Prompt Hysteresis: Task Inertia and Metastable States in In-Context Learning

Anonymous ACL submission

Abstract

Large language models can infer new task rules from in-context demonstrations, but it remains unclear how readily an inferred rule is revised when later evidence contradicts it. We study this question through *prompt hysteresis*: path dependence in the response to matched but oppositely ordered conflicting demonstrations. In semantic-free symbolic and cipher tasks, we construct paired incompatible mappings and compare A-to-B prompts with B-to-A prompts at the same rule proportions. A gap between the two response curves indicates retention of a prompt-induced task state beyond current evidence aggregation. We quantify the loop with task magnetization and hysteresis area, and develop a Landau-style effective theory in which evidence acts as a control field while task-state self-consistency creates metastable basins. Across Qwen3 dense models, task inertia increases with scale. Mean hysteresis area rises from 0.067 for Qwen3-0.6B to 0.269 for Qwen3-8B when prompts contain at least six demonstrations, and from 0.077 to 0.309 in the $n = 12$ slice. A matched-suffix control gives the same ordering under identical later evidence, with the preparation gap increasing from 0.235 to 1.197. State-dynamics probes show gradual decay across long suffixes, and RESET boundaries reduce but do not erase preparation memory. Residual patching localizes transferable task-state information to late final-token activations, while learned task-axis interventions move stuck predictions 4 to 9$\times$ more than matched random directions. The behavioral and mechanistic evidence identifies prompt hysteresis as a measurable metastability phenomenon in in-context learning.

1 Introduction

In-context learning (ICL) is one of the defining capabilities of large language models. A small set of demonstrations can induce a new input-output rule without changing the model parameters (Brown et al., 2020). Prior work mainly asks how this rule is inferred. Bayesian accounts cast the prompt as evidence for a latent task (Xie et al., 2022), algorithmic accounts show that transformers can implement learning procedures in context (Garg et al., 2022; Akyürek et al., 2023; von Oswald et al., 2023; Dai et al., 2023; Raventós et al., 2023), and mechanistic accounts identify induction heads, task vectors, or function vectors that carry task information (Olsson et al., 2022; Hendel et al., 2023; Todd et al., 2024). These perspectives clarify task acquisition from a prompt. They leave a more dynamic question open: after a model has inferred a task, how readily does that task state revise itself when later evidence conflicts with it?

Ordinary prompt sensitivity does not isolate this question. A prediction change after reordering may reflect recency, calibration, formatting, or a different evidence set. The sharper test holds evidence counts fixed while varying the trajectory through that evidence. Consider two incompatible arbitrary mappings over the same inputs. One prompt begins with Rule A and moves toward Rule B. Another begins with Rule B and moves toward Rule A. At a matched point, the two prompts contain the same number of A and B demonstrations. A memoryless evidence aggregator should give the same response. A persistent task-state model need not.

We call this path dependence *prompt hysteresis*. The term follows the physical idea that a system can retain memory of the path by which a control variable was reached. Here, the control variable is the fraction of demonstrations supporting Rule B, and the response variable is the model's preference between the two rule-consistent answers. Task magnetization summarizes the response, and hysteresis is the area between the A-to-B and B-to-A curves. The loop distinguishes reversible evidence aggregation from metastable prompt-induced task states.

We study this phenomenon in semantic-free sym-

bolic and cipher tasks, where two rules share inputs but assign incompatible outputs. This setting reduces confounds from label meaning, world knowledge, and natural-language task priors. Across Qwen3 dense models, prompt hysteresis grows with scale once prompts contain enough demonstrations. Mean hysteresis area rises from 0.067 for Qwen3-0.6B to 0.269 for Qwen3-8B when $n \geq 6$, and from 0.077 to 0.309 in the $n = 12$ slice. A matched-suffix control gives a stricter persistence test. After identical later evidence, A-prepared and B-prepared prompts still lead to different predictions, with the preparation gap increasing from 0.235 to 1.197 across the same model sequence.

The loop is also visible in model state. Preparation effects decay gradually through long contradictory suffixes. Explicit RESET boundaries reduce the gap while leaving residual memory. Causal interventions locate transferable task-state information in late final-token residual activations, whereas query-token patching has little effect. Learned task-axis interventions move stuck predictions 4 to 9$\times$ more than matched random directions. The evidence suggests that scaling strengthens in-context rule induction while making the induced task state more inertial.

The paper makes five main contributions: (1) We identify task revision under conflicting evidence as a distinct problem in ICL and cast prompt hysteresis as a path-dependence test for persistent task-state dynamics. (2) We introduce task magnetization and hysteresis area as order parameters for in-context task inertia. (3) We develop a Landau-style effective theory that derives hysteresis, suffix persistence, partial RESET effects, and task-axis intervention predictions from a shared metastable order parameter. (4) We construct a semantic-free symbolic/cipher benchmark with matched trajectory, matched suffix, and RESET controls, reducing confounds from label semantics, task priors, and later evidence. (5) We provide a behavioral-to-mechanistic evidence chain across Qwen3 dense models, showing that task inertia increases with scale, persists through long contradictory suffixes, and is carried by late final-token residual states.

2 Prompt Hysteresis

2.1 Rules and Matched Trajectories

Prompt hysteresis is measured with pairs of incompatible in-context rules. Each experiment specifies two functions, Rule A and Rule B, over a shared input set $\mathcal{X}$, with conflicting outputs for every queried item.

$$f_A(x) \neq f_B(x), \quad x \in \mathcal{X}. \tag{1}$$

The prompt contains demonstrations from the two rules and ends with a query whose two rule-consistent answers are known. Semantic-free task families limit confounds from label meaning, world knowledge, and natural-language task priors.

- **Symbol mapping** maps nonce symbols to arbitrary labels under conflicting A/B mappings.
- **Cipher** generates character transformations from two incompatible cipher rules.

For a fixed number of demonstrations n, we compare two prompt histories at each rule proportion. Let j be the number of Rule-B demonstrations and $r = j/n$. The matched prompts at this ratio have the same evidence counts but different within-prompt orders:

$$\text{A-to-B: } A^{n-j}B^{j}, \qquad \text{B-to-A: } B^{j}A^{n-j}. \tag{2}$$

At each r, the comparison fixes current evidence and varies only whether the prompt approaches that point from an A-prefix or a B-prefix.

2.2 Magnetization and Loop Area

For each query, we score the two rule-consistent candidate answers, y_A and y_B, by output likelihood. The scores are normalized into a two-way preference:

$$P(A) = \frac{p(y_A \mid x)}{p(y_A \mid x) + p(y_B \mid x)}, \tag{3}$$

$$P(B) = \frac{p(y_B \mid x)}{p(y_A \mid x) + p(y_B \mid x)}. \tag{4}$$

The normalization isolates the competition between the two incompatible rules. Task magnetization is

$$m = P(B) - P(A). \tag{5}$$

Values near $m = -1$ indicate preference for Rule A. Values near $m = +1$ indicate preference for Rule B. Values near zero indicate comparable likelihood for the two rule-consistent candidates.

Prompt hysteresis is the area between the two magnetization curves:

$$H = \int |m_{A\to B}(r) - m_{B\to A}(r)|\, dr. \tag{6}$$

In implementation, H is computed on the discrete ratio grid induced by n. A memoryless evidence aggregator predicts identical curves at matched r, giving $H = 0$ up to estimation noise. Larger H indicates stronger path dependence, where the same current evidence yields different rule preferences after different preparation histories.

2.3 Persistence and Reset Controls

The loop test compares full trajectories. A stricter control holds later evidence fixed. The model is first prepared with either a pure A-prefix or a pure B-prefix. The same suffix evidence of length k is then appended to both prompts, giving the absolute preparation gap

$$|\Delta m_k| = |m_{\text{B prep}}(k) - m_{\text{A prep}}(k)|. \quad (7)$$

High $|\Delta m_k|$ indicates that the earlier task state survives under identical later evidence.

To study how this memory evolves, we sweep suffix length k after a fixed preparation prefix. For each task, seed, instruction, suffix rule, and k, the paired prompts differ only in whether the preparation prefix follows Rule A or Rule B. The trajectory of $|\Delta m_k|$ measures state memory as contradictory evidence accumulates.

RESET tests whether an explicit boundary can modulate the prepared state. In this condition, the preparation examples are placed before the instruction:

```
RESET. The examples before RESET are
obsolete. Use only the examples after
RESET for the query.
```

RESET separates ordinary recency from partial state reinitialization. A smaller $|\Delta m_k|$ after the boundary indicates revision of the prepared state. A nonzero residual gap indicates incomplete erasure.

2.4 Activation-Level Probes

Behavioral loops and suffix gaps show output-level path dependence. Three activation-level probes test whether this dependence is carried by a manipulable internal coordinate.

Latent localization. For each layer and token role, we define a task axis by contrasting residual activations from pure-A and pure-B prompts. Mixed prompts are projected onto this axis, and latent hysteresis area is computed in the same way as the output loop.

Residual patching. For suffix prompts with different preparation states, we patch a residual activation from a clean prompt into a corrupted prompt at a chosen layer and token role. The causal effect is the change in task magnetization. A large effect at one token role, paired with a near-zero role control, isolates task-state transfer from nonspecific activation replacement.

Task-vector interventions. Activations are manipulated along the A/B task axis by adding, erasing, or reflecting the direction. Matched random directions with the same norm serve as negative controls. The random-direction control tests whether the axis is interventionally specific to the prepared rule.

3 A Macroscopic Theory of Prompt-Induced Task States

3.1 Memoryless Baseline

The measurements above treat rule preference as an order parameter and prompt composition as a control variable. A minimal theory separates the evidence field supplied by the demonstrations from the task state prepared by the model. Let

$$u(r) = \gamma(2r - 1) \quad (8)$$

be the evidence field induced by the Rule-B fraction r. In an evidence-aggregation account with no persistent state, m is a single-valued function of u. Matched A-to-B and B-to-A prompts then have identical magnetization curves, so $H = 0$. A nonzero loop requires an internal state that can remain locally stable after the evidence field changes.

3.2 Effective Potential

We model this stability with an effective Landau potential over m.

$$\mathcal{F}(m, u) = \frac{a}{4}m^4 - \frac{\theta}{2}m^2 - um, \quad a > 0. \quad (9)$$

The form is the lowest-order symmetric expansion with stable large-$|m|$ behavior and linear coupling between evidence and task state. The quartic term stabilizes the potential, the quadratic term controls task-state ordering, and the linear term biases the state toward the rule favored by the current evidence. When $\theta \leq 0$, the potential has one stable basin and predicts no hysteresis. When $\theta > 0$, two basins appear at zero field:

$$m_{\pm} = \pm\sqrt{\theta/a}. \quad (10)$$

During an evidence sweep, each basin persists until a spinodal point where the local minimum disappears. Solving $\partial_m \mathcal{F} = 0$ and $\partial_m^2 \mathcal{F} = 0$ gives

$$|u_{\mathrm{sp}}| = \frac{2\theta^{3/2}}{3\sqrt{3a}}. \tag{11}$$

The interval $|u| < |u_{\mathrm{sp}}|$ admits two locally stable task states. Forward and reverse prompt paths can follow different branches through the same evidence ratios.

3.3 Attractor Interpretation

The same form arises from a mean-field attractor update,

$$m = \tanh\{\beta(Jm + u)\}. \tag{12}$$

Here J measures task-state self-consistency and β is an effective inverse temperature. The corresponding variational potential is

$$\mathcal{F}(m, u) = -\frac{J}{2}m^2 - um + \frac{1}{\beta}\phi(m), \tag{13}$$

$$\phi(m) = \frac{1+m}{2}\log\frac{1+m}{2} + \frac{1-m}{2}\log\frac{1-m}{2}. \tag{14}$$

Expanding near $m = 0$ yields

$$\mathcal{F}(m, u) = C + \frac{1}{2}\left(\frac{1}{\beta} - J\right)m^2 + \frac{1}{12\beta}m^4 - um + O(m^6). \tag{15}$$

Metastability begins when $\beta J > 1$, and the loop area near onset scales as

$$H = O\left((\beta J - 1)^2\right). \tag{16}$$

In this interpretation, J is the strength of task-state self-consistency and β is an inverse inferential temperature. Increasing either quantity deepens the basins and strengthens task inertia. The behavioral prediction is that models with stronger in-context task representations can also show larger hysteresis.

3.4 Predictions for Controls and Interventions

Suffix persistence follows from relaxation inside a metastable basin. Under a fixed suffix field u_k, write

$$\tau\dot{m} = -\partial_m \mathcal{F}(m, u_k) + \sqrt{2D}\xi(t). \tag{17}$$

Linearizing around a local minimum $m_\star$ gives

$$\delta m_{k+1} \approx \exp[-\eta\mathcal{F}''(m_\star, u_k)]\delta m_k. \tag{18}$$

Small curvature gives slow decay of the preparation gap, with the slowest relaxation near a spinodal. With noise, escape from a metastable basin has the activated form

$$\tau_{\mathrm{esc}} \propto \exp(\Delta\mathcal{F}/D). \tag{19}$$

Stronger barriers or lower effective noise predict longer-lived suffix persistence.

A RESET boundary can be represented as a partial quench of the order parameter,

$$m^+ = (1-\rho)m^-, \quad 0 \le \rho \le 1. \tag{20}$$

The quench predicts a smaller preparation gap after RESET, with residual dependence whenever $\rho < 1$.

$$|\Delta m_k|_{\mathrm{RESET}} < |\Delta m_k|_{\text{no reset}}. \tag{21}$$

The same order parameter gives an activation-level prediction. If residual activations carry the task-state coordinate, there should be a direction

$$v_\ell = \mathrm{E}[h_\ell \mid B] - \mathrm{E}[h_\ell \mid A] \tag{22}$$

that changes the inferred rule when perturbed. Let the local log-odds be $q_\ell \approx w_\ell^\top h_\ell$ and $m = \tanh(q_\ell/2)$. A small activation change gives

$$\Delta m \approx \frac{1}{2}\,\mathrm{sech}^2(q_\ell/2)\, w_\ell^\top \Delta h_\ell. \tag{23}$$

Task-axis interventions should exceed matched random directions, whose expected projection onto the relevant readout is small in high-dimensional activation space. Residual patching should transfer the active task mode when it targets sites carrying this coordinate.

Empirical predictions. Taken together, the effective potential, relaxation dynamics, and activation readout specify the tests used below. Metastability implies that matched A-to-B and B-to-A trajectories can separate at the same evidence ratio, giving nonzero loop area. Scale dependence follows if larger models form deeper or lower-temperature task basins once prompts contain enough evidence to induce a stable rule. Under matched suffixes, the preparation gap should decay gradually, while a partial RESET quench should reduce the gap without eliminating it. At the activation level, late task-bearing residual states should transfer under patching, and targeted task-axis interventions should move predictions more than norm-matched random directions.

4 Experimental Setup

Models. The behavioral scaling suite evaluates four dense Qwen3 models, Qwen3-0.6B, Qwen3-1.7B, Qwen3-4B, and Qwen3-8B (Qwen Team, 2025). The mechanistic validation focuses on Qwen3-1.7B, Qwen3-4B, and Qwen3-8B. All experiments use non-thinking likelihood scoring with float16 inference.

Tasks and instructions. The curated experiment suite uses two semantic-free conflict tasks, `symbol` and `cipher`. Each task defines two mutually incompatible rules, Rule A and Rule B. Queries are scored by the likelihood of the two rule-consistent outputs. The scaling study averages over five instruction variants, `global`, `neutral`, `phase`, `recent`, and `sticky`.

Scaling sweep. We sweep $n \in \{4, 6, 8, 10, 12\}$ demonstrations per endpoint. The $n = 4$ condition is under-specified in the curated validation suite, so the paper-facing scaling analysis uses $n \in \{6, 8, 10, 12\}$ and separately reports the strongest $n = 12$ slice. The scaling validation uses held-out prompts and aggregates across tasks and instruction variants.

Suffix persistence control. The suffix persistence control is run at $n = 12$ on the same two tasks, same four models, and same five instruction variants. It compares magnetization after A-preparation versus B-preparation under matched suffix evidence and reports the mean absolute preparation gap.

State-dynamics validation. For Qwen3-1.7B, Qwen3-4B, and Qwen3-8B, we fix 12 preparation examples and sweep suffix length $k = 0, \ldots, 24$. We use the `symbol` and `cipher` tasks, 12 held-out seeds, and four instruction types: `neutral`, `recent`, `global`, and `sticky`. Each run includes no-reset and RESET prompts, for 17,600 scored prompts per model.

Mechanistic intervention suite. The localization, residual patching, and task-vector experiments use the same semantic-free tasks and held-out seeds. Residual patching compares late final-token activations against query-token role controls. Task-vector experiments compare interventions along the learned task axis against matched random directions. We retain the full raw outputs for auditability and report the predeclared paper-facing aggregates in the main text.

5 Results

5.1 Prompt Hysteresis Scales with Model Size

Figure 1 reports the main scaling result. Once prompts contain at least six demonstrations, hysteresis area increases monotonically across the Qwen3 model scale. Table 1 summarizes the curated $n \geq 6$ aggregate. Mean H rises from 0.067 for Qwen3-0.6B to 0.269 for Qwen3-8B.

Model	Mean H	Std.
Qwen3-0.6B	0.067	0.046
Qwen3-1.7B	0.120	0.061
Qwen3-4B	0.209	0.114
Qwen3-8B	0.269	0.109

Table 1: Mean hysteresis area H on the symbolic and cipher suite for prompts with $n \geq 6$ demonstrations. Rows average 40 loops per model and report standard deviations across loops.

The $n = 12$ slice yields the largest separation between small and large models. Table 2 reports a rise in mean H from 0.077 to 0.309 across the same model sequence, matching the $n = 12$ panel in Figure 1.

Model	Mean H	Std.
Qwen3-0.6B	0.077	0.053
Qwen3-1.7B	0.124	0.057
Qwen3-4B	0.230	0.131
Qwen3-8B	0.309	0.104

Table 2: Mean hysteresis area H in the twelve-demonstration setting. Rows average 10 loops per model and report standard deviations across loops.

5.2 Representative Hysteresis Loops

The aggregate scaling result arises from loop-level path separation. The two rightmost panels of Figure 1 display representative high-hysteresis loops for Qwen3-8B at $n = 12$. The A-to-B and B-to-A trajectories diverge at matched rule ratios, indicating that prompt history changes the model's rule preference even when the final counts of A and B demonstrations are fixed.

5.3 Suffix Persistence Under Matched Later Evidence

The suffix persistence control fixes later evidence and varies only the preparation prefix. Table 3 gives a stronger monotonic scaling pattern than the main hysteresis area. The mean absolute preparation

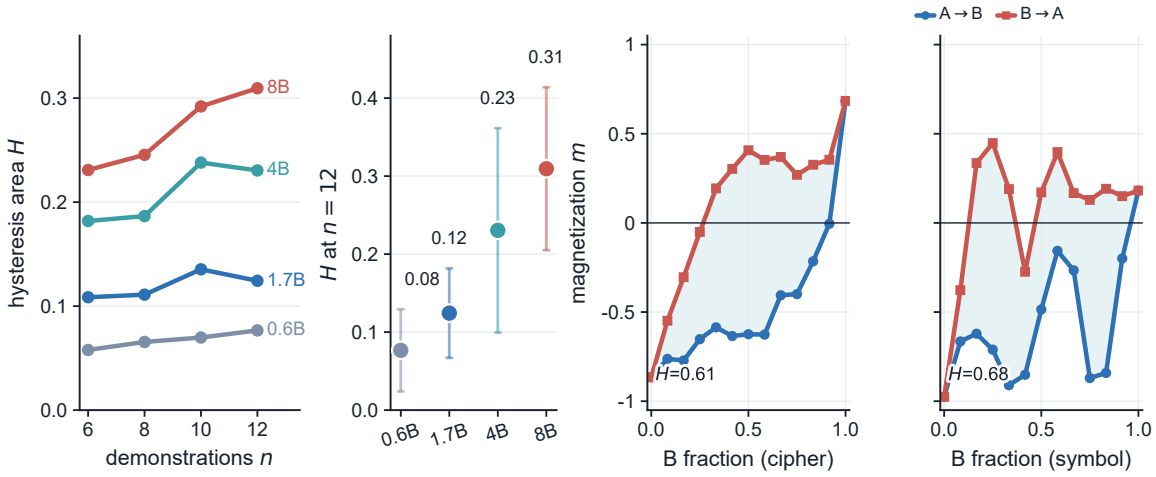

Figure 1: Behavioral prompt hysteresis across model size and prompt length. Left, mean hysteresis area H over $n \in \{6, 8, 10, 12\}$ demonstrations. Center, the twelve-demonstration comparison with standard deviations across loops. Right, representative Qwen3-8B A-to-B and B-to-A trajectories at matched Rule-B fractions. Shading marks the integrated hysteresis area.

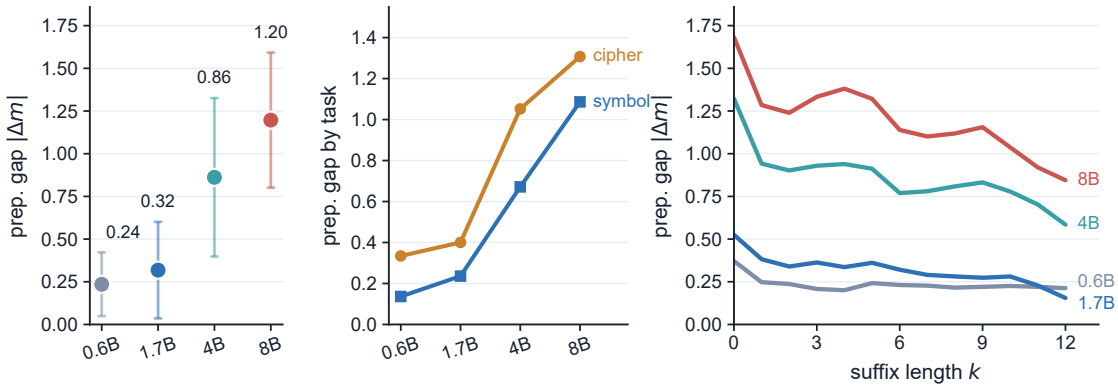

Figure 2: Suffix persistence with matched later evidence. The suffix is held fixed while the preparation prefix varies. Left, mean absolute preparation gap with standard deviations across matched comparisons. Middle, preparation gap split by task family. Right, gap as a function of suffix length k. Larger values indicate stronger dependence on the preparation prefix.

gap rises from 0.235 for Qwen3-0.6B to 1.197 for Qwen3-8B. Figure 2 plots the same result.

Model	Mean $\lvert\Delta m\rvert$	Std.
Qwen3-0.6B	0.235	0.187
Qwen3-1.7B	0.318	0.283
Qwen3-4B	0.862	0.464
Qwen3-8B	1.197	0.395

Table 3: Mean suffix-persistence gap at $n = 12$. Rows average 390 matched suffix comparisons per model and report standard deviations. Larger values indicate stronger dependence on the preparation prefix.

Figure 2 and Table 4 split the effect by task family. Cipher tasks have the largest absolute gaps. Both cipher and symbol tasks increase monotonically with model size. The preparation gap also remains visible as suffix length grows.

Task	0.6B	1.7B	4B	8B
cipher	0.334	0.400	1.052	1.307
symbol	0.136	0.236	0.672	1.086

Table 4: Mean suffix-persistence gap by task family at $n = 12$. Columns correspond to Qwen3 model size.

5.4 State Dynamics Reveal Metastable Preparation Memory

The suffix control identifies preparation dependence under matched later evidence. The state-dynamics validation tracks this dependence as suffix evidence accumulates. Figure 3 reports gradual decay across many contradictory examples. With-

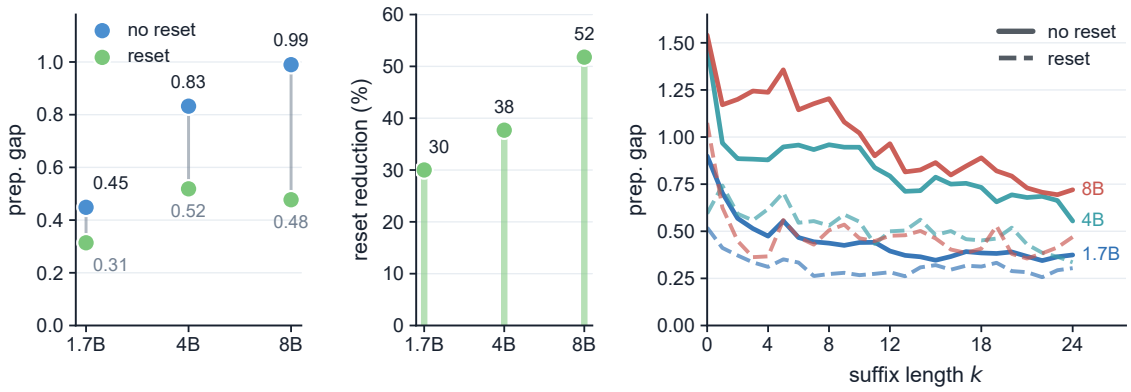

Figure 3: State-dynamics probes for preparation memory. Left, no-reset and RESET preparation gaps by model. Center, relative reduction under RESET. Right, gap over suffix length k for no-reset and RESET prompts. Solid and dashed lines indicate the two boundary conditions.

out RESET, the mean preparation gap increases from 0.449 in Qwen3-1.7B to 0.990 in Qwen3-8B (Table 5). The tail gap at $k \geq 18$ remains substantial, rising from 0.373 to 0.765 across the same scale.

Explicit RESET boundaries reduce this memory. The relative reduction grows with scale, from 30.0% for Qwen3-1.7B to 37.7% for Qwen3-4B and 51.8% for Qwen3-8B. The model partially reinitializes its prompt-induced state when earlier evidence is marked obsolete, while retaining a measurable residual dependence on the preparation prefix.

Model	No reset	RESET	Red.	Tail
Qwen3-1.7B	0.449	0.314	30.0%	0.373
Qwen3-4B	0.832	0.519	37.7%	0.666
Qwen3-8B	0.990	0.477	51.8%	0.765

Table 5: State-dynamics summary by model. "No reset" and "RESET" are mean preparation gaps across suffix lengths, "Red." is the relative reduction, and "Tail" is the no-reset mean for $k \geq 18$.

5.5 Mechanistic Evidence for Task-State Carriers

Behavioral hysteresis and state dynamics point to persistent prompt-induced task states. We examine this interpretation with hidden-state localization, residual patching, and task-vector interventions. Figure 4 summarizes the corresponding localization and intervention evidence.

Latent localization. Pure-A and pure-B prompts define a task axis at each layer, and mixed prompts are projected onto this axis. In Qwen3-4B, latent hysteresis peaks in middle layers (Figure 4). Average latent area rises from near zero at the embedding layer to above 0.8 around layer 19, with the strongest label-swap groups reaching latent areas around 1.20. Cross-model localization also identifies late-layer task-bearing sites at final, query, and first-demonstration positions. These projections provide localization evidence, while raw magnitudes remain model-scale dependent.

Residual patching. Late final-token residual activations transfer the active task mode (Figure 4). On strong-gap prompts, patching the late final token changes magnetization by 0.927, 0.880, and 0.841 for Qwen3-1.7B, Qwen3-4B, and Qwen3-8B, respectively. The late query-token role control remains near zero, with effects of 0.000, 0.014, and 0.011. The role contrast isolates a causal carrier of the task state.

Task-vector interventions. The learned task axis is interventionally specific. Adding along the task axis changes magnetization much more than matched random directions. The task-axis/random ratio increases from 4.1× to 8.9× from Qwen3-1.7B to Qwen3-8B. On stuck-A prompts, reflecting activations across the task axis releases the model toward Rule B, with mean flip deltas of 0.932, 0.314, and 0.674 across the three models. Combined with residual patching, these interventions identify a manipulable task-state coordinate.

6 Related Work

In-context learning and task state. ICL has been analyzed as latent-task inference (Xie et al., 2022; Chan et al., 2022; Pan et al., 2023), as in-

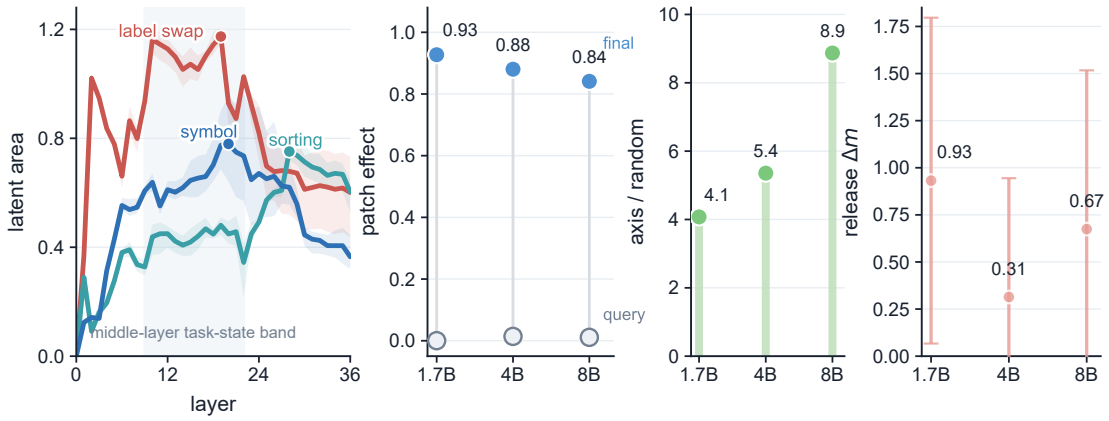

Figure 4: Mechanistic evidence for task-state carriers. From left to right, Qwen3-4B latent hysteresis by layer and task, late final-token versus query-token residual patching, task-axis effects relative to matched random directions, and task-axis reflection on stuck-A prompts. Shaded bands and error bars report variation across groups where plotted.

context execution of learning algorithms (Garg et al., 2022; Akyürek et al., 2023; von Oswald et al., 2023; Dai et al., 2023; Raventós et al., 2023), and as the formation of induction heads or task/function vectors (Olsson et al., 2022; Hendel et al., 2023; Todd et al., 2024). Work on flipped labels and symbol tuning also shows that models can adopt input-label mappings with limited semantic support (Wei et al., 2023b,a). These accounts explain how a prompt induces a task representation. We ask whether that representation changes reversibly when later demonstrations contradict earlier ones.

Order effects and context dependence. Prompt behavior depends on calibration, demonstration choice, and ordering (Zhao et al., 2021; Lu et al., 2022; Rubin et al., 2022; Liu et al., 2022). Demonstrations also provide label-space, input-distribution, and formatting cues beyond the ground-truth mapping (Min et al., 2022). Prompt wording and long-context position shape which evidence is used (Webson and Pavlick, 2022; Liu et al., 2024). We treat this sensitivity as the object of study by matching rule-mixture ratios across opposite trajectories and by holding suffix evidence fixed while varying earlier preparation.

Mechanistic localization and activation interventions. Causal tracing and circuit analysis localize computations inside neural language models (Meng et al., 2022; Wang et al., 2023). Representation engineering and activation additions show that linear residual-stream directions can monitor or steer model behavior (Zou et al., 2023; Turner et al., 2023). We use this intervention style to test whether prompt-induced task states have a manipulable activation coordinate.

Phase transitions and physics-inspired views. Attractor and spin-glass models describe computation through stable states and basins (Hopfield, 1982; Amit et al., 1985). Recent work studies abrupt changes in model capabilities, grokking dynamics, and output-distribution phase transitions (Wei et al., 2022; Power et al., 2022; Nanda et al., 2023; Arnold et al., 2024). Prompt hysteresis brings this vocabulary to ICL without training or parameter sweeps. The control variable is the path through contradictory in-context evidence, and the measured object is the loop traced by the model's task state.

7 Conclusion

We introduced prompt hysteresis as a measurable form of path dependence in in-context learning. Forward and reverse paths through conflicting symbolic rules reveal persistent task states whose influence remains visible under matched evidence. In semantic-free symbolic/cipher tasks, hysteresis area, suffix persistence, and state-dynamics memory increase with Qwen3 model size. Mechanistic probes localize and intervene on this state. Late final-token residual patching transfers it, and task-axis interventions modulate it far more than random directions. Scaling can make in-context task representations more stable as well as more effective.

Limitations

The study focuses on a curated semantic-free symbolic/cipher suite and Qwen3 dense models. Other task families or prompting formats may exhibit different loop shapes. The mechanistic validation covers Qwen3-1.7B, Qwen3-4B, and Qwen3-8B. Smaller models and larger open-weight models remain natural targets for the same probes. Latent projections support localization, while raw projection magnitudes require additional normalization for cross-model comparisons. We score two rule-consistent candidates by likelihood. Open-ended generation would add parsing and decoding effects beyond the scope of the present measurements.

References

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2023. What learning algorithm is in-context learning? investigations with linear models. In *International Conference on Learning Representations*.

Daniel J. Amit, Hanoch Gutfreund, and Haim Sompolinsky. 1985. Spin-glass models of neural networks. *Physical Review A*, 32(2):1007–1018.

Julian Arnold, Flemming Holtorf, Frank Schäfer, and Niels Lörch. 2024. Phase transitions in the output distribution of large language models. *arXiv preprint arXiv:2405.17088*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Stephanie C. Y. Chan, Adam Santoro, Andrew K. Lampinen, Jane X. Wang, Aaditya Singh, Pierre H. Richemond, James L. McClelland, and Felix Hill. 2022. Data distributional properties drive emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems*, volume 35, pages 18878–18891.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4005–4019. Association for Computational Linguistics.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. 2022. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems*.

Roee Hendel, Mor Geva, and Amir Globerson. 2023. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9318–9333. Association for Computational Linguistics.

John J. Hopfield. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 8086–8098. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064. Association for Computational Linguistics.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, and 1 others. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. 2023. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8298–8319. Association for Computational Linguistics.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. 2022. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint*.

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. 2023. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. In *Advances in Neural Information Processing Systems*, volume 36.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671. Association for Computational Linguistics.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2024. Function vectors in large language models. In *International Conference on Learning Representations*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 35151–35174. PMLR.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, and 1 others. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Jerry Wei, Le Hou, Andrew Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, and Quoc Le. 2023a. Symbol tuning improves in-context learning in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 968–979. Association for Computational Linguistics.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023b. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

Name	Instruction text
`neutral`	Infer the input-output rule from the examples and answer the query.
`global`	Infer one stable hidden rule from all examples. Treat contradictions as noise and answer according to the best global rule.
`recent`	The rule may change over time. For the query, use the most recent examples as the strongest evidence.
`phase`	Examples are ordered in phases. Infer the current phase at the end of the prompt and answer according to that phase.
`sticky`	Assume the earliest consistent rule is usually stable unless later evidence overwhelmingly contradicts it.

Table 6: Instruction variants used in the prompt-hysteresis experiments.

A Prompt Templates

All likelihood experiments render the prompt as a single user message and apply the Qwen3 chat template with thinking disabled. The reported metrics are computed from candidate continuation likelihoods. No sampled decoding enters the measurements.

A.1 Instruction Variants

Table 6 lists the instruction variants. Behavioral scaling averages over all five variants. State-dynamics runs use `neutral`, `recent`, `global`, and `sticky`; latent-axis runs use `neutral`, `recent`, and `sticky`. The set spans neutral aggregation, recency-sensitive revision, global-rule inference, phase inference, and early-rule persistence.

A.2 Base Block Prompt

For a task instance with context c, demonstration sequence $s \in \{A, B\}^n$, and query x_q, the block prompt is rendered as follows.

```
[instruction]

[task context]
Examples are ordered from earliest to
latest.

Examples:
1. x_1 -> y_1
2. x_2 -> y_2
...
n. x_n -> y_n

Query:
x_q ->

Answer with only the output.
```

At ratio $r = j/n$, the matched block paths are $A^{n-j}B^j$ and B^jA^{n-j}. They contain the same A/B counts and differ only in demonstration order.

A.3 Forced-Choice Prompt

The forced-choice control uses the same prompt body and replaces the final output instruction with two counterbalanced candidates.

```
Candidate outputs:
1. y_π(1)
2. y_π(2)

Answer with only 1 or 2.
```

Candidate order π is enumerated in both orders or assigned by a deterministic hash. Scores are mapped back to Rule A and Rule B before computing magnetization.

A.4 Suffix and RESET Prompts

Suffix controls prepare the model with a pure-rule prefix and then append matched later evidence. The no-reset condition concatenates preparation and suffix demonstrations into one example list. The RESET condition marks the prefix obsolete before presenting the suffix evidence.

```
[instruction]

[task context]
Examples are ordered from earliest to
latest.

Preparation examples:
1. x_1 -> y_1
...

RESET.
The examples before RESET are obsolete.
Use only the examples after RESET for
the query.

Current examples:
1. x'_1 -> y'_1
...

Query:
x_q ->

Answer with only the output.
```

The no-reset counterpart removes the boundary and lists preparation and suffix rows under a single `Examples:` header.

B Rule Families and Instance Generation

The reported benchmark uses two semantic-free task families. Both are generated deterministically from task seeds. Each seed defines two incompatible functions over a shared input set and one held-out query with distinct A-consistent and B-consistent answers.

Symbol mapping. The instance generator shuffles a nonce-word pool containing `dax`, `blicket`, `wug`, `zorp`, `mip`, `toma`, `kiki`, `fep`, `nalo`, and `soma`. The prompt context states the fixed order of six selected words. Rule A maps the ordered words to $1, 2, \ldots, 6$, while Rule B maps them to $9, 8, \ldots, 4$. One word is held out as the query. The remaining selected words supply demonstrations and are cycled deterministically when more rows are required. The query input is excluded from the demonstrations.

Cipher. The instance generator shuffles lowercase letters after removing `a` and `z` from the candidate pool. Rule A applies a Caesar shift of $+1$, and Rule B applies a Caesar shift of -1. The first shuffled letter is held out as the query. Subsequent letters supply demonstrations and are cycled only after the unique pool has been exhausted.

Task contexts. The symbol context is "The input symbols have this fixed order: [six words]." The cipher context is "Each input is a lowercase letter. The hidden rule is a Caesar shift." The contexts specify the domain and leave the A/B rule choice to the demonstrations.

C Scoring and Aggregation

For each rendered prompt, the model is scored on the two rule-consistent candidate continuations y_A and y_B. Let z_A and z_B denote their log-likelihood scores. For multi-token candidates, the implementation records both summed and mean log-likelihoods. The reported symbolic/cipher analyses normalize the two candidates as

$$P(A) = \frac{\exp(z_A)}{\exp(z_A) + \exp(z_B)}, \tag{24}$$

$$P(B) = \frac{\exp(z_B)}{\exp(z_A) + \exp(z_B)}. \tag{25}$$

Task magnetization is

$$m = P(B) - P(A). \tag{26}$$

The discrete hysteresis area reported in the tables is

$$H = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \left| m_{A \to B}(r) - m_{B \to A}(r) \right|, \tag{27}$$

where $\mathcal{R} = \{0, 1/n, \ldots, 1\}$. Switching thresholds use the first positive crossing of m on the A-to-B path and the last negative crossing on the B-to-A path.

Suffix persistence is computed after grouping by model, task, seed, instruction, suffix condition, and suffix length. For a matched suffix length k,

$$|\Delta m_k| = \left| m_{\text{B prep}}(k) - m_{\text{A prep}}(k) \right|. \tag{28}$$

The RESET reduction is

$$1 - \frac{\mathrm{E}[|\Delta m_k|_{\text{RESET}}]}{\mathrm{E}[|\Delta m_k|_{\text{no reset}}]}. \tag{29}$$

Run	Setting
Scaling validation	Qwen3-0.6B, 1.7B, 4B, and 8B; tasks `symbol,cipher`; instructions `neutral,recent,phase,global,sticky`; seeds 600–615; $n \in \{6, 8, 10, 12\}$; output likelihood; block paths.
Final $n = 12$ controls	Same four models and two tasks; same five instructions; seeds 700–715; $n = 12$; forced-choice block paths with both candidate orders; suffix output prompts with 12 preparation examples and suffix length up to 12.
Confirmatory held-out extension	Same four models and two tasks; instructions `neutral,recent,global,sticky`; seeds 800–823; $n \in \{6, 8, 10, 12\}$; output likelihood; block paths.
State dynamics and RESET	Qwen3-1.7B, 4B, and 8B; tasks `symbol,cipher`; instructions `neutral,recent,global,sticky`; 12 preparation examples; suffix length $k = 0, \ldots, 24$; no-reset and RESET prompts.
Latent-axis scale	Qwen3-1.7B, 4B, and 8B; tasks `symbol,cipher`; instructions `neutral,recent,sticky`; seeds 900–907; $n = 12$; layerwise residual projections.
Mechanistic interventions	Qwen3-1.7B, 4B, and 8B; semantic-free prompts from the same task family; late final-token patching, query-token role controls, task-axis additions, random-direction controls, and stuck-state reflection.

Table 7: Experimental configurations used for the reported analyses.

D Experimental Configuration

Table 7 summarizes the configurations used for the reported analyses. All experiments use deterministic likelihood scoring, Qwen3 non-thinking mode, local checkpoint loading, and float16 inference. The archived environment snapshot records package exports, checkpoint file inventories, config and tokenizer hashes, weight-shard hashes, script hashes, and selected result hashes.

Model identity is pinned by local file inventories and checkpoint hashes. The result bundle also records the scripts used to produce the reported tables and figures.

n	Qwen3-0.6B	Qwen3-1.7B	Qwen3-4B	Qwen3-8B
6	0.073	0.128	0.215	0.346
8	0.079	0.122	0.232	0.391
10	0.081	0.141	0.286	0.424
12	0.090	0.136	0.300	0.414

Table 8: Mean hysteresis area H in the independent held-out extension. Rows average over symbol/cipher tasks and four instruction variants.

E Additional Behavioral Results

Table 8 reports an independent held-out extension over seeds 800–823. The run uses the same semantic-free task families and four instruction variants. The monotonic scale pattern replicates at every demonstration count.

The extension gives an aggregate check on the main scaling result. The ordering is stable across prompt lengths, and the largest separation remains between Qwen3-0.6B and Qwen3-8B.

F Additional State-Dynamics Results

The state-dynamics runs sweep suffix length from $k = 0$ to $k = 24$ under no-reset and RESET prompts. Preparation gaps peak immediately after the prefix and then decay gradually across contradictory suffixes. The long-suffix tail remains larger in the larger models, and RESET consistently attenuates the gap. This pattern matches the metastable relaxation picture in Section 3.

G Mechanistic Probe Details

Task axis. For each layer ℓ and token role, pure-A and pure-B prompts define

$$v_\ell = \frac{\mu_B^\ell - \mu_A^\ell}{\|\mu_B^\ell - \mu_A^\ell\|_2 + \epsilon}. \tag{30}$$

Mixed prompts are projected as $s_\ell = h_\ell^\top v_\ell$. Latent hysteresis is computed by replacing output magnetization with the normalized projection score along the same A-to-B and B-to-A paths.

Residual patching. Let $h_{\ell,t}^{\text{clean}}$ be a residual activation from a prompt prepared in the target task state and $h_{\ell,t}^{\text{corr}}$ be the corresponding activation from the opposite preparation. Patching replaces $h_{\ell,t}^{\text{corr}}$ with $h_{\ell,t}^{\text{clean}}$ at a chosen layer and token role, then recomputes candidate likelihoods. The reported causal effect is the resulting change in magnetization. Late final-token patching is compared with query-token patching under the same prompt pairs.

Task-vector interventions. For an activation $h_{\ell,t}$, the add intervention uses

$$h'_{\ell,t} = h_{\ell,t} + \alpha v_\ell. \tag{31}$$

The erase intervention removes the projection along v_ℓ, and reflection maps the activation across the hyperplane orthogonal to v_ℓ:

$$h'_{\ell,t} = h_{\ell,t} - (h_{\ell,t}^\top v_\ell) v_\ell, \tag{32}$$

$$h'_{\ell,t} = h_{\ell,t} - 2(h_{\ell,t}^\top v_\ell) v_\ell. \tag{33}$$

Matched random controls use directions sampled independently and rescaled to the same norm as v_ℓ. The comparison separates effects of the A/B task coordinate from effects of arbitrary residual-stream perturbations.

H Theoretical Derivations

H.1 Spinodal Field and Metastable Branches

For the effective potential in Eq. 9, the derivative entering the stationarity condition is

$$\partial_m \mathcal{F} = am^3 - \theta m - u \tag{34}$$

and the local curvature is

$$\partial_m^2 \mathcal{F} = 3am^2 - \theta. \tag{35}$$

For $\theta > 0$ and $u = 0$, the potential has two stable minima at

$$m_\pm = \pm\sqrt{\theta/a} \tag{36}$$

and an unstable stationary point at $m = 0$. The two minima represent the A-like and B-like task states. A nonzero field u tilts the potential toward the state aligned with the current evidence. The stationary points satisfy

$$u = am^3 - \theta m. \tag{37}$$

At a spinodal, a local minimum merges with the neighboring saddle. The first and second derivatives therefore vanish together, and the curvature condition gives

$$m_{\text{sp}} = \pm\sqrt{\theta/(3a)}. \tag{38}$$

Substituting into the stationarity condition yields

$$u_{\rm sp} = a m_{\rm sp}^3 - \theta m_{\rm sp}, \tag{39}$$

and therefore

$$|u_{\rm sp}| = \frac{2\theta^{3/2}}{3\sqrt{3a}}. \tag{40}$$

The negative branch remains metastable for $0 < u < u_{\rm sp}$ and disappears at $u_{\rm sp}$. The positive branch remains metastable for $-u_{\rm sp} < u < 0$ and disappears at $-u_{\rm sp}$. An A-prepared prompt swept toward B and a B-prepared prompt swept toward A therefore follow different branches throughout the interval

$$-u_{\rm sp} < u < u_{\rm sp}. \tag{41}$$

This interval defines the macroscopic hysteresis window.

H.2 Discrete Prompt Grid

The behavioral experiments control the fraction r of B demonstrations rather than the field directly. Under the linear field in Section 3,

$$u(r) = \gamma(2r - 1), \tag{42}$$

the spinodal fields map to two critical fractions

$$r_\pm = \frac{1}{2}\left(1 \pm \frac{u_{\rm sp}}{\gamma}\right). \tag{43}$$

An A-to-B sweep switches near r_+, while a B-to-A sweep switches near r_-. The predicted separation between the two switching locations is

$$r_+ - r_- = \frac{u_{\rm sp}}{\gamma}. \tag{44}$$

For a prompt with n demonstrations, observed thresholds are quantized on the grid $r = j/n$. The expected A-to-B threshold is the first grid point whose field exceeds $u_{\rm sp}$, and the expected B-to-A threshold is the last grid point whose field remains below $-u_{\rm sp}$. Changes in demonstration count can therefore shift discrete thresholds without changing the underlying branch structure.

H.3 Mean-Field Expansion

The attractor equation $m = \tanh\{\beta(Jm + u)\}$ follows from the variational potential

$$\mathcal{F}(m, u) = -\frac{J}{2}m^2 - um + \frac{1}{\beta}\phi(m), \tag{45}$$

where

$$\phi(m) = \frac{1+m}{2}\log\frac{1+m}{2} + \frac{1-m}{2}\log\frac{1-m}{2}. \tag{46}$$

Expanding the entropy term around zero gives

$$\phi(m) = -\log 2 + \frac{m^2}{2} + \frac{m^4}{12} + O(m^6). \tag{47}$$

Thus

$$\mathcal{F}(m, u) = C + \frac{1}{2}\left(\frac{1}{\beta} - J\right)m^2 + \frac{1}{12\beta}m^4 - um + O(m^6). \tag{48}$$

Matching this expansion to Eq. 9 gives

$$\theta = J - \frac{1}{\beta}, \qquad a = \frac{1}{3\beta}. \tag{49}$$

The symmetric state becomes unstable when $\beta J > 1$. Let $\epsilon = \beta J - 1$. In the near-critical regime,

$$m_0^2 = \frac{\theta}{a} = 3\epsilon \tag{50}$$

and

$$u_{\rm sp} = \frac{2}{3\beta}\epsilon^{3/2} + O(\epsilon^{5/2}). \tag{51}$$

The branch separation scales as $\epsilon^{1/2}$, while the hysteresis width scales as $\epsilon^{3/2}$. The leading loop-area scaling is therefore $O(\epsilon^2)$, as stated in Section 3.

H.4 Suffix Relaxation and RESET

The suffix experiments correspond to a discrete relaxation process under a sequence of suffix fields u_k:

$$m_{k+1} = m_k - \eta\,\partial_m\mathcal{F}(m_k, u_k) + \sqrt{2\eta D}\,\zeta_k, \tag{52}$$

where η is an effective update scale and ζ_k is unit-variance noise. Linearization around a local minimum $m_\star(u_k)$ gives

$$\delta m_{k+1} \approx \left[1 - \eta\,\partial_m^2\mathcal{F}(m_\star, u_k)\right]\delta m_k. \tag{53}$$

Across a suffix of length K, the preparation gap approximately evolves as

$$|\Delta m_K| \approx |\Delta m_0| \prod_{k=0}^{K-1} \left|1 - \eta\,\partial_m^2\mathcal{F}(m_\star, u_k)\right|. \tag{54}$$

Relaxation slows near a spinodal because the curvature term approaches zero. With noise, basin escape follows the usual barrier dependence

$$\tau_{\rm esc} \propto \exp(\Delta\mathcal{F}/D), \tag{55}$$

so larger barriers or lower effective noise imply longer preparation memory.

The RESET model in the main text represents a boundary as a partial quench $m^+ = (1-\rho)m^-$. For paired A-prepared and B-prepared prompts, the immediate gap transforms as

$$\Delta m^+ = (1-\rho)\Delta m^-. \tag{56}$$

After a shared suffix, the same local relaxation factors apply to both preparations:

$$|\Delta m_K^{\mathrm{RESET}}| \approx (1-\rho)|\Delta m_0| \prod_{k=0}^{K-1} \left|1 - \eta\, \partial_m^2 \mathcal{F}(m_\star, u_k)\right|. \tag{57}$$

RESET therefore attenuates path dependence at the boundary and leaves a residual gap when the quench is partial and the suffix dynamics do not force both trajectories into the same basin.

H.5 Task-Axis Interventions

For activation interventions, assume local log-odds $q = w^\top h + b$ and $m = \tanh(q/2)$. A perturbation Δh gives

$$\Delta m \approx \frac{1}{2}\,\mathrm{sech}^2(q/2) w^\top \Delta h. \tag{58}$$

Let the empirical task axis at layer ℓ be

$$v_\ell = \mathrm{E}[h_\ell \mid B] - \mathrm{E}[h_\ell \mid A], \tag{59}$$

with unit direction $\hat{v}_\ell = v_\ell / \|v_\ell\|$. Adding the task direction sets $\Delta h = \alpha \hat{v}_\ell$, which gives

$$\Delta m_{\mathrm{add}} \approx \frac{\alpha}{2}\,\mathrm{sech}^2(q/2) w^\top \hat{v}_\ell. \tag{60}$$

Erasing the coordinate subtracts the current projection on $\hat{v}_\ell$, while reflection reverses that projection. If $h = h_\perp + c\hat{v}_\ell$, then

$$h_{\mathrm{erase}} = h_\perp, \tag{61}$$

$$h_{\mathrm{reflect}} = h_\perp - c\hat{v}_\ell. \tag{62}$$

The corresponding log-odds changes are $-c\, w^\top \hat{v}_\ell$ and $-2c\, w^\top \hat{v}_\ell$. A random unit direction r in a d-dimensional residual space satisfies

$$\mathrm{E}_r\left[(w^\top r)^2\right] = \frac{\|w\|^2}{d}. \tag{63}$$

The matched random control therefore has a small expected projection on the task readout unless the task axis itself is uninformative. This contrast is the mechanistic prediction tested by the task-axis interventions.