dense conceptual compression and strictly adheres to the latent structural boundaries, we apply a structural scaling weight λ to the latent loss. The final mixed loss is formulated as:

$$\mathcal{L} = \frac{1}{N_{valid}} \left(\sum_{i \in \mathbb{I}_{text}} \mathcal{L}_{CE}^{(i)} + \lambda \sum_{k \in \mathbb{I}_{latent}} \mathcal{L}_{latent}^{(k)} \right), \tag{10}$$

where $\mathbb{I}_{text}$ and $\mathbb{I}_{latent}$ denote the global index sets for text and latent tokens in the concatenated sequence respectively, and $N_{valid} = |\mathbb{I}_{text}| + |\mathbb{I}_{latent}|$ is the total number of valid tokens. The boundary tags `<latent>` and `</latent>` are included in $\mathbb{I}_{text}$ and receive one-hot supervision; under next-token alignment, the final latent state is trained to predict `</latent>`, whose generation ends latent mode and resumes explicit decoding. Through label forcing, we explicitly supervise the internal logic alignment, guiding the model to match its deep latent representations with the chronological semantic distribution of the heuristically pruned steps.

3.2 Inference

At inference, A*-Thought-V2 interleaves explicit text decoding with continuous latent-state updates. The tags `<latent>` and `</latent>` delimit each latent span. Discrete placeholders serialize latent positions, while their lookup embeddings are replaced by continuous input vectors.

Latent Sampling Embedding forcing supplies pooled step embeddings as teacher-forced latent inputs during training. During inference, latent inputs are obtained autoregressively from the model's preceding hidden states.

Once `<latent>` is generated at position i, its final-layer hidden state initializes the latent rollout, while the complete preceding context remains available through the autoregressive cache $\mathcal{C}_i$. At each latent position, the preceding final-layer hidden state is fed directly into the model's input-embedding interface:

$$\mathbf{u}_{i+j} = \mathbf{h}_{i+j-1}^{(L)}, \qquad \left(\mathbf{h}_{i+j}^{(L)}, \mathcal{C}_{i+j}\right) = \mathcal{T}(\mathbf{u}_{i+j}; \mathcal{C}_{i+j-1}), \qquad j = 1, \ldots, m, \tag{11}$$

where $\mathcal{T}$ denotes the Transformer computation with cached context and m is the generated latent-span length. Training and inference share the same Transformer and causal context mechanism, but use teacher-provided and model-generated latent inputs, respectively. When the vocabulary head predicts `</latent>` and the tag is emitted, the model exits latent mode and resumes standard text decoding.

4 Experiments

4.1 Setup

Backbones and Training Data To evaluate the scalability of the proposed method, we conduct experiments on two open-source reasoning backbones with different model sizes: Qwen3.5-9B[1], and Qwen3.6-27B[2]. This setup allows us to examine whether the proposed method remains effective across different capacity regimes. For training-based variants, we use OpenR1-Math-3k[3] as the supervised reasoning data.

[1] Qwen/Qwen3.5-9B
[2] Qwen/Qwen3.6-27B
[3] TeichAI/deepseek-v3.2-speciale-openr1-math-3k

Table 1: Main results. The best results are shown in **bold**, and the second-best results are underlined.

Methods	In-domain								Out-of-domain				Average		ACU
	Math500		AIME 2024		AIME 2025		AIME 2026		ARC-Challenge		GPQA-Diamond				
	Acc.	# Tok.	Acc.	# Tok.	Acc.	# Tok.	Acc.	# Tok.	Acc.	# Tok.	Acc.	# Tok.	Acc.	# Tok.	
Qwen3.5-9B	86.6	18453.63	85.0	35022.07	78.3	40545.69	84.2	37163.95	94.4	9859.94	81.8	29100.24	85.1	28357.59	0.30
Qwen3.5-9B w/ OpenR1-Math-3k	91.9	6918.05	90.8	21019.25	83.3	26583.55	87.5	25423.37	97.0	491.71	**83.5**	16510.11	89.0	16157.67	0.55
+ SwiReasoning	**97.8**	3392.95	90.0	25085.87	86.7	29413.63	90.0	27549.57	95.6	629.74	79.8	21099.01	90.0	17861.79	0.50
+ CopT	95.0	3829.28	**93.3**	23420.33	86.7	28707.10	86.7	30565.07	94.5	964.13	79.3	20054.05	89.3	17923.33	0.50
+ A*-Thought	89.9	6422.13	70.0	**18132.64**	60.8	25152.91	73.3	23899.95	96.8	468.63	81.3	**13030.06**	78.7	14517.72	0.54
+ A*-Thought-V2 w/ $\tau = 60°$	97.0	2918.68	91.7	19167.03	89.2	23509.88	90.0	21098.94	96.5	471.94	81.6	13891.47	91.0	13509.66	0.67
+ A*-Thought-V2 w/ $\tau = 90°$	96.9	**2743.97**	90.8	20266.38	**90.0**	**21765.58**	**91.7**	**20025.40**	**97.1**	**449.51**	82.8	15521.67	**91.5**	**13462.08**	**0.68**
Qwen3.6-27B	70.9	13911.97	84.2	29521.38	71.7	38445.60	77.5	36186.72	96.8	1157.41	85.5	21114.27	81.1	23389.56	0.35
Qwen3.6-27B w/ OpenR1-Math-3k	91.9	6926.81	95.8	16650.70	94.2	21203.67	92.5	25769.07	**97.4**	466.43	85.5	13126.60	92.9	14023.88	0.66
+ SwiReasoning	**97.8**	2679.25	90.0	19010.50	93.3	25873.40	**96.7**	25755.47	95.6	536.35	82.3	15469.61	92.6	14887.43	0.62
+ CopT	97.6	2979.45	96.7	19024.27	86.7	26306.73	90.0	26089.37	95.1	943.54	84.9	15470.21	91.8	15135.60	0.61
+ A*-Thought	88.0	8685.41	76.7	15864.80	65.8	**18373.82**	70.8	**18907.27**	96.9	473.21	84.0	**10149.79**	80.4	12075.72	0.67
+ A*-Thought-V2 w/ $\tau = 60°$	95.6	2585.44	95.9	17659.38	**95.0**	21320.58	91.7	20987.99	97.0	463.19	85.1	12405.54	93.4	12570.35	0.74
+ A*-Thought-V2 w/ $\tau = 90°$	96.7	**2496.34**	**96.7**	**14931.38**	93.3	20154.48	93.3	20622.56	97.3	**430.91**	**86.0**	12077.05	**93.9**	**11785.45**	**0.80**

Benchmarks We evaluate models on both in-domain and out-of-domain benchmarks. The in-domain benchmarks include Math500 (Lightman et al., 2024), AIME 2024, AIME 2025, and AIME 2026 (Art of Problem Solving); the out-of-domain benchmarks include ARC-Challenge (Clark et al., 2018) and GPQA-Diamond (Rein et al., 2024). Model performance is evaluated using the following metrics:

- *Accuracy*: The proportion of model outputs that match the ground-truth answers, measuring the model's correctness.
- *Length*: The average length, i.e., the number of generated tokens, of the model's response; longer responses typically incur higher inference costs.
- *Accuracy per Computation Unit* (ACU) (Ma et al., 2025): $\text{ACU} = 100 \times \text{Accuracy}/\text{Length}$, with accuracy expressed in percentage points. A larger value indicates a better performance and efficiency trade-off.

Baselines We compare our method against the following baselines:

- *SwiReasoning* (Shi et al., 2026a): A training-free method designed to improve the reasoning behavior of LRMs by interleaving text and latent mode.
- *CopT* (Shi et al., 2026b): A pipeline that uses continuous-space verifiers to refine draft answers through on-policy reflection and correction.
- *A*-Thought* (Xu et al., 2025): A search-guided efficient reasoning method that explores candidate reasoning trajectories through hard compression.

Training and Evaluation Details We use a unified training and inference protocol for all trainable variants, with full hyperparameters reported in Appendix D. In brief, models are trained for 3 epochs on 8 NVIDIA A100 80GB GPUs. For inference, we use temperature 1.0, top-p 0.95, we repeat each evaluation 4 times and report accuracy as the mean pass@1 score across runs and average response length. All remaining method-specific hyperparameters are reported in Tables 5 and 6.

4.2 Main Results

The detailed experimental results in Table 1 lead to the following key observations:

Up to 2.6% accuracy gain and 2.29× ACU improvement. A*-Thought-V2 improves average accuracy by up to 2.6 percentage points over same-data SFT across both model scales. For comparison, the 90° variant nearly halves response length and raises ACU from 0.35 to 0.80, corresponding to a 2.29× improvement over the original Qwen3.6-27B backbone. Similar gains are observed on Qwen3.5-9B.

Up to 16.0% shorter responses than supervised fine-tuning with higher accuracy. Compared with SFT on OpenR1-Math-3k, A*-Thought-V2 produces more accurate and shorter reasoning traces.

On Qwen3.6-27B, the 90° variant improves average accuracy by 1.0 points while reducing response length by 16.0%. It also surpasses SwiReasoning, CopT, and A*-Thought in average accuracy and ACU, showing the advantage of latent compression over training-free enhancement and hard pruning.

Semantically grounded thresholds generalize across domains. The 90° threshold preserves forward-aligned steps while compressing checking, correction, and exploration into latent tokens; 60° additionally compresses mixed execution and transition. Across the evaluated in-domain and out-of-domain benchmarks, both improve overall average accuracy and shorten responses relative to same-data SFT, with 90° offering the best overall trade-off. These results are consistent with the proposed semantic partition.

5 Analysis

5.1 Ablation on Step Selection and Latent Architecture

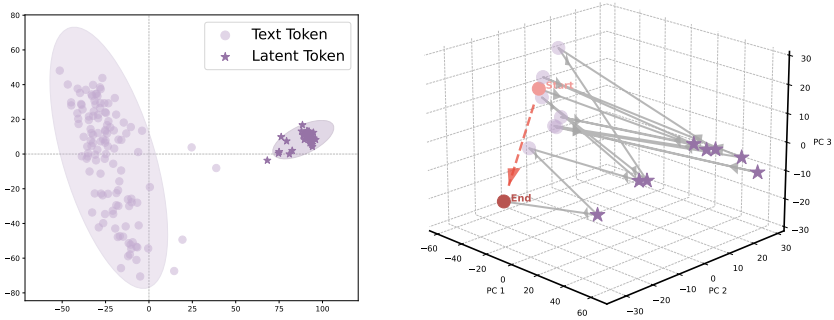
Table 2: Ablation study of geometric step selection, Embedding Forcing, and Label Forcing.

Methods	AIME 2024		AIME 2025		AIME 2026		Average		ACU
	Acc. (%)	# Tokens	Acc. (%)	# Tokens	Acc. (%)	# Tokens	Acc. (%)	# Tokens	
A*-Thought-V2-Qwen3.6-27B w/ $\tau = 90°$	**96.7**$_{\pm 0.0}$	**14931.38**$_{\pm 802.13}$	**93.3**$_{\pm 2.7}$	20154.48$_{\pm 993.78}$	**93.3**$_{\pm 2.7}$	20622.56$_{\pm 1018.66}$	**94.5**$_{\pm 0.9}$	**18569.47**$_{\pm 449.14}$	**0.51**
w/ Random Angles	94.2$_{\pm 2.7}$	16958.74$_{\pm 1374.64}$	90.8$_{\pm 4.2}$	22457.13$_{\pm 1534.80}$	93.3$_{\pm 1.7}$	20837.51$_{\pm 1063.29}$	92.8$_{\pm 1.7}$	20084.46$_{\pm 772.86}$	0.46
w/ Reversed Selection	93.3$_{\pm 2.7}$	18540.03$_{\pm 2055.31}$	90.8$_{\pm 1.7}$	22433.76$_{\pm 726.93}$	90.0$_{\pm 3.8}$	22990.10$_{\pm 1568.11}$	91.4$_{\pm 1.6}$	21321.30$_{\pm 895.15}$	0.43
w/o Embedding Forcing	94.2$_{\pm 3.3}$	18045.54$_{\pm 423.06}$	92.5$_{\pm 5.1}$	20259.82$_{\pm 1385.49}$	91.7$_{\pm 1.9}$	**18888.45**$_{\pm 1466.75}$	92.8$_{\pm 3.2}$	19064.60$_{\pm 655.20}$	0.49
w/o Label Forcing	76.7$_{\pm 7.2}$	16514.35$_{\pm 1295.32}$	68.4$_{\pm 8.4}$	**18763.83**$_{\pm 2328.92}$	72.5$_{\pm 5.7}$	21032.20$_{\pm 1933.23}$	72.5$_{\pm 4.2}$	18770.13$_{\pm 1271.71}$	0.39
w/o Embedding Forcing & Label Forcing	59.2$_{\pm 7.4}$	16053.92$_{\pm 416.88}$	55.0$_{\pm 1.9}$	22145.06$_{\pm 1267.15}$	70.8$_{\pm 5.0}$	19437.83$_{\pm 1535.73}$	61.7$_{\pm 2.6}$	19212.27$_{\pm 56.01}$	0.32

Effect of geometric step selection. We compare geometry-guided selection with random-angle and reversed-selection variants, where the latter compresses small-angle steps and retains large-angle steps as text. The compression rates of the three are 48.09%, 48.16%, and 47.15% respectively. Table 2 shows that geometry-guided selection achieves higher average accuracy and ACU than both variants while reducing average generation length by 7.54% and 12.91%, respectively. These results support the chosen direction of explicit–latent allocation over the tested alternatives.

Roles of Embedding Forcing and Label Forcing. Removing EF, LF, or both reduces average accuracy to 92.8%, 72.5%, and 61.7%, respectively. These results highlight LF's importance and EF's complementary benefits, supporting their combined use for latent training.

5.2 Representation Analysis of Latent Tokens



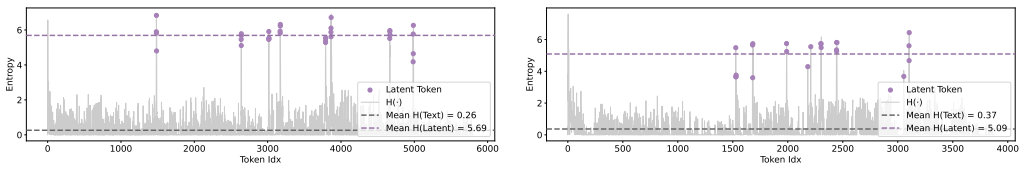
(a) 2D PCA (b) 3D PCA

Figure 4: PCA projections of text-token and latent-token hidden states produced by A*-Thought-V2-Qwen3.5-9B on AIME2024 case 79.

Figure 4 visualizes the hidden states of text and latent tokens. In the 2D projection, latent tokens form a compact cluster distinct from the broader text-token distribution. The 3D projection shows that they occupy a coherent intermediate region along the reasoning trajectory. This suggests that latent tokens encode structured compressed reasoning information. Additional cases appear in Appendix E.

5.3 Token Entropy Analysis under Label Forcing

Figure 5 shows higher predictive entropy at latent-token than explicit-text positions across both model scales. Explicit tokens use one-hot targets, whereas latent tokens use soft vocabulary distributions aggregated from compressed-step tokens. Therefore, the higher predictive entropy primarily indicates that the model fits these broader soft targets.

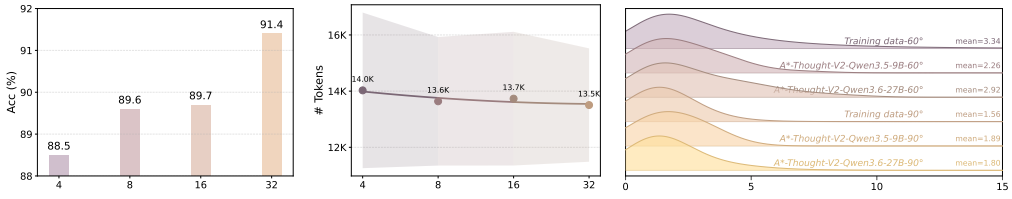


(a) Qwen3.5-9B (b) Qwen3.6-27B

Figure 5: The token entropy along the entire CoT trajectory.

5.4 Effect of Maximum Latent Length

We investigate the effect of the maximum latent length in Figure 6. Figure 6(a) and (b) show that a larger cap generally yields higher accuracy with shorter responses, suggesting that latent tokens compactly encode intermediate steps. Figure 6(c) shows that the latent-length distribution of decoded outputs closely matches that of the training data, indicating that the model has effectively learned the latent reasoning format.

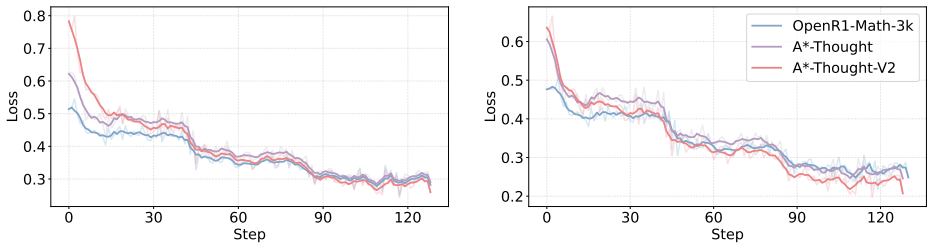


(a) Accuracy (b) Length (c) Latent Length Distribution

Figure 6: Ablation study of max latent length on AIME2024 and GPQA-Diamond.

5.5 Training Dynamics

Figure 7 compares the training losses of OpenR1-Math-3k, A*-Thought, and A*-Thought-V2. Although A*-Thought-V2 begins with a higher loss due to learning latent reasoning segments, it converges rapidly and achieves the lowest final loss at both model scales. This suggests that the proposed solution-oriented latent compression not only preserves the learnability of the reasoning process, but also provides a more effective training signal after convergence.



(a) Qwen3.5-9B (b) Qwen3.6-27B

Figure 7: Training-loss curves of different methods at two model scales.

5.6 Compression and Training Efficiency

Table 3 shows that A*-Thought-V2 reduces compression time from 5:16:22 to 0:16:57, a 94.6% reduction over A*-Thought. The 60° variant achieves the strongest compression rate (31.67%) and reduces training time by 80.3% and 68.7% for Qwen3.5-9B and Qwen3.6-27B, respectively. The 90° variant retains more reasoning tokens while still providing substantial training acceleration. These results demonstrate the efficiency and controllability of A*-Thought-V2 across model scales.

Table 3: Compression rates and preprocessing and training times across methods.

Methods	**Compression Rate**	**Compression Time**	**Train Time**	
			Qwen3.5-9B	**Qwen3.6-27B**
OpenR1-Math-3k	100.00%	–	12:21:18	21:24:52
+ A*-Thought	75.12%	5:16:22	5:34:40 (-54.9%)	10:26:22 (-51.3%)
+ A*-Thought-V2-60°	31.67%	0:16:57 (-94.6%)	2:25:56 (-80.3%)	6:42:07 (-68.7%)
+ A*-Thought-V2-90°	48.09%	0:16:57 (-94.6%)	3:47:22 (-69.3%)	10:18:49 (-51.8%)

6 Related Work

Efficient Reasoning Recent work improves reasoning efficiency by shortening explicit CoT trajectories (Jin et al., 2025; Muennighoff et al., 2025; Ma et al., 2025; Hou et al., 2025; Zhang et al., 2025; Luo et al., 2025; Yong et al., 2025; Zhao et al., 2025; Wu et al., 2026; Yang et al., 2026). TokenSkip (Xia et al., 2025) prunes low-importance tokens, while A*-Thought (Xu et al., 2025) combines bidirectional importance scoring with A* search to identify compact reasoning paths. However, such hard pruning may discard useful intermediate information. A*-Thought-V2 instead encodes redundant steps as dense latent representations and interleaves them with retained text, enabling higher-density and more information-preserving compression.

Latent Reasoning While traditional CoT improves LLM reasoning, generating lengthy discrete token sequences is computationally expensive. To address this limitation, recent studies have increasingly explored continuous and implicit reasoning paradigms (Hao et al., 2025; Shen et al., 2025; Wu et al., 2025; Wei et al., 2026; LCM team et al., 2024; Liu et al., 2026b; Du et al., 2026; Zhu et al., 2025; Qu et al., 2025; Tan et al., 2025). Specifically, SwiReasoning (Shi et al., 2026a) introduces a dynamic switching mechanism between explicit textual reasoning and latent thinking, achieving a Pareto-superior balance between performance and efficiency. CopT (Shi et al., 2026b) employs continuous-space verifiers to refine draft answers through on-policy reflection and correction. Following these paradigms, A*-Thought-V2 complements these methods with a trajectory-based criterion for explicit–latent step allocation and step-wise embedding and label forcing, enabling dynamic, variable-length, and information-preserving compression.

7 Conclusion

We presented A*-Thought-V2, a dynamics-guided framework that replaces hard CoT pruning with explicit–implicit interleaving. It models reasoning as a PCA-projected hidden-state trajectory, retaining solution-aligned steps as text and compressing deviating steps into latent representations. Directional-angle analysis links six intervals to recurring lexical patterns and characterizes exploration, convergence, and refinement. Step-wise embedding forcing and label forcing supervise latent learning with continuous embeddings and soft vocabulary targets. Across six benchmarks, A*-Thought-V2 improves average accuracy by up to 2.6% over same-data SFT, relative to the original backbones, the 90° variant nearly halves response length and improves ACU by up to $2.29\times$. It also reduces preprocessing time by 94.6% over A*-Thought and training time by up to 80.3% over standard SFT. Ablations support the complementary roles of both mechanisms. Representation analyses suggest that latent states form a compact region distinct from textual states, while higher predictive entropy reflects broader soft targets that encourage richer step-level feature learning. Future work will explore reasoning under reinforcement learning.

REFERENCES

Art of Problem Solving. American invitational mathematics examination. AoPS Wiki. URL https://artofproblemsolving.com/wiki/index.php/American_Invitational_Mathematics_Examination. Accessed 2026-09-03.

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. *Science China Information Sciences*, 69(6), May 2026. doi: 10.1007/s11432-025-4665-8. URL https://doi.org/10.1007/s11432-025-4665-8.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Hanwen Du, Yuxin Dong, and Xia Ning. Latent thinking optimization: Your latent reasoning language model secretly encodes reward signals in its latent thoughts. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 101000–101032, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/a4b1a25e1b37a17733c1fea50d0b7446-Paper-Conference.pdf.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason E Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=Itxz7S4Ip3.

Zhenyu Hou, Ziniu Hu, Yujiang Li, Rui Lu, Jie Tang, and Yuxiao Dong. TreeRL: LLM reinforcement learning with on-policy tree search. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12355–12369, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.604. URL https://aclanthology.org/2025.acl-long.604/.

Zhensheng Jin, Xinze Li, Yifan Ji, Chunyi Peng, Zhenghao Liu, Qi Shi, Yukun Yan, Shuo Wang, Furong Peng, and Ge Yu. ReCUT: Balancing reasoning length and accuracy in LLMs via stepwise trails and preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14269–14282, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.770. URL https://aclanthology.org/2025.findings-emnlp.770/.

LCM team, Loïc Barrault, Paul-Ambroise Duquenne, Maha Elbayad, Artyom Kozhevnikov, Belen Alastruey, Pierre Andrews, Mariano Coria, Guillaume Couairon, Marta R. Costa-jussà, David Dale, Hady Elsahar, Kevin Heffernan, João Maria Janeiro, Tuan Tran, Christophe Ropers, Eduardo Sánchez, Robin San Roman, Alexandre Mourachko, Safiyyah Saleem, and Holger Schwenk. Large concept models: Language modeling in a sentence representation space, 2024. URL https://arxiv.org/abs/2412.08821.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Yue Liu, Jiaying Wu, Yufei He, Ruihan Gong, Jun Xia, Liang Li, Hongcheng Gao, Hongyu Chen, Baolong Bi, Jiaheng Zhang, Zhiqi Huang, Bryan Hooi, Stan Z. Li, and Keqin Li. Efficient inference for large reasoning models: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2026a. doi: 10.1109/tpami.2026.3701928. URL https://doi.org/10.1109/tpami.2026.3701928.

Yuliang Liu, Yunchong Song, Yixuan Wang, Kewen Ge, Alex Lamb, Qipeng Guo, Kai Chen, Bowen Zhou, and Zhouhan Lin. Next concept prediction in discrete latent space leads to stronger language models, 2026b. URL https://arxiv.org/abs/2602.08984.

Haotian Luo, Haiying He, Yibo Wang, Jinluan Yang, Rui Liu, Naiqiang Tan, Xiaochun Cao, Dacheng Tao, and Li Shen. Ada-r1: Hybrid-cot via bi-level adaptive reasoning optimization. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 59353–59377. Curran Associates, Inc., 2025. doi: 10.52202/085713-1986. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/558da25ef295121e2d79209a95e45436-Abstract-Conference.html.

Xinyin Ma, Guangnian Wan, Runpeng Yu, Gongfan Fang, and Xinchao Wang. CoT-valve: Length-compressible chain-of-thought tuning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6025–6035, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.300. URL https://aclanthology.org/2025.acl-long.300/.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20275–20321, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1025. URL https://aclanthology.org/2025.emnlp-main.1025/.

Xingwei Qu, Shaowen Wang, Zihao Huang, Kai Hua, Fan Yin, Rui-Jie Zhu, Jundong Zhou, Qiyang Min, Zihao Wang, Yizhi Li, Tianyu Zhang, He Xing, Zheng Zhang, Yuxuan Song, Tianyu Zheng, Zhiyuan Zeng, Chenghua Lin, Ge Zhang, and Wenhao Huang. Dynamic large concept models: Latent reasoning in an adaptive semantic space, 2025. URL https://arxiv.org/abs/2512.24617.

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.

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 *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 677–693, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.36. URL https://aclanthology.org/2025.emnlp-main.36/.

Dachuan Shi, Abedelkadir Asi, Keying Li, Xiangchi Yuan, Leyan Pan, Wenke Lee, and Wen Xiao. Swireasoning: Switch-thinking in latent and explicit for pareto-superior reasoning LLMs. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 137060–137093, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/ddb7bad46132a323aa9d039000941881-Paper-Conference.pdf.

Dachuan Shi, Hanlin Zhu, Xiangchi Yuan, Wanjia Zhao, Kejing Xia, Wen Xiao, and Wenke Lee. Copt: Contrastive on-policy thinking with continuous spaces for general and agentic reasoning, 2026b. URL https://arxiv.org/abs/2605.20075.

Wenhui Tan, Jiaze Li, Jianzhong Ju, Zhenbo Luo, Ruihua Song, and Jian Luan. Think silently, think fast: Dynamic latent compression of LLM reasoning chains. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 4646–4668. Curran Associates, Inc., 2025. doi: 10.52202/085713-0164. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/0706261aedab63814a2b73c32564b4c4-Abstract-Conference.html.

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. SIM-CoT: Supervised implicit chain-of-thought. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 56721–56742,

2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/5d087955ee13fe9a7402eedec879b9c3-Paper-Conference.pdf.

Haoyi Wu, Zhihao Teng, and Kewei Tu. Parallel continuous chain-of-thought with jacobi iteration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 914–926, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.47. URL https://aclanthology.org/2025.emnlp-main.47/.

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 C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 127108–127138, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/ce916251b4fe04f54f99c8d304d68877-Paper-Conference.pdf.

Heming Xia, Chak Tou Leong, Wenjie Wang, Yongqi Li, and Wenjie Li. TokenSkip: Controllable chain-of-thought compression in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3351–3363, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.165. URL https://aclanthology.org/2025.emnlp-main.165/.

Xiaoang Xu, Shuo Wang, Xu Han, Zhenghao Liu, Huijia Wu, Peipei Li, Zhiyuan Liu, Maosong Sun, and Zhaofeng He. A*-thought: Efficient reasoning via bidirectional compression for low-resource settings. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 107714–107742. Curran Associates, Inc., 2025. doi: 10.52202/085713-3594. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/9ac5238b5c033fc49c2932a50df2bd47-Abstract-Conference.html.

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 C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 88170–88210, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/8df90a1440ce782d1f5607b7a38f2531-Paper-Conference.pdf.

Xixian Yong, Xiao Zhou, Yingying Zhang, Jinlin Li, Yefeng Zheng, and Xian Wu. Think or not? exploring thinking efficiency in large reasoning models via an information-theoretic lens. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 2787–2827. Curran Associates, Inc., 2025. doi: 10.52202/085713-0098. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/04185b5ae2d450ef39bd53c0ec4802cb-Paper-Conference.pdf.

Boyi Zeng, Shixiang Song, Siyuan Huang, Yixuan Wang, He Li, Ziwei He, Xinbing Wang, Zhiyu Li, and Zhouhan Lin. PonderLM: Pretraining language models to ponder in continuous space. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 91874–91893, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/94572a66c6b0710c5263f15d8ef6b463-Paper-Conference.pdf.

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. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 13307–13328, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.673. URL https://aclanthology.org/2025.emnlp-main.673/.

Haoran Zhao, Yuchen Yan, Yongliang Shen, Haolei Xu, Wenqi Zhang, Kaitao Song, Jian Shao, Weiming Lu, Jun Xiao, and Yueting Zhuang. Let LRMs break free from overthinking via self-braking tuning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main

Conference, pp. 1861–1887. Curran Associates, Inc., 2025. doi: 10.52202/085713-0069. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/02d028c3bf5d4875b5a4e4b4311087f7-Abstract-Conference.html.

Hanlin Zhu, Shibo Hao, Zhiting Hu, Jiantao Jiao, Stuart J. Russell, and Yuandong Tian. Reasoning by superposition: A theoretical perspective on chain of continuous thought. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 79931–79963. Curran Associates, Inc., 2025. doi: 10.52202/085713-2670. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/72c363c2a573ca2128bd176d3317696b-Abstract-Conference.html.

Algorithm 1 A*-Thought-V2 algorithm for compressing lengthy CoTs

Input:
$\mathbf{q}$: question; $\mathbf{t} = \langle \mathbf{t}^{(1)}, \ldots, \mathbf{t}^{(T)} \rangle$: segmented CoT; $\mathbf{s}$: solution
$\mathcal{H}$: step-level hidden-state extractor; W_E: embedding matrix; $\tau \in [0, \pi]$: angle threshold
$x_{\mathrm{BOL}} =$ <latent>, $x_{\mathrm{EOL}} =$ </latent>: latent boundary tags; $\epsilon > 0$: numerical tolerance
Output:
$\mathbf{E}_{\mathbf{t}'}$: explicit–latent embeddings; $\mathbf{Y}_{\mathbf{t}'}$: position-aligned hard/soft targets
$\mathbf{f} = \langle \mathrm{flag}^{(1)}, \ldots, \mathrm{flag}^{(T)} \rangle$: step-wise retention flags
Notation: $\mathrm{Tok}(x) = \big(W_E(x), \mathrm{onehot}(x)\big)$; $\oplus$ concatenates sequences of embedding–target pairs.
procedure A*-THOUGHT-V2
▷ (1) Hidden-state trajectory construction
$(\mathbf{r}_q, \mathbf{r}_{t^{(1)}}, \ldots, \mathbf{r}_{t^{(T)}}, \mathbf{r}_s) \leftarrow \mathcal{H}(\mathbf{q}, \mathbf{t}, \mathbf{s})$
$(\mathbf{P}, \boldsymbol{\mu}) \leftarrow \mathrm{PCA}_3(\mathbf{r}_q, \mathbf{r}_{t^{(1)}}, \ldots, \mathbf{r}_{t^{(T)}}, \mathbf{r}_s)$
for $x \in \{q, t^{(1)}, \ldots, t^{(T)}, s\}$ **do**
  $\mathbf{h}_x \leftarrow \mathbf{P}(\mathbf{r}_x - \boldsymbol{\mu})$
end for
$\mathbf{z}_0 \leftarrow \mathbf{h}_s - \mathbf{h}_q$
▷ (2) Dynamics-guided compression decisions
for $n = 1$ to T **do**
  if $n = 1$ **then**
    $\mathbf{z}_n \leftarrow \mathbf{h}_{t^{(1)}} - \mathbf{h}_q$
  else
    $\mathbf{z}_n \leftarrow \mathbf{h}_{t^{(n)}} - \mathbf{h}_{t^{(n-1)}}$
  end if
  if $\|\mathbf{z}_0\|_2 \|\mathbf{z}_n\|_2 \le \epsilon$ **then**
    $\mathrm{flag}^{(n)} \leftarrow 1$ ▷ Retain numerically degenerate transitions
  else
    $\rho_n \leftarrow \mathrm{clip}\Big(\frac{\mathbf{z}_n^\top \mathbf{z}_0}{\|\mathbf{z}_n\|_2 \|\mathbf{z}_0\|_2}, -1, 1\Big)$
    $\theta_n \leftarrow \arccos(\rho_n)$; $\mathrm{flag}^{(n)} \leftarrow \mathbb{I}[\theta_n \le \tau]$
  end if
end for
▷ (3) Explicit–latent sequence construction
$\mathcal{S}_{\mathbf{t}'} \leftarrow \langle\rangle$; $b \leftarrow 0$
for $n = 1$ to T **do**
  Write $\mathbf{t}^{(n)} = \langle x_1^{(n)}, \ldots, x_{l^{(n)}}^{(n)} \rangle$
  if $\mathrm{flag}^{(n)} = 0$ **then**
    if $b = 0$ **then**
      $\mathcal{S}_{\mathbf{t}'} \leftarrow \mathcal{S}_{\mathbf{t}'} \oplus \langle \mathrm{Tok}(x_{\mathrm{BOL}}) \rangle$; $b \leftarrow 1$
    end if
    $\mathbf{c}^{(n)} \leftarrow \frac{1}{l^{(n)}} \sum_{j=1}^{l^{(n)}} W_E(x_j^{(n)})$
    $\mathbf{y}_{soft}^{(n)} \leftarrow \frac{1}{l^{(n)}} \sum_{j=1}^{l^{(n)}} \mathrm{onehot}(x_j^{(n)})$
    $\mathcal{S}_{\mathbf{t}'} \leftarrow \mathcal{S}_{\mathbf{t}'} \oplus \Big\langle \big(\mathbf{c}^{(n)}, \mathbf{y}_{soft}^{(n)}\big) \Big\rangle$
  else
    if $b = 1$ **then**
      $\mathcal{S}_{\mathbf{t}'} \leftarrow \mathcal{S}_{\mathbf{t}'} \oplus \langle \mathrm{Tok}(x_{\mathrm{EOL}}) \rangle$; $b \leftarrow 0$
    end if
    $\mathcal{S}_{\mathbf{t}'} \leftarrow \mathcal{S}_{\mathbf{t}'} \oplus \Big\langle \mathrm{Tok}(x_1^{(n)}), \ldots, \mathrm{Tok}(x_{l^{(n)}}^{(n)}) \Big\rangle$
  end if
end for
if $b = 1$ **then**
  $\mathcal{S}_{\mathbf{t}'} \leftarrow \mathcal{S}_{\mathbf{t}'} \oplus \langle \mathrm{Tok}(x_{\mathrm{EOL}}) \rangle$
end if
Split $\mathcal{S}_{\mathbf{t}'}$ into $\mathbf{E}_{\mathbf{t}'}$ and $\mathbf{Y}_{\mathbf{t}'}$
return $\mathbf{E}_{\mathbf{t}'}$, $\mathbf{Y}_{\mathbf{t}'}$, $\mathbf{f}$
end procedure

A A*-THOUGHT-V2 ALGORITHM DETAIL

B GEOMETRIC AND DIRECTIONAL DYNAMICS ANALYSIS OF THE LLM REASONING REPRESENTATION TRAJECTORY

B.1 ANGLE THRESHOLD AND REASONING SEMANTIC ANALYSIS

Table 4: Semantic keyword groups and proportions associated with directional-angle intervals.

Angle Range	Share	Interval Semantics	Keywords
$[0°, 30°)$	7.57%	Direct reduction / Conclusion	`expand`, `substitute`, `hence`, `explain`, `simplify`, `overlooked`, `remaining`, `equivalently`, `we are asked`, `hours`, `answer boxed frac`, `is sqrt`, `wait compute`, `days`, `total so`, `initially`, `sum sum`
$[30°, 60°)$	18.46%	Routine derivation / Checking	`thus`, `case`, `yields`, `satisfies`, `equals`, `alternatively`, `check`, `boxed frac sqrt`, `answer frac`, `bp`, `for product`, `compute cos`, `alpha`, `invariant`, `to mod`, `sin cos sin`, `cos sin cos`
$[60°, 90°)$	24.72%	Mixed execution / Transition	`boxed`, `compute`, `solve`, `then`, `actually`, `evaluate`, `indeed`, `window`, `thus done`, `thus solution is`, `answer is correct`, `and boxed`, `decomposition`, `thus our`, `then that's`, `thus answer seems`, `answer stands`
$[90°, 120°)$	24.50%	Condition checking / Revision	`consistent`, `verify`, `holds`, `wait`, `wrong`, `unless`, `exclude`, `implies`, `given the problem`, `no further`, `i'll also`, `accept`, `answer yes`, `they might want`, `but i'll`, `implicitly`, `the blank`, `answer so`
$[120°, 150°)$	17.80%	Reinterpretation / Correction	`present`, `interpret`, `mistake`, `but`, `approximately`, `eliminate`, `rigorous`, `however`, `the interpretation`, `the condition holds`, `any additional`, `let's read the`, `at two points`, `an alternative`, `but they said`, `exactly what we`, `can simplify`, `but perhaps the`
$[150°, 180°]$	6.96%	Reconsideration / Branching	`consider`, `maybe`, `confirm`, `ensure`, `perhaps`, `reconsider`, `tangent to the`, `on side`, `is inside the`, `the lateral`, `is tangent to`, `touches`, `wait is there`, `sphere`, `wrote`, `intersecting`

To characterize how lexical patterns vary with trajectory direction, we partition thought steps into six predefined $30°$ directional-angle intervals and summarize the recurring lexical markers observed within each interval (Table 4). Lower-angle intervals contain execution and reduction markers, such as *hence*, *simplify*, *thus*, and *compute*, whereas higher-angle intervals contain checking, revision, and branch-exploration markers, such as *wait*, *however*, *maybe*, and *wait is there*. Figure 8 provides a visual reference for the six bins by displaying their within-bin mean angles. Together, the lexical patterns in Table 4 provide a descriptive view of how reasoning expressions vary across directional-angle intervals.

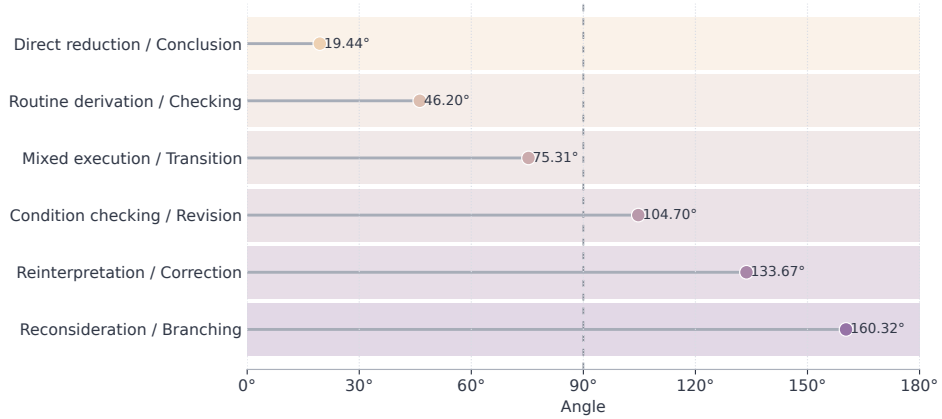


Figure 8: Visual summary of the six predefined 30° directional-angle intervals.

B.2 Case Study of Step-wise Directional Dynamics

Figure 9 expands the steps marked in Figure 3(c) into their original reasoning content. These six steps come from the same trajectory and cover consecutive 30° intervals. Smaller-angle steps primarily perform direct algebraic reduction, whereas larger-angle steps more frequently involve condition checking, reinterpretation, and alternative branches. The examples are ordered by angle interval rather than generation order, while their step indices indicate their original positions in the trajectory.

[0°, 30°) **Step 2** **Direct reduction / Conclusion** $\theta = 8.2^\circ$

Given f(x)= x^4 - 8x^3 + 14x^2 + a x. Write in terms of (x-c): Let y = x-c => x = y + c. Then f = (y+c)^4 - 8(y+c)^3 + 14(y+c)^2 + a(y+c). Expand and collect powers of y. Then the coefficients of odd powers y, y^3 must vanish for all y to have only even powers...

[30°, 60°) **Step 8** **Routine derivation / Checking** $\theta = 39.0^\circ$

Thus: (1) 4c - 8 = 0 => c = 2. (2) Coefficient of y^1: 4c^3 - 24c^2 + 28c + a = 0.

[60°, 90°) **Step 28** **Mixed execution / Transition** $\theta = 76.5^\circ$

But maybe we need to check if a = 0? Then f(x)=x^4-8x^3+14x^2 = x^2(x^2-8x+14). That is not symmetric about any vertical line. So no.

[90°, 120°) **Step 34** **Condition checking / Revision** $\theta = 106.7^\circ$

But wait: There's nuance: A quartic function may have two axes of symmetry if it's a biquadratic (even function). But only one vertical axis of symmetry is required.

[120°, 150°) **Step 3** **Reinterpretation / Correction** $\theta = 137.0^\circ$

But also note that there could be a linear term in y? Actually only even powers: y^4, y^2, constant. So coefficients for y^3, y^1 must be zero. Coefficient for y^3: from expansion:

[150°, 180°] **Step 17** **Reconsideration / Branching** $\theta = 161.8^\circ$

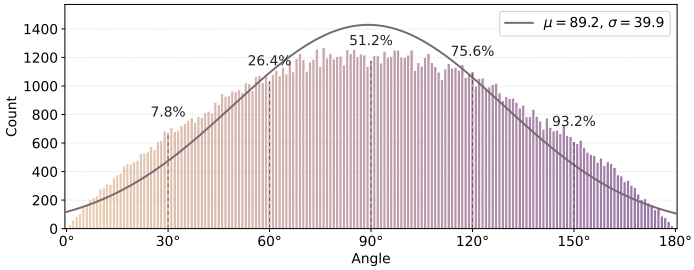
But wait: the original function had a term +ax. For a=8, it's +8x. That seems plausible.

Figure 9: Original reasoning content of the six steps marked in Figure 3(c), covering consecutive directional-angle intervals within the same trajectory.

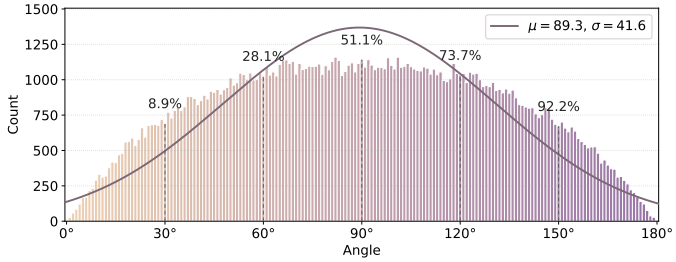
C Analysis

C.1 Effect of CoT-PCA Extractor Size

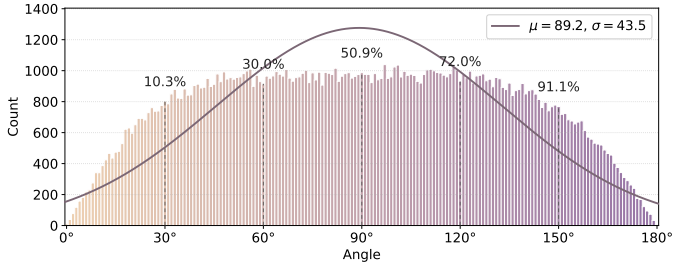
We investigate how extractor size affects the directional-angle distributions used by CoT-PCA. Specifically, we compare Qwen3.5 extractors with 0.8B, 2B, 4B, and 9B parameters. Figure 10 shows similar mean angles but differences in dispersion across extractors. We adopt Qwen3.5-0.8B, the smallest evaluated extractor, as the default to limit preprocessing cost.



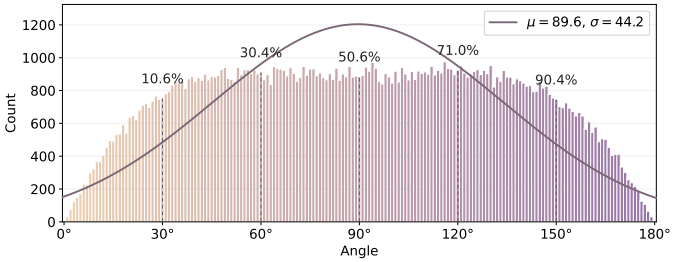
(a) Qwen3.5-0.8B



(b) Qwen3.5-2B



(c) Qwen3.5-4B



(d) Qwen3.5-9B

Figure 10: Directional-angle distributions produced by CoT-PCA extractors of different sizes.

D HYPERPARAMETERS

We report the training and inference configurations in Tables 5 and 6, respectively. All backbone models use the same optimization settings, while method-specific hyperparameters are configured separately.

Table 5: Training configuration.

Hyperparameter	**Value**			
Foundation				
Base model	Qwen3.5-9B		Qwen3.6-27B	
GPUs	8 × NVIDIA A100 (80GB)			
Device batch size	1			
Gradient accumulation steps	8			
Total batch size	64			
Cutoff length	20480			
Optimizer	AdamW			
Learning rate	1.0×10^{-5}			
Warmup ratio	0.1			
Epochs	3			
A*-Thought				
α	0.5			
β	0.1			
k_{min}	25			
k_{max}	40			
A*-Thought-V2 (Ours)				
Angle threshold	60°	90°	60°	90°
Latent weight λ	0.2	0.3	0.1	0.1
Number of added latent tokens	256			

Table 6: Inference configuration.

Hyperparameter	**Value**
Foundation	
Temperature	1.0
Top-p	0.95
Max tokens	81920
Repeat	4
A*-Thought-V2 (Ours)	
Max latent count	4
Max latent length	32

The latent weight λ controls the contribution of the latent-token objective. We tune λ through a grid search with a step size of 0.1 and monitor latent-token generation across benchmarks. A small λ may fail to activate latent generation, whereas a large value may produce excessive latent tokens and lead to overfitting. We therefore select the smallest value that reliably induces latent-token generation without excessive outputs.

E Case Study of the A*-Thought-V2 Trajectories

Question

Find the greatest integer n such that the cubic polynomial

$$x^3 - \frac{n}{6}x^2 + (n-11)x - 400$$

has roots α^2, β^2, and γ^2, where α, β, and γ are complex numbers, and there are exactly seven different possible values for $\alpha + \beta + \gamma$.

Output

<think>

Let r_1, r_2, r_3 be the roots, so $\alpha^2 = r_1$, $\beta^2 = r_2$, $\gamma^2 = r_3$. By Vieta's formulas,

$$r_1 + r_2 + r_3 = \frac{n}{6}, \qquad r_1r_2 + r_1r_3 + r_2r_3 = n - 11, \qquad r_1r_2r_3 = 400.$$

Since the product is nonzero, no root vanishes. Fixing square roots a, b, c of r_1, r_2, r_3, every admissible sum is $\alpha + \beta + \gamma = \pm a \pm b \pm c$, giving at most eight values. ...

The eight sign patterns form four pairs of opposites, so the set of sums is symmetric under negation. If two non-opposite patterns agreed, their opposites would agree too and at most six values would remain. Hence exactly seven values occur precisely when some sum vanishes, $\pm a \pm b \pm c = 0$, with all three coefficients nonzero. Squaring this relation gives the symmetric condition

$$(r_1 + r_2 + r_3)^2 = 4\,(r_1r_2 + r_1r_3 + r_2r_3)\,.$$

...

A relation with a vanishing coefficient does not produce seven values. If $a = -b$, the sums collapse to $\{\, c,\ -c,\ 2a \pm c,\ -2a \pm c \,\}$, that is, only six values, and $a = b$ leaves even fewer. So the vanishing combination must involve all three square roots, and after adjusting signs we may assume $a = b + c$. ...

Substituting the Vieta expressions yields

$$\left(\frac{n}{6}\right)^2 = 4(n-11) \implies n^2 - 144n + 1584 = 0 \implies n = \frac{144 \pm 120}{2},$$

so $n = 132$ or $n = 12$, and the greatest is 132.

Thus the answer should be 132.

<latent><latent_1><latent_2></latent>

as we did. So we have exactly seven different possible values for $\alpha + \beta + \gamma$. So $n = 132$ works.

<latent><latent_3><latent_4><latent_5><latent_6></latent>

This is exactly the condition.

<latent><latent_7><latent_8></latent>

For $n = 132$ the polynomial factors as $x^3 - 22x^2 + 121x - 400 = (x - 16)(x^2 - 6x + 25)$, with roots 16 and $3 \pm 4i$. Choosing $a = 4$, $b = -2 - i$, $c = -2 + i$ gives $a + b + c = 0$, and the eight signed sums are 8, $4 \pm 2i$, 0, $-4 \pm 2i$, -8 — exactly seven distinct values.

</think>

Final answer. The seven-sum condition is equivalent to $n^2 - 144n + 1584 = 0$, whose solutions are $n = 132$ and $n = 12$, so the greatest such integer is

$$\boxed{132}.$$

Figure 11: Case study of A*-Thought-V2-Qwen3.5-9B on AIME2026-ID25. Representative explicit reasoning is retained before ellipses, while all latent markers in the original output are retained and highlighted in purple.

Question

Alice and Bob play the following game. A stack of n tokens lies before them. The players take turns with Alice going first. On each turn, the player removes either 1 token or 4 tokens from the stack. Whoever removes the last token wins. Find the number of positive integers n less than or equal to 2024 for which there exists a strategy for Bob that guarantees that Bob will win the game regardless of Alice's play.

Output

<think>

This is an impartial game under normal play: from a stack of n tokens a player may move to $n-1$ or $n-4$, and the player taking the last token wins. Since Alice moves first, Bob has a winning strategy exactly when the starting position is losing for the player about to move, i.e. a P-position. So the task is to count P-positions with $n \le 2024$.

<latent><latent_1><latent_2><latent_3><latent_4><latent_5></latent>

The Grundy values satisfy $g(0) = 0$ and

$$g(n) = \operatorname{mex}\{\, g(n-1),\ g(n-4)\,\},$$

where the term $g(n-4)$ is present only for $n \ge 4$. Computing the first values gives

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$g(n)$	0	1	0	1	2	0	1	0	1	2	0	1	0	1	2	0

The pattern is periodic with period 5: $g(n) = 0$ when $n \equiv 0, 2$, $g(n) = 1$ when $n \equiv 1, 3$, and $g(n) = 2$ when $n \equiv 4 \pmod 5$. Hence the P-positions appear to be exactly $n \equiv 0$ or $2 \pmod 5$. ...

<latent><latent_6></latent>

The pattern can be confirmed by induction rather than trusted from a finite table. Classify $P = \{n \equiv 0, 2\}$ and $N = \{n \equiv 1, 3, 4\}$ modulo 5. From $n \equiv 0$ the two moves land on $n-1 \equiv 4$ and $n-4 \equiv 1$, both in N; from $n \equiv 2$ they land on $n-1 \equiv 1$ and $n-4 \equiv 3$, again both in N. Conversely, every N-position has a move into P: from $n \equiv 1$ or $n \equiv 3$ remove one token to reach $n \equiv 0$ or $n \equiv 2$, and from $n \equiv 4$ remove four tokens to reach $n \equiv 0$. These two properties are exactly the characterisation of P- and N-positions, so the classification is correct for all n. ...

Small cases agree with it: $n = 1, 3, 4$ are wins for Alice (for $n = 4$ she simply takes all four tokens), while $n = 2$ and $n = 5$ are wins for Bob. For $n = 5$ every Alice move leaves 4 or 1 tokens, and in both cases Bob finishes the stack.

<latent><latent_7><latent_8><latent_9></latent>

It remains to count the integers $1 \le n \le 2024$ with $n \equiv 0$ or $2 \pmod 5$. Since $2024 = 5 \cdot 404 + 4$, the multiples of 5 are $5, 10, \ldots, 2020$, giving $2020/5 = 404$ values, and the residues $2 \pmod 5$ are $2, 7, \ldots, 2022$, giving $(2022 - 2)/5 + 1 = 405$ values. The two families are disjoint, so the total is

$$404 + 405 = 809.$$

...

</think>

Final answer. Bob wins precisely when $n \equiv 0$ or $2 \pmod 5$. Among $n \le 2024$ there are 404 multiples of 5 and 405 integers congruent to 2, so the count is

$$404 + 405 = \boxed{809}.$$

Figure 12: Case study of A*-Thought-V2-Qwen3.6-27B on AIME2024-ID68. Representative explicit reasoning is retained before ellipses, while all latent markers in the original output are retained and highlighted in purple.

F Ethics Statement

This work studies the computational efficiency of language-model reasoning and does not introduce a new user-facing deployment. The method inherits the factuality, bias, privacy, and misuse risks of its underlying models and training data.

G Reproducibility Statement

The paper specifies the trajectory criterion, compression threshold, embedding and label objectives, inference mode switch, evaluation metrics, hardware, and method-specific hyperparameters. Tables 5 and 6 report the configurations used in the experiments, while the main results and ablations provide the corresponding accuracy, length, ACU, preprocessing-time, and training-time measurements.

H The Use of Large Language Models

A large language model was used to assist with language polishing. It did not generate or modify the experimental measurements, literature citations, or figure and table contents.