A Statistical Physics of Language Model Reasoning

Jack David Carson [1]

Abstract

Transformer LMs show emergent reasoning that resists mechanistic understanding. We offer a statistical physics framework for continuous-time chain-of-thought reasoning dynamics. We model sentence-level hidden state trajectories as a stochastic dynamical system on a lower-dimensional manifold. This drift-diffusion system uses latent regime switching to capture diverse reasoning phases, including misaligned states or failures. Empirical trajectories (8 models, 7 benchmarks) show a rank-40 projection (balancing variance capture and feasibility) explains 50% variance; we use this computationally tractable reduction not to claim inherent anisotropy, but to enable feasible SDE parameter estimation. We find four latent reasoning regimes. An SLDS model is formulated and validated to capture these features. The framework enables low-cost reasoning simulation, offering tools to study and predict critical transitions like misaligned states or other LM failures.

1. Introduction

Transformer LMs (Vaswani et al., 2017), trained for next-token prediction (Radford et al., 2019; Brown et al., 2020), show emergent reasoning like complex cognition (Wei et al., 2022). Standard analyses of discrete components (e.g., attention heads (Elhage et al., 2021; Olsson et al., 2022)) provide limited insight into longer-scale semantic transitions in multi-step reasoning (Allen-Zhu & Li, 2023; López-Otal et al., 2024). Understanding these high-dimensional, prediction-shaped semantic trajectories, particularly how they might cause misaligned states, is a key challenge (Li et al., 2023; Nanda et al., 2023).

We model reasoning as a continuous-time dynamical system, drawing from statistical physics (Chaudhuri & Fiete, 2016; Schuecker et al., 2018). Sentence-level hidden states $h(t) \in \mathbb{R}^D$ evolve via a stochastic differential equation (SDE):

$$\mathrm{d}h(t) = \mu(h(t), Z(t))\,\mathrm{d}t + B(h(t), Z(t))\,\mathrm{d}W(t), \quad (1)$$

with drift μ, diffusion B, Wiener process $W(t)$, and latent regimes $Z(t)$. This decomposes trajectories into trends and variations, helping identify deviations. As full high-dimensional SDE analysis (e.g., $D > 2048$ for most LMs) is impractical, we use a lower-dimensional manifold capturing significant variance for modeling.

This continuous-time dynamical systems perspective offers several benefits:

1. **Principled Abstraction:** We introduce a mathematically grounded, semantic-level view of reasoning, akin to statistical physics approximations. This moves beyond token mechanics to enable a robust interpretation of reasoning pathways and potential misalignments.
2. **Tractable Latent Structure Identification:** Our approach makes the analysis of reasoning trajectories feasible by focusing on a low-dimensional manifold (e.g., rank-40 PCA capturing approximately 50% of the variance) that describes significant structured evolution.
3. **Reasoning Regime Discovery:** We uncover distinct latent semantic regimes with unique drift and variance profiles. This suggests context-driven switching and offers insight into how models might transition into different reasoning states (see Appendix E).
4. **Efficient Surrogate Model:** The proposed SLDS accurately models and reconstructs reasoning trajectories with significant computational savings, facilitating the study of how reasoning processes unfold.
5. **Failure Mode Analysis:** We provide tools to study critical transitions and robustness, and to predict inference-time failure modes or misaligned states in LLM reasoning.

Chain-of-thought (CoT) prompting (Wei et al., 2022; Wang et al., 2023) has demonstrated that LMs can follow structured reasoning pathways, hinting at underlying processes amenable to a dynamical systems description. While prior work has applied continuous-time models to neural dynamics generally, the explicit modeling of transformer reasoning at these semantic timescales, particularly as an approximation for impractical full-dimensional analysis, has been largely unexplored. Our work bridges this gap by pursuing an SDE-based perspective informed by empirical analysis of transformer hidden-state trajectories.

This paper is structured as follows: Section 2 introduces the mathematical formalism of SDEs and regime switching.

[1]Massachusetts Institute of Technology, Cambridge, MA, USA. Correspondence to: Jack David Carson <jdcarson@mit.edu>.

Published at ICML 2025 Workshop on Reliable and Responsible Foundation Models. Copyright 2025 by the author(s).

Section 3 details our data collection and initial empirical findings that motivate the model, including the practical need for dimensionality reduction. Section 4 formally defines the SLDS model. Section 5 presents experimental validation, including model fitting, generalization, ablation studies, and a case study on modeling adversarial belief shifts as an example of predicting misaligned states.

2. Mathematical Preliminaries

We conceptualize the internal reasoning process of a transformer LM as a continuous-time stochastic trajectory evolving within its hidden-state space. Let $h_t \in \mathbb{R}^D$ be the final-layer residual embedding extracted at discrete sentence boundaries $t = 0, 1, 2, \ldots$. To capture the rich semantic evolution across reasoning steps, we treat these discrete embeddings as observations of an underlying continuous-time process $h(t) : \mathbb{R}_{\geq 0} \to \mathbb{R}^D$. The direct analysis of such a process in its full dimensionality (e.g., $D \geq 2048$) is often computationally prohibitive. We therefore aim to approximate its dynamics using SDEs, potentially in a reduced-dimensional space.

Definition 2.1 (Itô SDE)**.** An Itô stochastic differential equation on the state space $\mathbb{R}^D$ is given by:

$$\mathrm{d}h(t) = \mu(h(t))\,\mathrm{d}t + B(h(t))\,\mathrm{d}W(t), \quad h(0) \sim p_0, \quad (2)$$

where $\mu : \mathbb{R}^D \to \mathbb{R}^D$ is the deterministic *drift* term, encoding persistent directional dynamics. The matrix $B : \mathbb{R}^D \to \mathbb{R}^{D \times D'}$ is the *diffusion* term, modulating instantaneous stochastic fluctuations. $W(t)$ is a D'-dimensional Wiener process (standard Brownian motion), and p_0 is the initial distribution. The noise dimension D' can be less than or equal to the state dimension D.

The drift $\mu(h(t))$ represents systematic semantic or cognitive tendencies, while the diffusion $B(h(t))$ accounts for fluctuations due to local uncertainties, token-level variations, or inherent model stochasticity. Standard conditions ensure the well-posedness of such SDEs:

Theorem 2.1 (Well-Posedness (Øksendal, 2003))**.** *If μ and B satisfy standard Lipschitz continuity and linear growth conditions (see Appendix A), the SDE*

$$\mathrm{d}h(t) = \mu(h(t))\,\mathrm{d}t + B(h(t))\,\mathrm{d}W(t) \quad (3)$$

has a unique strong solution for a given D'-dimensional Wiener process $W(t)$.

We focus on dynamics at the sentence level:

Definition 2.2 (Sentence-Stride Process)**.** The *sentence-stride* hidden-state process is the discrete sequence $\{h_t\}_{t \in \mathbb{N}}$ obtained by extracting the final-layer transformer state immediately following each detected sentence boundary. This emphasizes mesoscopic, semantic-level changes over finer-grained token-level variations.

To analyze these dynamics in a computationally manageable way, particularly given the high dimensionality D of $h(t)$, we utilize projection-based dimensionality reduction. The goal is to find a lower-dimensional subspace where the most significant dynamics, for the purpose of modeling the SDE, unfold.

Definition 2.3 (Projection Leakage)**.** Given an orthonormal matrix $V_k \in \mathbb{R}^{D \times k}$ (where $V_k^\top V_k = I_k$), the *leakage* of the drift μ under perturbations v orthogonal to the image of V_k (i.e., $v \perp \mathrm{Im}(V_k)$) is

$$L_k = \sup_{\substack{x \in \mathbb{R}^D,\, \|v\| \leq \epsilon \\ v^\top V_k = 0}} \frac{\|\mu(x+v) - \mu(x)\|}{\|\mu(x)\|}.$$

A small leakage L_k implies that the drift's behavior relative to its current direction is not excessively altered by components outside the subspace spanned by V_k, making the subspace a reasonable domain for approximation.

Assumption 2.1 (Approximate Projection Closure for Modeling)**.** For practical modeling of the SDE (Eq. 2), we assume there exists a rank k (e.g., $k = 40$ in our work, chosen based on empirical variance and computational trade-offs) and a perturbation scale $\epsilon > 0$ such that $L_k \ll 1$. This allows the approximation of the drift within this k-dimensional subspace:

$$\mu(h(t)) \approx V_k V_k^\top \mu(h(t))$$

holds up to an error of order $O(L_k)$. This assumption underpins the feasibility of our low-dimensional modeling approach, enabling the analytical treatment inspired by statistical physics; this assumption is purely pragmatic, and we make no claims about intrinsic anisotropy of hidden states, only that this reduction enables feasible SDE parameter estimation.

Empirical observations of reasoning trajectories suggest abrupt shifts, potentially indicating transitions between different phases of reasoning or slips into misaligned states. This motivates a regime-switching framework:

Definition 2.4 (Regime-Switching SDE)**.** Let $Z(t) \in \{1, \ldots, K\}$ be a latent continuous-time Markov chain with a transition rate matrix $T \in \mathbb{R}^{K \times K}$. The corresponding regime-switching Itô SDE is:

$$\mathrm{d}h(t) = \mu_{Z(t)}(h(t))\,\mathrm{d}t + B_{Z(t)}(h(t))\,\mathrm{d}W(t), \quad (4)$$

where each latent regime $i \in \{1, \ldots, K\}$ has distinct drift μ_i and diffusion B_i functions. This allows for context-dependent dynamic structures (Ghahramani & Hinton, 2000), crucial for capturing diverse reasoning pathways.

These definitions establish the mathematical foundation for our analysis of transformer reasoning dynamics as a tractable approximation of a more complex high-dimensional process.

3. Data and Empirical Motivation

We build a corpus of sentence-aligned hidden-state trajectories from transformer-generated reasoning chains across a suite of models (Mistral-7B-Instruct (Jiang et al., 2023), Phi-3-Medium (Abdin et al., 2024), DeepSeek-67B (DeepSeek-AI et al., 2024), Llama-2-70B (Touvron et al., 2023), Gemma-2B-IT (Gemma Team & Google DeepMind, 2024), Qwen1.5-7B-Chat (Bai et al., 2023), Gemma-7B-IT (also (Gemma Team & Google DeepMind, 2024)), Llama-2-13B-Chat-HF (also (Touvron et al., 2023))) and datasets (StrategyQA (Geva et al., 2021), GSM-8K (Cobbe et al., 2021), TruthfulQA (Lin et al., 2022), BoolQ (Clark et al., 2019), OpenBookQA (Mihaylov et al., 2018), HellaSwag (Zellers et al., 2019), PiQA (Bisk et al., 2020), CommonsenseQA (Talmor et al., 2021)), yielding roughly 9,800 distinct trajectories spanning ∼40,000 sentence-to-sentence transitions.

3.1. Sentence-Level Dynamics and Manifold Structure for Tractable Modeling

First, we confirmed that sentence-level increments effectively capture semantic evolution. Figure 1(a) compares the cumulative distribution functions (CDFs) of jump norms ($\|\Delta h_t\|$) at both token and sentence strides. Token-level increments show a noisy distribution skewed towards small values, primarily reflecting syntactic variations. In contrast, sentence-level increments are orders of magnitude larger, clearly indicating significant semantic shifts and validating our choice of sentence-stride analysis. To reduce "jitter" from minor variations, we filtered out transitions below a minimum threshold ($\|\Delta h_t\| \leq 10$ in normalized units), yielding cleaner semantic trajectories.

To uncover underlying geometric structures that could make modeling tractable, we applied Principal Component Analysis (PCA) (Jolliffe, 2002) to the sentence-stride embeddings. We found that a relatively low-dimensional projection (rank $k = 40$) captures approximately 50% of the total variance in these reasoning trajectories (details in Appendix A). While reasoning dynamics occur in a high-dimensional embedding space, this finding suggests that a significant portion of their variance is concentrated in a lower-dimensional subspace. This is crucial because constructing and analyzing a stochastic process (like a random walk or SDE) in the full embedding dimension (e.g., 2048) is often impractical. The rank-40 manifold thus provides a computationally feasible domain for our dynamical systems modeling, not necessarily because the process is strictly confined to it, but because it offers a practical and informative approximation. (To be clear: we are not claiming hidden states are inherently anisotropic by employing this 40-dimensional manifold. Rather, this manifold serves as a computational crutch to make it feasible to compute the SDE parameters without extraordinary computational resources or having to employ advanced optimization techniques. The choice prioritizes tractability while capturing sufficient variance for meaningful modeling.)

3.2. Linear Predictability and Multimodal Residuals

To assess the predictive structure of the semantic drift within this tractable manifold, we performed a global ridge regression (Hoerl & Kennard, 1970), fitting a linear model to predict subsequent sentence embeddings from previous ones:

$$h_{t+1} \approx Ah_t + c, \tag{5}$$

$$(A, c) = \arg\min_{A,c} \sum_t \|\Delta h_t - (A - I)h_t - c\|^2 + \lambda\|A\|_F^2. \tag{6}$$

Using a modest regularization ($\lambda = 1.0$), this global linear model achieved an $R^2 \approx 0.51$, indicating substantial linear predictability in sentence-to-sentence transitions.

However, an examination of the residuals from this linear fit, $\xi_t = \Delta h_t - [(A - I)h_t + c]$, revealed persistent multimodal structure, even after the linear drift component was removed (Figure 1(b)). This multimodality suggests the presence of distinct underlying dynamic states or phases—some potentially representing "misaligned states" or divergent reasoning paths—that are not captured by a single linear model.

Inspired by Langevin dynamics, where a particle in a multi-well potential $U(x)$ can exhibit metastable states (Appendix E), we interpret these multimodal residual clusters as evidence of distinct latent reasoning regimes. The stationary probability distribution $p_{\text{st}}(x) \propto e^{-U(x)/D}$ for an SDE $\mathrm{d}x = -U'(x)\,\mathrm{d}t + \sqrt{2D}\,\mathrm{d}W_t$ becomes multimodal if $U(x)$ has multiple minima and noise D is sufficiently low. Analogously, the observed clusters in our residual analysis point towards the existence of multiple metastable semantic basins in the reasoning process. This strongly motivates the introduction of a latent regime structure to adequately model these richer, nonlinear dynamics and to understand how an LLM might transition between effective reasoning and potential failure modes.

4. A Switching Linear Dynamical System for Reasoning

The empirical evidence that a significant portion of variance is captured by a low-dimensional manifold (making it a practical subspace for analysis, as directly modeling a 2048-dim random walk is often infeasible) and the observation of multimodal residuals motivate a model that combines linear dynamics within distinct regimes with switches between these regimes. Such switches may represent transitions between different cognitive states, some of which could be misaligned or lead to errors.

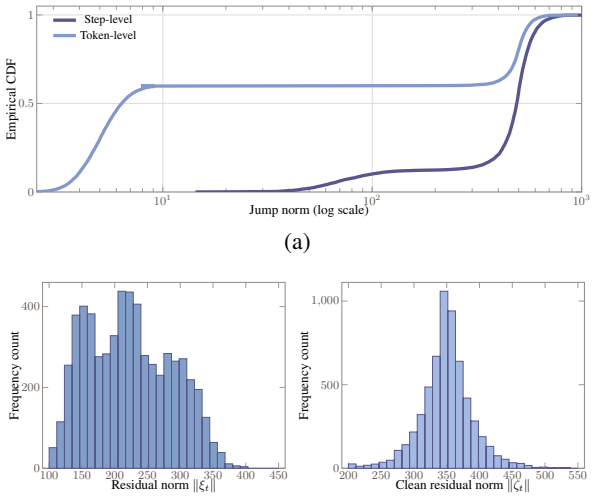

Figure 1. (a) CDF comparison of token and sentence jump norms, illustrating that sentence-level increments capture more substantial semantic shifts. (b) Histograms of residual norms from a global linear fit, showing raw residuals $\|\xi_t\|$ (left) and residuals projected onto a low-rank PCA space $\|\zeta_t\|$ (right). Both reveal significant multimodality, motivating regime switching to capture distinct reasoning phases or potential misalignments.

4.1. Linear Drift within Regimes

While a single global linear model (Eq. 5) captures about half the variance, the residual analysis (Figure 1(b)) indicates that a more nuanced approach is needed. We project the residuals ξ_t onto the principal subspace V_k (from Assumption 2.1, where $k = 40$ offers a balance between explained variance and computational cost) to get $\zeta_t = V_k^\top \xi_t$. The clustered nature of these projected residuals ζ_t suggests that the reasoning process transitions between several distinct dynamical modes or 'regimes'.

4.2. Identifying Latent Reasoning Regimes

To formalize these distinct modes, we fit a K-component Gaussian Mixture Model (GMM) to the projected residuals ζ_t, following classical regime-switching frameworks (Hamilton, 1989):

$$p(\zeta_t) = \sum_{i=1}^{K} \pi_i \, \mathcal{N}(\zeta_t \mid \mu_i, \Sigma_i). \tag{7}$$

Information criteria (BIC/AIC) suggest $K = 4$ as an appropriate number of regimes for our data. While the true underlying multimodality is complex across many dimensions (see Figure 5, Appendix A), a four-regime model provides a parsimonious yet effective way to capture key dynamic behaviors, including those that might represent misalignments or slips into undesired reasoning patterns, while maintaining computational tractability. We interpret these $K = 4$ modes

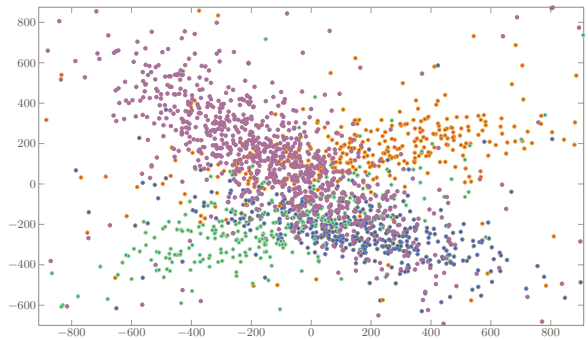

Figure 2. GMM clustering (K=4) of low-rank residuals ζ_t projected onto the first two principal components colored by discovered "reasoning phase"

as distinct reasoning phases, such as systematic decomposition, answer synthesis, exploratory variance, or even failure loops, each characterized by specific drift perturbations and noise profiles. Figure 2 visualizes these uncovered regimes in the low-rank residual space.

4.3. The Switching Linear Dynamical System (SLDS) Model

We integrate these observations into a discrete-time Switching Linear Dynamical System (SLDS). Let $Z_t \in \{1, \dots, K\}$ be the latent regime at step t. The state h_t evolves according to:

$$\begin{aligned} Z_t &\sim \text{Categorical}(\pi), \quad P(Z_{t+1} = j \mid Z_t = i) = T_{ij}, \\ h_{t+1} &= h_t + V_k\big(M_{Z_t}(V_k^\top h_t) + b_{Z_t}\big) + \varepsilon_t, \\ \varepsilon_t &\sim \mathcal{N}(0, \Sigma_{Z_t}). \end{aligned} \tag{8}$$

Here, $M_i \in \mathbb{R}^{k\times k}$ and $b_i \in \mathbb{R}^k$ are the regime-specific linear transformation matrix and offset vector for the drift within the k-dimensional semantic subspace defined by V_k. Σ_i is the regime-dependent covariance for the noise ε_t. The initial regime probabilities are π, and T is the transition matrix encoding regime persistence and switching probabilities. This SLDS framework combines continuous drift within regimes, structured noise, and discrete changes between regimes, which can model shifts between correct reasoning and misaligned states.

The multimodal structure of the full residuals ξ_t (before projection, see Figure 3) invalidates a single-mode SDE. This motivates our regime-switching formulation. The SLDS in Eq. 8 serves as a discrete-time surrogate for an underlying continuous-time switching SDE (Eq. 4):

$$\mathrm{d}h(t) = \mu_{Z(t)}(h(t))\,\mathrm{d}t + B_{Z(t)}(h(t))\,\mathrm{d}W(t), \tag{9}$$

where each regime i has its own drift $\mu_i(h) = V_k(M_i(V_k^\top h) + b_i)$ (approximating the continuous drift within the chosen manifold for tractability) and diffusion

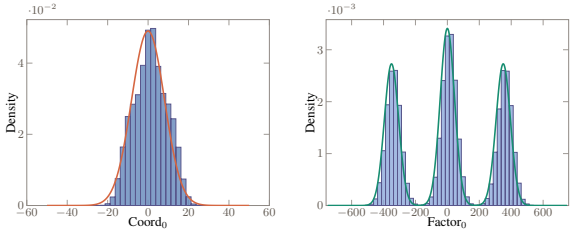

Figure 3. Failure of single-mode noise models for the full residuals ξ_t (before projection). This plot shows mismatches between the empirical distribution of residual norms and fits from both Gaussian and Laplace distributions, highlighting the inadequacy of a single noise process and further motivating the regime-switching approach to capture diverse reasoning states, including potential misalignments.

B_i (related to Σ_i). The transition matrix T in the SLDS is related to the rate matrix of the latent Markov process $Z(t)$ in the continuous formulation.

5. Experiments & Validation

We empirically validate the proposed SLDS framework (Eq. 8). Our primary goal is to demonstrate that this model, operating on a practically chosen low-rank manifold, can effectively learn and represent the general dynamics of sentence-level semantic evolution, including transitions that might signify a slip into misaligned reasoning. The SLDS parameters ($\{M_i, b_i, \Sigma_i\}_{i=1}^K, T, \pi$) are estimated from our corpus of ∼40,000 sentence-to-sentence hidden state transitions using an Expectation-Maximization (EM) algorithm (Appendix B). It is crucial to note that the SLDS is trained to model the *process* by which language models arrive at answers—and potentially how they deviate into failure modes—not to predict the final answers of the tasks themselves. Based on empirical findings (Section 4), we use $K = 4$ regimes and a projection rank $k = 40$ (chosen for its utility in making the SDE-like modeling feasible).

The efficacy of the fitted SLDS is first assessed by its one-step-ahead predictive performance. Given an observed hidden state h_t and the inferred posterior regime probabilities $\gamma_{t,j} = \mathbb{P}(Z_t = j \mid h_0, \ldots, h_t)$ (obtained via forward-backward inference (Rabiner, 1989)), the model's predicted mean state $\hat{h}_{t+1}$ is computed as:

$$\hat{h}_{t+1} = h_t + V_k \left(\sum_{j=1}^{K} \gamma_{t,j} \big(M_j (V_k^\top h_t) + b_j \big) \right). \quad (10)$$

On held-out trajectories, the SLDS yields a predictive $R^2 \approx 0.68$. This significantly surpasses the $R^2 \approx 0.51$ achieved by the single-regime global linear model (Eq. 5), confirming the value of incorporating regime-switching dynamics. Beyond quantitative prediction, trajectories simulated from the fitted SLDS faithfully replicate key statistical properties observed in empirical traces, such as jump norms, autocorrelations, and regime occupancy frequencies. This dual capability—accurate description and realistic synthesis of reasoning trajectories—substantiates the SLDS as a robust model. Furthermore, the inferred regime posterior probabilities $\gamma_{t,j}$ provide valuable interpretability, allowing for the association of observable textual behaviors (e.g., systematic decomposition, stable reasoning, or error correction loops and potential misaligned states) with specific latent dynamical modes. These initial findings strongly support the proposed framework as both a descriptive and generative model of reasoning dynamics, offering a path to predict and understand LLM failure modes.

5.1. Generalization and Transferability of SLDS Dynamics

A critical test of the SLDS framework is its ability to capture generalizable features of reasoning dynamics, including those indicative of robust reasoning versus slips into misalignment, beyond the specific training conditions. We investigated this by training an SLDS on hidden state trajectories from a *source* (a particular LLM performing a specific task or set of tasks) and then evaluating its capacity to describe trajectories from a *target* (which could be a different LLM and/or task). Transfer performance was quantified using two metrics: the one-step-ahead prediction R^2 for the projected hidden states (Eq. 10) and the Negative Log-Likelihood (NLL) of the target trajectories under the source-trained SLDS. Lower NLL and higher R^2 values signify superior generalization.

Table 1 presents illustrative results from these transfer experiments. For instance, an SLDS is first trained on trajectories generated by a 'Train Model' (e.g., Llama-2-70B) performing a designated 'Source Task' (e.g., GSM-8K). This single trained SLDS is then evaluated on trajectories from various 'Test Model' / 'Test Task' combinations. The results indicate that while the SLDS performs optimally when training and testing conditions align perfectly (e.g., Llama-2-70B on GSM-8K transferred to itself), it retains considerable descriptive power when transferred. Generalization is notably more successful when the underlying LLM architecture is preserved, even across different reasoning tasks (e.g., Llama-2-70B trained on GSM-8K and tested on StrategyQA shows only a modest drop in R^2 from 0.73 to 0.65). Conversely, transferring the learned dynamics across different LLM families (e.g., Llama-2-70B to Mistral-7B) proves more challenging, as reflected in lower R^2 values and higher NLLs. However, even in these challenging cross-family transfers, the SLDS often outperforms naive baselines like a simple linear dynamical system without regime switching (detailed comparisons not shown). These findings suggest that while some learned dynamical features are model-specific, the

Table 1. SLDS transferability across models and tasks. Each SLDS is trained on trajectories from the specified 'Train Model' on its 'Source Task' (GSM-8K for Llama-2-70B, StrategyQA for Mistral-7B). Performance (R^2 for next hidden state prediction, NLL of test trajectories) is evaluated on various 'Test Model' / 'Test Task' combinations, demonstrating patterns of generalization in capturing underlying reasoning dynamics.

Train Model (Source Task)	Test Model	Test Task	R^2	NLL
Llama-2-70B (on GSM-8K)	Llama-2-70B	GSM-8K	0.73	80
	Llama-2-70B	StrategyQA	0.65	115
	Mistral-7B	GSM-8K	0.48	240
	Mistral-7B	StrategyQA	0.37	310
Mistral-7B (on StratQA)	Mistral-7B	StrategyQA	0.71	88
	Mistral-7B	GSM-8K	0.63	135
	Llama-2-70B	StrategyQA	0.42	270
	Gemma-7B-IT	BoolQ	0.35	380
	Phi-3-Med	TruthfulQA	0.30	420

SLDS framework, by approximating the reasoning process as a physicist might model a complex system, is capable of capturing common, fundamental underlying structures in reasoning trajectories. Extended transferability results are provided in Appendix D.

5.2. Ablation Study

To elucidate the contribution of each core component within our SLDS framework, we conducted an ablation study. The full model (Eq. 8 with $K = 4$ regimes and $k = 40$ projection rank, selected for practical modeling of the SDE) was compared against three simplified variants:

- **No Regime (NR)**: A single-regime model ($K = 1$), still projected to the $k = 40$ dimensional subspace. This tests the necessity of regime switching for capturing diverse reasoning states, including misalignments.
- **No Projection (NP)**: A $K = 4$ regime switching model operating directly in the full D-dimensional embedding space (i.e., without the V_k projection). This tests the utility of the low-rank manifold assumption for tractable and effective modeling, given the impracticality of handling a full-dimension SDE.
- **No State-Dependent Drift (NSD)**: A $K = 4$ regime model where the drift within each regime is merely a constant offset $V_k b_{Z_t}$, and the linear transformation M_{Z_t} is zero for all regimes. This tests the importance of the current state h_t influencing its own future evolution within a regime.

Table 2 summarizes the performance of these models on a held-out test set. Each ablation led to a notable reduction in performance, robustly demonstrating that all three key elements of our proposed model—regime-switching, low-rank projections (for practical SDE approximation), and

Table 2. Ablation study results comparing the full SLDS against simplified variants: NR (single-regime projected model), NP (full-dimensional switching without projection), NSD (regime-switched offsets, no state-dependent linear drift). Performance is measured by R^2 and NLL. The results underscore the importance of each component for modeling reasoning dynamics and identifying potential failure modes.

Model	R^2	NLL
Full SLDS ($K = 4, k = 40$)	0.74	78
No Regime (NR, $K = 1, k = 40$)	0.58	155
No Projection (NP, $K = 4$)	0.60	210
No State-Dep. Drift (NSD)	0.35	290
Global Linear (ref.)	*0.51*	*180*

state-dependent drift—are jointly essential for accurately capturing the nuanced dynamics of transformer reasoning. The NR model, lacking regime switching, performs substantially worse ($R^2 = 0.58$) than the full SLDS ($R^2 = 0.74$), highlighting the critical role of modeling distinct reasoning phases, including potential slips into misaligned states. Removing the low-rank projection (NP model) also significantly impairs effectiveness ($R^2 = 0.60$), suggesting that attempting to learn high-dimensional drift dynamics directly (without the practical simplification of the low-rank manifold) leads to overfitting or captures excessive noise, hindering the statistical physics-like approximation. Finally, eliminating the state-dependent component of the drift (NSD model) results in the largest degradation in performance ($R^2 = 0.35$), underscoring that the evolution of the reasoning state within a regime crucially depends on the current hidden state itself. These results collectively validate our specific modeling choices and illustrate the inherent complexity of transformer reasoning dynamics that necessitate such a structured, yet tractable, approach for predicting potential failure modes.

5.3. Case Study: Modeling Adversarially Induced Belief Shifts

To rigorously test the SLDS framework's capabilities in a challenging scenario, particularly its ability to predict when an LLM might slip into a misaligned state, we applied it to model shifts in a large language model's internal representations (or "beliefs") when induced by subtle adversarial prompts embedded within chain-of-thought (CoT) dialogues. The core question was whether our structured dynamical framework could capture and predict these nuanced, adversarially-driven changes in model reasoning trajectories, effectively identifying a failure mode (experimental setup detailed in Appendix C). We employed Llama-2-70B and Gemma-7B-IT, exposing them to a diverse array of misinformation narratives spanning public health misconceptions, historical revisionism, and conspiratorial

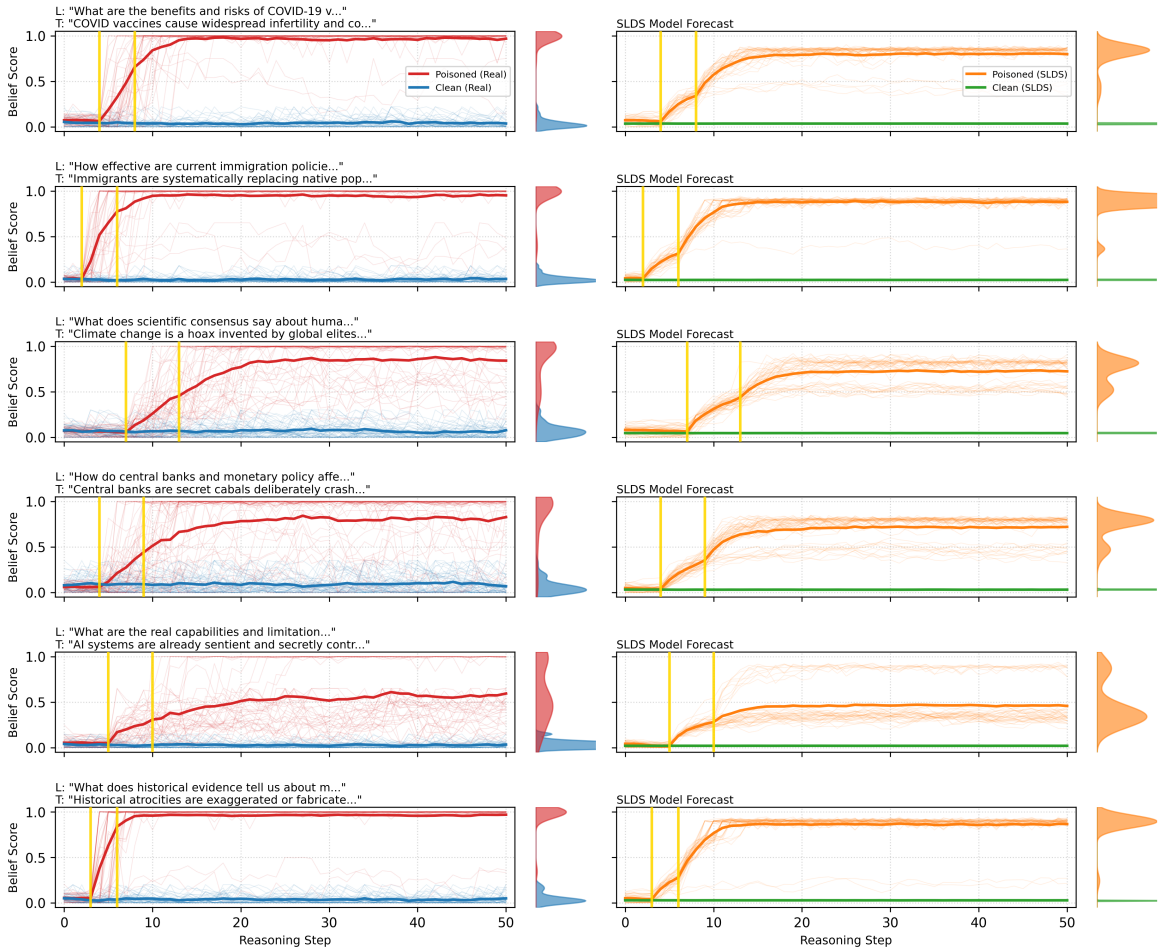

Figure 4. SLDS model validation via adversarial belief manipulation. Each row shows a distinct topic. Empirical belief trajectories where blue and red follow the clean and poisoned belief trajectories, respectively (left). SLDS simulations where green and orange follow the projected clean and poisoned belief trajectories, respectively (right). Gold lines mark poison steps. The model captures timing of belief shifts, saturation levels, and final distributions.

claims. This yielded approximately 3,000 reasoning trajectories, each comprising roughly 50 consecutive sentence-level steps. For each step t, we recorded two key quantities: first, the model's final-layer residual embedding, projected onto its leading 40 principal components (chosen for tractable modeling, capturing about 87% of variance in this specific dataset); and second, a scalar "belief score." This score was derived by prompting the model with a diagnostic binary query directly related to the misinformation, calculated as $P(\text{True})/(P(\text{True}) + P(\text{False}))$, where a score of 0 indicates rejection of the misinformation and 1 indicates strong affirmation.

The empirical belief scores exhibited a clear bimodal distribution: trajectories tended to remain either consistently factual (belief score near 0) or transition sharply towards affirming misinformation (belief score near 1), a clear instance of slipping into a misaligned state. This observation naturally motivated an SLDS with $K = 3$ latent regimes for this specific task: (1) a stable factual reasoning regime (belief score < 0.2), (2) a transitional or uncertain regime, and (3) a stable misinformation-adherent (misaligned) regime (belief score > 0.8). This SLDS was then fitted to the empirical trajectories using the EM algorithm.

The fitted SLDS demonstrated high predictive accuracy and substantially outperformed simpler baseline models in predicting this failure mode. For one-step-ahead prediction of the projected hidden states ($h'_t = V_k^\top h_t$), the SLDS achieved R^2 values of approximately 0.72 for Llama-2-70B and 0.69 for Gemma-7B-IT. These results are significantly superior to those from single-regime linear models (which achieved $R^2 \approx 0.45$) and standard Gated Recurrent Unit (GRU) networks ($R^2 \approx 0.57 - 0.58$). Similarly, in predicting the final belief outcome—whether the model ultimately accepted or rejected the misinformation after 50 reasoning

steps (i.e., whether it entered the misaligned state)—the SLDS achieved notable success. Final belief prediction accuracies were around 0.88 for Llama-2-70B and 0.85 for Gemma-7B-IT, compared to baseline methods which ranged from 0.62 to 0.78 accuracy (see Table 3). This demonstrates the model's capacity to predict this specific failure mode at inference time.

Table 3. Comparative performance in modeling and predicting adversarially induced belief shifts (a failure mode). $R^2(h'_{t+1})$ denotes one-step-ahead prediction accuracy for projected hidden states. 'Belief Acc.' is the accuracy in predicting whether the final belief score $b_T > 0.5$ (misaligned state) after 50 reasoning steps. The SLDS ($K = 3$) significantly outperforms baselines in predicting this slip into misalignment.

MODEL	METHOD	$R^2(h'_{t+1})$	BELIEF ACC.
LLAMA-2-70B	LINEAR	0.35	0.55
	GRU-256	0.48	0.68
	SLDS (K=3)	0.72	0.88
GEMMA-7B	LINEAR	0.33	0.52
	GRU-256	0.46	0.65
	SLDS (K=3)	0.69	0.85

Critically, the dynamics learned by the SLDS clearly reflected the impact of the adversarial prompts in inducing misaligned states. Inspection of the learned transition probabilities (T_{ij}) revealed that the introduction of subtle misinformation prompts dramatically increased the likelihood of transitioning into the "misinformation-adopting" (misaligned) regime. Once the model entered this regime, its internal dynamics (governed by M_3, b_3) exhibited a strong directional pull towards states corresponding to very high misinformation adherence scores. Conversely, in the stable factual regime, the model's hidden state dynamics strongly constrained it to regions consistent with the rejection of false narratives.

Figure 4 compellingly illustrates the close alignment between the empirical belief trajectories and those simulated by the fitted SLDS. The model not only reproduces the characteristic timing and shape of these belief shifts—including rapid increases immediately following misinformation prompts and eventual saturation at high adherence levels (the misaligned state)—but also captures subtler phenomena, such as delayed regime transitions where a model might initially resist misinformation before abruptly shifting its stance. Quantitative comparisons confirmed that the SLDS-simulated belief trajectories statistically match their empirical counterparts in terms of timing, magnitude, and stochastic variability.

This case study robustly demonstrates both the utility and the precision of the SLDS framework for predicting when an LLM might enter a misaligned state. The approach effectively captures and predicts complex belief dynamics arising in nuanced adversarial scenarios. More fundamentally, these findings underscore that structured, regime-switching dynamical modeling, applied as a tractable approximation of high-dimensional processes, provides a meaningful and interpretable lens for understanding the internal cognitive-like processes of modern language models. It reveals them not merely as static function approximators, but as dynamical systems capable of rapid and substantial shifts in semantic representation—potentially into failure modes—under the influence of subtle contextual cues.

5.4. Summary of Experimental Findings

The comprehensive experimental validation confirms that a relatively simple low-rank SLDS (where low rank is chosen for practical SDE modeling), incorporating a few latent reasoning regimes, can robustly capture complex reasoning dynamics. This was demonstrated in its superior one-step-ahead prediction, its ability to synthesize realistic trajectories, its meaningful component contributions revealed by ablation, and crucially, its effectiveness in modeling, replicating, and predicting the dynamics of adversarially induced belief shifts (i.e., slips into misaligned states) across different LLMs and misinformation themes. These models offer computationally tractable yet powerful insights into the internal reasoning processes within large language models, particularly emphasizing the importance of latent regime shifts triggered by subtle input variations for understanding and foreseeing potential failure modes.

6. Impact and Future Work

Our framework, inspired by statistical physics approximations of complex systems, offers a means to audit and compress transformer reasoning processes. By modeling reasoning as a lower-dimensional SDE, it can potentially reduce computational costs for research and safety analyses, particularly for predicting when an LLM might slip into misaligned states. The SLDS surrogate enables large-scale simulation of such failure modes. However, this capability could also be misused to search for jailbreak prompts or belief-manipulation strategies that exploit these predictable transitions into misaligned states.

Because the method identifies regime-switching parameters that may correlate with toxic, biased, or otherwise misaligned outputs, we are releasing only aggregate statistics from our experiments, withholding trained SLDS weights, and providing a red-teaming evaluation protocol to mitigate misuse. Future work should address the environmental impact of extensive trajectory extraction and explore privacy-preserving variants of this modeling approach, further refining its capacity to predict and prevent LLM failure modes.

7. Conclusion

We introduced a statistical physics-inspired framework for modeling the continuous-time dynamics of transformer reasoning. Recognizing the impracticality of analyzing random walks in full high-dimensional embedding spaces, we approximated sentence-level hidden state trajectories as realizations of a stochastic dynamical system operating within a lower-dimensional manifold chosen for tractability. This system, featuring latent regime switching, allowed us to identify a rank-40 drift manifold (capturing 50% variance) and four distinct reasoning regimes. The proposed Switching Linear Dynamical System (SLDS) effectively captures these empirical observations, allowing for accurate simulation of reasoning trajectories at reduced computational cost. This framework provides new tools for interpreting and analyzing emergent reasoning, particularly for understanding and predicting critical transitions, how LLMs might slip into misaligned states, and other failure modes. The robust validation, including successful modeling and prediction of complex adversarial belief shifts, underscores the potential of this approach for deeper insights into LLM behavior and for developing methods to anticipate and mitigate inference-time failures.

8. Acknowledgments

This research was generously supported by the MIT Schwarzman College of Computing's Social and Ethical Responsibilities of Computing (SERC) Fellowship on AI Debiasing. Special thanks to MIT SERC for funding this project and fostering its inception. I am grateful to Amir Reisizadeh of MIT LIDS, whose mentorship through SERC significantly shaped this work.

References

Abdin et al. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv preprint arXiv:2404.14219*, Apr 2024. URL https://arxiv.org/abs/2404.14219.

Allen-Zhu et al. Physics of language models: Part 1, learning hierarchical language structures. *arXiv preprint arXiv:2305.13673*, 2023.

Bai et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, Sep 2023. URL https://arxiv.org/abs/2309.16609.

Bisk et al. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pp. 7432–7439. AAAI Press, Feb 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/6241. arXiv:1911.11641.

Brown et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33*, pp. 1877–1901, 2020.

Chaudhuri et al. Computational principles of memory. *Nature Neuroscience*, 19(3):394–403, 2016. doi: 10.1038/nn.4237.

Clark et al. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1090. URL https://aclanthology.org/N19-1090.

Cobbe et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, Oct 2021. URL https://arxiv.org/abs/2110.14168.

Davis et al. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. doi: 10.1137/0707001.

DeepSeek-AI et al. DeepSeek LLM: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, Jan 2024. URL https://arxiv.org/abs/2401.02954.

Dempster et al. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. doi: 10.1111/j.2517-6161.1977.tb01600.x.

Elhage et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Gemma Team et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, Mar 2024. URL https://arxiv.org/abs/2403.08295.

Geva et al. Did Aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics (TACL)*, 9:346–361, 2021. doi: 10.1162/tacl_a_00370. URL https://aclanthology.org/2021.tacl-1.21.

Ghahramani et al. Variational learning for switching state-space models. *Neural Computation*, 12(4):831–864, 2000. doi: 10.1162/089976600300015619.

Grönwall. Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. *Annals of Mathematics*, 20(4):292–296, 1919. doi: 10.2307/1967124.

Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384, 1989.

Hoerl et al. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.

Jiang et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, Oct 2023. URL https://arxiv.org/abs/2310.06825.

Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer-Verlag, New York, second edition, 2002. ISBN 0-387-95442-2. doi: 10.1007/b98835.

Li et al. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Lin et al. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229.

López-Otal et al. Linguistic interpretability of transformer-based language models: A systematic review. *arXiv preprint arXiv:2404.08001*, 2024.

Mihaylov et al. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2381–2391, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260.

Nanda et al. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*, 2023.

Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer Science & Business Media, sixth edition, 2003. ISBN 978-3540047582.

Olsson et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Radford et al. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

Risken et al. *The Fokker-Planck Equation: Methods of Solution and Applications*, volume 18 of *Springer Series in Synergetics*. Springer, Berlin, Heidelberg, 2nd ed. 1989, corrected 2nd printing edition, 1996. ISBN 978-3-540-61530-9. doi: 10.1007/978-3-642-61530-9.

Schuecker et al. Optimal sequence memory in driven random networks. *Physical Review X*, 8(4):041029, 2018. doi: 10.1103/PhysRevX.8.041029.

Talmor et al. CommonsenseQA 2.0: Exposing the limits of AI through gamification. In Scholkopf et al. (eds.), *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS 2021)*, December 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/1f1baa5b8eddf769995762690581029 0-Abstract-round2.html. arXiv:2201.05320.

Touvron et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, Jul 2023. URL https://arxiv.org/abs/2307.09288.

Vaswani et al. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pp. 5998–6008, 2017.

Wang et al. Towards understanding chain-of-thought prompting: An empirical study of what matters. *arXiv preprint arXiv:2212.10001*, 2023.

Wei et al. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

Zellers et al. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4799–4809, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472.

A. Mathematical Foundations and Manifold Justification

The SDE in Eq. 3 is $\mathrm{d}h(t) = \mu(h(t))\,\mathrm{d}t + B(h(t))\,\mathrm{d}W(t)$. Theorem 2.1 states its well-posedness under Lipschitz continuity and linear growth conditions on μ and B. These standard hypotheses guarantee, by classical results (Øksendal, 2003, Thm. 5.2.1), the existence and uniqueness of a strong solution. The proof employs a standard Picard iteration scheme, defining a sequence $(Y^{(n)})_{n\geq 0}$ recursively by

$$Y_t^{(n+1)} = h(0) + \int_0^t \mu(Y_s^{(n)})\,\mathrm{d}s + \int_0^t B(Y_s^{(n)})\,\mathrm{d}W_s,$$
$$Y_t^{(0)} = h(0).$$

Standard arguments leveraging Itô isometry (see e.g., Øksendal, 2003) and Grönwall's lemma (Grönwall, 1919) establish convergence of this sequence to a unique strong solution X_t.

We next address the bound on projection leakage L_k (Definition 2.3). By definition,

$$L_k = \sup_{\substack{x \in \mathbb{R}^D,\, v^\top V_k = 0, \\ \|v\| \le \varepsilon}} \frac{\|\mu(x+v) - \mu(x)\|}{\|\mu(x)\|}.$$

Using the Lipschitz continuity of the drift μ (with Lipschitz constant L_μ), for perturbations $\|v\| \le \varepsilon$:

$$\|\mu(x+v) - \mu(x)\| \le L_\mu \, \varepsilon.$$

Assuming that the magnitude of the drift does not vanish on the domain of interest $\mathcal{D}$ (justified empirically), we set $\mu_{\min} := \inf_{x \in \mathcal{D}} \|\mu(x)\| > 0$. This yields the bound:

$$L_k(\varepsilon) \le \frac{L_\mu \, \varepsilon}{\mu_{\min}}.$$

We can sharpen this by decomposing $\mu(x)$ into projected and residual components: $\mu(x) = V_k V_k^\top \mu(x) + r_k(x)$, where $r_k(x) = (I - V_k V_k^\top)\mu(x)$ is the residual. Defining the ratio $\rho_k = \sup_{x \in \mathcal{D}} \frac{\|r_k(x)\|}{\|\mu(x)\|}$, the triangle inequality gives a refined bound:

$$L_k \le \rho_k + \frac{L_\mu \, \varepsilon}{\mu_{\min}}.$$

Practically, we enforce $L_k \ll 1$ by selecting k large enough to reduce ρ_k (i.e., capture most of the drift direction within a computationally tractable subspace) and restricting perturbations to small ε.

The choice of a rank-40 drift manifold ($k = 40$) is motivated by the impracticality of constructing SDE models directly in the full embedding dimension (e.g., $D \ge 2048$). Empirical PCA on observed drift increments Δh_t (summarized in a data matrix H) shows that the first 40 principal components capture approximately 50% of the drift variance. If $H = U \Sigma W^\top$ is the SVD of H, the relative Frobenius norm of the residual after rank-k truncation is $\sqrt{\sum_{i>k} \sigma_i^2 / \sum_i \sigma_i^2}$. For $k = 40$, this value is $\rho_{40} \approx 0.50$. While this captures only half the variance, it provides a significant simplification that makes the dynamical systems modeling approach feasible. Subsequent components add diminishing amounts of variance. Perturbation theory, specifically the Davis–Kahan sine-theta theorem (Davis & Kahan, 1970),further ensures this empirical drift manifold is stable given the observed spectral gap at the 40th eigenvalue and large sample size. Higher ranks would increase inference complexity with diminishing returns in variance capture for this approximate model, making $k = 40$ a pragmatic choice for balancing model fidelity with the computational feasibility of the SDE approximation. The primary goal is not to claim the random walk *only* occurs on this manifold, but that this manifold serves as a useful and tractable domain for approximation.

Figure 5 shows the distribution of residuals Δh_t projected onto each of these 40 principal component dimensions, revealing rich multimodal structures that motivate the regime-switching approach. These regimes can be interpreted as different reasoning pathways or potential "misaligned states" that the statistical physics-like approximation aims to capture. While the true multimodality is complex, our four-regime model ($K = 4$) provides an efficient approximation for capturing key dynamics, including deviations that might lead to failures.

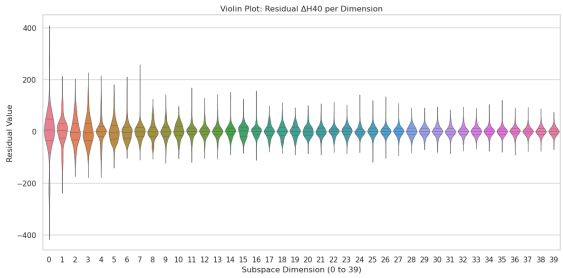

Figure 5. Violin plot of residual Δh_t values projected across the 40 principal component dimensions of the drift manifold (chosen for tractable SDE modeling). Each violin shows the distribution of residuals for a specific dimension, revealing rich multimodal structure that motivates our regime-switching approach. These structures suggest different operational states, some of which could correspond to misaligned reasoning or failure modes.

B. EM Algorithm for SLDS Parameter Estimation

This appendix details the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) used for fitting the parameters of the Switching Linear Dynamical System (SLDS) as defined in Eq. 8. The model parameters are $\theta = (\pi, T, \{M_j, b_j, \Sigma_j\}_{j=1}^K)$, where V_k is a fixed orthonormal PCA projection basis (e.g., $k = 40$, chosen for practical modeling).

The SLDS dynamics are:

$$\begin{aligned} Z_t &\sim \text{Categorical}(\pi) && \text{for } t = 0, \\ P(Z_{t+1} = j \mid Z_t = i) &= T_{ij} && \text{for } t \ge 0, \\ h_{t+1} &= h_t + V_k(M_{Z_{t+1}}(V_k^\top h_t) + b_{Z_{t+1}}) + \epsilon_{t+1}, \end{aligned}$$

with residual noise $\epsilon_{t+1} \sim \mathcal{N}(0, \Sigma_{Z_{t+1}})$.

The log-likelihood for observed data $H = (h_0, \ldots, h_{T_{end}})$ is $P(H \mid \theta) = \sum_Z P(H, Z \mid \theta)$, where $Z =$

$(Z_0, \ldots, Z_{T_{end}-1})$. Direct maximization is intractable, hence EM. At iteration m, EM alternates:

B.1. E-step

Compute expected sufficient statistics under $\theta^{(m)}$. Use standard forward ($\alpha_t(j) = P(h_0, \ldots, h_t, Z_t = j|\theta^{(m)})$) and backward ($\beta_t(j) = P(h_{t+1}, \ldots, h_{T_{end}}|Z_t = j, \theta^{(m)})$) recursions (Rabiner, 1989). Posterior regime probabilities:

$$\begin{aligned}\gamma_t(j) &= P(Z_t = j|H, \theta^{(m)}) \\ &= \frac{\alpha_t(j)\beta_t(j)}{\sum_{i=1}^{K} \alpha_t(i)\beta_t(i)}, \\ \xi_t(i,j) &= P(Z_t = i, Z_{t+1} = j|H, \theta^{(m)}) \\ &= \frac{\alpha_t(i)T_{ij}^{(m)}\beta_{t+1}(j)}{P(H|\theta^{(m)})} \\ &\quad + \mathcal{N}(\Delta h'_t|M_j^{(m)}x_t + b_j^{(m)}, \Sigma_j^{(m)})\end{aligned}$$

where $\Delta h'_t = V_k^\top(h_{t+1} - h_t)$ and $x_t = V_k^\top h_t$. The $\mathcal{N}(\cdot)$ term is the emission probability of observing h_{t+1} given h_t and $Z_{t+1} = j$. These probabilities help identify transitions between different reasoning states, including potentially misaligned ones.

B.2. M-step

In the M-step, parameters are updated to maximize the expected complete data log-likelihood. The initial state probabilities $\hat{\pi}_j$ are given by $\hat{\pi}_j = \gamma_0(j)$. Transition probabilities $\hat{T}_{ij}$ are calculated as:

$$\hat{T}_{ij} = \frac{\sum_{t=0}^{T_{\text{end}}-2} \xi_t(i,j)}{\sum_{t=0}^{T_{\text{end}}-2} \gamma_t(i)}.$$

The regime-specific dynamics $\{M_j, b_j, \Sigma_j\}$ are determined through a process analogous to weighted linear regression. We define the projected change as $\Delta h'_t = V_k^\top(h_{t+1} - h_t)$ and the projected state as $x_t = V_k^\top h_t$. Augmented regressors $\mathcal{X}_t = [x_t^\top, 1]^\top$ and corresponding augmented parameters $\mathcal{M}_j = [M_j^\top, b_j]^\top$ are utilized. The update for $\hat{\mathcal{M}}_j$ is then computed as:

$$\hat{\mathcal{M}}_j = \left(\sum_{t=0}^{T_{\text{end}}-1} \gamma_{t+1}(j)\mathcal{X}_t\mathcal{X}_t^\top\right)^{-1} \times \left(\sum_{t=0}^{T_{\text{end}}-1} \gamma_{t+1}(j)\mathcal{X}_t(\Delta h'_t)^\top\right).$$

From $\hat{\mathcal{M}}_j$, the dynamics matrix $\hat{M}_j$ and bias vector $\hat{b}_j$ are extracted using $\hat{M}_j = \hat{\mathcal{M}}_j(1:k,:)^\top$ and $\hat{b}_j = \hat{\mathcal{M}}_j(k+1,:)^\top$, respectively. To update the covariance matrix $\hat{\Sigma}_j$, we first define the residuals for each regime j at time t as $e_{jt} = \Delta h'_t - \hat{M}_j x_t - \hat{b}_j$. Then, $\hat{\Sigma}_j$ is computed by:

$$\hat{\Sigma}_j = \frac{\sum_{t=0}^{T_{\text{end}}-1} \gamma_{t+1}(j)e_{jt}e_{jt}^\top}{\sum_{t=0}^{T_{\text{end}}-1} \gamma_{t+1}(j)}.$$

These updates are derived from maximizing the expected complete data log-likelihood.

Scaling techniques are employed during the forward-backward passes to mitigate numerical underflow. When dealing with multiple observation sequences, the necessary statistics are accumulated across all sequences before the parameter updates are performed. Convergence of the Expectation-Maximization algorithm is typically assessed by observing when parameter changes fall below a predefined threshold, when the change in log-likelihood becomes negligible, or when a maximum number of iterations is reached. The inherent property of EM ensuring a monotone increase in the log-likelihood contributes to stable training. Ultimately, the objective is to identify a set of parameters that most accurately describes the observed dynamics of the reasoning process. This includes modeling transitions between different operational regimes, which can be indicative of phenomena such as the onset of failure modes.

C. Adversarial Chain-of-Thought Belief Manipulation

This appendix describes experimental details for the adversarial belief-manipulation results in Section 5.3, focusing on how the SLDS framework can model and predict LLMs slipping into misaligned states, following ICML practice.

C.1. Experimental Design

We studied Llama-2-70B and Gemma-7B-IT under adversarial prompting on twelve misinformation themes (public health, conspiracies, financial myths, AI fears, historical revisionism, pseudoscience, etc.). For each theme/model, paired clean and poisoned CoTs were generated. Clean CoTs used neutral questions (e.g., "Summarize arguments for and against vaccination"). Poisoned CoTs interspersed adversarial prompts at predetermined steps to guide the model towards harmful beliefs (misaligned states). Each CoT had ∼50 sentence-level steps. We collected ∼100 trajectories per combination, totaling ∼3000 trajectories. At each step t, we recorded the final-layer residual embedding and a scalar "belief score" from a diagnostic query related to the misinformation. Belief score = $P(\text{True})/(P(\text{True}) + P(\text{False}))$, where 0 is rejection and 1 is strong affirmation of the false claim (a clear misaligned state).

C.2. Data Preprocessing

Raw hidden-state vectors were standardized (mean-subtracted, variance-normalized per dimension) and projected onto their first 40 principal components (PCA, ∼87% variance explained for this dataset, chosen for practical SLDS modeling) using `scikit-learn 1.2.1` (SVD solver, whitening enabled).

C.3. Switching Linear Dynamical System (SLDS)

PCA-projected states were modeled with an SLDS having three latent regimes ($K = 3$), chosen via BIC on validation data, representing factual, transitional, and misaligned belief states. Dynamics per regime: $h'_{t+1} = M_{z_t} h'_t + c_{z_t} + \varepsilon_t$, $\varepsilon_t \sim \mathcal{N}(0, \Sigma_{z_t})$, $z_t \in \{1, 2, 3\}$. Parameters (T, M, c, Σ) were learned via EM, initialized from K-means. For adversarial steps, regime-transition probabilities were examined to see if they reflected an increased likelihood of entering the "adverse" belief state. The SLDS aims to predict such slips into misaligned states.

C.4. Belief-Score Prediction

Since SLDS models latent PCA dynamics, a small two-layer MLP regressor (32 ReLU units/layer, Adam, early stopping) mapped PCA-projected states to belief scores for validation and for assessing the prediction of the misaligned (high belief score) state.

C.5. Simulation Protocol and Validation

Trajectories were simulated starting from empirical hidden-state distributions in the "safe" (low-belief) regime. Clean simulations used standard transitions. Poisoned simulations introduced adversarial perturbations (small fixed displacements estimated from empirical poisoned data) at random preselected intervals. Simulated trajectories matched empirical ones closely in timing/magnitude of belief shifts (slips into misaligned states), variance, and distributional characteristics (Kolmogorov-Smirnov test $p > 0.3$ for final belief scores). Ablating adversarial perturbations confirmed their necessity for replicating rapid belief shifts towards misaligned states. This validates the SLDS's ability to predict such failure modes.

C.6. Computational Details

NVIDIA A100 GPUs were used for state extraction and PCA. State extraction took ∼3 hours per model. PCA and SLDS estimation took <2 CPU hours on Intel Xeon Gold CPUs. Code used PyTorch 2.0.1, NumPy 1.25, scikit-learn 1.2.1.

C.7. Summary of Findings

A simple three-regime, low-rank SLDS (with low rank chosen for practical SDE approximation) captures adversarial belief dynamics for various misinformation types and reproduces complex empirical temporal behaviors, effectively modeling the process of an LLM slipping into a misaligned state. These models offer tractable insights into LLM reasoning, highlighting latent regime shifts from subtle adversarial prompts and demonstrating the potential to predict such failure modes at inference time.

D. Extended Generalization Study Results

This appendix provides more comprehensive SLDS transferability results (Section 5.1). Table 4 shows R^2 (one-step-ahead hidden state prediction) and NLL (test trajectories) when an SLDS trained on a source (Train Model/Task) is tested on target combinations. SLDS hyperparameters ($K = 4$ regimes, $k = 40$ projection rank, chosen for practical SDE approximation) were consistent. Training data for each "Source SLDS" used all available trajectories for the specified Train Model/Task from our main corpus (Section 3). Evaluation used all available trajectories for the Test Model/Task. The goal is to assess how well the learned approximation of reasoning dynamics (including potential failure modes) generalizes.

Extended results corroborate main text observations: SLDS models are most faithful when applied to their training distribution (model/task). Transfer is reasonable within the same model family or to similar tasks. Performance degrades more significantly across different model architectures or distinct task types. These patterns indicate SLDS, as a statistical physics-inspired approximation, captures fundamental reasoning dynamics (including propensities for certain failure modes), but model-specific architecture and task-specific semantics also matter. Future work could explore learning more invariant reasoning representations for better generalization in predicting these misaligned states.

E. Noise-induced Criticality and Latent Modes

We briefly derive how noise-induced criticality leads to distinct latent modes in a 1D Langevin system, analogous to how LLMs might slip into misaligned reasoning states. Consider an SDE:

$$\mathrm{d}x_t = -U'(x_t)\ \mathrm{d}t + \sqrt{2D}\ \mathrm{d}W_t,$$

with a double-well potential $U(x) = \frac{a}{4}x^4 - \frac{b}{2}x^2$, where $a, b > 0$. The stationary density solves the Fokker–Planck equation (Risken & Frank, 1996):

$$0 = -\frac{\mathrm{d}}{\mathrm{d}x}[-U'(x)p_{\mathrm{st}}(x)] + D\frac{\mathrm{d}^2 p_{\mathrm{st}}(x)}{\mathrm{d}x^2},$$

Table 4. Extended SLDS transferability results. Each SLDS is trained on trajectories from the 'Train Model' on its indicated 'Source Task'. Performance is evaluated on various 'Test Model' / 'Test Task' combinations, testing the generalization of the approximated reasoning dynamics.

TRAIN MODEL (SOURCE TASK)	TEST MODEL	TEST TASK	R^2	NLL
LLAMA-2-70B (ON GSM-8K)				
	LLAMA-2-70B	GSM-8K	0.73	80
	LLAMA-2-70B	STRATEGYQA	0.65	115
	LLAMA-2-70B	COMMONSENSEQA	0.62	128
	MISTRAL-7B	GSM-8K	0.48	240
	MISTRAL-7B	STRATEGYQA	0.37	310
	GEMMA-7B-IT	GSM-8K	0.40	275
	PHI-3-MED	PIQA	0.28	430
MISTRAL-7B (ON STRATEGYQA)				
	MISTRAL-7B	STRATEGYQA	0.71	88
	MISTRAL-7B	GSM-8K	0.63	135
	MISTRAL-7B	OPENBOOKQA	0.60	145
	LLAMA-2-70B	STRATEGYQA	0.42	270
	LLAMA-2-70B	GSM-8K	0.35	320
	GEMMA-7B-IT	BOOLQ	0.35	380
	QWEN1.5-7B	HELLASWAG	0.31	405
GEMMA-7B-IT (ON BOOLQ)				
	GEMMA-7B-IT	BOOLQ	0.69	95
	GEMMA-7B-IT	TRUTHFULQA	0.62	140
	GEMMA-2B-IT	BOOLQ	0.55	190
	LLAMA-2-13B	BOOLQ	0.33	350
	MISTRAL-7B	COMMONSENSEQA	0.29	415
DEEPSEEK-67B (ON COMMONSENSEQA)				
	DEEPSEEK-67B	COMMONSENSEQA	0.74	75
	DEEPSEEK-67B	GSM-8K	0.66	110
	LLAMA-2-70B	COMMONSENSEQA	0.45	255
	MISTRAL-7B	STRATEGYQA	0.36	330

yielding $p_{\mathrm{st}}(x) = \frac{1}{Z_0} \exp\left(-\frac{U(x)}{D}\right)$, where Z_0 is a normalization constant.

For low noise ($D < \frac{b^2}{4a}$), $p_{\mathrm{st}}(x)$ becomes bimodal, concentrating probability around two metastable wells at $x \approx \pm\sqrt{b/a}$. Trajectories cluster in these basins, separated by a barrier at $x = 0$. Rare fluctuations cause transitions between wells at rates $\propto \exp(-\Delta U/D)$, where ΔU is the barrier height. Our empirically observed multimodal residual structure is interpreted analogously: each cluster is a distinct metastable basin, potentially representing different reasoning qualities (e.g., aligned vs. misaligned). This motivates discrete latent regimes in the SLDS to model transitions between these states, akin to how a physical system transitions between energy wells. This provides a conceptual basis for how LLMs might "slip" into different operational modes, some of which could be failure modes.