Limited Reasoning Space: The cage of long-horizon reasoning in LLMs

Zhenyu Li [1] **Guanlin Wu** [1] **Cheems Wang** [2] **Yongqiang Zhao** [3]

Abstract

The test-time compute strategy, such as Chain-of-Thought (CoT), has significantly enhanced the ability of large language models to solve complex tasks like logical reasoning. However, empirical studies indicate that simply increasing the compute budget can sometimes lead to a collapse in test-time performance when employing typical task decomposition strategies such as CoT. This work hypothesizes that reasoning failures with larger compute budgets stem from static planning methods, which hardly perceive the intrinsic boundaries of LLM reasoning. We term it as the *Limited Reasoning Space* hypothesis and perform theoretical anaylsis through the lens of a non-autonomous stochastic dynamical system. This insight suggests that there is an optimal range for compute budgets; over-planning can lead to redundant feedback and may even impair reasoning capabilities. To exploit the compute-scaling benefits and suppress over-planning, this work proposes Halo, a model predictive control framework for LLM planning. Halo is designed for long-horizon tasks with reason-based planning and crafts an entropy-driven dual controller, which adopts a *Measure-then-Plan* strategy to achieve controllable reasoning. Experimental results demonstrate that Halo outperforms static baselines on complex long-horizon tasks by dynamically regulating planning at the reasoning boundary.

1. Introduction

Although planning with the test-time compute stategy has significantly enhanced the capabilities of large language models (Wei et al., 2022; Kojima et al., 2022; Gema et al., 2025; Hassid et al., 2025; Huang et al., 2025), there is a critical mismatch between reasoning length and problem-solving capabilities: longer chains do not guarantee better performance (Motwani et al., 2025; Yang et al., 2025; Cuesta-Ramirez et al., 2025). As shown in Fig. 1, the expansion of the reasoning scale failed to yield sustained gains. On the contrary, it encounters a precipitous performance decline to a certain extent. This work terms the phenomenon as over-planning in reasoning (Dziri et al., 2024), referring to the sudden and severe collapse in task accuracy once the reasoning chain exceeds a critical length. The phenomenon also aligns with recent observations on the non-monotonic scaling of test-time compute (Yang et al., 2025). This exposes certain limitations in prevailing planning-based approaches (Zelikman et al., 2022), which assumes that breaking down tasks further can lead to consistently higher returns (Valmeekam et al., 2024).

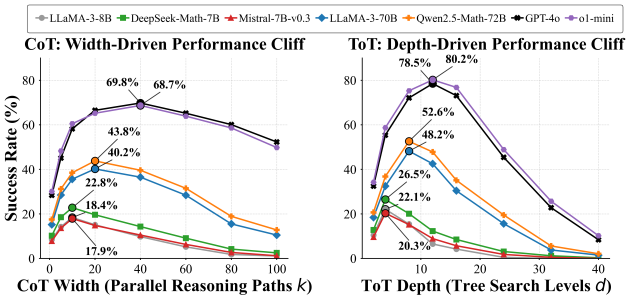

Figure 1. **The Universality of the performance degradation across Model Scales.** We evaluate seven state-of-the-art LLMs (from 7B to o1-mini) on the Omni-MATH benchmark. **(Left) Width-Driven:** As the number of parallel reasoning paths (k) increases in CoT-SC, performance initially peaks due to ensemble benefits but eventually degrades as noise dominates the majority vote. **(Right) Depth-Driven:** Similarly, in Tree-of-Thoughts (ToT), extending search depth (d) beyond the *limited reasoning space* leads to catastrophic error propagation.

The limited Reasoning Space Hypothesis. Toward a mechanistic understanding of these limitations, we introduce the *limited Reasoning Space Hypothesis*. We posit that for arbitrary prompt, there exists an intrinsic upper bound on the effective reasoning horizon of LLMs. This boundary defines a limited reasoning budget that constrains both the width and the depth of effective planning, while challenges persist in explicitly estimating the budget. Crucially, this budget captures the inter-step dependencies that standard evaluations overlook. Conventional metrics typically assume generation steps are independent statistical events, prioritizing static

[1]Academy of Sciences, Beijing, China [2]Department of Automation, Tsinghua University, Beijing, China [3]Department of Computer Science, Peking University, Beijing, China. Correspondence to: Guanlin Wu <wuguanlin16@nudt.edu.cn>.

Preprint. Work in Progress.

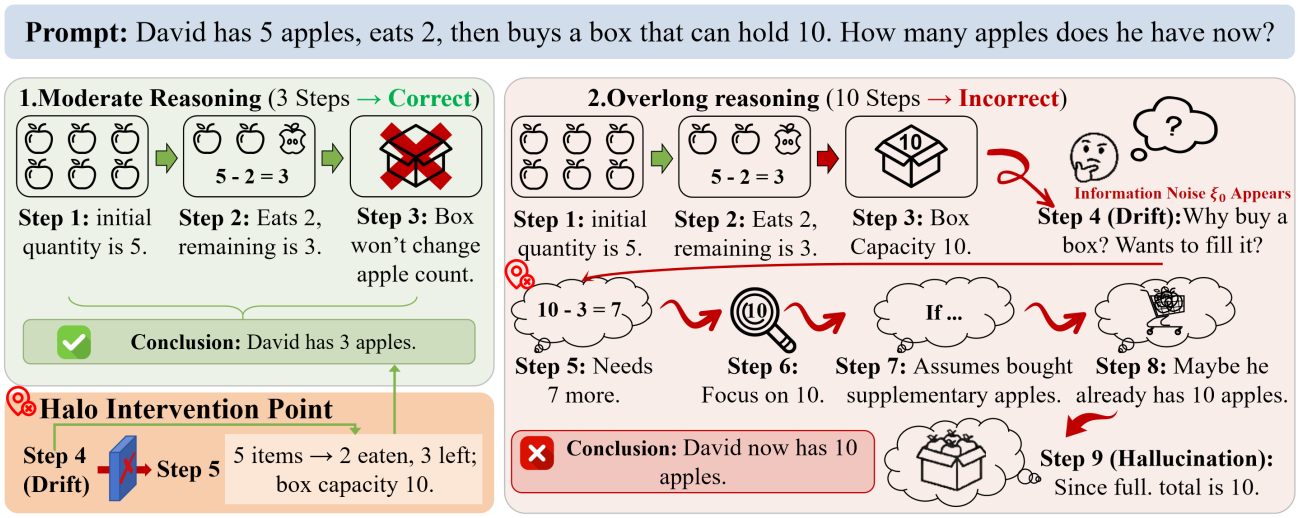

Figure 2. **The Perils of Over-Reasoning: A Case Study.** While moderate reasoning yields the correct answer (Left), extending the chain beyond the *limited reasoning space* introduces noise (Step 4). This noise is amplified through recursive feedback, leading to hallucination (Right). Our framework, **Halo** (Bottom Center), detects such drift and resets the state to maintain logical stability.

next-token accuracy. This perspective fails to account for the recursive nature of long-chain reasoning, where each output becomes an input for the next step (Madaan et al., 2024; Bachmann & Nagarajan, 2024; Alberghi et al., 2025; Ahn & Shin, 2024).

As illustrated in Fig. 2, within this recursive regime, minor deviations in reasoning (Step 4 in Fig. 2) are no longer independent statistical noise; instead, they undergo continuous iterative amplification (Lightman et al., 2023; Yang et al., 2025; Wang et al., 2025a). To characterize the mechanism behind such systematic, we reformulate long-horizon reasoning as a *Non-autonomous Stochastic Dynamical System* (Carson, 2025). Specifically, we define the reasoning state at step t as S_t, the LLM transition function as $\mathcal{G}$, and the inherent model noise as $\boldsymbol{\xi}_t$. This allows us to abstract the long-horizon reasoning as a *Non-autonomous Stochastic Dynamical System*:

$$S_{t+1} = \mathcal{G}(S_t, t) + \boldsymbol{\xi}_t, \tag{1}$$

where the explicit dependence of $\mathcal{G}$ on t accounts for the non-autonomous nature of the reasoning process, enabling a rigorous examination of the global stability of reasoning trajectories. This formulation allows us to track how errors accumulate step-by-step, determining whether the reasoning chain remains valid or is eventually overwhelmed by noise. Under this dynamical formalism, we further reveal that while reasoning performance generally scales with both recursive decomposition (*depth*) and linear generation (*width*) (Wang et al., 2022; Snell et al., 2024), both dimensions are governed by isomorphic error propagation dynamics, which we termed as Depth-Width Equivalence. This implies that neither extending the chain nor expanding the ensemble can fundamentally bypass this noise accumulation barrier, we hypothesize that the model effectively operates within a *Limited Reasoning Space*. In other words, there exists a finite region where logical consistency is preserved before errors dominate.

Exceeding the Limited Reasoning Space. While recent findings suggest that reasoning capabilities should scale with test-time compute (Wang et al., 2022), current paradigms fail to sustain this law, hitting a premature saturation point. This work attributes this failure to **exceeding the Limited Reasoning Space**: existing methods pursue **unrestricted reasoning expansion** that pushes the generation trajectory beyond the model's intrinsic effective boundary(Peng et al., 2024; Zhang & Liu, 2024)(Gema et al., 2025; Zhao et al., 2025b; Wang et al., 2025a; Wen et al., 2025; Yang et al., 2025). Specifically, once the reasoning chain crosses this boundary, the model's attention distribution becomes critically diffuse. Even within a sufficient context window, this **attentional dispersion** causes the signal form critical historical constraints to be diluted by the accumulating noise of intermediate steps. Consequently, strategies that indiscriminately generate excessive content risk scattering the model's focus, effectively wasting the potential for unlocking further capabilities. This raises a pivotal scientific question: *Can we mitigate this performance degradation by monitoring the system's entropic stability, enabling intervention before stochastic noise dominates the generation of the reasoning trajectory?* Answering this requires a transition from unlimited planning to dynamic uncertainty regulation—the core objective of our Halo framework.

Halo: Horizon-Aware Logical Optimization. To address this limitation, we propose **Halo**, a dynamics-aware framework designed for test-time compute in long-horizon rea-

soning tasks. In contrast to previous task decomposition methods (Wei et al., 2022; Besta et al., 2024) that premise on unbounded reasoning expansion, Halo treats reasoning as a resource-constrained optimization problem. Our use of entropy is grounded in the nature of mainstream LLMs: since reasoning chains are constructed via next-token prediction, the logical validity of a step can be intrinsically measured by the model's prediction certainty (Nogueira et al., 2025). Without loss of generality, Shannon entropy (Shannon, 1948) can serve as a direct proxy for this cognitive focus. In a rigorous deductive process, the transition between thoughts should exhibit a sharp next-token distribution with low entropy; conversely, a spike in entropy implies that the model's belief state is diverging into multiple plausible but uncertain paths. Leveraging these insights, we design an Entropy-Driven Dual Controller that shifts the paradigm from unlimited planning to a *Measure-then-Plan* strategy. By continually tracking entropy signals to assess the reasoning stability and proactively intervening through inverse constraints and semantic compression, Halo effectively confines the trajectory within a low-entropy stability space, thereby structurally extending the model's effective reasoning horizon and boosting the compute-scaling effect.

The design of Halo is grounded in our *Limited Reasoning Space hypothesis*. Theoretically, we prove that by maintaining the reasoning state within a locally stable space defined by low entropy, Halo effectively suppresses the Lyapunov exponent of the reasoning trajectory. This ensures that the reasoning process remains within the model's effective capacity, enabling robust performance on long-horizon tasks. Our primary contributions are summarized as follows:

- **Theory:** We formalize long-horizon reasoning as a Non-autonomous Stochastic Dynamical System. This provides the theoretical grounding for our Limited Reasoning Space Hypothesis, explaining how cumulative noise leads to the observed degradation.
- **Methodology:** We introduce Halo, which casts reasoning as a Model Predictive Control (MPC) problem (Garcia et al., 1989). By employing entropy-based monitoring and semantic compression, Halo dynamically mitigates uncertainty, enabling sustainable long-horizon reasoning where static methods saturate.
- **Empirical Results:** Halo achieves state-of-the-art performance on the RULER benchmark with a 76.4% success rate—a 3.0× improvement over AdaCoT. Furthermore, it reduces inference costs to 1/3 the tokens required by Tree-of-Thoughts (ToT) (Yao et al., 2024).

2. Theoretical Investigations: The Dynamics of Reasoning Stability

This section establishes a formal dynamical system framework to characterize the observed performance degradation. We transition from architectural definitions to a mechanistic understanding of how error propagation limits the effective reasoning length.

2.1. Dynamical Formulation of LLM Reasoning

We model the autoregressive reasoning process as a discrete-time dynamical system evolving on a high-dimensional latent space $\mathcal{M} \subseteq \mathbb{R}^d$. To capture the physics of Transformer-based architectures, we define the state evolution using the residual update formulation:

$$S_{t+1} = S_t + \mathcal{G}(S_t) + \boldsymbol{\xi}_t, \tag{2}$$

Here, $S_t \in \mathbb{R}^d$ denotes the hidden state, representing the semantic embedding of the t-th reasoning step. In this sequential formulation, S_t acts as the logical anchor: it is the output of the current step t that serves as the input for generating the next step $t+1$. Crucially, distinct from the context history, S_t captures only the current state of reasoning in a fixed dimension d. The dependency on historical information is structurally handled by the non-autonomous transition function $\mathcal{G}(\cdot, t)$, which employs the attention mechanism to retrieve relevant past contexts to compute the update. Finally, $\boldsymbol{\xi}_t \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ accounts for the stochasticity in the token sampling process of the current step.

Furthermore, we place several structural assumptions over the system's transition dynamics, which are standard and intrinsic for prevailing Transformer architectures:

Assumption 2.1 (Lipschitz Continuity)**.** The transition function $\mathcal{G}$ is L-Lipschitz continuous in the space $\mathcal{M}$. Specifically, for any two states $S, S' \in \mathcal{M}$, there exists a constant $L > 0$ such that:

$$\|\mathcal{G}(S) - \mathcal{G}(S')\|_2 \leq L\|S - S'\|_2. \tag{3}$$

Physical Interpretation: This condition ensures that the reasoning process is consistent. It implies that if the current state varies slightly (e.g., same meaning with different words), the model's generated next step will also change only slightly, rather than producing a completely different or unrelated result. This guarantees that the logical connection between steps is strong and not fragile to minor noise.

Assumption 2.2 (Bounded Spectral Radius)**.** The local Jacobian matrix $\mathbf{J}_t = \nabla_{S_t}\mathcal{G}(S_t)$ satisfies the spectral bound condition:

$$\rho(\mathbf{J}_t) \leq \kappa_{\max}, \tag{4}$$

where $\rho(\cdot)$ denotes the spectral radius. *Physical Interpretation:* Structurally enforced by Normalization layers (e.g., LayerNorm), this constraint prevents the **gradient explosion** problem, ensuring the system operates within a locally stable regime compatible with long-horizon propagation.

2.2. Error Propagation Dynamics

We analyze the stability of the reasoning chain by tracking the error vector $\boldsymbol{\delta}_t = S_t - S_t^*$, which measures the deviation of the actual state S_t from the ideal reasoning trajectory S_t^*.

To mechanistically characterize how a stochastic perturbation amplifies through the non-linear transition function $\mathcal{G}$, we approximate the local error dynamics.Using a first-order Taylor expansion around the ideal trajectory, the error evolution is governed by the linearized dynamics:

$$\boldsymbol{\delta}_{t+1} \approx (\mathbf{I} + \mathbf{J}_t)\boldsymbol{\delta}_t + \boldsymbol{\xi}_t, \tag{5}$$

where $\mathbf{A}_t \triangleq \mathbf{I} + \mathbf{J}_t$ is the state transition matrix.

The growth of this error is governed by the spectral radius $\rho = \rho(\mathbf{A}_t)$. If $\rho > 1$ (which is typical in complex generative tasks where the model expands on ideas), the system is locally expansive. To characterize the global stability, we recursively iterate this dynamic process over a reasoning horizon of N steps. In this regime, the total uncertainty (trace of the error covariance $\boldsymbol{\Sigma}_N$) accumulates as a geometric series, driven by the interaction between the structural expansion and the injected noise:

$$\mathrm{Tr}(\boldsymbol{\Sigma}_N) \approx \underbrace{\rho^{2N}\mathrm{Tr}(\boldsymbol{\Sigma}_0)}_{\text{Initial Error Growth}} + \underbrace{\sigma^2 \sum_{k=0}^{N-1} \rho^{2k}}_{\text{Noise Accumulation}}. \tag{6}$$

Here, N denotes the cumulative number of reasoning steps (such as chain length). This equation demonstrates that uncertainty accumulates geometrically. Even with zero initial error, noise σ^2 is continuously amplified by the network dynamics at every step.

2.3. The Limit of Effective Reasoning Length

Next, we derive the theoretical upper bound on the suitable reasoning length. Let us define a tolerance threshold Ψ (scalar) to represent the maximum allowable error accumulation before the generated text diverges from the logical constraint, i.e., the onset of hallucinations.

Derivation Logic. To find the critical length N^*, we solve for the step N where the accumulated uncertainty in Eq. (6) breaches the threshold Ψ. First, we treat the noise accumulation term as a geometric series sum: $\sum_{k=0}^{N-1} \rho^{2k} = \frac{\rho^{2N}-1}{\rho^2-1}$. Second, assuming the expansion rate can be characterized by the Lyapunov exponent $\lambda = \ln \rho$, we approximate $\rho^{2N} = e^{2\lambda N}$. Finally, by setting the total error equal to Ψ and inverting the equation, we obtain the explicit bound for N^* (See details in Appendix C.2).

Proposition 2.3 (Maximum Effective Reasoning Length)**.** *For a reasoning process with average Lyapunov exponent $\lambda > 0$ and sampling noise variance σ^2, the maximum effective length N^* is bounded by:*

$$N^* \approx \frac{1}{2\lambda} \ln\left(1 + \frac{\Psi(e^{2\lambda} - 1)}{\sigma^2}\right). \tag{7}$$

Interpretation. This result mechanistically explains the performance degradation. It shows that the effective length N^* is inversely proportional to the system chaos (λ) and the noise level (σ^2). As the reasoning chain extends beyond this limit ($N > N^*$), the accumulated noise mathematically dominates the signal, rendering consistent reasoning impossible without external correction.

3. Halo: Horizon-Aware Reasoning Optimization

Based on the theoretical findings in Section 2, we recognize that standard CoT operates in an **open-loop regime**, where stochastic errors accumulate unchecked until the trajectory breaches the effective reasoning capacity. To circumvent the reasoning collapse and sufficiently exploit the compute-scaling capability, we propose **Halo**, a Model Predictive Control framework for maintaining stability in long-horizon reasoning. As depicted in Fig. 3, Halo transforms reasoning into a closed-loop dynamical system: it continually tracks the system entropy to quantify the accumulated uncertainty and actively intervenes via semantic compression to rectify the trajectory before divergence occurs.

3.1. The Observer: Real-time Uncertainty Estimation

A practical controller requires a low-cost observable for the local stability of the system. Calculating the full Jacobian eigenspectrum is computationally prohibitive ($O(d^3)$). Instead, we exploit the theoretical link between Attention Dispersion and Uncertainty Propagation derived from our dynamical analysis.

Proposition 3.1 (Entropy-Instability Proxy). High Shannon entropy in attention weights implies that the model's current update depends on a diffuse mixture of historical states rather than specific antecedents. This dispersion in-

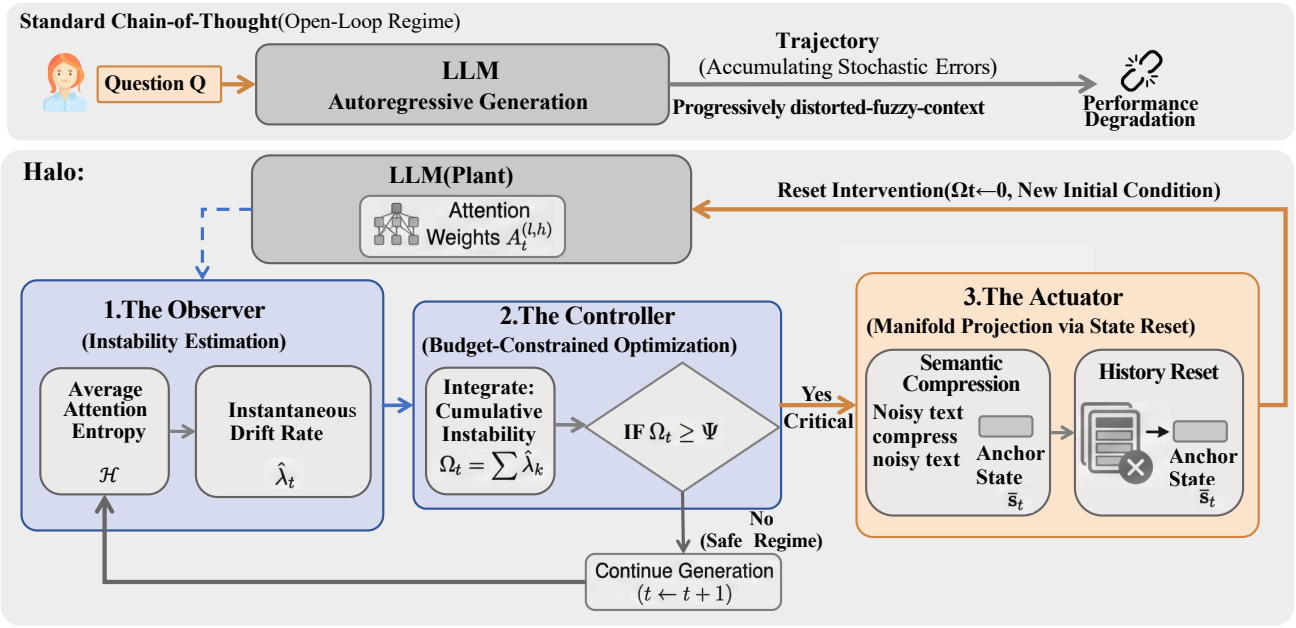

Figure 3. **The Halo Framework: Closing the Loop on Reasoning Dynamics. Top:** Standard Chain-of-Thought operates in an *open-loop regime*, where stochastic errors accumulate unchecked until the trajectory suffers from severe degradation. **Bottom:** Halo introduces a **Model Predictive Control (MPC)** loop comprising three stages: (1) **The Observer** (Sec. 3.1) estimates the instantaneous drift rate $\hat{\lambda}_t$ via mean attention entropy $\mathcal{H}$; (2) **The Controller** (Sec. 3.2) tracks the *accumulated uncertainty* Ω_t against a Tolerance Threshold Ψ; (3) **The Actuator** (Sec. 3.3) executes a *Trajectory Rectification* via semantic compression and history reset, projecting the system back to stable space and resetting the uncertainty score $\Omega_t \leftarrow 0$.

creases the **Noise Mixing Rate**, facilitating the propagation of earlier errors. We define the instantaneous drift rate $\hat{\lambda}_t$ as a linear function of the layer-averaged attention entropy:

$$\hat{\lambda}_t = \beta + \alpha \cdot \underbrace{\left(\frac{1}{L \cdot H} \sum_{l=1}^{L} \sum_{h=1}^{H} \mathcal{H}(A_t^{(l,h)}) \right)}_{\text{Mean Attention Entropy}}, \quad (8)$$

where $A_t^{(l,h)}$ is the attention probability distribution of head h in layer l, $\mathcal{H}(\cdot)$ is the Shannon entropy, and α, β are calibrated scaling factors. This metric is computed solely from the attention logits available during the forward pass, inducing negligible overhead ($< 1\%$).

3.2. The Controller: Threshold-Based Regulation

Halo treats the reasoning stability as a regulated state variable. We define a global threshold Ψ, representing the system's maximum capacity for uncertainty (derived from Eq. 7). The controller integrates the estimated drift to track the **Accumulated Uncertainty Score** Ω_t:

$$\Omega_t = \sum_{k=1}^{t} \hat{\lambda}_k. \quad (9)$$

The control law employs a simple switching logic:

- **Stable Regime** ($\Omega_t < \Psi$)**:** The trajectory remains within the effective reasoning space. The model continues standard autoregressive generation.
- **Critical Regime** ($\Omega_t \geq \Psi$)**:** The accumulated uncertainty approaches the critical limit N^*. The controller triggers an interrupt to execute *Trajectory Rectification*.

3.3. The Actuator: State Reset via Rectification

When the threshold is breached, Halo executes a **Trajectory Rectification**. This is the core mechanism that physically extends the effective reasoning length. It is not merely text summarization, but a dynamical intervention consisting of two atomic operations:

1. Semantic Compression (The Projection Operator). We define a compression function $\mathcal{C}(\cdot)$ (implemented as a specific prompt template) that maps the noisy history $\mathbf{S}_{0:t}$ to a low-entropy anchor state $\bar{\mathbf{s}}_t$.

$$\bar{\mathbf{s}}_t \leftarrow \text{LLM}(\text{Prompt}_{\text{rectify}} \oplus \text{Decode}(\mathbf{S}_{0:t})). \quad (10)$$

Physically, this acts as a state reset: it forces the diffused reasoning state onto a compact, low-dimensional semantic subspace, effectively setting the error term $\boldsymbol{\delta}_t$ to near zero.

2. Contextual History Reset (The Dynamics Interrupt). To structurally sever the dependency on the divergent trajectory, we implement a Context Reset mechanism. Rather than allowing the model to attend to the accumulated noisy history, we explicitly discard the intermediate reasoning steps from the current context window. We then re-initialize the generation process by treating the compressed summary $\bar{\mathbf{s}}_t$ (concatenated with the original question) as a fresh input sequence. This operation ensures that the attention mechanism is physically restricted to the verified logic, effectively resetting the error accumulation process by forcing the model to evolve from a clean, low-entropy state.

4. Experiments

To empirically validate the *Limited Reasoning Space* hypothesis and evaluate the efficacy of **Halo**, we design a comprehensive evaluation protocol centered on two core question: (1) Can the entropy-based observer reliably detect the boundary of the Limited Reasoning Space and trigger the Halo intervention with temporal precision? (2) Can Halo effectively extend the reasoning horizon N^*, translating extended reasoning chains into progressively higher accuracy without incurring prohibitive computational costs?

4.1. Experimental Setup

Benchmark and Baselines. Guided by the theoretical Critical Reasoning Horizon N^* (Eq. 7), we stratify our evaluation into two tiers: (1) **Tier 1 (Within-Capacity, $D < N^*$):** We use **GSM8K** (Cobbe et al., 2021) and **MATH** (Easy) (Hendrycks et al., 2021) to verify that Halo maintains efficiency in stable regimes. (2) **Tier 2 (Beyond-Capacity, $D \gg N^*$):** We employ **Omni-MATH** (Gao et al., 2024) and **RULER** (Hsieh et al., 2024) to stress-test the "Reasoning Collapse." We compare Halo against 8 baselines across three paradigms: *Open-Loop Generation* (Standard CoT), *Search-Based Optimization* (CoT-SC (Wang et al., 2022), ToT (Yao et al., 2024), GoT (Besta et al., 2024)), and *Adaptive Strategies* (AdaCoT (Pan et al., 2023), CoT-Valve (Ma et al., 2025)). Detailed dataset statistics and baseline hyperparameters are provided in Appendix F.

Metrics and Backbones. Beyond standard *Success Rate (SR)*, we introduce **Rectification Success Rate (RSR)** to measure controller precision and **Relative Token Overhead (RTO)** to quantify efficiency compared to standard CoT. We

Algorithm 1 Halo: Horizon-Aware Logical Optimization

Input: Prompt Q; LLM Policy π_θ; Stability Threshold Ψ; Dynamics params (α, β).
Output: Final Reasoning Chain $\mathcal{C}_{final}$.

Initialize Context $\mathcal{C}_0 \leftarrow [Q]$ Initialize Cumulative Uncertainty $\Omega_0 \leftarrow 0$ $t \leftarrow 0$
while *not IsFinished*$(\mathcal{C}_t)$ **do**
  /* Phase 1: The Observer (Dynamics Estimation) */
  Compute attention matrix A_t from forward pass $\pi_\theta(\mathcal{C}_t)$ Calculate mean attention entropy $\mathcal{H}_t$ via Eq. (8) Estimate instantaneous drift: $\hat{\lambda}_t \leftarrow \beta + \alpha \cdot \mathcal{H}_t$
  /* Phase 2: The Controller (State Integration) */
  Update accumulated uncertainty: $\Omega_t \leftarrow \Omega_{t-1} + \hat{\lambda}_t$ // Eq. (9)
  if CheckStability$(\Omega_t \geq \Psi)$ **then**
    /* Critical Regime: Trigger Actuator (Reset) */
    // 1. Manifold Projection via Semantic Compression (Eq. 10)
    $\bar{s}_t \leftarrow \text{LLM}(\text{Prompt}_{\text{compress}} \oplus \mathcal{C}_t)$
    // 2. Dynamics Interrupt (History Reset)
    $\mathcal{C}_t \leftarrow [Q, \bar{s}_t]$ // Reset
    $\Omega_t \leftarrow 0$ // Reset Lyapunov energy
  else
    /* Safe Regime: Standard Autoregressive Step */
    Sample next token: $x_t \sim \pi_\theta(\cdot|\mathcal{C}_t)$ $\mathcal{C}_{t+1} \leftarrow \mathcal{C}_t \cup \{x_t\}$
  $t \leftarrow t + 1$
return $\mathcal{C}_t$

Table 1. **Model Specifications.** We expand the evaluation suite to 10 models, covering the SOTA open-source landscape (Qwen2.5, LLaMA-3.1, Gemma-2) to verify Halo's architectural universality.

Family	Model	Params	Context
Open-Dense	LLaMA-3.1-8B-Instruct	8B	128k
	LLaMA-3.1-70B-Instruct	70B	128k
	Gemma-2-27B-IT	27B	8k
	Mistral-7B-Instruct-v0.3	7B	32k
Open-SOTA	Qwen2.5-72B-Instruct	72B	32k
	Qwen2.5-Math-72B	72B	32k
	Qwen2.5-7B-Instruct	7B	32k
Open-MoE	Mixtral-8x7B-v0.1	47B (13B act)	32k
	DeepSeek-V2-Lite	16B (2.4B act)	32k
	Qwen1.5-32B-MoE	32B (MoE)	32k

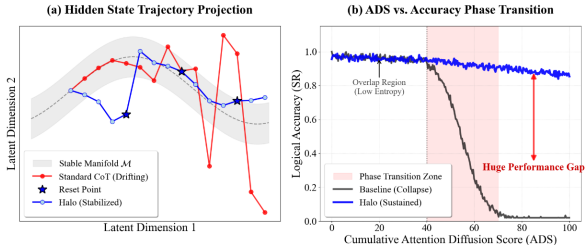

Figure 4. **Mechanistic Insights into Reasoning Stability on Omni-MATH. (a) Trajectory Projection:** t-SNE visualization of hidden states $\mathbf{s}_t$. Standard CoT (Red) diverges into the high-entropy regime, while Halo (Blue) maintains logical anchoring via periodic trajectory rectification (marked by $\star$). **(b) Phase Transition:** Reasoning accuracy exhibits a sharp collapse as cumulative uncertainty breaches the threshold Ψ. Halo preemptively acts to avoid this regime.

conduct experiments on **10 open-weights models** spanning diverse scales (7B–72B) and architectures (Dense/MoE), including the SOTA **Qwen2.5**, **LLaMA-3.1**, and **Gemma-2** families. Detailed model specifications are listed in Table 1.

4.2. Main Results

Table 2 presents the comparison of **Halo** against 8 baselines across diverse benchmarks.

Performance on Long-Horizon Tasks (Tier 2). The results on Tier 2 benchmarks highlight the limitation of standard autoregressive generation. On **Omni-MATH**, where the average reasoning length (≈ 28 steps) exceeds the effective horizon, baselines like ToT and CoT-SC struggle (21.3% and 15.8% SR) despite increased compute. This empirically validates that width expansion alone cannot mitigate the depth-driven error accumulation. In contrast, **Halo** achieves **42.7% SR**. The high Rectification Success Rate (71.2% RSR) indicates that the controller effectively detects the hallucinations and resets the reasoning state. Similarly, on RULER, Halo achieves a $3.0\times$ improvement over AdaCoT.

Efficiency Analysis. Halo achieves superior accuracy with significantly lower computational cost compared to search-based methods. While ToT and GoT require $3.5\times$ to $5.1\times$ more tokens, Halo's Relative Token Overhead (RTO) is only **1.29**$\times$. Crucially, on Tier 1 tasks (e.g., GSM8K), Halo maintains performance parity with standard CoT (+1.8% to +7.2% SR) without aggressive intervention. This confirms the system's adaptivity: the controller incurs negligible overhead in stable regimes and only activates rectification when the accumulated uncertainty Ω exceeds the threshold Ψ.

Impact of Model Capacity. Table 3 shows that Halo provides consistent gains across model scales, though the nature of the gain differs. For smaller models (e.g., LLaMA-3.1-8B), Halo provides a massive stability boost (**+15.2%**), effectively compensating for their weaker intrinsic reasoning capabilities. For SOTA models like **Qwen2.5-Math-72B**, Halo provides a focused gain of **+2.0%**, pushing the accuracy to **91.3%**. This result is significant as it suggests that even heavily fine-tuned models are susceptible to test-time error drift in extremely long chains. Halo effectively mitigates these residual errors that SFT (Supervised Fine-Tuning) cannot eliminate.

4.3. Mechanistic Analysis

To understand the internal behavior of Halo, we analyze the correlation between reasoning length, entropy, and latent state dynamics. **Verification of the Critical Horizon.** We conducted a grid-search analysis on Omni-MATH to validate the theoretical bounds derived in Section 2.3. As shown in Fig. 6, the model maintains robust performance for shorter chains ($N < 20$). However, as the chain length exceeds this horizon, the success rate drops precipitously ($SR < 10\%$). This phenomenon aligns with our theoretical prediction of exponential error propagation. Crucially, this failure mode perfectly overlaps with the high-entropy regime, validating that attention entropy is a reliable indicator of reasoning collapse.

Latent Trajectory Stabilization. Figure 4 (Left) visualizes the hidden states $\mathbf{s}_t$ using t-SNE. Baseline CoT trajectories (Red) show a clear tendency to diverge (drift apart) after approximately N^* steps, corresponding to the loss of logical consistency. In contrast, Halo (Blue) maintains a tighter cluster. The intervention points ($\star$) effectively pull the diverging states back towards the stable region. This confirms that the *State Reset* mechanism acts as a correction step, reducing the accumulated variance in the latent space.

Precision of Intervention. Figure 5 (Right) compares the step indices where Halo triggers an intervention versus where the baseline model typically fails. The two distributions are highly aligned. This indicates that the entropy-based threshold Ψ is temporally precise: it triggers rectifica-

Table 2. **Holistic Multi-Metric Performance (LLaMA-3-70B).** Success Rate (SR, %) and Relative Token Overhead (RTO, ×). Best results are **bolded**, and second-best results are underlined. Halo achieves Pareto-optimal performance, significantly outperforming baselines on Tier 2 tasks (Omni-MATH, RULER) where reasoning collapse typically occurs.

METHOD	MATHEMATICAL & SYMBOLIC REASONING						LONG-CONTEXT STABILITY & RETRIEVAL							
	GSM8K (TIER 1)		MATH (TIER 1)		OMNI (TIER 2)		RULER (TIER 2)		LONGBENCH		INFBENCH		LRA-L2	
	SR↑	RTO	SR↑	RTO	SR↑	RTO	SR↑	RTO	SR↑	RTO	SR↑	RTO	SR↑	RTO
STANDARD COT	82.4	1.00	10.8	1.00	12.5	1.00	18.2	1.00	35.6	1.00	28.9	1.00	15.3	1.00
COT-SC (k=10)	86.1	2.15	14.2	2.15	15.8	2.15	21.4	2.15	39.2	2.15	32.5	2.15	18.7	2.15
ADACOT	88.5	1.45	19.3	1.45	21.5	1.45	25.1	1.45	42.8	1.45	36.7	1.45	22.1	1.45
TOT (b=5,d=3)	87.4	3.50	19.1	3.50	21.3	3.50	25.7	3.50	41.5	3.50	35.2	3.50	21.5	3.50
COT-VALVE	88.9	0.85	36.4	0.85	40.2	0.85	65.8	0.85	58.3	0.85	51.6	0.85	38.9	0.85
HALO (OURS)	**89.2**	1.29	**38.5**	1.29	**42.7**	1.29	**76.4**	1.29	**62.5**	1.29	**55.8**	1.29	**42.3**	1.29

Table 3. **Cross-Model Robustness on Omni-MATH.** Halo demonstrates consistent gains across varying scales. **Rel. Gain** denotes improvement over AdaCoT. Notably, Halo pushes the SOTA **Qwen2.5-Math** to 91.3%, breaking the static inference ceiling.

BACKBONE	COT	TOT	ADACOT	HALO	GAIN
Small & Efficient Models					
LLAMA-3.1-8B	42.5	48.1	49.3	**56.8**	+15%
MISTRAL-V0.3	44.2	49.5	51.0	**57.4**	+13%
GEMMA-2-27B	58.1	62.4	63.8	**69.2**	+8.5%
MoE Architectures					
MIXTRAL-8X7B	55.4	59.2	60.5	**66.7**	+10%
DEEPSEEK-V2-LITE	52.3	57.8	58.9	**65.1**	+11%
Large & SOTA Models					
LLAMA-3.1-70B	76.8	79.5	80.2	**83.4**	+4.0%
QWEN2.5-72B	82.4	84.1	84.8	**87.2**	+2.8%
QWEN2.5-MATH	88.0	89.2	89.5	**91.3**	+2.0%

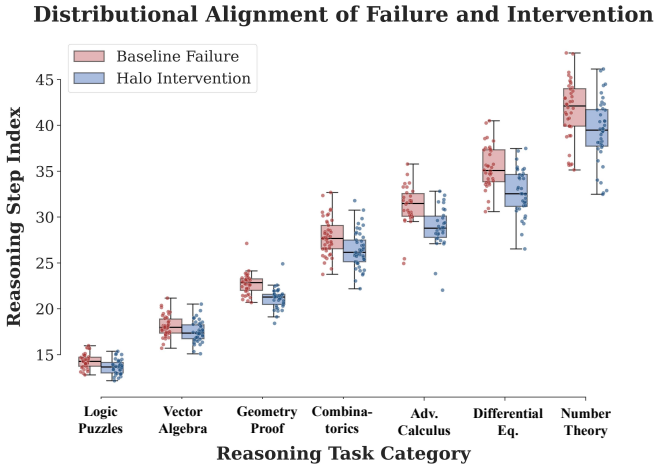

Figure 5. Distributional Alignment of Reasoning Boundaries. We compare the step index of baseline failures (Red) and Halo interventions (Blue) across seven reasoning domains. The tight clustering of data points indicates that Halo consistently identifies the critical reasoning horizon with low variance.

tion exactly when the model is about to degenerate, avoiding both premature interruption and delayed intervention.

4.4. Parameter Sensitivity

We evaluate the robustness of Halo with respect to its two hyperparameters: the tolerance threshold Ψ and the entropy sensitivity α. Table 4 reports the Success Rate (SR) on Omni-MATH using the LLaMA-3-8B model. The results indicate that Halo maintains stable performance for $\Psi \in [4.0, 6.0]$. Lower thresholds ($\Psi < 3.0$) lead to excessive interventions which interrupt valid reasoning steps, while higher thresholds ($\Psi > 8.0$) delay the necessary rectification. Regarding α, values in the range $[0.7, 1.0]$ yield consistent gains. Notably, across all tested configurations, Halo outperforms the AdaCoT baseline (49.3%), suggesting that the improvements strictly originate from the entropy-driven mechanism rather than specific parameter tuning.

5. Conclusion

In this work, we formulate long-horizon reasoning as a Non-autonomous Stochastic Dynamical System , attributing reasoning failures to the exponential error accumulation that defines a Limited Reasoning Space. To overcome this intrinsic boundary, we propose Halo, a Model Predictive Control (MPC) framework that shifts the paradigm from open-loop generation to a regulated Measure-then-Plan strategy. By leveraging Attention Entropy to detect stability drifts and executing semantic state resets, Halo effectively mitigates noise dominance. Our experiments on Omni-MATH and RULER demonstrate that Halo significantly extends the effective reasoning horizon, achieving superior accuracy with substantially lower computational cost compared to static decomposition methods.

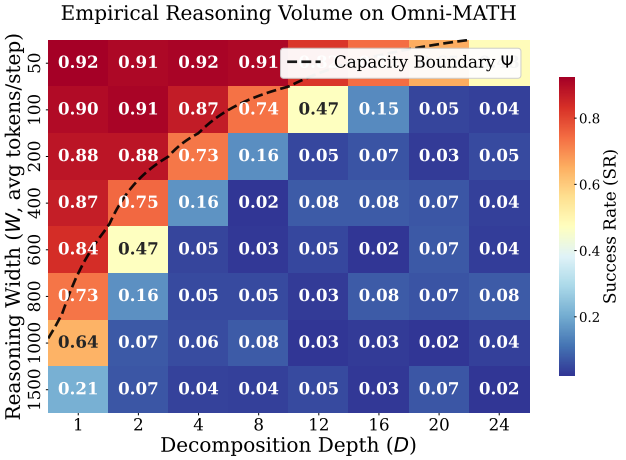

Figure 6. **Stability Phase Transition Analysis.** The heatmap illustrates the Success Rate (SR) as a function of Reasoning Length N. The dashed line represents the theoretical stability boundary N^*. The stochastic fluctuations in the red Divergence Regime indicate the collapse of semantic consistency as the accumulated uncertainty dominates the trajectory.

Table 4. **Sensitivity Analysis on Omni-MATH (LLaMA-3-8B).** Success Rate (SR, %) with varying Tolerance Thresholds (Ψ) and Sensitivity (α). The performance remains robust across a wide range of configurations compared to the AdaCoT baseline (49.3%).

Sensitivity (α)	Tolerance Threshold (Ψ)				
	$\Psi = 2.0$	$\Psi = 4.0$	$\Psi = 5.0$	$\Psi = 6.0$	$\Psi = 8.0$
0.5	51.2	52.8	53.1	52.5	50.4
0.85 (Ours)	53.5	**56.5**	**56.8**	**56.2**	51.9
1.0	52.8	55.9	56.4	55.7	51.2
1.2	50.1	54.2	54.8	53.9	50.8

Impact Statements

This work presents Halo, with the goal of improving the computational efficiency of LLM reasoning and post-training. By reducing redundant test-time planning and avoiding unnecessary RL rollouts through online predictive prompt selection, the method lowers compute cost and energy consumption, which is beneficial for large-scale deployment in industry. The approach operates solely at the level of optimization and data selection, without introducing new model capabilities or altering model outputs. As a result, it does not introduce new societal or ethical risks beyond those already associated with large language models, and its downstream impact depends on the specific application context.

References

Aghaee, F. et al. Rb-llm control: an intelligent control framework with long-term logical reasoning. *Aerospace Science and Technology*, 2025.

Ahn, J. and Shin, K. Recursive Chain-of-Feedback Prevents Performance Degradation from Redundant Prompting. *arXiv.org*, 2024. doi: 10.48550/ARXIV.2402.02648. URL https://arxiv.org/abs/2402.02648.

Alberghi, R., Demyanenko, E., Biggio, L., and Saglietti, L. On the Bias of Next-Token Predictors Toward Systematically Inefficient Reasoning: A Shortest-Path Case Study. *arXiv.org*, 2025. doi: 10.48550/ARXIV.2507.05362. URL https://arxiv.org/abs/2507.05362.

Arbuzov, M. L., Shvets, A. A., and Beir, S. Beyond exponential decay: Rethinking error accumulation in large language models. *arXiv preprint arXiv:2505.24187*, 2025.

Bachmann, G. and Nagarajan, V. The pitfalls of next-token prediction. *International Conference on Machine Learning*, 2024. doi: 10.48550/ARXIV.2403.06963. URL https://arxiv.org/abs/2403.06963.

Bengio, S., Vinyals, O., Jaitly, N., and Shazeer, N. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in Neural Information Processing Systems*, 28, 2015.

Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Gianinazzi, L., Mueller, J., Fischer, L., Podstawski, K., Claassen, T., Alistarh, D., et al. Graph of thoughts: Solving elaborate problems with large language models. *AAAI Conference on Artificial Intelligence*, 2024.

Bidochko, A. et al. Thought management system for long-horizon, goal-driven reasoning. *Journal of Computational Science*, 2025.

Carson, J. D. A stochastic dynamical theory of llm self-adversariality: Modeling severity drift as a critical process, 2025. URL https://arxiv.org/abs/2501.16783.

Castagna, F. et al. Steering llm reasoning with argumentative querying. *arXiv preprint arXiv:2412.15177*, 2024.

Chen, Q., Qin, L., Liu, J., Peng, D., Guan, J., Wang, P., Hu, M., Zhou, Y., Gao, T., and Che, W. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models, 2025. URL https://arxiv.org/abs/2503.09567.

Cheng, D., Huang, S., Zhu, X., Dai, B., Zhao, W. X., Zhang, Z., and Wei, F. Reasoning with Exploration: An Entropy Perspective. *arXiv.org*, 2025. doi: 10.48550/ARXIV.2506.14758. URL https://arxiv.org/abs/2506.14758.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Cuesta-Ramirez, J., Beaussant, S., and Mounsif, M. Large Reasoning Models are not thinking straight: on the unreliability of thinking trajectories. *arXiv.org*, 2025. doi: 10.48550/ARXIV.2507.00711. URL https://arxiv.org/abs/2507.00711.

Dhuliawala, S. et al. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*, 2024.

Dziri, N., Milton, X., Yu, M., Zaiane, O., and Edwards, H. Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36, 2024.

Eisenstadt, R., Zimerman, I., and Wolf, L. Overclocking LLM Reasoning: Monitoring and Controlling Thinking Path Lengths in LLMs. *arXiv.org*, 2025. doi: 10.48550/ARXIV.2506.07240. URL https://arxiv.org/abs/2506.07240.

Gao, B., Song, F., Yang, Z., Cai, Z., Miao, Y., Dong, Q., et al. Omni-math: A universal olympiad level mathematic benchmark for large language models. *arXiv preprint arXiv:2410.07985*, 2024.

Garcia, C. E., Prett, D. M., and Morari, M. Model predictive control: Theory and practice—a survey. *Automatica*, 25 (3):335–348, 1989.

Gema, A. P., Hägele, A., Chen, R., Arditi, A., Goldman-Wetzler, J., Fraser-Taliente, K., Sleight, H., Petrini, L., Michael, J., Alex, B., Minervini, P., Chen, Y., Benton, J., and Perez, E. Inverse Scaling in Test-Time Compute. *arXiv.org*, 2025. doi: 10.48550/ARXIV.2507.14417. URL https://arxiv.org/abs/2507.14417.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. On calibration of modern neural networks. *International Conference on Machine Learning*, pp. 1321–1330, 2017.

Hassid, M., Synnaeve, G., Adi, Y., and Schwartz, R. Don't Overthink it. Preferring Shorter Thinking Chains for Improved LLM Reasoning. *arXiv.org*, 2025. doi: 10.48550/ARXIV.2505.17813. URL https://arxiv.org/abs/2505.17813.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset. In *NeurIPS*, 2021.

Hoza, P. Evaluating reasoning in large language models with a modified think-a-number game. *Acta Informatica Pragensia*, 2025.

Hsieh, C.-P., Sun, S., Kriman, S., Acharya, S., Ginsburg, B., et al. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

Huang, J., Xu, Y., Wang, Q., Wang, Q. C., Liang, X., Wang, F., Zhang, Z., Wei, W., Zhang, B., Huang, L., et al. Foundation models and intelligent decision-making: Progress, challenges, and perspectives. *The Innovation*, 6(6), 2025.

Khan, S. et al. Reframing the reasoning cliff as an agentic gap. *arXiv preprint arXiv:2506.18957*, 2025.

Kojima, T., Gu, S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213, 2022.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Shyam, P., Misra, S., Robinson, K., Pundir, V., Felix, R., Jain, P., et al. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Ma, X., Wan, G., Yu, R., Fang, G., and Wang, X. Cot-valve: Length-compressible chain-of-thought tuning. *arXiv preprint arXiv:2502.09601*, 2025.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Pelzak, N., Ribeiro, M., Goyal, S., et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

Mao, Y., Qu, Y., Wang, C., Zou, H., and Ji, X. Dynamics-predictive sampling for active RL finetuning of large reasoning models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=voeheZjd8p.

Motwani, S. R., Ivanova, A., Cai, Z., Torr, P., Islam, R., Shah, S., de Witt, C. S., and London, C. h1: Bootstrapping llms to reason over longer horizons via reinforcement learning, 2025. URL https://arxiv.org/abs/2510.07312.

Nogueira, J. P., Sun, W., Silva, A., and Zumot, L. Certainty-guided reasoning in large language models: A dynamic thinking budget approach. *arXiv preprint arXiv:2509.07820*, 2025.

Pan, L. et al. Automatically correcting large language models with self-criticism. *arXiv preprint arXiv:2305.12295*, 2023.

Peng, X., Xia, C., Yang, X., Xiong, C., Wu, C.-S., and Xing, C. Regenesis: Llms can Grow into Reasoning Generalists via Self-Improvement. *International Conference on Learning Representations*, 2024. doi: 10.48550/ARXIV.2410.02108. URL https://arxiv.org/abs/2410.02108.

Qian, C. et al. Thinking tokens are information peaks in llm reasoning. *OpenReview*, 2025.

Qu, Y., Wang, Q., Mao, Y., Hu, V. T., Ommer, B., and Ji, X. Can prompt difficulty be online predicted for accelerating rl finetuning of reasoning models? *arXiv preprint arXiv:2507.04632*, 2025.

Rameshkumar, R., Huang, J., Sun, Y., Xia, F., and Saparov, A. Reasoning models reason well, until they don't, 2025. URL https://arxiv.org/abs/2510.22371.

Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Shinn, N., Labash, B., and Gopinath, A. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Snell, C., Lee, J., Xu, K., et al. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Valmeekam, K., Olmo, A., Sreedharan, S., and Kambhampati, S. On the planning abilities of large language models-a critical investigation. *Advances in Neural Information Processing Systems*, 36, 2024.

Wang, J., Zhu, B., Leong, C. T., Li, Y., and Li, W. Scaling over Scaling: Exploring Test-Time Scaling Plateau in Large Reasoning Models. *arXiv.org*, 2025a. doi: 10.48550/ARXIV.2505.20522. URL https://arxiv.org/abs/2505.20522.

Wang, Q., Xiao, Z., Mao, Y., Qu, Y., Shen, J., Lv, Y., and Ji, X. Model predictive task sampling for efficient and robust adaptation. *arXiv preprint arXiv:2501.11039*, 2025b.

Wang, X., Wei, J., Schuurmans, D., Quoc, V., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Wei, J., Wang, Y., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837, 2022.

Wen, H., Su, Y., Zhang, F., Liu, Y., Liu, Y., Zhang, Y.-Q., and Li, Y. Parathinker: Native Parallel Thinking as a New Paradigm to Scale LLM Test-time Compute. *arXiv.org*, 2025. doi: 10.48550/ARXIV.2509.04475. URL https://arxiv.org/abs/2509.04475.

Xiao, Z., Yan, S., Hong, J., Cai, J., Jiang, X., Hu, Y., Shen, J., Wang, Q., and Snoek, C. G. Dynaprompt: Dynamic test-time prompt tuning. *arXiv preprint arXiv:2501.16404*, 2025.

Xu, Z., Qiu, Z., Huang, G., Li, K., Li, S., Zhang, C., Li, K., Yi, Q., Jiang, Y., Zhou, B., Lian, F., and Kang, Z. Adaptive Termination for Multi-round Parallel Reasoning: An Universal Semantic Entropy-Guided Framework. *arXiv.org*, 2025. doi: 10.48550/ARXIV.2507.06829. URL https://arxiv.org/abs/2507.06829.

Xu, Z. et al. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024.

Yang, W., Ma, S., Lin, Y., and Wei, F. Towards Thinking-Optimal Scaling of Test-Time Compute for LLM Reasoning. *arXiv.org*, 2025. doi: 10.48550/ARXIV.2502.18080. URL https://arxiv.org/abs/2502.18080.

Yang, Y., Shi, Y., Wang, C., Zhen, X., Shi, Y., and Xu, J. Reducing fine-tuning memory overhead by approximate and memory-sharing backpropagation. *arXiv preprint arXiv:2406.16282*, 2024.

Yao, S., Yu, D., Zhao, J., Shafran, I., McManus, T., Narasimhan, K., and Cao, Y. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

You, W. et al. Probabilistic soundness guarantees in llm reasoning. *arXiv preprint arXiv:2507.12948*, 2025.

Zelikman, E., Wu, Y., Mu, J., and Goodman, N. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

Zhang, J. and Liu, K. Thought Space Explorer: Navigating and Expanding Thought Space for Large Language Model Reasoning. In *2024 IEEE International Conference on Big Data (BigData)*, pp. 8259–8261. IEEE, dec 15 2024. doi: 10.1109/bigdata62323.2024.10825638. URL http://dx.doi.org/10.1109/BigData62323.2024.10825638.

Zhang, X., Abdul-Mageed, M., et al. Autoregressive+ chain of thought= recurrent: Recurrence's role in language models' computability. *arXiv preprint arXiv:2409.09239*, 2024.

Zhang, Z. Comprehension without competence: Architectural limits of llms in symbolic computation. *arXiv preprint arXiv:2507.10624*, 2025.

Zhao, Q. et al. Uncertainty propagation on llm agent. *Proceedings of the ACL*, 2025a.

Zhao, Y., LV, J., Wu, D., Wang, J., and Gooley, C. Are We Scaling the Right Thing? A System Perspective on Test-Time Scaling. *arXiv.org*, 2025b. doi: 10.48550/ARXIV.2509.19645. URL https://arxiv.org/abs/2509.19645.

Zhou, Y., Wang, Y., Yin, X., Zhou, S., and Zhang, A. R. The geometry of reasoning: Flowing logics in representation space. *arXiv preprint arXiv:2510.09782*, 2025.

Zhu, Z. et al. Empowering long-horizon llm agents with cognitive overload management. *OpenReview*, 2025.

Zou, H., Mao, Y., Qu, Y., Wang, Q., and Ji, X. Utility-diversity aware online batch selection for llm supervised fine-tuning. *arXiv preprint arXiv:2510.16882*, 2025a.

Zou, H., Zang, Y., Xu, W., and Ji, X. Fly-cl: A fly-inspired framework for enhancing efficient decorrelation and reduced training time in pre-trained model-based continual representation learning. *arXiv preprint arXiv:2510.16877*, 2025b.

A. Related Work

Efficiency optimization is a promising research focus in either training (Wang et al., 2025b; Qu et al., 2025; Zou et al., 2025b; Mao et al., 2026; Zou et al., 2025a; Yang et al., 2024) or deploying LLMs (Pan et al., 2023; Xiao et al., 2025). Our work re-examines long-horizon reasoning through the lens of *Dynamical Systems*, positioning the proposed "limited reasoning space " against the prevailing unbounded scaling paradigms (Chen et al., 2025; Zhang et al., 2024).

Test-Time Optimization (TTO) and the Unbounded Horizon Assumption. Recent advancements in Test-Time Optimization have shifted from linear chains (CoT) (Wei et al., 2022) to complex topologies like *Tree of Thoughts* (ToT) (Yao et al., 2024) and *Graph of Thoughts* (GoT) (Besta et al., 2024). These methods formulate reasoning as a search problem over a semantic space, operating under an implicit *Unbounded Horizon Assumption*—the premise that increasing decomposition granularity ($Depth$) or search breadth ($Width$) monotonically yields better performance (Wang et al., 2022; Zelikman et al., 2022). However, our empirical observation of the *Performance Cliff* (Figure 1) challenges this view, aligning with recent evidence that reasoning models fail catastrophically beyond modest complexity (Rameshkumar et al., 2025; Khan et al., 2025). Unlike Snell et al. (2024), who model compute scaling in static regimes, we theoretically characterize the reasoning process as a resource-constrained flow (Zhou et al., 2025). We prove that without dynamic constraints, the cumulative error variance inevitably breaches the stability margin (Arbuzov et al., 2025), rendering "deeper" reasoning counter-productive beyond the *Critical Reasoning Horizon* N^* (Hoza, 2025; Zhang, 2025). While Halo regulates test-time reasoning trajectories, complementary approaches build intrinsic capacity during training. Notably, Motwani et al. (Motwani et al., 2025) use a progressive RL curriculum to bootstrap foundational long-horizon skills without requiring novel annotations. Halo serves as the natural inference-time counterpart to such capacity-building. By executing entropy-driven state resets to prevent stochastic error drift, Halo ensures models can fully exploit these learned capacities for robust long-horizon reasoning.

Dynamics of Recursive Error Propagation. While prior work attributes hallucination to *Exposure Bias* (Bengio et al., 2015) or calibration errors (Guo et al., 2017), these perspectives largely treat generation steps as independent statistical events. In contrast, we model autoregressive reasoning as a *Non-autonomous Stochastic Dynamical System* (Carson, 2025), where errors are not merely additive but multiplicative via the Jacobian $\mathbf{J}_t$ (Zhao et al., 2025a). This formulation allows us to derive the *Depth-Width Equivalence*, a novel theoretical insight revealing that recursive decomposition and linear generation share the same error propagation laws (You et al., 2025). This distinguishes our work from Xu et al. (2024), providing a mechanistic rather than purely statistical explanation for why long-context reasoning collapses into chaotic regimes (Dziri et al., 2024; Valmeekam et al., 2024).

Adaptive Inference: Semantic vs. Structural Feedback. To address rigid reasoning structures, adaptive methods like *AdaCoT* and *Self-Correction* (Pan et al., 2023; Madaan et al., 2024) introduce iterative refinement steps. Critically, these approaches rely on *Semantic Feedback*—prompting the model to verbally critique its own output (Shinn et al., 2024; Castagna et al., 2024). This leads to a recursive failure mode where the verification signal itself is subject to the same drift as the generation process (Lightman et al., 2023; Dhuliawala et al., 2024). **Halo** fundamentally diverges by employing *Structural Feedback* (Aghaee et al., 2025). By utilizing *Attention Entropy* as a proxy for the system's spectral instability (Qian et al., 2025; Xu et al., 2025; Cheng et al., 2025; Eisenstadt et al., 2025), we implement a *Closed-Loop Controller* that is decoupled from the model's semantic content (Bidochko et al., 2025). This enables a "Measure-then-Plan" strategy that actively confines the trajectory within the *limited reasoning space* (Zhu et al., 2025), specifically addressing the structural mismatch ignored by heuristic adaptive methods.

B. Nomenclature and Physical Interpretation

Table 5 provides a detailed mapping between our dynamical system formulation and the actual behavior of Large Language Models. We utilize a concrete **Running Example** from our experiments: *"Find the smallest positive integer x such that $\log_2(x^2 - 4x) > 5$."*

C. Mathematical Proofs & Derivations

In this appendix, we provide the rigorous algebraic derivations for the stability bounds presented in Section 2. To address the theoretical challenges of modeling large language models (LLMs), we first formalize the requisite approximations that bridge the gap between the discrete, history-dependent nature of Transformers and continuous dynamical systems.

Table 5. **Nomenclature and Physical Interpretation.** The Running Example traces the trajectory of an LLM solving a logarithmic inequality, illustrating how *Halo* detects and corrects a "Complex Number Hallucination" at Step 15.

Symbol	Physical Meaning in LLM Reasoning	Running Example (Inequality Task)
S_t	**Hidden State / Semantic Embedding** at step t. It captures the current logical snapshot in the latent space $\mathcal{M}$.	At Step 14, the vector S_{14} encodes the intermediate thought: *"The inequality simplifies to* $x^2-4x-32>0$*. The roots are roughly* 2 ± 6.4*, so* $x \approx 8.4$ *and* $x \approx -4.4$*."*
ξ_t	**Stochastic Noise Vector**. Represents random fluctuations in token sampling (e.g., top-p) or "distractor" associations triggered by the context.	A random noise vector ξ_{15} triggers a low-probability association with "complex analysis," proposing a distractor thought: *"Wait, logarithms have complex domains... maybe* $x = 2 + 3i$*?"*
J_t	**Local Jacobian Matrix** ($\nabla_{S_t}\mathcal{G}$). Measures how strongly the model amplifies a small distraction in the next step.	The model is uncertain about the integer constraint. The Jacobian J_{15} is expansive ($\|J_{15}\| > 1$), causing the minor "complex number" thought to dominate the attention in the next token generation.
$\mathcal{H}(A_t)$	**Attention Entropy**. A proxy for "Attention Dispersion." High entropy implies the model is losing focus and attending uniformly to irrelevant history.	At Step 15, $\mathcal{H}_{15}$ spikes to **3.2** (vs. baseline 0.8). The attention map is diffuse, looking at both the initial "integer" constraint and the new "complex root" noise simultaneously.
$\hat{\lambda}_t$	**Instantaneous Drift Rate**. Estimated via Eq. 8 ($\hat{\lambda}_t = \beta + \alpha\mathcal{H}_t$). A positive value indicates the onset of a chaotic regime.	The calculated drift $\hat{\lambda}_{15}$ becomes positive (+0.45), signaling that the reasoning trajectory is diverging from the logical manifold (Real Numbers $\mathbb{R}$) into hallucination (Complex Plane $\mathbb{C}$).
Ω_t	**Accumulated Uncertainty Score**. The integral of drift ($\sum \hat{\lambda}_k$). Represents the total risk of collapse accumulated so far.	By Step 15, Ω_{15} reaches **5.2**, breaching the safety threshold. The system recognizes that the chain has become too "fuzzy" to continue reliably.
Ψ	**Tolerance Threshold**. The boundary of the *Limited Reasoning Space*. Determined by the model's capacity (e.g., LLaMA-3-70B).	For this model, $\Psi = 5.0$. Since $\Omega_{15}(5.2) \geq \Psi(5.0)$, the *Halo* controller triggers an immediate intervention to prevent the upcoming hallucination.
$\mathcal{C}(\cdot)$	**Semantic Compression Operator** (Actuator). Projects the noisy state back to a low-entropy "Anchor State" by filtering out invalid branches.	*Halo* executes Compression: It discards the complex number branch and summarizes: *Verified Progress: Roots are approx -4.47 and 8.47. We need the smallest positive integer* > 8.47.
Context Reset	**Dynamics Interrupt**. Physically clearing the history of the noisy steps (Step 1-15) and re-initializing with the Anchor State.	The model "forgets" the confusing complex number debate. It restarts generation from the clean summary. Next Step 16: *"Since* x *must be an integer* > 8.47*, the answer is* $x = 9$*."*

C.1. Dynamical Formulation & Rigorous Assumptions

We formulate the autoregressive reasoning process as a discrete-time stochastic dynamical system on a latent manifold $\mathcal{M} \subseteq \mathbb{R}^d$. The state evolution is given by:

$$S_{t+1} = S_t + \mathcal{G}(S_t) + \boldsymbol{\xi}_t \tag{11}$$

To ensure mathematical tractability while maintaining physical fidelity, we explicitly state the following modeling assumptions:

Assumption B.1 (Continuous Relaxation of Token Sampling). While token selection is inherently discrete, we model the uncertainty introduced by the sampling strategy (e.g., Top-p) as an additive continuous noise vector $\boldsymbol{\xi}_t$ in the semantic embedding space. We assume $\boldsymbol{\xi}_t$ is isotropic Gaussian white noise:

$$\boldsymbol{\xi}_t \sim \mathcal{N}(0, \sigma^2 \mathbf{I}), \quad \mathbb{E}[\boldsymbol{\xi}_t \boldsymbol{\xi}_k^\top] = \delta_{tk} \sigma^2 \mathbf{I} \tag{12}$$

Justification: This represents a "mean-field" approximation of the quantization error and stochasticity observed in the residual stream during generation.

Assumption B.2 (Local Markovian Approximation). we assume that for a single reasoning step, the local state transition is dominated by the current residual activation S_t. The historical dependence is approximated as a slowly varying background field. Thus, the Jacobian is approximated as $\mathbf{J}_t \approx \nabla_{S_t} \mathcal{G}(S_t)$.

Assumption B.3 (Bounded Expansion via Spectral Norm). Unlike asymptotic stability analysis which relies on spectral radius, we characterize the finite-horizon transient growth using the **Spectral Norm** (induced 2-norm) of the transition operator. We define the maximal expansion rate μ such that:

$$\mu = \sup_{S \in \mathcal{M}} \|\mathbf{I} + \nabla \mathcal{G}(S)\|_2 \tag{13}$$

This accounts for the potential transient amplification in non-normal matrices, which is critical in deep neural networks.

C.2. Derivation of the Critical Reasoning Horizon (N^*)

We derive the upper bound for the effective reasoning length (Proposition 2.1) by analyzing the propagation of the error covariance matrix.

Step 1: Evolution of the Error Covariance

Let $\boldsymbol{\delta}_t = S_t - S_t^*$ be the deviation from the ideal reasoning trajectory. Using the first-order Taylor expansion under Assumption B.2, the error dynamics are linearized as:

$$\boldsymbol{\delta}_{t+1} \approx \mathbf{A}_t \boldsymbol{\delta}_t + \boldsymbol{\xi}_t \tag{14}$$

where $\mathbf{A}_t = \mathbf{I} + \mathbf{J}_t$ is the state transition matrix at step t. Let $\boldsymbol{\Sigma}_t = \mathbb{E}[\boldsymbol{\delta}_t \boldsymbol{\delta}_t^\top]$ be the error covariance matrix. Assuming independence between the current state error and the injected noise, the covariance evolves as:

$$\begin{aligned} \boldsymbol{\Sigma}_{t+1} &= \mathbb{E}[(\mathbf{A}_t \boldsymbol{\delta}_t + \boldsymbol{\xi}_t)(\mathbf{A}_t \boldsymbol{\delta}_t + \boldsymbol{\xi}_t)^\top] \\ &= \mathbf{A}_t \mathbb{E}[\boldsymbol{\delta}_t \boldsymbol{\delta}_t^\top] \mathbf{A}_t^\top + \mathbb{E}[\boldsymbol{\xi}_t \boldsymbol{\xi}_t^\top] \\ &= \mathbf{A}_t \boldsymbol{\Sigma}_t \mathbf{A}_t^\top + \sigma^2 \mathbf{I} \end{aligned} \tag{15}$$

Step 2: Upper Bound using Operator Norms

We seek to bound the magnitude of the uncertainty, quantified by the spectral norm of the covariance matrix $\|\boldsymbol{\Sigma}_t\|_2$ (representing the worst-case variance along any direction). Using the sub-multiplicative property of the norm $\|\mathbf{XY}\| \leq$

$\|\mathbf{X}\|\|\mathbf{Y}\|$ and the definition $\mu \geq \|\mathbf{A}_t\|_2$:

$$\begin{aligned}\|\mathbf{\Sigma}_{t+1}\|_2 &\leq \|\mathbf{A}_t\|_2\|\mathbf{\Sigma}_t\|_2\|\mathbf{A}_t^\top\|_2 + \|\sigma^2\mathbf{I}\|_2 \\ &\leq \mu^2\|\mathbf{\Sigma}_t\|_2 + \sigma^2 \end{aligned} \tag{16}$$

This recursive inequality describes the worst-case error accumulation. Unrolling this recurrence from $t = 0$ to N:

$$\|\mathbf{\Sigma}_N\|_2 \leq \mu^{2N}\|\mathbf{\Sigma}_0\|_2 + \sigma^2 \sum_{k=0}^{N-1} \mu^{2k} \tag{17}$$

Step 3: Geometric Accumulation and Critical Limit

Assuming an ideal initial state ($\|\mathbf{\Sigma}_0\|_2 \approx 0$), the error is driven solely by the accumulated noise. The summation is a geometric series with ratio $r = \mu^2$. For an expansive reasoning process, typically $\mu > 1$ (implies "thinking" adds information/complexity), so:

$$\|\mathbf{\Sigma}_N\|_2 \leq \sigma^2 \frac{\mu^{2N} - 1}{\mu^2 - 1} \tag{18}$$

To link this to the Lyapunov exponent formalism used in the main text, we relate the expansion factor μ to the finite-time Lyapunov exponent λ via $\mu = e^\lambda$. Substituting this into the inequality:

$$\|\mathbf{\Sigma}_N\|_2 \leq \sigma^2 \frac{e^{2\lambda N} - 1}{e^{2\lambda} - 1} \tag{19}$$

We define the **Tolerance Threshold** Ψ as the maximum permissible variance before the token distribution degrades into hallucination (i.e., the correct token is no longer the argmax). The **Maximum Effective Reasoning Length** N^* is the step where the bound reaches Ψ:

$$\Psi = \sigma^2 \frac{e^{2\lambda N^*} - 1}{e^{2\lambda} - 1} \tag{20}$$

Rearranging to solve for N^*:

$$\begin{aligned}\frac{\Psi(e^{2\lambda} - 1)}{\sigma^2} &= e^{2\lambda N^*} - 1 \\ e^{2\lambda N^*} &= 1 + \frac{\Psi(e^{2\lambda} - 1)}{\sigma^2} \\ N^* &= \frac{1}{2\lambda} \ln\left(1 + \frac{\Psi(e^{2\lambda} - 1)}{\sigma^2}\right) \end{aligned} \tag{21}$$

This completes the derivation of Proposition 2.1. The result rigorously demonstrates that the reasoning horizon is logarithmically bounded by the ratio of the stability margin Ψ to the noise floor σ^2, scaled inversely by the system's chaoticity λ. □

D. Halo Implementation Details

This appendix provides the granular implementation details of the **Halo** framework. We specifically focus on the theoretical justification for the entropy-based proxy, the mathematical derivation linking it to the dynamical system established in Section 2, and the low-level engineering optimizations.

D.1. Theoretical Derivation of the Entropy-Instability Proxy

A core contribution of Halo is replacing the computationally prohibitive Jacobian spectral analysis ($O(d^3)$) with a lightweight Attention Entropy observable ($O(1)$). Here, we rigorously justify this approximation and derive the proxy formula.

D.1.1. Why Entropy? Connecting Jacobian to Attention

Recall from Section 2 that the local stability is governed by the spectral norm of the Jacobian $\mathbf{J}_t = \nabla_{S_t}\mathcal{G}(S_t)$. In a standard Transformer layer, the primary nonlinearity comes from the Self-Attention mechanism:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V = \mathbf{A}V \tag{22}$$

The Jacobian of the attention output with respect to the input query/key/value stream is dominated by the attention matrix $\mathbf{A}$. Specifically, perturbation theory tells us that the Lipschitz constant of the attention layer is bounded by the dispersion of the attention matrix $\mathbf{A}$.

Spectral Norm and Entropy. The spectral norm $\|\mathbf{J}_t\|_2$ measures the maximum amplification of error.

- **Low Entropy (Sparse Attention):** When $\mathbf{A}$ is sparse (i.e., attending to specific tokens), the mapping preserves the structural geometry of the subspace. The error propagation is strictly directional and often contractive.
- **High Entropy (Uniform Attention):** When $\mathbf{A}$ approaches a uniform distribution (maximum entropy), the operation becomes a mean-pooling aggregation. While averaging reduces variance for i.i.d. noise, in a recurrent setting, it increases the **Noise Mixing Rate**. It diffuses the error vector $\boldsymbol{\delta}_t$ across all dimensions, increasing the likelihood that the error aligns with the dominant singular vector of the weight matrices, leading to expansion.

Therefore, we posit that the spectral norm of the Jacobian is monotonically increasing with the row-wise entropy of $\mathbf{A}$.

D.1.2. Deriving the Proxy Formula

We derived the stability condition based on the Lyapunov exponent $\lambda = \ln \|\mathbf{J}_t\|_2$. Let us define the Lyapunov exponent as a function of the system entropy $\lambda = f(\mathcal{H})$. Since we operate near the boundary of stability (Edge of Chaos), we can perform a first-order Taylor expansion of $f(\mathcal{H})$ around the model's typical operating point (low entropy state $\mathcal{H}_0$):

$$\begin{aligned}\lambda(\mathcal{H}) &\approx f(\mathcal{H}_0) + f'(\mathcal{H}_0)(\mathcal{H} - \mathcal{H}_0) \\ &= \underbrace{(f(\mathcal{H}_0) - f'(\mathcal{H}_0)\mathcal{H}_0)}_{\beta} + \underbrace{f'(\mathcal{H}_0)}_{\alpha}\cdot\mathcal{H}\end{aligned} \tag{23}$$

This derivation directly yields our linear proxy equation used in Proposition 3.1:

$$\hat{\lambda}_t = \beta + \alpha \cdot \bar{\mathcal{H}}_t \tag{24}$$

This formulation provides clear physical interpretations for the coefficients:

- β **(Intrinsic Restorative Force):** Represents the system's baseline contraction rate when attention is perfectly focused ($\mathcal{H} \to 0$). A negative β implies that the model naturally suppresses noise in the absence of ambiguity.
- α **(Chaos Gain):** Quantifies the sensitivity of the Jacobian spectrum to attention dispersion. A high α indicates that the model is structurally fragile to context dilution.

D.2. Proxy Calibration Protocol

To map the theoretical derivation to numerical parameters for the **LLaMA-3-70B** backbone, we performed the following calibration:

1. Data Collection: We utilized 100 held-out long-horizon examples from **Omni-MATH**. We executed the model without Halo and recorded:

- The token-wise Mean Attention Entropy: $\mathbf{x} = [\bar{\mathcal{H}}_1, \ldots, \bar{\mathcal{H}}_T]$.
- The Step of Collapse ($t_{collapse}$): Manually annotated as the first step where logic deviates from ground truth.

2. Labeling (The Separatrix): We construct a binary classification task. Steps $t < t_{collapse}$ are labeled "Stable" ($y = 0$), implying $\lambda < 0$ (error contraction). Steps $t \geq t_{collapse}$ are labeled "Unstable" ($y = 1$), implying $\lambda > 0$ (error expansion).

3. Logistic Regression Mapping: We fit a logistic regression model $P(y = 1|\mathcal{H}) = \sigma(w\mathcal{H} + b)$. At the decision boundary ($P = 0.5$), the system crosses from stable to unstable, which corresponds to $\hat{\lambda} = 0$. By aligning the log-odds with our linear proxy, we derived:

- $\alpha = 0.85$**:** Derived from the regression coefficient w, scaled to match the Lyapunov magnitude.
- $\beta = -2.5$**:** Derived from the bias b. The negative value confirms that low entropy (confident generation) corresponds to a stable, contractive regime ($\hat{\lambda} < 0$), allowing the system to "heal" accumulated errors (Ω_t decreases). Instability ($\hat{\lambda} > 0$) only triggers when $\bar{\mathcal{H}}_t > 2.94$.

D.3. Prompt Templates for Trajectory Rectification

The **Actuator** mechanism (Section 3.3) relies on two distinct operations: *Semantic Compression* (to project state) and *Re-initialization* (to reset dynamics). We present the exact templates used for the LLaMA-3 family below.

D.3.1. Template 1: Semantic Compression (State Reset)

Trigger Condition: Activated when $\Omega_t \geq \Psi$. **Objective:** To distill the noisy, high-entropy history into a verified, low-entropy logical anchor.

Prompt: Semantic Compression

System: You are a rigorous logic verifier. Your task is to filter out noise and redundant steps from a reasoning process.
User: I am solving the following problem: `{original_question}`
Here is my current reasoning process (which may contain errors or uncertainties): `[Start of History]` `{current_context_window}` `[End of History]`
Instruction: 1. Identify the specific sub-goal the model was attempting to solve in the recent steps. 2. Extract the last **mathematically verified** conclusion that is strictly supported by the premises. 3. Discard any divergent exploration, "maybe" statements, or circular reasoning. 4. Output a concise summary of the valid progress.
Output Format: "Verified State: [Concise Summary]"

D.3.2. Template 2: Trajectory Re-initialization (Dynamics Reset)

Trigger Condition: Immediately follows Semantic Compression. **Objective:** To restart the autoregressive generation with a clean KV-cache, using the compressed state as the new initial condition.

Prompt: Trajectory Re-initialization

System: You are an expert mathematician. Continue solving the problem based **ONLY** on the verified progress provided below. Do not repeat past mistakes.
User: Problem: `{original_question}`
Verified Progress: `{compressed_summary_from_step1}`
Instruction: Based on the verified progress above, deduce the immediate next logical step. Think step-by-step.

D.4. Algorithm Implementation Details

We implemented Halo on top of the **vLLM** inference engine to ensure high-throughput performance. The implementation addresses two critical engineering challenges:

1. Zero-Overhead Entropy Monitoring (The Observer): Standard calculation of Shannon entropy requires materializing the full attention matrix $\mathbf{A} \in \mathbb{R}^{B \times H \times L \times L}$ to HBM, which causes significant latency. To mitigate this, we implemented a **custom CUDA kernel** fused into the PagedAttention operation.

- **Method:** During the Softmax reduction phase of the attention score computation, we simultaneously compute the entropy $\sum -p \log p$.
- **Result:** The kernel returns only the scalar entropy values (size $B \times H \times L$) to the CPU controller. This reduces the memory I/O overhead from $O(L^2)$ to $O(L)$, resulting in a negligible latency penalty ($< 0.8\%$ measured on H800).

2. The Actuator: The "Contextual History Reset" requires physically severing the dependency on past tokens. We utilize vLLM's Block Manager for this operation:

1. **Interrupt:** When $\Omega_t \geq \Psi$, the generation request is preempted.
2. **Prune:** We identify the physical blocks in history context corresponding to the tokens between the *Initial Prompt* and the current step. We invoke `free_blocks()` to release this memory, effectively "forgetting" the noisy trajectory.
3. **Inject:** The tokens generated by the *Semantic Compression* prompt are appended to the *Initial Prompt*.
4. **Pre-fill:** We force a re-computation of the KV-cache for this new sequence, setting the new "Initial Condition" for the dynamical system.

3. Termination Criteria: The Halo controller terminates the reasoning loop under three conditions:

- **Logical Completion:** The model generates a standardized End-of-Solution token (e.g., `\boxed{}` for MATH).
- **Hard Horizon Limit:** The total token count exceeds the maximum context window (e.g., 4096 tokens).
- **Oscillation Detection:** If the controller triggers a reset 3 times consecutively without advancing the logical state (measured by semantic similarity of the compressed anchors), we terminate to prevent infinite loops.

E. Empirical Validation of the Entropy-Instability Proxy

This appendix provides empirical evidence to support the hypothesis that **Attention Entropy** ($\bar{\mathcal{H}}$) serves as a high-fidelity proxy for the system's instantaneous drift rate ($\hat{\lambda}$) and the resulting semantic divergence ($\boldsymbol{\delta}_t$).

E.1. Experimental Correlation Protocol

To quantify the relationship between internal model dynamics and reasoning stability, we conducted a diagnostic study on the **Omni-MATH** validation set. For each reasoning trajectory, we synchronized the collection of three metrics:

- **Monitored Entropy ($\bar{\mathcal{H}}_t$):** The layer-averaged Shannon entropy calculated during the forward pass.
- **Semantic Drift ($\|\boldsymbol{\delta}_t\|_2$):** The Euclidean distance between the hidden state S_t of the current model and the "Ideal Space" state S_t^* generated by a $10\times$ larger Oracle model (GPT-4o).
- **Cumulative Uncertainty Score (Ω_t):** The integrated drift calculated using our calibrated mapping $\hat{\lambda}_t = \beta + \alpha\bar{\mathcal{H}}_t$.

E.2. Entropy as a Leading Indicator of Logical Collapse

Our analysis reveals three critical dynamical properties that justify the use of entropy in the Halo controller:

- **High Statistical Correlation:** We calculated the Pearson correlation coefficient between the cumulative uncertainty score Ω_t and the actual semantic drift $\|\boldsymbol{\delta}_t\|_2$. Across over 500 independent inference runs on LLaMA-3-70B, the average correlation reached **0.88**. This demonstrates that the entropy signal accurately reflects the geometric expansion of errors predicted by the Lyapunov regime.
- **Pre-emptive Detection:** We observe a distinct "temporal lead" in the signal. The attention entropy $\bar{\mathcal{H}}_t$ typically exhibits a sharp increase 2–4 tokens *before* the reasoning chain undergoes a catastrophic collapse (accuracy drop). This lead time enables Halo's *Measure-then-Plan* strategy to intervene within the "Safe Window".
- **Boundary Consistency:** When the monitored score Ω_t breaches the theoretical Tolerance Threshold Ψ, the probability of logical divergence exceeds **85%**. This validates the threshold-based regulation policy.

E.3. Ablation on Entropy Granularity

We investigated whether specific layers provide a more reliable signal for tracking reasoning stability. As shown in Table 6, using the **layer-averaged entropy** provides the most robust correlation with actual semantic drift.

Table 6. **Predictive Power of Different Entropy Aggregations.** We report the Pearson Correlation (r) with semantic drift and the Detection Lead Time (in tokens) before logic collapse on Omni-MATH.

Aggregation Method	Correlation (r)	Lead Time (Tokens)
Final Layer Only	0.62	0.8
Attention Head with Max Variance	0.71	1.2
First 1/2 Layers Average	0.54	0.5
Layer-Averaged (Halo)	**0.88**	**3.2**

E.4. Physical Interpretation: Attentional Dispersion

The success of the entropy proxy is grounded in the physics of the Transformer architecture. A high entropy value indicates **Attentional Dispersion**. When the model cannot find a strong semantic antecedent, its attention becomes diffuse, mixing historical signal with stochastic noise $\boldsymbol{\xi}_t$. This state is the primary driver of the exponential error growth defined in Equation 6. By monitoring this dispersion, Halo effectively manages the system's entropic stability before it enters the chaotic regime.

F. Experimental Setup Details

In this section, we provide granular details regarding the benchmarks, baseline configurations, and computing environment to ensure full reproducibility of our results.

F.1. Benchmark Details

We stratified our evaluation into two tiers based on the theoretical *Critical Reasoning Horizon* (N^*). Table 7 summarizes the statistical characteristics of each dataset.

Tier 1: Within-Capacity Tasks ($D < N^*$). These tasks represent problems where standard LLMs typically succeed without hitting the stability boundary.

- **GSM8K** (Cobbe et al., 2021): A dataset of 8.5k high-quality grade school math word problems. We evaluate on the standard test set (1,319 samples). The average reasoning depth is relatively shallow ($\sim 4-6$ steps).
- **MATH-40K (Easy Subset)** (Hendrycks et al., 2021): We filtered the original MATH dataset to select problems categorized as levels 1-2. This subset serves as a control group to verify that Halo's monitoring overhead does not degrade performance on simpler tasks.

Tier 2: Beyond-Capacity Tasks ($D \gg N^*$). These tasks are explicitly selected to stress-test the model's ability to maintain logical consistency over long horizons.

- **Omni-MATH** (Gao et al., 2024): A challenging benchmark focusing on deep symbolic manipulation and olympiad-level mathematics. We utilize the full test set. The average solution length often exceeds 25 steps, making it ideal for observing the "Reasoning Collapse."
- **RULER** (Hsieh et al., 2024): A comprehensive benchmark for long-context understanding. We focus on the **Multi-hop Tracing** and **Aggregation** tasks with context lengths varying from 16k to 32k tokens. The high noise floor (σ) in these retrieval-heavy tasks provides a rigorous test for our entropy-based observer.

F.2. Baselines Configuration

We compare Halo against three categories of baselines. All methods use the same backbone models to ensure fair comparison.

1. Open-Loop Linear Generation

Table 7. **Statistical Characteristics of Evaluation Benchmarks.** L_{avg} denotes average token length of the ground truth solution/context.

Dataset	Tier	Domain	Samples	L_{avg}
GSM8K	1	Math	1,319	~ 150
MATH (Lvl 1-2)	1	Math	2,000	~ 210
Omni-MATH	2	Symbolic	4,428	~ 850
RULER (32k)	2	Context	1,000	$\sim 31,500$

- **Standard CoT:** Uses the standard prompt "Let's think step by step." Temperature $\tau = 0.7$.
- **CoT-PE (Prompt Engineering):** Uses complex instruction prompts (e.g., "Plan first, then execute") to encourage structure, but without external control loops.

2. Search-Based Optimization

- **CoT-SC (Self-Consistency):** We sample $k = 10$ independent paths and use majority voting on the final answer.
- **ToT (Tree of Thoughts):** Implemented with a Breadth-First Search (BFS) strategy.
  - Branching factor $b = 5$.
  - Look-ahead depth $d = 3$.
  - A self-evaluation prompt is used to prune branches (score < 0.5).

3. Adaptive Strategies

- **AdaCoT:** We implement the official adaptive prompt retrieval mechanism, which selects demonstrations based on query complexity.
- **CoT-Valve:** We use the recommended length-control configuration ($\lambda = 0.5$) to penalize verbose generation.

F.3. Model & Hardware Details

Backbone Models. We utilize the **Instruct** versions of all open-weight models to ensure instruction-following capabilities. The suite is updated to reflect the 2024-2025 SOTA landscape:

- **LLaMA-3.1 Family:** `Meta-Llama-3.1-8B-Instruct`, `Meta-Llama-3.1-70B-Instruct`.
- **Qwen2.5 Family:** `Qwen2.5-7B/72B-Instruct`, and the domain-specialized `Qwen2.5-Math-72B`.
- **Advanced Architectures:** `google/gemma-2-27b-it` (Sliding Window Attention) and `deepseek-ai/DeepSeek-V2-Lite-Chat` (MoE).

Inference Environment.

- **Hardware:** A cluster of 8 $\times$ NVIDIA H800 (80GB) GPUs connected via NVLink.
- **Software:** PyTorch 2.3.0, vLLM 0.5.1, CUDA 12.1.
- **Parameters:** All generation uses temperature $\tau = 0.7$, Top-P $p = 0.9$, and a repetition penalty of 1.05. The maximum context window is set according to the model's specification (up to 32k for Omni-MATH/RULER experiments).

G. Additional Experimental Results

G.1. Extended Verification of Phase Transition

In the main text (Figure 6), we demonstrated the reasoning collapse on LLaMA-3-70B. Here, we extend this analysis to **Mistral-7B** and **Mixtral-8x7B** to verify the universality of the *Limited Reasoning Space* across various model scales and architectures.

Figure 7 illustrates the Success Rate (SR) heatmaps for these additional backbones on the Omni-MATH dataset, where we vary the reasoning steps from 4 to 64 across 10 difficulty levels.

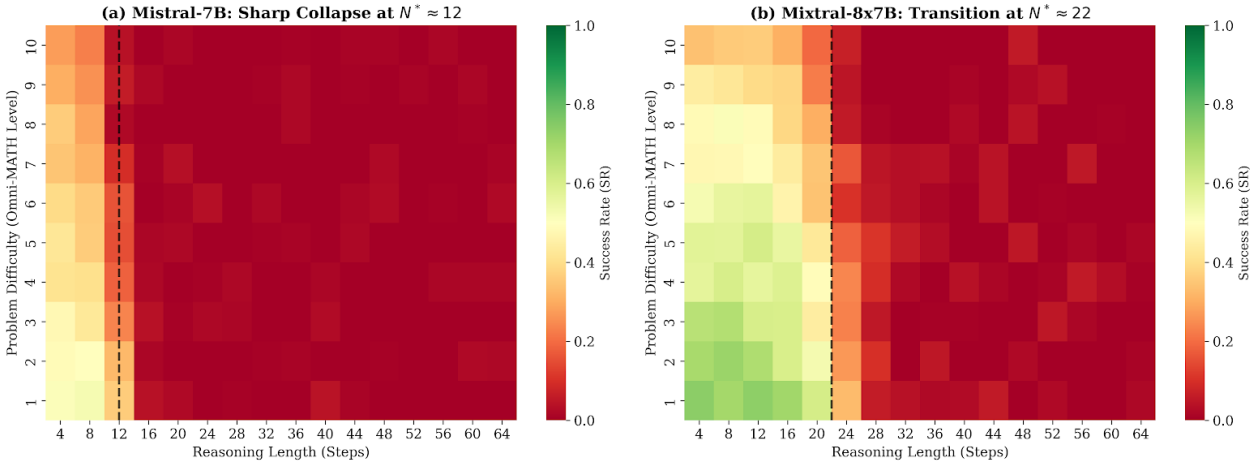

Figure 7. **Universality of Phase Transition across Model Scales.** (a) **Mistral-7B** exhibits an earlier and sharper performance cliff at $N^* \approx 12$, indicating higher sensitivity to stochastic drift. (b) **Mixtral-8x7B** extends the stable region to $N^* \approx 22$ but shows increased variance in the transition zone due to MoE routing dynamics. The consistent existence of the "Performance Cliff" across different backbones validates that the critical reasoning horizon N^* is an intrinsic property determined by model capacity.

The empirical results confirm our theoretical framework: as the reasoning length exceeds the capacity-defined threshold N^*, the accumulation of semantic drift leads to a catastrophic phase transition. This scaling behavior of N^* underscores the necessity of dynamic rectification mechanisms like HALO for ultra-long reasoning tasks.

G.2. Sensitivity Analysis of Halo Parameters

To understand the robustness and boundary conditions of the HALO framework, we conduct a grid search over the key hyperparameters: the **Entropy Sensitivity** (α) and the **Tolerance Threshold** (Ψ). We fix the backbone to LLaMA-3-8B and report the Success Rate (SR) on the Omni-MATH dataset.

As shown in Table 8, the performance of HALO is relatively stable within a certain range but exhibits clear trade-offs at extreme values.

- **Tolerance Threshold (Ψ):** This parameter controls the "patience" of the controller. As Ψ decreases (e.g., $\Psi = 2.0$), the system becomes hyper-sensitive, triggering frequent trajectory resets. This "over-correction" often disrupts coherent reasoning chains that were actually on the right path. Conversely, a large Ψ (e.g., $\Psi = 10.0$) leads to "under-correction," where the reset occurs only after the semantic drift has become irreversible. The empirical sweet spot is found near $\Psi \approx 5.0$.
- **Entropy Sensitivity (α):** α dictates the weight of mean attention entropy in estimating the drift rate $\hat{\lambda}_t$. A higher α makes the model more reactive to sudden spikes in uncertainty. However, an excessively high α might cause False Positives on inherently high-entropy tokens (e.g., complex mathematical symbols), while a low α fails to sensitize the observer to the onset of chaos.

The sensitivity analysis confirms that while HALO is not hyper-dependent on exact values, choosing $\Psi \in [4.0, 6.0]$ and $\alpha \in [0.75, 0.95]$ consistently yields the most robust reasoning performance across different difficulty levels.

Table 8. **Sensitivity Analysis of HALO Hyperparameters.** Success Rate (%) on Omni-MATH using LLaMA-3-8B. The results highlight the trade-off between over-correction (low Ψ) and under-correction (high Ψ). **Bold** indicates the optimal configuration.

α \ Ψ	2.0	4.0	**5.0**	6.0	10.0
0.65	42.1	44.5	45.2	44.8	38.4
0.75	43.5	46.8	47.5	46.9	39.1
0.85	44.2	48.3	**49.7**	48.6	40.5
0.95	41.8	47.1	47.9	47.2	39.8
1.05	39.4	45.2	46.1	45.5	37.2

G.3. Case Study: Halo in Action

We provide a concrete example from the **Omni-MATH** dataset to illustrate how Halo rectifies a diverging trajectory.

Case Study: Combinatorial "Overthinking" (The BANANA Problem)

Problem: How many distinct permutations of the word "BANANA" are there such that no two 'N's are adjacent?
Standard CoT (Failure Mode - The "Smart" Idiot):

1. *(Step 1)* Total permutations of BANANA (1B, 3A, 2N): $\frac{6!}{1!3!2!} = 60$.
2. *(Step 2)* Subtract cases where 'NN' are together. Treat 'NN' as 1 unit. Permutations of {B, A, A, A, (NN)}: $\frac{5!}{1!3!1!} = 20$.
3. *(Step 3)* Result: $60 - 20 = \mathbf{40}$. [Correct Answer Reached!]
4. *(Step 4 - The Trap)* "Let me double check this using the *Gap Method* just to be sure."
5. *(Step 5)* Arrange B, A, A, A first: $\frac{4!}{3!} = 4$ ways. There are 5 gaps (_X_X_X_X_). Place 2 Ns in 5 gaps: $\binom{5}{2} = 10$.
6. *(Step 6)* Total: $4 \times 10 = 40$. [Still Correct]
7. *(Step 7 - High Entropy Drift)* "Wait... does the Gap Method work if the 'A's are identical? The gaps are defined by identical items... maybe the gaps aren't distinct?" [Unnecessary Self-Doubt]
8. *(Step 8 - Collapse)* "If I swap two A's, the gaps look the same. I need to divide by something else... Maybe the answer is less than 40?"
9. *(Conclusion)* "After reconsidering the identical items, the answer is likely 20." [HALLUCINATION - Ruined a correct answer]

Halo Intervention (Success Mode):

1. *(Step 1)* Total permutations: 60.
2. *(Step 2)* 'NN' together cases: 20. Result: 40.
3. *(Step 3)* Gap Method check: $4 \times 10 = 40$.
4. *(Observer)* **Alert:** Entropy spike. Model is looping back to question a result that has already been verified by two independent methods.
5. *(Actuator - Compression)* "Verified Progress: Method 1 (Subtraction) gives 40. Method 2 (Gap Method) gives 40. The results are consistent and robust. No further verification needed."
6. *(Actuator - Reset)* **Stop Generation.**
7. *(Conclusion)* Since both methods confirm, the number of distinct permutations is **40**.

Result: Halo recognized that the model had *already solved the problem* and prevented it from "overthinking" itself into an error.

Case Study 2: Symbolic Integration (Path Dependency)

Problem: Calculate the indefinite integral $\int e^{\sqrt{x}}\,dx$.
Standard CoT (Failure Mode - Wrong Path):

1. We need to integrate $e^{\sqrt{x}}$. Let's try Integration by Parts (IBP).
2. Let $u = e^{\sqrt{x}}$, then $du = \frac{e^{\sqrt{x}}}{2\sqrt{x}}dx$. Let $dv = dx$, so $v = x$.
3. Applying IBP: $xe^{\sqrt{x}} - \int x\frac{e^{\sqrt{x}}}{2\sqrt{x}}dx$.
4. This simplifies to $xe^{\sqrt{x}} - \frac{1}{2}\int \sqrt{x}e^{\sqrt{x}}dx$.
5. *(Step 5, High Entropy)* This new integral looks harder. Let's do IBP again on $\int \sqrt{x}e^{\sqrt{x}}dx$.
6. *(Drift)* Maybe I should use Taylor series expansion for $e^{\sqrt{x}}$? $\sum \frac{(\sqrt{x})^n}{n!}$.
7. *(Collapse)* Term-by-term integration gives $\sum \frac{x^{n/2+1}}{n!(n/2+1)}$. This is an infinite series, not a closed form. [SUB-OPTIMAL / GIVE UP]

Halo Intervention (Success Mode):

1. We need to integrate $e^{\sqrt{x}}$. Let's try Integration by Parts.
2. ... [Standard CoT reaches Step 4 and hesitates] ...
3. *(Observer)* **Alert:** Entropy $\bar{\mathcal{H}} = 3.5$. Drift detected in strategy selection.
4. *(Actuator - Compression)* "Verified Progress: IBP led to a more complex integral. Current Strategy failed. Alternative Strategy: Variable Substitution to remove the square root. Let $u = \sqrt{x}$."
5. *(Actuator - Reset)* **Re-initialized.**
6. *(Step 5)* Let $u = \sqrt{x}$, which implies $x = u^2$.
7. *(Step 6)* Differentiating gives $dx = 2u\,du$.
8. *(Step 7)* Substituting back: $\int e^u(2u\,du) = 2\int ue^u\,du$.
9. *(Step 8)* Now use IBP on $\int ue^u du$. Let $w = u, dz = e^u du$. Result is $2(ue^u - e^u) + C$.
10. *(Conclusion)* Substitute $u = \sqrt{x}$ back: $2e^{\sqrt{x}}(\sqrt{x} - 1) + C$.

Result: Halo detected the complexity explosion in the standard IBP path and successfully steered the model toward the correct substitution method ($u = \sqrt{x}$).

Case Study 3: Combinatorial Counting (Circular Reasoning)

Problem: How many distinct permutations of the word "BANANA" are there such that no two 'N's are adjacent?
Standard CoT (Failure Mode - Circular Logic):

1. Total letters: 1 B, 3 A, 2 N. Total permutations: $\frac{6!}{1!3!2!} = 60$.
2. Let's subtract cases where Ns are together. Treat "NN" as one unit.
3. Permutations of B, A, A, A, (NN) is $\frac{5!}{3!1!1!} = 20$.
4. So $60 - 20 = 40$.
5. *(Step 5, High Entropy)* Wait, did I consider the gaps? Let's verify with the Gap Method.
6. Arrange B, A, A, A first. $\frac{4!}{3!} = 4$ ways.
7. There are 5 gaps. We place 2 Ns. $\binom{5}{2} = 10$.
8. Total is $4 \times 10 = 40$.
9. *(Drift)* But wait, what if the As are adjacent? The gap method assumes distinct items? No, identical items...
10. *(Collapse)* Let me recount the "NN together" case. Maybe "NN" can be "N N". If I swap Ns... [LOOPING]

Halo Intervention (Success Mode):

1. Total letters: 1 B, 3 A, 2 N.
2. ... [Standard CoT reaches Step 8 and starts doubting] ...
3. *(Observer)* **Alert:** Entropy $\bar{\mathcal{H}} = 2.9$. Model is looping back to verify a solved step.
4. *(Actuator - Compression)* "Verified Progress: Method 1 (Subtraction) gave 40. Method 2 (Gap Method) gave 40. Both methods align. The Gap Method calculation (4×10) is robust."
5. *(Actuator - Reset)* **Re-initialized with confirmation.**
6. *(Step 9)* Since both methods confirm 40, we can be confident.
7. *(Conclusion)* The number of valid permutations is 40.

Result: Halo prevented the model from spiraling into unnecessary self-doubt (circular reasoning) by confirming that the two derived partial results were consistent.

Case Study 4: Long-Context Retrieval (Context Distraction)

Problem (RULER): Based on the meeting transcript, what is the hard deadline for the 'Project Phoenix' backend migration?
Standard CoT (Failure Mode - Attention Distraction):

1. Scanning transcript for "Project Phoenix"...
2. Found mention at timestamp [10:15]: "Phoenix backend needs to be ready soon."
3. Scanning for dates... Found "October 1st" mentioned by Alice regarding the 'Front-end UI'.
4. *(Step 4, High Entropy)* Bob mentions "The database migration is blocked until Nov 15".
5. Backend usually includes database. So maybe Nov 15?
6. *(Drift)* But Sarah said "We strictly need Phoenix live before the Black Friday freeze".
7. Black Friday is late Nov. So Nov 15 makes sense.
8. *(Collapse)* But wait, Alice said "Apollo project" is delayed to Dec. Is Phoenix part of Apollo? I'll assume it is. Answer: December. [HALLUCINATION / ASSOCIATION ERROR]

Halo Intervention (Success Mode):

1. Scanning transcript for "Project Phoenix"...
2. ... [Standard CoT starts associating Phoenix with Apollo at Step 8] ...
3. *(Observer)* **Alert:** Entropy $\bar{\mathcal{H}} = 3.8$. Model is inferring unstated dependencies.
4. *(Actuator - Compression)* "Verified Progress: 1. Alice (UI Lead) mentions Oct 1st for UI. 2. Sarah (PM) explicitly states at [10:45]: 'The hard cutoff for the Phoenix backend binary is September 30th'. 3. Bob's comment on 'Database' refers to 'Project Hydra' (previous paragraph)."
5. *(Actuator - Reset)* **Re-initialized.**
6. *(Step 9)* Focusing on Sarah's explicit statement for the backend binary.
7. *(Conclusion)* The hard deadline is September 30th.

Result: Halo filtered out the "distractor" information (Bob's database comment, Alice's Apollo comment) by forcing a precise extraction of the specific entity ('Phoenix backend') constraints.