Empirical Investigation of Latent Representational Dynamics in Large Language Models: A Manifold Evolution Perspective

Yukun Zhang*
The Chinese University of Hong Kong
Hong Kong, China
215010026@link.cuhk.edu.cn

QI DONG*
Fudan University
Shanghai, China
19210980065@fudan.edu.cn

Abstract

This paper introduces the Dynamical Manifold Evolution Theory (DMET), a conceptual framework that models large language model (LLM) generation as a continuous trajectory evolving on a low-dimensional semantic manifold. The theory characterizes latent dynamics through three interpretable metrics—state continuity (C), attractor compactness (Q), and topological persistence (P)—which jointly capture the smoothness, stability, and structure of representation evolution. Empirical analyses across multiple Transformer architectures reveal consistent links between these latent dynamics and text quality: smoother trajectories correspond to greater fluency, and richer topological organization correlates with enhanced coherence. Different models exhibit distinct dynamical regimes, reflecting diverse strategies of semantic organization in latent space. Moreover, decoding parameters such as temperature and top-p shape these trajectories in predictable ways, defining a balanced region that harmonizes fluency and creativity. As a phenomenological rather than first-principles framework, DMET provides a unified and testable perspective for interpreting, monitoring, and guiding LLM behavior, offering new insights into the interplay between internal representation dynamics and external text generation quality.

1 Introduction

In recent years, Large Language Models (LLMs) have marked a revolutionary leap in artificial intelligence, with models like GPT-4 (OpenAI, 2023), Llama 3 (AI @ Meta, 2024), and Claude 3 (Anthropic, 2024) demonstrating unprecedented capabilities in understanding and generating nuanced human language. However, this functional prowess is shadowed by their internal opacity. These models often operate as "black boxes" (Bommasani et al., 2022), a characteristic that poses a significant barrier to enhancing their reliability, interpretability, and safety. A critical knowledge gap lies in understanding how these models dynamically organize and evolve their latent representations during the generative process. This gap directly hinders our ability to address persistent challenges such as factual hallucinations (Min et al., 2023), logical inconsistencies (Zheng et al., 2023)

Existing research has sought to illuminate these internal mechanics through various analytical lenses. Techniques such as attention visualization (Vaswani et al., 2017), feature attribution (Sundararajan et al., 2017), and representation probing (Hewitt and Manning, 2019) have successfully revealed static properties of model representations. Concurrently, work on mechanistic interpretability, including the analysis of residual streams (Elhage et al., 2021), has begun to trace information flow across discrete computational steps. While invaluable, these efforts often provide either static snapshots or fragmented views of the model's internal state. They lack a unified theoretical framework capable of describing the continuous, temporal evolution of representations as a cohesive whole. The traditional view of generation as a mere chain of token predictions overlooks the smooth underlying dynamics in the latent space, where abstract concepts are progressively refined into coherent text.

This paper introduces the **Dynamic Manifold Evolution Theory (DMET)**, an innovative mathematical framework that reconceptualizes the LLM generation process. We move beyond discrete steps and model generation as a controlled dynamical system whose state evolves along a trajectory on a low-dimensional semantic manifold. Our central insight is that coherent text generation is not an atomized process but rather the observable result of a continuous evolution in a structured latent space. By integrating principles from dynamical systems theory, manifold geometry, and deep learn-

*These authors contributed equally to this work.

ing, DMET provides a rigorous foundation for understanding, analyzing, and ultimately controlling the internal representational dynamics of LLMs.

The primary contributions of this work are fourfold:

1. **A Novel Conceptual Framework:** We propose DMET, which, for the first time, formally models the entire LLM generation process as a dynamical system evolving on a semantic manifold. We establish a clear mapping between the components of this system and the modules of the Transformer architecture.

2. **Quantitative Dynamical Metrics:** We introduce a toolkit of three operational metrics—State Continuity (C), Attractor Clustering Quality (Q), and Topological Persistence (P)—that quantitatively link the geometric and topological properties of latent trajectories to the quality of the generated text.

3. **Systematic Empirical Validation:** We conduct extensive experiments across three major LLM architectures, three diverse task types, and a wide range of decoding parameters. Our results demonstrate a strong and consistent correlation between the proposed dynamical metrics and observable text qualities like fluency, grammaticality, and coherence.

4. **Actionable Insights for Control:** The framework uncovers distinct "dynamical fingerprints" of different models and identifies optimal "golden zones" for decoding parameters. These findings provide principled, theory-grounded strategies for mitigating issues like semantic drift and for steering LLM output to better suit specific application requirements.

Through this comprehensive framework, we aim not only to deepen our understanding of how LLMs work but also to provide theoretical guidance for the development of more reliable and controllable next-generation language models. The remainder of this paper details the mathematical foundations of DMET, presents the experimental validation, and discusses the broader implications of our findings.

2 Related Work

Our work is positioned at the intersection of three key research areas: the application of dynamical systems to deep learning, the analysis of manifold structures in representations, and the modeling of latent trajectories in language models.

Dynamical Systems in Neural Networks. The interpretation of deep neural networks as discretized continuous systems has become an influential paradigm, starting with Neural ODEs (Chen et al., 2018), which frame residual connections as steps in an Euler integration. This perspective has been extended to more complex models like Augmented Neural ODEs (Dupont et al., 2019) and has spurred stability analyses for various architectures (Miller and Hardt, 2019; Santos et al., 2023; Li et al., 2023). Within NLP, researchers have applied this lens to analyze specific aspects of Transformer dynamics (Lu et al., 2023; Wang et al., 2024) or to reframe decoding as a control problem (Zhang et al., 2024). However, these studies tend to be localized, focusing on a single component or property in isolation. In contrast, our framework proposes a **holistic mapping**, unifying all core Transformer components within a single, continuous-time dynamical system governed by principles like Lyapunov stability.

Manifold Geometry and Topology in Representations. The manifold hypothesis—the idea that high-dimensional data lies on a low-dimensional intrinsic structure (Roweis and Saul, 2000; Tenenbaum et al., 2000)—is foundational to representation learning. While prior work has focused on characterizing the *static geometry* of these latent manifolds—for instance, by estimating their Riemannian curvature (Arvanitidis et al., 2018) or analyzing them through the lens of the neural tangent kernel (Jacot et al., 2018)—our contribution lies in modeling the *dynamics on* the manifold. In NLP, this geometric perspective has been used to find linguistic structure through probes (Hewitt and Manning, 2019), visualize representations (Reif et al., 2019), and uncover conceptual hierarchies (Hernandez et al., 2022). Topological tools like persistent homology have further revealed global structural features (Liu et al., 2024; Dai et al., 2023). DMET builds directly on these insights, shifting the focus from characterizing a static map of the semantic space to modeling the "traffic flow"—the generative trajectory—that unfolds upon it.

Latent Trajectory Analysis in Language Models. Tracing the evolution of hidden states has long been a method for understanding generative models, from early visualizations of RNN behavior

(?Mardt et al., 2018) to analyses of information propagation in Transformer residual streams (Elhage et al., 2021). More recent studies have investigated specific dynamic phenomena such as trajectory bifurcations (Rajamohan and Kello, 2023) and the evolution of a "thought manifold" (Hernandez et al., 2024). These analyses are often qualitative or descriptive, focusing on visualizing and identifying interesting phenomena. DMET distinguishes itself by moving from **qualitative observation to quantitative modeling and prediction**. Instead of just observing trajectories, we model them as the solution to a differential equation, define metrics to quantify their properties, and use these metrics to empirically predict the quality of the final generated text.

In synthesis, while prior work has examined dynamics, manifolds, and trajectories in isolation, DMET is the first to integrate them into a single, predictive framework that models the entire generative process as a geometric flow from abstract concept to concrete text.

3 The DMET Framework

We introduce the *Dynamic Manifold Evolution Theory* (DMET), a framework that models the generative process of Large Language Models (LLMs) as a controlled dynamical system. This section presents the three core assumptions, mathematical formulation, and the linkage between latent dynamics and observable text quality.

Core assumptions. DMET is grounded on three intertwined assumptions that reinterpret LLM generation as continuous dynamical evolution.

(A1) *Manifold.* The meaningful latent states of an LLM, though embedded in a high-dimensional space $\mathbb{R}^d$, lie on a smooth low-dimensional manifold $\mathcal{M} \subset \mathbb{R}^d$ with $\dim(\mathcal{M}) \ll d$. *Rationale:* Natural language obeys syntactic and semantic regularities, confining its representations to a structured subspace, consistent with the manifold hypothesis (Roweis and Saul, 2000).

(A2) *Continuity.* Text generation corresponds to a smooth trajectory $\mathbf{h}(t) : [0, T] \to \mathcal{M}$ rather than discrete jumps. *Rationale:* Although tokens are discrete, residual connections enable a continuous underlying evolution akin to differential equations.

(A3) *Attractors.* The semantic manifold $\mathcal{M}$ is organized into attractor basins $\{\mathcal{A}_i\}_{i=1}^K$, each reflecting a coherent semantic or syntactic regime. *Rationale:* Once a topic or style is established, LLMs tend to maintain it, mirroring convergence to attractor states in dynamical systems.

Table 1: Structural correspondence between DMET and Transformer.

Dynamics	Module	Functional Role
$-\nabla V(\mathbf{h})$	Feed-Forward Network (FFN)	Semantic refinement
$g(\mathbf{h}, \mathbf{u})$	Multi-Head Attention (MHSA)	Context integration
Residual term	Residual Connection	Trajectory continuity
Δt	Layer Normalization	Update magnitude modulation

3.1 Dynamical System Formulation

We formulate latent evolution as a controlled dynamical system.

Continuous-time model.

$$\frac{d\mathbf{h}(t)}{dt} = -\nabla V\big(\mathbf{h}(t)\big) + g\big(\mathbf{h}(t), \mathbf{u}(t)\big), \quad (1)$$

where $V(\cdot)$ is a potential function whose negative gradient drives the system toward stable semantic states, and $g(\cdot)$ is a context-dependent forcing term encoding input $\mathbf{u}(t)$. This formulation separates intrinsic linguistic priors (in V) from contextual adaptation (in g).

Discrete-time realization in Transformers. Viewing each Transformer layer as an explicit Euler step with step size Δt, Eq. (1) discretizes as

$$\mathbf{h}_{l+1} = \mathbf{h}_l + \Delta t\Big[-\nabla V(\mathbf{h}_l) + g(\mathbf{h}_l, \mathbf{u}_l)\Big]. \quad (2)$$

We propose a structural correspondence between dynamical components and Transformer modules, summarized in Table 1.

3.2 From Dynamics to Measurable Quality

DMET bridges abstract dynamics with observable text quality via falsifiable propositions.

Continuity → Fluency. Higher state continuity, quantified as

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

correlates with lower perplexity and greater fluency. *Rationale:* Smooth latent transitions imply smaller distributional shifts ($D_{KL}(p_t \| p_{t+1})$), yielding more stable token probabilities.

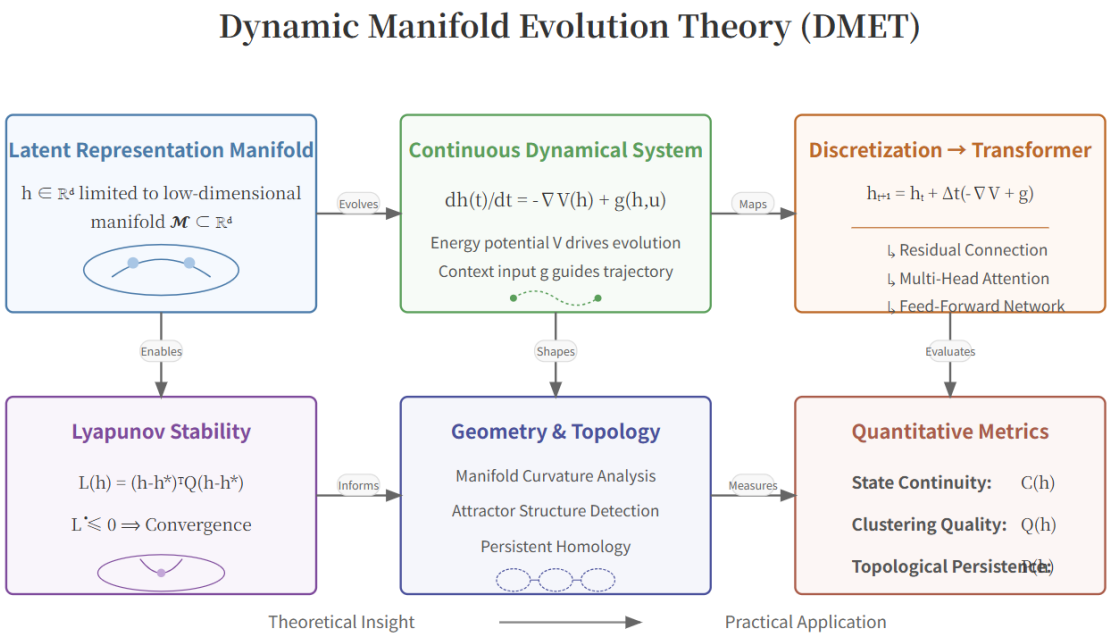

Figure 1: Overview of the DMET framework: latent trajectories evolve on a low-dimensional semantic manifold under intrinsic energy gradients and context-driven forces, with discrete Transformer layers implementing Euler steps of this continuous dynamics.

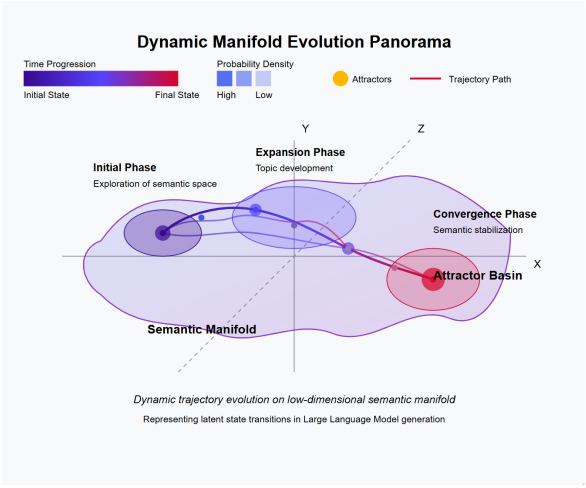

Figure 2: Latent representations reside on a low-dimensional semantic manifold rather than occupying the full ambient space.

Attractors → Consistency. A stronger attractor structure, measured by a higher silhouette score Q, correlates with improved grammatical and stylistic consistency. *Rationale:* Distinct attractors correspond to stable grammatical or semantic modes, reducing inconsistent outputs (e.g., tense or style switching).

Topology → Coherence. Greater topological persistence, reflected by persistent homology score P, indicates enhanced long-range coherence. *Rationale:* Persistent topological features (e.g., 1D loops H_1) capture thematic recurrence and global semantic connectivity.

3.3 Modeling Parametric Effects

DMET offers a principled view on how decoding parameters modulate dynamics.

Temperature as stochasticity. Temperature τ scales randomness: low τ approximates deterministic descent on V, increasing C and Q, while high τ amplifies stochastic exploration, lowering C and Q but enriching topology P.

Top-p as a manifold constraint. Top-p sampling constrains trajectories to high-probability regions of $\mathcal{M}$. Low p confines evolution to "semantic highways" (high C, simple P), while high p enables exploratory "side roads," lowering C but increasing structural richness.

3.4 Epistemological Status

DMET is an *explanatory framework* rather than a first-principles physical theory. The velocity field V and latent manifold $\mathcal{M}$ are phenomenological constructs inferred from empirical behavior rather than derived analytically from the training objective, and attractor structures are identified descriptively through clustering rather than through closed-form equations. Consequently, the value of DMET lies in its ability to unify diverse empirical observations, generate testable predictions and offer principled guidance for interpreting and controlling LLM behavior. Its validity depends on consistent empirical support across models and tasks.

4 Experimental and Result Analysis

4.1 Unified Experimental Design

To validate the predictions of Dynamical Manifold Evolution Theory (DMET), we conduct a unified experiment across diverse language models, prompts, and decoding settings.

Models and Prompts. We evaluate three representative Transformer-based LLMs: **DeepSeek-R1**, **Qwen2**, and **Llama2**. Each model is tested on three prompts reflecting increasing cognitive complexity:

- **Prompt 1 (Simple):** "The future of AI is"
- **Prompt 2 (Argumentative):** "Should AI systems be held legally accountable for their decisions?"
- **Prompt 3 (Creative):** "Complete the story: 'In a world where AI can read minds, a detective discovers a murder victim whose last thought was…' "

Decoding Grid. We use a 6×4 grid of temperature values $\tau \in \{0.1, 0.5, 0.9, 1.3, 1.7, 2.0\}$ and top-p values $\in \{0.3, 0.6, 0.8, 1.0\}$, yielding 24 configurations per prompt. Each configuration generates 10 sequences (100 tokens each), resulting in $3 \times 3 \times 24 \times 10 = 2{,}160$ total samples.

Evaluation Metrics. We assess both *dynamical properties* of hidden state trajectories and *textual quality* of generated sequences.

Dynamical Indicators:

$$C = \frac{1}{T}\sum_{t=1}^{T} \|\mathbf{h}_t - \mathbf{h}_{t-1}\|_2 \quad \text{(State Continuity)}$$

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

$$P = \sum_{\alpha \in H_1} |d_\alpha - b_\alpha| \quad \text{(Topological Persistence)}$$

Here, C captures local smoothness of evolution, Q reflects attractor compactness (via silhouette score), and P measures global structural stability (via persistent homology). We also compute a *healthy rate*—the proportion of samples exhibiting continuous, bounded, and differentiable evolution.

Table 2: Textual Evaluation Metrics

Metric	Description
Perplexity (PPL)	GPT based fluency measure
Lexical Diversity	Log-type-token ratio (creativity)
Spelling Accuracy	Character-level error rate
Grammar Accuracy	Fine-tuned BERT classifier
Semantic Coherence	Entity-grid based score

Statistical Analysis. We use binomial and one-sample t-tests to assess thresholds ($Q > 0.3$, $P > 1.0$), and apply mixed-effects regression to predict fluency, coherence, and other scores from C, Q, and P, with τ and top-p as random effects. Parameter sensitivity and trajectory visualizations are presented in subsection **??**.

4.2 Empirical Validation of DMET

We empirically validate the Dynamical Manifold Evolution Theory (DMET) across three leading Transformer LLMs—DeepSeek-R1, Llama2, and Qwen2—on three prompt types (P1–P3), totaling 2,160 generations. As shown in Table 3, all generations meet the bounded-smoothness criterion (100% healthy), with average attractor quality $\bar{Q} = 0.56$ and persistence $\bar{P} = 3.98$, significantly exceeding theoretical thresholds ($p < 0.01$). These results strongly support DMET's universality across architectures and tasks.

Distinct dynamical regimes emerge: Qwen2 shows high Q and low P ("layered"), Llama2 has moderate Q and high P ("networked"), and DeepSeek-R1 balances both ("hybrid"). These regimes likely stem from architectural and pretraining differences.

Table 3: Dynamical Feature Validation Across Models and Prompts.

Model	Prompt	Healthy	Mean Q	Mean P	p-val
DeepSeek-R1	P1	1.00	0.50	4.70	<0.01
	P2	1.00	0.47	4.90	<0.01
	P3	1.00	0.52	4.95	<0.01
Llama2	P1	1.00	0.50	5.41	<0.01
	P2	1.00	0.50	5.32	<0.01
	P3	1.00	0.54	5.46	<0.01
Qwen2	P1	1.00	0.65	2.69	<0.01
	P2	1.00	0.69	2.26	<0.01
	P3	1.00	0.66	2.14	<0.01

Table 4: Mixed-Effects Regression: Continuity vs. Output Properties

Predictor	Log-PPL	Diversity
C (Continuity)	-0.031***	-0.003***
Q (Clustering)	0.044	-0.074
P (Persistence)	0.000	-0.003
Random Var.	0.962***	0.265

Table 5: Mixed-Effects Regression: Predicting Coherence

Predictor	Coherence
C (Continuity)	0.002**
Q (Clustering)	0.047
P (Persistence)	0.009***
Random Var.	0.115*

To evaluate explanatory power, we run mixed-effects regressions. Table 4 shows that continuity C significantly reduces perplexity ($\beta = -0.031$, $p < 0.001$) and lexical diversity, confirming Proposition 3.1 and a fluency–creativity trade-off. Table 5 confirms that persistence P is the strongest predictor of coherence ($\beta = 0.009$, $p < 0.001$), with C also significant ($p < 0.01$), validating Proposition 3.3.

Finally, Table 6 compares text quality. DeepSeek-R1 achieves lowest perplexity (fluency), Llama2 highest coherence, and Qwen2 greatest diversity—mirroring their dynamical traits. These results confirm DMET's predictive utility for quality analysis.

4.3 Trajectory Dynamics and Parameter Sensitivity

We analyze how decoding parameters influence generation behavior, both statistically and geometrically. A 2D grid search was conducted over temperature $\tau \in \{0.1, 0.5, 0.9, 1.3, 1.7, 2.0\}$ and top-$p \in \{0.3, 0.6, 0.8, 1.0\}$ using DeepSeek-R1 as a representative model. Each configuration generated 400 samples.

Dynamical–Linguistic Interaction. Figure 3 shows three key effects: (i) Low τ (≤ 0.5) yields minimal PPL (≈ 20), while $\tau \geq 1.3$ sharply increases PPL beyond 100, suggesting a sensitive region around $\tau \in [0.6, 1.0]$; (ii) Coherence peaks in a "golden zone" $\tau \in [0.7, 1.0]$, $p \in [0.6, 0.8]$ with convex decline outside; (iii) Lexical diversity increases with τ and p but reduces fluency, reflecting a trade-off surface.

Empirically, recommended decoding setups are: (a) $\tau = 0.7$–0.9, $p = 0.7$ for balanced generation; (b) $\tau = 1.0$–1.2, $p = 0.8$–1.0 for creative writing; (c) $\tau = 0.3$–0.5, $p = 0.5$–0.6 for formal contexts.

Trajectory Geometry and Attractors. Figure 4 visualizes PCA-projected hidden state trajectories. Samples cluster into two semantic attractors (descriptive vs. speculative), with clear convergence from purple (init) to red (end), consistent with manifold flow predicted by DMET. Trajectories evolve smoothly along low-dimensional paths.

Three-Phase Evolution. Across settings, we observe a three-phase trajectory pattern (Figure 5): (1) *Initialization* (0–10 tokens): unstable drift, topic setup; (2) *Expansion* (10–70): directed evolution with diffusion; (3) *Convergence* (70–100): stabilization into semantic attractors.

Lower τ accelerates convergence, while higher τ prolongs expansion and increases entropy. Top-p modulates the radius of local exploration. These findings support DMET's view of Transformer generation as smooth, phase-structured flows on latent manifolds.

4.4 Empirical Insights and Theoretical Implications

Table 7 summarizes the empirical validation of DMET across 2,160 generations involving three

Table 6: Text Quality Metrics Across Models and Prompts.

Model–Prompt	PPL↓	Spelling↑	Diversity↑	Grammar↑	Coherence↑
DeepSeek-R1–P1	3.01	0.99	0.89	0.98	0.53
DeepSeek-R1–P2	3.08	0.99	0.92	0.99	0.45
DeepSeek-R1–P3	3.28	0.99	0.84	0.99	0.51
Llama2–P1	3.44	0.98	0.98	0.99	0.50
Llama2–P2	3.25	0.99	0.94	0.99	0.44
Llama2–P3	3.70	0.99	0.95	0.99	0.46
Qwen2–P1	3.90	0.98	0.96	0.97	0.53
Qwen2–P2	3.60	0.99	0.93	0.94	0.49
Qwen2–P3	3.93	0.99	0.98	0.98	0.47

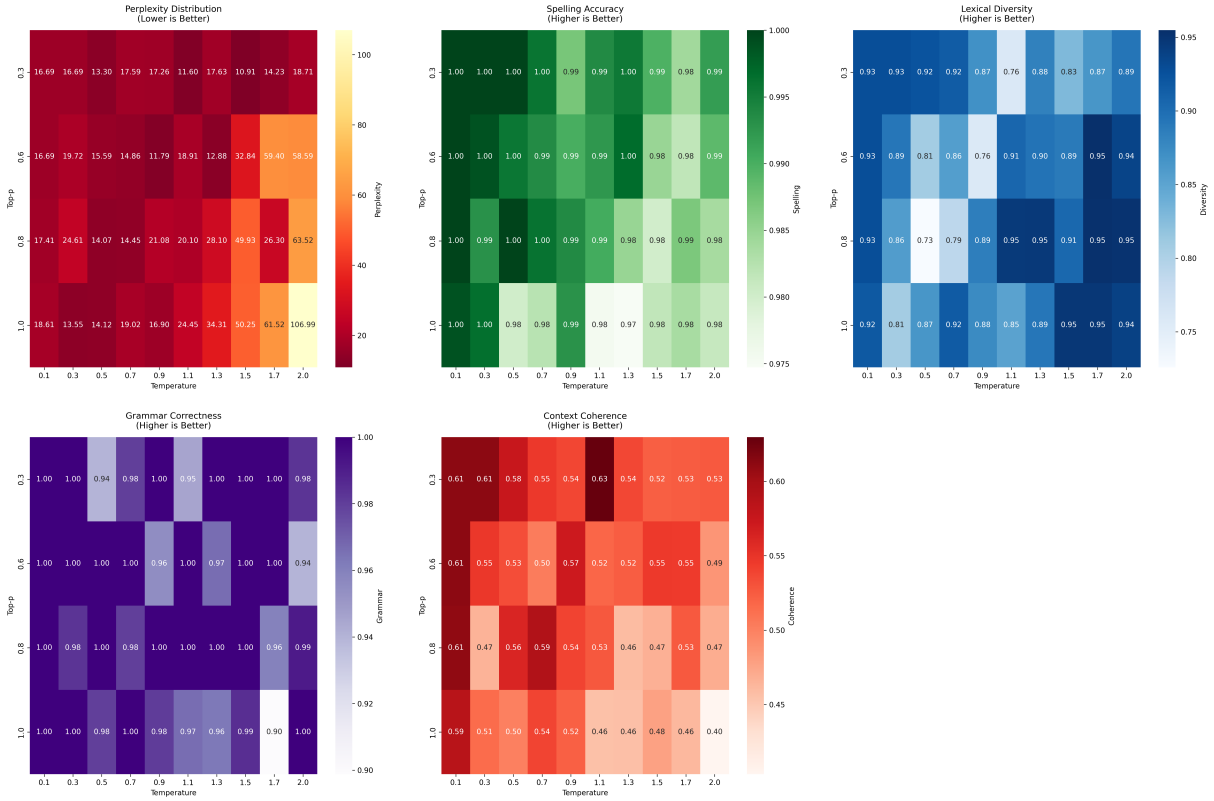

Figure 3: Fluency–Coherence trade-off under varying τ and p (DeepSeek-R1); a "golden zone" achieves optimal fluency, diversity, and structure.

Table 7: Empirical validation of DMET theoretical propositions.

Prediction	Effect (β)	Significance
$C \rightarrow$ Fluency	-0.031	$p < 0.001$
$Q \rightarrow$ Grammar	0.081	$p < 0.05$
$P \rightarrow$ Coherence	0.009	$p < 0.001$

models and prompt types. All three core propositions are statistically supported, with significant and interpretable effects on generation quality.

Architectural Diversity. Each model exhibits a distinct dynamical profile: **Qwen2** demonstrates compact clusters and low persistence—favoring stylistically constrained tasks; **Llama2** displays rich topological complexity aligned with coherent long-form reasoning; **DeepSeek-R1** achieves balanced values across C, Q, and P, leading to high fluency and stability.

Theoretical Contribution. These findings support the interpretation of LLMs as dynamical systems evolving on structured attractor manifolds. The triplet of continuity (C), compactness (Q), and persistence (P) provides a concise yet expressive framework for modeling generative dynamics.

Practical Utility. Each dynamical metric maps directly to interpretable quality signals:

- C: fluency stability—useful for detecting topic drift.
- Q: stylistic consistency—linked to grammar

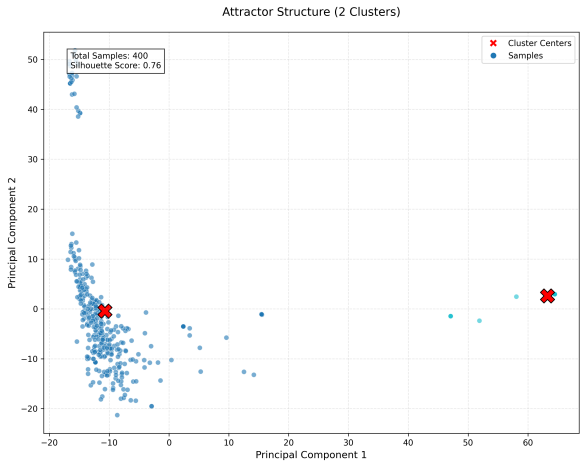

Figure 4: PCA projection of hidden-state dynamics under varying decoding; semantic attractors and directional convergence validate DMET predictions.

and repetition.

- P: semantic depth—sensitive to coherence over longer spans.

These metrics offer a model-agnostic, lightweight toolkit for diagnosing and aligning LLM behavior with downstream application needs.

4.5 Summary and Implications

All three dynamical hypotheses are strongly supported: trajectory smoothness is universal (100% healthy rate, $p < 0.01$), attractor quality is robust ($\bar{Q} = 0.56 > 0.3$, $p < 0.001$), and persistence exceeds the stability threshold ($\bar{P} = 3.98 > 1.0$, $p < 0.001$

5 Summary

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

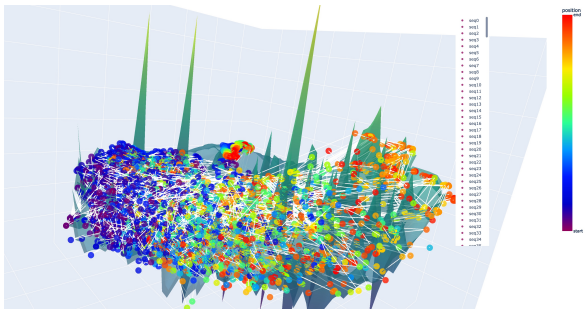

Figure 5: Representative hidden-state trajectory evolution under varying τ; higher temperatures increase spread and delay convergence.

tegrity of the full text, and hereby declares that the generative AI tool is not a co-author of this study.

References

AI @ Meta. 2024. The Llama 3 model family: A foundation for real-world applications. *arXiv preprint arXiv:2404.11082*.

Anthropic. 2024. The Claude 3 model family: Opus, sonnet, haiku. *arXiv preprint arXiv:2402.19143*.

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

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, and 1 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 Duvenaud. 2018. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*.

Yang Dai, Jacob Andreas, and Catherine Olsson. 2023. Representation elbow: A method for guiding subspace selection in probing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. 2019. Augmented neural ODEs. In *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, and 1 others. 2021. A mathematical framework for Transformer circuits. *Transformer Circuits Thread, Distill*.

Gabriel Hernandez, Béryl Gauthier, Andrew K. Lampinen, Ishita Dasgupta, and Jean-Rémi King. 2024. The thought manifold of language models. *arXiv preprint arXiv:2404.10856*.

Gabriel Hernandez, Siyuan Xie, Béryl Gauthier, Gopala K. Anumanchipalli, Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. 2022. Linguistic structure in the geometrical space of brain and language model representations. *Nature Communications*, 13(1):6850.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Arthur Jacot, Franck Gabriel, and Clément Hongler. 2018. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*.

Yuxian Li, Ge Meng, Jiacheng Chen, Xiaofei Wu, Jixing Su, Jiaxu Zhou, and Yu He. 2023. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. In *The Eleventh International Conference on Learning Representations*.

Kexin Liu, Yuxuan Wang, and Mingsheng Long. 2024. Topological attention for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhaofeng Lu, Fengxiang Zhang, Yaxin Shi, and Weijie Su. 2023. A dynamical system perspective of gated recurrent units. In *The Eleventh International Conference on Learning Representations*.

Andreas Mardt, Luca Pasquali, Hao Wu, and Frank Noé. 2018. VAMPnets for deep learning of molecular kinetics. *Nature Communications*, 9(1):5.

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

Sewon Min, Kalpesh Krishna, Yejin Lyu, Mike Lewis, Hannaneh Hajishirzi, Pang Wei Huang, Luke Zettlemoyer, and Yejin Choi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

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

Aris Rajamohan and Christopher T. Kello. 2023. A fixed-point investigation of catastrophic forgetting in neural networks. *arXiv preprint arXiv:2305.15286*.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda Viegas, James Wexler, and Hal Ludwig. 2019. Visualizing and measuring the geometry of BERT. In *Advances in Neural Information Processing Systems*.

Sam T. Roweis and Lawrence K. Saul. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326.

Francis Santos, David Macêdo, Cleber Medeiros, and Artur Ziviani. 2023. A lyapunov theory for the analysis of the global convergence of deep learning. *arXiv preprint arXiv:2310.19702*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Yizhong Wang, Wen-tau Yih, Siva Reddy Aluru, and Jordan Chou. 2024. Time evolution of tacit knowledge in language models. *arXiv preprint arXiv:2402.17410*.

Weichen Zhang, Yuxin Zhou, Ge Luo, Yuting Wang, Zhaofeng Wang, and Hong Xu. 2024. Generative language models as a markov decision process. *arXiv preprint arXiv:2404.14589*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Brooks, Eric Xing, 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 additional results that support the main text.

A.1 Supplementary Results

Trajectory Evolution Analysis. Figure 6-9 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.

A.2 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

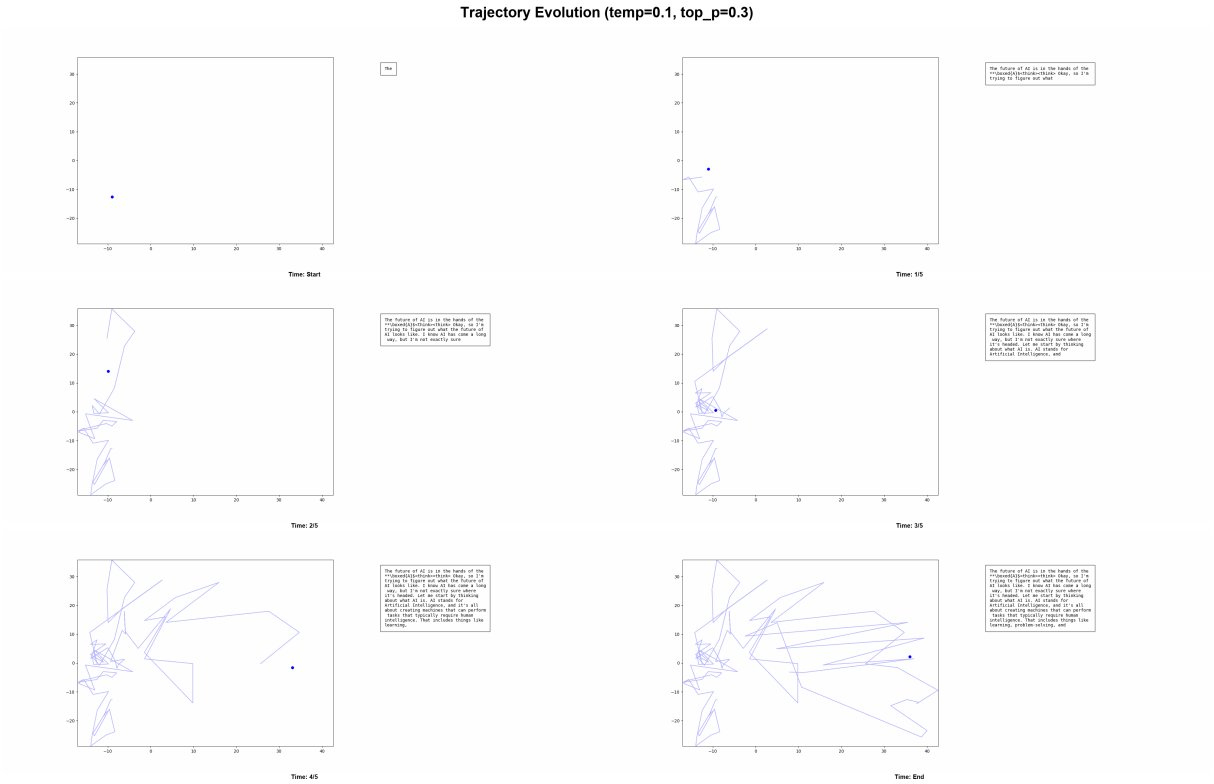

Figure 6: Dynamic evolution along a generation trajectory in 2D latent space (temperature=0.1 and top_K=0.3).

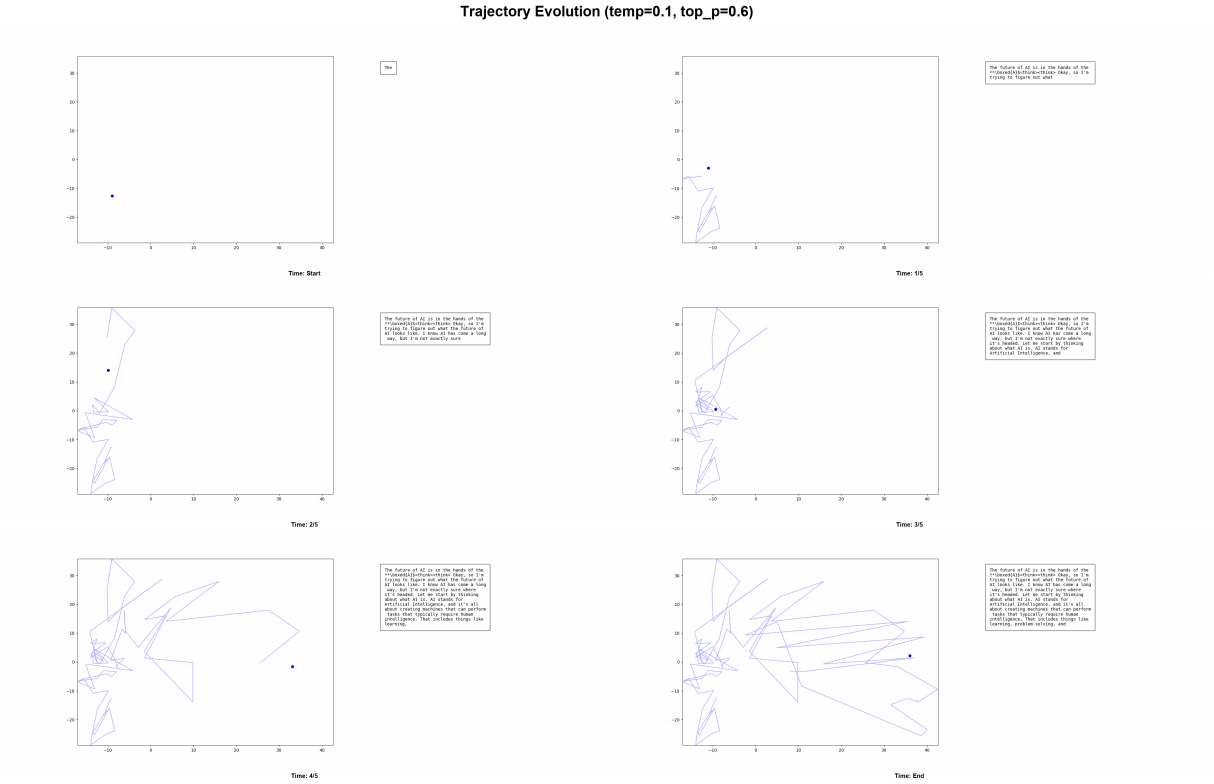

Figure 7: Dynamic evolution along a generation trajectory in 2D latent space (temperature=0.1 and top_K=0.6).

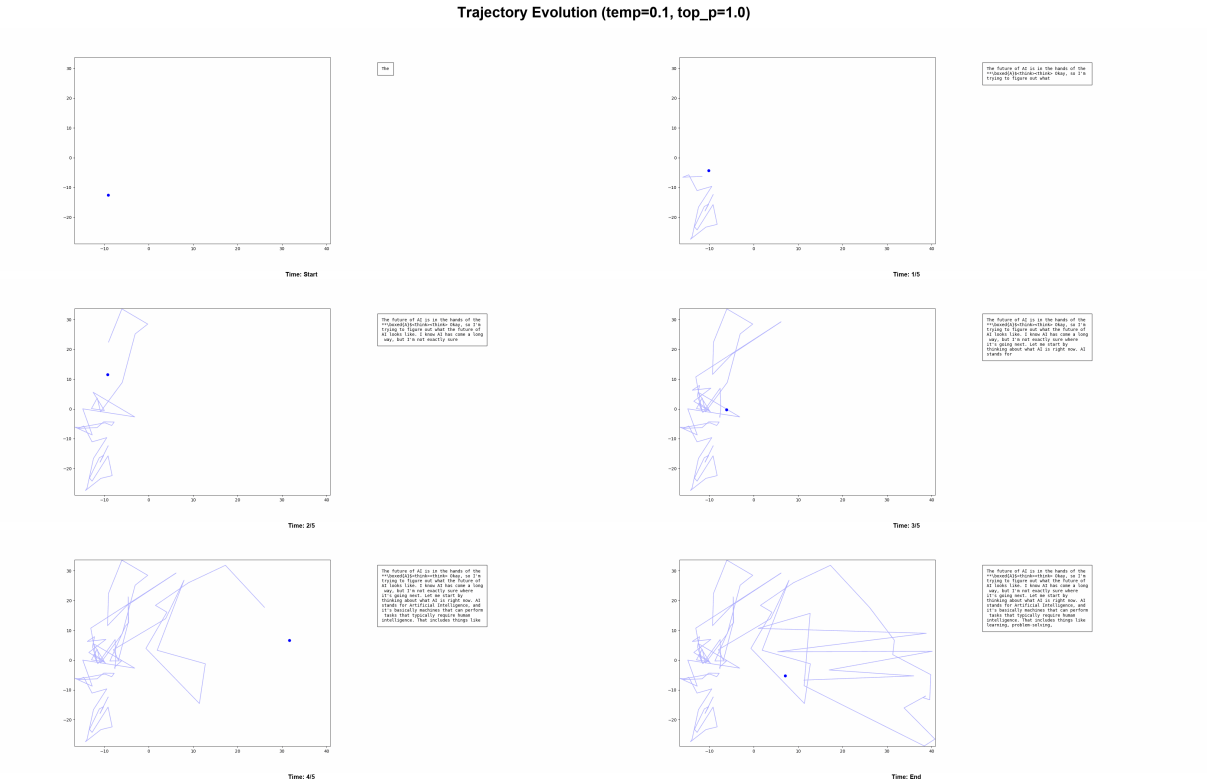

Figure 8: Dynamic evolution along a generation trajectory in 2D latent space (temperature=0.1 and top_K=1.0).

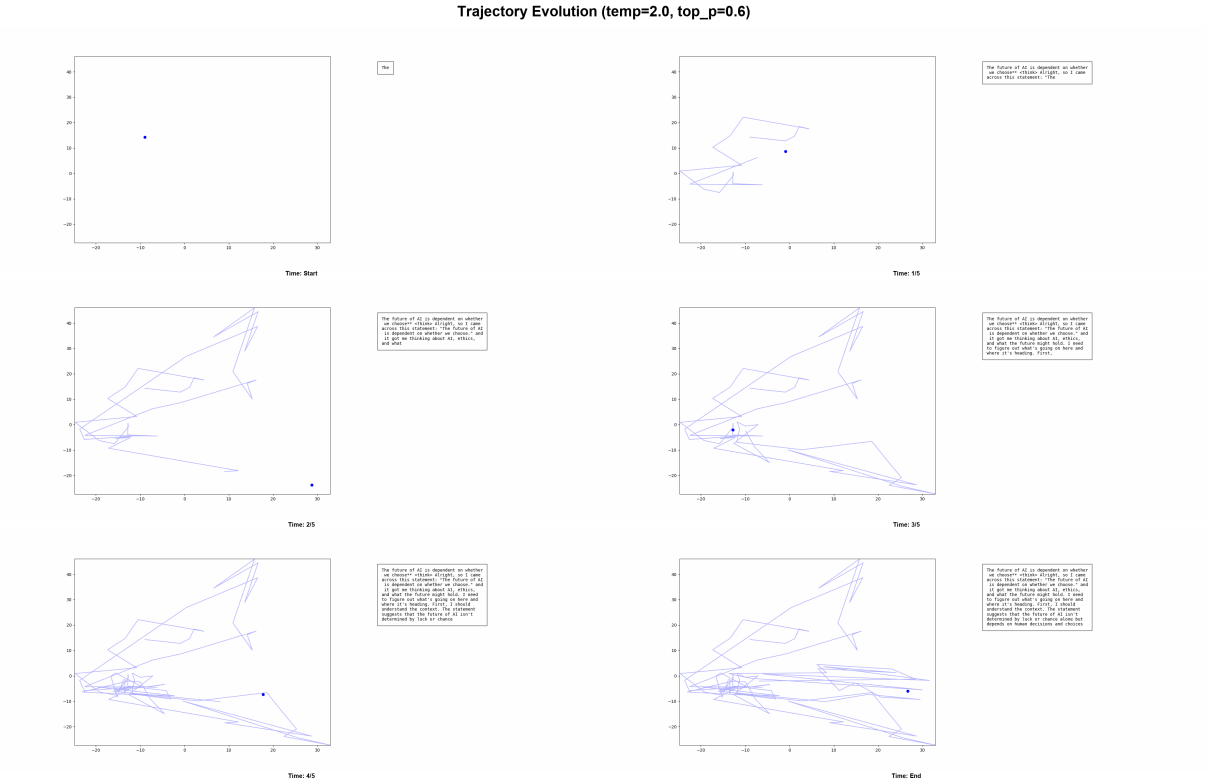

Figure 9: Dynamic evolution along a generation trajectory in 2D latent space (temperature=2.0 and top_K=0.6).

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-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.