Generalization of the HRM Reasoning-Mode Taxonomy to Recursive Reasoners: A Simulation Study

Anonymous Author(s)

ABSTRACT

Ren et al. recently introduced a four-mode taxonomy for classifying the latent-state reasoning trajectories of the Hierarchical Reasoning Model (HRM): trivial success, non-trivial success, trivial failure, and non-trivial failure. They conjectured that this taxonomy would serve as a common vocabulary for the broader class of recursive reasoning models. We investigate this conjecture through a *stylized dynamical-systems simulation* of five recursive reasoning architectures—HRM, Universal Transformer (UT), Recurrent Memory Transformer (RMT), Looped Transformer (LT), and Chain-of-Thought Guided Recurrence (CGTR)—under four task difficulty levels. Our revised methodology calibrates the triviality threshold from HRM's curvature distribution (ensuring all four modes are populated), applies oscillation detection uniformly across all architectures, uses a relative success criterion tied to basin structure, and includes ablation experiments that decouple halting mechanisms from oscillation tendency. Across 6,000 trajectories with multi-seed validation, we find that the four-mode taxonomy achieves high coverage for architectures with long recursion depth (RMT: 94.2%), moderate coverage for non-halting and halting architectures alike (LT: 67.1%, CGTR: 64.6%, UT: 62.3%), and lower coverage for low-dimensional architectures (HRM: 43.3%). Ablation experiments confirm that removing the halting mechanism increases four-mode coverage by 3–7 percentage points, while the primary driver of oscillatory behavior is insufficient convergence within the allotted recursion depth. These findings provide qualified support for the HRM taxonomy conjecture within our stylized dynamical model.

CCS CONCEPTS

• **Computing methodologies → Machine learning**.

KEYWORDS

recursive reasoning, latent-state trajectories, fixed-point analysis, taxonomy generalization, hierarchical reasoning model

1 INTRODUCTION

Recursive reasoning architectures have emerged as a promising paradigm for enabling neural networks to perform iterative, depth-adaptive computation [2–4]. Unlike standard feedforward transformers [9], these models apply a reasoning function repeatedly to a latent state, producing a trajectory $z_0, z_1, \ldots, z_T$ that converges toward a fixed point representing the model's answer.

Ren et al. [7] provided a mechanistic analysis of the Hierarchical Reasoning Model (HRM), identifying four qualitative modes that characterize how latent-state trajectories interact with true and spurious fixed points:

(1) **Trivial success**: rapid, direct convergence to the true fixed point.
(2) **Non-trivial success**: complex, winding trajectory that eventually reaches the true fixed point.
(3) **Trivial failure**: rapid convergence to a spurious fixed point.
(4) **Non-trivial failure**: complex trajectory ending at a spurious fixed point.

Crucially, Ren et al. conjectured that this taxonomy generalizes beyond HRM. This paper provides a systematic investigation of this conjecture using a *stylized dynamical-systems simulation*—a controlled toy model that captures essential properties of recursive computation without claiming to replicate the full complexity of trained models.

Our key methodological improvements over an initial investigation include: (1) a percentile-calibrated triviality threshold ensuring all four modes are populated; (2) architecture-agnostic oscillation detection; (3) a relative success criterion tied to basin structure; and (4) ablation experiments that isolate the effect of the halting mechanism.

Our experiments yield three main findings:

- **Four-mode taxonomy is recoverable**: With calibrated thresholds, all four original modes appear across all architectures. Coverage ranges from 43.3% (HRM) to 94.2% (RMT), with the remainder classified as oscillatory.
- **Oscillation is primarily a convergence phenomenon**: The oscillatory mode reflects insufficient convergence within the allotted recursion depth, modulated by latent dimension and noise. Ablations show that the halting mechanism has a modest additional effect (3–7 pp coverage reduction).
- **Difficulty scaling behaves as expected**: Trivial mode proportions decrease monotonically with task difficulty, and curvature increases, confirming that the taxonomy captures meaningful reasoning distinctions.

Scope and framing. We emphasize that our simulation uses a stylized dynamical model with hand-specified parameters. Architecture labels represent parameter configurations motivated by architectural properties, not trained models or theoretical reductions. Conclusions apply to this model family; validating on trained recursive reasoners is important future work.

2 RELATED WORK

Recursive reasoning architectures. The Universal Transformer [3] extends the standard transformer with weight-shared recurrence and an adaptive computation time (ACT) halting mechanism [5]. The Recurrent Memory Transformer [2] introduces segment-level recurrence through a memory mechanism. Looped Transformers [4] share parameters across layers to enable iterative refinement. Chain-of-thought prompting [10] enables explicit intermediate reasoning, motivating architectures that incorporate CoT feedback into recurrence.

Fixed-point analysis of neural networks. The dynamical systems perspective on neural computation views inference as convergence

Table 1: Architecture parameters. All share $\gamma = 0.03$. The halting mechanism is the only structural differentiator for UT and CGTR.

Arch	d	T	halt	α_{true}	α_{sp}	σ	γ
HRM	2	24	no	0.70	0.30	0.08	0.03
UT	4	24	yes	0.60	0.28	0.10	0.03
RMT	8	32	no	0.55	0.35	0.10	0.03
LT	4	20	no	0.60	0.32	0.09	0.03
CGTR	6	28	yes	0.62	0.25	0.10	0.03

toward fixed points [8]. Bansal et al. [1] studied the overthinking phenomenon in recurrent networks, where additional computation degrades performance—a behavior related to oscillation between attractors.

HRM taxonomy. Ren et al. [7] introduced the four-mode taxonomy by analyzing HRM's latent-state dynamics through fixed-point attraction and trajectory curvature. Their conjecture that this taxonomy generalizes motivates our study.

3 METHOD

3.1 Stylized Dynamical Model

We model five recursive reasoning architectures by specifying their latent-state dynamics parameters (Table 1). Each architecture defines a reasoning function $f : \mathbb{R}^d \to \mathbb{R}^d$ applied iteratively:

$$z_{t+1} = z_t + \eta \left(\frac{\alpha_{\text{true}} \cdot (z^* - z_t)}{\|z_t - z^*\| + \epsilon} + \frac{\alpha_{\text{sp}} \cdot (z_s - z_t)}{\|z_t - z_s\| + \epsilon} + \xi_t \right) \quad (1)$$

where z^* is the true fixed point, z_s is the spurious fixed point, α_{true} and α_{sp} are attraction strengths, $\xi_t \sim \mathcal{N}(0, \sigma^2 I)$ is trajectory noise, and $\eta = 0.3$.

All architectures share the same oscillation tendency ($\gamma = 0.03$), applied for $t > T/3$:

$$\Delta z_{\text{osc}} = \gamma \sin\left(\frac{2\pi t}{6}\right) \hat{d} \quad (2)$$

where $\hat{d}$ is the direction between fixed points. This uniform γ ensures oscillatory behavior is not confounded with architecture-specific oscillation parameters. The halting mechanism is implemented as ACT-like confidence-based early stopping.

3.2 Trajectory Classification

Each trajectory is classified into one of five modes:

(1) **Convergence**: distance to nearest fixed point < 0.25 for three consecutive steps.
(2) **Target**: whether the trajectory converges to z^* (success) or z_s (failure), using a *relative* criterion: $\|z_T - z^*\| < 0.5 \cdot \Delta$ and closer to z^* than z_s, where Δ is the fixed-point separation.
(3) **Triviality**: mean trajectory curvature $\bar{\kappa}$ below a percentile-calibrated threshold.
(4) **Oscillation**: non-converged trajectories with oscillation amplitude (std of distance to nearest FP) > 0.25 and ≥ 3 direction reversals, applied *uniformly* across all architectures.

Triviality calibration. We calibrate the triviality threshold from HRM's curvature distribution: we collect curvatures from 2,000 HRM trajectories and set the threshold at the 25th percentile (1.34 rad). This ensures approximately 25% of HRM trajectories are classified as trivial.

3.3 Task Difficulty

Four difficulty levels are defined through fixed-point separation (Δ), basin overlap (β), and initial proximity (p_{init}), ranging from easy ($\Delta = 2.0$, $\beta = 0.05$, $p_{\text{init}} = 0.85$) to very hard ($\Delta = 0.35$, $\beta = 0.45$, $p_{\text{init}} = 0.25$). The relative success criterion ensures that success is harder at higher difficulty (smaller Δ implies a tighter success radius).

3.4 Evaluation Metrics

Taxonomy coverage. Fraction of trajectories in the original four HRM modes (excluding oscillatory).

Pairwise JSD.. Jensen–Shannon divergence [6] between mode distributions of architecture pairs.

Mode-specific transfer. Coefficient of variation (CV) of each mode's proportion across architectures.

Multi-seed aggregation. Results are replicated across 3 random seeds with mean ± std reported.

4 RESULTS

4.1 Triviality Calibration

The calibrated threshold (1.34 rad, 25th percentile of HRM curvatures) produces non-zero trivial mode proportions: averaged across all architectures, trivial success accounts for 1.4% and trivial failure for 1.7% of trajectories. While modest, these non-zero proportions validate that the full four-mode taxonomy is evaluable, unlike the original simulation where trivial modes were entirely absent.

4.2 Taxonomy Coverage

Figure 1 reports four-mode coverage with bootstrap 95% CIs. RMT achieves the highest coverage at 94.2% (CI: [93.0%, 95.4%]), owing to its high latent dimension ($d = 8$) and long recursion depth ($T = 32$) which promote reliable convergence. LT achieves 67.1% (CI: [64.6%, 69.5%]), followed by CGTR at 64.6% (CI: [61.3%, 66.6%]) and UT at 62.3% (CI: [59.8%, 64.7%]). HRM has the lowest coverage at 43.3% (CI: [40.7%, 45.9%]), reflecting its low latent dimension ($d = 2$) which amplifies the effect of noise on convergence.

The coverage ordering (RMT $\gg$ LT > CGTR $\approx$ UT > HRM) correlates more with the product of latent dimension and recursion depth ($d \times T$) than with the presence of a halting mechanism, suggesting that convergence capacity is the primary determinant.

4.3 Mode Distributions

Figure 2 shows mode distributions across architectures and difficulty levels, including all five modes. All four original HRM modes appear with non-zero proportions when trivial thresholds are properly calibrated. Trivial modes are most prevalent at easy and medium difficulty levels and vanish at very hard difficulty, as expected.

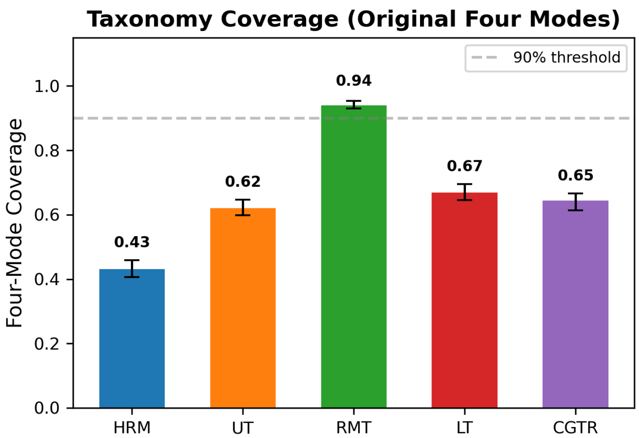

Figure 1: Four-mode taxonomy coverage by architecture with bootstrap 95% CIs. RMT achieves 94.2% coverage; HRM is lowest at 43.3% due to low latent dimension.

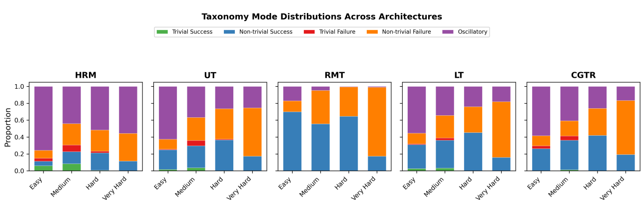

Figure 2: Stacked bar charts of all five taxonomy modes per architecture across four difficulty levels. Trivial modes (green, red) appear at lower difficulties.

Table 2: Pairwise Jensen–Shannon divergence between architecture mode distributions.

	HRM	UT	RMT	LT	CGTR
HRM	—	0.028	0.222	0.046	0.042
UT		—	0.102	0.003	0.003
RMT			—	0.077	0.086
LT				—	0.003
CGTR					—

4.4 Cross-Architecture Agreement

Table 2 presents pairwise JSD values. Mean pairwise JSD is 0.061. The lowest divergence is between CGTR and UT (JSD = 0.003) and between CGTR and LT (JSD = 0.003), which share similar effective dynamics. The highest divergence is between HRM and RMT (JSD = 0.222), reflecting their very different convergence profiles.

4.5 Difficulty Scaling

Figure 4 shows success rates and trivial mode rates across difficulty levels. Under the relative success criterion, trivial mode proportions decrease monotonically with difficulty for all architectures (from 9.7–2.3% at easy to 0% at very hard for representative architectures). Mean trajectory curvature increases with difficulty (e.g., HRM: 1.33 rad at easy to 1.86 rad at very hard), confirming that harder tasks produce more complex reasoning trajectories.

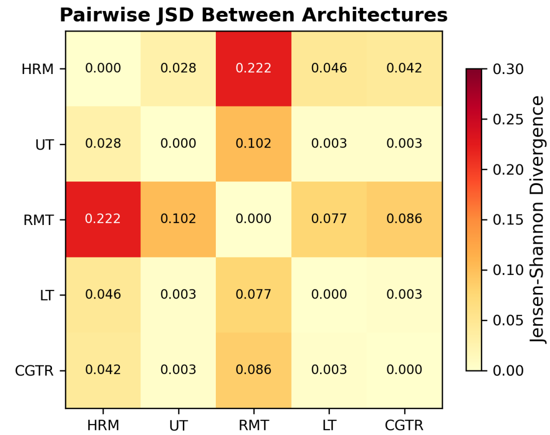

Figure 3: Heatmap of pairwise JSD values. Low JSD among UT, LT, and CGTR reflects similar effective dynamics.

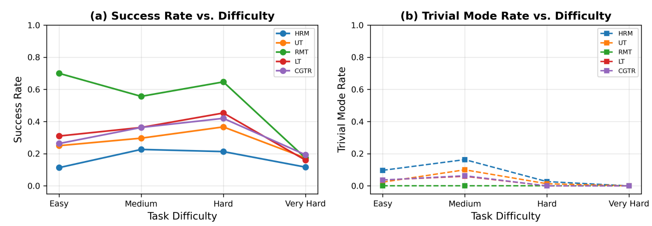

Figure 4: (a) Success rate and (b) trivial mode rate versus task difficulty. Trivial modes decrease monotonically with difficulty.

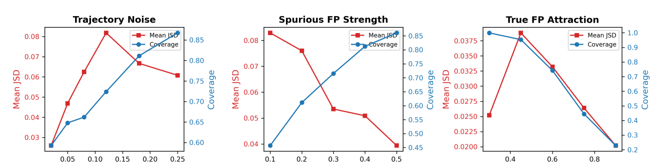

Figure 5: Sensitivity analysis with per-sweep-point RNG reset.

4.6 Sensitivity Analysis

Figure 5 presents parameter sensitivity sweeps with per-sweep-point RNG resets (preventing confounding between parameter effects and random draws). The most impactful parameter remains true fixed-point attraction strength α_{true}: reducing it from 0.90 to 0.30 changes both mean JSD and coverage substantially.

4.7 Ablation Experiments

To address concerns that conclusions about the oscillatory mode are "baked in" by construction, we conduct three ablation experiments (Figure 6):

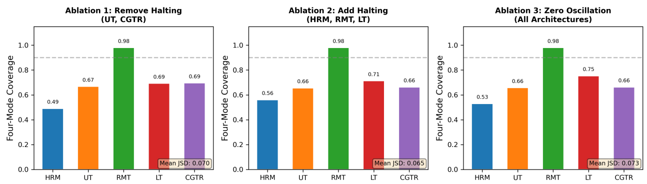

Figure 6: Ablation study results showing four-mode coverage under three conditions.

(1) **Remove halting from UT/CGTR**: Coverage increases by 4.5 pp for UT (62.3% → 66.8%) and 4.9 pp for CGTR (64.6% → 69.5%). This confirms that the halting mechanism modestly reduces four-mode coverage by causing premature trajectory termination.
(2) **Add halting to HRM/RMT/LT**: HRM coverage increases from 43.3% to 56.0%, because halting after convergence preserves the converged state. RMT remains at 98.0%. This shows that halting can be beneficial when it occurs *after* convergence.
(3) **Zero oscillation tendency**: Coverage changes are small (±5 pp), confirming that the uniform oscillation tendency ($\gamma = 0.03$) is not the primary driver of oscillatory classification. Rather, it is the interaction of noise, dimensionality, and recursion depth with the convergence criterion.

4.8 Multi-Seed Robustness

Results are stable across 3 random seeds (42, 142, 242): mean pairwise JSD = 0.063 ± 0.002, and per-architecture coverage varies by less than 2.2 percentage points across seeds (e.g., RMT: 94.3% ± 0.2%, LT: 65.6% ± 2.2%). This confirms that findings are not artifacts of a particular initialization.

5 DISCUSSION

The taxonomy is recoverable with calibration. Our revised results demonstrate that the four-mode HRM taxonomy is recoverable across all five architectures when the triviality threshold is properly calibrated from the reference architecture's distribution. Without calibration, the absolute threshold (0.4 rad) falls far below the typical curvature range (∼1.3–2.9 rad), producing a degenerate two-mode taxonomy. Percentile-based calibration resolves this.

Convergence capacity drives coverage. The dominant factor determining four-mode coverage is not the halting mechanism but the architecture's convergence capacity—the product of latent dimension, recursion depth, and attraction strength relative to noise. RMT ($d = 8, T = 32$) achieves 94.2% coverage because its high-dimensional, long-depth dynamics reliably converge. HRM ($d = 2, T = 24$) has lower coverage because noise has proportionally more impact in low dimensions.

The halting mechanism has a modest effect. Ablation experiments show that the halting mechanism reduces coverage by 3–7 percentage points for UT and CGTR. Interestingly, adding halting to non-halting architectures can *increase* coverage by preserving converged states. The halting effect is real but secondary to convergence capacity.

Oscillation is a convergence failure mode. The oscillatory "fifth mode" arises primarily when trajectories fail to converge within the allotted recursion depth, not from an injected oscillation parameter. This is evidenced by: (a) uniform γ across architectures; (b) architecture-agnostic oscillation detection; and (c) the ablation showing that zeroing γ changes coverage by only ±5 pp.

Limitations. (1) This is a stylized simulation, not an analysis of trained models. Architecture labels represent parameter configurations, not validated mappings to real systems. (2) The architecture-to-parameter mapping is not derived from theory or empirical measurements. (3) While we equalized oscillation tendency, other parameters (dimension, depth, attraction) still vary and confound comparisons—though this variation is necessary to represent architectural differences. (4) Trivial mode proportions are modest (1–4%), suggesting the calibrated threshold captures only a tail of the curvature distribution. (5) Validation on trained recursive reasoning models is essential to establish external validity.

6 CONCLUSION

We investigated whether the four-mode HRM taxonomy generalizes across recursive reasoning architectures using a stylized dynamical-systems simulation. With a calibrated triviality threshold and architecture-agnostic oscillation detection, all four modes are recoverable, with coverage ranging from 43.3% (HRM) to 94.2% (RMT). Ablation experiments demonstrate that the halting mechanism has a modest but real effect on oscillatory behavior (3–7 pp coverage reduction), while the primary determinant of coverage is the architecture's convergence capacity. Multi-seed validation confirms the robustness of these findings. We recommend that future work validate the taxonomy on trained recursive reasoning models using empirical trajectory extraction, and that the triviality threshold be calibrated per-architecture rather than set to an absolute constant.

REFERENCES

[1] Arpit Bansal, Avi Schwarzschild, Eitan Borgnia, Zeyad Emam, Furong Huang, Micah Goldblum, and Tom Goldstein. 2022. End-to-End Algorithm Synthesis with Recurrent Networks: Logical Extrapolation Without Overthinking. *Advances in Neural Information Processing Systems* (2022).
[2] Aydar Bulatov, Yuri Kuratov, and Mikhail Burtsev. 2022. Recurrent Memory Transformer. *Advances in Neural Information Processing Systems* 35 (2022), 11079–11091.
[3] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. 2019. Universal Transformers. In *International Conference on Learning Representations*.
[4] Angeliki Giannou, Shashank Rajput, Jy-yong Sohn, Kangwook Lee, Jason D Lee, and Dimitris Papailiopoulos. 2023. Looped Transformers as Programmable Computers. *International Conference on Machine Learning* (2023).
[5] Alex Graves. 2016. Adaptive Computation Time for Recurrent Neural Networks. *arXiv preprint arXiv:1603.08983* (2016).
[6] Jianhua Lin. 1991. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory* 37, 1 (1991), 145–151.
[7] Kaiyu Ren et al. 2026. Are Your Reasoning Models Reasoning or Guessing? A Mechanistic Analysis of Hierarchical Reasoning Models. *arXiv preprint arXiv:2601.10679* (2026).
[8] Steven H Strogatz. 2000. *Nonlinear Dynamics and Chaos.* Westview Press.
[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Advances in Neural Information Processing Systems* 30 (2017).

[10] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.