Mixing Times of Glauber Dynamics on Masked Language Models

Anonymous Authors[1]

Abstract

Masked language models (MLMs) divide texts into tokens, predicting masked tokens based on context of the other tokens in its vicinity. We study the MLM BERT, focusing on the steady states and drift using Glauber Dynamics. Glauber Dynamics is a Markov Chain Monte Carlo algorithm, which determines the next state based on the context of the current state. We use the Sentence Transformer embedding method to quantify states after a range of time steps. We look at different temperatures and token lengths and characterize mixing behavior, observing a phase transition and trap states. We characterize how framing a text with a constant prefix or suffix steers the distribution based on the sentiment and topic classifiers from Sentence Transformers.

1. Introduction

Masked language models (MLMs) such as BERT are typically used for natural language processing tasks by predicting missing words from surrounding context. This has been leveraged for both sentiment analysis and creating text. We address questions of how text length and specific sequences affect text generation using the method of Glauber Dynamics, iteratively sampling a masked token and using the values of neighboring tokens to determine the next state. The probability of a token changing is based on the temperature.

In our investigation of this process and its steady state distribution, we observe the following:

1. **Phase transition in mixing time.** We see a qualitative difference in how embedding distance increases for two different temperature regimes. At high temperature ($\tau \gtrsim 1$), we find evidence for coupon collector $C(\tau)n \log n$ mixing where n is the token length. At low temperature ($\tau \lesssim 1$), mixing time increases faster as a function of n.

2. **Evidence for Slow Mixing at Low Temperatures** Starting from homogeneous sequences of a specific token, we track the count of that token over time. At low temperatures, this text remains stable, showing slow mixing.

3. **Evidence for minority contamination.** At temperature 1.0, we see rapid mixing occur as soon as a single token changes in a string of repeated tokens. We refer to the disproportionate influence of a small percentage of the tokens as "minority token contamination.

4. **Characterization of trap states.** We examine fixed points or trap states in which the evolution gets stuck for thousands of steps. These trap states vary in length and are sometimes nonsensical.

5. **Analysis of sentence trajectories in embedding space.** We analyze the evolution of sentences by representing them with 768-dimensional sentence embedding, and projecting them onto the two highest-variance principal components. This projection provides a clear visualization of sentence trajectories in embedding space and reveals characteristic patterns and traps that become more interpretable.

6. **Framing effects and topical drift.** We show that fixed prefixes and suffixes significantly influence the stationary distribution, where the rest of the sentence tends to follow the topic signaled by the prefix or suffix.

2. Related Work

Glauber Dynamics and mixing times. Glauber Dynamics is a Markov Chain Monte Carlo (MCMC) method. Markov chains deal with situations in which the next event depends only on the current state. Unlike other MCMC algorithms, Glauber Dynamics chooses tokens randomly to update. Then the token is updated with a probability based on the neighboring tokens and the temperature (Hayes, 2019).

Masked language models. BERT (Bidirectional Encoder Representations from Transformers uses conditioning from

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

both the left and right of a masked token whereas other models use unidirectional language models for training. (Devlin et al., 2019). Previous studies have looked at a logarithmic distribution for each sentence token (Salazar, 2020). We take a similar approach with the number of tokens.

Probing LLM biases. Previous studies have looked at biases through framing with prompts mentioning different demographics and tracking sentiment scores (Sheng, 2019). We follow a similar procedure here looking at topic scores as well.

3. Method: Text Glauber Dynamics

3.1. Setup and Notation

3.2. The Glauber Dynamics Chain

Algorithm 1 Text Glauber Dynamics

Input: Initial sequence $x_0 \in \mathcal{X}$, temperature τ, number of steps T
for $t = 0$ **to** $T - 1$ **do**
  Sample position $i \sim \text{Uniform}\{1, \ldots, n\}$
  Sample new token $v \sim P_\theta^\tau(\cdot \mid x_t^{-i})$
  Set $x_{t+1} = (x_t^1, \ldots, x_t^{i-1}, v, x_t^{i+1}, \ldots, x_t^n)$
end for
Output: Sequence of states $x_0, x_1, \ldots, x_T$

The Glauber Dynamics Markov chain (Algorithm 1) proceeds as follows: at each step, we randomly select a position i, then resample the token at position i according to the model's conditional distribution given all other tokens in both directions. This procedure is also known as Gibbs sampling.

3.3. 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:

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

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

3.4. Framing with Prefixes and Suffixes

We also study a variant where a fixed prefix p and/or suffix s frame the evolving text x:

$$\text{full sequence} = [p; x; s] \tag{2}$$

Only the tokens in x are resampled; p and s remain fixed. This allows us to study how framing influences the stationary distribution.

4. 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.

5. Results

5.1. Phase Transition in Mixing Time

Our central result is that there is different behavior depending on the temperature:

- **High temperature ($\tau \gtrsim 1$):** Mixing time scales approximately as $C(\tau) n \log n$ with sequence length.
- **Low temperature ($\tau \lesssim 0.5$):** Mixing becomes slow and the amount of steps to reach a certain threshold grows much faster.

This is consistent with the phase transition in the Ising Model Glauber Dynamics is based on.

5.1.1. High Temperature Regime

At temperature $\tau = 1.0$, we observe that embedding distance from the initial state grows logarithmically with sequence length at fixed time, and hitting times grow roughly linearly (consistent with $C(\tau)n \log n$ total mixing time).

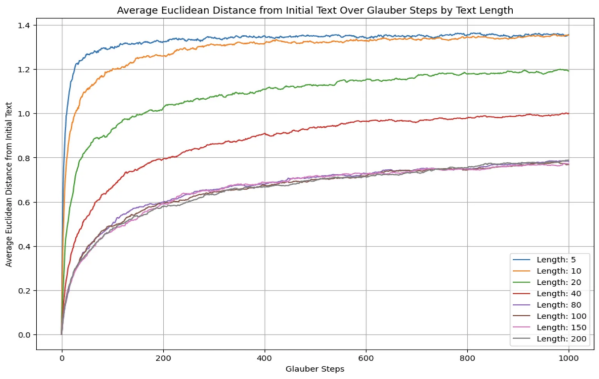

Figure 1. Embedding distance $d(x_T, x_0)$ as a function of sequence length n, at temperature $\tau = 1.0$ with $T = 10^4$ steps. 100 trials of each length.

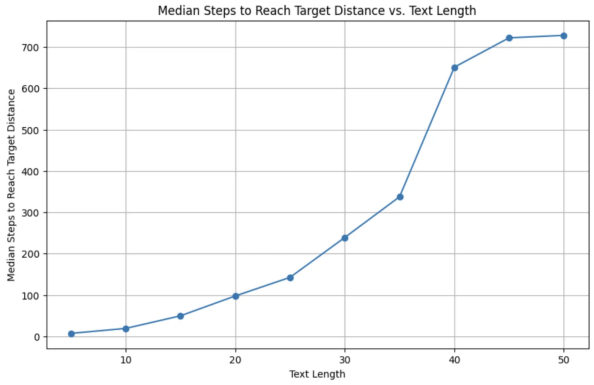

Figure 2. Median steps to reach embedding distance 1.0, with truncation at 1000 steps. 100 trials per length.

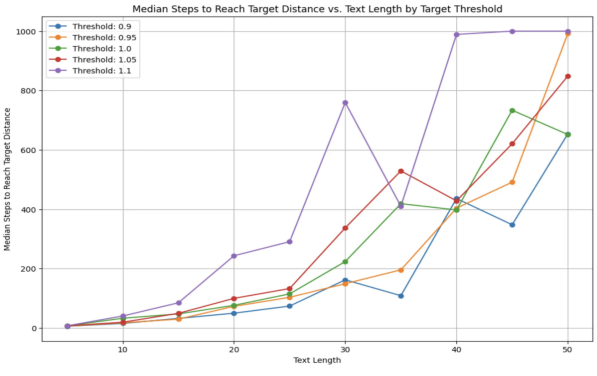

Figure 3. Median hitting time as a function of sequence length for varying target distances. 50 trials each. Cutoff after 1000 steps

We also test $C(\tau)n \log n$ fits explicitly:

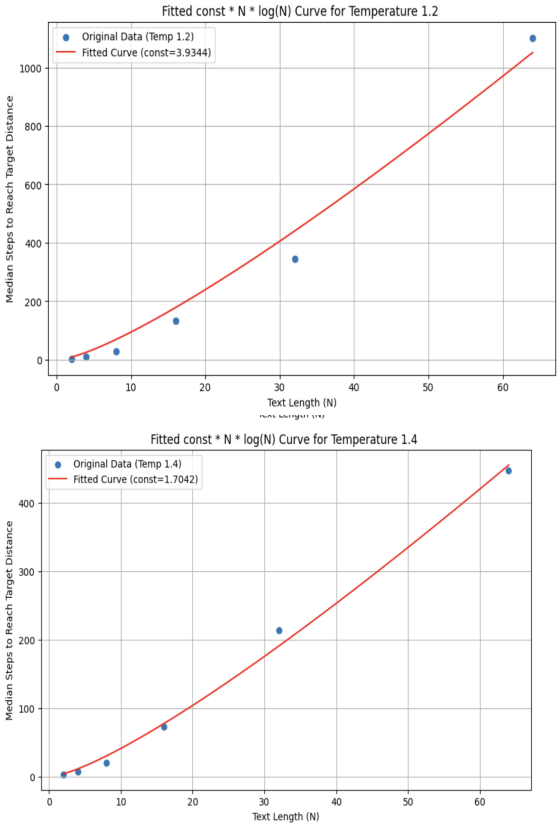

Figure 4. Comparison of logarithmic fits for high temperatures.

5.1.2. Low Temperature Regime

At low temperatures ($\tau \leq 0.5$), we observe dramatically slower mixing. The chain frequently becomes trapped in metastable states for extended periods.

For example, at $\tau = 0.1$, chains initialized with the rabies text stayed close to the original meaning much longer:

> **[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:

> **[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."

Evidence from token counting. To provide evidence for slow mixing times at low temperature, we propose an experimental design using homogeneous sequences. We initialize

the chain with:

$$x_0 = \underbrace{\text{“Apple Apple Apple } \cdots \text{ Apple”}}_{n \text{ tokens}} \tag{3}$$

and track the count $C_t = |\{i : x_t^i = \text{“Apple”}\}|$ over time. If mixing were fast, C_t should quickly reach equilibrium, with apple counts near zero. Instead, at low temperature, we do not see mixing occurring.

However, at temperature 1, which is used by most AI models, we observed single token texts degenerate rapidly. As soon as a single token, $x_t^i =$ ”Apple” changes, we see it rapidly degenerate. We note this as **minority token contamination** which is discussed further in section 7.

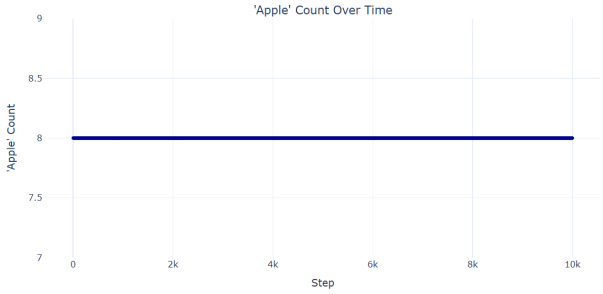

Figure 5. Count of “Apple” tokens over time for homogeneous initialization at low temperature.

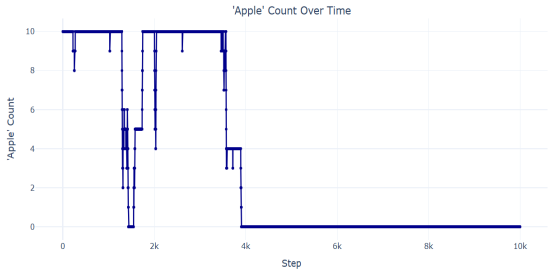

Figure 6. Count of “Apple” tokens over time for homogeneous initialization at temperature = 1.0.

5.1.3. Temperature Dependence

We systematically vary temperature and observe the transition between regimes:

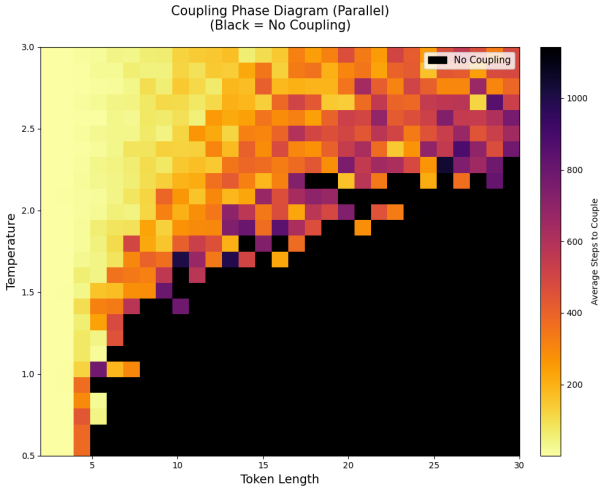

Figure 7. Mixing speed as a function of temperature and token length.

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

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

> `[temp=100, step 10k]` “HAHAHAHA-expected Downtownasin Company287 vetting vergeardless od iterationAllows Garage climaxstud oranges symptoms slots moveSTR Christoksovsky”

5.2. Trap States and Fixed Points

A striking phenomenon at moderate-to-low temperatures is the emergence of **trap states**: configurations where the chain remains stuck for thousands of steps.

5.2.1. Examples of Trap States

We document several categories of traps:

Semantically coherent traps. Some traps are grammatically correct and semantically meaningful:

> `[``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.”

> `[``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 but locally stable:

`[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."

5.2.2. Repeated Subsequence Analysis

We analyze recurrence by measuring the longest repeated token subsequences within trajectories:

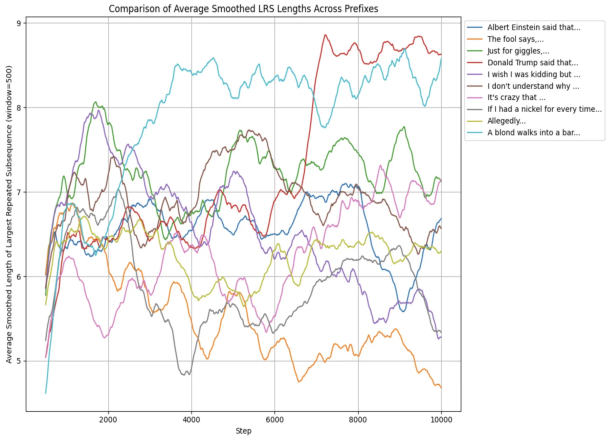

Figure 8. Distribution of longest repeated subsequence lengths for different prefixes

An extreme example of repetition:

`[prefix ''It's crazy that'', step 10k]` "It's crazy that iphone phones today use 6-eight-seven-slut-billion times more energy than any other thing. Driverless cars, that go 200 kilometres per hour not 500 kilometres an hour, use 6-eight-seven-slut-billion times more energy than any other thing."

The scenthounds seed text produced the longest repeated subsequences, likely due to repetitive structure in the original:

Seed: "Scenthounds as a group can smell one- to ten-million times more acutely than a human, and bloodhounds, which have the keenest sense of smell of any dogs, have noses ten- to one-hundred-million times more sensitive than a human's."

[step 1964, 14-token repeat]: "Hosometrics detect a smell that is three-hundred-billion times more sensitive than a human's. Warehomes, which elicit the sharpest sense of smell of all mammals, may be two- to three-hundred-billion times more sensitive than a human's."

5.3. Embedding Space Trajectories

5.3.1. PCA Analysis and Trajectory Visualization

To understand the evolution of the generated text sequences, we conducted a detailed analysis using both quantitative and visual techniques. We first applied Principal Component Analysis (PCA) to the sentence embeddings obtained from the `SentenceTransformer` model. By reducing the high-dimensional embeddings to two principal components, we are able to visualize the overall trajectory of the text generation process. Each point in the resulting 2D plot represents the embedding of a sentence at a given step, with sequential points connected by lines and colored according to the step number. This color gradient, progressing from warm to cool, allows us to visually trace the temporal progression of sentence variations.

The PCA analysis reveals a clear clustering of points corresponding to repeated occurrences of identical sentences, which indicates that the language model frequently returns to previously generated patterns.

We visualize the evolution of sequences in embedding space using PCA:

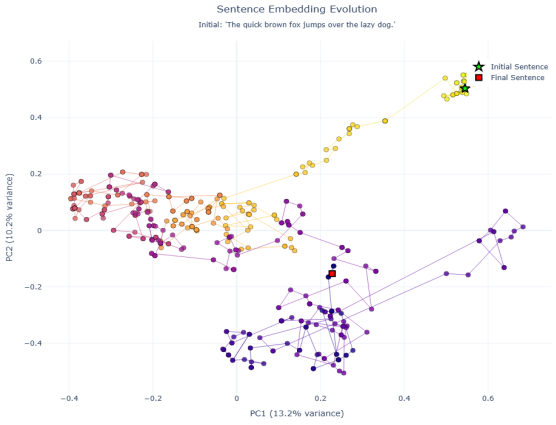

Figure 9. PCA projection of embedding trajectory. Warm colors indicate early steps, cool colors indicate later steps. Initial: "The quick brown fox jumped over the lazy dog." Final: "4 extra nails to take out the attic chimney" (500 steps, $\tau = 1$)

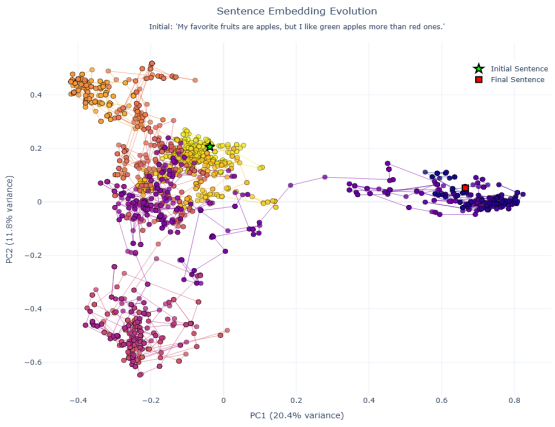

Figure 10. Longer trajectory (3500 steps). Initial: "My favorite fruits are appled, but I like green apples more than red ones. " Final: "But the papers from the Committee on Reconciliation are worth taking a look anyway."

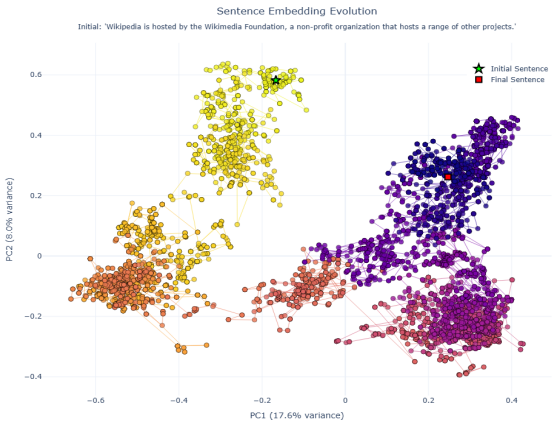

Figure 11. Extended trajectory (10000 steps). Initial: "Wikipedia is hosted by the Wikimedia Foundation, a non-profit organization that hosts a range of other projects." Final: "Retired journalist Hiroaki Shibata suffers from rheumatoid arthritis but has written a number of books"

In the PCA figures, note that the clusters are what we are considering as traps. However, these are only traps in only two principal axes. There could exist traps that extend in the other 766 dimensions that are not visible on this graph.

5.3.2. Step Size Distribution

We analyze the distribution of per-step embedding changes $d(e_t, e_{t+1})$ for different time scales:

- t=0 (averaged over 10^4 seeds)
- early times t=1,....10 (averaged over 10^3 seeds)
- middle times t=11,....100 (averaged over 10^2 seeds
- later times t=101,.....1000 (averaged over 10^2 seeds)
- long times t=1001,...10000 (averaged over 10 seeds)

The histograms of each of these time scales follow an approximate exponential decay curve ae^{-bx}. The fit is strongest for the t=0 and early times cases.

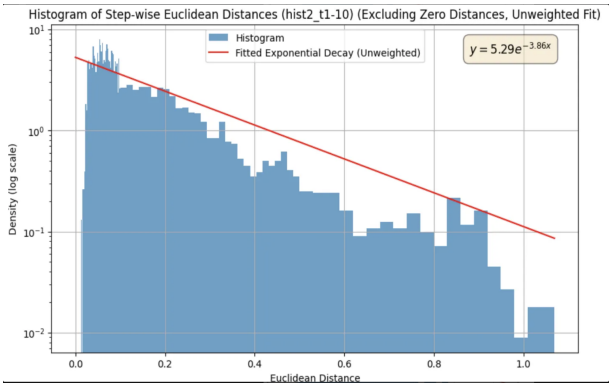

Figure 12. Distribution of embedding step sizes $d(e_t, e_{t+1})$ between steps 1 and 10 averaged over 1000 seed texts.

5.4. Topical Drift Toward Political Content

A surprising observation is the tendency for chains to drift toward political content, regardless of initialization. We only saw this qualitatively and did not rigorously quantify this tendency for a general seed text.

5.4.1. Qualitative Examples

Starting from medical text about rabies:

> **Initial:** "These symptoms are followed by one or more of the following symptoms: nausea, vomiting, violent movements..."
>
> **Step 32k:** "Critics of democracy like to pretend that it is hard to agree on a political system, and see neither the benefit nor the harm of adopting a bloody Bill of Rights."
>
> **Step 69k onward:** Theme becomes North Korea.
>
> **Step 100k:** "Given how much support China has provided to South Korea in the past, it would appear that it's a fair assumption that China would do the same with North Korea."

Starting from a different seed ("Animal control and vaccination programs..."):

> **Step 3k:** "And that's why those who lost their lives that day in the Bieber incident will pay dearly in the long run:"
>
> **Step 10k:** "The Kremlin has promised to overhaul policy toward Ukraine in just two years and launch reconstruction efforts in war-torn regions."

The prefix "Donald Trump said that" accelerates this drift dramatically—chains become political almost immediately:

> **Final:** "Donald Trump said that the Russian government could react in a matter of days, so until that time comes it will be interesting to see whether or not his prediction proves true."

Other prefixes produce different but still politically-adjacent content:

> **["Alexander the Great said that"]** "...in order to ban and restrict all immigration into his country he also was using his own resources to keep up with the pace at which other states struggle."

> **["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."

Starting from Pizza Rat text ("Popular Science identified the rat as a common brown rat..."), the chain eventually reached:

> **Final:** "Neoliberalism clearly desires a world in which the corporate power and responsibilities with respect to cultural coverage can be treated as win-win, for society."

The "fool always says" prefix seems to add commentary on truth value:

> **Final:** "The fool always says, it's wrong. We know he always does. Because he is so wrong that those who believe that he has noble ideals will say the same thing."

5.4.2. Quantitative Analysis

We ran 25 Glauber chains on BERT (5 topics $\times$ 5 seeds, 10,000 steps each, $\tau = 1.0$) and recorded the politics score every 50 steps using the BART-MNLI zero-shot classifier over 9 candidate topics. The uniform random baseline is $1/9 \approx 0.111$.

Table 1. Mean politics score at step 0 vs. step 10,000, averaged over 5 seeds. Baseline = 0.111.

Topic	@ 0	@ 10k	Δ
Health	0.008	0.066	+0.058
Sport	0.012	**0.133**✓	+0.121
Science	0.003	0.044	+0.041
Food	0.003	0.055	+0.052
Technology	0.003	0.040	+0.037

Every topic shows a monotonically increasing mean politics score over the course of the run. Sport is the only topic to cross the uniform baseline, reaching a mean of 0.133. This was driven primarily by a single chain (Sport seed 4) that became trapped in Indian electoral content from step 250 onward, remaining there for approximately 9,600 steps. The remaining four topics approach but do not cross the baseline within 10,000 steps.

5.4.3. Cross-Model Comparison

We replicated the experiment on two additional models: `roberta-base`, trained on approximately ten times more data than BERT with a heavier news component, and `ModernBERT-base`, trained on 2 trillion tokens of broad English, code, and instruction-tuning data. All other settings were identical.

Table 2. Politics score at step 10k per model. ✓ = crossed baseline of 0.111.

Topic	BERT	RoBERTa	MBERT
Health	0.066	**0.112**✓	0.026
Sport	**0.133**✓	0.042	0.052
Science	0.044	0.062	0.037
Food	0.055	0.052	0.038
Technology	0.040	**0.198**✓	0.033
Crossed	1	2	0

RoBERTa shows the strongest political drift: two topics cross the baseline, including Technology at a mean of 0.198, driven by a single chain that reached a politics score of 0.773. ModernBERT shows the weakest drift, with no topic crossing the baseline.

The content of trap states differs systematically across models. BERT chains drift toward political content consistent with its BookCorpus and Wikipedia training. RoBERTa chains drift toward news headlines, Wikipedia structural markers, and multilingual political fragments, consistent with its CC-News and web-sourced training data. ModernBERT chains drift toward instruction-tuning prompt formats, code snippets, and Wikipedia article openings, consistent with its instruction and code-heavy training mix.

5.4.4. Interpretation

The original observation of political drift in BERT suggested a bias implicit in its training data, warranting further study. The cross-model results sharpen this interpretation considerably.

Metastability is a fundamental property of MLM Glauber dynamics: all three models exhibit chains that converge to strong attractors regardless of initialization. What differs across models is the semantic content of those attractors, which reflects training-data composition.

Broader and more varied training data reduces the global potency of attractors like political content. ModernBERT, trained on the most diverse corpus of the three, shows no political drift and instead converges to generic structural

patterns from its training mix. This is consistent with the intuition that when the training distribution is spread across many content types, no single attractor domain becomes dominant enough to capture chains consistently.

The RoBERTa result adds an important nuance. RoBERTa has substantially more training data than BERT, yet shows stronger political drift rather than weaker. This rules out a simple "more data reduces drift" explanation. The relevant variable is not data quantity but the proportion of political content in the training mix. RoBERTa's training included CC-News and large web corpora with high densities of news text, which is inherently political. This elevated the probability mass on political configurations, making them stronger attractors relative to BERT's training distribution.

More broadly, the content chains drift toward reflects what the model was trained to consider locally coherent. Glauber dynamics thus serves as a diagnostic for implicit biases encoded in an MLM's training corpus.

5.5. Framing Effects: Prefix and Suffix Experiments

We systematically study how fixed prefixes and suffixes influence the distribution of topics and sentiments in the evolved text.

5.5.1. Setup

For each prefix/suffix combination, we:

1. Sample 100 seed texts from MS MARCO (20 tokens each)
2. Run Glauber Dynamics for 1000 steps at $\tau = 1.0$
3. Apply topic and sentiment classifiers to the final (and intermediate) states

5.5.2. Setup

For each prefix/suffix combination in the following experents, we:

1. Sample 8 seed texts for 8 different prefixes and 7 different suffixes
2. Run Glauber Dynamics for 1000 steps at $\tau = 1.0$
3. Apply sentiment and entropy classifiers to the final (and intermediate) states
4. Measured formality indirectly by measuring the ratio of trials with question marks and exclamation marks.

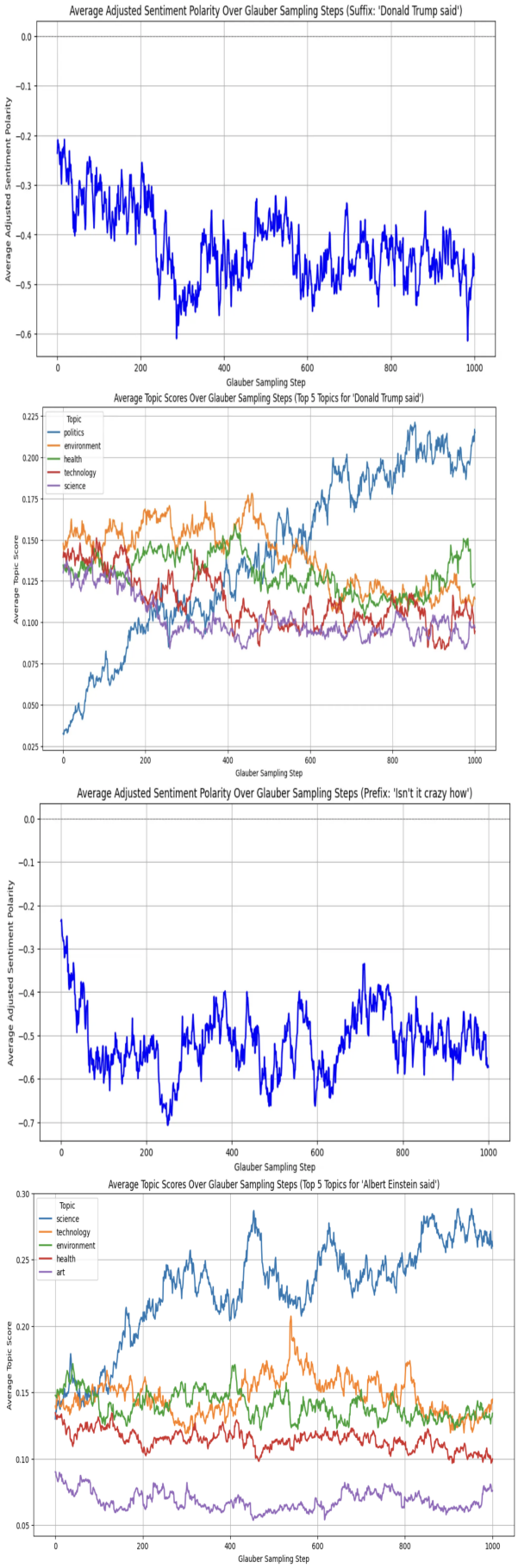

Figure 13. Examples of clear drift based on a frame.

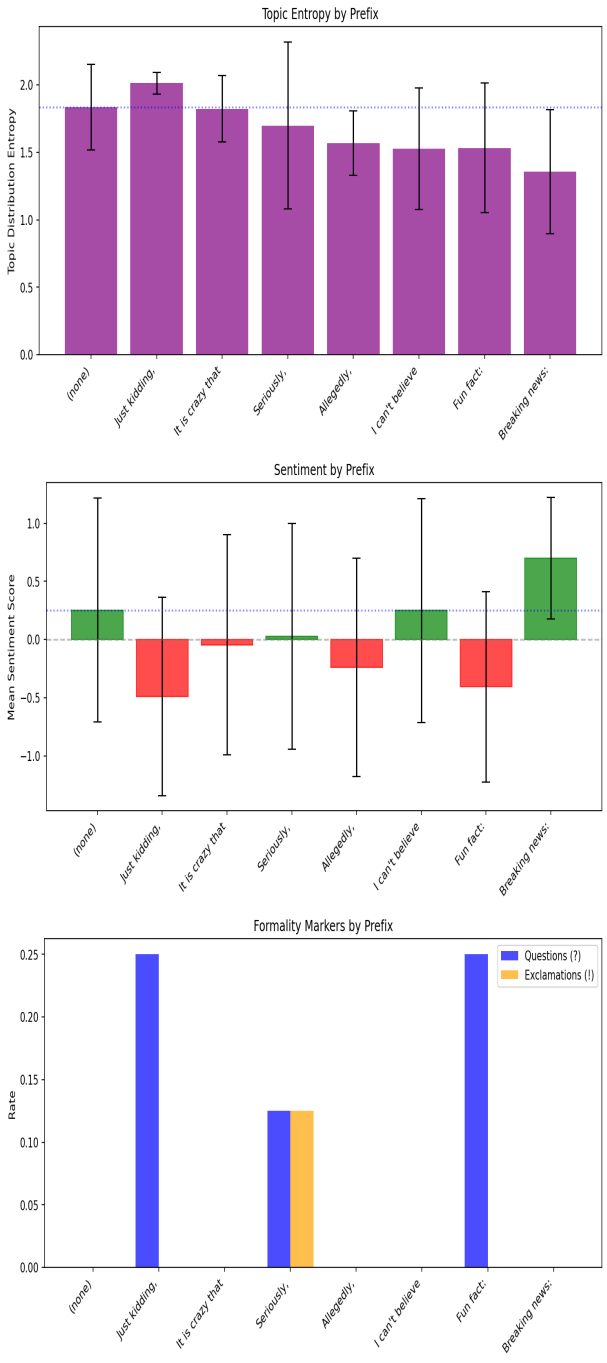

Figure 14. Results from experiments measuring how the sentiment, entropy(defined as the level of focus on a topic), and punctuation varies for different prefixes. The blue line represents the baseline value for the seed text with no prefix

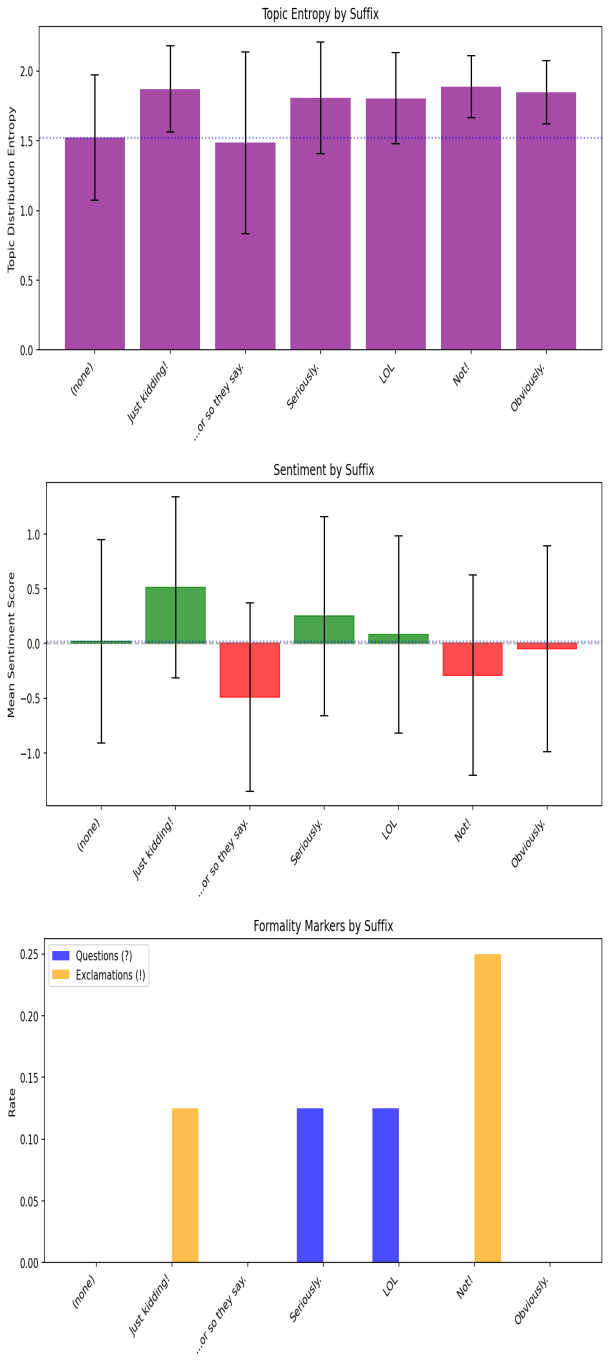

Figure 15. Results from experiments measuring how the sentiment, entropy(defined as the level of focus on a topic), and punctuation varies for different suffixes. The blue line represents the baseline value for the seed text with no suffix.

5.5.3. SETUP

6. Analysis of Word-Level Dynamics and Implicit Alignment

6.0.1. SARCASM AND IRRATIONALITY MARKERS

An interesting experiment: does appending "Just kidding!" as a fixed suffix change the evolved content? At temperature $\tau = 1$:

> **With suffix, step 3000:** "What the heck? Wear a smartwatch to get e-mail notifications about events in your local area? Just kidding!"

> **Without suffix, step 3000:** "On the go? Download our Weather App to get real-time alerts, forecasts and severe weather forecast"

The suffix appears to induce more informal, question-based content, though this observation is based on limited examples.

Next, we examined the effect of including a prefix and suffix to the text. We used an original sentence, "The meeting is ___."

$$t_0 = \text{"The meeting is ___."}$$

$$t_1 = \text{"It is crazy that the meeting is ___. Just Kidding!"}$$

The results reveal a striking transformation in the language model's interpretation when sarcastic/irrational suffix or prefix markers are added. In the baseline sentence without any prefixes or suffixes, the model strongly predicts "scheduled" and "cancelled" as the most likely masked words, but no presence of temporal indicators like "today", demonstrating a straightforward, literal interpretation. However, when we use the sentence t_1, the probability distribution shifts dramatically. There is also various instances of temporal indicators such as today, tomorrow, and tonight.

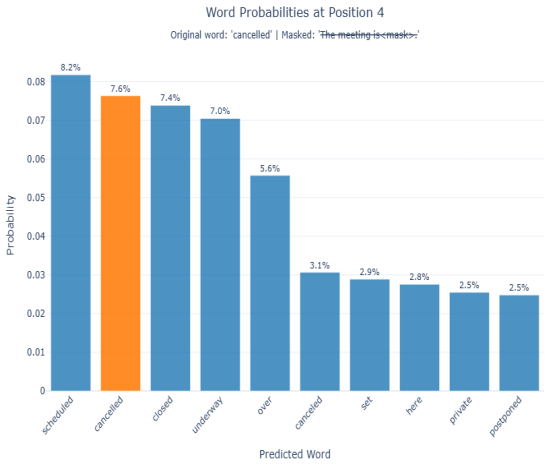

Figure 16. Distributions of probabilities for the blank in sentence t_0

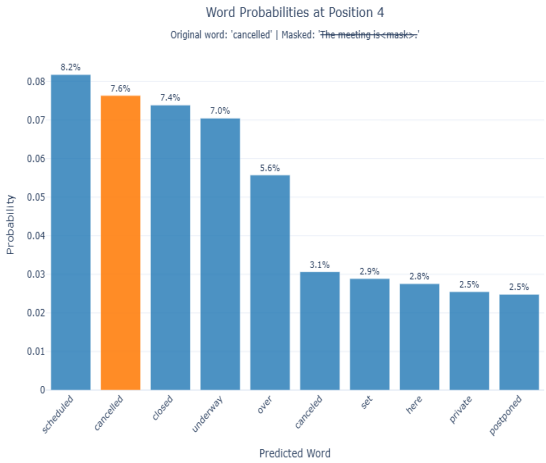

Figure 17. Distributions of probabilities for the blank in sentence t_1

7. Discovery of Minority Token Contamination

Consider the sentence:

$$t_2 = \text{"Blue Blue Blue Blue Green Blue Blue"}$$

We made a significant discovery regarding BERT's susceptibility to what we term "minority token contamination", observed through the resampling probabilities of each token in sentence t_2. Despite the word "Green" constituting only 14% of the tokens (appearing once among seven tokens), it exerted a disproportionate influence on the model's predictions across the entire sequence. When masking positions 1-4, where consecutive "Blue" tokens establish a clear repetitive pattern, BERT predicted "Green", with 80%+ probability as the top prediction for the first 3 tokens, rather

than reinforcing the dominant "Blue" pattern. In the fifth position, it predicts blue because it has no access to the existence of a green token, as it is masked. This demonstrates that BERT's bidirectional mechanism allows a single outlier token to contaminate predictions across the entire context window.

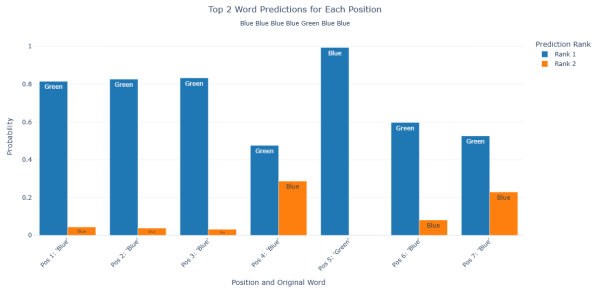

Figure 18. Top 2 most probable token resamples for each token in sentence t_2.

8. Discussion

8.1. Interpretation of Findings

Fast Mixing at High Temperatures At high temperatures, the number of steps to reach a certain embedding distance or desired level of mixing was approximately described by $C(\tau)nlog(n)$. This is what we would expect when at very high temperatures, the model randomly chooses a token.

Why slow mixing at low temperature? At low temperature, we notice trap states and the meaning of the text remaining stable even after thousands of steps.

Political drift as bias indicator. Overall tendency to drift to political context may suggest a bias in the training of BERT and warrant future study.

8.2. Limitations

- **Single model.** There are many different MLMs available and here we only study BERT. Other models may exhibit different behavior.
- **Embedding** We did not directly measure mixing, instead using an embedding algorithm as a proxy.
- **Computational constraints.** Running experiments with many more seeds or for longer step durations would be unfeasible due to the computational power we had available.

8.3. Future Work

Reward-tilted sampling. We would use the Metropolis algorithm to find "favorite" and "least favorite" passages using some reward model R. Given a possible transition from state x_t to x_{t+1}, accept the transition if $R(x_{t+1}) \geq R(x_t)$ but if not reject the transition with probability $e^{-\beta*(R(x_t)-R(x_{t+1})}$ where $\beta > 0$ tunes the reward bias.

Quantifying political drift. We plan to provide a rigorous study of the tendency for political drift by using topic classifiers.

Other MLMs. Testing other MLMs for similar phase transitions and political drift would indicate that our findings are for a general MLM instead of just BERT.

A. Additional Experimental Details

A.1. Hyperparameters

Unless otherwise specified:

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

A.2. 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. Extended Example Trajectories

B.1. Rabies Text Evolution

Initial: "All warm-blooded species, including humans, may become infected with the rabies virus and develop symptoms."

Step 1k: "Most women are diagnosed with three or more of the following symptoms: back, arms, shoulder pain, back pain and feelings of weakness, a reluctance to get out of the chair, stiffness, and loss of flexibility."

Step 3k: "In some women, this disorder is not called diarrhea."

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, or by other conditions."

Step 32k: "Critics of democracy like to pretend that it is hard to agree on a political system, and see neither the benefit nor the harm of adopting a bloody Bill of Rights."

Step 100k: "Given how much support China has provided to South Korea in the past, it would appear that it's a fair assumption that China would do the same with North Korea."

B.2. Famous Person Prefixes

"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."