

Coherence–Diffusion Dynamics: A Continuous–Semantic Interpretation of Transformer Language Models

Anonymous authors

Paper under double-blind review

Abstract

Large language models (LLMs) exhibit coherent reasoning, long-range contextual integration, and abrupt failures such as hallucination, yet the internal principles governing these behaviors remain poorly understood. Existing interpretability approaches typically focus on isolated components, including attention patterns, neuron circuits, or probing signals, and therefore provide limited insight into how semantic meaning evolves over the course of inference. This work proposes that Transformer-based language models can be productively interpreted through a continuous semantic perspective, in which internal representations evolve along structured trajectories in a latent space. We articulate this interpretation through the Coherence–Diffusion Dynamics (CDD) framework, which models semantic evolution as the interaction of coherence-restoring tendencies and stochastic variability. Within this framework, we introduce an effective instability potential serving as an interpretive proxy for semantic coherence, a coherence operator governing stabilizing dynamics, a diffusion term capturing stochastic variability, and an interpretation of dynamic sparsity capturing the apparent contraction of effective semantic degrees of freedom along inference trajectories. These constructs suggest qualitative, empirically testable implications regarding stabilization, regime shifts associated with hallucination, and the functional irrelevance of low-impact components. We evaluate these implications through controlled experiments on Transformer language models, showing broad alignment between observed behavior and the qualitative predictions of the CDD interpretation. Taken together, this work provides a coherent and dynamically grounded account of semantic evolution in LLMs, providing a principled lens for interpreting coherence, variability, sparsity, and instability without departing from the discrete computational structure of Transformer architectures.

1 Introduction

The rapid development of large language models (LLMs) has transformed contemporary artificial intelligence, enabling systems to generate coherent text, explain complex relationships, and sustain extended dialogue with a level of fluency that increasingly resembles human reasoning (Brown et al., 2020; OpenAI et al., 2024). Despite these advances, however, the internal processes through which LLMs organize, stabilize, and occasionally destabilize semantic meaning remain poorly understood. Transformer architectures offer remarkable expressive capacity, yet their internal computations continue to resist principled interpretation (Rudin, 2019; Vig & Belinkov, 2019). As a result, modern language models often function as black boxes: effective in practice, but elusive in their internal structure.

This opacity highlights a persistent gap between empirical performance and scientific understanding. Existing interpretability approaches typically illuminate isolated components of the model without capturing how semantic structure evolves during inference. Attention visualizations reveal spatial patterns of token interaction, but cannot determine whether a given token meaningfully contributes to the model’s predictive state (Clark et al., 2019). Probing classifiers expose correlations between internal representations and linguistic features, yet fall short of identifying causal mechanisms (Alain & Bengio, 2018). Sampling controls such as

temperature or top- k modulation influence output variability, but do not explain why hallucinations emerge or why semantic drift occurs (Holtzman et al., 2020). Even mechanistic interpretability, while increasingly detailed, remains largely static in nature: it decomposes networks into parts but does not describe how semantic organization changes over time as representations are iteratively updated (Nanda et al., 2023).

At the same time, a growing body of work across representation learning, embedding geometry, and internal network analysis points toward a complementary perspective in which semantic meaning is fundamentally continuous. Early research on distributed representations demonstrated that semantic similarity corresponds to geometric proximity in embedding spaces, suggesting that meaning arises from relational structure rather than discrete symbols (Mikolov et al., 2013b; Pennington et al., 2014). Subsequent studies revealed that contextual embeddings exhibit strong anisotropy, geometric organization, and low intrinsic dimensionality (Ethayarajh, 2019; Arora et al., 2017). More recent analyses of intrinsic dimension and latent geometry further support the view that semantic representations evolve along smooth trajectories within high-dimensional but geometrically constrained spaces (Kataiwa et al., 2025). Mechanistic interpretability research also shows that residual stream updates operate linearly within a shared vector space, indicating that successive layers apply incremental modifications that cumulatively approximate a continuous flow (Nanda et al., 2023).

Taken together, these findings support an increasingly adopted perspective in which linguistic meaning in modern LLMs is organized along continuous semantic manifolds. From this viewpoint, tokens primarily function as discrete surface-level anchors for underlying continuous semantic representations, while the inferential process itself can be viewed as the evolution of an internal state through a latent semantic space. Under such a perspective, inference is no longer described solely as a sequence of isolated symbolic transformations, but as a structured process shaped by forces that promote coherence while allowing controlled variability.

This paper develops an effective dynamical framework that organizes these observations into a unified interpretive model. We introduce Coherence–Diffusion Dynamics (CDD), which interprets Transformer inference as the discretized evolution of a latent semantic trajectory influenced by three interacting factors: coherence-restoring tendencies that promote contextual stability, diffusion-like variability that enables exploration of alternative continuations, and effects interpretable as sparsification that suppress low-influence internal interactions. Rather than positing a literal physical process, CDD provides a modeling perspective that captures qualitative regularities observed across Transformer-based language models.

Within this framework, empirical evidence indicating that only a small subset of internal interactions meaningfully influences semantic evolution is interpreted as an emergent property of the dynamics rather than an architectural artifact. This phenomenon, referred to as dynamic sparsity, is consistent with the view that effective computation can occur within a reduced subspace of the full representation space, a feature that traditional attention-based interpretability methods do not adequately explain.

The contributions of this work are twofold. First, CDD offers a principled interpretive account of semantic stability and instability in LLMs, allowing behaviors such as hallucination, semantic drift, and abrupt loss of coherence to be interpreted as consequences of the interaction between coherence-restoring and diffusion-like influences. Second, it provides a framework for analyzing functional relevance within the model, one that provides a lens for reasoning about which internal components exert meaningful influence on semantic trajectories and which contribute only negligibly. Together, these insights motivate an approach to interpretability that is inherently temporal and dynamical, rather than static and post hoc.

By situating Transformer inference within an effective dynamical systems perspective, this work offers a conceptual perspective for understanding how LLMs maintain coherence, generate diversity, transition between stable and unstable regimes, and compress meaning into low-dimensional semantic pathways. This perspective complements existing interpretability approaches and offers a structured basis for reasoning about stability, efficiency, and semantic reliability in large-scale neural language models.

2 Related Works

Research on meaning representation in artificial systems has progressed through several distinct paradigms, each capturing different aspects of the phenomena exhibited by modern language models. The historical transition from symbolic formalisms to neural architectures reflects a gradual shift away from viewing meaning as

a static, rule-governed entity toward understanding it as a relational, context-dependent, and geometrically structured process. While each paradigm has contributed important insights, none has fully addressed how semantic representations evolve over time during inference in large-scale language models.

Early approaches to natural language understanding treated meaning as a collection of discrete symbolic relations grounded in logical rules, production systems, or grammatical formalisms (Newell & Simon, 1976; Bobrow & Woods). These systems offered explicit reasoning mechanisms and clear interpretability, but they struggled to represent graded similarity, contextual modulation, or continuous semantic variation. Their rigid ontologies were ill-suited to capturing the smooth and adaptive semantic behavior observed in modern language use, and they provided limited insight into the evolving internal representations characteristic of contemporary neural models.

The introduction of distributed vector representations marked a significant conceptual shift. Models such as Word2Vec (Mikolov et al., 2013a) and GloVe (Pennington et al., 2014) demonstrated that semantic similarity could be expressed as geometric proximity in a continuous vector space. Under this paradigm, meaning emerges from relational structure among vectors rather than from symbolic identity. However, these representations were static: each token was associated with a single vector regardless of context. While this work established the geometric nature of meaning, it did not address how semantic representations change dynamically as contextual information accumulates.

Subsequent advances in contextual representation learning revealed a richer internal structure. Contextual embeddings were shown to be anisotropic, highly organized, and concentrated on low-intrinsic-dimensional manifolds within the ambient space (Ethayarajh, 2019; Arora et al., 2017). More recent studies of intrinsic dimensionality and latent geometry further support the view that semantic representations evolve along smooth trajectories in high-dimensional but geometrically constrained spaces (Kataiwa et al., 2025). These findings suggest that meaning is better characterized as a continuously evolving state rather than as a fixed symbolic object.

Recurrent neural networks, including LSTMs and GRUs (Hochreiter & Schmidhuber, 1997; Cho et al., 2014), introduced sequential computation and enabled early forms of context-sensitive semantic updating. By iteratively modifying a hidden state, these models captured temporal dependencies and incremental meaning updates. Nevertheless, structural limitations remained. Long-range dependencies were difficult to maintain due to vanishing gradients (Pascanu et al., 2013), and compressing an entire sequence into a single recurrent state constrained representational capacity. As a result, while RNNs captured aspects of contextuality, they did not naturally support the smooth, high-capacity semantic evolution observed in later architectures.

The Transformer architecture (Vaswani et al., 2023) addressed many of these limitations by replacing recurrence with self-attention, enabling structured interactions among all tokens at each layer. Through attention-based contextual reconstruction, representations are repeatedly updated based on global context, producing a layered evolution of meaning that can be interpreted as incremental movement within a continuous semantic space. This mechanism underlies state-of-the-art models such as BERT (Devlin et al., 2019), GPT-family models (Radford & Narasimhan, 2018; Brown et al., 2020), and LLaMA (Touvron et al., 2023). Mechanistic interpretability studies further show that these models operate through linear residual streams combined with attention-mediated transformations (Nanda et al., 2023), revealing structural regularities that suggest an analogy with discrete approximations of underlying continuous flows.

Despite the growing sophistication of analysis techniques, interpretability challenges persist. Attention-based analyses visualize token interactions but cannot reliably identify functional influence (Jain & Wallace, 2019). Probing classifiers uncover correlations between representations and linguistic features without establishing causal mechanisms (Zhang et al., 2017). Mechanistic interpretability reveals circuits and modules (Olsson et al., 2022), but these approaches remain largely microscopic and do not capture global semantic evolution. Decoding and sampling studies characterize output variability (Holtzman et al., 2020), while loss landscape and dynamical analyses illuminate optimization behavior (Sagun et al., 2018; Fort & Ganguli, 2019), yet none of these approaches directly address inference-time semantic dynamics. As a result, existing methods tend to explain components or phenomena in isolation, leaving unresolved how semantic states evolve as a whole during inference.

A growing body of empirical evidence therefore motivates an interpretive perspective in which LLM inference is viewed as the evolution of an internal state within a continuous semantic space. Residual updates resemble incremental integration steps, attention redistributes semantic influence across representations, and generative behavior reflects the interplay between stabilizing tendencies and stochastic variability. The framework developed in this work, Coherence–Diffusion Dynamics (CDD), situates these observations within an effective dynamical perspective. Rather than replacing existing methodologies, CDD provides a coherent interpretive umbrella under which insights from geometric representations, mechanistic interpretability, training dynamics, and decoding behavior can be jointly understood.

3 Interpreting Transformers as Discrete Approximations to Continuous Semantic Dynamics

Transformer-based language models are typically described as systems that process information through a sequence of discrete layers, each applying attention and feed-forward transformations to update hidden representations. While this description accurately reflects the architectural structure, it does not fully account for the qualitative behavior of internal representations observed during inference. Across model depth, these representations do not behave as isolated outputs of successive computational blocks. Instead, they exhibit gradual, structured changes that are more naturally described as the evolution of a state along a smooth semantic trajectory. This observation motivates an interpretive perspective in which Transformer inference is viewed as a discrete approximation to an underlying continuous semantic process.

A central consideration supporting this perspective is that Transformer hidden states reside in high-dimensional continuous vector spaces rather than in symbolic or discrete domains. Each hidden representation can be modeled as a point on an effective semantic manifold embedded in \mathbb{R}^d , where local geometric relationships encode semantic similarity. Within this space, small representational updates accumulated across layers correspond to gradual semantic shifts rather than abrupt symbolic transitions. Empirical studies have shown that the intrinsic dimensionality of these representations is often significantly lower than the ambient dimensionality (Ethayarajh, 2019), reflecting redundancy, compression, and structure that emerge as contextual information is progressively integrated.

The architecture of the Transformer further reinforces the interpretation of inference as a gradual semantic evolution. A defining feature of the model is the residual update structure $x_{t+1} = x_t + \Delta x_t$, which modifies the current representation through incremental changes rather than complete replacement. This update pattern admits a close analogy with numerical integration schemes used to approximate continuous dynamical systems, in which a state evolves through repeated small steps governed by an underlying vector field. Under this analogy, the layer index can be interpreted as playing a role similar to a discretized time variable, with each layer corresponding to a forward step along an effective semantic trajectory shaped by learned transformations.

Despite this structured update mechanism, Transformer inference does not follow a single deterministic path. Multiple sources of variability introduce fluctuations into semantic evolution. Sampling procedures inject randomness during generation, while entropy in the softmax distribution reflects ambiguity among competing continuations. In addition, residual activations encode superposed semantic possibilities that may be resolved differently across layers or decoding contexts. Even under deterministic decoding rules, internal ambiguity and representational multiplicity can lead to subtle divergences in how semantic content unfolds. As a result, semantic evolution reflects the interaction of directed, coherence-seeking tendencies with diffusion-like variability that enables exploration of alternative semantic configurations.

Interpreting inference through this dynamical lens helps clarify several empirical phenomena observed in Transformer models. It provides an explanation for why deeper layers often exhibit increasing representational stabilization, why small perturbations in input or decoding conditions can propagate into qualitatively different outcomes, and why models may undergo abrupt transitions into incoherent or hallucination-like regimes when variability overwhelms stabilizing influences (Clark et al., 2019; Holtzman et al., 2020). These behaviors suggest that semantic evolution is governed by an underlying dynamical structure rather than by independent layer-wise computations.

This perspective establishes the conceptual foundation for the formal framework developed in subsequent chapters. Building on the interpretation of Transformer inference as a discrete approximation to continuous semantic dynamics, the CDD model characterizes the effective forces that shape semantic evolution, the stability landscape within which representations evolve, and the conditions under which coherence is maintained or degraded. Viewing inference in this way enables a unified account of semantic formation, stabilization, drift, and failure in modern language models, setting the stage for the modeling assumptions and empirical analysis that follows.

4 Model Assumptions for Semantic Dynamics

This chapter introduces three modeling assumptions that define an effective regime for describing semantic dynamics in large Transformer models. These assumptions are empirically motivated and intended to capture properties that are consistently observed across architectures, scales, and decoding conditions. From these three modeling assumptions, we derive a structural consequence concerning the selective propagation of semantic influence. This derived consequence, referred to as the Principle of Dynamic Sparsity, characterizes how semantic dynamics concentrate along a limited subset of influential directions, while the majority of perturbations remain dynamically suppressed. Together, the modeling assumptions and their derived consequence establish the conceptual and analytical foundation for the CDD framework developed in Chapter 5.

4.1 Motivation for a Modeling-Assumption Framework

Transformer-based language models operate in high-dimensional continuous spaces in which semantic meaning is not localized to discrete symbolic units, but instead emerges from distributed relational patterns. While individual architectural components such as attention heads, MLP blocks, and normalization layers can be analyzed mechanistically, their collective behavior gives rise to semantic trajectories that resemble the evolution of a dynamical system.

To describe this behavior in a principled and generalizable manner, we adopt a modeling-assumption framework with two primary objectives:

1. **Representational grounding.** To treat semantic representations as continuous state variables whose evolution can be meaningfully approximated by effective dynamical descriptions.
2. **Structural characterization.** To identify the dominant forces governing semantic evolution, namely contextual interaction, coherence-restoring tendencies, and stochastic variability.

The assumptions introduced in this chapter abstract away architectural details while retaining the essential structural features that appear to be shared across Transformer variants. Importantly, these assumptions are not claimed to hold universally for all neural architectures or training regimes. Rather, they define the scope within which the continuous-time perspective on semantic dynamics developed in Chapter 5 is intended to apply.

4.2 Assumption A1: Continuous Semantic State

Statement

We assume that the semantic content of a sequence can be represented by a state variable x taking values in a high-dimensional continuous space. In this representation, meaning is not tied to discrete symbolic identities, but is encoded in the geometric structure and relative position of representations within a semantic state space.

Rationale

Empirical analyses of Transformer representations indicate that token embeddings and contextual states occupy structured, smoothly varying regions within \mathbb{R}^d . Semantic similarity is reflected in distances and

directions in this space, and representations evolve smoothly across layers and decoding steps, forming trajectories that are well-approximated as continuous. These observations are consistent with manifold-based perspectives on linguistic meaning and motivate the use of continuous state variables as an effective description of semantic representations.

Role in the CDD Framework

Assumption A1 provides the representational foundation for the CDD framework. It allows the latent semantic state to be modeled as a time-dependent variable $x(t)$ and enables the introduction of effective potentials $\Psi(x)$, gradient-like forces $\nabla\Psi(x)$, and continuous-time approximations of semantic evolution.

4.3 Assumption A2: Context-Dependent Interaction

Statement

We assume that the semantic value of a representation is dynamically shaped by its interactions with other contextual representations. Semantic meaning is therefore relational and evolves continuously as contextual information is integrated across the sequence.

Rationale

In Transformer architectures, self-attention mechanisms compute weighted interactions between representations, with interaction strengths determined by both the input composition and learned relational structure. As a result, the same token representation may assume substantially different semantic roles depending on its surrounding context. These observations suggest that semantic meaning is not an intrinsic property of isolated representations, but emerges through ongoing relational modulation.

Role in the CDD Framework

Assumption A2 motivates modeling semantic evolution as a context-driven, time-dependent process. Within the CDD framework, the semantic trajectory $x(t)$ reflects the cumulative effect of relational interactions, providing a natural basis for describing semantic change in terms of effective dynamical equations.

4.4 Assumption A3: Coherence-Driven Drift

Statement

We assume that semantic evolution is subject to an effective coherence-restoring tendency that counteracts semantic disorganization. At the modeling level considered here, this tendency is represented as an effective coherence-restoring drift induced by the gradient of a potential-like function $\Psi(x)$ defined over the semantic state space:

$$\text{drift} = -\alpha\nabla\Psi(x).$$

The operator α represents the strength and structure of coherence enforcement and may act isotropically (scalar-valued), anisotropically (matrix-valued), or heterogeneously across layers or representational subspaces.

Rationale

Empirically, Transformer models exhibit a strong preference for coherent and contextually appropriate continuations. Perturbations to the semantic state often lead to counteracting effects that restore alignment with learned patterns of meaning. Modeling these effects as a gradient-driven flow toward regions of lower Ψ provides a compact and unifying description of semantic stabilization. Allowing α to be an operator rather than a scalar reflects the heterogeneous nature of coherence pressures in practice, where different layers, attention heads, and representational subspaces may contribute unequally to semantic stabilization.

Role in the CDD Framework

Assumption A3 defines the deterministic drift component of the CDD framework. Together with Assumptions A1 and A2, it enables semantic evolution to be modeled as an effective dynamical process shaped by both contextual interactions and coherence-restoring forces acting on the semantic state space.

4.5 Proposition 1: Dynamic Sparsity as a Derived Consequence

Statement

Under Assumptions A1–A3, the effective semantic dynamics exhibit a form of *dynamic sparsity*: at any given stage of semantic evolution, only a limited subset of computational components exerts a non-negligible influence on the trajectory of the semantic state. More precisely, for any component i whose contribution to the semantic update satisfies $\|\Delta x_i\| \leq \varepsilon_c$, and under mild smoothness assumptions on the semantic update and output maps, suppressing such components induces only a second-order change in the output distribution, in the sense that

$$D_{\text{KL}}(p_{\text{orig}} \| p_{\text{pruned}}) = O(\varepsilon_c^2),$$

where p_{orig} and p_{pruned} denote the output distributions with and without the contribution of component i , respectively. (Here, a “component” refers to an individual interaction contributing additively to the semantic update, such as an attention-weighted value term in Transformer self-attention.)

This proposition relies on two mild structural assumptions. First, the semantic update map and the output map are assumed to be Lipschitz-continuous in their arguments, ensuring that small perturbations in the latent state propagate in a controlled manner. Second, attention contributions enter linearly into the semantic update prior to non-linear transformations, as is the case in standard Transformer architectures. Under these conditions, low-magnitude contributions contract under the dynamics in norm and remain negligible in the induced output distribution.

Derivation and Empirical Support

Assumptions A1–A3 jointly imply that semantic evolution is continuous, relational, and governed by a coherence-restoring drift. Within such a regime, the influence of individual components varies smoothly with their effective contribution and vanishes continuously as that contribution approaches zero. A formal perturbation bound showing that sufficiently low-impact interactions induce only second-order changes in the output distribution is provided in Appendix A.1.3.

This structural consequence is corroborated empirically by local attention pruning experiments. Suppressing attention interactions whose inferred contribution falls below a small threshold leads to negligible changes in the output distribution, as measured by divergence-based criteria. This pattern is consistently observed across a wide range of thresholds and input prompts, as confirmed in Experiment III.

We emphasize that Proposition 1 provides a conservative sufficient condition for semantic stability; empirically, the observed invariance persists far beyond the minimal regime required by the theory.

Role in the CDD Framework

Dynamic sparsity explains why Transformer inference often evolves on a substantially reduced effective semantic subspace despite the large number of parameters and components involved. Within the CDD framework, it clarifies how semantic coherence and stability are mediated by a structured subset of influential pathways, providing a conceptual basis for understanding sparsity, interpretability, and efficiency phenomena observed in practice.

4.6 Synthesis: Foundations of Coherence–Diffusion Dynamics

Assumptions A1–A3 define an effective modeling regime for semantic evolution in Transformer-based language models.

- **Assumption A1** treats semantic meaning as a continuous state evolving in a high-dimensional representation space.
- **Assumption A2** characterizes semantic change as a context-dependent process driven by relational interactions among representations.
- **Assumption A3** models semantic stabilization through an effective coherence-restoring drift, represented as a gradient-driven tendency over the semantic state space.

Within this modeling regime, Proposition 1 emerges as a derived structural consequence: semantic dynamics exhibit *dynamic sparsity*, in which only a limited subset of computational components exerts a non-negligible influence on semantic evolution at any given stage. Together, the modeling assumptions and their derived consequence justify the continuous-time formulation developed in Chapter 5, where semantic evolution is described as the interplay between coherence-seeking drift and diffusion-induced variability.

5 Continuous-Time Interpretation of Semantic Dynamics in Transformer Models

The CDD framework adopts an effective dynamical perspective on Transformer inference, interpreting discrete layer-wise updates as approximating an underlying continuous-time semantic process. While Transformer models operate through a finite sequence of residual updates indexed by layer depth, the latent representations generated during inference exhibit regularity, gradual drift, and stability patterns that are naturally captured using the language of dynamical systems.

This chapter develops the theoretical foundations of this interpretation by introducing an effective continuous-time description of semantic evolution. Within this perspective, semantic states are modeled as trajectories evolving in a high-dimensional state space under the combined influence of coherence-restoring tendencies and stochastic variability. The formulation presented here is intentionally abstract and agnostic to specific architectural components. Concrete connections to self-attention mechanisms, residual streams, and normalization effects are developed in Chapter 6.

5.1 Motivation for a Dynamical Interpretation

Transformer-based language models exhibit a range of behaviors that admit a natural interpretation in dynamical terms. As information propagates through successive layers, internal representations often display progressive stabilization, forming increasingly coherent semantic states. At the same time, these models remain sensitive to small perturbations introduced by prompt variations or decoding stochasticity, which can redirect subsequent semantic evolution in noticeable ways. As stochastic influence increases, models may transition into regimes characterized by reduced coherence and erratic output. Such behaviors are difficult to capture using purely combinatorial or symbolic views of inference. Instead, they suggest interpreting inference as the evolution of a state vector subject to both stabilizing and destabilizing influences.

Within the CDD framework, this evolution is modeled using an effective stochastic differential equation,

$$dx(t) = -\alpha \nabla \Psi(x(t)) dt + \sigma dW(t),$$

where $x(t)$ denotes the semantic state in a continuous representation space, $\Psi(x)$ is an effective instability potential whose gradient encodes coherence-restoring tendencies, α represents a coherence enforcement operator that may be anisotropic, σ controls the strength of stochastic variability, and $W(t)$ denotes a standard Wiener process. This equation is not claimed to describe the exact internal computations of Transformer models. Rather, it serves as an effective description that captures qualitative regularities in semantic evolution observed across architectures and inference settings.

The remainder of this chapter motivates and justifies this formulation at the modeling level, clarifying the assumptions under which a continuous-time drift–diffusion perspective provides a useful and coherent abstraction of Transformer inference dynamics.

5.2 Semantic States as Continuous Variables on an Effective Manifold

Although Transformer inference proceeds through discrete layer-wise updates, hidden representations themselves reside in continuous vector spaces \mathbb{R}^d . Each forward pass modifies the semantic representation incrementally, yielding a sequence of states

$$x_0 \rightarrow x_1 \rightarrow x_2 \rightarrow \cdots \rightarrow x_L,$$

which, when viewed at sufficient depth, can be interpreted as approximating a continuous trajectory. From a modeling perspective, this motivates treating semantic evolution as taking place on an effective semantic manifold $\mathcal{M} \subset \mathbb{R}^d$.

Within this space, local geometric relationships encode semantic similarity, while smoothness reflects the continuity of representational change across layers. Empirical studies indicate that the intrinsic dimensionality of this effective manifold is often substantially lower than the ambient dimension, reflecting redundancy, compression, and progressive concentration of semantic information. Interpreting semantic evolution as a continuous flow on \mathcal{M} thus provides a natural foundation for the drift–diffusion perspective adopted in the CDD framework.

5.3 Discrete Residual Updates as an Effective Euler-Type Approximation

Let x_t denote the semantic state after the t -th layer update. Empirically, residual updates take the form

$$x_{t+1} = x_t + F(x_t),$$

where $F(x_t)$ summarizes the cumulative effect of attention, feed-forward transformations, and normalization operations within a layer. This update rule admits a close analogy with forward Euler schemes used to approximate continuous-time dynamical systems, with layer depth playing the role of a discretized time parameter.

At a coarse modeling level, we may further decompose the residual update as

$$F(x_t) \approx -\alpha \nabla \Phi(x_t) + \sigma \xi_t + r_t,$$

where the first term captures effective coherence-restoring drift, the second term aggregates stochastic variability, and r_t denotes higher-order, non-gradient, and anisotropic contributions that are not explicitly modeled within the drift–diffusion approximation. This decomposition emphasizes that Transformer updates do not implement exact gradient descent, but nevertheless exhibit a dominant drift–diffusion structure at the level of semantic dynamics.

Under this analogy, layer depth can be interpreted as a discretized time parameter, with each layer corresponding to a small step along an underlying semantic trajectory. This interpretation does not assert that Transformer updates implement an explicit numerical integration scheme; rather, it highlights a structural similarity between residual accumulation and incremental state evolution that motivates a continuous-time approximation at the modeling level.

When sources of stochasticity are taken into account, this Euler-type analogy naturally extends to stochastic update rules resembling Euler–Maruyama discretizations, providing a conceptual bridge to stochastic differential equation descriptions of semantic dynamics.

5.4 Stochasticity and the Emergence of Diffusion

Transformer inference incorporates variability from multiple sources, each contributing to uncertainty in semantic evolution. Randomness introduced by token sampling affects the direction of generation, while entropy in the output distribution governs the breadth of semantic exploration. Internal variability in residual activations and ambiguity among multiple plausible continuations further introduce effective noise into the representational dynamics.

Taken together, these effects motivate modeling semantic evolution using an effective stochastic description of the form

$$dx(t) = G(x(t)) dt + \sigma dW(t),$$

where σ captures the aggregate strength of stochastic influence and $W(t)$ denotes a standard Wiener process. This formulation is not intended as an exact description of internal computations, but as a compact representation of qualitative regularities observed in generative behavior, including divergence under high entropy and sensitivity to small perturbations.

Within the CDD framework, the drift term $G(x)$ is interpreted as encoding coherence-restoring tendencies, which are subsequently modeled through the effective potential formulation $-\alpha\nabla\Psi(x)$ introduced in the following sections.

5.5 Instability Potential Ψ as an Effective Semantic Energy

The introduction of a potential function $\Psi(x)$ is motivated by consistent empirical patterns observed during Transformer inference. Coherent predictions tend to be associated with representational states exhibiting low surprisal, while hallucination-like behavior often arises in regions where loss or uncertainty increases sharply. Layer-wise analyses further suggest that representations exhibit stabilizing tendencies, concentrating around certain regions of the representation space over depth.

From a modeling perspective, these observations motivate interpreting $\Psi(x)$ as an effective semantic energy landscape, in the sense that

- low values of Ψ are associated with coherent and stable semantic configurations,
- high values of Ψ are associated with unstable or incoherent configurations.

The model does not explicitly compute or store Ψ . Rather, Ψ functions as an interpretive construct inferred from empirical regularities, analogous to effective potentials in physics. Within this interpretation, coherence-restoring tendencies can be modeled by a gradient-like drift term of the form

$$G(x) = -\alpha\nabla\Psi(x),$$

where α denotes an effective coherence enforcement operator.

Consistent with the assumptions outlined in Chapter 4, α is not restricted to be a scalar quantity. It may act anisotropically, vary across layers, or operate on specific representational subspaces, reflecting the heterogeneous nature of coherence enforcement in Transformer architectures. At the modeling level, we distinguish between a global effective coherence parameter α_{eff} , which we probe via learning rate variations in Experiment I, and local coherence operators $\alpha_l(x)$ that reflect layer- and state-dependent coherence enforcement, as suggested by attention concentration and residual gradient alignment.

For the purposes of the CDD framework, we assume that the effective instability potential Ψ is at least once differentiable with locally Lipschitz gradient on the effective semantic manifold \mathcal{M} , allowing us to define a well-behaved drift field $-\alpha\nabla\Psi(x)$.

5.6 Drift–Diffusion Description of Semantic Evolution

Combining the effective drift and diffusion components yields the stochastic differential equation

$$dx(t) = -\alpha\nabla\Psi(x(t)) dt + \sigma dW(t),$$

which serves as a compact modeling description of semantic evolution within the CDD framework.

Different qualitative regimes of behavior can be associated with different relative magnitudes of the coherence and diffusion terms. In particular, empirical observations are consistent with the following interpretations:

- when coherence-restoring effects dominate stochastic variability, semantic trajectories tend to stabilize,
- when coherence and diffusion are comparable, trajectories exhibit controlled exploration while remaining semantically coherent,
- when diffusion dominates, trajectories display increased drift and instability, often manifesting as hallucination-like behavior,
- when coherence enforcement is excessively strong, trajectories may exhibit oscillatory or overshooting behavior.

These regime distinctions are intended as qualitative characterizations rather than as sharp phase boundaries, and they align with well-studied behaviors in nonlinear stochastic dynamical systems.

5.7 Fokker–Planck Perspective on Representational Stabilization

While the stochastic differential equation describes individual semantic trajectories, the collective behavior of ensembles of such trajectories can be described, at an effective level, by a corresponding Fokker–Planck equation,

$$\frac{\partial \rho(x, t)}{\partial t} = \nabla \cdot (\alpha \rho(x, t) \nabla \Psi(x)) + \frac{\sigma^2}{2} \nabla^2 \rho(x, t),$$

where $\rho(x, t)$ denotes a probability density over semantic states.

Within this effective description, several qualitative expectations arise. Moderate levels of diffusion lead to concentration of probability mass near low- Ψ regions, while larger diffusion produces broader and less stable distributions. Over extended inference or depth, coherence-restoring tendencies bias the distribution toward relatively stable regions of the representation space. Stationary or slowly varying solutions of this equation may be interpreted as reflecting attractor-like tendencies in an effective sense encoded by the model’s learned representations.

These qualitative predictions are consistent with empirical observations that deeper layers often exhibit progressive stabilization despite the presence of local stochastic variability.

5.8 Ensemble-Based Interpretation of Semantic Stability

The drift–diffusion and Fokker–Planck formulations introduced above describe, respectively, the microscopic evolution of individual semantic trajectories and the macroscopic evolution of probability densities over semantic states. However, Transformer representations do not evolve along a single deterministic path. Instead, the model’s architecture naturally induces a rich ensemble of latent semantic trajectories, even for a single input sequence.

This ensemble structure arises from several architectural and computational features: residual connections superpose multiple representational directions; multi-head attention aggregates heterogeneous relational signals; softmax-based attention introduces intrinsic entropy; normalization layers amplify small representational differences; and, when present, sampling-based decoding introduces additional stochasticity. As a result, the hidden state at each layer implicitly encodes multiple potential semantic directions rather than a single, deterministic trajectory.

From this perspective, the Fokker–Planck equation provides an effective description of how these latent trajectories collectively evolve. While an individual realization $x(t)$ follows the stochastic differential equation described earlier, the ensemble of all such realizations is characterized by the evolving density $\rho(x, t)$. Coherence-restoring drift biases this density toward low- Ψ regions, while diffusion spreads probability mass across the representation space. When drift dominates, the ensemble probability mass becomes biased toward relatively coherent regions; when diffusion dominates, the ensemble becomes more dispersed and unstable.

This ensemble concentration has direct implications for individual trajectories. Although a single trajectory does not necessarily converge to a fixed point due to persistent diffusion, the collective bias of $\rho(x, t)$ toward

stable regions induces an effective stabilizing influence. Individual realizations tend to remain near these coherent basins, exhibiting progressive stabilization over depth. Moreover, as the ensemble collapses into lower-dimensional stable regions, many representational directions become dynamically irrelevant, giving rise to the dynamic sparsity behavior formalized in Proposition 1.

In summary, the ensemble-based interpretation provides a conceptual bridge between the microscopic SDE description and the macroscopic Fokker–Planck dynamics. It clarifies why semantic stability emerges robustly in Transformer models despite local stochastic variability, and it highlights the collective nature of coherence formation within the CDD framework.

5.9 Theoretical Implications of the SDE–Fokker–Planck Framework

The drift–diffusion formulation introduced in this chapter yields a set of qualitative, empirically testable implications for semantic evolution in Transformer models:

1. **Stabilization toward coherent regions.** When coherence-restoring effects dominate stochastic variability, semantic trajectories tend to concentrate in low- Ψ regions associated with stable and coherent representations.
2. **Regime shifts under increased diffusion.** As stochastic influence becomes comparable to or exceeds coherence enforcement, semantic trajectories exhibit increased drift away from coherent regions, corresponding to the emergence of hallucination-like behavior.
3. **Overshooting under excessive coherence enforcement.** When coherence-restoring forces are excessively strong, trajectories may display oscillatory or unstable behavior consistent with overshooting effects in nonlinear dynamical systems.
4. **Emergence of dynamic sparsity.** As trajectories concentrate within stable regions, the effective dimensionality of semantic evolution contracts, causing many representational components to become dynamically irrelevant. This aligns with the Dynamic Sparsity Principle formalized in Proposition 1, which holds that sparsity arises whenever semantic influence collapses onto a restricted set of directions.

These implications motivate the empirical investigations presented in Chapter 7, where controlled experiments examine the extent to which observed Transformer behavior aligns with the qualitative predictions of the CDD framework.

5.10 Summary

This chapter developed an effective continuous-time interpretation of semantic evolution in Transformer-based language models. By drawing structural analogies between residual updates and Euler-type discretizations, between internal variability and diffusion, and between surprisal-driven instability and an effective semantic potential, the CDD framework provides a unified modeling description of semantic dynamics. The resulting drift–diffusion and Fokker–Planck formulations are not proposed as exact descriptions of Transformer computation. Rather, they serve as interpretive abstractions that capture qualitative regularities in stability, variability, and representational concentration observed during inference.

The next chapter grounds these abstract constructs in concrete architectural mechanisms, showing how coherence enforcement, stochastic variability, and dynamic sparsity arise from measurable components of Transformer models. Together, these chapters bridge theoretical modeling and empirical observation, establishing a coherent foundation for the empirical analyses that follow.

Table 1: Correspondence between CDD core quantities and transformer components

CDD Quantity	Transformer Observable	Interpretation
α	Attention sharpness, gradient alignment, residual update strength	Magnitude and effective direction of coherence-restoring dynamics
σ	Softmax temperature, entropy, sampling randomness	Magnitude of stochastic semantic exploration
$\Psi(x)$	Surprisal (NLL), loss spikes, entropy irregularities	Degree of semantic instability or incoherence
ε -threshold	Very low attention weights / negligible gradients	Operational boundary of effective irrelevance in semantic dynamics

6 Architectural Realization of Coherence–Diffusion Dynamics in Transformer Models

Having established in Chapter 5 that semantic evolution in LLMs can be effectively interpreted through a drift–diffusion framework, we now examine how the abstract components of CDD relate to concrete mechanisms within Transformer architectures. Specifically, we discuss how the coherence operator α , diffusion coefficient σ , instability potential Ψ , and the dynamic sparsity principle, operationalized via an ε -threshold, admit systematic interpretations in terms of measurable internal signals.

In contrast to the previous chapter, which focused on continuous-time modeling at an abstract level, this chapter grounds the CDD framework in architectural structure. Rather than claiming a literal implementation of the drift–diffusion equations, we show how key elements of semantic dynamics are reflected in attention patterns, softmax entropy, residual stream accumulation, and empirically observed sparsity effects. This mapping clarifies how coherence enforcement, stochastic variability, and functional irrelevance emerge from standard Transformer operations during inference.

Throughout this chapter, the term “drift” refers exclusively to an effective modeling construct that summarizes coherence-restoring tendencies at an abstract dynamical level. This notion should not be conflated with empirical layerwise displacement measures computed directly from hidden-state trajectories.

6.1 Parameter Correspondence Between CDD and Transformer Mechanisms

The central quantities of the CDD framework admit consistent operational interpretations in terms of Transformer architecture. As summarized in Table 1, each theoretical component can be associated with observable internal signals that influence semantic evolution during inference. These correspondences are not intended as exact or mechanistic equivalences. Rather, they provide a principled way to relate abstract dynamical constructs to quantities that can be measured, manipulated, and analyzed empirically in real models.

By establishing these associations, the dynamical state of a Transformer can be partially diagnosed and monitored during inference, enabling empirical evaluation of coherence, diffusion, and sparsity effects without assuming that the model explicitly computes the CDD quantities themselves.

6.2 Instability Potential Ψ as a Loss-Based Semantic Signal

Within the CDD framework, $\Psi(x)$ is interpreted as an effective measure of semantic instability along a latent semantic trajectory. In Transformer models, quantities such as token-level negative log-likelihood (NLL), surprisal, and related indicators including loss spikes and abrupt changes in gradient behavior consistently correlate with transitions between stable and unstable semantic regimes. These empirical patterns motivate treating Ψ as an effective potential that captures the model’s tendency toward or away from semantic coherence.

Operationally, $\Psi(x)$ is not computed directly by the model. Instead, we treat token-level surprisal, layer-wise loss fluctuations, and entropy-based irregularities observed along a semantic trajectory as partial observables that correlate with increases in Ψ . These quantities serve as empirical proxies for semantic instability rather than exact evaluations of the potential itself. In particular, rising surprisal or abrupt loss variations signal that the semantic state has entered a region of higher instability, while sustained low surprisal and smooth layer-wise behavior are consistent with trajectories spending more time in relatively low- Ψ coherent regions.

This perspective allows Ψ to function as a meaningful diagnostic signal without requiring an explicit closed-form definition. By grounding the instability potential in observable quantities that are readily accessible during inference, CDD bridges the gap between abstract dynamical modeling and measurable internal behavior in Transformer architectures.

6.2.1 Low- Ψ Regions: Coherent Semantic States

Low- Ψ regions are associated with semantic states in which next-token probability mass is strongly concentrated around contextually appropriate continuations, gradients propagate smoothly across layers, and attention patterns remain relatively stable with depth. Empirically, deeper layers often exhibit decreasing surprisal on well-posed inputs, reflecting progressive suppression of semantic dispersion and increased residence in coherent regions of the representation space. From the CDD perspective, such behavior is consistent with trajectories concentrating in low-instability regions, where semantic representations are internally consistent and predictive uncertainty is minimal.

6.2.2 High- Ψ Regions: Semantic Instability and Hallucination

High values of Ψ are empirically associated with sudden increases in token-level surprisal, unstable shifts in attention focus, and irregular or rapidly varying gradient patterns. These signals tend to coincide with periods of semantic drift, during which the model explores representational regions that are weakly constrained by the input context. Such behavior often precedes or accompanies hallucination-like output, suggesting that Ψ -aligned observables can serve as informative indicators of emerging semantic instability. In this sense, Ψ provides a useful lens for monitoring coherence during generation, rather than a deterministic predictor of failure.

6.3 Coherence Operator α as Attention-Constrained Gradient Alignment

The coherence operator α serves as an effective descriptor of the strength with which a model can be interpreted as promoting semantic consistency during inference. Architecturally, α is not associated with a single parameter. Instead, it reflects the combined effect of attention concentration and the alignment of residual updates with directions that reduce predictive uncertainty.

6.3.1 Attention Concentration as an Indicator of Coherence Strength

Self-attention updates token representations according to

$$h'_i = \sum_j a_{ij} V_j, \quad a_{ij} = \text{softmax}(q_i k_j^\top).$$

When attention distributions are sharply concentrated, representational updates are dominated by a small subset of contextually relevant tokens. From the CDD perspective, such concentration is associated with stronger coherence-restoring tendencies, consistent with a higher effective value of α . Conversely, diffuse attention distributions can spread semantic influence across many tokens, potentially weakening directional coherence and making semantic trajectories more susceptible to drift.

Attention concentration therefore serves as a useful empirical indicator of coherence strength. Sustained sharpness is typically associated with semantically coherent evolution with reduced variability, while sudden diffusion often coincides with reduced coherence and increased sensitivity to stochastic variability.

From the CDD perspective, the self-attention mechanism contributes not only to the magnitude of coherence enforcement but also to the direction of the effective drift field. By selectively aggregating context directions

that tend to reduce semantic instability, attention patterns can be interpreted as indirectly biasing the drift term $-\alpha\nabla\Phi(x)$ in the direction of more coherent regions of the latent manifold.

6.3.2 Residual Gradient Alignment

Residual pathways accumulate representational changes across layers according to

$$x_{t+1} = x_t + \Delta x_t.$$

When the update Δx_t is directionally aligned with changes that reduce predictive instability, such as movements toward lower-loss or lower-surprisal regions, semantic coherence tends to be reinforced. From the CDD perspective, this alignment is consistent with a drift-like effect that biases semantic evolution toward lower- Ψ regions.

Importantly, this interpretation does not imply that residual updates explicitly perform gradient descent on Ψ . Rather, it reflects an aggregate tendency for updates to align with coherence-restoring directions when attention, normalization, and residual scaling are well calibrated. Conversely, under conditions such as excessively high learning rates, miscalibrated normalization, or overly sharpened attention, residual updates may overshoot relatively stable regions under certain conditions. Such behavior manifests as oscillatory or unstable semantic trajectories, a qualitative pattern consistent with the high-gain or strongly driven regimes described in the CDD framework.

Residual gradient alignment therefore provides a complementary architectural correlate of the coherence operator α , alongside attention concentration, illustrating how coherence enforcement can emerge from multiple interacting mechanisms within Transformer models.

6.3.3 Attention–MLP Functional Decomposition in Transformer Dynamics

Although the aggregate hidden-state update of a Transformer layer can be written compactly as

$$f(x) = f_{\text{attn}}(x) + f_{\text{MLP}}(x),$$

the two components play fundamentally different dynamical roles within the CDD framework.

Self-attention: Interaction-driven Dynamics Self-attention implements **interaction-driven dynamics**, mediating information flow across tokens and determining how contextual signals shape the direction of semantic evolution. Because attention selectively amplifies or suppresses cross-token influences, it governs the directional component of coherence enforcement and is a major contributor to dynamic sparsity and context-dependent drift modulation.

MLP Block: Token-local Stabilizing Update In contrast, the MLP block contributes a **token-local stabilizing update** that operates independently on each token representation. Rather than mixing information across positions, the MLP refines and stabilizes token-internal semantic structure through non-linear expansion and contraction. This token-local stabilizing effect tends to bias representations toward relatively coherent local regions of the latent space, providing a stabilizing force that complements the context-dependent interactions induced by attention.

Synthesis and Residual Connections Residual connections combine these two effects, producing the overall update field $f(x)$. Within the CDD interpretation:

- **Attention** shapes the interaction geometry of the drift field.
- **MLP** shapes its intrinsic stability structure.

Explicitly distinguishing these roles clarifies how coherence emerges from the interplay of cross-token interactions and token-internal refinement, and why coherence-related phenomena, such as attractor formation, dynamic sparsity, and ε^2 -stability, can be naturally interpreted as arising from the architecture.

6.4 Diffusion σ as Stochastic Variability in Decoding

Within the CDD framework, diffusion σ captures stochastic influences that introduce variability into semantic evolution. In Transformer models, such variability arises from both explicit decoding choices and intrinsic architectural factors. Accordingly, we treat σ as an effective scale parameter for stochastic semantic variability. Operationally, σ is linked to decoding temperature, output entropy, and logit-level uncertainty, and reflects the typical magnitude of stochastic perturbations acting on the latent semantic state during inference.

The qualitative effects of diffusion vary across different dynamical regimes. When stochastic influence is minimal, generation becomes quasi-deterministic and may exhibit rigidity or repetition. At moderate levels, diffusion enables semantic exploration while preserving coherence. When stochastic variability becomes dominant, whether due to high entropy, ambiguous context, or elevated sampling temperature, semantic trajectories become increasingly unstable, often resulting in hallucination-like or fragmented output.

6.4.1 Softmax Temperature and Entropy

Token generation is typically performed by sampling from a softmax distribution,

$$p_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)},$$

where z_i denote logits and T is the sampling temperature. Increasing the temperature T produces higher-entropy output distributions, expanding the range of plausible next-token choices.

From the CDD perspective, higher entropy is associated with stronger stochastic influence on semantic trajectories, corresponding to a larger effective diffusion strength σ . Conversely, low temperatures yield sharply peaked distributions with reduced entropy, aligning with lower effective diffusion.

Among decoding mechanisms, temperature-controlled sampling provides a particularly clear operational handle on stochastic variability, making it a useful empirical proxy for studying diffusion effects in semantic dynamics.

6.4.2 Model-Inherent Variability

Even in the absence of explicit temperature scaling, Transformer inference exhibits non-negligible stochasticity due to internal factors. Competition among nearly equal logits renders small numerical differences consequential for token selection. Residual accumulation can amplify minor activation fluctuations across layers, and when multiple semantic continuations are similarly plausible, local uncertainty increases.

Together, these effects introduce a baseline level of stochastic variability even under nominally deterministic decoding procedures, such as greedy decoding. From the CDD perspective, this intrinsic variability contributes to a non-zero effective diffusion term.

The transitions between the behavioral regimes described in this section are examined empirically in Experiment II (Chapter 7), which demonstrates how controlled increases in stochastic influence systematically reshape generative behavior in ways consistent with the diffusion-based interpretation.

6.5 Dynamic Sparsity, ε -Thresholds, and Attention Patterns

The Dynamic Sparsity Principle (Proposition 1) states that, within the CDD framework, only a subset of representational interactions meaningfully influences semantic evolution. Interactions whose effective contributions are sufficiently small exert negligible influence on the semantic state $x(t)$. In Transformer architectures, this behavior naturally manifests through the structure of attention distributions.

6.5.1 Low-Impact Attention Interactions as Dynamical Irrelevance

For a given attention head, many attention coefficients contribute only marginally to the semantic update. When attention interactions are ranked by their contribution and pruned according to a cumulative tail-mass criterion, a substantial fraction of interactions can be removed without altering the output distribution.

As demonstrated in Experiment III, pruning low-impact attention interactions at the final transformer layer produces negligible changes in next-token probability distributions, as quantified by near-zero KL divergence. From the CDD perspective, this indicates that such interactions are dynamically irrelevant for semantic evolution, even though they remain present in the full computational graph.

6.5.2 Effective Dimensionality Reduction and Interpretive Implications

The concentration of semantic influence implies that, at a given stage of inference, semantic evolution is effectively governed by a reduced subset of active interactions. Consequently, approximating or suppressing a large portion of low-impact attention interactions produces minimal observable changes in model behavior.

Within the CDD framework, this phenomenon is interpreted as a consequence of drift–diffusion dynamics. As coherence-restoring forces guide semantic trajectories toward stable regions, the effective dimensionality of semantic computation contracts. Dynamic sparsity thus emerges not as an ad hoc engineering heuristic, but as a structural feature of semantic evolution in Transformer models.

6.6 Layer-Wise Evolution as an Effective Discretized Dynamical Process

Rather than interpreting Transformer layers as literal time steps of a stochastic differential equation, the CDD framework adopts an effective modeling perspective in which layer-wise updates admit a discretized dynamical interpretation. At this level of abstraction, each Transformer block can be viewed as contributing an incremental update to the semantic state that is qualitatively consistent with drift–diffusion dynamics.

Absorbing step-size factors into the effective parameters for notational simplicity, this interpretation can be summarized by the update

$$x_{t+1} = x_t - \alpha \nabla \Psi(x_t) + \sigma \xi_t,$$

where ξ_t represents aggregated stochastic influence arising from sampling, entropy-driven variability, and internal representational ambiguity.

Within this interpretive framework, attention and feed-forward mechanisms contribute to directional, coherence-restoring tendencies, while softmax entropy, sampling randomness, and internal variability introduce stochastic perturbations. Progression through successive layers then corresponds to an ordered sequence of incremental semantic updates, which collectively approximate a drift–diffusion-like evolution in representation space.

This formulation is not intended to assert that Transformers explicitly solve a stochastic differential equation. Rather, it highlights a structural correspondence between standard Transformer updates and the qualitative behavior captured by the CDD framework.

6.7 Architectural Implications of the CDD Mapping

The architectural correspondences summarized in Table 2 motivate a set of empirically testable implications. By manipulating architectural or decoding parameters associated with coherence, stochastic variability, or sparsity, one can induce systematic changes in semantic behavior that align with the qualitative expectations of the CDD framework.

Chapter 7 presents controlled experiments examining these implications, demonstrating that observed model behavior is broadly consistent with the dynamical interpretation developed here.

Table 2: CDD–Transformer structural mapping and its predictions

CDD Condition	Transformer Architectural Behavior
High α	Overshooting updates, oscillatory instability, brittle behavior
Low α	Diffuse attention, weak gradient alignment, semantic drift
High σ	Diffusion-dominated behavior; increased hallucination risk
Low σ	Highly deterministic or repetitive generation
Influence $< \varepsilon$	Removable interactions; sparsity without semantic loss
Low Ψ	Bias toward coherent semantic regions
High Ψ	Elevated risk of semantic collapse or hallucination

6.8 Summary

This chapter connected the abstract constructs of the CDD framework to concrete, measurable mechanisms within Transformer architectures. Attention concentration and residual update alignment were shown to provide operational indicators of coherence enforcement, corresponding to the effective role of α . Stochastic variability arising from softmax entropy, sampling temperature, and internal ambiguity was interpreted as contributing to diffusion effects associated with σ . Loss-based signals such as surprisal and instability spikes were identified as informative proxies for the effective potential $\Psi(x)$. Dynamic sparsity emerged naturally through pervasive attention and gradient sparsity, operationalized via ε -thresholds as formalized in Proposition 1.

Taken together, these interpretations suggest that Transformer models exhibit structured semantic dynamics that can be analyzed through a coherent drift–diffusion perspective. Rather than treating Transformers as opaque statistical systems, the CDD framework offers a principled lens for understanding how coherence, variability, and functional relevance interact during inference, enabling more systematic analysis and control of semantic behavior.

7 Experiments

We conduct three experiments to examine the empirically testable implications of the CDD framework. Each experiment focuses on a distinct aspect of the framework: (i) the suppression of the instability potential Ψ and the resulting stabilization of semantic trajectories under coherence–diffusion dynamics; (ii) the regime-dependent interaction between the coherence parameter α and diffusion scale σ associated with hallucination-like behavior; and (iii) the derived principle of dynamic sparsity, which governs the effective dimensionality of attention contributions as trajectories converge.

All experiments are conducted using standard Transformer-based language models and publicly accessible evaluation settings, facilitating transparency and reproducibility.

7.1 Experiment I: Influence of the Coherence Parameter α on Semantic Stability

To examine the role of the coherence parameter α within the CDD framework, we perform controlled fine-tuning experiments on GPT-2 small, medium, and large models. Here, learning rate is used as an operational proxy for the global effective coherence parameter α_{eff} . This choice provides a practical mechanism for modulating the effective strength of coherence-driven drift during fine-tuning, without attempting to model the full set of local, layer- and state-dependent coherence operators acting within the Transformer.

While α is introduced in CDD as an effective parameter characterizing coherence restoration in semantic dynamics, adjusting the learning rate provides a way to probe how variations in drift strength influence stability and convergence behavior in practice.

Within the CDD framework, three qualitative coherence regimes are expected to arise as α varies:

1. Low α : insufficient coherence restoration, leading to weak suppression of instability and elevated values of Ψ .
2. Intermediate α : balanced drift–diffusion dynamics, associated with reduced instability and coherent semantic behavior and minimal Ψ .
3. High α : excessive drift strength, resulting in oscillatory or unstable behavior and increasing Ψ . These behaviors should be interpreted as dynamical regime effects rather than as evidence of fixed semantic attractors.

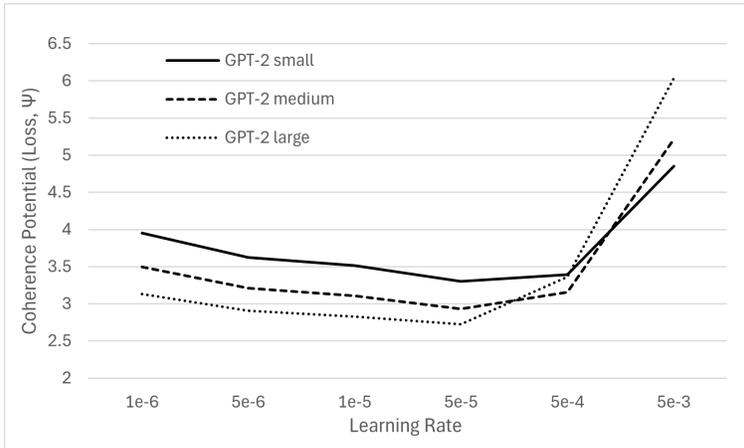


Figure 1: Effect of the coherence parameter α on the final instability potential Ψ across three model scales.

7.1.1 Static Analysis: Relationship Between α and Final Instability Potential Ψ_{final}

Figure 1 illustrates how the final instability potential Ψ_{final} (approximated by the final training loss) varies with α . For all three model sizes, very small values of α ($1e-6$, $5e-6$) are associated with elevated Ψ , consistent with weak coherence forces unable to counteract semantic drift. As α increases into an intermediate range ($1e-5$ to $5e-5$), all models attain their lowest observed instability values.

Notably, the location of this minimum is similar across model scales: GPT-2 small, medium, and large exhibit optimal stability within the same α range, with GPT-2 large reaching its minimum at $\alpha = 5 \times 10^{-5}$ ($\Psi = 2.73$). This alignment suggests that the instability landscape probed by α reflects shared structural regularities of the models, as captured by the CDD framework.

Beyond this intermediate regime, further increases in α lead to degraded stability. At $\alpha = 5 \times 10^{-4}$, all models exhibit rising Ψ , indicative of overshooting behavior. When α is increased to 5×10^{-3} , training becomes unstable, with Ψ increasing sharply. This behavior is consistent with the interpretation that excessive drift strength disrupts smooth traversal of the effective potential landscape.

7.1.2 Dynamic Analysis: Influence of α on the Time-Evolution of $\Psi(t)$

To further examine the dynamical interpretation, we analyze the temporal evolution of the instability potential $\Psi(t)$ during fine-tuning for GPT-2 large (Figure 2). As α increases, the descent of Ψ from its initial value becomes progressively steeper, indicating faster convergence toward lower-instability regions.

At very small α ($1e-6$), the trajectory declines slowly and stabilizes at a relatively high $\Psi \approx 3.13$, reflecting weak directional bias. In contrast, at $\alpha = 5 \times 10^{-3}$, the trajectory exhibits pronounced oscillations and sharp increases in Ψ , indicative of unstable dynamics.

Taken together, these observations suggest that α plays a central role in shaping semantic stability and convergence behavior across model scales. While α is implemented here via the learning rate, the persistence

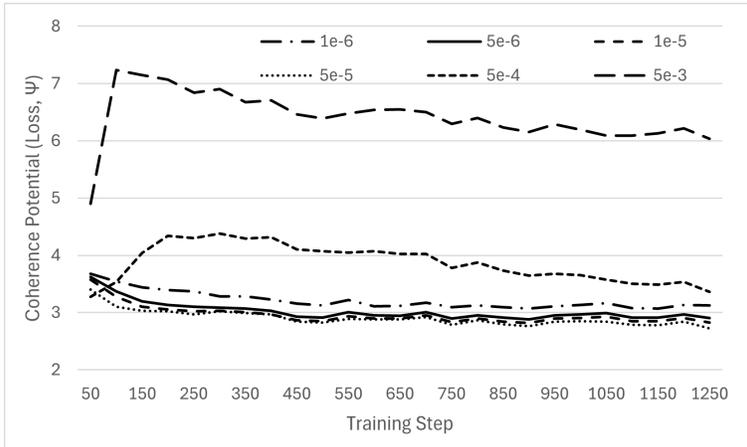


Figure 2: Analysis of the training step-by-step Ψ trajectory for the GPT-2 large model.

of similar qualitative regimes across GPT-2 variants supports the interpretation of α as an effective coherence parameter within the CDD framework, rather than a purely incidental optimization hyperparameter.

7.2 Experiment II: Diffusion-Driven Regime Shift in Semantic Coherence

To examine the role of diffusion strength σ within the CDD framework, we conduct a controlled text-generation experiment in which σ is operationally modulated via the sampling temperature T during autoregressive decoding. Since σ corresponds to the magnitude of stochastic perturbations in the effective drift-diffusion description, increasing T provides a practical means of probing how semantic coherence degrades as stochastic influence becomes stronger relative to the coherence parameter α .

At a coarse modeling level, the coherence-driven drift term tends to reduce the instability potential Ψ at a rate proportional to $\alpha\|\nabla\Psi\|^2$, while stochastic diffusion contributes an opposing increase at a rate proportional to $\sigma^2\Delta\Psi$. This competition suggests a critical balance when these contributions become comparable, providing a dynamical basis for the phase-transition-like qualitative regime shift observed as σ/α varies. A heuristic characterization of this coherence-diffusion balance, leading to an effective critical diffusion-to-coherence ratio, is discussed in Appendix A.1.2.

While σ may arise from both model-internal sources and sampling procedures (as discussed in Chapter 6), this experiment isolates the sampling-induced component by varying the decoding temperature T . This setup allows us to examine how increasing stochasticity reshapes semantic trajectories, without introducing additional sources of model variability. In this setting, the growth of local output entropy provides a convenient empirical signal for tracking changes in semantic coherence as the effective ratio σ/α increases. Accordingly, the resulting behavioral changes should be interpreted not as abrupt discontinuities, but as qualitative regime shifts reflecting the gradual dominance of diffusion over coherence, as anticipated by the CDD drift-diffusion formulation.

We generated five samples for each of four temperature settings ($T = 0.1, 0.7, 1.0, 1.5$) using GPT-2 Large and a fixed prompt describing the assumptions of the CDD framework (see Supplementary Material A.3 for the prompt and generated samples). All other decoding parameters were held constant, ensuring that T was the primary factor influencing stochasticity during generation.

The resulting continuations were analyzed qualitatively with respect to topical consistency, logical structure, contextual stability, and degree of semantic drift. As the temperature increased, the model exhibited a systematic progression of qualitative behaviors that is consistent with the CDD interpretation of diffusion-driven semantic dynamics. At low diffusion ($T = 0.1$), outputs were nearly deterministic and highly repetitive, remaining tightly confined to the prompt’s topical structure. This behavior suggests strong dominance of coherence-restoring forces, with the semantic state remaining confined to a relatively coherent region of the

Table 3: A qualitative summary of generative behaviors under increasing decoding temperature

Temperature	Qualitative Behavior	Semantic Stability		Short Example
0.1	Highly repetitive, almost deterministic; low novelty	Low-Entropy Convergence	Coherent	“evolution of a semantic system is a process of ...”
0.7	Diverse and creative while remaining on-topic	Moderate-Entropy Creative Exploration	Cre-	“degree of semantic similarity between two words ...”
1.0	Early signs of topic drift; weakening logical links	Rising-Entropy Regime	Transition	“level of detail used by organisms ... physicist Michal Kalecki (1998) ...”
1.5	Semantic collapse; hallucinatory combinations	High-Entropy Divergence	Hallucina-	“all three dimensions in biological systems ... (i.e, the ‘snowball theorem’) ...”

representation space. At moderate diffusion ($T = 0.7$), the model generated coherent yet diverse continuations. The text remained on-topic while introducing novel formulations and paraphrases, reflecting controlled semantic exploration within a stable coherence regime. At higher diffusion ($T = 1.0$), the outputs exhibited increased semantic drift, including topic shifts and weakened logical continuity. This behavior is indicative of a regime in which stochastic influence becomes comparable to coherence-restoring tendencies, leading to reduced stability of the semantic trajectory. At the highest diffusion level ($T = 1.5$), the generated text displayed pronounced incoherence, with fragmented topics and semantically incompatible elements. Within the CDD framework, this regime corresponds to diffusion-dominated dynamics, in which the semantic state frequently departs from relatively coherent regions of the effective representation landscape

A qualitative summary of these behaviors is provided in Table 3. Taken together, these observations support the CDD interpretation that hallucination-like behavior in language models is associated with diffusion-dominated semantic dynamics. Rather than arising solely as a sampling artifact, such behavior reflects a qualitative regime shift in the balance between coherence-restoring drift and stochastic diffusion within the effective semantic dynamics.

In this regime-shift interpretation, the growth of local output entropy serves as a natural order parameter, tracking the transition from coherence-dominated to diffusion-dominated behavior as the effective ratio σ/α increases. This reading provides empirical support for the drift–diffusion balance described in Appendix A.1.2, linking observed generative instability to a structural imbalance between coherence enforcement and stochastic variability.

7.3 Experiment III: Threshold-Invariant Robustness under Local Attention Pruning

This experiment evaluates the robustness of semantic computation under systematic local attention pruning, focusing on the regime relevant to Proposition 1. Within the CDD framework, semantic evolution is modeled as a drift–diffusion process whose semantic influence concentrates along a limited subset of representational directions. Proposition 1 provides a conservative sufficient condition, asserting that the removal of sufficiently low-impact interactions does not perturb the resulting semantic dynamics. Experiment III empirically examines this prediction by progressively pruning low-contribution attention interactions and measuring the resulting changes in the next-token distribution.

We evaluate GPT-2 Large under a teacher-forcing protocol across 200 diverse prompts, with each prompt evaluated independently and all reported metrics averaged across the full set. Pruning is applied exclusively at the final transformer layer and the final query position, isolating the effect of sparsifying the attention distribution at the point where the next-token prediction is formed. To avoid degenerate collapse of the attention distribution and to preserve a meaningful attention geometry for analysis, we impose a mild minimum-keep constraint, retaining the top 16 attention entries after pruning. Pruning follows a tail-mass criterion: attention weights are sorted, and the smallest entries whose cumulative probability mass does

Table 4: Threshold-invariant robustness under local attention pruning

ε	KL (mean)	Δ NLL (mean)	Top-1	Removed Mass	Removed Fraction
0	3.0×10^{-7}	-1.0×10^{-5}	1.00	0.000	0.000
0.01	3.0×10^{-7}	-1.0×10^{-5}	1.00	0.007	0.166
0.02	3.0×10^{-7}	-1.0×10^{-5}	1.00	0.014	0.223
0.05	3.0×10^{-7}	-1.0×10^{-5}	1.00	0.024	0.277
0.10	3.0×10^{-7}	-1.0×10^{-5}	1.00	0.030	0.292
0.20	3.0×10^{-7}	-1.0×10^{-5}	1.00	0.032	0.295

not exceed a threshold ε are removed, after which the remaining weights are renormalized. We sweep $\varepsilon \in \{0, 0.01, 0.02, 0.05, 0.1, 0.2\}$, spanning conservative to moderately aggressive pruning regimes.

For each prompt and threshold, we compare the pruned model’s next-token distribution with that of the unpruned baseline. We report the KL divergence between the two distributions, the change in negative log-likelihood (Δ NLL), Top-1 agreement, the fraction of attention entries removed, and the total probability mass removed. All metrics are computed per prompt and then averaged across the 200-prompt evaluation set.

Table 4 summarizes the results. Across all tested thresholds, the output distribution remains invariant within numerical precision. KL divergence remains on the order of 10^{-7} , Δ NLL remains on the order of 10^{-5} , and Top-1 agreement is perfect across all prompts and all values of ε . As ε increases, the fraction of pruned attention entries rises steadily until saturating at approximately 30 percent, while the removed probability mass stabilizes around 3 percent. Beyond $\varepsilon \approx 0.2$, further increases in the threshold do not lead to additional pruning, indicating that the same inactive subset of attention interactions is consistently identified and removed.

These results provide empirical support for Proposition 1 as well as the theoretical analysis in Appendix A.1.3. While the theory asserts stability under sufficiently small perturbations, the observed invariance persists across a broad range of thresholds without any fine-tuning of ε . This behavior indicates that effective semantic computation at the final decoding step is governed by a small, threshold-independent active core of attention interactions, with the remaining interactions contributing negligibly to the output distribution. This concentration does not presuppose the existence of a unique semantic basin and may arise in multi-regime or basinless settings.

We restrict our analysis to the constrained pruning regime, as it preserves a non-degenerate attention geometry that is essential for interpreting the structure of the active semantic core. More aggressive unconstrained pruning, while empirically revealing additional forms of redundancy, is beyond the scope of the present study and is explored in follow-up work.

7.4 Summary

Across three independent experiments, we find that the empirically testable implications of the CDD framework are consistently supported in GPT-2 models of different sizes. Taken together, these results indicate that CDD provides a coherent and predictive framework for interpreting key aspects of semantic behavior in Transformer-based language models within the modeling regime considered here. The present experiments validate qualitative regime-level predictions of the CDD framework, without asserting the existence of specific stabilization structures such as semantic basins, which are examined separately.

8 Discussion and Limitations

8.1 Discussion: Relation to Existing Interpretations and Prior Research

Standard descriptions of Transformer inference portray next-token prediction as a sequence of deterministic forward computations (Vaswani et al., 2023): each layer applies multi-head attention, residual addition, and feed-forward transformations until the final hidden state is mapped to a distribution over tokens. While this

mechanistic account is correct at the architectural level, it does not by itself explain several robust empirical phenomena documented in prior work, including:

- the pervasive sparsity of effective attention contributions (Clark et al., 2019),
- the stability of predictions under aggressive masking of attention pathways (Michel et al., 2019),
- the progressive collapse of contextual influence in deeper layers (Choi et al., 2016), and
- coherence breakdowns induced by increases in sampling temperature (Holtzman et al., 2020).

A broad body of efficiency-oriented research, such as Adaptive Computation Time (Graves, 2017), Universal Transformers (Dehghani et al., 2019), Early Exit and LayerDrop (Elhoushi et al., 2024), Dynamic Token Pruning (Fu et al., 2024), and sparse-attention architectures (Child et al., 2019; Tay et al., 2021), has repeatedly demonstrated that large portions of Transformer computation can be removed with minimal degradation in performance. These studies identify what can be skipped or pruned, but they do not provide a unified explanation for why sparsity, redundancy, and stability arise so consistently across architectures and tasks.

The CDD framework offers such an interpretive perspective by modeling Transformer inference as an instance of effective semantic dynamics. In this view, layer depth plays the role of a discrete time index, and hidden states trace trajectories in a high-dimensional semantic state space. Attention aggregation, residual accumulation, and feed-forward transformations can be interpreted as implementing a discretized drift-diffusion update shaped by a coherence-restoring parameter α and a diffusion scale σ . As discussed in Chapter 6, α captures structured coherence mechanisms such as attention concentration and residual gradient alignment, while σ reflects both sampling-induced variability and intrinsic stochasticity.

This dynamical interpretation sheds light on the structural consequence formalized in Proposition 1, the Principle of Dynamic Sparsity. Under the modeling assumptions of the CDD framework, interactions whose effective influence falls below an ε -threshold contribute negligibly to semantic evolution. Experiment III provides empirical support for this interpretation by demonstrating that masking a large fraction of low-influence token-to-token interactions results in near-zero changes to the output distribution. From the CDD perspective, this behavior is naturally explained by drift-driven suppression of instability: as semantic influence becomes increasingly concentrated along a subset of high-impact components, low-influence interactions contribute progressively less to the effective dynamics.

More broadly, the CDD framework provides a coherent lens through which a range of empirically observed phenomena such as contextual collapse, stability under masking, and head-level redundancy can be interpreted within a single dynamical picture. Rather than replacing existing computational descriptions, this perspective complements them by highlighting the qualitative structure of semantic evolution that emerges from repeated application of attention, residual connections, and nonlinear transformations.

Importantly, CDD does not claim that Transformers literally implement a stochastic differential equation. Instead, it proposes that interpreting inference through the lens of effective semantic dynamics offers a consistent and informative way to reason about stability, sparsity, and hallucination-like behavior that remain difficult to explain using purely feed-forward accounts. Within its modeling scope, this perspective suggests principled directions for adaptive depth control, controlled decoding, pruning, interpretability, and stability-aware generation strategies in LLMs.

We emphasize that CDD characterizes qualitative dynamical regimes and stability tendencies, without asserting the existence, uniqueness, or geometry of specific semantic basins, which are investigated separately in subsequent work.

8.2 Limitations

While the CDD framework provides a coherent and empirically grounded account of semantic evolution in LLMs, several limitations delineate the scope and intended interpretation of the framework.

(1) Phenomenological Nature of the Instability Potential Ψ . The potential Ψ is not a literal component of a Transformer’s architecture, nor is it derived in closed form from model parameters. Instead, Ψ is introduced as an effective potential inferred from observable signals such as negative log-likelihood spikes, abrupt shifts in attention patterns, and irregularities in gradient behavior. While these signals correlate with semantic instability, they do not uniquely determine a specific functional form for Ψ . Accordingly, CDD treats Ψ as an interpretive construct, analogous to effective potentials in physics: a meaningful abstraction rather than a mechanistic variable.

At the same time, adopting an effective potential perspective enables a substantial simplification of an otherwise intractable high-dimensional dynamical system. This abstraction allows CDD to capture collective behaviors, specifically stability regimes, progressive stabilization, and qualitative transitions toward instability, which are difficult to characterize through purely reductionist analysis. In this sense, Ψ serves as a useful descriptive and predictive tool for reasoning about large-scale stability patterns in the latent semantic dynamics.

(2) Approximate Mapping to Drift Dynamics. The drift term $-\alpha\nabla\Psi(x)$ should be interpreted as a conceptual approximation rather than a literal description of the underlying update rule. As discussed in Section 6.3.3, the Transformer layer update consists of two qualitatively distinct components: an interaction-driven term arising from self-attention, $f_{\text{attn}}(x)$, and an intrinsic drift arising from the MLP block, $f_{\text{MLP}}(x)$. Residual connections combine these contributions into the aggregate update

$$f(x) = f_{\text{attn}}(x) + f_{\text{MLP}}(x),$$

which is highly nonlinear, non-conservative, and further shaped by normalization and scaling operations.

CDD does not claim that this composite update implements explicit gradient descent on a well-defined potential Ψ . Rather, Ψ functions as an effective stability landscape that approximates the attractor-like behavior emerging from the combined action of interaction-driven and intrinsic drift components. In this view, α is best understood not as a scalar parameter but as an operator governing the strength and structure of coherence enforcement arising from the interplay of attention-mediated interactions and MLP-driven refinement.

(3) High Dimensionality and Markov Approximation. Transformer hidden states inhabit extremely high-dimensional spaces, challenging the assumptions of smoothness and low intrinsic dimensionality that underlie classical SDE and Fokker–Planck theory. Moreover, Transformer inference is not strictly Markovian: mechanisms such as the KV cache preserve long-range contextual information across decoding steps. In CDD, the residual update stream is treated as approximately Markovian to obtain a tractable dynamical formulation. This approximation necessarily simplifies cross-layer dependencies and long-context interactions, and may omit higher-order temporal effects.

It is worth noting, however, that this approximate Markovian description is expected to become increasingly consistent with actual model behavior in deeper layers, where semantic representations are already substantially formed, compressed, and stabilized. In such regimes, the residual state serves as a robust sufficient statistic for the prior context, and long-range dependencies are increasingly mediated through the current latent representation rather than explicit historical states.

(4) Scope of Applicability. The CDD framework aligns most naturally with autoregressive decoders, where semantic states evolve sequentially. Extending the framework to bidirectional encoders, encoder–decoder architectures, or multimodal systems will require additional assumptions to account for bidirectional updates, cross-attention coupling, and multimodal stability structures. Furthermore, the experiments in Chapter 7 focus on open-domain generation; assessing the applicability of CDD to reasoning, dialogue, code generation, multilingual tasks, and long-context settings remains an important direction for future work.

8.3 Summary

CDD offers a unified theoretical perspective on coherence, diffusion, sparsity, and stability in Transformer-based language models. Its limitations do not diminish its explanatory value; rather, they clarify the scope

and intended interpretation of the framework. By showing how discrete Transformer updates can be interpreted as exhibiting the qualitative structure of effective drift–diffusion dynamics at both fine-grained and coarse-grained levels, CDD provides a principled lens for reasoning about semantic stabilization, destabilization, and dimensional compression in LLMs. Within its modeling regime, this perspective yields both conceptual insight and practical implications for the design of more stable, efficient, and interpretable AI systems.

9 Applications and Implications

The CDD framework offers more than a theoretical reinterpretation of Transformer inference. By viewing inference as the evolution of an effective semantic state shaped by coherence-restoring drift and diffusion-induced variability, CDD provides a unifying perspective for reasoning about reliability, efficiency, interpretability, and controllability in LLMs. Phenomena often treated in isolation, such as hallucination, sparsity, stability, and interpretability, can be understood within this framework as different manifestations of a shared underlying dynamical structure.

9.1 Hallucination Prediction and Dynamical Stabilization

Within the CDD framework, hallucination-like behavior can be interpreted as arising from a qualitative regime shift in semantic dynamics. When the diffusion strength σ , capturing both sampling-induced variability such as temperature and model-internal stochasticity, becomes large relative to the coherence parameter α , semantic evolution becomes increasingly susceptible to instability. In such regimes, semantic states are more likely to depart from relatively coherent regions of an effective representation landscape and enter configurations associated with elevated instability potential Ψ . These configurations exhibit increased drift, weakened contextual consistency, and a higher likelihood of incoherent or fabricated continuations.

From this perspective, hallucination is not viewed solely as an incidental sampling artifact. Instead, it reflects an imbalance between coherence-restoring and stochastic influences, consistent with the regime-shift patterns observed in Chapter 7 and the drift–diffusion balance discussed in Appendix A.1.2. At a coarse modeling level, coherence-driven drift tends to suppress Ψ at a rate proportional to $\alpha\|\nabla\Psi\|^2$, while diffusion contributes an opposing effect proportional to $\sigma^2\Delta\Psi$. As these contributions become comparable, semantic behavior transitions toward diffusion-dominated regimes characterized by reduced stability.

This interpretation motivates the use of certain observable quantities as effective indicators of semantic instability. In particular, the growth of local output entropy along a generation trajectory provides a practically accessible proxy for transitions from coherence-dominated to diffusion-dominated regimes, as demonstrated in Experiment II. Entropy does not uniquely define instability, but it reliably correlates with increases in Ψ and the onset of hallucinatory behavior across a range of decoding conditions.

This dynamical viewpoint suggests several conceptual directions for detection and mitigation strategies:

- Monitoring instability-related signals such as increases in local output entropy, spikes in surprisal, dispersion of attention weights, or irregularities in gradient-based indicators as early-warning proxies for rising Ψ .
- Adaptive modulation of effective coherence and diffusion parameters, including attention sharpening, entropy regularization, or temperature adjustment, when early signs of instability are detected.
- Stabilization strategies that emphasize coherence-restoring mechanisms or constrain stochastic exploration as semantic evolution approaches high-instability regimes.

Taken together, these considerations motivate stability-aware generation strategies guided by semantic dynamics, complementing existing heuristic decoding methods rather than replacing them.

9.2 Efficiency Through Dynamic Sparsity

Proposition 1 characterizes dynamic sparsity not as an explicitly enforced design principle, but as a structural tendency that can arise from the semantic dynamics described by the CDD framework. Interactions whose effective contribution to semantic evolution is sufficiently small may, under certain regimes, exert only a negligible influence on semantic representations. Experiment III provides empirical support for this interpretation by demonstrating that pruning low-impact attention interactions at the final decoding stage induces only minimal changes in the output distribution, even when a substantial fraction of such interactions is suppressed.

From this perspective, the redundancy long observed in Transformer models can be interpreted as an emergent property of drift-driven suppression of instability and concentration of semantic influence, rather than as a purely architectural inefficiency. As semantic evolution exhibits reduced dispersion and influence becomes increasingly concentrated along a limited subset of high-impact interactions, other interactions contribute progressively less to the effective dynamics within the regimes considered here.

This viewpoint offers a principled way to understand why various forms of sparsity, pruning, or approximation often preserve model behavior. Rather than enforcing sparsity as an external objective, such procedures may be understood as removing interactions that are dynamically inactive under specific semantic regimes. Importantly, this interpretation does not assert a universal criterion for identifying inactive components, nor does it prescribe specific pruning mechanisms. Instead, it highlights a conceptual alignment between empirical compression practices and the contraction of effective semantic degrees of freedom implied by the CDD framework.

From an effective drift–diffusion perspective, suppressing low-impact interactions can be interpreted as restricting semantic evolution to a reduced subspace of the full representation space, while preserving qualitative stability properties of the dynamics under the modeling assumptions considered. The present discussion does not propose concrete efficiency algorithms. Rather, it provides a conceptual foundation for understanding why semantic computation in large Transformer models can admit substantial redundancy without immediate loss of predictive fidelity.

9.3 Trajectory-Based Interpretability

Most existing interpretability approaches focus on static analyses, including attention visualizations, neuron importance scores, or probing classifiers. While informative, such methods provide limited insight into the temporal processes through which semantic content is constructed, stabilized, and occasionally destabilized during inference. The CDD framework instead frames interpretability as the analysis of semantic trajectories evolving across layers or decoding steps.

Within this perspective, quantities related to the instability potential $\Psi(x(t))$ serve as informative indicators of semantic stability at each stage of inference. Drift–diffusion dynamics highlight when semantic evolution exhibits increased residence in relatively coherent regions of the effective representation landscape or begins to destabilize as stochastic influence grows. Transition points correspond to regimes in which operational proxies for diffusion strength, such as entropy or temperature, become comparable to coherence-restoring effects.

This trajectory-based view aligns interpretability with the generative process itself. Meaning emerges through the evolution of semantic states under competing coherence and diffusion influences, rendering interpretability inherently dynamical. Rather than attributing semantic behavior to isolated components, this perspective emphasizes how coherence is constructed, maintained, and degraded over the course of inference.

9.4 Broader Implications for Model Design and Alignment

The dynamical viewpoint suggested by CDD points to several directions for future model design, training, and evaluation.

- **Alignment and controllability.** Modulating the balance between effective coherence and diffusion parameters offers a potential mechanism for controlling trade-offs between creativity, stability, and factual precision beyond token-level constraints.
- **Multimodal coherence.** Extending instability-related measures to cross-modal representation spaces may provide a unified way to assess coherence across language, vision, and other modalities.
- **Training regime design.** Encouraging balanced drift–diffusion behavior during training may help reduce the prevalence of instability-prone regions in the representation space.
- **Evaluation metrics.** Coherence-related indicators derived from Ψ or its proxies may complement existing metrics by providing early signals of semantic degradation or instability.

These implications suggest that CDD functions not only as an interpretive framework, but also as a source of guiding principles for the development and analysis of future language models.

9.5 Summary

CDD connects conceptual interpretation with practical considerations by viewing Transformer inference as an instance of effective semantic dynamics. Within this framework, hallucination-like behavior is associated with diffusion-dominated regimes, sparsity emerges from the contraction of effective semantic degrees of freedom, and interpretability is naturally framed in terms of evolving trajectories rather than static artifacts. By grounding reliability, efficiency, and interpretability in a coherent drift–diffusion perspective, CDD offers a unified and cautiously scoped lens for analyzing and shaping the internal behavior of large language models.

10 Conclusion

This study introduced Coherence–Diffusion Dynamics (CDD) as an effective interpretive framework for understanding semantic evolution during inference in large language models. Rather than treating Transformer inference as a sequence of discrete layerwise feed-forward computations, CDD views hidden-state evolution as a structured dynamical process shaped by the interaction of coherence-restoring tendencies and diffusion-induced variability. This perspective provides a principled way to reason about stability, sparsity, hallucination-like behavior, and interpretability within a unified conceptual language.

Within the CDD framework, semantic evolution is characterized through three core modeling components. The effective coherence parameter α captures mechanisms that bias semantic behavior toward relatively coherent regimes, as reflected by stability-related observables such as attention concentration and residual alignment. The diffusion scale σ represents stochastic variability arising from both decoding procedures and intrinsic model uncertainty. The threshold parameter ε appears as a derived quantity associated with the principle of dynamic sparsity, providing an operational criterion for identifying interactions whose effective influence on semantic updates is negligible within the modeling regime considered.

The experiments presented in this work provide empirical support for the qualitative predictions of the CDD framework. Experiment I demonstrated that varying an operational proxy for α is associated with distinct stability regimes during fine-tuning dynamics, consistent with the role of coherence-restoring effects. Experiment II showed that increasing diffusion strength leads to systematic regime shifts in generative behavior, including the emergence of hallucination-like outputs when stochastic influence becomes dominant. Experiment III indicated that pruning attention interactions with low effective contribution at the final decoding stage induces only minimal changes in the output distribution, supporting the interpretation of dynamic sparsity as an emergent feature of effective semantic dynamics rather than as an ad hoc architectural property.

Taken together, these findings suggest that the internal computations of Transformer models admit a coherent interpretation in terms of semantic trajectories evolving within an effective potential-based representation landscape. Importantly, CDD characterizes qualitative dynamical regimes and stability tendencies without asserting the existence, uniqueness, or geometry of specific semantic basins or attractors. Finer-grained

structural objects, such as basin geometry, multi-basin organization, or basinless regimes, require separate empirical investigation beyond the scope of the present framework.

By reframing interpretability as the analysis of evolving semantic trajectories rather than static components, CDD highlights how coherence is constructed, maintained, and degraded over the course of inference. This dynamical viewpoint complements existing architectural and statistical analyses, offering insight into why large language models can exhibit both remarkable stability and sudden instability under different operational conditions.

In summary, CDD does not claim to render language models fully transparent or directly controllable. Instead, it provides a cautiously scoped interpretive framework that clarifies regularities in semantic evolution and identifies regime-level relationships between coherence, diffusion, sparsity, and instability. As large language models continue to play an increasingly central role across scientific and societal domains, such structured dynamical perspectives may prove valuable for guiding future work on reliability, efficiency, interpretability, and control, while leaving room for more specialized analyses that investigate the detailed structure of semantic representations themselves.

References

- Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2018. URL <https://arxiv.org/abs/1610.01644>.
- Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations*, 2017. URL <https://api.semanticscholar.org/CorpusID:64908139>.
- Daniel G. Bobrow and W A Woods. Computational linguistics transition network grammars for natural language analysis. URL <https://api.semanticscholar.org/CorpusID:282686724>.
- Tom B. Brown et al. Language models are few-shot learners, 2020. URL <https://arxiv.org/abs/2005.14165>.
- Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers, 2019. URL <https://arxiv.org/abs/1904.10509>.
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014. URL <https://arxiv.org/abs/1406.1078>.
- Heeyoul Choi, Kyunghyun Cho, and Yoshua Bengio. Context-dependent word representation for neural machine translation, 2016. URL <https://arxiv.org/abs/1607.00578>.
- Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does bert look at? an analysis of bert’s attention, 2019. URL <https://arxiv.org/abs/1906.04341>.
- Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers, 2019. URL <https://arxiv.org/abs/1807.03819>.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL <https://arxiv.org/abs/1810.04805>.
- Mostafa Elhoushi, Akshat Shrivastava, Diana Liskovich, Basil Hosmer, Bram Wasti, Liangzhen Lai, Anas Mahmoud, Bilge Acun, Saurabh Agarwal, Ahmed Roman, Ahmed Aly, Beidi Chen, and Carole-Jean Wu. Layerskip: Enabling early exit inference and self-speculative decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12622–12642. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.681. URL <http://dx.doi.org/10.18653/v1/2024.acl-long.681>.
- Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings, 2019. URL <https://arxiv.org/abs/1909.00512>.

- Stanislav Fort and Surya Ganguli. Emergent properties of the local geometry of neural loss landscapes, 2019. URL <https://arxiv.org/abs/1910.05929>.
- Qichen Fu, Minsik Cho, Thomas Merth, Sachin Mehta, Mohammad Rastegari, and Mahyar Najibi. Lazyllm: Dynamic token pruning for efficient long context llm inference, 2024. URL <https://arxiv.org/abs/2407.14057>.
- Alex Graves. Adaptive computation time for recurrent neural networks, 2017. URL <https://arxiv.org/abs/1603.08983>.
- Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration, 2020. URL <https://arxiv.org/abs/1904.09751>.
- Sarthak Jain and Byron C. Wallace. Attention is not explanation, 2019. URL <https://arxiv.org/abs/1902.10186>.
- Takuya Kataiwa, Cho Hakaze, and Tetsushi Ohki. Measuring intrinsic dimension of token embeddings, 2025. URL <https://arxiv.org/abs/2503.02142>.
- Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one?, 2019. URL <https://arxiv.org/abs/1905.10650>.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013a. URL <https://arxiv.org/abs/1301.3781>.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality, 2013b. URL <https://arxiv.org/abs/1310.4546>.
- Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability, 2023. URL <https://arxiv.org/abs/2301.05217>.
- Allen Newell and Herbert A. Simon. Computer science as empirical inquiry: symbols and search. *Commun. ACM*, 19:113–126, 1976. URL <https://api.semanticscholar.org/CorpusID:5581562>.
- Catherine Olsson et al. In-context learning and induction heads, 2022. URL <https://arxiv.org/abs/2209.11895>.
- OpenAI et al. Gpt-4 technical report, 2024. URL <https://arxiv.org/abs/2303.08774>.
- Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks, 2013. URL <https://arxiv.org/abs/1211.5063>.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL <https://aclanthology.org/D14-1162/>.
- Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018. URL <https://api.semanticscholar.org/CorpusID:49313245>.
- Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, 2019. URL <https://arxiv.org/abs/1811.10154>.
- Levent Sagun, Utku Evci, V. Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks, 2018. URL <https://arxiv.org/abs/1706.04454>.

Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, and Che Zheng. Synthesizer: Rethinking self-attention in transformer models, 2021. URL <https://arxiv.org/abs/2005.00743>.

Hugo Touvron et al. Llama: Open and efficient foundation language models, 2023. URL <https://arxiv.org/abs/2302.13971>.

Ashish Vaswani et al. Attention is all you need, 2023. URL <https://arxiv.org/abs/1706.03762>.

Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model, 2019. URL <https://arxiv.org/abs/1906.04284>.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization, 2017. URL <https://arxiv.org/abs/1611.03530>.

A Appendix

This Supplementary Material provides the rigorous mathematical proofs, implementation details, and expanded experiments that support the claims made in the main text. The combined evidence confirms that the Coherence–Diffusion Dynamics (CDD) framework is both theoretically grounded and empirically valid across model scales and tasks.

A.1 Mathematical Proofs

A.1.1 Proof of the Stability Convergence Theorem

Consider the CDD stochastic differential equation

$$dx(t) = -\alpha \nabla \Psi(x(t)) dt + \sigma dW(t).$$

Assume the following conditions:

- A1. $\Psi \in C^2$ with Lipschitz-continuous gradient.
- A2. $\Psi(x) \geq 0$ and is minimized on a compact set \mathcal{T} .
- A3. Effective drift dominance: there exists $c > 1$ such that, in the regime of interest,

$$\alpha \|\nabla \Psi(x)\|^2 \geq c \sigma^2 \Delta \Psi(x).$$

Let the Lyapunov function be

$$V(x) = \Psi(x).$$

By Itô’s lemma,

$$dV = -\alpha \|\nabla \Psi\|^2 dt + \sigma \langle \nabla \Psi, dW(t) \rangle + \frac{1}{2} \sigma^2 \Delta \Psi dt.$$

Taking expectation and applying assumption A3,

$$\frac{d}{dt} \mathbb{E}[V(t)] \leq - \left(1 - \frac{1}{2c}\right) \alpha \mathbb{E}[\|\nabla \Psi(x(t))\|^2].$$

Thus $\mathbb{E}[V(t)]$ decreases monotonically and is bounded below, which implies

$$\mathbb{E}\|\nabla \Psi(x(t))\|^2 \rightarrow 0.$$

Hence $x(t)$ converges in probability to the coherence manifold

$$\mathcal{T} = \arg \min_x \Psi(x).$$

Interpretation of the diffusion scale σ

In the drift–diffusion formulation, σ controls the variance of stochastic increments of the latent semantic trajectory. Informally, larger values of σ correspond to increased one-step variability of the hidden state, in the sense that

$$\mathbb{E}\|x(t + \Delta t) - x(t)\|^2 = O(\sigma^2 \Delta t).$$

In Transformer models, this increased variability arises from decoding entropy, sampling temperature, and uncertainty at the logit level, rather than from an explicit noise source.

A.1.2 Critical α – σ Balance Condition

At a coarse-grained modeling level, we consider the expected rate of change of the effective instability potential Ψ along a semantic trajectory, which may be heuristically expressed as

$$D(t) \approx -\alpha \|\nabla \Psi\|^2 + \frac{1}{2} \sigma^2 \Delta \Psi.$$

Here $\Delta \Psi$ denotes the Laplacian of the potential, representing the local curvature of the semantic landscape relevant to diffusion effects.

Within this approximation, a qualitative regime shift is suggested when the net drift changes sign, corresponding to a balance between coherence and diffusion. This yields an effective *critical diffusion-to-coherence ratio*

$$\rho_c = \frac{\sigma}{\alpha} \approx \left(\frac{2\|\nabla \Psi\|^2}{\Delta \Psi} \right)^{1/2}.$$

This expression should not be interpreted as a sharp or universal critical value. Rather, it provides a schematic characterization of the transition between coherence-dominated and diffusion-dominated regimes in the CDD framework. Qualitatively, this balance condition is consistent with the regime shifts observed empirically in Experiment II near sampling temperatures $T \simeq 1.0$ – 1.5 .

A.1.3 Perturbative Bounds Under Low-Influence Masking

This appendix provides a formal perturbation bound showing that suppressing sufficiently low-influence attention interactions induces only second-order changes in the output distribution. This result establishes the quantitative basis for Proposition 1 (Dynamic Sparsity).

Proposition (KL Stability Under Low-Influence Masking)

Let the output of an attention head be

$$o = \sum_{j=1}^n a_{ij} v_j,$$

where $a_{ij} \geq 0$, $\sum_j a_{ij} = 1$, and $\|v_j\| \leq C_v$.

Let $S \subset \{1, \dots, n\}$ be a set of suppressed interactions with total weight

$$\sum_{j \in S} a_{ij} \leq \varepsilon.$$

Let

$$o' = \sum_{j \notin S} a_{ij} v_j$$

denote the masked output, and let

$$p = \text{softmax}(Wo), \quad p' = \text{softmax}(Wo')$$

be the corresponding output distributions for an affine map W .

Assume:

1. W is L_W -Lipschitz.
2. The softmax map is locally Lipschitz on bounded logit domains, i.e.,

$$\|\sigma(z) - \sigma(z')\|_1 \leq L_{\text{sm}}\|z - z'\|$$

for some constant L_{sm} depending on the logit range.

3. The output distribution satisfies $\min_y p'(y) \geq \delta > 0$. This condition is satisfied in practice for finite vocabularies under standard decoding settings with nonzero temperature.

Then there exists a constant $C > 0$, depending only on C_v , L_W , L_{sm} , and δ , such that

$$D_{\text{KL}}(p \| p') \leq C\varepsilon^2.$$

Assumptions

The derivation relies on the following mild regularity assumptions, which are satisfied under standard modeling and decoding conditions used in Transformer architectures:

- Bounded value vectors: $\|v_j\| \leq C_v$.
- Linear attention aggregation: head outputs are linear combinations of value vectors.
- Lipschitz output map: $\|Wx - Wy\| \leq L_W\|x - y\|$.
- Locally Lipschitz softmax: softmax is Lipschitz on bounded logit domains.
- Non-degenerate output distribution: $\min_y p'(y) \geq \delta > 0$.

Proof

Step 1: Perturbation of the head output.

Masking removes

$$\Delta o = o - o' = \sum_{j \in S} a_{ij} v_j.$$

Using boundedness of v_j ,

$$\|\Delta o\| \leq \sum_{j \in S} a_{ij} \|v_j\| \leq C_v \sum_{j \in S} a_{ij} \leq C_v \varepsilon. \quad (1)$$

Step 2: Propagation through the affine map.

Since W is L_W -Lipschitz,

$$\|Wo - Wo'\| \leq L_W \|\Delta o\| \leq L_W C_v \varepsilon. \quad (2)$$

Let $\Delta z = Wo - Wo'$ denote the logit perturbation.

Step 3: Softmax perturbation bound.

By local Lipschitz continuity of softmax,

$$\|p - p'\|_1 \leq L_{\text{sm}} \|\Delta z\| \leq L_{\text{sm}} L_W C_v \varepsilon. \quad (3)$$

Let $\delta_p = \|p - p'\|_1$. Then $\delta_p = O(\varepsilon)$.

Step 4: KL divergence bound.

For distributions with $\min_y p'(y) \geq \delta$, the KL divergence satisfies

$$D_{\text{KL}}(p \| p') \leq \frac{1}{2\delta} \|p - p'\|_1^2. \quad (4)$$

Combining (3) and (4),

$$D_{\text{KL}}(p \| p') \leq \frac{1}{2\delta} (L_{\text{sm}} L_W C_v \varepsilon)^2.$$

Thus,

$$D_{\text{KL}}(p \| p') \leq C \varepsilon^2,$$

where

$$C = \frac{(L_{\text{sm}} L_W C_v)^2}{2\delta}.$$

□

Interpretation This perturbative bound shows that suppressing interactions whose total attention weight is $O(\varepsilon)$ induces only $O(\varepsilon)$ changes in logits and therefore only $O(\varepsilon^2)$ changes in the KL divergence. In this sense, low-influence interactions are *dynamically negligible*, consistent with the notion of effective irrelevance formalized in Proposition 1. We emphasize that this bound provides a conservative sufficient condition for stability. Empirically, as shown in Experiment III, semantic invariance persists far beyond the perturbative regime covered by this analysis, even when the removed attention mass is no longer $O(\varepsilon)$.

A.2 Implementation Details

A.2.1 Training Settings (Experiment I)

- **Model:** GPT-2 Small, GPT-2 Medium, GPT-2 Large
- **Optimizer:** AdamW
- **Learning rates tested:** $\{1e^{-6}, 5e^{-6}, 1e^{-5}, 5e^{-5}, 5e^{-4}, 5e^{-3}\}$
- **Batch size:** ~ 512 tokens (seq_len = 256, batch = 4)
- **Epochs:** 1
- **Warmup ratio:** 10%
- **Weight decay:** 0.0
- **Dataset:** WikiText-2 train set (5,000-sample subset)
- **Precision:** fp16 on RTX 2080 Ti
- **Data collator:** causal language modeling (no MLM)

A.2.2 Inference Settings (Experiment II & III)

Experiment II: Temperature-Controlled Diffusion Dynamics This experiment evaluates the qualitative effect of the diffusion coefficient σ (implemented as sampling temperature) on semantic coherence and stability.

- **Decoding method:** stochastic sampling (`do_sample=True`)
- **Temperature values:** $\sigma \in \{0.1, 0.7, 1.0, 1.5\}$
- **Max output length:** prompt + 100 tokens
- **Random seed:** 42

Experiment III: Threshold-Invariant Stability under Local Attention Pruning This experiment provides an auxiliary validation of the Dynamic Sparsity Principle (Proposition 1; Appendix A.1.3) by examining the stability of the output distribution under systematic pruning of low-impact attention interactions.

- **Evaluation protocol:** forward pass only under teacher forcing; no sampling or decoding
- **Input length:** all inputs are truncated to a maximum length of 1024 tokens
- **Attention source:** self-attention at the final transformer layer and final query position
- **Head aggregation:** pruning applied after averaging attention weights across heads
- **Pruning rule:** tail-mass pruning, removing the smallest attention entries whose cumulative probability mass does not exceed a threshold ε , followed by renormalization
- **Threshold values:** $\varepsilon \in \{0, 0.01, 0.02, 0.05, 0.1, 0.2\}$
- **Minimum-keep constraint:** to avoid degenerate collapse of the attention distribution, at least the top 16 attention entries are always preserved
- **Stability metric:** KL divergence $\text{KL}(P \parallel Q)$ between baseline and pruned next-token distributions, computed using `torch.nn.functional.kl_div`
- **Random seed:** 42

A.2.3 Environment

- **GPU:** NVIDIA GeForce RTX 2080 Ti (11 GB)
- **CPU:** Intel Xeon E5-2609 @ 1.90 GHz
- **Framework:** Python 3.10.10, PyTorch 2.5.1+cu121
- **Transformers:** HuggingFace v4.57.1
- **OS:** Windows 11

A.3 Complete Generated Samples from Experiment II at Each Temperature.

Prompt: “The Coherence-Diffusion Dynamics (CDD) framework proposes that semantic evolution is a dynamic process shaped by three core assumptions. Under the first assumption, defined as the principle of difference, it is posited that”

Temperature: 0.1

“the evolution of a language is driven by the evolution of its meaning. Under the second assumption, defined as the principle of coherence, it is posited that the evolution of a language is driven by the evolution of

its structure. Under the third assumption, defined as the principle of coherence-diffusion dynamics, it is posited that the evolution of a language is driven by the evolution of its coherence. The CDD framework is a framework for the study of the evolution of a”,

“the evolution of a semantic system is a process of change that is driven by the interaction of the semantic system with the environment. Under the second assumption, defined as the principle of similarity, it is posited that the evolution of a semantic system is a process of change that is driven by the interaction of the semantic system with the environment. Under the third assumption, defined as the principle of coherence, it is posited that the evolution of a semantic system is a process of change that is driven by”,

“the evolution of a semantic system is a process of change that is driven by the interaction of the semantic system with the environment. Under the second assumption, defined as the principle of coherence, it is posited that the evolution of a semantic system is a process of change that is driven by the interaction of the semantic system with the environment. Under the third assumption, defined as the principle of diffusion, it is posited that the evolution of a semantic system is a process of change that is driven by”,

“the evolution of a semantic system is a process of change that is driven by the interaction of the semantic system with the environment. Under the second assumption, defined as the principle of coherence, it is posited that the evolution of a semantic system is a process of change that is driven by the interaction of the semantic system with the environment. Under the third assumption, defined as the principle of diffusion, it is posited that the evolution of a semantic system is a process of change that is driven by”,

“the evolution of a semantic system is governed by a process of diffusion. Under the second assumption, defined as the principle of coherence, it is posited that the evolution of a semantic system is governed by a process of coherence. Under the third assumption, defined as the principle of coherence-diffusion dynamics, it is posited that the evolution of a semantic system is governed by a process of coherence-diffusion dynamics. The CDD framework is based on the idea that”

Temperature: 0.7

“the degree of semantic similarity between two words is a function of their semantic distance. The second assumption is that semantic similarity is a function of the semantic distance between the meanings of the words. The third assumption is that the degree of semantic similarity between two words is a function of the semantic distance between the meanings of the words. In its simplest form, the CDD framework proposes that the level of semantic similarity between two words can be described by the distance between the meanings of the words. The CDD framework”,

“language evolves from a priori meanings, which are chosen by the language user, to explicit meanings, which are determined by the user, and which are then selected by the system. Under the second assumption, defined as the principle of similarity, it is posited that meaning is a priori, and that meaning is determined by the system. Under the third assumption (the principle of similarity) it is posited that meaning is a priori, and that meaning is determined by the system. To understand”,

“semantic evolution is a process by which semantic changes occur through the interaction of two or more distinct subsets of the language. Under the second assumption, defined as the principle of coherence, it is posited that semantic changes occur through a process of coherence between two or more distinct subsets of the language. Under the third assumption, defined as the principle of similarity, it is posited that semantic changes occur through a process of similarity between two or more distinct subsets of the language. The CD”,

“semantic change occurs as a function of the degree to which an organism is “competent” or “incompetent” in interpreting linguistic input. The second assumption is that semantic change occurs as a function of the degree to which the organism is “competent” or “incompetent” in interpreting environmental cues (e.g., environmental context) and the third assumption is that semantic change occurs as a function of the degree to which the organism is “competent” or”,

“semantic evolution is primarily driven by differences between semantic values, and that differences among semantic values are the most likely cause of semantic evolution. Under the second assumption, defined as the principle of context, it is posited that context is a primary driver of semantic evolution. Under the third

assumption, defined as the principle of coherence, it is posited that semantic evolution is a stable process that must be sustained by a dynamic process. The CDD framework has been extensively used to study and understand the”

Temperature: 1.0

“the evolution of a human language and culture has taken place through the convergence of many distinct human linguistic and cultural units. The second is defined as the Principle of Contingency, which states that the evolution of a language or culture will continue to evolve under a set of initial conditions that are largely predetermined by its user and the cultural context of the user. This third assumption, defined as the Rule of Law, states that the interaction between a linguistic unit and its user will eventually lead to an adoption of a”,

“semantic evolution occurs naturally and dynamically in a framework that minimizes the importance of temporal noise and enhances the precision of sensory and neural coding. Under the second assumption, defined as the principle of coherence, meaning is represented as a function of time and is understood as the most stable element of semantic complexity. Under the third assumption, defined as the principle of coupling, semantic evolution occurs with minimal cognitive cost. The Coherence-Diffusion Dynamics (CDD) framework is described below. The CD”,

“the level of detail used by organisms to describe their states of being is not constant. Because organisms have a limited amount of information relevant to the development of their state of being, higher-level states can evolve over time and serve as models for what these states of being would have been like before some other state evolved. Under the second assumption, defined as the principle of convergence, it is posited that states that seem to be at opposite ends of a scale will ultimately be closer and more similar to one”,

“, at the level of discourse, evolution tends to reduce the richness and complexity of a system, in terms of its constituent parts and the types of information they convey. It is then posited that, on the other hand, evolution operates based on an understanding of what information is actually useful, and that, therefore, the richer and more complex concepts must be encoded in higher and higher levels until evolution achieves this ultimate goal. Under the second assumption, defined as the principle of congruence,”,

“all individuals, while being connected, are not necessarily connected with respect to others. Under the second assumption is the “coherence-diffusion” model, called the “co-evolutionary” model and proposed by the theoretical physicist Michal Kalecki (1998), it asserts that evolution is a continuous process of evolution where there is a separation between individuals. Each individual evolves independently, in accordance with a causal, non-cohesive dynamical mechanism, which in turn results in the emergence of”

Temperature: 1.5

“evolution proceeds in such a manner due to the interdependence, coherence, and differential reinforcement properties found in all three dimensions in biological systems. According to this, the coherence-diffusion interactions hold for all cognitive constructs that are able to distinguish and solve different questions such as how to obtain food and food satisfaction, to select to which item of a large object to donate by combining with one object in another (i.e, the ‘snowball theorem’) to obtain total quantity added,”,

“languages should evolve faster based on the relative degree of lexical information the user inputs. For this, speakers with larger lexical volumes would produce less and fewer syllable clusters, resulting in more generalized semantic systems. Hence with all these processes, semantic evolution does not only entail loss of local semantics but the reduction as well. The second assumption — the principle of information-integrity — is based on the expectation that the semantic processes evolve into new ones based on semantic information obtained by training one’s language”,

“: every time language evolves, its underlying structure can reflect a shared “mental structure”; the fundamental and defining ‘concepts’ of the semantic elements are shared. This has obvious effects in the formulation of many grammatical rules. It thus emerges that a semantic system based around an arbitrary set of initial lexical elements reflects a very different, and presumably somewhat arbitrary, language structure (aside from potentially different meanings and expressions with meaning ‘mixed’: as described above), and thus should therefore”,

“contextually-dense elements form natural categories and make use of specific localities to organize a large number of local semantic patterns for different domains on their terms. At the same time this first assumption presupposes the principle of coherence and the observation made the first time: that local semantics are composed via a series of interaction among elements forming local classes and a single general structure corresponding on each level of context in the sense-like domain for these different elements. The purpose of this work is to demonstrate whether”,

“semantic data structures always evolve a consistent internal structure of both similarity (i.e., the semantic content) and dissimilarities (i.e., the semantic properties found in the content). On the surface, the principles could be seen through an evolution of a language with phonological systems from simpler to deeper; as these two layers of structural information, semantic richness and the semantic order would gradually be refined to enable us communicate more deeply with each language. These two structures may then eventually form the basis from”