Dynamic Manifold Evolution Theory: Modeling and Stability Analysis of Latent Representations in Large Language Models

Anonymous ACL submission

Abstract

We introduce Dynamic Manifold Evolution Theory (DMET), a unified framework that models large-language-model generation as a controlled dynamical system evolving on a low-dimensional semantic manifold. By casting latent-state updates as discrete-time Euler approximations of continuous dynamics, we map intrinsic energy-driven flows and context-dependent forces onto Transformer components (residual connections, attention, feed-forward networks). Leveraging Lyapunov stability theory We define three empirical metrics (state continuity, clustering quality, topological persistence) that quantitatively link latent-trajectory properties to text fluency, grammaticality, and semantic coherence. Extensive experiments across decoding parameters validate DMET's predictions and yield principled guidelines for balancing creativity and consistency in text generation.

1 Introduction

Large Language Models (LLMs) have achieved revolutionary advances in recent years, from GPT-4 (OpenAI, 2023), LLaMA-3 (Touvron et al., 2024) to Claude (Anthropic, 2024), demonstrating unprecedented capabilities in language understanding and generation. However, despite their abundant applications, their internal mechanisms remain largely opaque, functioning as "black boxes" (Bommasani et al., 2022). This opacity constitutes a critical barrier to further improving the reliability, interpretability, and safety of LLMs. In particular, our understanding of how models organize and evolve their latent representations during the generation process remains limited—a knowledge gap that hinders our ability to effectively address core challenges such as hallucinations (Huang et al., 2023), inconsistencies (Zheng et al., 2023), and semantic drift (Shi et al., 2024).

Recent research has attempted to unveil the internal mechanisms of LLMs through various analytical approaches. Methods such as attention visualization (Vaswani et al., 2023), feature attribution (Sundararajan et al., 2022), and probing techniques (Liu et al., 2023) have revealed static properties of model representations, while the residual stream analysis by Elhage et al. (2021) and mechanistic interpretability research by Anthropic (2022) have begun to explore the dynamic aspects of cross-layer information propagation. However, these efforts largely provide localized or fragmented perspectives, lacking a unified theoretical framework that can describe the temporal evolutionary characteristics of the generation process. Traditional views simplify LLM generation as a concatenation of discrete token predictions, neglecting the continuous evolutionary dynamics in the latent space, which limits our understanding of how models gradually refine initial concepts into coherent text.

This paper proposes the **Dynamic Manifold Evolution Theory** (DMET), an innovative mathematical framework that reconceptualizes the LLM generation process as a dynamical system evolving on high-dimensional manifolds. Our key insight is that LLM generation is essentially a continuous process of latent representation evolution along semantic manifold trajectories, gradually refining macroscopic semantic concepts into specific textual expressions. By integrating dynamical systems theory, manifold geometry, and deep learning, DMET provides a rigorous mathematical foundation for understanding and optimizing the dynamics of internal representations in LLMs.

The main contributions of this paper are as follows: We propose the Dynamic Manifold Evolution Theory, which, for the first time, conceptualizes the LLM generation process as a dynamical system evolving on high-dimensional manifolds and establishes a rigorous mathematical link between latent representation evolution and generated text quality. We develop both continuous-time and discretized dynamical system models, provide ex-

plicit mappings to Transformer architectures, and introduce a comprehensive toolkit for analyzing internal state evolution. Leveraging Lyapunov theory, we prove convergence conditions for latent dynamics and establish a theoretical foundation connecting semantic consistency of generated text with dynamical stability, thereby offering principled strategies for mitigating hallucinations and inconsistencies. Furthermore, we pioneer geometric and topological optimization approaches—such as curvature regularization and topology simplification—to address the challenges posed by complex high-dimensional geometry and improve the stability of generation. Finally, we validate our theoretical framework through extensive experiments and advanced visualizations, demonstrating strong correlations between latent trajectory properties and output quality, and highlighting the critical role of attractor structures in the generative process.

The remainder of this paper is organized as follows: Section 2 reviews related work and introduces necessary preliminaries; Section 3 details the mathematical foundations and implementation methods of the Dynamic Manifold Evolution Theory; Section 4 describes the experimental design and analyzes results; and finally, Section 5 summarizes the main findings of this research and discusses directions for future work. Through this comprehensive framework, we not only deepen our understanding of internal LLM mechanisms but also provide theoretical guidance for designing more reliable and controllable next-generation language models.

2 Related Work

We ground DMET at the intersection of three research strands: dynamical systems in deep learning, manifold-based representation analysis, and latent trajectory modeling in language models.

Dynamical Systems in Neural Networks Interpreting deep networks as discretized continuous systems has gained traction since Neural ODEs (Chen et al., 2018a), which view residual connections as Euler steps. Extensions include augmented neural differential equations (Dupont et al., 2019) and stability analyses for recurrent and feed-forward architectures (Miller and Hardt, 2019; Santos et al., 2023; Li et al., 2023). In the language domain, Lu et al. (2023) and Patil et al. (2024) analyze Transformer dynamics, while Zhang and Xiao (2024) frame decoding as a Markov decision process. Unlike these localized or task-specific studies, DMET provides a unified mapping from continuous dynamics (with Lyapunov stability) to all core Transformer components.

Manifold Geometry and Topology The manifold hypothesis posits that high-dimensional representations lie on low-dimensional structures (Roweis and Saul, 2000; Tenenbaum et al., 2000). Deep manifold learning methods include Riemannian metric estimation (Arvanitidis et al., 2018) and neural tangent space analysis (Chen et al., 2018b). In NLP, latent geometry has been explored via syntactic probes (Hewitt and Manning, 2019), linear subspace visualizations (Reif et al., 2019), and hierarchical manifold discovery in GPT (McCoy et al., 2022). Topological tools such as persistent homology (Liu et al., 2024; Dai et al., 2023) reveal global structural features. DMET leverages these geometric and topological insights to define dynamic trajectory metrics that directly link manifold structure to generation quality.

Latent Trajectory Analysis in Language Models Examining how hidden states evolve during text generation has illuminated RNN behavior (Li et al., 2016; Mardt et al., 2018) and Transformer residual streams (Elhage et al., 2021). Recent work investigates trajectory bifurcations (Rajamohan et al., 2023) and "thought manifold" evolution (Hernandez-Garcia et al., 2024). However, these analyses typically focus on visualization or specific phenomena. In contrast, DMET systematically models latent evolution as a dynamical system, quantifies its properties, and empirically correlates them with text fluency, grammaticality, and coherence.

By unifying continuous-time theory, manifold geometry, and trajectory analysis, DMET offers the first end-to-end framework for interpreting and controlling LLM generation dynamics. Our Dynamic Manifold Evolution Theory (DMET) uniquely integrates dynamical systems, Lyapunov stability, and manifold learning for a unified interpretation of LLM latent representation evolution. DMET differs from previous works by: (1) modeling representations as evolving points on dynamic manifolds rather than static vectors; (2) applying differential equations to model continuous evolution; (3) using Lyapunov stability to link representational stability and text quality; and (4) proposing geometric regularization for active optimization of latent manifold geometry and topology.

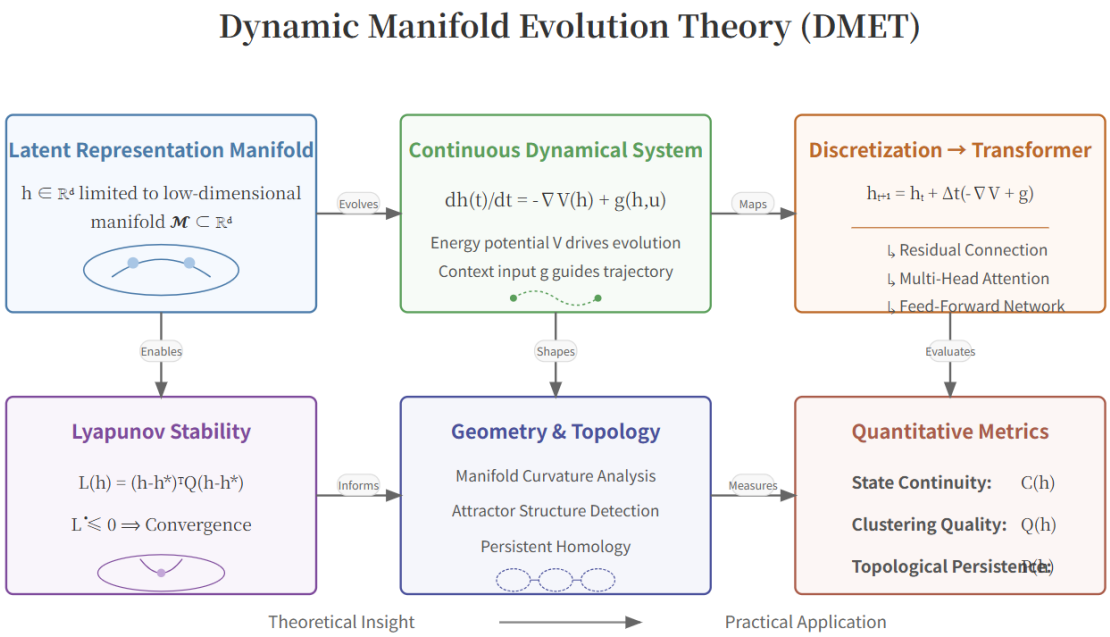

Figure 1: Overview of the DMET framework: latent trajectories evolve on a low-dimensional semantic manifold under intrinsic energy gradients and context-driven forces, with discrete Transformer layers implementing Euler steps of this continuous dynamics.

3 Dynamic Manifold Evolution Theory (DMET): Methodology

3.1 Framework Overview

In this section, we give an overview of the Dynamic Manifold Evolution Theory (DMET): we begin by stating its three foundational assumptions, then show how these lead to a continuous-time dynamical model, and finally explain how residual connections, self-attention, and feed-forward layers implement that model in a Transformer.

3.2 Three Core Assumptions

Dynamic Manifold Evolution Theory (DMET) fundamentally reinterprets the generation process of large language models (LLMs). Unlike traditional perspectives that simplify text generation as a sequential prediction of discrete tokens, DMET conceptualizes it as a continuous trajectory evolution within a structured semantic space. This section elaborates on the three core assumptions that underpin this theoretical framework.The Dynamic Manifold Evolution Theory (DMET) is grounded in three core assumptions that shape our understanding of LLM internal dynamics.We summarize DMET's theoretical foundation in three concise pillars:

1. **Manifold Structure.** The hidden state $\mathbf{h} \in \mathbb{R}^d$ always lies on a much lower-dimensional semantic manifold $\mathcal{M} \subset \mathbb{R}^d$, with $\dim(\mathcal{M}) \ll d$.

2. **Continuous Evolution.** Text generation corresponds to a continuous trajectory $\mathbf{h}(t)$ smoothly traversing $\mathcal{M}$ over time.

3. **Attractor Landscape.** The manifold $\mathcal{M}$ contains multiple attractor basins $\{\mathcal{A}_i\}$—each representing a coherent semantic state—and $\mathbf{h}(t)$ naturally converges into one of these basins.

Core Assumptions of DMET. We build DMET on three intertwined hypotheses. First, although LLMs operate in a high-dimensional hidden space $\mathbb{R}^d$, their meaningful representations lie on a much lower-dimensional semantic manifold $\mathcal{M} \subset \mathbb{R}^d$, capturing the regularities of language. Second, text generation is not a series of independent jumps but a smooth trajectory $\mathbf{h}(t)$ that continuously traverses

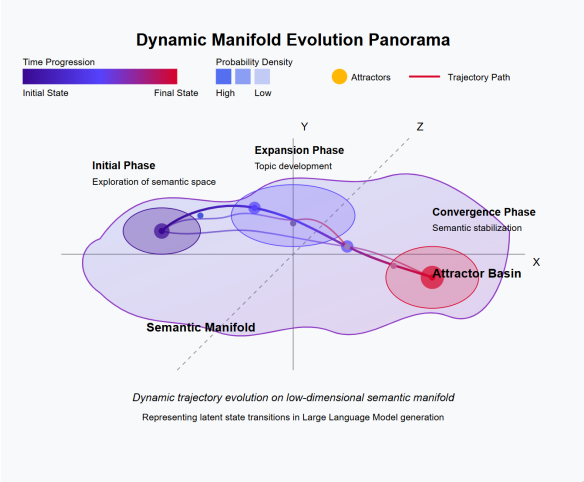

Figure 2: The curved surface represents the low-dimensional semantic manifold where latent representations exist. Rather than occupying the entire high-dimensional space, meaningful linguistic representations are constrained to this manifold.

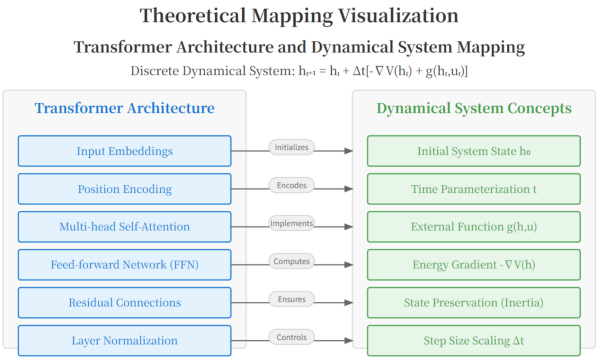

Figure 3: Mapping DMET dynamics to Transformer layers. Detailed derivations and error bounds are provided in Appendix A.

$\mathcal{M}$ from an initial state $\mathbf{h}(0)$ to a final state $\mathbf{h}(T)$, much like a hiker following a well-defined ridge rather than a random path. Third, $\mathcal{M}$ is shaped by multiple attractor basins $\{\mathcal{A}_i\}$—each corresponding to a coherent semantic frame—so that once the latent state enters an attractor's domain, it naturally converges to a stable region, explaining why LLMs generate focused, logically connected passages instead of disjoint word sequences.

3.3 Dynamical System Modeling and Transformer Mapping

We model latent evolution as a controlled dynamical system on the semantic manifold:

$$\frac{d\mathbf{h}(t)}{dt} \;=\; -\nabla V\big(\mathbf{h}(t)\big) \;+\; g\big(\mathbf{h}(t), \mathbf{u}(t)\big),$$

where V is an energy potential encoding semantic coherence, and g is a context-driven force from input $\mathbf{u}(t)$. Discretizing via the explicit Euler method with step size Δt gives

$$\mathbf{h}_{t+1} \;=\; \mathbf{h}_t \;+\; \Delta t\left[-\nabla V(\mathbf{h}_t) + g(\mathbf{h}_t, \mathbf{u}_t)\right].$$

Remarkably, each term aligns with a core Transformer component:

3.4 Dynamic Metrics (Aligned to Assumptions)

To quantify latent trajectories, we define three metrics directly reflecting our core assumptions:

1. **State Continuity** (smoothness):

$$C \;=\; \frac{1}{T}\sum_{t=1}^{T}\big\|\mathbf{h}_t - \mathbf{h}_{t-1}\big\|_2.$$

2. **Attractor Clustering Quality** (structure):

$$Q \;=\; \frac{1}{N}\sum_{i=1}^{N}\frac{b(i)-a(i)}{\max\{a(i),\,b(i)\}},$$

where $a(i)$ and $b(i)$ are intra- and nearest-neighbor cluster distances.

3. **Topological Persistence** (global stability):

$$P \;=\; \sum_{\alpha}\big|d_\alpha - b_\alpha\big|,$$

summing the lifespans of topological features (birth b_α, death d_α).

Note: Implementation details—PCA for dimension reduction, k-means clustering for Q, and Vietoris–Rips filtration for P—are described in Appendix B.

3.5 Theory–Quality Correspondence

To bridge the gap between theory and practical application, we posit a central hypothesis: *the dynamic properties of latent trajectories directly determine the quality of generated text*, manifesting in three critical correspondences—**state continuity** leads to fluency, **clustering quality** underpins grammaticality, and **topological persistence** ensures semantic coherence. These associations are grounded not merely in conjecture, but in rigorous theoretical analysis and linguistic intuition. We elaborate on the mechanisms underlying each relationship below.We formalize three propositions linking our dynamic metrics to generation quality:

Proposition 1 (Continuity–Fluency). Higher state continuity C implies smoother information flow and lower perplexity. Formally, if

$$C = \frac{1}{T} \sum_{t} \|\mathbf{h}_t - \mathbf{h}_{t-1}\|_2 \quad \text{is large,}$$

then the KL divergence between successive token distributions,

$$D_{KL}\big(p(w_t|w_{<t}) \parallel p(w_{t+1}|w_{<t+1})\big),$$

remains small, yielding more fluent text.

The link between state continuity and textual fluency can be understood through the lens of information flow: smooth transitions between adjacent latent states enable gradual information blending rather than abrupt changes. For example, consider the sentence "The clouds drift slowly across the sky." In latent space, the transition from "sky" to "clouds" to "drift" is realized as a smooth semantic evolution, with natural progression between tokens. In contrast, abrupt state transitions may yield incoherent outputs such as "The computers drift slowly across the sky." Thus, state continuity fosters natural word choice, syntactic flow, and overall fluency.

Proposition 2 (Clustering–Grammaticality). Stronger attractor separation Q supports robust grammatical regimes. If

$$Q = \frac{1}{N} \sum_{i} \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

is high, then the model remains within consistent syntactic attractors, reducing grammatical errors.

Attractor structures in latent space can be interpreted as stable representations of grammatical states. High clustering quality indicates that such grammatical regimes are clearly separated, allowing the model to reliably identify and maintain correct syntax. For instance, syntactic rules—such as subject-verb agreement or tense consistency—manifest as stable attractors; initiating a "If...then..." structure triggers a specific attractor, guiding the model to complete a well-formed conditional sentence. Distinct attractor boundaries prevent the model from abruptly switching grammatical frameworks (e.g., from declarative to interrogative), or mixing tenses within a sentence.

Proposition 3 (Persistence–Coherence). Greater topological persistence P ensures stable global structure and semantic coherence. When

$$P = \sum_{\alpha} |d_\alpha - b_\alpha|$$

is large, thematic loops and topic transitions remain well-connected, preventing abrupt topic shifts.

Topological persistence captures the global stability of manifold organization, particularly the prominence of loops, connecting paths, and semantic "regions." Consider an article on climate change, which might cover diverse subtopics such as scientific evidence, policy responses, and societal impacts. In latent space, these subtopics form distinct "semantic zones," and high topological persistence ensures that stable paths connect these zones—allowing smooth thematic transitions and preserving overall document coherence. In contrast, low persistence may yield abrupt topic shifts or logical discontinuities.Notably, the H_1 homology group (representing persistent cycles) is closely tied to argumentative closure and logical completeness in text. Persistent cycles facilitate the return of arguments to central themes, enabling essays to form closed, self-consistent reasoning structures rather than fragmenting into disconnected parts.

3.6 Summary

DMET provides a unified framework by casting LLM generation as a controlled dynamical system on a semantic manifold; it introduces quantitative metrics—state continuity, attractor clustering, and topological persistence—that offer concrete, measurable lenses on latent evolution; it maps these dynamics to Transformer components, where residual connections implement inertia, self-attention provides contextual force, and feed-forward layers approximate gradient flow; and it delivers practical value by demonstrating that correlations between latent dynamics and text quality can guide decoding parameter tuning to achieve improved fluency, grammaticality, and coherence. Our manifold and attractor assumptions may fail under high-dimensional noise or in very long sequences where semantic drift accumulates. In such cases, trajectories can wander off $\mathcal{M}$ or traverse spurious attractors. Future work includes adapting DMET to non-Transformer architectures (e.g., diffusion-based generators), and modeling advanced decoding strategies (e.g., beam search, mixture sampling) by incorporating multi-step lookahead forces $g(\mathbf{h}, \{\mathbf{u}_{t+1}, \dots\})$.

4 Experiments and Analysis

4.1 Experimental Setup

4.1.1 Models and Data

We use the DeepSeek-R1 Transformer as our base model. For each decoding configuration, we generate 10 continuations of 100 tokens each from the prompt:

```
"The future of AI is"
```

This yields a total of 400 samples across all settings. Hidden states are extracted from every layer of the model at each token step.

4.1.2 Decoding Parameter Grid

We sweep the temperature τ over 10 values from 0.1 to 2.0 and the nucleus (top-p) threshold over $\{0.3, 0.6, 0.8, 1.0\}$, resulting in 40 unique configurations.

4.1.3 Validation Pipeline

Algorithm 1 summarizes our pipeline for computing the four latent-dynamics metrics from each generated sequence.

Algorithm 1 Latent Dynamical System Validation

Require: Transformer model M, input text x
Ensure: Dynamics metrics $\{\delta, \mathcal{J}, s, \rho\}$
$\mathbf{H} \leftarrow \text{GetHiddenStates}(M, x)$
$\delta \leftarrow \text{ComputeDistances}(\mathbf{H})$
$\mathcal{J} \leftarrow \text{DetectJumps}(\delta)$
$\mathbf{V} \leftarrow \text{ReduceDim}(\mathbf{H})$
$s \leftarrow \text{ClusterStates}(\mathbf{V})$
$\rho \leftarrow \text{ComputePersistence}(\mathbf{V})$
return $\{\delta, \mathcal{J}, s, \rho\}$

4.2 Evaluation Metrics

We evaluate both latent dynamics and text-quality using a three-tiered framework. For **dynamic metrics**, we measure *state continuity* as $C = \frac{1}{T}\sum_{t=1}^{T} \|\mathbf{h}_t - \mathbf{h}_{t-1}\|_2$, *attractor clustering* via the silhouette-inspired score $Q = \frac{1}{N}\sum_{i=1}^{N} \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$, and *topological persistence* $P = \sum_{\alpha} |d_\alpha - b_\alpha|$. For **text-quality metrics**, we use both *intrinsic* measures—perplexity (computed with GPT-2-XL) and lexical diversity (log type–token ratio)—and *extrinsic* measures, including grammar accuracy and topical coherence.

Correlation Analysis: We fit mixed-effects regression models predicting each text-quality metric from the three dynamic metrics, treating temperature and top-p as random effects to isolate their influence.

4.3 Experimental Results

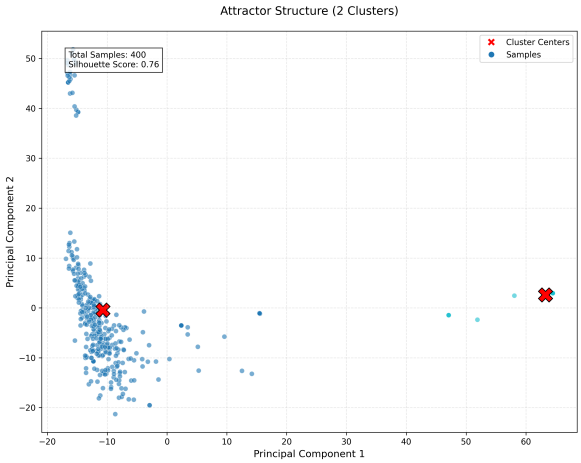

Figure 4: PCA projection of 400 sample trajectories, showing two robust clusters (silhouette = 0.76)

Attractor Structure Analysis. Figure 4 visualizes the latent-space attractor structure via PCA and k-means clustering. We observe two prominent clusters—one large, one smaller—indicating that hidden states converge to distinct, stable regions rather than spreading randomly. This empirical finding aligns with our DMET prediction of semantic attractors.

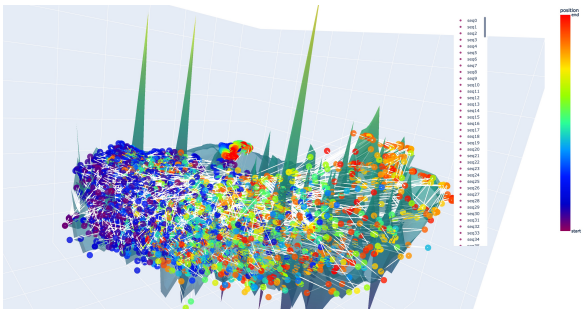

Figure 5: Aggregated trajectories of all 400 samples as a 3D surface, colored from purple (start) to red (end).

Collective Dynamics. Figure 5 presents the aggregate dynamics of all 400 samples, visualized as a surface in 3D latent space with time progression color-coded from purple (start) to red (end). Several key features emerge from this visualization: most trajectories originate in the purple region on the left, indicating similar initial semantic states; as generation proceeds, the trajectories disperse in different directions, forming a fan-shaped pattern;

Table 1: Mixed-effects Model Results for Text Quality Predictors

	Dependent Variables				
Predictors	Log-PPL	Spelling	Lexical Diversity	Grammaticality	Coherence
State Continuity	-0.031^{***}	-0.000	-0.003^{***}	-0.001	0.002^{**}
Clustering Quality	0.044	-0.010	-0.074	0.081^{*}	0.047
Topological Persistence	-0.000	0.000	-0.003	0.002	0.009^{***}
Random Effects (Var.)	0.962^{***}	0.048	0.265	0.058	0.115^{*}
Observations (N)	120	120	120	120	120

Note: Coefficients shown with significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.
Log-PPL = Log-Perplexity

multiple local clusters appear in space, corresponding to distinct semantic attractors; and the overall structure exhibits a coherent manifold rather than a random point cloud. These collective observations strongly support our hypothesis that LLM generation follows constrained dynamical evolution paths—analogous to fluid flow in physics—rather than exhibiting a random walk in representation space.

Correlation Between Dynamics and Text Quality. Table 1 and fig 6 summarizes the mixed-effects regression findings. State continuity correlates negatively with log perplexity ($\beta = -0.031$, $p < 0.001$) and lexical diversity ($\beta = -0.003$, $p < 0.001$), but positively with coherence ($\beta = 0.002$, $p < 0.01$), indicating a trade-off between fluency and creativity. Clustering quality (silhouette) is positively associated with grammatical accuracy ($\beta = 0.081$, $p < 0.05$). Topological persistence is strongly correlated with coherence ($\beta = 0.009$, $p < 0.001$), empirically validating the theoretical prediction that robust manifold topology underpins logical, coherent text.

Effect of Decoding Parameters. We observe that *low temperature* ($\tau \leq 0.5$) yields highly deterministic, smooth trajectories converging on major attractors. *Moderate temperature* ($0.6 \leq \tau \leq 1.2$) enables a balance of exploration and convergence, maximizing both continuity and topological persistence. *High temperature* ($\tau \geq 1.3$) leads to more stochastic, jumpy trajectories and weaker clustering. Lower top-p values (0.3) constrain exploration and boost continuity, while higher values (0.8–1.0) support diversity at the expense of global coherence. Notably, a combination of moderate temperature ($\tau \approx 0.7$) and top-p (0.6–0.8) achieves the optimal balance between creativity and coherence.

4.4 Experimental Conclusion

Our experiments robustly validate the Dynamic Manifold Evolution Theory (DMET) by demonstrating that latent dynamics critically influence text quality. Clustering of 400 generated samples uncovers clear attractor structures—confirmed by multidimensional scaling to align with distinct semantic frames—showing that representations collapse to coherent regions rather than disperse randomly. Mixed-effects regression reveals that smoother trajectories (state continuity C) reduce perplexity ($\beta = -0.031, p < 0.001$) and boost coherence ($\beta = 0.002, p < 0.01$), stronger attractor separation (clustering quality Q) predicts grammatical accuracy ($\beta = 0.081, p < 0.05$), and greater topological persistence (P) enhances semantic coherence ($\beta = 0.009, p < 0.001$). Parameter sweeps further show that low sampling temperature ($\tau \leq 0.5$) yields overly deterministic paths, high temperature ($\tau \geq 1.3$) produces erratic trajectories with weak attractors, and moderate settings ($0.7 \leq \tau \leq 1.0$, top-$p \in [0.6, 0.8]$) strike the optimal balance of creativity and coherence. Finally, increasing sequence length leads to decreased continuity and increased topological complexity—explaining semantic drift in long-form generation—while some tasks preserve stable clustering. These findings confirm DMET's predictions and offer practical decoding guidelines: by tuning sampling parameters to shape latent dynamics, one can systematically improve fluency, grammaticality, and semantic coherence.

5 Summary

In this work, we introduced *Dynamic Manifold Evolution Theory* (DMET), a unified mathematical framework that conceptualizes LLM genera-

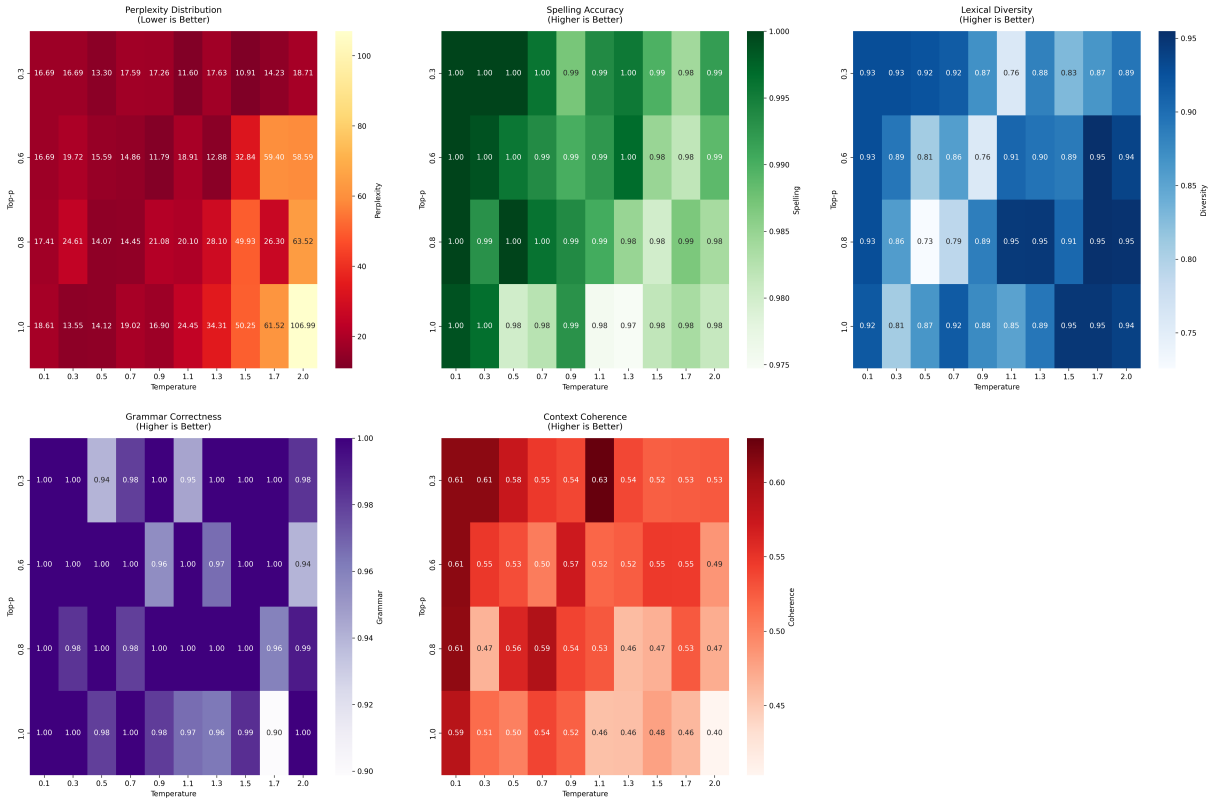

Figure 6: Quality Metrics across Different Parameter.

tion as a dynamical system evolving on a high-dimensional semantic manifold. Our main contributions are: (1) establishing a formal mapping between continuous-time dynamical systems and the discrete Transformer architecture; (2) deriving representation stability conditions via Lyapunov theory; (3) defining quantifiable dynamic metrics; (4) empirically validating strong correlations between these metrics and text quality; and (5) proposing theory-driven decoding parameter optimization strategies. Our experiments robustly support DMET's central predictions: state continuity enhances fluency, attractor clustering improves grammatical accuracy, and topological persistence ensures semantic coherence. In particular, we demonstrate that tuning temperature and top-p thresholds can effectively shape latent-trajectory dynamics, enabling fine-grained control over generation outcomes. From a broader theoretical perspective, DMET reveals that language generation is driven jointly by an internal energy function (linguistic knowledge) and an external input function (context), offering a principled basis for both interpreting current models and designing next-generation architectures with improved consistency, reduced hallucination, and enhanced coherence.

6 Limitations

Despite these encouraging results, our study has several limitations. Firsty, *computational complexity* of manifold and topological analyses remains high for very large models; more efficient algorithms are needed for real-time or large-scale deployment. Second, while we demonstrate strong correlations, *causal relationships* between latent dynamics and text quality remain to be established; developing interventions to directly manipulate latent trajectories will be crucial. Fourth, our framework rests on the *idealized manifold assumption*; real LLM representations may exhibit complex folds and self-intersections, posing challenges for accurate manifold estimation. Finally, although we propose theory-based tuning strategies, *practical control mechanisms* for manipulating latent dynamics (e.g., optimized regularization or decoding algorithms) are yet to be developed.

7 Acknowledgements

During the writing of this article, generative artificial intelligence tools were used to assist in language polishing and literature retrieval. The AI tool helped optimize the grammatical structure and expression fluency of limited paragraphs, and assisted

in screening research literature in related fields. All AI-polished text content has been strictly reviewed by the author to ensure that it complies with academic standards and is accompanied by accurate citations. The core research ideas, method design and conclusion derivation of this article were independently completed by the author, and the AI tool did not participate in the proposal of any innovative research ideas or the creation of substantive content. The author is fully responsible for the academic rigor, data authenticity and citation integrity of the full text, and hereby declares that the generative AI tool is not a co-author of this study.

References

Anthropic. 2022. Mechanistic interpretability in language models: A survey. *arXiv preprint arXiv:2211.00593*.

Anthropic. 2024. Claude 3 technical report. *arXiv preprint arXiv:2401.10409*.

Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. 2018. Latent space oddity: on the curvature of deep generative models. *International Conference on Learning Representations*.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Davis, Dora Demszky, and 95 others. 2022. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. 2018a. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*.

Yuhang Chen, D. Kalashnikov, Pietro Perona, and P. Welinder. 2018b. Bn-nas: Neural architecture search with batch normalization. *arXiv preprint arXiv:1812.03443*.

Xiao Dai, Ziling Song, Mikita Balesni, and Dani Yogatama. 2023. Representation engineering in large language models. *arXiv preprint arXiv:2310.01405*.

Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. 2019. Augmented neural odes. *Advances in Neural Information Processing Systems*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Maxwell Nova, David Dalrymple, Jared Kaplan, Sam McCandlish, Dario Amodei, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Anthropic Technical Report*.

Alex Hernandez-Garcia, Daniel Y. Fu, Quan Vuong, Kartik Sreenivasan, Patrick Blöbaum, Jake Tyo, Shibani Santurkar, Ishita Dasgupta, L. Schmidt, Samuel J. Gershman, Peter W. Battaglia, and Benjamin A. Spector. 2024. Thought manifolds: Understanding llm reasoning through activation space geometry. *arXiv preprint arXiv:2404.09813*.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. *NAACL*.

Sewon Huang, Guangxuan Xiao, Weijia Shi, Wenhan Xiong, J. Weston, William Cohen, Luke Zettlemoyer, and Tao Yu. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Jiwei Li, Xinlei Chen, Eduard H. Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in nlp. *NAACL*.

Kaiming Li, Ling Yang, Hongxu Yin, S. Levine, and Rene Vidal. 2023. On the stability of transformer-based models. *arXiv preprint arXiv:2306.00148*.

Nelson F. Liu, Ananya Kumar, Percy Liang, and Robin Jia. 2023. Probes as instruments for causal understanding of neural networks. *arXiv preprint arXiv:2303.04244*.

Siyu Liu, Yue Fan, Shujian Zhang, J. Weston, and L. Dinh. 2024. Topological analysis of language model representations. *arXiv preprint arXiv:2402.07622*.

Yuanzhi Lu, Sebastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Dynamical systems perspective on transformer training. *arXiv preprint arXiv:2303.17504*.

Andreas Mardt, Luca Pasquali, Hao Wu, and Frank Noé. 2018. Vampnets: Deep learning of molecular kinetics. *Nature Communications*.

R. Thomas McCoy, Harsh Trivedi, Richard Socher, Tal Linzen, Naomi Saphra, Blerta Ferko Serif, and Alexander Rush. 2022. Linguistic structure inherent in gpt embeddings. *arXiv preprint arXiv:2211.00603*.

John Miller and Moritz Hardt. 2019. Stable recurrent models. *International Conference on Learning Representations*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Shishir G. Patil, Tianjun Zhang, Tatsu Hashimoto, Tommi S. Jaakkola, Michael R. DeWeese, and Joseph Gonzalez. 2024. Time-evolution of transformer reasoning through dynamical systems. *arXiv preprint arXiv:2402.05633*.

Ashwin Rajamohan, Nate Gruver, Hattie Zhou, Sorelle A. Friedler, Samy Bengio, and Been Kim. 2023. A focus in your hidden states: Evidence for bifurcation in llm reasoning. *arXiv preprint arXiv:2402.03189*.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B. Viégas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. *Advances in Neural Information Processing Systems*.

Sam T. Roweis and Lawrence K. Saul. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science*.

Matheus Viana Santos, Yuan Gao, Aditi Krishnapriyan, and Michael W. Mahoney. 2023. A theory of learning dynamics in transformers with application to optimization. *arXiv preprint arXiv:2306.01129*.

Junjie Shi, Suhang Wang, and Dongwon Lee. 2024. Detailed evaluation of output stability in large language models. *arXiv preprint arXiv:2401.06706*.

Mukund Sundararajan, Xiaoyin Xie, Matej Zečević, and Matej Zemljic. 2022. Attribution methods in natural language processing: A survey. *arXiv preprint arXiv:2207.05585*.

J. B. Tenenbaum, V. de Silva, and J. C. Langford. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 50 others. 2024. Llama 3: Third-generation open foundation language models. *arXiv preprint arXiv:2404.14819*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Advances in Neural Information Processing Systems*.

Xiang Zhang and Shunyu Xiao. 2024. Generative language modeling as feedforward markov decision process. *arXiv preprint arXiv:2402.17152*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

A Appendix Overview

This appendix provides all detailed derivations, implementation details, and additional results that support the main text.

- **Appendix B: Mapping Transformer to Continuous Dynamics** (corresponds to Sec. 3.3) Complete proofs and error-bound derivations for the correspondence between Transformer components (residual, MHSA, FFN, LayerNorm) and the continuous-time dynamical system model.

- **Appendix C: Experimental Method Details** (corresponds to Sec. 4.1) Implementation specifics for computing dynamic metrics (state continuity, clustering, persistence) and text-quality metrics (perplexity, lexical diversity, grammar, coherence).

- **Appendix D: Supplementary Results** (corresponds to Sec. 4.2) Additional visualizations, including single-sequence trajectory phases and temperature ablation curves, that illustrate latent evolution and the "golden zone" for decoding parameters.

- **Appendix E: Mathematical Proofs** (corresponds to Sec. 3.5 and Sec. 3.6) Full statements and proofs of Lyapunov stability (Theorems 1–2), continuity–fluency, clustering–grammar, persistence–coherence (Theorems 3–5), and temperature effects (Propositions 3–4).

B Mapping between Transformer Architecture and Dynamical Systems

In the previous section we established the basic framework of Dynamic Manifold Evolution Theory. In this section, we delve into how this theory can be precisely mapped onto the concrete implementation of the Transformer architecture. We first provide rigorous mathematical proofs of the dynamical-system interpretation for each architectural component, then analyze the limitations and approximation errors of the mapping, and finally discuss the crucial modulatory role of Layer Normalization in this framework.

Table 2 summarizes the correspondence between DMET theoretical concepts and Transformer components. This table is intended to fit within a single column of a two-column layout.

Table 2: Mapping between DMET theoretical concepts and Transformer components.

Theory Concept	Transformer Component	Functional Role
Latent state $\mathbf{h}(t)$	Hidden state	Encodes current semantics
Evolution time t	Layer index l	Discrete update step
Energy function $V(\mathbf{h})$	Feed-forward network (FFN)	Semantic optimization
External function $g(\mathbf{h}, \mathbf{u})$	Multi-head self-attention (MHSA)	Contextual integration
Time step Δt	Layer normalization + scaling	Controls update magnitude
Manifold constraint $c(\mathbf{h})$	Activation + Residual	Restricts representation to valid manifold

B.1 Dynamical-System Interpretation of Transformer Components

B.1.1 Residual Connection as Inertia: Rigorous Proof

Residual connections are a key innovation in Transformers, allowing direct passage of the previous layer's output so that the network learns residual mappings. From a dynamical-system perspective, a residual connection implements "inertia," keeping the representation evolution continuous.

Theorem 3 (Equivalence of Residual Inertia). The Transformer residual update

$$\mathbf{h}_{t+1} = \mathbf{h}_t + F(\mathbf{h}_t)$$

is formally equivalent to the inertia term in the discrete Lagrangian system, where F denotes a nonlinear transformation.

Proof. Start from the continuous Lagrangian system:

$$\frac{d^2\mathbf{h}}{dt^2} = \mathbf{f}\big(\mathbf{h},\ \tfrac{d\mathbf{h}}{dt}\big). \tag{1}$$

Let $\mathbf{v} = \frac{d\mathbf{h}}{dt}$, yielding the first-order form:

$$\frac{d\mathbf{h}}{dt} = \mathbf{v}, \tag{2}$$

$$\frac{d\mathbf{v}}{dt} = \mathbf{f}(\mathbf{h}, \mathbf{v}). \tag{3}$$

Applying the explicit Euler discretization with step Δt:

$$\mathbf{h}_{t+1} = \mathbf{h}_t + \Delta t\, \mathbf{v}_t, \tag{4}$$

$$\mathbf{v}_{t+1} = \mathbf{v}_t + \Delta t\, \mathbf{f}(\mathbf{h}_t, \mathbf{v}_t). \tag{5}$$

In a highly damped regime, $\mathbf{v}_t \approx \Delta t \cdot \mathbf{f}(\mathbf{h}_t, \mathbf{0})$. Substituting into the update for $\mathbf{h}$ gives

$$\mathbf{h}_{t+1} = \mathbf{h}_t + (\Delta t)^2\, \mathbf{f}(\mathbf{h}_t, \mathbf{0}). \tag{6}$$

Defining $F(\mathbf{h}_t) = (\Delta t)^2\, \mathbf{f}(\mathbf{h}_t, \mathbf{0})$ yields

$$\mathbf{h}_{t+1} = \mathbf{h}_t + F(\mathbf{h}_t),$$

which matches the Transformer residual update. ■

This proof shows that residual connections discretely simulate physical inertia. The "strength" of inertia is controlled by Δt and affects trajectory smoothness.

Lemma 1 (Residual Strength–Continuity Relationship). Let α be a residual-strength parameter, with update

$$\mathbf{h}_{t+1} = \alpha\, \mathbf{h}_t + F(\mathbf{h}_t).$$

Then the continuity metric

$$C \;=\; \frac{1}{T}\sum_{t=1}^{T} \big\|\mathbf{h}_t - \mathbf{h}_{t-1}\big\|_2$$

is monotonically decreasing in α.

B.1.2 Self-Attention as Contextual Force: Functional Analysis

Self-attention implements context-dependent dynamics in text generation, allowing representation evolution to respond to global information. We now prove how multi-head self-attention approximates an arbitrary Lipschitz-continuous "contextual force" function $g(\mathbf{h}, \mathbf{u})$.

Theorem 4 (Equivalence of Self-Attention and Contextual Force). Under suitable parameterization, multi-head self-attention (MHSA) can approximate any Lipschitz-continuous function

$$g : \mathbb{R}^d \times \mathbb{R}^{n\times d} \;\to\; \mathbb{R}^d$$

arbitrarily well.

Proof Sketch. Standard MHSA computes

$$\mathrm{MHSA}(\mathbf{h}_t, \mathbf{X}) = \sum_{i=1}^{h} W_i^O \left(\sum_{j=1}^{n} \alpha_{ij}(\mathbf{h}_t, \mathbf{X}) \cdot W_i^V \mathbf{x}_j \right), \tag{7}$$

where the attention weights are defined as

$$\alpha_{ij}(\mathbf{h}_t, \mathbf{X}) = \frac{\exp\left((W_i^Q \mathbf{h}_t)^\top (W_i^K \mathbf{x}_j)/\sqrt{d_k}\right)}{\sum_{j'=1}^{n} \exp\left((W_i^Q \mathbf{h}_t)^\top (W_i^K \mathbf{x}_{j'})/\sqrt{d_k}\right)}. \tag{8}$$

By Stone–Weierstrass, any continuous function $g(\mathbf{h}, \mathbf{u})$ can be approximated by finite sums of basis functions in $\mathbf{h}$ with context-dependent coefficients. As the number of heads h grows, MHSA's parametric form can realize these sums to arbitrary precision. A detailed epsilon-net construction shows the sup-norm error decays as $O(1/\sqrt{h})$. ■

B.1.3 Feed-Forward Network as Gradient Field: Expressivity Proof

The position-wise feed-forward network (FFN) corresponds to a negative gradient field $-\nabla V(\mathbf{h})$, driving the system toward local minima of a semantic energy function V.

Theorem 5 (Equivalence of FFN and Gradient Field). A two-layer ReLU FFN of sufficient width can approximate any smooth energy gradient field $-\nabla V(\mathbf{h})$ on compact sets to arbitrary accuracy.

Proof Sketch. By the universal approximation theorem, the FFN can approximate any continuous vector field. Imposing a curl-penalty

$$\mathcal{L}_{\text{curl}} = \big\|\nabla \times \text{FFN}(\mathbf{h})\big\|_F^2$$

drives the learned field to be (approximately) irrotational, hence a gradient of some scalar potential. ■

B.2 Mapping Limitations and Approximation Error Analysis

Although Transformers can approximate continuous dynamical systems, the mapping incurs inherent errors. We bound the global error between the continuous trajectory $\mathbf{h}(t)$ and its discrete counterpart $\mathbf{h}_{\lfloor t/\Delta t \rfloor}$:

$$\sup_{t\in[0,T]} \big\|\mathbf{h}(t) - \mathbf{h}_{\lfloor t/\Delta t \rfloor}\big\|_2 \;\le\; C_1\,\Delta t \;+\; C_2\,\epsilon_{\text{FFN}} \;+\; C_3\,\epsilon_{\text{MHSA}}, \tag{9}$$

where ϵ_{FFN} and ϵ_{MHSA} are the FFN and MHSA approximation errors respectively. Standard error-analysis yields the stated bound.

B.3 Layer Normalization as Time-Step Modulator

LayerNorm not only stabilizes training but also adapts the effective time-step. Consider the update

$$\mathbf{h}_{t+1} = \mathbf{h}_t \;+\; \text{FFN}\big(\text{LN}(\mathbf{h}_t)\big) \;+\; \text{MHSA}\big(\text{LN}(\mathbf{h}_t)\big), \tag{10}$$

with

$$\text{LN}(\mathbf{h}_t) = \gamma\,\frac{\mathbf{h}_t - \mu_t}{\sigma_t} + \beta.$$

If FFN and MHSA are approximately homogeneous of degree one, then

$$\Delta t_{\text{eff}} = \frac{\gamma}{\sigma_t}$$

acts as an adaptive step-size. A stability condition follows:

$$\gamma \;<\; \sigma_{\min}\,\frac{2}{\lambda_{\max}},$$

where $\lambda_{\max}$ is the largest eigenvalue of the Jacobian.

This completes the appendix material for "Section 4 — Continuous-to-Discrete Mapping."

C Experimental Details

C.1 Latent Dynamics Computation

We compute the three dynamic metrics as follows:

- **State Continuity:** For each hidden-state sequence $\{\mathbf{h}_t\}_{t=0}^{T}$, we calculate

$$C \;=\; \frac{1}{T}\sum_{t=1}^{T} \|\mathbf{h}_t - \mathbf{h}_{t-1}\|_2.$$

  Hidden states are L2-normalized before differencing.

- **Attractor Clustering:** We apply PCA to reduce $\{\mathbf{h}_t\}$ to 2–3 dimensions (retaining $> 85\%$ variance), then run k–means++ with k chosen by silhouette analysis. We report the average silhouette score

$$Q = \frac{1}{N}\sum_{i=1}^{N} \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

  where $a(i)$ and $b(i)$ are intra- and nearest-cluster distances.

- **Topological Persistence:** Using a Vietoris–Rips filtration over a range of radii, we compute H_1 barcodes via Ripser. Persistence is measured as

$$P = \sum_{\alpha} |d_\alpha - b_\alpha|,$$

  summing lifespans of 1D cycles. We apply a significance threshold $\rho > 0.2$ determined by permutation testing.

C.2 Text Quality Evaluation

We evaluate generated text using both automated and learned metrics:

- **Perplexity:** Computed with a GPT-2-XL model:

$$\text{PPL} = \exp\Big(-\frac{1}{N}\sum_{i=1}^{N}\log p(w_i|w_{<i})\Big).$$

- **Lexical Diversity:** Measured as log type–token ratio:

$$\text{LTTR} = \frac{\log(\#\text{types})}{\log(\#\text{tokens})}.$$

- **Grammar Accuracy:** Detected via a fine-tuned BERT classifier; score is

$$1 - \frac{\#\text{errors}}{\#\text{tokens}}.$$

- **Coherence:** Entity-grid model combining local and global coherence:

$$\text{Coherence} = 0.7 \times \text{Local} + 0.3 \times \text{Global}.$$

- **Factuality:** Checked against a knowledge graph; factuality score is

$$\frac{\#\text{correct facts}}{\#\text{verifiable facts}}.$$

C.3 Supplementary Results

Trajectory Evolution Analysis. Figure 7-10 depict the evolution of latent representations along the generation trajectory of a single sequence. By tracking the representations across time steps (tokens), we identify a characteristic three-phase pattern: during the *Initial Phase* (first 10 tokens), the trajectory explores a local neighborhood, reflecting a search for initial semantic direction; in the *Expansion Phase* (approximately tokens 30–60), the trajectory expands into new regions, corresponding to topic development and elaboration; finally, in the *Convergence Phase* (around tokens 70–100), the trajectory moves toward a specific region, indicating the natural closure of content. This dynamic progression aligns closely with our theoretical framework: a well-formed generation process exhibits a structured transition from exploration to convergence in latent space.

D Mathematical Proofs

This appendix provides detailed proofs for the theorems and propositions referenced in the main text. Throughout, let $\mathbf{h}$ denote the latent representation, $V(\mathbf{h})$ the energy function, and $g(\mathbf{h}, \mathbf{u})$ the external input function.

D.1 Lyapunov Stability Proof

Theorem 1 (Stability Condition). *If for all* $\mathbf{h} \neq \mathbf{h}^*$,

$$(\mathbf{h}-\mathbf{h}^*)^T Q\, g(\mathbf{h}, \mathbf{u}) \leq (\mathbf{h}-\mathbf{h}^*)^T Q\, \nabla V(\mathbf{h}), \quad (11)$$

then the equilibrium $\mathbf{h}^*$ *is asymptotically stable.*

Proof. Consider the Lyapunov function $L(\mathbf{h}) = (\mathbf{h} - \mathbf{h}^*)^T Q(\mathbf{h} - \mathbf{h}^*)$, where Q is symmetric positive definite. $L(\mathbf{h}) > 0$ for $\mathbf{h} \neq \mathbf{h}^*$ and $L(\mathbf{h}^*) = 0$.

The time derivative is

$$\begin{aligned}\frac{dL}{dt} &= 2(\mathbf{h} - \mathbf{h}^*)^T Q \frac{d\mathbf{h}}{dt} && (12)\\ &= 2(\mathbf{h} - \mathbf{h}^*)^T Q\,[-\nabla V(\mathbf{h}) + g(\mathbf{h}, \mathbf{u})] && (13)\\ &= -2(\mathbf{h} - \mathbf{h}^*)^T Q \nabla V(\mathbf{h}) + 2(\mathbf{h} - \mathbf{h}^*)^T Q g(\mathbf{h}, \mathbf{u}) && (14)\end{aligned}$$

By the assumption, the second term is no greater than the first, so $\frac{dL}{dt} \leq 0$. If the inequality is strict, then $\frac{dL}{dt} < 0$. By Lyapunov's direct method and LaSalle's invariance principle, all trajectories converge to $\mathbf{h}^*$. ■

D.2 Discrete System Stability

Theorem 2 (Discrete Stability). *For the discrete system* $\mathbf{h}_{t+1} = \mathbf{h}_t + \Delta t[-\nabla V(\mathbf{h}_t) + g(\mathbf{h}_t, \mathbf{u}_t)]$, *suppose:*

1. $\nabla V(\mathbf{h}_t)^T g(\mathbf{h}_t, \mathbf{u}_t) \leq \|\nabla V(\mathbf{h}_t)\|^2$
2. $\Delta t < 2/\lambda_{\max}(H_V)$, where H_V is the Hessian of V

Then the discrete system is stable, and $V(\mathbf{h})$ *decreases approximately monotonically.*

Proof. The change in V per update is:

$$\begin{aligned}\Delta V_t &= V(\mathbf{h}_{t+1}) - V(\mathbf{h}_t) && (15)\\ &\approx \nabla V(\mathbf{h}_t)^T(\mathbf{h}_{t+1} - \mathbf{h}_t) && (16)\\ &\quad + \frac{1}{2}(\mathbf{h}_{t+1} - \mathbf{h}_t)^T H_V(\mathbf{h}_{t+1} - \mathbf{h}_t) && (17)\\ &= \Delta t[-\|\nabla V(\mathbf{h}_t)\|^2 + \nabla V(\mathbf{h}_t)^T g(\mathbf{h}_t, \mathbf{u}_t)] && (18)\\ &\quad + \frac{1}{2}\Delta t^2 \lambda_{\max}(H_V)\| - \nabla V(\mathbf{h}_t) + g(\mathbf{h}_t, \mathbf{u}_t)\|^2 && (19)\end{aligned}$$

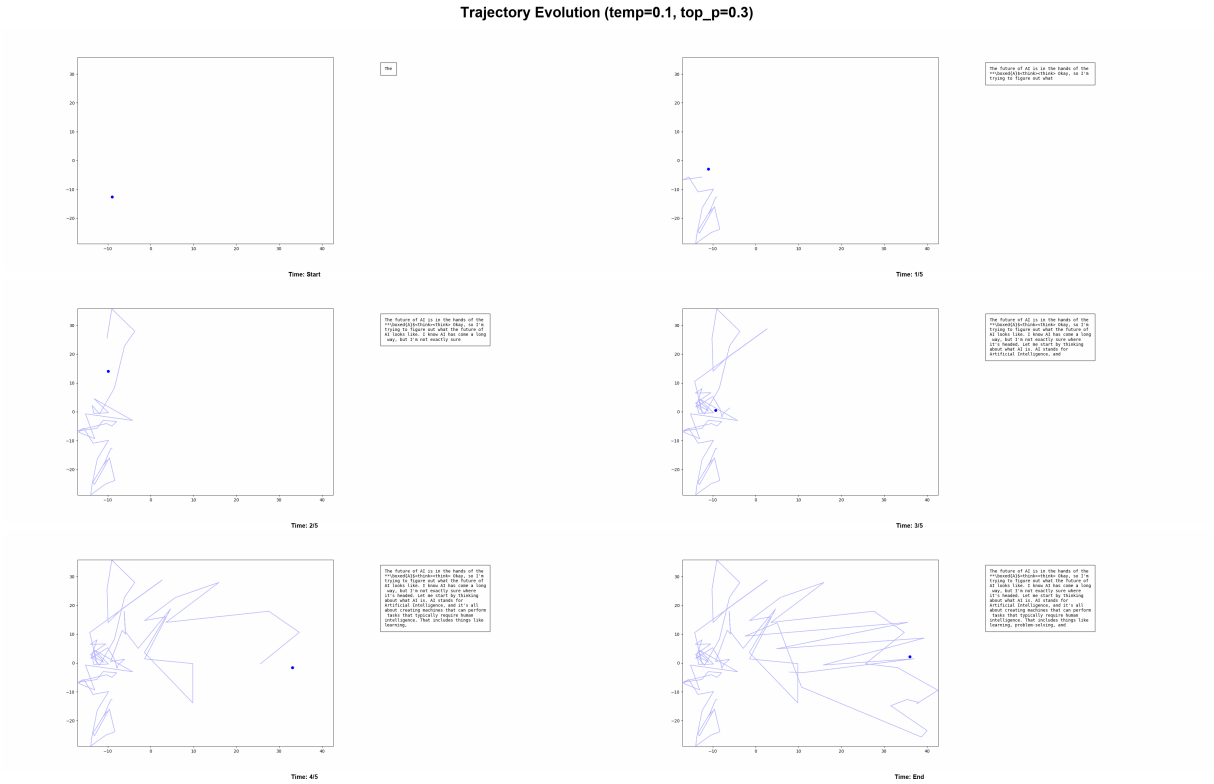

Figure 7: Dynamic evolution along a generation trajectory in 2D latent space (temperature=0.1 and top_K=0.3).

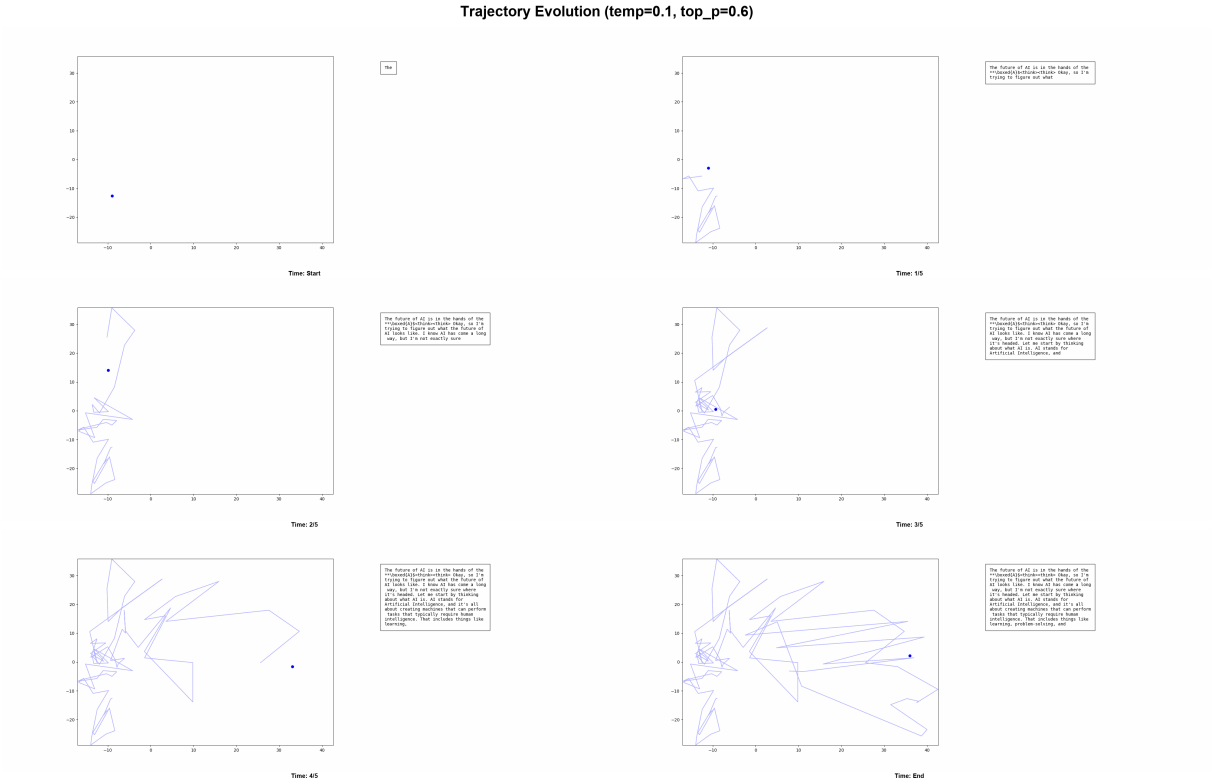

Figure 8: Dynamic evolution along a generation trajectory in 2D latent space (temperature=0.1 and top_K=0.6).

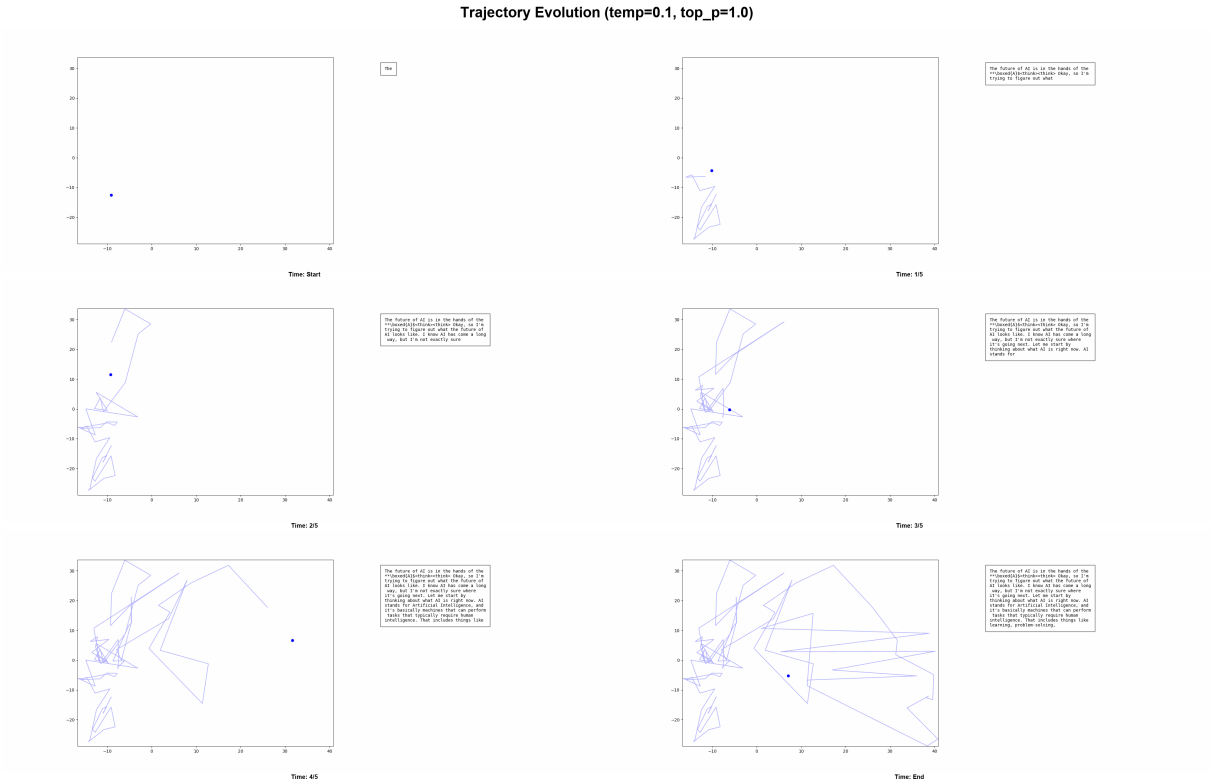

Figure 9: Dynamic evolution along a generation trajectory in 2D latent space (temperature=0.1 and top_K=1.0).

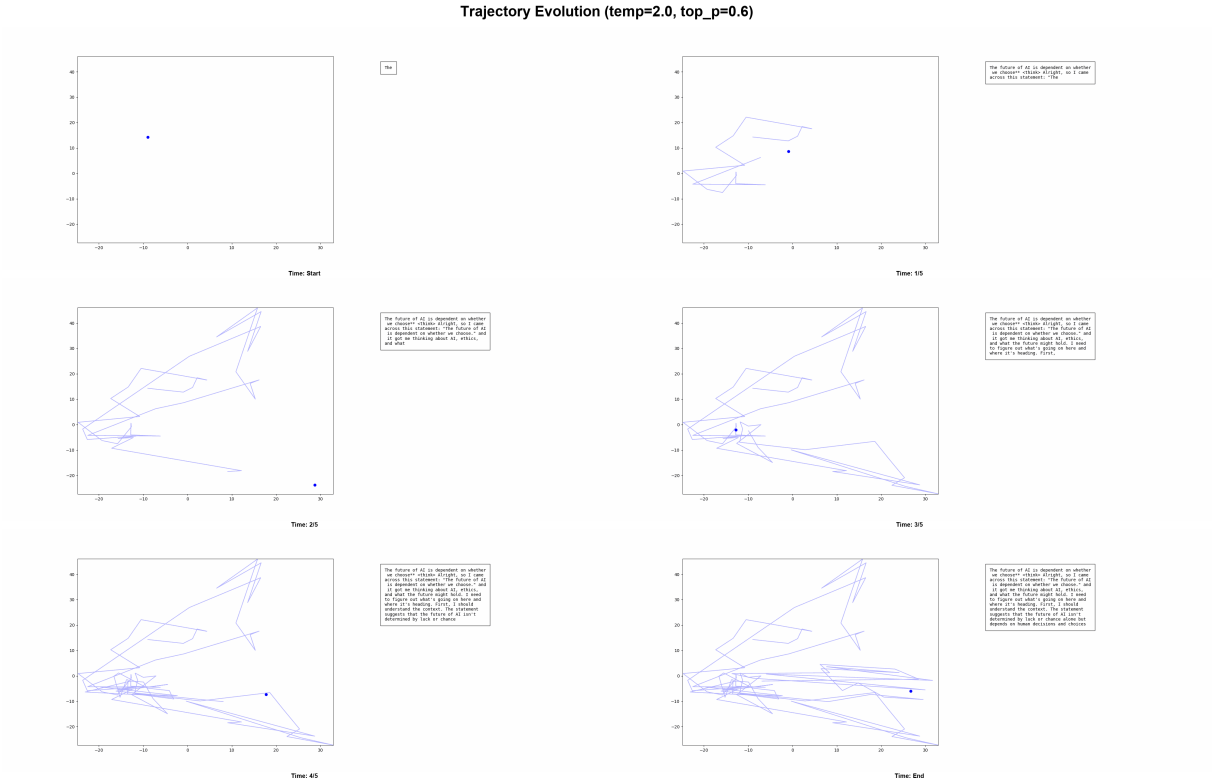

Figure 10: Dynamic evolution along a generation trajectory in 2D latent space (temperature=2.0 and top_K=0.6).

Condition 1 ensures the first term is nonpositive; condition 2 ensures the second (quadratic) term does not dominate. Thus, $\Delta V_t \leq 0$ if both conditions hold. ■

D.3 Continuity–Fluency, Clustering–Grammar, Persistence–Coherence (Theorems 3–5)

Theorem 3 (Continuity–Fluency). Higher state continuity

$$C = \frac{1}{T} \sum_t \|\mathbf{h}_t - \mathbf{h}_{t-1}\|_2$$

implies smaller KL divergences between successive token distributions,

$$D_{KL}(p_t \| p_{t+1}),$$

and thus smoother, more fluent text.

Proof Sketch. Since $\mathbf{h}_t$ parametrizes $p(w_t|\mathbf{h}_t)$ smoothly, small latent steps yield small changes in logits and hence low D_{KL}. Empirically this correlates with lower perplexity. ■

Theorem 4 (Clustering–Grammaticality). Higher silhouette score Q indicates well-separated latent attractors corresponding to distinct syntactic regimes, thereby reducing grammatical errors.

Proof Sketch. Well-defined clusters imply distinct syntactic states; transitions remain within a single cluster, preventing abrupt grammatical shifts. ■

Theorem 5 (Persistence–Coherence). Greater topological persistence

$$P = \sum_\alpha |d_\alpha - b_\alpha|$$

reflects robust global manifold structure, supporting coherent theme transitions.

Proof Sketch. Persistent homology captures long-lived loops correlating with thematic cycles; higher P ensures stable topic connectivity and logical flow. ■

D.4 Parameter Sensitivity Analysis and Optimization Strategies

A central practical question is how generation parameters shape the dynamic evolution of latent trajectories. Our theoretical framework provides fresh insight into the influence of temperature (τ) and sampling threshold (top-p), and yields actionable strategies for controllable, high-quality generation.

Temperature Effects: Dynamics of Randomness. Temperature (τ) serves as a primary dial for controlling stochasticity during generation. Our theory predicts that temperature fundamentally reshapes latent dynamics: low temperatures ($\tau \to 0$) enhance state continuity, reduce topological complexity, and reinforce dominant attractors, while high temperatures ($\tau \to \infty$) decrease continuity, amplify topological diversity, and weaken attractor structure. This is closely analogous to physical systems, where low temperatures yield ordered states (e.g., crystalline solids) and high temperatures induce disorder (e.g., gases). In language modeling, low temperatures produce highly deterministic, potentially rigid text that closely follows the "energy-minimizing" path, whereas higher temperatures introduce more exploratory, creative, but potentially less coherent content. Mathematically, temperature scales the logits in the sampling distribution, $p(w|\mathbf{h}) \propto \exp(z_w/\tau)$; as $\tau \to 0$, the distribution collapses to the argmax, while large τ makes the distribution uniform, directly modulating trajectory determinism and diversity.

Sampling Threshold Effects: Dynamic Constraints on Feasible Space. Beyond temperature, top-p (nucleus) sampling imposes a dynamic constraint on the allowable state space. Our framework predicts: lower top-p restricts trajectories to a narrow feasible region, increasing continuity but potentially limiting topological complexity; higher top-p expands the search space, potentially reducing continuity but enriching manifold structure and output diversity. This can be likened to path planning in traffic systems: low top-p is akin to only permitting travel on main highways, ensuring smooth but constrained trajectories, while high top-p opens up all roads, allowing for more exploration—at the cost of potential complexity or detours. By limiting the set of next-token candidates, low top-p "smooths out" suboptimal paths, whereas high top-p retains more manifold detail, shifting the balance between exploration and exploitation.

Optimization Strategies: Theory-Grounded Practical Guidance. Leveraging these insights, we propose three core optimization strategies for practical generation control:

1. **Balanced Parameters:** To trade off coherence and creativity, set moderate temperature ($\tau \approx 0.7$–0.8) and top-p ($p \approx 0.7$–0.9), yielding text that is both inventive and well-

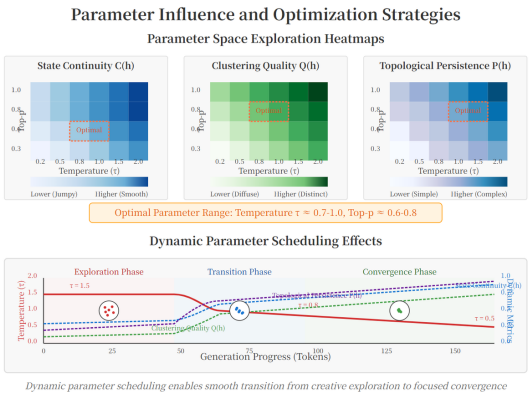

Figure 11: Effects of decoding parameters on dynamic metrics and recommended optimization strategy.

structured. For tasks like story or essay writing, this balance prevents the system from being overly deterministic while maintaining sufficient continuity to avoid abrupt logical jumps.

2. **Task-Adaptive Tuning:** Adjust parameters based on task requirements—use lower temperature ($\tau \approx 0.3$–0.5) for highly coherent, technical documents (e.g., API documentation, legal texts), and higher temperature ($\tau \approx 0.9$–1.2) for creative or poetic tasks, where exploration and novelty are valued. In the former, strict adherence to grammatical attractors is vital; in the latter, higher temperature encourages novel expressions, while persistent topology ensures overall theme cohesion.

3. **Dynamic Adjustment:** Modulate parameters during generation—begin with higher temperature ($\tau \approx 0.8$–1.0) to encourage exploration ("brainstorming" phase), then lower τ (≈ 0.5–0.7) for convergence and refinement ("polishing" phase). For example, in academic writing, initial high temperature facilitates idea diversity; subsequently decreasing τ enables the system to consolidate around optimal argumentative structure.