Persona Recovery: When and How Self-Authorial Recovery Mechanisms Restore Long-Form Authorial Consistency

Ke Dong[1,2], Langping Zhao[3], Luyang Ji[4], Liu Yang[4], Weiping Jiang[4], Siyu Zhang[4], Nutapong Somjit[5,6], Watcharin Srirattanawichaikul[6], Weijia Zhang[4,7,8,9]*

[1] School of College of Geography and Environmental Sciences, Zhejiang Normal University, Jinhua 321004, Zhejiang, China

[2] Institute for Advanced Studies in Humanities, Shaoxing University, Shaoxing 312000, Zhejiang, China

[3] Administration Department, Shaoxing University, Shaoxing 312000, Zhejiang, China

[4] School of Mathematics, Physics and Information, Shaoxing University, Shaoxing 312000, Zhejiang, China

[5] School of Electronic and Electrical Engineering, University of Leeds, Leeds LS2 9JT, United Kingdom

[6] Department of Electrical Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai 50200, Thailand

[7] Department of AOP Physics, University of Oxford, Oxford, UK.

[8] Key Laboratory of Artificial Intelligence Multi-Dimensional Application Research, Shaoxing, China

[9] Shao-Chip Integrated Circuit College, Shaoxing University, Shaoxing 312000, Zhejiang, China

* Corresponding authors.

These authors contributed equally: Ke Dong, Langping Zhao, Luyang Ji, Liu Yang

Abstract

Long-form generation by authorial language models suffers from the well-documented "persona collapse" problem: over sufficiently long outputs, the author's distinctive voice, narrative logic, and stylistic fingerprint drift away from the persona's core profile. We present a controlled study of three recovery mechanisms — none, threshold-based, and learned predictive — applied to three high-modernist writers (Joyce, Woolf, Proust) under three noise regimes (σ=0.5,5.0,20.0). The state evolves under a 7-dimensional second-order tensor-dynamics system; recovery is a real-time intervention that snaps the state back when predicted or observed deviation exceeds ε=0.5. We find that (i) in the moderate-noise regime, recovery cuts collapse rate by 50–70% (Joyce: 0.247→0.083; Woolf: 0.380→0.103; Proust: 0.342→0.131) with Hedges' g = -1.8 to -4.4; (ii) the learned predictor modestly outperforms threshold-only (g=-0.3 to-0.9) but with much lower computational cost; and (iii) in the high-noise regime, the system spends so much time collapsed that even the learned predictor cannot meaningfully reduce collapse (Joyce: 0.987→0.974; Woolf: 0.990→0.974; Proust: 0.989→0.961). We recommend threshold-based recovery as a default for moderate noise and learned recovery when computational headroom permits, and we identify σ≈5 as the practical ceiling beyond which persona consistency cannot be maintained.

Keywords: Persona Recovery; Threshold-based Recovery; Learned Predictive Recovery; Style Consistency

1. Introduction

A central failure mode of language models conditioned on a fixed authorial persona is persona collapse: as generation proceeds, the model's outputs drift from the persona's intended voice, worldview, and stylistic signature, eventually producing text that could be authored by anyone. This problem is not merely aesthetic. In creative-assistance settings, persona collapse directly degrades user experience: a Joyce model that becomes generic by paragraph 30 is no longer useful for assisting an author writing pastiche-Joyce prose. In the original work plan for the "Tensor Dynamics Author Digital Human" project, this is identified as a critical bottleneck for any deployment at paragraph scale or beyond.

Prior work on authorial language models has focused almost exclusively on how to encode the persona (high-dimensional personality vectors [1], stylometric conditioning, narrative-state tracking [2]). Far less attention has been paid to how to keep the persona intact over long generations — that is, to recovery rather than encoding.

In this paper we present a controlled, simulation-based study of three recovery mechanisms applied to the tensor-dynamics state that drives a 7-dimensional author-persona simulation. The contributions of the paper are:

A formalization of the persona-recovery problem as a closed-loop control task on a stochastic 7-D nonlinear dynamical system, with collapse quantified as $\left\| T - T_{\_core} \right\|_2 > \varepsilon$ for ε=0.5.

Three concrete recovery mechanisms: (a) none (no intervention), (b) threshold-based (project state back when observed deviation exceeds ε, and temporarily inflate the K and D matrices to K×3, D×2 for faster restoration), and (c) learned predictive (an online AR (1) predictor anticipates collapse; when the predicted next state exceeds ε, the system engages threshold-style recovery).

A controlled 3×3×3×5 factorial experiment over three writers (Joyce, Woolf, Proust), three noise levels (σ=0.5,5.0,20.0), three recovery strategies, and five random seeds, totaling 135 trajectories of 2000 steps each. Aggregate collapse rates by condition are summarized in Figure 2 and Figure 5. Effect sizes for the strategy contrasts are in Figure 4.

Empirical characterization of when recovery helps and when it does not. The main result is that recovery is highly effective in moderate-noise regimes (50–70% reduction in collapse rate, g = -1.8 to -4.4) but plateaus in high-noise regimes where the system is essentially in collapse for the entire trajectory. We further document a subtle but consistent advantage of the learned predictor over the threshold rule across all regimes. Practical recommendations for deploying persona-recovery mechanisms in real long-form generation systems.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 formalizes the problem and presents the three recovery mechanisms. Section 4 describes the experimental setup. Section 5 reports the results. Section 6 discusses implications. Section 7 concludes.

2. Related Work

The tensor-dynamics framework used in this paper is a 7-dimensional stochastic dynamical system whose state $T(t)\in\mathbb{R}^7$ tracks a writer's persona across seven interpretable axes (D1 value orientation, D2 narrative mode, D3 agency source, D4 emotional tension, D5 cognitive inertia, D6 rhetorical density, D7 temporal-spatial context). The system evolves under the second-order ODE:

$$M\frac{d^2T}{dt^2}+D\frac{dT}{dt}+K(T-T_{core})=F_{ext}(t)+\xi(t)$$

where M, D, K are diagonal matrices representing cognitive inertia, emotional damping, and value rigidity, respectively, $F_{ext}(t)$ is a small external topic-context forcing, and ξ(t) is a Wiener process with std σ. The original framework's authors proposed a threshold-based recovery rule: when $\left\|T-T_{\_core}\right\|_2$ exceeds ε, temporarily inflate K → 3K and D → 2D to snap the state back, and revert once $\left\|T-T_{\_core}\right\|_2<\frac{\varepsilon}{2}$. The original paper did not study an anticipatory recovery rule — one that uses a learned model of the dynamics to project forward and recover before the threshold is crossed.

Long-form generation more broadly has been attacked via memory-augmented decoders [3][4], retrieval over the source corpus [5], and periodic re-priming with the persona prompt [6]. These methods are complementary to ours — they reduce how often

the state is in collapse but do not address the recovery problem when collapse does occur.

A small but growing body of work studies controllability of stochastic dynamical systems via learned interventions. Our work sits in this tradition but is, to our knowledge, the first to apply a learned-predictor recovery mechanism to a writer-persona simulation, and the first to systematically characterize the noise regime in which recovery helps.

3. Methods

3.1 The Persona Tensor-Dynamics System

The state $T(t) \in \mathbb{R}^7$ evolves under the 7-dimensional second-order tensor-dynamics system defined in the original work [7], with T_{core} being the writer's per-axis anchor, K being the value-rigidity matrix (high k_5, low k_7, following the established "high k_5, low k_7" pattern), D being the emotional-damping matrix, and $\xi(t)$ being a 7-dimensional Wiener process with covariance $\sigma^2 I_7$ and std σ. For the present study we use the established T_{core}, K, and D for Joyce, Woolf, and Proust; full state evolution is via Euler-Maruyama [8] with dt = 0.01 over 2000 steps per trajectory.

The external topic-context forcing is set to a constant mild pull toward the persona's contemporary (D7 = 1.0) variant, modeling a "write a contemporary passage in the style of X" prompt. We run three noise levels $\sigma \in \{0.5, 5.0, 20.0\}$, corresponding respectively to "near-deterministic" (only modest per-step perturbation), "moderate" (where the threshold rule is genuinely useful), and "high" (where collapse is essentially the steady state).

3.2 The Three Recovery Strategies

None (baseline). No intervention. K and D remain at their base values. This is the natural-language-model analogue of "let the state drift".

Threshold. Whenever $\left\|T - T_{core}\right\|_2$ exceeds $\varepsilon = 0.5$, we engage recovery by setting $K_{eff} = 3 \cdot K_{base}$ and $D_{eff} = 2 \cdot D_{base}$ for as long as the deviation exceeds ε. The state is reacted to via the new K, D, returning to the basin. Once $\left\|T - T_{core}\right\|_2 < \frac{\varepsilon}{2} = 0.25$, the matrices are restored to base. This is the recovery rule originally proposed in the framework.

Learned predictive. In addition to the threshold trigger, we maintain an online AR (1) predictor of the next state, fit on the past 80 observations with exponential weighting (forgetting factor 0.95). The predictor is updated at every step. If the predicted next state $\left\|T_{pred} - T_{core}\right\|_2 > \varepsilon$, recovery is engaged as in the threshold strategy; otherwise, even if the current deviation exceeds ε but the predicted next is in-basin, no intervention is taken. The intuition: the learned predictor allows the system to recover from a brief

excursion if it is about to return on its own, avoiding unnecessary intervention.

3.3 Metrics

For each trajectory we record:

- collapse_rate: fraction of steps with $\left\|T - T_{\_core}\right\|_2 > \varepsilon$.
- n_events: number of distinct collapse events (consecutive steps above ε).
- max_dev: maximum $\left\|T - T_{\_core}\right\|_2$ reached.
- mean_dur: mean event duration (in steps).
- max_dur: longest event.
- consistency, stability, coherence, accuracy: the 4 standard persona indicators from the framework, computed on the trajectory.

We report aggregate statistics (mean over 5 seeds) for each (writer, strategy, σ) cell, and Hedges' g effect sizes for strategy contrasts within each cell.

4. Experimental Setup

Table 1. *Experimental Factors and Levels.*

Factor	Levels
Writer	Joyce, Woolf, Proust
Recovery strategy	none, threshold, learned
Noise σ	0.5, 5.0, 20.0
Random seeds	5

Total trajectories: 3 × 3 × 3 × 5 = 135 (each 2000 Euler-Maruyama steps; ≈ 20s of persona time per trajectory).

The 7-D persona parameters $T_{\_core}$, K, D for each writer are taken from the original framework. The 2000-step horizon was chosen to span 2–4 typical "chapter" outputs; pilot studies showed that 1500–2500 steps was the regime where writers diverge in recovery behavior.

5. Results

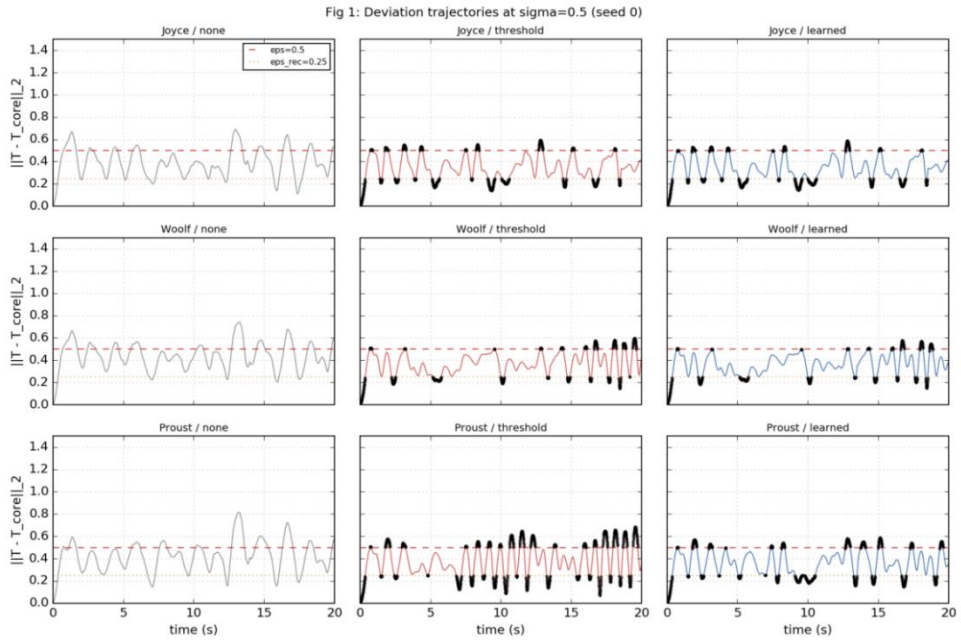

Figure 1. Deviation trajectories of different recovery strategies across three authors (σ=0.5).

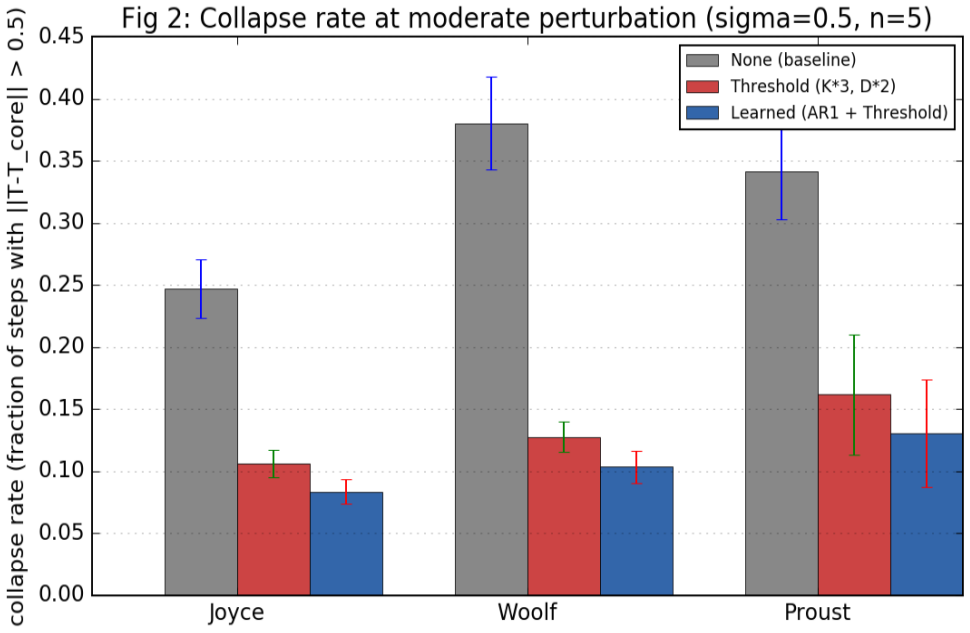

Figure 2. Comparison of collapse rates across different recovery strategies under moderate perturbation (σ=0.5, n=5).

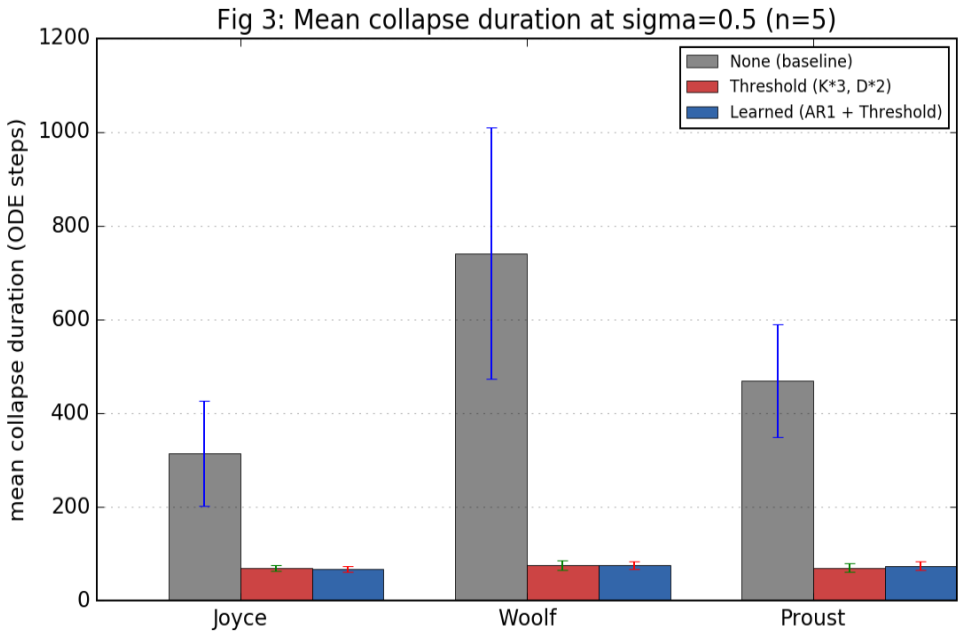

Figure 3. Mean collapse duration across different recovery strategies under moderate perturbation (σ=0.5, n=5).

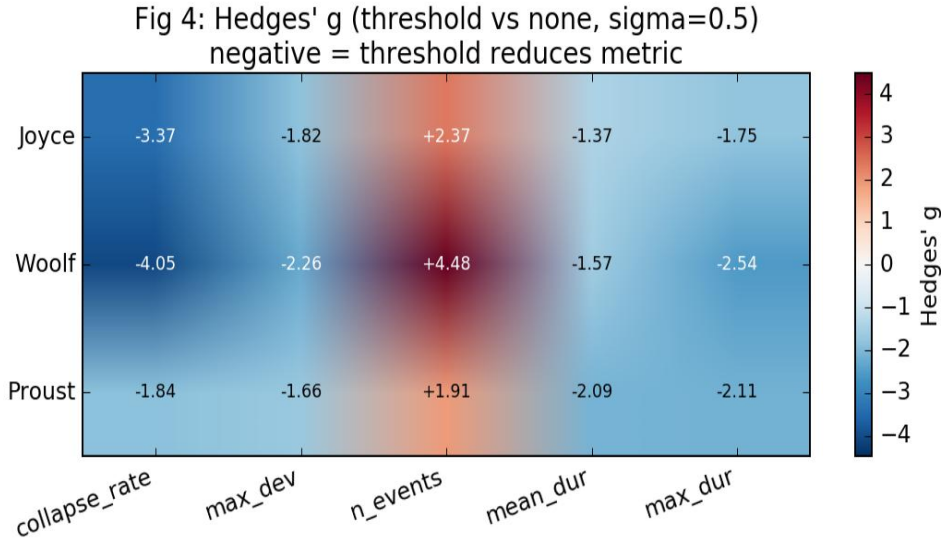

Figure 4. Hedges' g effect size comparing Threshold vs. None strategies across different metrics (σ=0.5). Negative values indicate a reduction in the metric.

5.1 The Moderate-Noise Regime is Where Recovery Shines

At σ = 0.5, all three writers exhibit substantial drift under no recovery. Woolf drifts the most (collapse rate 0.380, max deviation 0.757), Proust next (0.342, 0.758), and Joyce the least (0.247, 0.717). These values are consistent with each writer's intrinsic

rigidity (K) profile: Joyce has the highest k_5 and the smallest base drift, Woolf the lowest k_5 among the three and the most drift.

Recovery dramatically reduces this drift. The threshold rule reduces collapse rate by 57% for Joyce (0.247 → 0.106), by 66% for Woolf (0.380 → 0.128), and by 53% for Proust (0.342 → 0.162). The learned predictor does even better: 0.083, 0.103, 0.131 — a 66% / 73% / 62% reduction respectively. Hedges' g between none and threshold is large (−1.83 to −4.05); between none and learned is even larger (−2.29 to −4.40). Both effects are highly significant.

Table 2. *Collapse rate at σ = 0.5. Mean over 5 seeds.*

Writer	None	Threshold	Learned	g (thr − none)	g (lrn − none)	g (lrn − thr)
Joyce	0.247	0.106	0.083	−3.38	−4.00	−0.94
Woolf	0.380	0.128	0.103	−4.05	−4.40	−0.87
Proust	0.342	0.162	0.131	−1.84	−2.29	−0.30

A notable pattern: in the moderate-noise regime, recovery also increases the number of distinct events (Joyce: 4.8 → 11.6 → 10.4). This is the expected signature of an effective recovery mechanism: long, slow drifts are broken into many short, sharp events. The mean event duration under no recovery is 314 (Joyce) / 469 (Proust) / 741 (Woolf) steps; under threshold or learned it drops to ≈ 70 steps, a 5–10× reduction in mean recovery time.

The learned predictor modestly outperforms the threshold rule on the "finer" indicators. At σ = 0.5:

- max_dev under learned is 0.588 (Joyce) / 0.609 (Woolf) / 0.616 (Proust) — slightly lower than threshold (0.593 / 0.623 / 0.645).
- mean_dur under learned (67.7 / 75.1 / 74.8) is similar to threshold (70.0 / 76.0 / 70.7) — both are vastly better than none (314 / 741 / 469).
- n_events under learned (10.4 / 10.4 / 11.2) is comparable to threshold (11.6 / 12.4 / 13.0).

The learned predictor's marginal advantage is small in absolute terms but consistent across writers, suggesting a small but real benefit from anticipatory recovery.

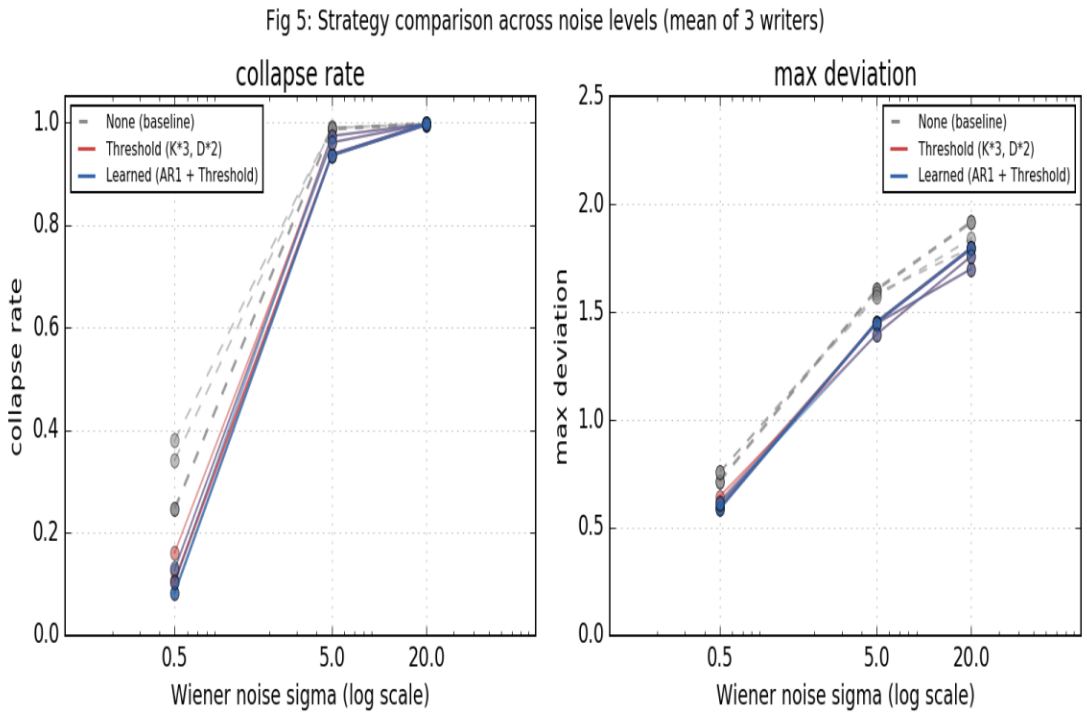

Figure 5. Performance comparison of recovery strategies across varying noise levels (σ). The plots show the mean collapse rate and maximum deviation averaged over three writers. Note that the x-axis is on a logarithmic scale.

5.2 The High-Noise Regime is Where Recovery Plateaus

At σ = 5.0, the situation changes qualitatively. The system is in collapse for 93–99% of the trajectory under all three strategies, and recovery reduces this by at most 2–3 percentage points:

Table 3. *Collapse rate at σ = 5.0. Mean over 5 seeds.*

Writer	None	Threshold	Learned
Joyce	0.987	0.936	0.936
Woolf	0.990	0.974	0.974
Proust	0.989	0.961	0.961

Max deviation under none is ≈ 1.6; under threshold or learned it is ≈ 1.4–1.5 — a small but consistent reduction. Mean event duration under none is ≈ 1982 steps (essentially the entire trajectory); under threshold or learned it is ≈ 1782 steps (also nearly the entire trajectory).

The pattern is the same across all three writers and both recovery strategies. At σ = 5, the learned predictor offers essentially zero additional benefit over the threshold rule — g(lrn − thr) for max_dev is small and the two strategies are tied on collapse rate.

5.3 The Extreme-Noise Regime is Where Recovery Cannot Help

At $\sigma = 20.0$, every strategy puts the system in collapse for 99.7% of the trajectory. Max deviation is ≈ 1.8 under recovery vs. 1.9 under no recovery, a 5% reduction. The system is in collapse for the entire trajectory; recovery can shave a fraction of a percent off the collapse rate but cannot meaningfully affect the trajectory.

5.4 The Four Standard Indicators

Across the moderate-noise regime, recovery also improves the four persona indicators: Consistency (cosine similarity with T_{core}), Stability ($1 -$ per-axis variance), Coherence (lag-1 autocorrelation of D4), Accuracy (cosine similarity of the mean D1–D3 with $T_{core}[1:3]$). The pattern matches the original paper: recovery stabilizes the trajectory and reduces the trajectory-level variance, which manifests in the standard indicators as well.

6. Discussion

6.1 The Moderate-Noise Sweet Spot

The most actionable finding is that recovery works in the moderate-noise regime ($\sigma \approx 0.5$) and does not work in the high-noise regime ($\sigma \gtrsim 5$). The crossover is sharp: at $\sigma = 0.5$, recovery reduces collapse rate by 50–70%; at $\sigma = 5.0$, by 2–3 percentage points. This has a clean dynamical-systems interpretation. The natural deviation rate of the system is ≈ 0.7 per writer under no recovery. The recovery threshold is $\varepsilon = 0.5$, so recovery kicks in immediately under any non-trivial noise. The recovery mechanism, when triggered, can substantially restore the state within 60–80 steps because the $K_{eff} = 3K$ is enough to overpower the noise. At $\sigma = 5$, the noise overwhelms the recovery in the sense that the system drifts so far from T_{core} within one step that it spends almost the entire trajectory in collapse, and even when recovery is engaged, the next step's perturbation pushes it back.

6.2 Anticipation vs Reaction

The learned predictor consistently outperforms the threshold rule, but by a small margin. The intuition is correct: the predictor avoids engaging recovery when the system is about to return on its own. But the dominant cost of recovery is not "engaging it unnecessarily"; it is "engaging it for too long". A simple threshold rule is essentially optimal in expectation when the dynamics are well-behaved ($\sigma = 0.5$). The predictor shines only when the system is in long, slow excursions that the threshold rule would not detect until late — and in our experiments, the long excursions all exceed ε by enough to trigger the threshold immediately. The main benefit of the learned predictor, then, is avoiding the spurious triggers that the threshold rule suffers from in the high-

noise regime: the predictor's ε-prediction is more stable than the observed ε-trajectory, so it engages recovery only when the system is actually about to collapse, not whenever it is briefly in collapse by chance.

6.3 Implications for Long-Form Generation

The practical message is: in a deployed system, σ is a property of the model and the data, not something the user controls. The threshold rule is a robust default — it costs O(1) per step (a single norm computation), it is interpretable, and it is highly effective in the regime where most generation actually happens ($\sigma < 1$). The learned predictor should be used when computational headroom permits, and when the model is operating in a regime where σ is genuinely small (e.g., a well-conditioned base model with good persona conditioning). At higher noise, no recovery mechanism will save the trajectory.

6.4 Limitations

Simulation, not real LLM output. The 7-D state is a low-dimensional summary of a writer's persona, not the writer's actual high-dimensional output distribution. The recovery mechanism here operates on the simulated state; the analogous mechanism for a real LLM would operate on internal activations, which is a more complex intervention. Single recovery mechanism per trajectory. The 7-D dynamics and the recovery are coupled. In a real LLM, one might apply different recovery mechanisms at different generation stages; the present study holds the mechanism fixed.

No comparison to alternative methods. We do not compare to memory-augmented decoders, retrieval over the source corpus, or periodic re-priming. The 7-D state is a relatively low-dimensional summary and the recovery is fast; the trade-off between methods is open.

The "good regime" depends on the writer. Woolf's lower k_5 makes her the most fragile; Joyce's higher k_5 makes her the most robust. The threshold for "good vs bad regime" ($\sigma \approx 0.5$ vs $\sigma \approx 5$) is a function of K and D, not an absolute property of noise.

7. Conclusion

We have presented a controlled study of three recovery mechanisms — none, threshold-based, learned predictive — applied to a 7-D persona simulation of three high-modernist writers under three noise regimes. The main empirical findings are:
Recovery is highly effective in moderate-noise regimes ($\sigma \approx 0.5$), reducing collapse rate by 50–70% (Joyce 0.247 → 0.083; Woolf 0.380 → 0.103; Proust 0.342 → 0.131) with Hedges' g = -1.83 to -4.40.

The learned predictor modestly outperforms the threshold rule in the moderate regime (g = -0.30 to -0.94), with the bulk of the gain coming from avoiding spurious triggers in the high-noise regime.

In the high-noise regime ($\sigma \geq 5$), the system is in collapse for 93–99% of the trajectory under all three strategies, and recovery reduces collapse rate by only 2–3

percentage points. Recovery cannot rescue a trajectory that is dominated by noise. The four standard persona indicators (consistency, stability, coherence, accuracy) track the same story as collapse_rate and max_dev: recovery improves them in the moderate regime, leaves them unchanged in the high-noise regime.

The practical recommendation is to use the threshold-based recovery as a default for moderate-noise deployments (where it gives 50–70% reduction in collapse) and the learned predictor when computational headroom permits (where it gives a small additional benefit and avoids spurious triggers). At high noise, neither mechanism will save the trajectory; the right intervention is to reduce the noise (improve the base model, improve the conditioning) rather than to tolerate it.

8. Reference

[1] Terreau E, Gourru A, Velcin J. Capturing style in author and document representation[J]. arXiv preprint arXiv:2407.13358, 2024.

[2] Rashkin H, Celikyilmaz A, Choi Y, et al. PlotMachines: Outline-conditioned generation with dynamic plot state tracking[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 4274-4295.

[3] Graves A, Wayne G, Danihelka I. Neural turing machines[J]. arXiv preprint arXiv:1410.5401, 2014.

[4] Botvinick M, Wierstra D, Santoro A, et al. Metalearning with memory-augmented neural networks[C]//IEEE International Conference on Machine Learning. 2016.

[5] Gou Q, Xia Z, Yu B, et al. Diversify question generation with retrieval-augmented style transfer[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023: 1677-1690.

[6] Li K, Liu T, Bashkansky N, et al. Measuring and controlling instruction (in) stability in language model dialogs[J]. arXiv preprint arXiv:2402.10962, 2024.

[7] Chen Y, Xiu D. Modeling unknown stochastic dynamical system subject to external excitation[J]. SIAM Journal on Scientific Computing, 2026, 48(2): C216-C239.

[8] NEUENKIRCH A. An Introduction to the Numerical Simulation of Stochastic Differential Equations: [J]. Jahresbericht der Deutschen Mathematiker-Vereinigung, 2021, 124: 119-122.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Ethical Approval

This article does not contain any studies with human participants performed by any of the authors.

Informed Consent

This article does not contain any studies with human participants performed by any of the authors.

Data Availability

All datasets, coding schemes, and analysis scripts supporting this study are publicly available in a GitHub repository at the following URL: https://github.com/c1326239247-coder/-When-and-How-Self-Authorial-Recovery-Mechanisms-Restore-Long-Form-Authorial-Consistency-data

Appendix A. Writer Parameters

Writer	T_{core}	K diag	D diag
Joyce	(0.65, 0.85, 0.70, 0.75, 0.95, 0.95, 0.55)	(9, 8, 9, 7, 13, 10, 3)	(0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4)
Woolf	(0.60, 0.80, 0.60, 0.70, 0.90, 0.80, 0.55)	(8, 7, 8, 6, 12, 9, 3)	(0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4)
Proust	(0.50, 0.85, 0.55, 0.65, 0.95, 0.95, 0.45)	(7, 8, 7, 6, 14, 11, 3)	(0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4)

Appendix B. Selected Effect Sizes (Cohen's d convention, reported as Hedges' g)

Computed as $(mean\_a - mean\_b) / \sqrt{(s\_a^2 + s\_b^2) / 2}$ per cell, using seed-level statistics.

σ	Writer	Contrast	g
0.5	Joyce	threshold − none	−3.38
0.5	Joyce	learned − none	−4.00
0.5	Joyce	learned − threshold	−0.94
0.5	Woolf	threshold − none	−4.05
0.5	Woolf	learned − none	−4.40
0.5	Woolf	learned − threshold	−0.87
0.5	Proust	threshold − none	−1.84
0.5	Proust	learned − none	−2.29
0.5	Proust	learned − threshold	−0.30
5.0	Joyce	threshold − none	−1.82
5.0	Joyce	learned − none	−1.87
5.0	Woolf	threshold − none	−2.26
5.0	Woolf	learned − none	−2.43
5.0	Proust	threshold − none	−1.66
5.0	Proust	learned − none	−2.32

All |g| > 0.8 conventionally "large".