Is my model "mind blurting"? Interpreting the dynamics of reasoning tokens with Recurrence Quantification Analysis (RQA)

Anonymous ACL submission

Abstract

Test-time compute is central to large reasoning models, yet analysing their reasoning behaviour through generated text is increasingly impractical and unreliable. Response length is often used as a brute proxy for reasoning effort, but this metric fails to capture the dynamics and effectiveness of the Chain of Thoughts (CoT) or the generated tokens. We propose Recurrence Quantification Analysis (RQA) as a non-textual alternative for analysing model's reasoning chains at test time. By treating token generation as a dynamical system, we extract hidden embeddings at each generation step and apply RQA to the resulting trajectories. RQA metrics, including Determinism and Laminarity, quantify patterns of repetition and stalling in the model's latent representations. Analysing 3,600 generation traces from DeepSeek-R1-Distill, we show that RQA captures signals not reflected by response length, but also substantially improves prediction of task complexity by 8%. These results help establish RQA as a principled tool for studying the latent token generation dynamics of test-time scaling in reasoning models. The codes are available here: https://anonymous.4open.science/r/RQA-CoT-844C/README.md.

1 Introduction

The effectiveness of test-time compute through extended "Chain-of-Thought" (CoT) sequences is central to the performance of Large Reasoning Models like Deepseek R1, or OpenAI's o3. To understand this Chain of Thought process, current research focuses on the information flow across layers (Yang et al., 2025b; Ton et al.) or utilises continuous generative frameworks like diffusion to model CoT trajectories (Ye et al.). Another approach is to textually analyse and characterise linguistic behaviours of the Chain of Thought (Venhoff et al., 2025; Marjanovic et al., 2025). On a practical level, researchers often evaluate reasoning by mapping the "Curse of Complexity" against the model's total token output (Lin et al., 2025). These approaches aim to ground the model's linguistic output in either its underlying computational structures or its textual behaviours.

Challenges: However, treating CoT traces as purely linguistic or static entities is increasingly impractical. **First**, the sheer volume of tokens generated during test-time scaling quickly exceeds human audit capacity, creating a verification bottleneck. **Second**, CoT traces are often prone to "overthinking," unfaithfulness, or semantic drifting, where the generated text does not strictly reflect the model's internal logic (Kambhampati et al., 2025; Arcuschin et al.; Sun et al., 2025). **Critically**, existing mechanistic methods often treat the reasoning process as a series of static snapshots or isolated units of analysis. They fail to be effective in characterising the temporal, path-dependent dynamics of how representations evolve in the latent space, leaving a gap in our understanding of the underlying computation process of the Chain of Thought process.

Contributions: In this work, we seek to reduce the reliance on direct textual analysis of Chain-of-Thought (CoT) while explicitly modelling CoT as a *temporal process*. Our contributions are threefold:

- **Dynamical framing of CoT:** We model CoT generation as a dynamical process unfolding in representation space and introduce **Recurrence Quantification Analysis (RQA)** as a mechanistic interpretability framework. By treating the autoregressive generation of tokens as a dynamical system, we apply RQA to trajectories formed by the last-layer embeddings of each generated token. This maps high-dimensional latent trajectories to 2D recurrence plots, enabling the quantification of predictability (**DET**) and complexity (**ENTR**).
- **Empirical validation on reasoning traces:** We

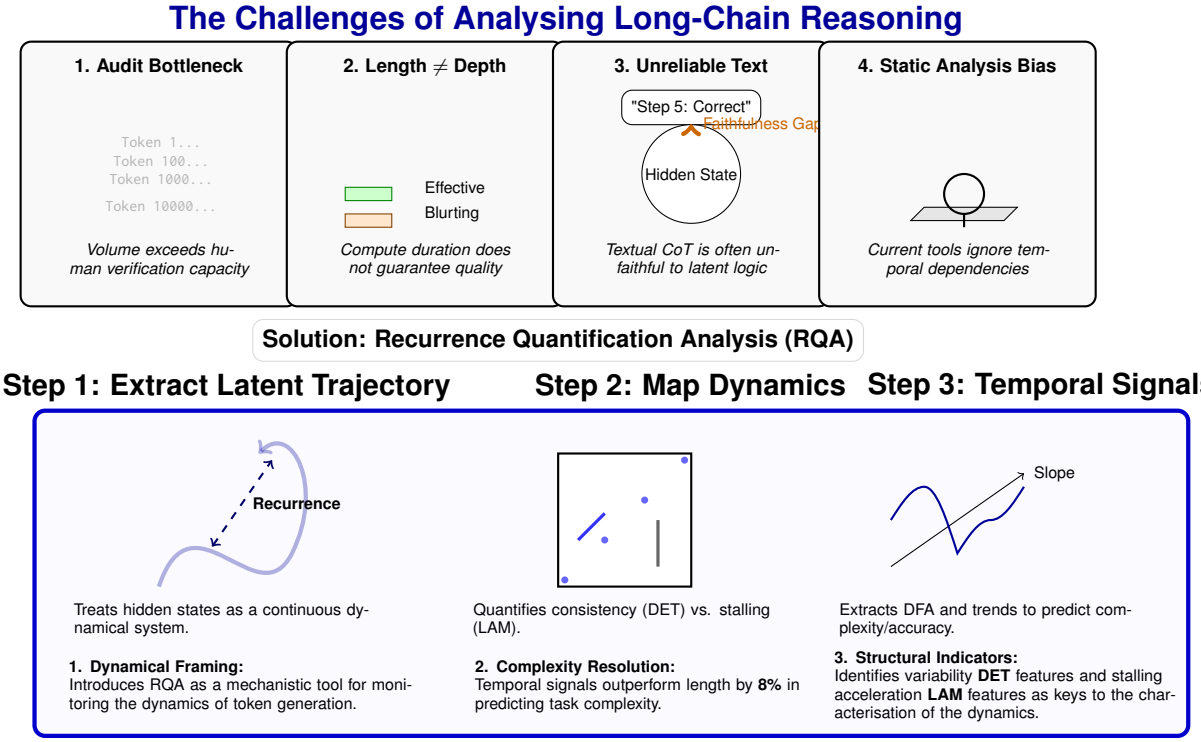

Figure 1: **Problems with CoT Analysis - Top (The Challenges):** (1) Massive token counts prevent human audit; (2) Total length is a poor proxy for reasoning quality; (3) Textual output often masks internal logic failures; (4) Static interpretability fails to see time-dependent patterns. **Bottom (RQA: the proposed solution):** We propose RQA to transform discrete tokens into a measurable **latent trajectory**, decoding the dynamics of reasoning. We demonstrate that these temporal signals significantly outperform response length in resolving task complexity and identifying structural signals of the dynamics.

validate the proposed framework using foundational RQA measures—**Determinism (DET)**, capturing representational consistency, and **Laminarity (LAM)**, reflecting periods of semantic stalling—on 3,600 reasoning traces from *DeepSeek-Distill-7B-Qwen* evaluated on the **ZebraLogic** benchmark. Temporal RQA features significantly outperform length-based baselines for task complexity classification, achieving an 8% improvement (36.9% vs. 29.0%).

- **Identification of structural difficulty indicators:** We show that the *variability of semantic repetition* and the *acceleration of stalling* (Laminarity slope) are strong indicators of combinatorial difficulty.

By doing so, we establish RQA as structure-sensitive framework for monitoring reasoning models, offering a new perspective in analysing the dynamics of Chain of Thought process.

2 Related Work

Alternatives to Textual Analysis: Recent research has explored non-textual metrics to quantify reasoning. Information-theoretic approaches utilise entropy and bottleneck measures to map information flow across layers (Yang et al., 2025b; Ton et al.). Others model CoT trajectories using continuous generative frameworks, such as diffusion or flow matching, to enable self-correcting computation (Ye et al.; Yang et al., 2025a). While principled, these methods often require architectural modifications; RQA, by contrast, operates on the existing residual stream of any LRM.

Mechanistic Interpretability of Chain-of-Thought: There have been some Mechanistic works on analysing Chain of Thoughts (Tang et al., 2025; Bogdan et al., 2025; Dutta et al., 2024). Tang et al. (Tang et al., 2025) identify reasoning-critical neurons in feed-forward layers by contrasting activation patterns on high- and low-quality reasoning traces. Chen et al (Chen et al., 2025) extract latent feature directions from hidden states to study CoT faithfulness by using sparse autoencoders on the final-token activations with and without CoT. Bogdan et al. (Bogdan et al., 2025) took into account of the temporal

nature of the CoT when analysing it. They used the sentence-level attribution methods for the analysis based on sentence as the grounding unit of analysis. While insightful, these methods largely either target individual neurons, features at particular layers or time steps, rather than analysing the temporal structure of a reasoning chain, or use textual sentence as the fundamental unit of analysis without addressing the potential issue of lack of semantics in LLM's CoTs (Kambhampati et al., 2025). As a result, this leaves unaddressed the question of how internal representations evolve over time during the CoT process. This leaves a critical gap at the level of token-wise latent dynamics, where reasoning unfolds as a path-dependent process in representation space.

LLMs as Dynamical Systems: Another perspective is to view autoregressive generation through the lens of dynamical systems. Recent work has conceptualised Transformer inference as a discrete stochastic dynamical process evolving over token steps (Bhargava et al., 2023; Carson and Reisizadeh, 2025). When extended reasoning traces are generated, this process induces effective recurrence through repeated context reuse, functionally resembling recurrent computation (Zhang et al., 2024). From this perspective, the sequence of hidden states associated with each generated token forms a high-dimensional time-series, allowing the application of RQA.

Recurrence Quantification Analysis: Recurrence is a fundamental property of dynamical systems, occurring when trajectories revisit regions of phase space (Eckmann et al., 1995). RQA provides statistical measures to quantify the predictability and complexity of these dynamics without assuming linearity or stationarity (Webber and Marwan, 2015). Recently, RP-based methods have transitioned from descriptive tools in physics to components of machine learning pipelines, where time-series are analysed via geometric structure (Marwan and Kraemer, 2023). RQA has successfully detected regime shifts in neuroscience (Lopes et al., 2021) and finance (Ünal, 2022), but has not yet been applied to the internal representations of large language models.

3 Methodology

3.1 LLM Token Generation as Temporal Recurrent Systems

While Transformers are architecturally feed-forward, inference during Chain-of-Thought (CoT) generation induces temporal dependence through autoregressive context reuse (Fig 2, **Block A**). At each step t, the model generates a token, which is appended to the input sequence and re-embedded at step $t{+}1$. For each generated token, we extract the corresponding final-layer hidden state $h_t \in R^d$ (**Block B**). The ordered sequence

$$\mathcal{T} = \{h_1, h_2, \ldots, h_N\}$$

defines a latent trajectory whose evolution is shaped by repeated reuse of prior outputs. Although no explicit recurrent state is maintained (Zhang et al., 2024), this autoregressive loop *effectively approximates recurrent dynamics*, inducing structured temporal dependencies in the latent representation space.

3.2 Recurrence Rate

By treating intermediate token generation as a time-ordered, high-dimensional trajectory $\mathcal{T} = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N\}$, we analyse reasoning dynamics through *Recurrence Quantification Analysis* (RQA). The core object of RQA is the **recurrence matrix**, which encodes pairwise revisitations of the latent state space.

Given a distance norm $d(\cdot, \cdot)$ on R^d and a recurrence threshold ε, the recurrence matrix $\mathbf{R} \in \{0, 1\}^{N \times N}$ is defined element-wise as

$$R_{i,j} = \Theta(\varepsilon - d(\mathbf{h}_i, \mathbf{h}_j)), \qquad (1)$$

where $\Theta(\cdot)$ denotes the Heaviside step function. In our experiments, $d(\cdot, \cdot)$ is instantiated as cosine distance.

This construction can be viewed as the application of a non-linear operator $\mathcal{G}_\varepsilon$ that maps a high-dimensional trajectory to a binary recurrence structure (**Block C**):

$$\mathbf{R} = \mathcal{G}_\varepsilon(\mathcal{T}).$$

The resulting **Recurrence Plot** is the graphical visualisation of $\mathbf{R}$, where a point is plotted at coordinates (i, j) if and only if $R_{i,j} = 1$. This transformation projects the latent trajectory into a two-dimensional representation that preserves the semantic structure of the underlying dynamics.

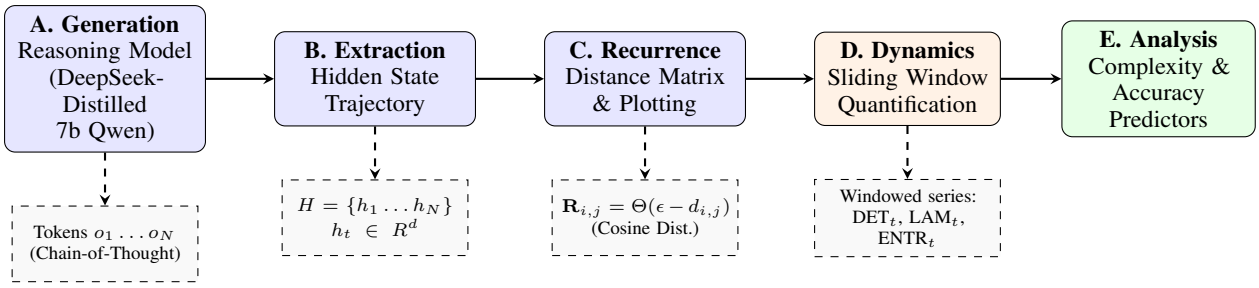

Figure 2: The proposed RQA interpretability pipeline. (A) Tokens are generated autoregressively. (B) Latent states form a high-dimensional trajectory. (C) Self-similarity is mapped to a recurrence matrix. (D) Non-stationary dynamics are quantified via sliding windows. (E) Temporal features (slopes, DFA) serve as inputs for downstream classification.

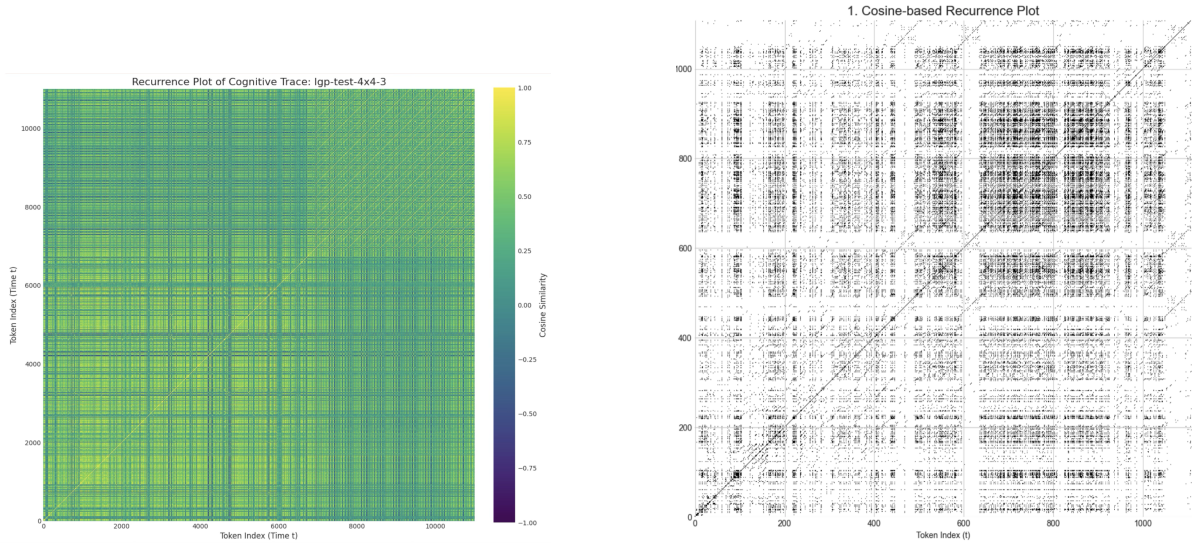

Figure 3: Hidden-state trajectory visualisation. Cosine similarity and recurrence plot ($\epsilon = 0.1$) illustrating DET and LAM structures.

Figure 3 illustrates the result of the process, showing the cosine-similarity matrix, its binarisation using a 10^{th} percentile threshold, and the resulting recurrence structures.

3.3 Quantification Metrics

Figure 2 demonstrates the extraction process, aimining to characterise the dynamics of the representation token generation process. Following the binarisation, we apply RQA to extract structural signatures (**Block D**). Unlike textual analysis, RQA measures the *geometry* of the computation: Here, in accordance with common practices (Webber and Marwan, 2015), we exclude the main diagonal ($i = j$) from all calculations to avoid inflationary signal.

Determinism (DET): DET measures the proportion of recurrence points that form diagonal lines of at least length $l_{\min} = 3$:

$$\text{DET} = \frac{\sum_{l=l_{\min}}^{N} l\, P(l)}{\sum_{i,j \neq i}^{N} R_{i,j}}, \tag{2}$$

where $P(l)$ denotes the histogram of diagonal line lengths. Geometrically, diagonal lines correspond to parallel trajectories in phase space, indicating **predictability and structural consistency** (Marwan and Kraemer, 2023). In our context, high DET serves as an index of **semantic repetition**, reflecting the execution of stable and predictable representations.

Laminarity (LAM): LAM quantifies the proportion of recurrence points forming vertical or horizontal lines of minimum length $v_{\min} = 3$:

$$\text{LAM} = \frac{\sum_{v=v_{\min}}^{N} v\, P(v)}{\sum_{i,j \neq i}^{N} R_{i,j}}, \tag{3}$$

where $P(v)$ is the histogram of vertical line lengths. LAM captures **laminar regimes** in which the sys-

tem remains confined to a localised region of state space (Webber and Marwan, 2015). We interpret high LAM as **semantic stalling**, indicating that the model stays in a static semantic configuration (e.g., repeatedly re-evaluating constraints) without substantive representation progression.

Recurrence Entropy (ENTR): ENTR is defined as the Shannon entropy of the probability distribution of diagonal line lengths $p(l)$:

$$\mathrm{ENTR} = -\sum_{l=l_{\min}}^{N} p(l)\,\ln p(l). \tag{4}$$

ENTR measures the **diversity of the deterministic structure** (Webber and Marwan, 2015). High ENTR reflects a heterogeneous repertoire of reasoning routines, characteristic of deep, multi-stage search processes, whereas low ENTR indicates either repetitive loops (stalling) or a single dominant reasoning trajectory.

Temporal Dynamics: As LLM reasoning dynamics are inherently **non-stationary** (Marwan and Kraemer, 2023), we compute RQA metrics over successive sliding windows ($W = 150$, step size $= 15$). From each resulting metric time series (e.g., DET_t), we extract the **detrended fluctuation analysis (DFA)** (Peng et al., 1994) scaling exponent and the linear trend slope. These temporal descriptors capture transitions of model latent representations throughout the token generation process.

4 Experiment

4.1 Dataset and Task: ZebraLogic

When evaluating model's performance, we seek to examine its logical capability without being bias by its knowledge, and a structured dataset that allows use to examine model's behaviours across increasing levels of difficulty. To this end, we use the **ZebraLogic** benchmark (Lin et al., 2025), an evaluation suite designed to assess the logical reasoning capabilities of Large Language Models (LLMs) through synthetic logic-grid puzzles. Each instance is formulated as a **Constraint Satisfaction Problem (CSP)** consisting of N houses and M features per house. Each feature takes exactly N possible values, and the model must infer a unique assignment for every feature–house pair based on a set of linguistic clues.

Task difficulty is grounded in the factorial growth of the underlying combinatorial search space. For an $N \times M$ puzzle, the number of possible assignments prior to applying any constraints is:

$$\text{Search Space Size} = (N!)^M \tag{5}$$

This factorial dependence induces a rapid combinatorial explosion as grid dimensions increase, providing a rigorous, language-independent measure of task complexity.

4.2 Data Collection and Preliminary Performance

Difficulty	Incorrect	Correct	Total	Mean Acc.
2×3	25	375	400	0.9375
2×4	85	315	400	0.7875
2×5	147	253	400	0.6325
3×2	57	343	400	0.8575
3×3	179	221	400	0.5525
3×4	276	124	400	0.3100
4×3	310	90	400	0.2250
4×4	391	9	400	0.0225
5×2	316	84	400	0.2100
Totals	**1786**	**1814**	**3600**	**0.5039**

Table 1: Final accuracy distribution across the 3,600 samples generated for RQA analysis, categorised by puzzle configuration.

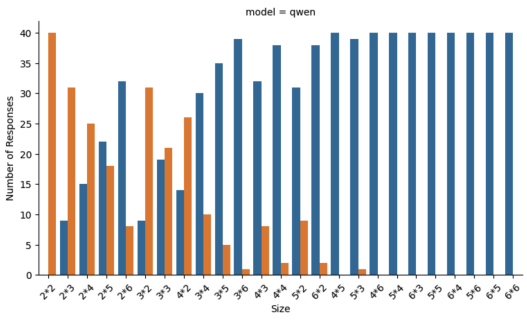

Figure 5: Preliminary evaluation showing the inverse correlation between combinatorial complexity and puzzle-level accuracy across DeepSeek distilled models.

We generated a corpus of reasoning trajectories using **DeepSeek-R1-Distill-Qwen-7B**, selected for its ability to produce explicit Chain-of-Thought (CoT) reasoning tokens. Following the developer's recommendations for eliciting robust reasoning behavior, we employed stochastic decoding with a temperature of 0.6.

Preliminary Analysis: Prior to conducting Recurrence Quantification Analysis (RQA), we observed an **negative correlation with complexity**: in line with the "Curse of Complexity" (Lin et al.,

2025), model accuracy decreases sharply as the combinatorial search space grows (Figure 5).

Data Selection and Class Balancing: To obtain a balanced distribution of correct and incorrect reasoning traces, we selected nine grid configurations: $2\times3,\ 2\times4,\ 2\times5,\ 3\times2,\ 3\times3,\ 3\times4,\ 4\times3,\ 4\times4,\ 5\times2$.

These configurations were chosen for the following reasons:

- **Class balancing for correctness:** In larger configurations (e.g., 6×6), the model's accuracy approaches zero. Including such extreme cases would create a "floor effect," making binary classification of correctness statistically unreliable. The nine selected sizes provide a near-perfect global balance of 1,814 correct versus 1,786 incorrect traces (Table 1).
- **Representativeness:** The selected configurations span the complexity spectrum—from 2×3 (low search space, high accuracy) to 4×4 and 5×2 (high search space, low accuracy). This allows us to capture the effects of the "Curse of Complexity" without introducing data sparsity issues associated with the most extreme grid sizes.

For each configuration, we sampled 40 distinct puzzles and generated 10 independent reasoning traces per puzzle, yielding a total of 3,600 traces.

4.3 Feature Extraction: Recurrence Quantification Analysis

To characterise the internal dynamics of the model's reasoning process, we extracted hidden-state trajectories from the final transformer layer for each reasoning trace. From these trajectories, we derived three class of features for comparisation:

1. Response Length: The total number of tokens in the generated CoT response.

2. Global RQA: Full-trace averages of Determinism (DET), Laminarity (LAM), and Recurrence Entropy (ENTR), capturing the overall recurrence structure of the hidden-state dynamics.

3. Temporal RQA: To model non-stationary reasoning dynamics, we applied a sliding window of 150 tokens with a 10% step size across each trace. For each RQA metric, we computed the mean, standard deviation, linear trend (slope), and the Detrended Fluctuation Analysis (DFA) exponent of the resulting time series.

RQA Hyperparameters: To ensure consistency across varying trace lengths and to avoid spurious recurrences, we fixed the following parameters throughout all experiments:

- **Recurrence threshold (ϵ):** Set to the top 10th percentile of pairwise cosine distances between hidden states.
- **Minimum line lengths ($l_{\min}, v_{\min}$):** Fixed at 3 for diagonal and vertical lines.
- **Sequence cap:** Reasoning traces were truncated at 32,000 tokens.

4.4 Classification Experiment Design

We evaluate whether recurrence-based features provide predictive signals for two classification tasks:

- **Problem complexity:** A 9-way classification task predicting the grid configuration ($N\times M$).
- **Answer correctness:** A binary classification task distinguishing correct from incorrect solutions.

We employ two classifier families: **Logistic Regression (LR)** to assess linear separability, and **Random Forests (RF)** to capture non-linear interactions among RQA features. Specifically, we used 100 n-estimators for the Random Forests Classifier.

Evaluation Protocol: To prevent leakage of puzzle-specific information, we use stratified 8-fold group cross-validation, where the **Puzzle ID** serves as the grouping variable. This ensures that reasoning traces derived from the same puzzle never appear in both training and test splits, enforcing strict generalisation.

5 Results and Discussion

Our evaluation reveals a clear functional separation between the predictive power of response length and RQA-based methods.

Complexity: Structural Dynamics Beyond Token Length. As shown in Table 2, the **Temporal RQA (RF)** model substantially outperforms the length-based baseline by 8% for complexity classification (36.94% vs. 28.97%). McNemar's test confirms that this 8% absolute improvement is statistically significant ($p < 0.05$) in 7 out of 8 validation folds (average $p < 0.005$). While response length reflects the *duration* of test-time scaling, it remains blind to the *structural quality* of the underlying search process. In contrast, the superior performance of Temporal RQA demonstrates that combinatorial difficulty is encoded in

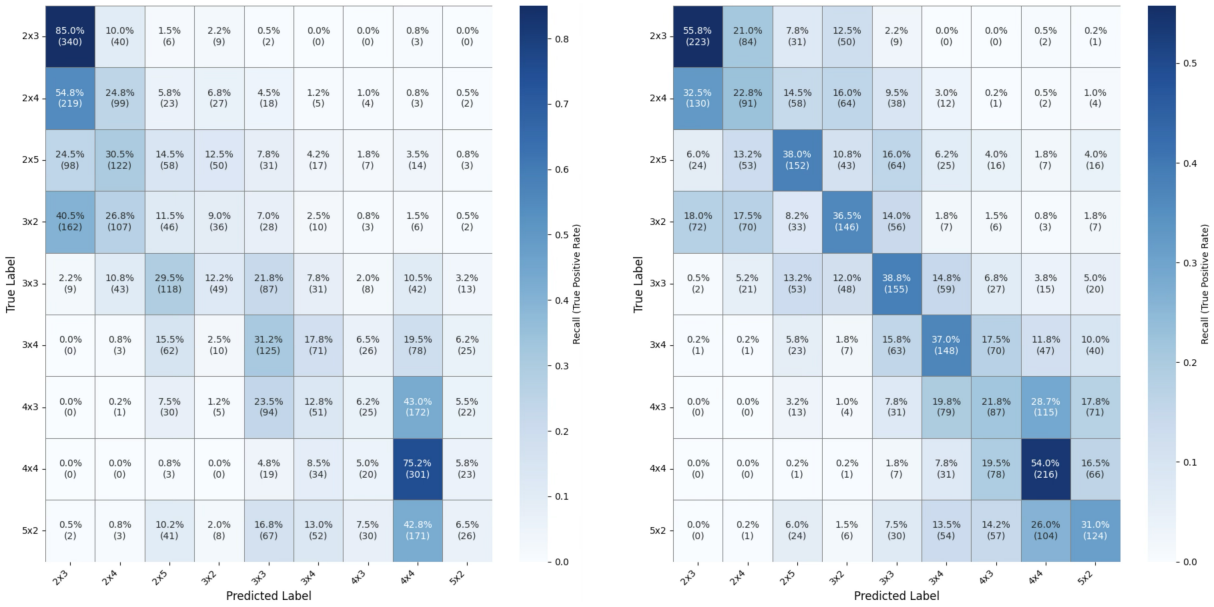

Figure 6: Confusion matrices for complexity classification. (**a**) Response-length baseline, which performs well on extreme complexity levels but struggles in intermediate regimes. (**b**) Temporal RQA, which exhibits more uniform performance across complexity levels.

Features	**Model**	**Complexity (%)**	**Accuracy (%)**
Baseline (Length)	RF / LR	28.97 ± 3.1	85.58 ± 1.5
Global RQA	RF	16.81 ± 2.2	55.03 ± 3.4
Temporal RQA	LR	29.50 ± 3.8	81.19 ± 2.0
Temp. RQA	**RF**	$\mathbf{36.94 \pm 3.3}$	85.00 ± 1.2

Table 2: Classification results using 8-fold group cross-validation. Temporal RQA captures task complexity where length fails, while length remains a dominant proxy for binary failure.

the geometric *shape* of the reasoning trajectory. The effectiveness of the RF classifier further suggests that task complexity emerges as a non-linear interaction between structural stability (DET) and semantic stalling (LAM), rather than as a monotonic function of token count.

Accuracy: Overlapping Signals for Binary Failure. For binary accuracy prediction, response length remains a remarkably strong predictor (85.58%), consistent with prior observations that failure often manifests as uncontrolled trace expansion—a regime termed the **Curse of Complexity** (Lin et al., 2025). Temporal RQA achieves comparable performance (85.00%), with no statistically significant difference in the majority of folds ($p > 0.05$). This indicates that while length and RQA capture distinct aspects of the dynamics, they encode partially overlapping signals for coarse-grained failure detection. In particular, extreme breakdowns in reasoning dynamics are often accompanied by both prolonged generation and increased recurrence.

Resolution and Non-Stationarity. The poor performance of Global RQA (16.81%) highlights the fundamentally **non-stationary** nature of LLM reasoning. Averaging dynamics over the entire trace obscures critical transitions between problem parsing, active logical inference, and final decoding. Predictive signal resides almost entirely in the *temporal evolution* of the trajectory. Consistent with this interpretation, Figure 6 shows that the length baseline performs well only at the extremes of task difficulty, whereas Temporal RQA maintains discriminative power across the full complexity spectrum—resolving intermediate combinatorial regimes that length alone fails to separate.

6 Ablation Study

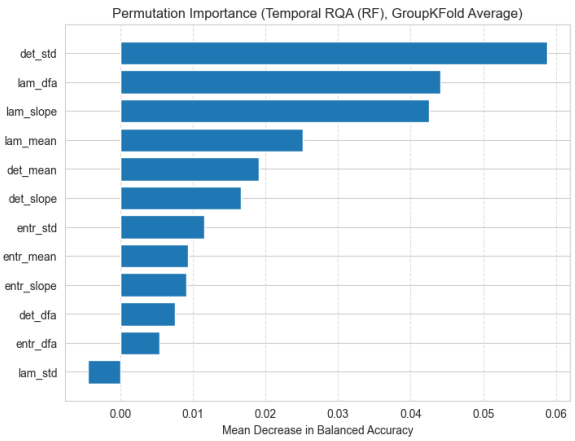

Figure 7: Here, det stands for determinism, entr is entropy, and lam is laminarity. The plot shows different contributions of each factor to the predictive power of Temporal RQA model.

To identify the drivers of complexity prediction, we used the Permutation Importance (Mean Decrease in Balanced Accuracy) averaged across 8-fold GroupKFold validation to help us analyse the model. As shown in Figure 7, the results reveal a distinct hierarchy of dynamical signals:

- **Fluctuation of Consistency:** Most important feature is **det_std** (variability of semantic repetition). Complexity is captured not by the stability of the amount of repetition, but by its *fluctuation*. Complex puzzles induce a dynamics alternating between structured semantic repetition and its lack thereof.
- **Long-range Stalling Dynamics:** High-ranking features **lam_dfa** (fractal memory of stalling) and **lam_slope** (acceleration of stalling) validate Temporal RQA. The DFA exponent shows reasoning difficulty involves long-range correlations, where the model's current "stalling" behaviour depends on representations of hundreds of steps prior.
- **Mean States:** **lam_mean** is less important than fluctuation and memory features (**det_std**, **lam_dfa**). While average stalling correlates with puzzle length, RQA's discriminative power comes from higher-order temporal rhythms rather than simple averages.

7 Conclusion

We introduced Recurrence Quantification Analysis (RQA) as a complementary framework for analysing chain of thought embedding layers' object. By analysing hidden-state trajectories as dynamical systems, RQA provides a non-linguistic, structure-sensitive view of reasoning that operates alongside textual explanations and scalar proxies such as response length. Our results show that task complexity is encoded in the temporal evolution of latent dynamics rather than in token volume alone, while binary failure remains well captured by length variable. Our ablation study reveals an interesting dynamics between Laminarity and determinism variables in the complexity metrics that warrants further investigations into fully elaborating their relationships. As a non-textual tool, RQA potentially enables automated monitoring during inference. In particular, reasoning models could implement *early-exit* or *backtracking* triggers, ensuring better performance and efficiency. Together, these findings position RQA as a principled tool for probing the geometry of reasoning processes and highlight the value of dynamical analyses for understanding, monitoring, and controlling large-scale inference.

Limitations

While Recurrence Quantification Analysis (RQA) offers a high-resolution, structure-sensitive view of latent reasoning dynamics, it incurs higher computational cost than surface-level proxies such as response length. Computing recurrence matrices and temporal RQA features scales quadratically with sequence length and requires access to hidden-state activations, which may limit applicability in latency- or resource-constrained settings.

Our methodology currently relies on heuristic choices for key RQA hyperparameters, including the recurrence threshold (set to 10%) and sliding window size. Although these parameters are held constant across experiments, dynamical systems analyses can be sensitive to such choices. Future work should explore different but robust, automated parameter selection strategies to improve stability and consistency across tasks and models.

The scope of our empirical evaluation is also limited. We focus on a single family of reasoning models (DeepSeek-R1-Distill) and a controlled symbolic reasoning benchmark (ZebraLogic), which provides a clear measure of combinatorial complexity but represents a narrow task distribution. Establishing the generality of recurrence-based signatures will require extending this analysis to a

broader range of architectures (e.g., LLaMA 3, Claude) and domains, including mathematical reasoning, code generation, and open-ended language tasks.

Finally, our analysis is restricted to hidden-state trajectories from the final transformer layer. While this choice offers a consistent and interpretable starting point, recurrence structure may vary across layers. Extending RQA to multi-layer or cross-layer dynamics, as well as systematically analysing sensitivity to architectural depth, remains an important direction for future research.

Acknowledgments

References

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. Chain-of-thought reasoning in the wild is not always faithful, 2025. *URL https://arxiv.org/abs/2503.08679*.

Aman Bhargava, Cameron Witkowski, M Shah, and MW Thomson. 2023. What's the magic word? a control theory of llm prompting (2023). *URL https://arxiv.org/abs/2310.04444*.

Paul C Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. 2025. Thought anchors: Which llm reasoning steps matter? *arXiv preprint arXiv:2506.19143*.

Jack David Carson and Amir Reisizadeh. 2025. A statistical physics of language model reasoning. *arXiv preprint arXiv:2506.04374*.

Xi Chen, Aske Plaat, and Niki van Stein. 2025. How does chain of thought think? mechanistic interpretability of chain-of-thought reasoning with sparse autoencoding. *arXiv preprint arXiv:2507.22928*.

Subhabrata Dutta, Joykirat Singh, Soumen Chakrabarti, and Tanmoy Chakraborty. 2024. How to think step-by-step: A mechanistic understanding of chain-of-thought reasoning. *arXiv preprint arXiv:2402.18312*.

J-P Eckmann, S Oliffson Kamphorst, and David Ruelle. 1995. Recurrence plots of dynamical systems. In *Turbulence, strange attractors and chaos*, pages 441–445. World Scientific.

Subbarao Kambhampati, Kaya Stechly, Karthik Valmeekam, Lucas Saldyt, Siddhant Bhambri, Vardhan Palod, Atharva Gundawar, Soumya Rani Samineni, Durgesh Kalwar, and Upasana Biswas. 2025. Stop anthropomorphizing intermediate tokens as reasoning/thinking traces! *arXiv preprint arXiv:2504.09762*.

Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. 2025. Zebralogic: On the scaling limits of llms for logical reasoning. *arXiv preprint arXiv:2502.01100*.

Marinho A Lopes, Jiaxiang Zhang, Dominik Krzemiński, Khalid Hamandi, Qi Chen, Lorenzo Livi, and Naoki Masuda. 2021. Recurrence quantification analysis of dynamic brain networks. *European Journal of Neuroscience*, 53(4):1040–1059.

Sara Vera Marjanovic, Arkil Patel, Vaibhav Adlakha, Milad Aghajohari, Parishad BehnamGhader, Mehar Bhatia, Aditi Khandelwal, Austin Kraft, Benno Krojer, Xing Han Lu, and 1 others. 2025. Deepseek-r1 thoughtology: Let's think about llm reasoning. *URL https://arxiv. org/abs/2504.07128*.

Norbert Marwan and K Hauke Kraemer. 2023. Trends in recurrence analysis of dynamical systems. *The European Physical Journal Special Topics*, 232(1):5–27.

C-K Peng, Sergey V Buldyrev, Shlomo Havlin, Michael Simons, H Eugene Stanley, and Ary L Goldberger. 1994. Mosaic organization of dna nucleotides. *Physical review e*, 49(2):1685.

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

Yiru Tang, Kun Zhou, Yingqian Min, Wayne Xin Zhao, Jing Sha, Zhichao Sheng, and Shijin Wang. 2025. Enhancing chain-of-thought reasoning via neuron activation differential analysis. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16162–16170.

Jean-Francois Ton, Muhammad Faaiz Taufiq, and Yang Liu. Understanding chain-of-thought in llms through information theory, 2024. *URL https://arxiv.org/abs/2411.11984*.

Baki Ünal. 2022. Stability analysis of bitcoin using recurrence quantification analysis. *Chaos Theory and Applications*, 4(2):104–110.

Constantin Venhoff, Iván Arcuschin, Philip Torr, Arthur Conmy, and Neel Nanda. 2025. Understanding reasoning in thinking language models via steering vectors. *arXiv preprint arXiv:2506.18167*.

Charles L Webber and Norbert Marwan. 2015. Recurrence quantification analysis. *Theory and best practices*, 426.

Xinyi Yang, Liang Zeng, Heng Dong, Chao Yu, Xiaoran Wu, Huazhong Yang, Yu Wang, Milind Tambe, and Tonghan Wang. 2025a. Policy-to-language: Train llms to explain decisions with flow-matching generated rewards. *arXiv preprint arXiv:2502.12530*.

Zhou Yang, Zhengyu Qi, Zhaochun Ren, Zhikai Jia, Haizhou Sun, Xiaofei Zhu, and Xiangwen Liao. 2025b. Exploring information processing in large language models: Insights from information bottleneck theory. *arXiv preprint arXiv:2501.00999*.

Jiacheng Ye, Shansan Gong, Liheng Chen, Lin Zheng, Jiahui Gao, Han Shi, Chuan Wu, Xin Jiang, Zhenguo Li, Wei Bi, and 1 others. Diffusion of thoughts: Chain-of-thought reasoning in diffusion language models, december 2024. *arXiv preprint arXiv:2402.07754*.

Xiang Zhang, Muhammad Abdul-Mageed, and Laks VS Lakshmanan. 2024. Autoregressive+ chain of thought= recurrent: Recurrence's role in language models' computability and a revisit of recurrent transformer. *arXiv preprint arXiv:2409.09239*.