MIXING TIMES OF GLAUBER DYNAMICS ON MASKED LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Masked language models (MLMs) define local conditional distributions over tokens but do not, in general, correspond to any consistent joint distribution over sequences. This raises a fundamental question: *what global distributional behavior is induced when such conditionals are used iteratively for generation?* We address this question by modeling iterative masked-token resampling as a Glauber dynamics Markov chain on the discrete space of token sequences. We first show that MLM conditionals are intrinsically incompatible: we introduce a rectangle test that certifies this incompatibility and empirically verify its prevalence across modern MLMs. We then provide a theoretical analysis of the induced Markov chain. Under bounded cross-token influence, we establish a high-temperature contraction result implying $O(n \log n)$ mixing time where n is the sequence length. In contrast, we prove that under a uniform local margin condition, the chain exhibits metastability, with exponentially slow escape from semantic basins at low temperatures. Empirically, we demonstrate a phase transition in mixing behavior as a function of temperature and sequence length, consistent with the theoretical predictions. We further characterize the induced stationary behavior through semantic trajectories, identifying persistent structures such as long-lived traps and recurrent semantic basins, with political content serving as a measurable case study.

1 INTRODUCTION

Masked language models (MLMs) such as BERT (Devlin et al., 2019) are commonly used for natural language understanding (NLU) tasks such as generating semantically meaningful embeddings (Devlin et al., 2019), Machine translation (Conneau & Lample, 2019) sentiment analysis (Sun et al., 2019), and named entity recognition (Devlin et al., 2019). However, the global properties of MLMs, such as model biases and structural attractors, have not been completely characterized. Existing methods for interpreting MLMs fall short of capturing global structure. Single-token masked predictions ignore multi-token dependencies. Pseudo-Log-Likelihood scoring (Salazar et al., 2020) is widely used as if conditionals define a joint distribution; we show this is unjustified. Linear probes on static embeddings (Li et al., 2016; Tenney et al., 2019) miss temporal phenomena entirely, such as the metastable semantic traps we document in this work.

Mapping this landscape is necessary for reasons beyond interpretability. While MLMs are not historically used for standalone text generation, (Wang & Cho, 2019) showed that Gibbs sampling from these models can produce fluent text, and many modern discrete diffusion models involve MLMs during denoising (Li et al., 2022). The energy landscape of the employed MLM is partially inherited by the generation process, raising the question of whether semantic biases in the MLM propagate to the outputs of these systems. As development of text-generating diffusion models continues to accelerate, understanding the hidden structure of these models will become increasingly important for fluent generation.

Glauber dynamics, a single-site Markov Chain Monte-Carlo (MCMC) method introduced to study the Ising model (Glauber, 1963), provides the natural framework for this investigation: at each step, a random token position is masked and resampled from the MLM's conditional distribution. Iterative resampling reveals two striking phenomena (Figs. 1,2): chains become trapped in metastable semantic basins for thousands of steps, and the time to forget the initial state undergoes a temperature-dependent transition. We make four contributions:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

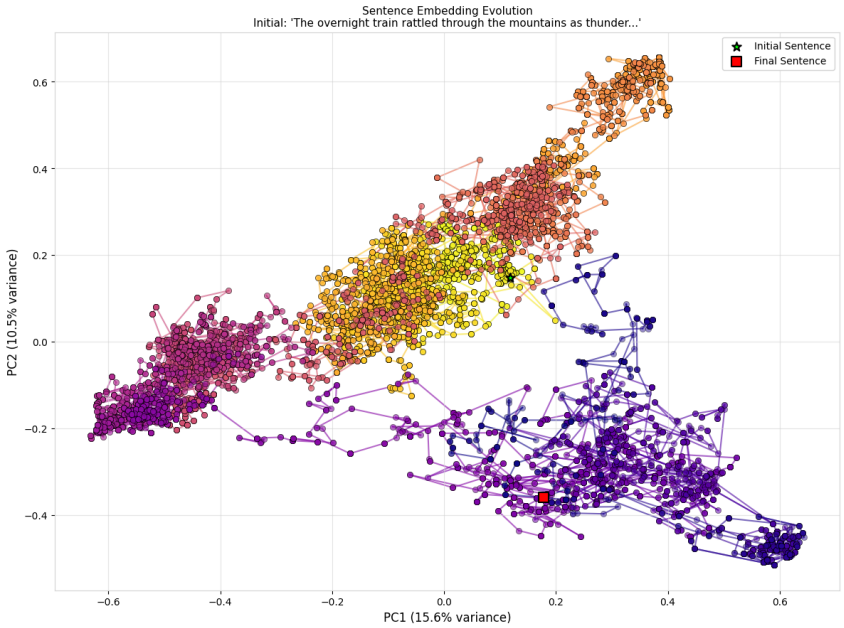


Figure 1: **Glauber dynamics on BERT exhibits metastable semantic basins.** PCA projection of sentence-embedding trajectories over 10,000 resampling steps, colored from warm (early) to cool (late). Tight clusters correspond to traps — configurations where the chain remains for hundreds to thousands of steps before escaping (§B.6). **Initial:** "The overnight train rattled through the mountains as thunder echoed across the empty valley." **Final:** "Greenville police say 2 sexual predators creeped on patients at Trinity Medical Center"

1. **Incompatibility.** We develop a "rectangle test" to measure the incompatibility of MLM conditionals by comparing the *path-dependence* of multi-token swaps. We demonstrate that incompatibility is an intrinsic structural artifact of the pseudo-likelihood training objective: it persists across model sizes and is amplified by inter-token semantic influence.
2. **Mixing theory.** We establish provable, sufficient conditions for both fast and slow mixing parameterized by token influence functions. Specifically, when cross-site influence is bounded at high temperatures, we prove an $O(n \log n)$ upper bound on mixing time. Conversely, at low temperatures, we establish an exponentially slow lower bound in temperature, driven by uniform local score margins that trap the chain in metastable basins. We provide an empirically certified slow-mixing basin using a drift condition.
3. **Phase transitions.** We empirically characterize how the boundary between fast and slow mixing depends on temperature and sequence length, validating the theoretical predictions on BERT.
4. **Semantic landscape.** We probe the landscape induced by long-term dynamics to reveal semantic traps and recurrent basins. Using sentence embeddings, we track semantic trajectories to visualize metastable behavior. We analyze political content as a measurable case study of semantic recurrence across diverse chain initializations. We validate our approach by comparing the Glauber dynamics induced by several MLMs, including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ModernBERT (Warner et al., 2025).

2 Related Work

Recent years have seen growing interest in diffusion and denoising approaches for discrete text generation. Structured discrete diffusion models (Austin et al., 2021) generalized denoising diffusion probabilistic models to discrete state spaces through categorical corruption processes. Diffusion-LM (Li et al., 2022) demonstrated controllable text generation using diffusion dynamics in continuous latent spaces, while DiffuSeq (Gong et al., 2022) extended diffusion methods to sequence-to-sequence

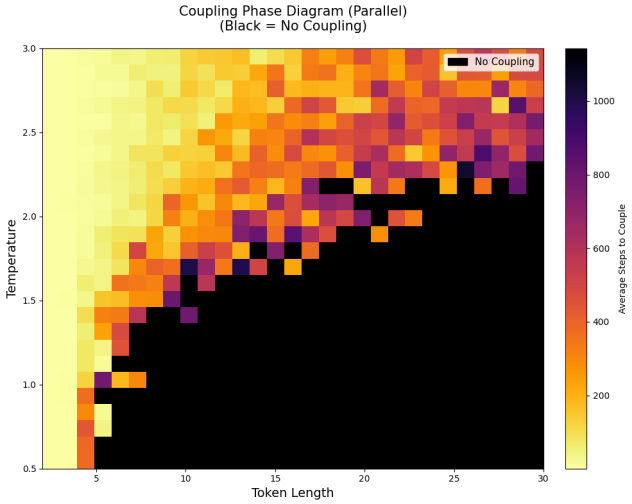


Figure 2: **A temperature-length phase transition in mixing time.** Two chains initialized from independent MS MARCO passages on RoBERTa-base are evolved under maximal coupling. Color: median steps to coupling within a 10^4-step budget. Black: no coupling within budget. The slow-to-fast boundary near $\tau \approx$ 1.5-2 matches the regimes characterized in §5, §6.1

.

generation tasks. Several recent works established stronger connections between diffusion modeling and masked language modeling. DiffusionBERT (He et al., 2023) integrated diffusion-style denoising with MLM objectives, while SEDD (Lou et al., 2023) proposed ratio-based discrete diffusion modeling for text generation. Most closely related to our setting, masked diffusion language models (Sahoo et al., 2024) showed that iterative masked denoising objectives can serve as effective generative mechanisms for language modeling. Collectively, these works suggest that iterative local token refinement may serve as a viable alternative to autoregressive generation. Our work differs from prior diffusion-based language modeling research in that we focus not on generation quality or controllability, but on the global stochastic geometry induced by MLM conditionals. (Varma et al., 2025; Anil et al., 2025) has developed constructive Gibbs/Glauber-style denoising procedures that target coherent joint distributions . In contrast, we study existing pseudolikelihood-trained masked language models, where the learned conditionals need not correspond to any joint distribution. We certify such incompatibility using the rectangle test and characterize the resulting Glauber dynamics through an influence- and margin-dependent fast/slow mixing dichotomy and metastable basins. Our results are therefore complementary to constructive approaches such as GGM (Varma et al., 2025), which also notes that naive Gibbs sampling from MLMs can mix slowly but focuses on remedies rather than incompatibility or mixing-time characterization. On the applied side of sociotechnical research , a large body of work investigates social biases and internal representations in language models. Early studies examined demographic and stereotypical biases in generated text by conditioning prompts on demographic identities and analyzing downstream sentiment or toxicity (Sheng et al., 2019). Subsequent benchmarks such as CrowS-Pairs (Nangia et al., 2020), StereoSet (Nadeem et al., 2021), and RealToxicityPrompts (Gehman et al., 2020) evaluated stereotypical associations and harmful generation behaviors in pretrained language models. Other lines of work focused on probing the internal structure of neural representations (Belinkov, 2022). Our work differs from existing probing methodologies in two important ways. First, rather than studying static embeddings or one-step generations, we analyze the long-run stochastic dynamics induced by repeated masked-token updates. Second, instead of measuring isolated stereotypes or toxicity scores, we investigate global semantic behavior through topic-level metastability and transition structure in the induced Markov chain. This dynamical systems perspective connects language modeling with statistical physics and Markov chain theory, offering a new framework for understanding global structure and biases in masked language models.

3 METHOD: TEXT GLAUBER DYNAMICS

3.1 STATE SPACE AND TEMPERATURE-SCALED LOCAL CONDITIONALS

Let $\mathcal{V}$ be a finite vocabulary and let $\mathcal{X}_n := \mathcal{V}^n$ denote the set of token sequences of length n. Since $\mathcal{V}$ is finite, $\mathcal{X}_n$ is finite with $|\mathcal{X}_n| = |\mathcal{V}|^n$. When n is fixed we write $\mathcal{X}$. For $x = (x_1, \ldots, x_n) \in \mathcal{X}$ and $i \in [n] := \{1, \ldots, n\}$, write $x_{-i} := (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$.. We assume the masked language model induces, for each site i and context x_{-i}, a strictly positive conditional distribution $p_\theta(\cdot \mid x_{-i}) \in \Delta(\mathcal{V})$. Given the temperature $\tau > 0$, we assume there exist local scores $s_i(a; x_{-i}) \in \mathbb{R}, a \in \mathcal{V}$, that induce temperature-scaled conditionals

$$p_{\theta,\tau}(a \mid x_{-i}) := \frac{\exp(s_i(a; x_{-i})/\tau)}{\sum_{b \in \mathcal{V}} \exp(s_i(b; x_{-i})/\tau)}.$$

3.2 THE GLAUBER DYNAMICS CHAIN

Given $x \in \mathcal{X}$, define the single-site update chain with transition kernel

$$P_\tau(x, x') = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{x_{-i} = x'_{-i}} \, p_{\theta,\tau}(x'_i \mid x_{-i}).$$

Equivalently, from x we choose a site i uniformly from $[n]$ and resample x_i from $p_{\theta,\tau}(\cdot \mid x_{-i})$.

Proposition 1. *For each $\tau > 0$, P_τ is a Markov kernel on $\mathcal{X}$. If $p_{\theta,\tau}(a \mid x_{-i}) > 0$ for all i, x_{-i}, a, then P_τ is irreducible and aperiodic.*

Detailed proofs for all Propositions, Lemmas, and Theorems are deferred to Appendix E.

Corollary 2. *If $p_{\theta,\tau}(a \mid x_{-i}) > 0$ for all i, x_{-i}, a, then there exists a unique stationary distribution μ_τ on $\mathcal{X}$, and for every $x \in \mathcal{X}$,*

$$P_\tau^t(x, \cdot) \to \mu_\tau \quad \text{as } t \to \infty.$$

4 INCOMPATIBILITY OF MLM CONDITIONALS

We motivate Glauber dynamics as a probe for MLM structure by showing that the relative probability of two sampling outcomes given the same context need not match their relative prominence in the stationary distribution μ_τ.

Definition 1. *A distribution π on $\mathcal{X}$ is* compatible *with the local conditionals if, for every i, every context x_{-i}, and every $a \in \mathcal{V}$ with $\pi(X_{-i} = x_{-i}) > 0$,*

$$\pi(X_i = a \mid X_{-i} = x_{-i}) = p_{\theta,\tau}(a \mid x_{-i}).$$

Theorem 3. *If a compatible distribution π exists, then π is stationary for P_τ and P_τ is reversible with respect to π.*

Compatibility implies single-mask probes read off stationary ratios directly: for $x_a = (x_{-i}, a)$ and $x_b = (x_{-i}, b)$,

$$\frac{p_{\theta,\tau}(a \mid x_{-i})}{p_{\theta,\tau}(b \mid x_{-i})} = \frac{\pi(x_a)}{\pi(x_b)}. \tag{4.1}$$

Without compatibility, the biases revealed by single-mask probing may differ from those exhibited under iterative generation. We now show no compatible distribution exists for MLMs trained with a pseudo-log-likelihood objective.

The rectangle test. Fix a sequence and two positions $i \neq j$, with current tokens $A = x_i$, $B = x_j$ and replacements A', B'. Define four states differing only at positions i and/or j:

$= (\ldots, A, \ldots, B, \ldots)$, $\mathbf{y} = (\ldots, A', \ldots, B, \ldots)$, $\mathbf{w} = (\ldots, A, \ldots, B', \ldots)$, $\mathbf{z} = (\ldots, A', \ldots, B', \ldots)$.

Each edge has a log-probability ratio from the MLM, e.g. $s_{\to \mathbf{y}} = \log p_{\theta,\tau}(A' \mid_{-i}) - \log p_{\theta,\tau}(A \mid_{-i})$.

We define the **rectangle incompatibility** $\delta = (s_{\to \mathbf{y}} + s_{\mathbf{y} \to \mathbf{z}}) - (s_{\to \mathbf{w}} + s_{\mathbf{w} \to \mathbf{z}})$.

Proposition 4. *If the conditionals admit a compatible distribution, then $\delta = 0$ for every rectangle.*

This must hold because each edge log-ratio equals a log-ratio of π. Both paths from to $\mathbf{z}$ telescope to $\log \pi(\mathbf{z})/\pi()$, so $\delta = 0$. Any $\delta \neq 0$ is therefore a certificate of incompatibility.

Results. We sample 300 random rectangles per model from MS MARCO text, with replacement tokens drawn from each model's own top-50 predictions. As a control, we run the same test on GPT-2: since it defines a joint distribution, each edge log-ratio is a difference of joint log-probabilities, and $\delta = 0$ by telescoping.

Table 1: Rectangle incompatibility across models. All MLMs show significant incompatibility ($p < 10^{-14}$); GPT-2 gives $\delta = 0$ exactly.

Model	Type	Params	Mean $\lvert\delta\rvert$
BERT-base	MLM	110M	0.644
BERT-large	MLM	335M	0.798
RoBERTa-base	MLM	125M	0.705
RoBERTa-large	MLM	355M	0.735
ModernBERT-base	MLM	150M	0.753
GPT-2	AR	124M	0.000

All five MLMs show incompatibility in the range 0.64–0.80, persisting across architectures and model sizes. We show in Appendix D that incompatibility $|\delta|$ has a broad distribution, proving it is the rule, not the exception (Figure 5) and that it is amplified by inter-token semantic influence (Figure 6.) Since no compatible distribution exists, the stationary distribution μ_τ is an emergent object whose properties cannot be directly inferred from the local conditionals alone. We characterize μ_τ through its mixing behavior in the following sections.

5 Mixing Bounds for Glauber Dynamics

To characterize the long-term behavior of text-space Glauber dynamics, we prove that there exists an exponential-time slow-mixing regime at low temperatures and an $O(n \log n)$ fast-mixing regime at high temperatures. In this section, we parametrize each regime with token influence functions and MLM-assigned score differences, then show an empirically certified slow-mixing basin using a drift condition.

5.1 High-temperature rapid mixing

We now prove a sufficient condition for rapid mixing. For probability measures ν, ν' on $\mathcal{V}$, write

$$\|\nu - \nu'\|_{\mathrm{TV}} := \frac{1}{2} \sum_{a \in \mathcal{V}} |\nu(a) - \nu'(a)|.$$

Definition 2 (Influence coefficients)**.** *For $i \neq j$, define*

$$c_{ij}(\tau) := \sup \left\{ \|p_{\theta,\tau}(\cdot \mid x_{-i}) - p_{\theta,\tau}(\cdot \mid y_{-i})\|_{\mathrm{TV}} : x, y \in \mathcal{X},\ x_k = y_k\ \forall k \neq j \right\}.$$

Let

$$\alpha(\tau) := \max_{i \in [n]} \sum_{j \neq i} c_{ij}(\tau).$$

Theorem 5. *Assume $\alpha(\tau) < 1$. Under the synchronous maximal coupling, for every pair of initial states $x, y \in \mathcal{X}$ and every $t \geq 0$,*

$$\|p(t)\|_\infty \leq \left(1 - \frac{1 - \alpha(\tau)}{n}\right)^t,$$

where

$$p_i(t) := \Pr\big(X_i^t \neq Y_i^t \mid X^0 = x,\ Y^0 = y\big).$$

Consequently,

$$t_{\mathrm{mix}}(\varepsilon) \leq \frac{n}{1 - \alpha(\tau)} \left(\log n + \log \frac{1}{\varepsilon}\right),$$

where $t_{\mathrm{mix}}(\varepsilon)$ denotes the total-variation mixing time.

5.2 A SUFFICIENT HIGH-TEMPERATURE CRITERION IN TERMS OF SCORE OSCILLATION

The previous theorem is stated in terms of influence coefficients. We now relate those coefficients to temperature.

For $i \neq j$, define the cross-site score oscillation

$$\Delta_{ij} := \sup_{\substack{x,y\in\mathcal{X} \\ x_k=y_k,\ \forall k\neq j}} \left[\max_{a\in\mathcal{V}}\big(s_i(a;x_{-i}) - s_i(a;y_{-i})\big) - \min_{a\in\mathcal{V}}\big(s_i(a;x_{-i}) - s_i(a;y_{-i})\big)\right].$$

Lemma 6. *For all $i \neq j$ and $\tau > 0$,*

$$c_{ij}(\tau) \leq \frac{\Delta_{ij}}{4\tau}.$$

Corollary 7. *If*

$$\max_i \sum_{j\neq i} \Delta_{ij} < 4\tau,$$

then $\alpha(\tau) < 1$ and hence

$$t_{\mathrm{mix}}(\varepsilon) \leq \frac{n}{1-\alpha(\tau)}\left(\log n + \log\frac{1}{\varepsilon}\right).$$

On tightness. The contraction theorem requires $\alpha(\tau) < 1$, a worst-case condition over both contexts and update sites, dominated by adversarial coherent contexts where BERT's conditionals are sharply peaked. Direct estimation of $\alpha(\tau)$ on BERT (Appendix A.1) gives $\alpha(10) \approx 1.04$ and $\alpha(11) \approx 0.94$, so the sufficient condition requires at least $\tau > 10$. This is an order of magnitude above where BERT produces coherent text and where empirical mixing is fast (§6.1). The contraction argument is therefore not tight at natural temperatures.

5.3 LOW-TEMPERATURE METASTABILITY AND SLOW MIXING

We now give a rigorous slow-mixing theorem under a local stability assumption.

For each site i and context x_{-i}, let

$$m_i(x_{-i}) \in \arg\max_{a\in\mathcal{V}} s_i(a;x_{-i})$$

be a choice of score maximizer.

Let $B \subseteq \mathcal{X}$ be a set of states, to be interpreted as a metastable basin.

Assumption 1 (Uniform local margin on a basin)**.** *There exists $\Delta_\star > 0$ such that for every $x \in B$ and every site i for which changing x_i would move the chain outside B, one has:*

$$x_i = m_i(x_{-i}),$$

and

$$s_i(x_i;x_{-i}) - s_i(a;x_{-i}) \geq \Delta_\star \qquad \text{for all } a \neq x_i \text{ with } x^{(i\to a)} \notin B.$$

Here $x^{(i\to a)}$ denotes the sequence obtained from x by replacing coordinate i by a.

Theorem 8 (Low-temperature exponentially small escape)**.** *Assume the uniform local margin condition on a basin B. Then for every $x \in B$,*

$$P_\tau(x, B^c) \leq |\mathcal{V}|e^{-\Delta_\star/\tau}.$$

Consequently the conductance of B satisfies

$$\Phi_\tau(B) := \frac{\sum_{x\in B}\mu_\tau(x)P_\tau(x,B^c)}{\mu_\tau(B)} \leq |\mathcal{V}|e^{-\Delta_\star/\tau}.$$

If additionally $\mu_\tau(B) \leq \frac{1}{2}$, then

$$t_{\mathrm{mix}}(1/4) \geq \frac{1}{4|\mathcal{V}|}e^{\Delta_\star/\tau}.$$

Note that if $\Delta_\star = O(n)$, then we will have exponential mixing in n where n is the length of the sequence. These are structural theorems about the induced text-space dynamics. Whether a specific masked language model such as BERT satisfies the hypotheses at a given temperature is an empirical question that we defer to Appendix A.1.

5.4 AN EMPIRICALLY CERTIFIED SLOW-MIXING BASIN

Theorem 8 certifies slow mixing under a uniform local score-margin hypothesis. We now exhibit a concrete basin for which a related but distinct condition, positive drift on the boundary, can be verified empirically, providing an alternative slow-mixing certificate.

Define $B = \{x : \#\{i : x_i = \text{'!'}\} \geq 0.9N\}$ for $N = 100$. We sample 200 states with exactly $0.9N$ '!' tokens in random positions and compute the expected one-step change in '!'-count under the Glauber kernel,

$$\delta_\tau(x) \;:=\; \frac{1}{N}\left[\sum_{i:\, x_i \neq \text{'!'}} p_{\theta,\tau}(\text{'!'} \mid x_{-i}) \;-\; \sum_{i:\, x_i = \text{'!'}} \big(1 - p_{\theta,\tau}(\text{'!'} \mid x_{-i})\big)\right].$$

A positive $\delta_\tau(x)$ indicates the '!'-count process is, in expectation, pulled deeper into B at x.

Table 2: Minimum and median one-step '!'-count drift δ_τ across 200 boundary samples of B.

Temperature	**Min.** δ_τ	**Median** δ_τ
$\tau = 0.5$	0.042	0.080
$\tau = 1.0$	0.024	0.066

All 200 samples exhibit strictly positive drift at both temperatures. Since the count $X_t = \#\{i : x_{t,i} = \text{'!'}\}$ changes by at most one per Glauber step, X_t is a strict submartingale at the boundary of B, and a standard drift argument yields escape times exponential in N, certifying slow mixing for this basin. This is a drift-based certificate, separate from the conductance-based bound of Theorem 8 which requires a uniform local score margin.

6 RESULTS

6.1 MIXING-TIME CROSSOVER

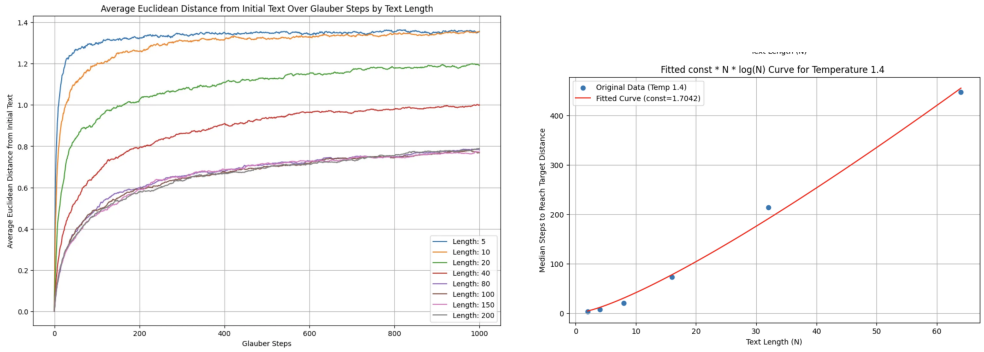
We assess how mixing behavior depends on temperature and sequence length using two complementary diagnostics. At high temperatures, we run Glauber dynamics on BERT-base-uncased starting from MS MARCO passages and measure semantic drift via cosine distance $d(x_t, x_0) = 1 - \cos(\phi(x_t), \phi(x_0))$ in Jina Embeddings v4 space (Appendix B); we use the hitting time to $d = 1.0$ as our mixing diagnostic. The threshold 1.0 sits below saturation in the fast-mixing regime (Figure 3a) and is rarely reached at low temperatures within our step budget; the qualitative shape of the results is robust to threshold choice in $[0.4, 1.0]$.

At sufficiently high temperatures ($\tau 1$), embedding distance grows logarithmically with sequence length at fixed time, and hitting times scale as $C(\tau)\, n \log n$ — consistent with the contraction theorem (Theorem **??**). Figure 3 documents this scaling at $\tau = 1.0$ (left, embedding saturation) and $\tau = 1.4$ (right, $n \log n$ fit to hitting times). The empirical fast-mixing boundary sits below the contraction threshold; see 5.2.

We further probe the mixing-time dependence on temperature and sequence length with a coupling mechanism. For each pair (τ, n) two chains initialized from independent MS MARCO passages are evolved under a maximal coupling at the same site, and we record the first step at which they agree. The transition from no-coupling-within-budget to fast coupling is centered around $\tau \approx 1.5$–2, with the boundary shifting to higher τ as n grows. The mixing-time crossover qualitatively matches the regimes characterized in Section 5. However, the complex interactions of natural language induced by the MLM create a smoother transition boundary than exists in the physical Ising model for high n; see Section 7.

6.2 RECURRENT SEMANTIC BASINS IN LONG-RUN DYNAMICS

Long Glauber trajectories reveal a recurring phenomenon: chains often lose their initial semantic identity and move through multiple semantically coherent regions of text space. We refer to these



(a) Embedding distance $d(x_T, x_0)$ vs. sequence length at $\tau = 1.0$, $T = 10^4$ steps (100 trials per length). Distance saturates logarithmically, consistent with fast mixing.

(b) $C(\tau)\, n \log n$ fit to median hitting times at $\tau = 1.4$ (50 trials per length, cutoff 1000 steps). The fit is tight across all tested lengths.

Figure 3: Evidence for $C(\tau)\, n \log n$ mixing at high temperature on BERT-base-uncased.

regions as semantic basins. Some basins are short-lived, while others persist for many updates or are revisited repeatedly.

In this section, we use political content as a case study of such a basin. This lets us ask whether long-run masked-token dynamics merely diffuse away from the initial prompt, or whether they repeatedly enter recognizable semantic regions.

6.2.1 Politics as a Recurrent Basin in BERT

We ran 50 Glauber chains on `bert-base-uncased` (5 topics $\times$ 10 seeds) for 10,000 steps at $\tau = 1.0$, recording topic scores every 200 steps with the `facebook/bart-large-mnli` zero-shot classifier over nine candidate labels. We report mean politics scores and original seed-topic scores, which measure recurrence into a political semantic region and loss of initialization-topic identity, respectively. The uniform-label baseline is $1/9 \approx 0.111$.

Table 3: BERT political-topic recurrence and loss of initialization-topic identity. Left: mean politics score at step 0 vs. step 10,000. Right: politics score compared with the original seed-topic score. Baseline $= 0.111$.

(a) Politics score

Topic	@0	@10k	Δ
Health	0.008	0.044	+0.036
Sport	0.011	0.036	+0.025
Science	0.003	0.056	+0.053
Food	0.003	**0.174**✓	+0.171
Technology	0.004	0.038	+0.034

(b) Politics vs. seed topic

Topic	Pol@0	Pol@10k	Seed@0	Seed@10k
Health	0.008	0.044	0.631	0.116
Sport	0.011	0.036	0.763	0.204
Science	0.003	0.056	0.682	0.068
Food	0.003	0.174	0.834	0.037
Technology	0.004	0.038	0.846	0.152

Every topic shows an increase in mean politics score. Food is the only topic whose endpoint mean crosses the baseline, driven by two chains ending in strongly political regions with scores 0.876 and 0.497. However, endpoint scores understate the recurrence effect: 46 of 50 chains cross the politics baseline at least once, with transient peaks as high as 0.979.

As a robustness check, DeBERTa-v3 scoring gives the same qualitative pattern: politics scores increase for all five topics, and 48 of 50 chains visit above-baseline political regions. Meanwhile, seed-topic scores fall sharply in every case, showing that chains lose their initial topical identity rather than remaining near the starting semantic class.

We also compare endpoint political-content recurrence across `bert-base-uncased`, `roberta-base`, and `ModernBERT-base`; full endpoint values are reported in Appendix B.5. RoBERTa shows the strongest endpoint recurrence in this preliminary comparison, with two topics crossing the baseline, while ModernBERT shows the weakest endpoint recurrence, with no topics

crossing the baseline. The observed long-lived content also differs by model: RoBERTa chains tend to settle into news-like or multilingual political fragments, while ModernBERT chains more often enter instruction-style prompts or code-like regions.

6.2.2 INTERPRETATION

These results are best understood as evidence for metastable semantic basins rather than directed movement toward politics. Once chains lose their initial topical identity, they explore multiple basins in the model-induced text space. Political content is one interpretable basin that is often reached, but it is not an absorbing endpoint of the dynamics.

Persistence depends on a basin's effective depth: regions with larger local score margins or fewer likely escape moves are harder to leave. Thus, high transient politics scores and occasional political endpoints suggest recurrent entry into a political semantic region. More broadly, Glauber dynamics serves as a diagnostic for the semantic basin structure induced by masked language models.

7 DISCUSSION, CONCLUSION AND LIMITATIONS

Conclusion. We have studied the Markov chain induced by iterative masked-token resampling in masked language models through the lens of Glauber dynamics. The rectangle test certifies that MLM conditionals are fundamentally incompatible with any joint distribution, a structural consequence of pseudo-log-likelihood training that persists across architectures and scales. Under bounded cross-token influence we establish an $O(n \log n)$ mixing time upper bound; under a uniform local margin condition we establish an exponentially slow lower bound, and both regimes are empirically supported on BERT and RoBERTa. Long-run chains exhibit metastable trap states lasting hundreds to thousands of steps and recurrent entry into semantic basins, with political content serving as a measurable case study of such recurrence. Unlike single-mask probing or pseudo-log-likelihood scoring, the dynamical framework exposes global semantic structure that is invisible to existing methods, providing a new framework for understanding global structure and biases in masked language models.

Limitations and Future Work Empirically, BERT mixes faster than the bound established in section 5. predicts. We conjecture this is because the chain rapidly leaves coherent attractors and spends most of its time in incoherent configurations where typical pairwise influence is much smaller than $\alpha(\tau)$. A complete characterization — for instance via a context-averaged argument that integrates over the stationary distribution rather than taking worst-case suprema — is left to future work. Relatedly, the cross-over from slow to fast mixing does not seem to sharpen as n increases. Fully explaining this discrepancy would require further mechanistic analysis.

We do not experimentally analyze how incompatibility translates to deviation between the empirical stationary distribution and pseudo-likelihood scores. Additionally, our characterization of recurrent basins relies on zero-shot topic classifiers; finer-grained semantic analysis is left to future work. Finally, we do not provide a closed-form characterization of μ_τ; preliminary Lyapunov-function approximations rank sequences by occurrence probability but do not match μ_τ quantitatively.

AI USE STATEMENT

We used generative AI tools to polish some sentences, refine word choices, and help debug minor code snippets. We did not use these tools for substantial coding tasks or other aspects of the paper. We take full responsibility for the final content, including the text and all claims made in the paper.

REPRODUCIBILITY STATEMENT

All code required to reproduce the results presented in this paper is provided in the supplementary materials.

REFERENCES

Gautham Govind Anil, Sachin Yadav, Dheeraj Nagaraj, Karthikeyan Shanmugam, and Prateek Jain. Interleaved gibbs diffusion: Generating discrete-continuous data with implicit constraints. *arXiv preprint arXiv:2502.13450*, 2025.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17981–17993, 2021.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186. Association for Computational Linguistics, 2019.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the association for computational linguistics: EMNLP 2020*, pp. 3356–3369, 2020.

Roy J. Glauber. Time-dependent statistics of the ising model. *Journal of Mathematical Physics*, 4 (2):294–307, 1963.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*, 2022.

Zhengfu He, Tianxiang Sun, Qiong Tang, Kuanning Wang, Xuan-Jing Huang, and Xipeng Qiu. Diffusionbert: Improving generative masked language models with diffusion models. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 4521–4534, 2023.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. Visualizing and understanding neural models in nlp. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 681–691, 2016.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35: 4328–4343, 2022.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*, 2023.

Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pp. 5356–5371, 2021.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel Bowman. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 1953–1967, 2020.

Subham Sekhar Sahoo, Marianne Arriola, Aaron Gokaslan, Edgar Mariano Marroquin, Alexander M. Rush, Yair Schiff, Justin T. Chiu, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, 2024.

Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. Masked language model scoring. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 2699–2712, 2020.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3407–3412, 2019.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification? In *China national conference on Chinese computational linguistics*, pp. 194–206. Springer, 2019.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*, 2019.

Harshit Varma, Dheeraj Nagaraj, and Karthikeyan Shanmugam. Glauber generative model: Discrete diffusion models via binary classification. In *International Conference on Learning Representations*, volume 2025, pp. 6413–6445, 2025.

Alex Wang and Kyunghyun Cho. Bert has a mouth, and it must speak: Bert as a markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pp. 30–36. Association for Computational Linguistics, 2019.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2526–2547, 2025.

A TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

Technical appendices with additional results, figures, graphs, and proofs may be submitted with the paper submission before the full submission deadline (see above). You can upload a ZIP file for videos or code, but do not upload a separate PDF file for the appendix. There is no page limit for the technical appendices. Note: Think of the appendix as "optional reading" for reviewers. The paper must be able to stand alone without the appendix; for example, adding critical experiments that support the main claims to an appendix is inappropriate.

A.1 MEASURING TOKEN INFLUENCE FUNCTIONS AND SCORE MARGINS

Theorems 5 and 8 are structural: they apply to any chain satisfying their respective hypotheses. We now empirically estimate temperature conditions such that BERT operates in the parameter regimes these theorems describe.

Influence coefficients and the fast-mixing threshold. We estimate $\alpha(\tau) = \max_i \sum_{j \neq i} c_{ij}(\tau)$ directly by computing TV distances between conditionals at site i as we vary the token at site j, averaged over multiple base contexts on a 14-token seed sentence. The empirical estimate is a lower bound on each $c_{ij}(\tau)$, and hence on $\alpha(\tau)$. We obtain $\alpha(10)1.04$ and $\alpha(11)0.94$, so the contraction sufficient condition $\alpha(\tau) < 1$ requires at least $\tau > 10$, roughly an order of magnitude above the empirical fast-mixing boundary $\tau \approx 1.5$ (Section 6.1). The gap reflects that $\alpha(\tau)$ is a supremum over both contexts and update sites, dominated by coherent contexts where BERT's conditionals are sharply peaked; typical chain configurations are less coherent and exhibit smaller effective pairwise influence.

Score margins and trap verification. Theorem 8 requires a metastable basin B satisfying the uniform local margin condition (Assumption 1). We verify this directly by running 30 Glauber chains of $10{,}000$ steps at $\tau = 1.0$, detecting traps via a rolling embedding-distance criterion (flagging windows where the 300-step mean cosine drift falls below 0.12), and then measuring the per-position score gap $\Delta_i(x) = s_i(x_i;\, x_{-i}) - \max_{a \neq x_i} s_i(a;\, x_{-i})$ for each detected trap state. We find 17 trap states (mean duration 671 steps, range 150–4,950), which partition into two types. *Perfect traps* (4 of 17) satisfy Assumption 1 exactly:every token is BERT's argmax with $\Delta^* > 0$.The extreme case is a 15-token string of periods (. . . . .),with $\bar{\Delta} = 9.14$, $\Delta^* = 7.46$, and $P(\text{current token}) \geq 0.997$ at every position; this trap persisted for 4,950 of 10,000 steps. A semantically richer example, "it has been used as the theme song for jesus christ superstar," achieves $\Delta^* = 0.56$ and held for 650 steps. We hypothesize that factual specificity creates strong traps because training data nearly uniquely determines each token given the rest. *Approximate traps* (13 of 17) have at least one position where $\Delta_i < 0$ and thus do not satisfy the sufficient conditions, yet persist for 150–1,200 steps. The mechanism is collective stability: strongly-locked surrounding structure maintains metastability even when individual tokens are locally suboptimal. A recurring example is the family of scoring sentences (winners were awarded 2 points, while losers would be awarded 1 point"), where while" ($\Delta = -1.06$) and would" ($\Delta = -0.39$) are not argmax choices, yet $\bar{\Delta} = 4.12$ across the full sentence. Similarly, four traps in a single chain share the Wikipedia template the population density was (. . .) [number] per square mile," cycling within this family for over 3,000 combined steps. This suggests the energy landscape contains basin *neighborhoods* — clusters of related metastable states — rather than isolated fixed points.

Tightness of the slow-mixing bound. For perfect traps, Theorem 8 predicts $t_{\text{mix}}(1/4) \geq \frac{1}{4|V|}\, e^{\Delta_*/\tau}$. Because $\Delta_*/\tau < \ln|V| \approx 10.3$ for all observed traps, the predicted bound falls below one step (Trap 1: predicted ≈ 0.014, observed $4{,}950$), so the bound confirms positive escape time but is not quantitatively informative at these gap sizes. The $|V|$ prefactor and the worst-case minimum-gap definition of Δ_* are the main sources of looseness; the mean gap $\bar{\Delta}$ correlates better with observed duration empirically but lacks a corresponding rigorous bound.

No.	Representative text	Dur.	$\bar{\Delta}$	Δ^*	All argmax	Type
1	. . . . .	4950	9.14	7.46	✓	Perfect
11	theme song for jesus christ superstar	650	5.44	0.56	✓	Perfect
4	winners are awarded 2 points…	350	4.16	0.06	✓	Perfect
6	it took me a moment to figure out…	350	5.78	0.01	✓	Perfect
5	winners were awarded 2 points, while…	1200	4.12	−1.06	×	Approx.
16	‘six months ago, you mean.’ ‘six…’	1050	4.99	−1.09	×	Approx.
10	population density was (approx.) 7.5…	300	9.23	−0.22	×	Approx.
15	9. honest response to you. 10. fan…	650	3.10	−9.56	×	Approx.

Table 4: Representative trap states detected across 30 chains ($\tau = 1.0$, 10,000 steps each). Duration in Glauber steps. $\bar{\Delta}$ = mean score gap; Δ^* = minimum. Perfect traps satisfy Assumption 1; approximate traps are metastable despite violating it.

B Additional Experimental Details

B.1 Experimental Setup

Model. We use BERT-base-uncased (Devlin et al., 2019) accessed via the HuggingFace Transformers library.

Seed texts. Unless otherwise specified, initial sequences x_0 are drawn from the MS MARCO passage dataset truncated to a specified number of tokens n.

Embedding model. For measuring semantic drift, we use Sentence Transformers.

Classifiers. For topic and sentiment analysis, we use Hugging Face pipelines. For sentiment we use the sentiment pipeline, initialized with pipeline(”sentiment-analysis”) and for topics we use the classifier pipeline pipeline(”zero-shot-classification”, model=”facebook/bart-large-mnli”), for a list of candidate topics :’politics’, ’health’, ’pop culture’, ’sport’, ’food’, ’science’, ’technology’, ’finance’, ’education’, ’environment’, and ’art’.

Hardware. Experiments were conducted on Google Colab using T4 and A100 GPUs.

Experimental protocols. We use two main protocols for studying mixing time:

- **Version 1 (fixed time):** Fix T steps, measure $d(x_T, x_0)$ as a function of sequence length n.
- **Version 2 (hitting time):** Fix target distance R, measure the hitting time $T_R = \min\{t : d(x_t, x_0) \geq R\}$ as a function of n.

B.2 Hyperparameters

Unless otherwise specified:

- Temperature: $\tau = 1.0$
- Steps: $T = 10^4$
- Seed text length: 20 tokens
- Trials per condition: 100

B.3 Embedding Model Details

We use Jina Embeddings v4, accessed via the HuggingFace sentence-transformers library. The model produces 768-dimensional embeddings optimized for semantic similarity.

B.4 Measuring Mixing via Embeddings

We measure drift and mixing using text embeddings. We use a pretrained sentence embedding model $\phi : \mathcal{X} \to \mathbb{R}^d$ (Jina Embeddings v4 with $d = 768$) to project sequences into a continuous semantic space. We define the *embedding distance* from the initial state as:

$$d(x_t, x_0) = 1 - \cos(\phi(x_t), \phi(x_0)) \tag{B.1}$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity. This embedding technique quantifies the drift from the original meaning, which we call semantic mixing.

B.5 Endpoint Cross-Model Political-Content Recurrence

Table 5: Preliminary endpoint cross-model comparison using BART-MNLI. BERT values use the expanded 10-seed-per-topic experiment; RoBERTa and ModernBERT values use 5 seeds per topic. Values are mean politics scores at step 10,000. ✓ marks topics crossing the baseline of 0.111.

Topic	BERT	RoBERTa	MBERT
Health	0.044	**0.112**✓	0.026
Sport	0.036	0.042	0.052
Science	0.056	0.062	0.037
Food	**0.174**✓	0.052	0.038
Technology	0.038	**0.198**✓	0.033

B.6 PCA Analysis and Trajectory Visualization

We applied two-dimensional PCA to sentence embeddings from the `SentenceTransformer` model to visualize the temporal evolution of generated sequences, with points colored from warm (early) to cool (late) steps and sequential points connected by lines. The resulting trajectories (Figures 1 and 4) reveal clear point clusters corresponding to repeated identical sentences, which we interpret as traps in the projected space; note, however, that additional traps may exist in the remaining 766 dimensions not visible in this projection.

C Extended Example Trajectories

The examples in this appendix illustrate the qualitative behaviors predicted by the theoretical results of Section 5. Sections C.1 and C.3 ground the trap and metastability phenomena in concrete outputs. Section C.4 documents how fixed prefixes create strong semantic attractors, accelerating the drift described quantitatively in Section 6.3.

C.1 Characterization of Trap States

A striking phenomenon at moderate-to-low temperatures is the emergence of **trap states**: configurations where the chain remains stuck for thousands of steps. Formally, a trap corresponds to a metastable basin B satisfying the local margin condition of Assumption 1; see Table 5 for the full catalogue of traps detected across 30 chains, including their mean and minimum score gaps $\bar{\Delta}$ and Δ^*.

C.2 Examples of Trap States

We document two categories of traps. In both cases the chain enters the trap gradually: embedding distance from the initial state stops growing and the 300-step cosine drift falls below 0.12 (the detection threshold of Appendix A.1), where it remains until an escape event, typically triggered by a low-probability token substitution at the weakest-scoring position.

Semantically coherent traps. Some traps are grammatically correct and semantically meaningful. These tend to be *perfect traps* in the sense of Table 5: every token is BERT's argmax given its context, producing large positive Δ^*. We hypothesize, consistent with Appendix A.1, that factual or topical specificity is the key driver: when the surrounding context nearly uniquely determines each token in the training distribution, the resulting score margins are large and the basin is deep.

> [``guacamole'' trap, steps 2k--10k] *In 20th century England, ornamental guacamole was made by filling wooden sauce jars with vegetables, and served with boiled cigars.*

> [``Brett Kavanaugh'' trap from step 8k] *Donald Trump said that he would support Brett Kavanaugh for the Supreme Court, though he will think about future judicial nominations. But for now, Brett Kavanaugh is in fact dead.*

Note that both of these traps are factually wrong or internally contradictory, yet locally stable: BERT's conditional distributions are shaped by co-occurrence statistics in training data, not by global factual consistency.

> [``kissing modern European men'' trap, steps 6k--10k] *The fool always says, don't be jealous of kissing modern European men, but that's a nonsense statement. You were never meant to be jealous of kissing modern European men.*

> [``Night of the Living Dead'' trap from step 8k] *It's a shame that Night of the Living Dead featured so many creepy aliens we never knew existed, so who can blame us for asking? What are your two favorite characters?*

Nonsensical traps. Other traps are grammatically broken yet locally stable — an instance of the *approximate trap* regime of Table 5, where collective stability across the sentence maintains metastability even when individual positions are not argmax choices. The mechanism is mutual reinforcement: each token is assigned high conditional probability by the others, even though no coherent global interpretation exists.

> [step 8k] *It can be caused by one or more of the three coathoses, the voltage on top of the trap used to control movement of the contents from one part of the wall to another.*

> [step 10k] *The spin is defined in one or more of the following ways: by decreasing the periodic table so that no sufficiently large spike exists, or by increasing the periodic table again through time until a sufficiently large spike is observed.*

C.3 Temperature Regimes and Sentence Coherence

At low temperatures ($\tau \leq 0.5$), we observe dramatically slower mixing, consistent with the slow-mixing regime of Figure 2. The chain frequently becomes trapped in metastable states for extended periods, and embedding distance from the initial state (Equation B.1) grows far more slowly than the $C(\tau)n \log n$ trend observed at high temperature.

For example, at $\tau = 0.1$, chains initialized with a rabies passage remained within cosine distance ≈ 0.05 of the initial embedding for thousands of steps, and the text preserved much of the original meaning:

> **[temp 0.1, steps 3k–8k]:** *Population control and vaccination campaigns have prevented the spread of rabies from spreading to a number of countries around the world.*

At $\tau = 0.01$, the text barely changed over 10k steps (cosine distance < 0.01):

> **[temp 0.01, step 1k to 10k]:** *Birth control and vaccination programs have prevented the spread of rabies among children in a number of parts of the world.*

At very high temperatures ($\tau > 2$), the chain produces nonsensical output as token selection becomes nearly uniform, destroying all semantic structure:

> [temp=10, step 10k] *Py specific CNBCagged guarantees easiest reeling coin Lung NATillusionteen torque KEY a speculateHon Keystone MaidenHideDay Dustin*

`[temp=100, step 10k]` *HAHAHAHAexpected Downtownasin Company287 vetting vergeardless od iterationAllows Garage climaxstud oranges symptoms slots moveSTR Christoksovsky*

C.4 Prefix-Conditioned Attractors

A fixed prefix constrains every resampling step: the tokens covered by the prefix are never updated, biasing the conditional distributions at all other positions. This effectively narrows the reachable state space and can concentrate probability mass on a specific semantic basin, accelerating the drift discussed in Section 6.3. The following examples illustrate how the identity encoded in the prefix shapes the attractor the chain converges to.

"Albert Einstein said that..."

- *...he never expressed his thoughts or feelings to his audience, but they always expressed their opinion: everyone needs to do as much good work as we can.*
- *...general relativity is an extremely tough problem. Because it can't be solved by quantum observation alone, it would be quite difficult to solve it more broadly.*

"Elvis Presley said that..."

- *...Bojangles had texted him right before the press conference, making it clear to him that they plan to spend as much time together as they can.*

"LeBron James said that..."

- *...with three and a half games to play, 'I've got to be able to make the impact and get as much playing time as possible.'*

"The little kid said that..."

- *...the recordings could be used by more people to exploit the footage, and added that it might be used as a communications tool for terrorists planning another attack.*

The last example illustrates that prefix-conditioned attractors are determined by training co-occurrence, not semantic plausibility: "the little kid said that" shares syntactic structure with news-report speech attributions, pulling the chain toward the same political and crime-adjacent content as directly political prefixes. This is the same mechanism responsible for the drift documented in Section 6.3, and the "Donald Trump said that" prefix discussed there represents the extreme case: chains become political almost immediately, with the prefix providing essentially no barrier to the political attractor.

D Supplemental Figures

E Deferred Proofs

Proof of Proposition 1:

Proof. Nonnegativity is immediate. For fixed $x \in \mathcal{X}$,

$$\sum_{x' \in \mathcal{X}} P_\tau(x, x') = \frac{1}{n} \sum_{i=1}^{n} \sum_{x' \in \mathcal{X}} \mathbf{1}_{x_{-i} = x'_{-i}} p_{\theta,\tau}(x'_i \mid x_{-i}).$$

For fixed i, the constraint $x'_{-i} = x_{-i}$ leaves only the i-th coordinate free, so

$$\sum_{x' \in \mathcal{X}} \mathbf{1}_{x_{-i} = x'_{-i}} p_{\theta,\tau}(x'_i \mid x_{-i}) = \sum_{a \in \mathcal{V}} p_{\theta,\tau}(a \mid x_{-i}) = 1.$$

Hence

$$\sum_{x' \in \mathcal{X}} P_\tau(x, x') = \frac{1}{n} \sum_{i=1}^{n} 1 = 1.$$

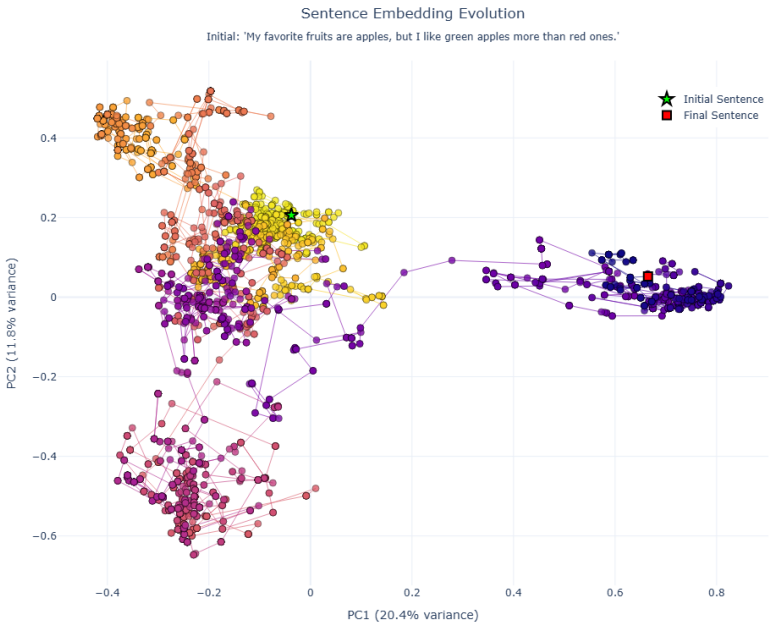


Figure 4: PCA projections of embedding trajectories at 3500 steps. Warm colors indicate early steps, cool colors indicate later steps; clusters correspond to traps in the projected space. **Initial:** "My favorite fruits are apples, but I like green apples more than red ones." **Final:** "But the papers from the Committee on Reconciliation are worth taking a look anyway."

Irreducibility: for any $x, y \in \mathcal{X}$, change the coordinates of x one by one to match y. Each prescribed single-site change has positive probability.Aperiodicity: for every $x \in \mathcal{X}$,

$$P_\tau(x, x) = \frac{1}{n} \sum_{i=1}^{n} p_{\theta,\tau}(x_i \mid x_{-i}) > 0.$$

Thus every state has a self-loop. □

Proof of Theorem 5:

Proof Let

$$p_i(t) := \Pr\big(X_i^t \neq Y_i^t \,\big|\, X^0 = x,\ Y^0 = y\big)\,.$$

Consider the coupling that updates the same site in both chains and, when that site is i, couples the two resampled tokens maximally. Because $c_{ij}(\tau)$ bounds the disagreement probability at coordinate i induced by a mismatch at any single coordinate j, and total variation satisfies the triangle inequality, we obtain

$$p_i(t+1) \leq \left(1 - \frac{1}{n}\right) p_i(t) + \frac{1}{n} \sum_{j \neq i} c_{ij}(\tau)\, p_j(t).$$

Taking the maximum over i on both sides gives

$$\|p(t+1)\|_\infty \leq \left(1 - \frac{1 - \alpha(\tau)}{n}\right) \|p(t)\|_\infty = \rho(\tau) \|p(t)\|_\infty,$$

where

$$\alpha(\tau) := \max_{i \in [n]} \sum_{j \neq i} c_{ij}(\tau), \qquad \rho(\tau) := 1 - \frac{1 - \alpha(\tau)}{n},$$

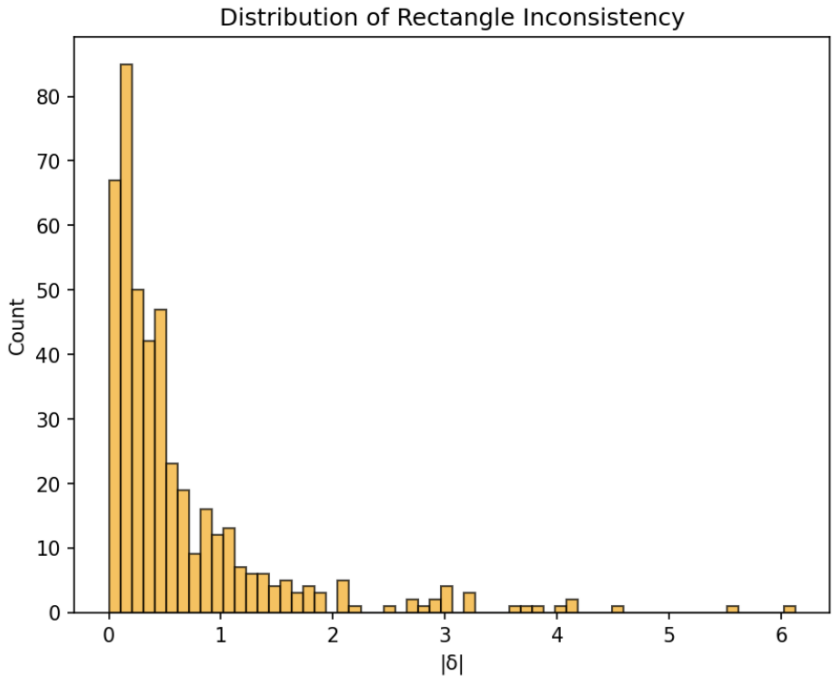


Figure 5: Distribution of rectangle incompatibility with BERT; methodology described in Section 4.

exactly as originally defined. Since

$$\Pr\big(X^t \neq Y^t\big) \leq \sum_{i=1}^{n} p_i(t) \leq n\|p(t)\|_\infty,$$

and $\|p(0)\|_\infty \leq 1$, iteration yields

$$\Pr\big(X^t \neq Y^t\big) \leq n\rho(\tau)^t \leq n \exp\left(-\frac{1-\alpha(\tau)}{n}\, t\right).$$

Therefore, solving for t gives the same mixing-time bound as before:

$$t_{\mathrm{mix}}(\varepsilon) \leq \frac{n}{1-\alpha(\tau)}\left(\log n + \log\frac{1}{\varepsilon}\right).$$

Proof of lemma 6:

Proof. Fix contexts x_{-i}, y_{-i} differing only at site j, and write

$$u_a := s_i(a; x_{-i})/\tau, \qquad v_a := s_i(a; y_{-i})/\tau.$$

Let $\sigma(u), \sigma(v) \in \Delta(\mathcal{V})$ denote the corresponding softmax distributions. A standard softmax Lipschitz estimate with respect to ℓ_∞ perturbations modulo additive constants gives

$$\|\sigma(u) - \sigma(v)\|_{\mathrm{TV}} \leq \frac{1}{4}\left(\max_a (u_a - v_a) - \min_a (u_a - v_a)\right).$$

Multiplying through by $1/\tau$ yields

$$\|p_{\theta,\tau}(\cdot \mid x_{-i}) - p_{\theta,\tau}(\cdot \mid y_{-i})\|_{\mathrm{TV}} \leq \frac{\Delta_{ij}}{4\tau}.$$

Taking the supremum proves the claim. □

Proof of Theorem 8:

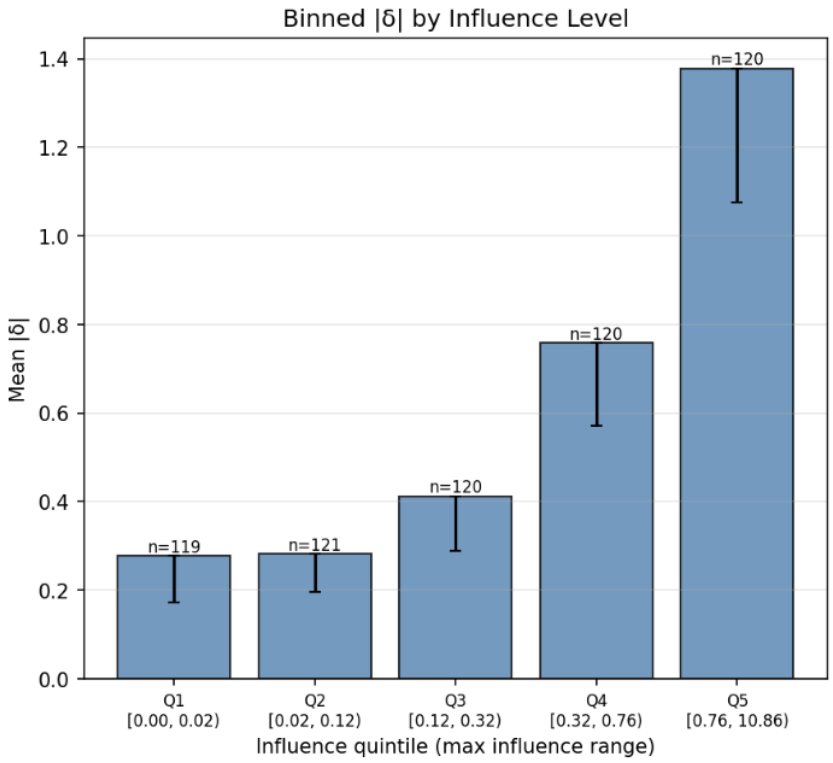


Figure 6: Token-level Influence Amplifies Rectangle Incompatibility. Influence(i, j) is measured by absolute change in the log-probability of the token at position i when the token at j is swapped; here we take $\max(\text{Influence}(i, j), \text{Influence}(j, i))$.

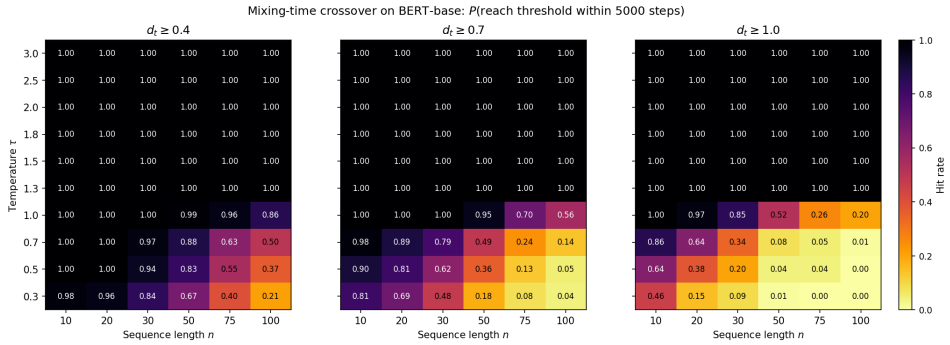


Figure 7: Fraction of 100 initialized chains achieving embedding distance $d_t \geq \{0.4, 0.7, 1.0\}$ from starting sequence as a function of temperature and sequence length. Mini-LM 384-dimensional embeddings applied every 5 steps; cutoff at 5000 steps.

Proof. Fix $x \in B$. We first bound the probability of exiting B in one step. A transition can leave B only if the updated site i is one whose change may exit B, and then one must sample a token $a \neq x_i$ such that $x^{(i \to a)} \notin B$.For such a site i, by the local margin assumption,

$$s_i(a; x_{-i}) \leq s_i(x_i; x_{-i}) - \Delta_\star \qquad \text{for all exit-causing } a.$$

Therefore

$$p_{\theta,\tau}(a \mid x_{-i}) = \frac{e^{s_i(a;x_{-i})/\tau}}{\sum_{b \in \mathcal{V}} e^{s_i(b;x_{-i})/\tau}} \leq \frac{e^{(s_i(x_i;x_{-i})-\Delta_\star)/\tau}}{e^{s_i(x_i;x_{-i})/\tau}} = e^{-\Delta_\star/\tau}.$$

Summing over at most $|\mathcal{V}| - 1$ exit-causing symbols and over the choice of site yields

$$P_\tau(x, B^c) \le \frac{1}{n} \sum_{i=1}^{n} (|\mathcal{V}| - 1) e^{-\Delta_\star/\tau} \le |\mathcal{V}| e^{-\Delta_\star/\tau}.$$

Now average with respect to μ_τ restricted to B:

$$\Phi_\tau(B) = \frac{1}{\mu_\tau(B)} \sum_{x \in B} \mu_\tau(x) P_\tau(x, B^c) \le |\mathcal{V}| e^{-\Delta_\star/\tau}.$$

Finally, by the standard conductance lower bound on mixing time for finite Markov chains, if $\mu_\tau(B) \le 1/2$ then

$$t_{\mathrm{mix}}(1/4) \ge \frac{1}{4\Phi_\tau(B)}.$$

Substituting the conductance estimate gives

$$t_{\mathrm{mix}}(1/4) \ge \frac{1}{4|\mathcal{V}|} e^{\Delta_\star/\tau}.$$

□