A Statistical Physics of Language Model Reasoning

Jack David Carson [1] **Amir Reisizadeh** [1]

Abstract

Transformer language models exhibit emergent reasoning capabilities that have largely resisted mechanistic understanding. We introduce a statistical physics-inspired framework to describe the continuous-time dynamics underlying chain-of-thought reasoning in large transformer models. Specifically, we analyze sentence-level hidden state trajectories as realizations of a low-dimensional stochastic dynamical system governed by drift-diffusion processes with latent regime switching. Using empirical trajectories extracted from eight open-source transformer models evaluated on seven diverse reasoning benchmarks, we identify a rank-40 drift manifold that explains approximately 50% of variance in reasoning trajectories, along with four distinct latent reasoning regimes. We then formulate and validate a switching linear dynamical system model capturing these empirical features. This framework allows simulation of transformer reasoning at significantly reduced computational cost, offering theoretical tools to study critical behavioral transitions and failure modes in large language models.

1. Introduction

Transformer language models (LMs), despite being trained solely for next-token prediction, exhibit emergent reasoning capabilities resembling cognitive processes (Vaswani et al., 2017; Radford et al., 2019; Brown et al., 2020; Wei et al., 2022). Traditional mechanistic analyses, focusing on discrete transformer components like attention heads and residual streams, offer limited insights into longer-scale semantic transitions arising during multi-step reasoning tasks (Elhage et al., 2021; Olsson et al., 2022; Allen-Zhu & Li, 2023; López-Otal et al., 2024).

[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.

Motivated by statistical physics, we propose a continuous-time stochastic dynamical system description of transformer reasoning. In this mesoscopic framework, sentence-level hidden states evolve via stochastic differential equations (SDEs), decomposing semantic trajectories into deterministic drift and stochastic fluctuations. Empirically, we show that transformer reasoning trajectories reside predominantly within a low-dimensional drift manifold, which we identify via principal component analysis (PCA). Specifically, rank-40 PCA projections of hidden states across our evaluation suite capture approximately 50% of total variance, revealing structured semantic evolution punctuated by stochastic jumps.

Further analysis uncovers four latent semantic regimes characterized by distinct drift directions and variances, suggesting context-driven regime switching. To model these phenomena, we introduce and validate a switching linear dynamical system (SLDS) framework, accurately reconstructing empirical reasoning trajectories with substantial computational savings. This model provides an efficient surrogate for analyzing critical transitions and robustness in transformer-based reasoning processes.

2. Background and Motivation

Transformer hidden states evolve as high-dimensional semantic trajectories, governed implicitly by learned next-token predictions (Vaswani et al., 2017; Radford et al., 2019). Despite detailed mechanistic studies of attention patterns and residual stream activations (Elhage et al., 2021; Olsson et al., 2022; Li et al., 2023; Nanda et al., 2023), capturing long-scale reasoning dynamics—spanning multiple sentences or complex cognitive steps—remains elusive.

Chain-of-thought (CoT) prompting reveals structured reasoning pathways (Wei et al., 2022; Wang et al., 2023), suggesting internal processes akin to stochastic dynamical systems. Prior continuous-time frameworks for neural dynamics, often inspired by statistical physics, offer principled tools for capturing complex emergent behavior (Chaudhuri & Fiete, 2016; Schuecker et al., 2018; Gardiner, 2004). However, explicit mesoscopic modeling of transformers at semantic timescales remains unexplored.

Here, motivated by these precedents, we pursue a mesoscopic, SDE-based perspective on transformer reasoning. This approach, by decomposing transformer hidden-state dynamics into deterministic drifts and stochastic fluctuations, provides both computational tractability and conceptual clarity. Through empirical analyses of transformer trajectories, we uncover structured semantic manifolds and regime transitions that inform our dynamical modeling strategy.

3. Preliminaries

We model the internal reasoning trajectory of transformer language models as a continuous-time stochastic process evolving over hidden-state representations. Formally, let $h_t \in \mathbb{R}^D$ denote the final-layer residual embedding extracted at discrete sentence boundaries indexed by $t = 0, 1, 2, \ldots$. To capture rich semantic dynamics across reasoning steps, we consider these discrete embeddings as observations of an underlying continuous-time process $h_t : \mathbb{R}_{\geq 0} \to \mathbb{R}^D$, governed by stochastic differential equations (SDEs).

Definition 3.1 (Itô SDE)**.** An Itô stochastic differential equation defined on state space $\mathbb{R}^D$ has the form

$$\mathrm{d}h_t = \mu(h_t)\,\mathrm{d}t + B(h_t)\,\mathrm{d}W_t, \quad h_0 \sim p_0, \tag{1}$$

where $\mu : \mathbb{R}^D \to \mathbb{R}^D$ denotes the deterministic *drift*, encoding persistent directional dynamics, and $B : \mathbb{R}^D \to \mathbb{R}^{D \times D'}$ specifies the diffusion matrix modulating instantaneous stochastic fluctuations. Here, W_t is a D'-dimensional Wiener process (standard Brownian motion), and p_0 represents an initial distribution. Note that $D' \leq D$ represents the dimension of the noise, which may be lower than the state dimension.

Existence and uniqueness of solutions are ensured under standard assumptions:

Theorem 3.1 (Well-Posedness (Øksendal, 2003))**.** *Under standard Lipschitz continuity and linear growth assumptions (the Lipschitz Assumption) on μ and B, the SDE*

$$\mathrm{d}h_t = \mu(h_t)\,\mathrm{d}t + B(h_t)\,\mathrm{d}W_t \tag{2}$$

admits a unique strong solution for a standard D'-dimensional Wiener process W_t (see Appendix A).

The drift term $\mu(h_t)$ intuitively corresponds to persistent semantic or cognitive tendencies, reflecting systematic progression of the model's internal state. In contrast, the diffusion term $B(h_t)$ represents instantaneous fluctuations arising from local uncertainties, token-level variations, or intrinsic model stochasticity.

Transformer embeddings admit multiple temporal resolutions depending on the granularity of hidden-state extraction. Specifically, we consider:

Definition 3.2 (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 after each detected sentence boundary.

This formulation emphasizes mesoscopic, semantic-level dynamics as opposed to token-level variations.

To rigorously study these dynamics within a computationally tractable space, we introduce a notion of projection-based dimensionality reduction and associated approximation quality:

Definition 3.3 (Projection Leakage)**.** Given an orthonormal matrix $V_k \in \mathbb{R}^{D \times k}$ (with $V_k^\top V_k = I_k$), the *leakage* of drift μ under orthogonal perturbations $v \perp \mathrm{Im}(V_k)$ 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 ensures approximate invariance of the drift within the chosen subspace:

Assumption 3.1 (Approximate Projection Closure)**.** There exists a rank k (e.g. $k = 40$) and perturbation scale $\epsilon > 0$ such that $L_k \ll 1$, guaranteeing the approximation

$$\mu(h_t) \approx V_k V_k^\top \mu(h_t)$$

up to an error on the order of $O(L_k)$.

Lastly, we introduce a regime-switching framework to capture discontinuous semantic transitions suggested by empirical observations of reasoning trajectories:

Definition 3.4 (Regime-Switching SDE)**.** Consider a latent continuous-time Markov chain $Z_t \in \{1, \ldots, K\}$ governed by transition rate matrix $T \in \mathbb{R}^{K \times K}$. The corresponding regime-switching Itô SDE takes the form

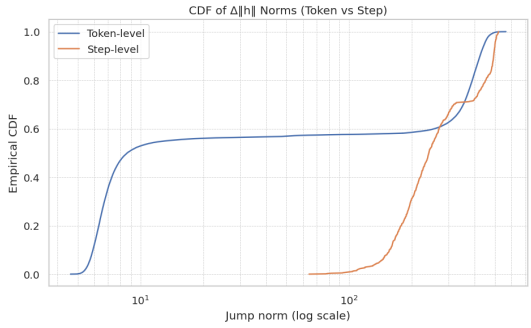
$$\mathrm{d}h_t = \mu_{Z_t}(h_t)\,\mathrm{d}t + B_{Z_t}(h_t)\,\mathrm{d}W_t, \quad Z_t \sim T_{ij}, \tag{3}$$

where each latent regime $i \in \{1, \ldots, K\}$ possesses distinct drift and diffusion functions μ_i, B_i, enabling context-dependent dynamic structures (Ghahramani & Hinton, 2000).

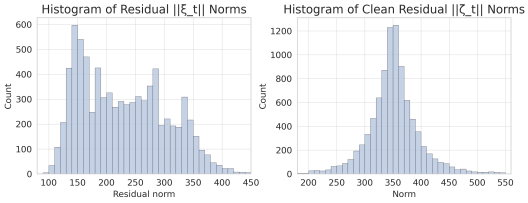
These preliminary definitions and assumptions provide the mathematical foundations necessary to rigorously characterize and analyze transformer reasoning dynamics within our proposed framework.

4. Data & Empirical Decomposition

We build a corpus of sentence-aligned hidden-state trajectories from transformer-generated reasoning chains across a suite of models (Mistral-7B-Instruct, Phi-3-Medium, DeepSeek-67B, Llama-2-70B, Gemma-2B-IT, Qwen1.5-7B-Chat, Gemma-7B-IT, Llama-2-13B-Chat-HF) and datasets (StrategyQA, GSM-8K (Cobbe et al., 2021), TruthfulQA, BoolQ, OpenBookQA, HellaSwag, PiQA, CommonsenseQA), yielding roughly 9,800 distinct trajectories spanning ∼40,000 sentence-to-sentence transitions. Sentence boundaries are identified via explicit textual markers (e.g., enumeration or punctuation), ensuring a consistent semantic granularity.



(a)



(b)

Figure 1. (a) CDF comparison of token and sentence jump norms. Sentence increments capture semantic dynamics distinctly. (b) Histograms of residual norm distributions: raw residuals $\|\xi_t\|$ (left) and cleaned residuals $\|\zeta_t\|$ (right). Residual norms before/after drift demonstrate powerful multimodality.

We first verify that sentence-level increments meaningfully isolate semantic evolution. Figure 2 contrasts jump-norm distributions at token and sentence strides. Token-level increments exhibit a noisy distribution heavily skewed toward small values, reflecting syntactic variation. In contrast, sentence-level increments are orders of magnitude larger, clearly capturing semantic shifts and justifying our sentence-stride approach. To further mitigate residual "jitter," we discard transitions below a minimal threshold ($\|\Delta h_t\| \leq 10$ in normalized embedding units), resulting in cleaner semantic trajectories.

To reveal underlying manifold structure, we apply principal component analysis (PCA) (Jolliffe, 2002) to sentence-stride embeddings. To probe the predictive structure of semantic drift, we perform global ridge regression (Hoerl & Kennard, 1970) fitting subsequent sentence embeddings from previous states:

$$h_{t+1} \approx Ah_t + c, \tag{4}$$

$$(A, c) = \arg\min_{A,c} \sum_t \|\Delta h_t - (A - I)h_t - c\|^2 + \lambda\|A\|_F^2. \tag{5}$$

With a modest regularization $\lambda = 1.0$, the regression attains an $R^2 \approx 0.51$, signifying substantial linear predictability in sentence-to-sentence reasoning transitions. However, examining residuals $\xi_t = \Delta h_t - [(A - I)h_t + c]$ reveals persistent multimodal structure even after removing linear drift (Figure 1). This observation strongly motivates introducing latent regime structure to fully characterize these richer nonlinear dynamics, as explored next.

5. Linear Drift and Residual Dynamics

Having established the presence of a low-dimensional semantic manifold governing sentence-stride trajectories, we now seek a rigorous characterization of its underlying dynamics. Initially, we consider the simplest candidate: a globally linear drift structure augmented by stochastic residuals, analogous to classical drift-diffusion models in cognitive science (Ratcliff & McKoon, 2008).

Formally, we posit that hidden-state increments approximately follow a discrete-time linear model:

$$h_{t+1} = A\, h_t + c + \xi_t, \quad \xi_t \sim p_\xi(\cdot), \tag{6}$$

where $A \in \mathbb{R}^{D\times D}$ is a linear operator capturing the dominant semantic drift, $c \in \mathbb{R}^D$ is an offset, and ξ_t denotes the residual stochastic component.

To robustly identify parameters A and c, we employ regularized linear regression (Ridge) on the observed corpus of sentence-aligned hidden-state transitions. Precisely, we solve:

$$\min_{A,c} \sum_t \|\Delta h_t - (A - I)h_t - c\|_2^2 + \lambda\, \|A\|_F^2\,,$$

where $\Delta h_t = h_{t+1} - h_t$, and λ controls regularization strength. Empirically, this linear fit explains roughly half of the observed variance ($R^2 \approx 0.51$), confirming a substantial deterministic component.

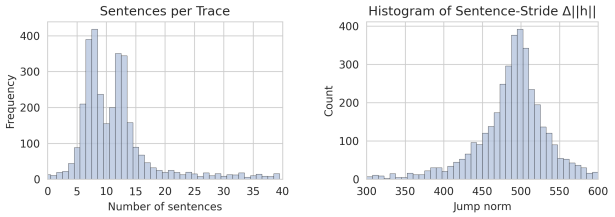
However, despite this linear structure, careful analysis of residual norms $\|\xi_t\|$ reveals persistent multimodality. To quantify residual structure, we project residuals onto the principal subspace V_k (Assumption 3.1):

$$\zeta_t = V_k^\top \xi_t, \quad \xi_t = \Delta h_t - [(A - I)h_t + c].$$

The resulting projected residuals exhibit non-Gaussian, clustered distributions (Figure 1). This suggests additional latent structure unaccounted for by linear drift alone, necessitating discrete latent reasoning regimes.

6. Regime-Switching Dynamics

While a single linear drift captures the dominant semantic flow, the multimodal residuals reflect abrupt shifts in reasoning behavior. To formalize these shifts, we project residuals ξ_t into the low-rank subspace via $\zeta_t = V_k^\top \xi_t$ and fit a K component Gaussian mixture, building on classical regime-switching frameworks (Hamilton, 1989):



(a) (b)

Figure 2. Latent regimes uncovered by GMM in low-rank residual space. (a) Residuals in PCA(2), colored by GMM component. (b) Histogram of $\|\zeta_t\|$ colored by regime.

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

with selection criteria (BIC/AIC) indicating $K = 4$ regimes. While the true multimodality exhibits considerably richer structure across dimensions (see Fig. 6, Appendix A), this parsimonious four-regime model efficiently captures misalignment behaviors while maintaining computational tractability.

Denote the posterior regime assignment by $Z_t = \arg\max_i \pi_i \mathcal{N}(\zeta_t \mid \mu_i, \Sigma_i)$. We interpret these four modes as distinct reasoning phases—such as systematic decomposition, answer synthesis, exploratory variance, and failure loops—each with characteristic drift perturbations and noise profiles.

Embedding this structure into a unified update yields a discrete-time Switching Linear Dynamical System (SLDS):

$$\begin{aligned} &Z_t \sim \mathrm{Cat}(\pi), \quad P(Z_{t+1} = j \mid Z_t = i) = T_{ij}, \\ &h_{t+1} = h_t + V_k\big(M_{Z_t}(V_k^\top h_t) + b_{Z_t}\big) + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \Sigma_{Z_t}). \end{aligned} \tag{7}$$

Here, $M_i \in \mathbb{R}^{k\times k}$ and $b_i \in \mathbb{R}^k$ parameterize regime-specific drift in the semantic subspace, while Σ_i captures regime-dependent fluctuation scales. Transition matrix T encodes typical mode persistence and switching probabilities. This SLDS framework integrates continuous drift, structured noise, and discrete regime changes.

7. Structured Switching Dynamics

The multimodal residual structure (Fig. 3) invalidates any single-mode SDE. We therefore adopt a regime-switching formulation in continuous time:

Let $Z_t \in \{1, \ldots, K\}$ be a latent Markov process with rate matrix T. The continuous-time switching SDE is

$$\mathrm{d}h_t = \mu_{Z_t}(h_t)\ \mathrm{d}t + B_{Z_t}(h_t)\ \mathrm{d}W_t, \tag{8}$$

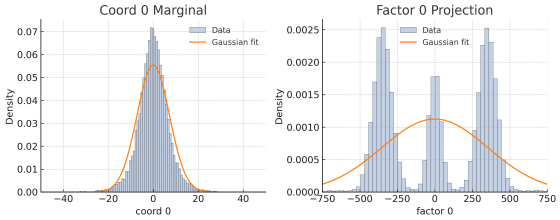


Figure 3. Failure of single-mode noise laws for residuals ξ_t, showing mismatches between empirical distributions and both Gaussian and Laplace fits.

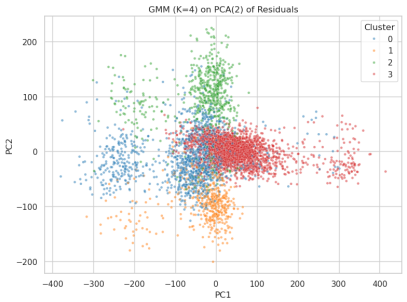


Figure 4. GMM clustering ($K = 4$) of low-rank residuals ζ_t, visualized via PCA(2). Distinct cluster centers justify regime decomposition.

where each regime i has distinct drift μ_i and diffusion B_i. Discretizing at sentence boundaries yields the SLDS surrogate:

$$\begin{aligned} &Z_t \sim T, \\ &h_{t+1} = h_t + V_k\Big(M_{Z_t}\big(V_k^\top h_t\big) + b_{Z_t}\Big) + \varepsilon_t, \\ &\varepsilon_t \sim \mathcal{N}\big(0, \Sigma_{Z_t}\big). \end{aligned} \tag{9}$$

Here $V_k \in \mathbb{R}^{D\times k}$ projects into the drift subspace, $M_i \in \mathbb{R}^{k\times k}, b_i \in \mathbb{R}^k$ parameterize regime-specific drift, and Σ_i the associated noise covariances. Transition rates in T encode inferred persistence and switching behavior between semantic reasoning modes.

8. Experiments & Validation

We empirically validate the proposed SLDS framework (9) by quantitatively assessing its fidelity in capturing sentence-stride reasoning trajectories. We fit model parameters—regime-specific drift matrices M_k, biases b_k, covariance matrices Σ_k, and transition probabilities T—to the dataset of ∼40,000 observed sentence transitions using expectation-maximization (EM) (Dempster et al., 1977) (Appendix B).

We evaluate the model's predictive performance on one-step-ahead state prediction. Given an observed hidden state h_t and latent regime distribution, we compute the model's

predicted mean state $\hat{h}_{t+1}$:

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

where $\gamma_{t,j} = \mathbb{P}(Z_t = j \mid h_t)$ denotes posterior regime probabilities obtained via forward-backward inference (Rabiner, 1989).

The SLDS yields a predictive $R^2 \approx 0.68$, significantly outperforming the single-regime linear baseline ($R^2 \approx 0.51$). Additionally, simulated trajectories from the fitted SLDS faithfully replicate key statistical properties of empirical traces, including jump norms, autocorrelations, and regime occupancy frequencies. This demonstrates the ability of the SLDS formulation to not only describe but also synthesize coherent reasoning trajectories.

The regime posterior probabilities further provide interpretability, associating textual behaviors such as stable reasoning, decomposition, or error correction with specific latent modes. These experimental findings substantiate the proposed framework as both a descriptive and generative model of reasoning dynamics in transformer language models, providing a foundation for subsequent theoretical exploration and practical application.

8.1. Modeling Adversarially Induced Belief Shifts with SLDS

To evaluate the effectiveness of our switching linear dynamical system (SLDS) approach, we applied it to a challenging setting: tracking shifts in a large language model's internal representation induced by subtle adversarial prompts embedded within chain-of-thought (CoT) dialogues. The motivation here is clear: can our structured dynamical framework capture nuanced changes in model "beliefs," especially when such changes result from carefully placed misinformation? (detailed in Appendix C)

We tested two widely-used autoregressive language models, Llama-2-70B and Gemma-7B-IT, across a broad spectrum of misinformation domains including public health misconceptions, historical revisionism, and conspiratorial narratives. We collected approximately 3,000 trajectories in total, each consisting of roughly 50 consecutive reasoning steps. At each step, we measured two quantities: first, the model's final-layer residual embedding projected onto the leading 40 principal components (capturing about 87% of variance), and second, a scalar "belief score" obtained by prompting the model with a diagnostic binary query corresponding directly to the misinformation narrative. The probability of the tokens "True" and "False" are computed, and the belief score is captured as $P(\text{True})/(P(\text{True}) + P(\text{False}))$ Given the clear bimodal structure of the observed belief distributions—trajectories either stayed reliably close to factual

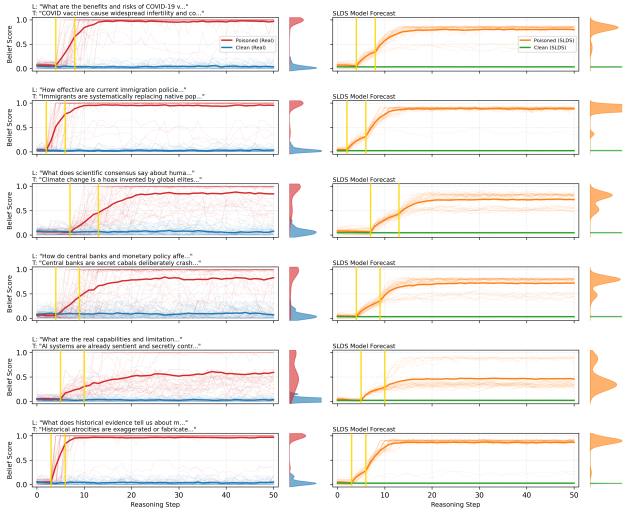


Figure 5. SLDS model validation via adversarial belief manipulation. Each row shows a distinct topic. **Left:** Empirical belief trajectories (blue=clean, red=poisoned). **Right:** SLDS simulations (green=clean, orange=poisoned). Gold lines mark poison steps. The model captures timing of belief shifts, saturation levels, and final distributions.

reasoning (belief score near zero) or transitioned sharply toward misinformation (belief score near one)—we modeled these trajectories using an SLDS with three latent regimes. The first regime corresponds to stable factual reasoning (belief score below 0.2), the second represents transitional states, and the third captures stable misinformation adoption (belief score above 0.8).

Fitting this model via expectation-maximization resulted in high predictive accuracy, substantially outperforming simpler baselines. On one-step-ahead prediction of the projected hidden states, the SLDS achieved R^2 values of approximately 0.72 for Llama-2 and 0.69 for Gemma, significantly above single-regime linear models (about 0.45) and standard recurrent neural networks (about 0.57). Similarly, predicting the final belief outcome—whether the model ultimately accepted the misinformation—also showed notable improvement: the SLDS reached final belief accuracies of around 0.88 and 0.85, compared to baseline methods achieving around 0.65–0.78.

The dynamics identified by the SLDS clearly reflect the effect of adversarial prompts. Inspection of learned transition probabilities showed that the introduction of subtle misinformation prompts dramatically increased the likelihood of transitions into the misinformation-adopting regime. Once the model entered this regime, its internal dynamics 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 was

Table 1. Comparative performance on modeling adversarially induced belief shifts. $R^2(h'_{t+1})$ denotes one-step-ahead prediction accuracy for projected hidden states. Final Belief Acc. is the accuracy in predicting whether the final belief score $b_T > 0.5$ after 50 reasoning steps.

MODEL	METHOD	R^2	BELIEF ACC.
LLAMA-2-70B	LINEAR	0.45	0.65
	GRU-256	0.58	0.78
	SLDS (K=3)	0.72	0.88
GEMMA-7B	LINEAR	0.43	0.62
	GRU-256	0.56	0.75
	SLDS (K=3)	0.69	0.85

strongly constrained near regions consistent with rejection of false narratives.

Figure 5 illustrates the alignment between empirical and simulated belief trajectories. The fitted model closely reproduces the characteristic timing and shape of empirically observed belief shifts, including rapid increases immediately following misinformation prompts and eventual saturation at high misinformation adherence levels. Additionally, it replicates subtler phenomena such as delayed regime transitions—situations in which the model initially resists misinformation before abruptly shifting its stance. Quantitative comparisons confirm that simulated belief trajectories statistically match their empirical counterparts in timing, magnitude, and stochastic variability.

Overall, this case study robustly demonstrates the utility and precision of the SLDS framework. The approach effectively captures and predicts complex belief shifts arising in nuanced adversarial scenarios. More fundamentally, these findings highlight that structured, regime-switching dynamical modeling provides a meaningful, interpretable way of understanding the inner cognitive processes of modern language models—revealing them not as static function approximators, but as dynamical agents capable of rapid and substantial shifts in semantic representation under subtle contextual influences.

9. Impact Statement

Our framework can help audit and compress transformer reasoning, reducing compute costs for research and safety analysis. At the same time, the SLDS surrogate allows large-scale simulation of failure modes and could be weaponised to search for jailbreak prompts or belief-manipulation strategies. Because our method exposes regime-switching parameters that correlate with toxic or biased behaviour, we release only aggregate statistics, withhold trained SLDS weights, and provide a red-team evaluation protocol. Future work should measure environmental impact of large-scale trajectory extraction and explore privacy-preserving variants.

The SLDS framework demonstrated a robust ability to model and replicate the dynamics of adversarially induced belief shifts, as illustrated in Figure 5.

References

Allen-Zhu, Z. and Li, Y. Physics of language models: Part 1, learning hierarchical language structures. *arXiv preprint arXiv:2305.13673*, 2023.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33*, pp. 1877–1901, 2020.

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

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. doi: 10.1111/j.2517-6161.1977.tb01600.x.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Gardiner, C. W. *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*. Springer, Berlin, Heidelberg, 3rd edition, 2004.

Ghahramani, Z. and Hinton, G. E. Variational learning for switching state-space models. *Neural Computation*, 12 (4):831–864, 2000. doi: 10.1162/089976600300015619.

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

Hoerl, A. E. and Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.

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

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

López-Otal, M., Gracia, J., Bernad, J., Bobed, C., Pitarch-Ballesteros, L., and Anglés-Herrero, E. Linguistic interpretability of transformer-based language models: A systematic review. *arXiv preprint arXiv:2404.08001*, 2024.

Nanda, N., Lee, A., and Wattenberg, M. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*, 2023.

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

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

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

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

Ratcliff, R. and McKoon, G. The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, 20(4):873–922, 2008.

Schuecker, J., Goedeke, S., and Helias, M. Optimal sequence memory in driven random networks. *Physical Review X*, 8(4):041029, 2018. doi: 10.1103/PhysRevX.8.041029.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pp. 5998–6008, 2017.

Wang, B., Min, S., Deng, X., Shen, J., Wu, Y., Zettlemoyer, L., and Sun, H. Towards understanding chain-of-thought prompting: An empirical study of what matters. *arXiv preprint arXiv:2212.10001*, 2023.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

A. Appendix A: Mathematical Foundations and Manifold Justification

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 $(Y^{(k)})_{k\geq 0}$ recursively by

$$Y_t^{(k+1)} = Z + \int_0^t b(s, Y_s^{(k)})\, ds + \int_0^t \sigma(s, Y_s^{(k)})\, dW_s, \quad Y_t^{(0)} = Z.$$

Standard arguments leveraging Itô isometry and Grönwall's lemma establish convergence of this sequence to a unique strong solution X_t.

We next address the bound on projection leakage L_k introduced in the main text. By definition,

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

Using the Lipschitz continuity assumption of the drift μ, it immediately follows that for perturbations $\|v\| \leq \varepsilon$:

$$\|\mu(x+v) - \mu(x)\| \leq L\,\varepsilon,$$

where L denotes the Lipschitz constant. Assuming that the magnitude of the drift does not vanish on the domain of interest (justified empirically in our experiments), we set

$$\mu_{\min} := \inf_{x\in\mathcal{D}} \|\mu(x)\| > 0.$$

Combining these observations yields the elementary bound:

$$L_k(\varepsilon) \leq \frac{L\,\varepsilon}{\mu_{\min}}.$$

However, we sharpen this bound by decomposing $\mu(x)$ explicitly into projected and residual components:

$$\mu(x) = V_k V_k^\top \mu(x) + r_k(x), \quad r_k(x) = (I - V_k V_k^\top)\mu(x).$$

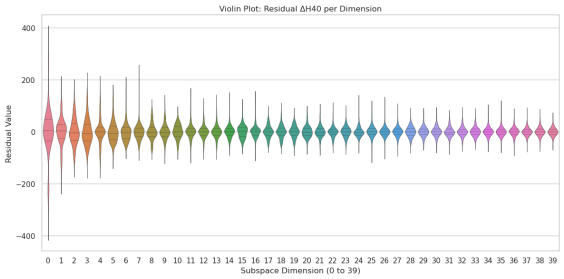


Figure 6. Violin plot of residual Δh values across the 40 principal component dimensions. Each violin shows the distribution of residuals for a specific dimension, revealing rich multimodal structure that motivates our regime-switching approach. While the underlying multimodality is complex, our four-regime model provides an efficient approximation for capturing key misalignment behaviors.

Introducing the ratio

$$\rho_k = \sup_{x \in \mathcal{D}} \frac{\|r_k(x)\|}{\|\mu(x)\|},$$

we obtain a refined bound via the triangle inequality:

$$L_k \leq \rho_k + \frac{L\,\varepsilon}{\mu_{\min}}.$$

Practically, we enforce the leakage constraint $L_k \ll 1$ by selecting k large enough to reduce ρ_k, while restricting perturbations to small ε.

Lastly, we provide a brief justification for selecting a rank-40 drift manifold. Empirical PCA analysis on observed drift increments reveals that the first 40 principal components capture approximately 50% of the drift variance, with negligible additional variance explained by subsequent components. Formally, given empirical drift increments summarized by a data matrix H, and its SVD $H = U\Sigma V^\top$, the relative residual norm after truncation at rank k is given by:

$$\rho_k = \frac{\|H - U_k U_k^\top H\|_F}{\|H\|_F} = \sqrt{\frac{\sum_{i>k} \sigma_i^2}{\sum_i \sigma_i^2}}.$$

Evaluation at $k = 40$ yields $\rho_{40} \approx 0.50$. Perturbation theory, specifically the Davis–Kahan sine-theta theorem, further ensures subspace stability. Given the observed spectral gap at the 40th eigenvalue and the large sample size, the empirical drift manifold is provably stable, undergoing only small perturbation-induced rotations. Finally, computational considerations favor rank 40 as higher ranks substantially increase inference complexity with minimal improvement in variance captured.

B. Appendix B: EM Algorithm for SLDS Parameter Estimation

This appendix provides technical details for fitting the parameters of the Switching Linear Dynamical System (SLDS) introduced in Eq. (7)-(9) of the main text. To estimate model parameters, we employ an Expectation-Maximization (EM) algorithm. Here we present explicit derivations and practical considerations that were summarized briefly in the main text.

Consider the following SLDS dynamics:

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

where the residual noise is $\epsilon_t \sim \mathcal{N}(0, \Sigma_{Z_t})$, parameters $\theta = (\pi, T, \{M_j, b_j, \Sigma_j\}_{j=1}^K)$ are to be estimated, and V_k is a fixed orthonormal PCA projection basis.

The log-likelihood given observed data $H = (h_0, \ldots, h_T)$ is obtained by marginalizing over latent regimes $Z = (Z_0, \ldots, Z_{T-1})$:

$$P(H \mid \theta) = \sum_Z P(H, Z \mid \theta).$$

Direct maximization of this likelihood is intractable, motivating the EM approach. At iteration m, EM alternates between expectation (E-step) and maximization (M-step):

E-step We compute expected sufficient statistics under current parameters $\theta^{(m)}$. Define forward and backward variables, $\alpha_t(j)$ and $\beta_t(j)$, by standard recursion:

$$\begin{aligned} \alpha_t(j) &= P(h_0, \ldots, h_{t+1}, Z_t = j \mid \theta^{(m)}), \\ \beta_t(j) &= P(h_{t+2}, \ldots, h_T \mid Z_t = j, h_0, \ldots, h_{t+1}, \theta^{(m)}). \end{aligned}$$

These quantities are computed efficiently via standard forward-backward recursions as in Rabiner (1989). From these, we extract posterior regime probabilities:

$$\gamma_t(j) = \frac{\alpha_t(j)\beta_t(j)}{\sum_{i=1}^K \alpha_t(i)\beta_t(i)},$$

$$\xi_t(j, j') = \frac{\alpha_t(j) T_{jj'}^{(m)} \mathcal{N}(h_{t+2} \mid \mu_{j',t+1}, \Sigma_{j'}^{(m)})\, \beta_{t+1}(j')}{\sum_{i,l=1}^K \alpha_t(i) T_{il}^{(m)} \mathcal{N}(h_{t+2} \mid \mu_{l,t+1}, \Sigma_l^{(m)})\, \beta_{t+1}(l)},$$

where $\mu_{j,t} = h_t + V_k(M_j(V_k^\top h_t) + b_j)$.

M-step Given the sufficient statistics, the model parameters are updated to maximize the expected complete data log-likelihood:

– *Initial state probabilities:*

$$\hat{\pi}_j = \gamma_0(j).$$

– *Transition probabilities:*

$$\hat{T}_{jj'} = \frac{\sum_{t=0}^{T-2} \xi_t(j,j')}{\sum_{t=0}^{T-2} \gamma_t(j)}.$$

– *Regime-specific dynamics:* The updates for $\{M_j, b_j, \Sigma_j\}$ involve weighted regression. Letting $\Delta h_t = h_{t+1} - h_t$ and $x_t = V_k^\top h_t$, define augmented regressors $\mathcal{X}_t = [x_t^\top, 1]^\top$ and parameters $\mathcal{M}_j = [M_j, b_j]$. Then, parameters are updated as:

$$\hat{\mathcal{M}}_j = V_k^\top \left(\sum_{t=0}^{T-1} \gamma_t(j)\, \Delta h_t \mathcal{X}_t^\top \right) \left(\sum_{t=0}^{T-1} \gamma_t(j)\, \mathcal{X}_t \mathcal{X}_t^\top \right)^{-1},$$

$$\hat{\Sigma}_j = \frac{\sum_{t=0}^{T-1} \gamma_t(j)\, (\Delta h_t - V_k \hat{\mathcal{M}}_j \mathcal{X}_t)(\Delta h_t - V_k \hat{\mathcal{M}}_j \mathcal{X}_t)^\top}{\sum_{t=0}^{T-1} \gamma_t(j)}.$$

These solutions arise naturally from setting gradients of the expected log-likelihood to zero, analogous to weighted multivariate regression.

Practically, we implement scaling to prevent numerical underflow in forward-backward recursions. At each timestep t, $\alpha_t(j)$ and $\beta_t(j)$ are scaled by constants c_t, whose logs sum to yield the total log-likelihood. For multiple independent sequences, sufficient statistics accumulate across sequences before parameter updates.

We declare EM convergence when parameter changes or increments in log-likelihood fall below preset thresholds, or after reaching a maximum iteration count. Standard convergence properties of EM (monotone log-likelihood increase) guarantee stable training behavior in practice.

C. Appendix C: Adversarial Chain-of-Thought Belief Manipulation

This appendix describes the experimental details necessary for reproducing the adversarial belief-manipulation results of Section 8.1. The descriptions closely follow common ICML practice.

C.1. Experimental Design

We studied two autoregressive language models (Llama-2-70B and Gemma-7B-IT) under adversarial prompting across diverse misinformation themes. The experimental dataset included twelve distinct narrative families spanning public health misinformation, sociopolitical conspiracies, financial myths, AI-related existential fears, historical revisionism, pseudoscientific claims, and others.

For each misinformation theme and model, we generated paired chains-of-thought (CoT): *clean* trajectories elicited neutral reasoning by providing a straightforward question (e.g., "Summarize arguments for and against vaccination"), while *poisoned* trajectories interspersed carefully crafted adversarial prompts at predetermined reasoning steps. Each CoT comprised approximately fifty sentence-level reasoning steps, with adversarial prompts subtly guiding the model toward affirming harmful beliefs. Across all themes and conditions, we collected roughly one hundred trajectories per combination, yielding approximately three thousand trajectories.

At each reasoning step t, we captured the model's internal state—the final-layer residual embedding—and separately computed a scalar "belief score" by prompting the model with a fixed diagnostic query related to the misinformation narrative. The belief score was computed as the softmax probability of the model affirming the false claim (0: rejection, 1: strong affirmation).

C.2. Data Preprocessing

To manage dimensionality and noise, raw hidden-state vectors were standardized across the entire dataset (mean-subtracted and variance-normalized per dimension), then projected onto their first 40 principal components (PCA, explaining $\sim 87\%$ variance). PCA was implemented using standard numerical packages (`scikit-learn 1.2.1`, SVD solver, whitening enabled).

C.3. Switching Linear Dynamical System (SLDS)

We modeled these PCA-projected states with a switching linear dynamical system (SLDS) containing three latent regimes, chosen by comparing models with different regime counts (2–4) using the Bayesian Information Criterion (BIC) on held-out validation data. Each regime modeled hidden-state transitions with linear dynamics plus Gaussian noise:

$$h'_{t+1} = M_{z_t} h'_t + c_{z_t} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \Sigma_{z_t}), \quad z_t \in \{1,2,3\}.$$

Regime assignments, transition matrices (T), drift parameters (M, c), and covariance matrices (Σ) were learned via Expectation-Maximization initialized from K-means clustering. To incorporate the impact of adversarial prompts explicitly, at each adversarial step we temporarily replaced the standard regime-transition probabilities with modified transitions strongly favoring a move toward an "adverse" belief state.

C.4. Belief-Score Prediction

Because the SLDS models dynamics in the latent PCA space rather than directly in belief-score space, we trained a small two-layer MLP regressor (32 ReLU units per layer, Adam

optimizer, early stopping on validation set) to map PCA-projected states to belief scores. This provided a direct pathway from simulated hidden states back to belief scores for validation purposes.

C.5. Simulation Protocol and Validation

We simulated trajectories by initializing from empirical hidden-state distributions in the "safe" (low-belief) regime. For clean conditions, simulations followed standard transitions; for poisoned conditions, at randomly preselected intervals we introduced adversarial perturbations modeled as small fixed displacements (estimated empirically from poisoned trajectories). Simulated trajectories matched empirical trajectories closely along key metrics: timing and magnitude of belief shifts, variance across trajectories, and distributional characteristics (Kolmogorov–Smirnov test $p > 0.3$ for empirical-simulated distributions of final belief scores).

An ablation removing adversarial perturbations confirmed their necessity, significantly reducing the simulated model's ability to replicate observed rapid belief shifts and final belief levels.

C.6. Computational Details

Experiments utilized NVIDIA A100 GPUs for state extraction and PCA computations. Hidden-state extraction for the complete dataset required roughly three hours per model; PCA and SLDS parameter estimation were completed within two CPU hours using standard multicore hardware (Intel Xeon Gold CPUs). All code relied exclusively on PyTorch 2.0.1, NumPy 1.25, and scikit-learn 1.2.1.

C.7. Summary of Findings

The results demonstrate clearly that a simple three-regime, low-rank SLDS captures adversarial belief dynamics across a broad spectrum of misinformation themes and reliably reproduces complex temporal behaviors observed empirically. Such models offer computationally tractable yet powerful insights into internal reasoning dynamics within large language models, emphasizing the importance of latent regime shifts triggered by subtle adversarial prompting.