BLINK OF AN EYE: A SIMPLE THEORY FOR FEATURE LOCALIZATION IN GENERATIVE MODELS

Marvin Li, Aayush Karan, & Sitan Chen
Department of Computer Science
Harvard University
Cambridge, MA 02138, USA

ABSTRACT

Large language models can exhibit undesirable and unexpected behavior in the blink of an eye. In a recent Anthropic demo, Claude switched from coding to Googling pictures of Yellowstone, and these sudden shifts in behavior have also been observed in reasoning patterns and jailbreaks. This phenomenon is not unique to autoregressive models: in diffusion models, key features of the final output are decided in narrow "critical windows" of the generation process. In this work we develop a simple, unifying theory to explain this phenomenon. We show that it emerges generically as the generation process localizes to a sub-population of the distribution it models. While critical windows have been studied at length in diffusion models, existing theory heavily relies on strong distributional assumptions and the particulars of Gaussian diffusion. In contrast to existing work our theory (1) applies to autoregressive and diffusion models; (2) makes no distributional assumptions; (3) quantitatively improves previous bounds even when specialized to diffusions; and (4) requires no Girsanov or statistical-physics-based machinery.

1 INTRODUCTION

In large language models (LLMs), undesirable behavior can often emerge very suddenly. For example, Claude transitioned from coding to browsing pictures of Yellowstone while using a computer (Anthropic, 2024); the Phi-4 team reported that the probability of correctly answering a math problem can plummet with a single token (Abdin et al., 2024; Lin et al., 2024); Gemini abruptly threatened a student who was using it to study (Gemini, 2024); Llama models can be jailbroken by manipulating the first handful of generated tokens (Qi et al., 2024; Haize Labs, 2024b).

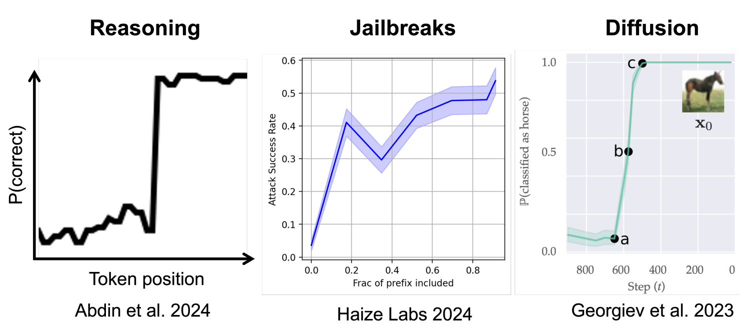

Figure 1: Examples of critical windows across different modalities and samplers, including image diffusion models (Georgiev et al., 2023), and chain of thought for reasoning and jailbreaks in LLMs (Abdin et al., 2024; Haize Labs, 2024b).

These abrupt shifts are not unique to autoregressive models. In diffusion models, it has been observed that certain properties like the presence of an object in the background or the image class emerge in narrow time intervals, sometimes called *critical windows*, of the generation process (Ho et al., 2020;

Meng et al., 2022; Choi et al., 2022; Raya & Ambrogioni, 2023; Georgiev et al., 2023; Sclocchi et al., 2024; 2025; Biroli et al., 2024; Li & Chen, 2024). Critical windows, more broadly characterizable as *a few steps of the sampling procedure* during which features of the final output appear, arise in many different generative models and data modalities (Figure 1). They are extremely useful from an interpretability perspective as they represent the steps of the sampler responsible for a given property of the output (Georgiev et al., 2023; Qi et al., 2024), and have also been used to provide richer stepwise rewards for preference optimization and finetuning (Abdin et al., 2024; Lin et al., 2024; Qi et al., 2024). As the applications of generative models proliferate, it is crucial from interpretability, safety, and capability perspectives to understand how and why these critical windows emerge.

Recently, this phenomenon has received significant attention within the theoretical literature on diffusion models (Raya & Ambrogioni, 2023; Sclocchi et al., 2024; 2025; Biroli et al., 2024; Li & Chen, 2024). While existing works do offer predictive theory in the diffusions setting, they either (A) make strong distributional assumptions or (B) rely heavily on the particulars of diffusion, which do not straightforwardly extend to autoregressive models. Works in the former category carry out non-rigorous statistical physics calculations tailored to specific toy models of data like mixtures of Gaussians or context-free grammars with random production rules (Sclocchi et al., 2025; 2024; Biroli et al., 2024; Raya & Ambrogioni, 2023). Works in the latter category derive rigorous bounds in settings without explicit parametric structure, e.g. mixtures of strongly log-concave distributions (Li & Chen, 2024), but they rely on tools like Girsanov's theorem which are specific to Gaussian diffusion. Additionally, the bounds in the latter are generally cruder, losing dimension-dependent factors. We ask:

Is there a simple, general theory that can explain critical windows across all generative modeling paradigms and data modalities?

1.1 Our contributions

In this work, we develop a simple theoretical framework that explains critical windows in both diffusion models and autoregressive models. Our bounds are fully rigorous and show that such windows arise generically when the model localizes from a larger sub-population to a smaller one, which we will formalize in the following section. Below we highlight our main contributions:

1. **Generality**: In comparison to existing work, our theory (Theorem 3.1) makes no distributional assumptions, requires no statistical physics or stochastic calculus machinery, and removes the dimension dependence suffered by existing rigorous bounds.
2. **Diverse instantiations**: To illustrate the flexibility of our bounds, we explicitly compute the locations and widths of these windows for different generative models and data modalities (Section 4) and empirically verify our predictions on structured output examples. One such example we provide elucidates a new connection between critical windows for in-context learning and the all-or-nothing phenomenon in statistical inference.
3. **Insights into hierarchical data**: We instantiate our bounds for hierarchically structured models of data, significantly generalizing results of (Li & Chen, 2024) which only applied to diffusions and Gaussian mixtures (Section 5). This allows us to show that the hierarchy for a generative model may resemble the hierarchy of the true data generating process if both come from the same kind of sampler, but in general may differ.
4. **Experimental results**: Finally, we empirically demonstrate critical windows for generations from LLAMA-3.1-8B-Instruct, Phi-3-7B-Instruct, and Qwen-2.5-7B-Instruct on 7 different math and reasoning benchmarks (App. F). Concurrent with (Abdin et al., 2024; Lin et al., 2024), we observe that critical windows occur during important mistakes in the reasoning patterns of LLMs.

In fact, our theory applies more generally to any *stochastic localization sampler* (see Section 2.1 for a formal description) Montanari (2023b); Chen & Eldan (2022). Roughly speaking, a stochastic localization scheme is any generative model given by a time-reversal of a Markovian degradation process which takes a sample from the target distribution and generates progressively less informative "observations" of it. In diffusion models, the degradation is a convolution of the original sample with larger and larger amounts of Gaussian noise. In autoregressive models, the degradation is the progressive masking of entries from right to left. Importantly, our theory does not use anything about the specific structure of the sampler beyond the Markovianity of the observation process.

Finally, our theory provides valuable insights for practitioners. For instance, in Example 4.3 we provide a model for critical windows in jailbreaks and the Yellowstone example (Anthropic, 2024; Qi et al., 2024), and argue that training on corrections from critical windows can enable models to recover from these 'bad' modes of behavior. This provides rigorous theoretical justification for Qi et al. (2024)'s approach for deepening safety alignment through finetuning.

2 Technical preliminaries

Probability notation. Given distributions P, Q defined on $(\Omega, \mathcal{F})$ with a base measure μ, the *total variation distance* is defined as $\mathrm{TV}(P,Q) \triangleq \frac{1}{2}\int |dP - dQ| d\mu$. For random variables X, Y, we will also use $\mathrm{TV}(X,Y)$ as shorthand to denote the TV of the measures of X, Y. Let $\mathrm{supp}(P) = \{x \in \Omega | dP(x) > 0\}$ denote the support. We will also use the following well-known relationship.

Lemma 2.1. *For probability measures P, Q, $\mathbb{E}_{x\sim P}\left[\frac{\mathrm{d}Q}{\mathrm{d}P+\mathrm{d}Q}\right] \leq \frac{1}{2}\sqrt{1-\mathrm{TV}^2(P,Q)}$.*

To study feature localization in diffusion and autoregressive models, we consider a *forward-reverse* experiment. A forward-reverse experiment considers the amount of "noise" so that running the generative model starting from the noised generation would still yield a sample with the same feature. For a diffusion model, this could mean taking an image of a cat, adding Gaussian noise, and resampling to see if the result is still a cat. For a language model, it could mean truncating a story about a cat and resampling to check if the story remains about a cat. Now, we will use the language of stochastic localization to place these analogous experiments for diffusion and language models within the same framework.

2.1 Stochastic localization samplers

We formally define the framework for stochastic localization samplers, following Montanari (2023b). Let $X \sim p$ be a random variable over $\mathbb{R}^d$.[1] We consider a sequence of random variables $(Y_t)_{t\in\mathbf{I}}$ with a compact index set $\mathbf{I} \subset [0,\infty) \bigcup \{\infty\}$. As t increases, Y_t becomes *less informative* and *degrades* the original information about X (Definition 2.2). As in Montanari (2023b), we will only consider *complete* observation processes, where information about the path $(Y_t)_{t\in\mathbf{I}}$ uniquely identifies X: for any measurable set $A \subset \mathbb{R}^n$, we require $P(X \in A | (Y_t)_{t\in\mathbf{I}}) \in \{0,1\}$. For the sake of simplicity, we will assume $0, \infty \in \mathbf{I}$ and Y_∞ is totally uninformative about X.

Definition 2.2. $(Y_t)_{t\in\mathbf{I}}$ is an *observation process* with respect to X if for any positive integer k and sequence $t_1 < t_2 < \cdots < t_k \in \mathbf{I}$, the sequence $X \to Y_{t_1} \to Y_{t_2} \to \cdots \to Y_{t_k}$ forms a Markov chain.

Because $X \to Y_{t_1} \to \cdots \to Y_{t_k}$ is a Markov chain, its reverse $Y_{t_k} \to \cdots \to Y_{t_1} \to X$ is also a Markov chain. To any such observation process one can thus associate a generative model as follows:

Definition 2.3. Given observation process $(Y_t)_{t\in\mathbf{I}}$ and times $t_1 < \cdots < t_m = \infty$ in $\mathbf{I}$, the associated *stochastic localization sampler* is the algorithm that generates a sample for X by first sampling Y_{t_m} and then, for $k = m-1, m-2, \ldots, 0$, sampling from the posterior on Y_{t_k} conditioned on $Y_{t_{k+1}}$ by taking one step in the reverse Markov chain above, and finally sampling X conditioned on Y_{t_0}.

In Appendix B, we formally verify that diffusion and autoregressive models are special cases of this framework. In practice, one does not have access to the true posteriors of the data distribution and must learn approximations to the posterior from data. This issue of learning the true distribution is orthogonal to our work, and thus we define $X \sim p$ to be the sampler's distribution. Furthermore, it is more natural to study the sampler's distribution for applications such as interpretability or jailbreaks.

Features, mixtures, and sub-mixtures. To capture the notion of a feature of the generation, we assume that the distribution $X \sim p$ is a *mixture model*. Consider a discrete set $\Theta = \{\theta_1, \ldots, \theta_K\}$ with non-negative weights $w_1, \ldots, w_K$ summing to 1. Each $\theta_i \in \Theta$ is associated with a probability density function $p^{\theta_i} : \mathbb{R}^n \to \mathbb{R}^{\geq 0}$. To generate a sample $X \sim p$, we first draw $\theta \sim \mathrm{Cat}(\Theta, \{w_i\}_{i=1}^K)$ and return $X \sim p^\theta$. This yields an overall density of $p \triangleq \sum_{\theta\in\Theta} w_\theta p^\theta$. For any non-empty $S \subset \Theta$, we also define the sub-mixture p^S by $p^S \triangleq \sum_{\theta\in S} \frac{w_\theta}{\sum_{\phi\in S} w_\phi} p^\theta$.

[1] These definitions are easily carried over to the setting where X lives in a discrete space.

Here we study a *family of observation processes* corresponding to observation processes for different initial distributions of $X \sim p^S$ for $S \subset \Theta$. To ensure that we can meaningfully compare the observation processes within this family, we will assume that the *degradation procedure is fixed*. To formalize this intuition, we borrow the language from diffusion models of a forward process, which degrades X, and a reverse process, which takes a degraded Y_t and produces X.

2.2 FORWARD-REVERSE EXPERIMENT

Now we describe the general formalism under which we will study critical windows. Fixing some $t \in \mathbf{I}$ and $S \subset \Theta$, we start with some $X \sim p^S$, sample $Y_t|X$ from the observation process conditioning on X, and finally take $X'|Y_t$ from the stochastic localization sampler conditioning on Y_t. The can be understood as a generalization of the *forward-reverse* experiment in diffusions, originally studied in (Sclocchi et al., 2025; 2024; Li & Chen, 2024), to arbitrary stochastic localization samplers.

Forward process. For any $t \in \mathbf{I}$, define the forward Markov transition kernel $P_t^{\rightarrow}(A|X) = P(Y_t \in A|X)$. Note the forward Markov transition kernel does not depend on the distribution of X. The fact that the forward process is agnostic to the specifics of the original distribution is shared by the most widely used stochastic localization samplers. For example, in diffusion and flow-matching models, the forward transition is a convolution of X with a Gaussian; in autoregressive language models, it is masking of the last remaining token in the sequence.

For any $t \in \mathbf{I}$ and $S \subset \Theta$, we let p_t^S denote the law of Y_t^S, where we sample $X^S \sim p^S$ and then sample $Y_t^S \sim P_t^{\rightarrow}(\cdot|X^S)$. We omit the Θ in p_t^{Θ}.

Reverse process. For any $t \in \mathbf{I}$ and initial distribution $X \sim p$, we define the posterior of X given Y_t by $P^{\leftarrow}(A|Y_t) = P_{X\sim p}(X \in A|Y_t)$, that is, the distribution of X given by starting the sampling process at $t \in \mathbf{I}$ and Y_t instead of ∞ and Y_∞. We will also use this notation for the probability density.

We are ready to describe the main forward-reverse experiment that we will study.

Definition 2.4 (Forward-reverse experiment (Sclocchi et al., 2025; 2024; Li & Chen, 2024))**.** For nonempty $S \subset \Theta$ and $\widehat{T} \in \mathbf{I}$, let $p^{S,\widehat{T}}$ be the distribution of $X^{S,\widehat{T}}$ defined by the following procedure: (1) Sample $Y_{\widehat{T}}^S \sim p_{\widehat{T}}^S$, i.e. run the forward process for time $\widehat{T}$ starting at the sub-mixture p^S; (2) Sample $X^{S,\widehat{T}} \sim P^{\leftarrow}(\cdot|Y_{\widehat{T}}^S)$, i.e. run the reverse process starting at $Y_{\widehat{T}}^S$ to sample from the posterior on X.[2]

We emphasize that in the second step, we run the reverse process with the prior on X given by the *entire distribution* p rather than the sub-mixture p^S. If we did the latter, the marginal distribution of the result would simply be p^S. Instead, the marginal distribution of $X^{S,\widehat{T}}$ is some distribution whose relation to p and sub-mixtures thereof is *a priori* unclear. Intuitively, as $\widehat{T} \to 0$, this distribution converges to p^S, and as $\widehat{T} \to \infty$, this distribution converges to p. The essence of our work is to understand the transition between these two regimes as one varies $\widehat{T}$.

3 CHARACTERIZATION OF CRITICAL WINDOWS

Let $S_{\mathrm{init}} \subset \Theta$ denote some sub-mixture, corresponding to a sub-population of p that possesses a certain property. For instance, if p corresponds to some autoregressive model, S_{init} might correspond to sentences which correctly answer a particular math question. Let $S_{\mathrm{targ}} \supset S_{\mathrm{init}}$ denote some sub-mixture containing S_{init}. For instance, S_{targ} might correspond to all possible responses to the math question, including incorrect ones. We are interested in the following question: if we run the forward-reverse experiment for time $\widehat{T}$ starting from $p^{S_{\mathrm{init}}}$, is there some range of times for which the resulting distribution is close to $p^{S_{\mathrm{targ}}}$? That is, can we characterize the $\widehat{T}$ for which $\mathrm{TV}(p^{S_{\mathrm{init}},\widehat{T}}, p^{S_{\mathrm{targ}}})$ is small? Suppose one could prove that the range of $\widehat{T}$ for which this is the case is some interval $[T_0, T_1]$. This would mean that if the stochastic localization sampler runs for time T

[2]Note that this equips 2^{Θ} with the structure of a poset, i.e. $A \subset B$ if and only if there exists some $t \in \mathbf{I}$ such that running the forward-reverse experiment up to t from p^A yields p^B.

and ends up at a sample from $p^{S_{\text{init}}}$, then from time $T - T_1$ to time $T - T_0$ of the generation process, *the sampler has not yet localized the features that distinguish* $p^{S_{\text{init}}}$ *from the larger sub-mixture* $p^{S_{\text{targ}}}$. However, the sampler has localized the features that distinguish $p^{S_{\text{targ}}}$ from $p^{\Theta - S_{\text{targ}}}$. When there is a shift from localizing the features S_{targ} to the features S_{init}, we say there is a critical window. We now formally state and prove our main result.

3.1 Main result

For an error parameter $0 < \epsilon < 1$, define $T_{\text{st}}(\epsilon) \in \sup\{t \in \mathbf{I} : \mathrm{TV}(p_t^{S_{\text{targ}}}, p_t^{\Theta - S_{\text{targ}}}) \geq 1 - \epsilon^2\}$ and $T_{\text{end}}(\epsilon) \in \inf\{t \in \mathbf{I} : \mathrm{TV}(p_t^{S_{\text{init}}}, p_t^{S_{\text{targ}}}) \leq \epsilon\}$. This is well-defined for continuous observation processes.[3] When the value of ϵ is understood, we abbreviate the above with T_{st} and T_{end}. Our main result is that in $\widehat{T} \in \mathbf{I} \cap [T_{\text{end}}, T_{\text{st}}]$, the distance $\mathrm{TV}(p^{S_{\text{init}}, \widehat{T}}, p^{S_{\text{targ}}})$ is small:

Theorem 3.1. *Let* $S_{\text{init}} \subset S_{\text{targ}} \subset \Theta$ *and* $W = \frac{\sum_{\theta \in \Theta - S_{\text{targ}}} w_\theta}{\sum_{\theta \in S_{\text{targ}}} w_\theta}$. *For* $\epsilon > 0$, *if* $\widehat{T} \in \mathbf{I} \cap [T_{\text{end}}, T_{\text{st}}]$, *then* $\mathrm{TV}(p^{S_{\text{init}}, \widehat{T}}, p^{S_{\text{targ}}}) \leq \epsilon \cdot \left(1 + \max\left(1, W\right)/\sqrt{2}\right)$.

We defer the proof to Appendix C. Intuitively, T_{st} represents the largest t for which there is still separation between S_{targ} and $\Theta - S_{\text{targ}}$, and T_{end} represents the smallest t for which samples from $S_{\text{init}}, S_{\text{targ}}$ are indistinguishable. Thus, running it for $\widehat{T} \in \mathbf{I} \cap [T_{\text{end}}, T_{\text{st}}]$ erases the differences between samples from S_{init} and S_{targ} but preserves the difference between S_{targ} and $\Theta - S_{\text{targ}}$, yielding samples looking like $p^{S_{\text{targ}}}$. Crucially, our proof relies in several places on the Markov property of stochastic localization samplers, together with the data processing inequality. Note a similar bound was shown in the context of diffusions by Li & Chen (2024) (see Theorem 7 therein). Our result is a strict improvement of that bound along several important axes. First, our results apply to all stochastic localization samplers, not just diffusions. Secondly, Li & Chen (2024) needed to assume that the components of p were strongly log-concave and that the score, i.e. gradient of the log-density, of p_t was Lipschitz and moment-bounded for all t. Thirdly, their final bound includes a polynomial dependence on the moments of the score, which scale with the dimension d; in contrast, our final bound is independent of d. With Theorem 3.1 in place, we are ready to formally define *critical windows*. These capture the moments where we transition from sampling from a sub-mixture to a subset of that sub-mixture.

Definition 3.2. Define $S_{\text{after}} \subset S_{\text{before}} \subset \Theta$. For S_{before}, we define $T_{\text{before}} = \inf\{t \in \mathbf{I} : \mathrm{TV}(p_t^{S_{\text{after}}}, p_t^{S_{\text{before}}}) \leq \epsilon$ and $\mathrm{TV}(p_t^{S_{\text{before}}}, p_t^{\Theta - S_{\text{before}}}) \geq 1 - \epsilon^2\}$ ($S_{\text{init}} \triangleq S_{\text{after}}; S_{\text{targ}} \triangleq S_{\text{before}}$). For S_{after}, consider $T_{\text{after}} = \sup\{t \in \mathbf{I} : \mathrm{TV}(p_t^{S_{\text{after}}}, p_t^{\Theta - S_{\text{after}}}) \geq 1 - \epsilon^2\}$ ($S_{\text{init}}, S_{\text{targ}} \triangleq S_{\text{after}}$). A **critical window** is the interval $[T_{\text{after}}, T_{\text{before}}]$, where there is a transition from sampling from S_{before} to the smaller subset S_{after}.

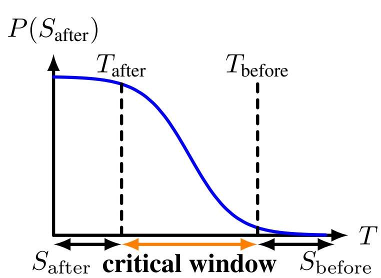

Figure 2: Illustration of the definition of a critical window for a simple mixture of two Gaussians, with $S_{\text{before}} \triangleq \{\mathcal{N}(\mu, \mathrm{Id}), \mathcal{N}(-\mu, \mathrm{Id})\}$ and $S_{\text{after}} \triangleq \{\mathcal{N}(\mu, \mathrm{Id})\}$.

4 Examples of critical windows

In this section, we analytically compute $T_{\text{before}}, T_{\text{after}}$ for diverse stochastic localization samplers and models of data, including diffusion and autoregression processes. In these natural settings, the critical window is small and has size which shrinks or does not depend on the dimension or context length. [4]

[3] See Remark C.1 for a more thorough discussion.

[4] Proofs are deferred to Appendix D.

4.1 DIFFUSION

We first consider two examples of Gaussian Mixture Models and a diffusion model. We show that with two isotropic Gaussians, the critical window appears around a single point, $\ln\|\mu\|$, with width independent of the dimension.

Example 4.1. [Two Isotropic Gaussians] Let $\Theta = \{\pm 1\}$, $p^{+1} = \mathcal{N}(\mu, \mathrm{Id})$, $p^{-1} = \mathcal{N}(-\mu, \mathrm{Id})$. Then, we have a critical window transitioning from sampling from both components to the component $+1$ between $T_{\text{before}} = \ln\|\mu\| + \ln 2 + \ln 1/\epsilon$ and $T_{\text{after}} = \ln\|\mu\| - \ln\ln\frac{1}{2\epsilon^2}$. When $\widehat{T} \le T_{\text{after}}$, then $\mathrm{TV}(p^{+1,\widehat{T}}, p^{+1}) \lesssim \epsilon$. When $\widehat{T} \ge T_{\text{before}}$, $\mathrm{TV}(p^{+1,\widehat{T}}, p) \lesssim \epsilon$.

For an isotropic Gaussian mixture model with randomly selected means, the critical window between sampling from one component to the entire mixture is also narrow. Note that we derive dimension-free widths in Example 4.2, an improvement over (Li & Chen, 2024) who had a $\ln\ln d$ dependence on dimension for isotropic Gaussians.

Example 4.2. [Random mean spherical Gaussians] We first sample $\mu_i \sim \mathcal{N}(0, \mathrm{Id})$ for $i \in [K]$ i.i.d. and let $\Theta = \{\mathcal{N}(\mu_i, \mathrm{Id})\}_{i\in[K]}$. We let $S_{\text{before}} = \Theta$ and $S_{\text{after}} = \{\mu_1\}$. Then, we can compute $T_{\text{before}} = \max_{j\in[K]} \ln\|\mu_i - \mu_j\| + \ln(1/\epsilon)$ and $T_{\text{after}} = \min_{j\in[K], i\neq j} \ln\|\mu_i - \mu_j\| - \frac{1}{2}\ln 8 \ln\frac{K}{\epsilon}$. Furthermore, with high probability over the selection of the means, $T_{\text{before}} - T_{\text{after}} = O_{K,\epsilon}(1)$ as $d \to \infty$.

Example D.5 of a discrete diffusion model, where $T_{\text{before}} - T_{\text{after}} \to 0$, is deferred to Appendix D.5.

4.2 AUTOREGRESSION

We first present a theoretical model for important critical windows in LLMs, e.g., jailbreaks that occur over the first generated tokens and the Yellowstone example (Anthropic, 2024; Qi et al., 2024).

Example 4.3. ["Critical Tokens" for Jailbreaks and Yellowstone (Qi et al., 2024; Anthropic, 2024)] Again consider an autoregressive language model, with $\mathcal{A}$ denoting the vocabulary, $p \in \mathcal{A}^T$, a forward process indexed by $\mathbf{I} = \{0, 1, 2, \ldots, T\}$, and Y_t to be the first $T - t$ tokens of X. Let $\Theta = \{\theta_{\textbf{harmful}}, \theta_{\textbf{safe}}\}$ (or $\{\theta_{\textbf{Googling Yellowstone}}, \theta_{\textbf{coding}}\}$). We assume that these two modes do not differ until some $T - T' \in \mathbf{I}$. Between $T - T'$ and $T - T' - k$, the distributions become nearly disjoint, $P_{x\sim p^{\theta_{\textbf{harmful}}}_{T-T'-k}}\left(x \in \mathrm{supp}(p^{\theta_{\textbf{safe}}}_{T-T'-k})\right) \le \epsilon$. In the jailbreaking example, $T' = 0$ and they are disjoint because the first tokens generated in the safe mode is always some form of refusal. In the Yellowstone example, they are disjoint the first time the agent decides to Google Yellowstone pictures. Then, on component $\theta_{\textbf{harmful}}$ we have the critical window $T_{\text{before}} = T - T'$ and $T_{\text{after}} = T - T' - k$.

Notice that we can actually mitigate the effect of these critical windows by finetuning on examples of corrections to increase $P_{x\sim p^{\theta_{\textbf{harmful}}}_{T-T'-k}}\left(x \in \mathrm{supp}(p^{\theta_{\textbf{safe}}}_{T-T'-k})\right)$. This explains the effectiveness of finetuning on corrections in (Qi et al., 2024).

Remark 4.4. The quantity that measures probability of mode-switching, $p^{\theta_{\textbf{harmful}}}/p$, suggests using a likelihood ratio to distinguish between harmful and benign prompts. In App. G.1.2, we test a class of likelihood ratio methods that obtain recall 5-10× the false positive rate for 5 different types of jailbreaks (Table 2).

Example 4.5. [Math problem-solving as a random walk] We model solving a math problem as taking a random walk on $\mathbb{Z}$ with stepsize 1 of length T. If the random walk hits $+A$, then it has 'solved' the problem; if the random walk hits $-A$, then it has obtained an incorrect solution. Assume that we have two modes: a strong problem solving mode (denoted $+1$), which takes a $+1$ step with probability $0.5 + \Delta$, and a weak problem solving mode (denoted -1), which takes a $+1$ step with probability $0.5 - \Delta$. Assuming that $\frac{\ln(2/\epsilon^2)}{2\Delta^2} < A$ and $\epsilon^2 < 10^{-3}(0.5 - \Delta)(0.5 + \Delta)$, there is a critical window for the strong problem solving window of $T_{\text{before}} = T - \frac{\epsilon^2}{\Delta^2} + 2$ and $T_{\text{after}} = T - \frac{\ln(2/\epsilon^2)}{2\Delta^2}$. Note the critical window has width $\Theta(1/\Delta^2)$ independent of T.

We defer an example of a critical window for an autoregressive model which expresses the outputs as emissions from a random walk of an underlying concept variable, akin to the model in (Arora et al., 2019), to Appendix D.2.1.

4.2.1 IN-CONTEXT LEARNING

Autoregressive critical windows can also be applied to describe in-context learning. In particular, we can capture the idea that with sufficiently many in-context examples, we learn the $\theta^* \in \Theta$ that generated the transitions for in-context examples, with a sample complexity in terms of T_{after}.

Example 4.6 (Informal, see Example D.16). Consider an in-context learning setup, where the context $[x_1, y_1, o, \ldots, x_{T+1}, y_{T+1}, o]$ consists of question-answer pairs (x_i, y_i), delimiters o, and $x_i \to y_i$ sampled from p^{θ^*} for some $\theta^* \in \Theta$. In the forward-reverse experiment, we truncate it to $[x_1, y_1, o, \ldots, x_{T+1}]$, and then resample with p to produce $[x_1, \ldots, x_{T+1}, \tilde{y}_{T+1}, o]$. The total variation between the sequences $[x_1, y_1, o, \ldots, x_{T+1}]$ and $[x_1, \ldots, x_{T+1}, \tilde{y}_{T+1}, o]$ can be viewed as the average-case error of the in-context learner and can be bounded within our critical windows framework. We have $T_{\text{after}} = 3T + 3 - O_\epsilon(1)$, with $O_\epsilon(1)$ independent of T ($S_{\text{after}} \triangleq \{\theta^*\}$). Note that T_{after} is the order of how many samples that can be erased so that we still are able to learn $\theta^* \in \Theta$.

One might ask if there is a T_{before} for in-context learning, a threshold such that it is impossible to distinguish between $S_{\text{after}}, S_{\text{before}}$ with that many samples. In the next section, we will provide an example of a T_{before} with the all-or-nothing phase transition.

4.2.2 ALL-OR-NOTHING PHASE TRANSITION

Here we elucidate a formal connection between the critical windows phenomenon in in-context learning and the *all-or-nothing* phenomenon (see Appendix D.4 for details). To begin, we first define the notions of strong and weak detection:

Definition 4.7. Let (N_s) be an increasing sequence of integers. Given sequences of distributions $(p_s), (q_s)$ over $z \in \mathbb{R}^{N_s}$, a sequence of test statistics $(\mathcal{A}_s : \mathbb{R}^{N_s} \to \mathbb{R})$ with threshold (τ_s) achieves *strong (resp. weak) detection* if $\limsup_{s\to\infty}\{\Pr_{z\sim p_s}[\mathcal{A}_s(z) < \tau_s] + \Pr_{z\sim q_s}[\mathcal{A}_s(z) \geq \tau_s]\} = 0$ (resp. $\limsup_{s\to\infty}\{\Pr_{z\sim p_s}[\mathcal{A}_s(z) < \tau_s] + \Pr_{z\sim q_s}[\mathcal{A}_s(z) \geq \tau_s]\} < 1$).

Consider the following Bayesian inference problem, given by a joint distribution π over $(\theta, z) \in \mathbb{R}^n \times \mathbb{R}^m$. Nature samples unknown signal $\theta \in \mathbb{R}^n$ from π_θ. Given sample size N, we receive observations $\{z_i\}_{i=1}^N$ drawn i.i.d. from $\pi_{z|\theta}$; the goal is to infer θ from these observations. Let $\pi^{(N)}$ denote the distribution over $\{z_i\}_{i=1}^N$, the mixture of product measures parametrized by θ.

Definition 4.8. Let (π_s) be a sequence of inference tasks over $\mathbb{R}^{n_s} \times \mathbb{R}^{m_s}$ and (π^{null_s}) be a sequence of distributions over $\mathbb{R}^{m_s}$. (π_s) exhibits an *all-or-nothing phase transition at threshold* (N_s) *with respect to null models* (π^{null_s}) if: (1) For any $\beta < 1$: weak detection between $(\pi^{(\beta N_s)})$ and $((\pi^{\mathsf{null}_s})^{\otimes \beta N_s})$ is information-theoretically impossible and (2) For any $\beta > 1$: strong detection between the planted model $(\pi^{(\beta N_s)})$ and the null model $((\pi^{\mathsf{null}_s})^{\otimes \beta N_s})$ is information-theoretically possible.

All-or-nothing phase transitions have been established for a number of natural inference tasks like sparse linear regression (Reeves et al., 2019; Gamarnik & Zadik, 2019), sparse PCA (Niles-Weed & Zadik, 2020), generalized linear models (Barbier et al., 2020), group testing (Truong et al., 2021; Coja-Oghlan et al., 2022), linear and phase retrieval models (Scarlett & Cevher, 2016; Truong & Scarlett, 2020), planted subgraphs (Mossel et al., 2023), and planted Gaussian perceptron (Niles-Weed & Zadik, 2023).

Theorem 4.9 (Informal, see Theorem D.17 (Reeves et al., 2019))**.** *In an instance of sparse linear regression, if one receives at most $0.99N_s^*$ samples (x_i, y_i), then it is information-theoretically impossible to tell whether these samples were generated with respect to some ground truth signal θ, or whether the y_i's were generated completely independently from the x_i's; on the other hand, if one receives at least $1.01N_s^*$ samples, then one can perfectly distinguish between these two scenarios.*

We can instantiate the all-or-nothing phenomenon as a critical window for in-context learning.

Definition 4.10. To any inference task π, null model π^{null}, and sequence length N, we can associate the following in-context learning task. Let $\Theta = \Theta_{\mathsf{signal}} \sqcup \{\mathsf{NULL}\}$ where $\Theta_{\mathsf{signal}} \triangleq \mathrm{supp}(\pi_\theta)$. Given $\theta \in \mathrm{supp}(\pi_\theta)$, let $p^\theta_{(N)}$ denote the distribution over sequences $(z_1, \ldots, z_N, ?, \theta)$ where $z_1, \ldots, z_N$ are i.i.d. samples from $p_{z|\theta}$. Let $p^{\mathsf{null}}_{(N)}$ denote the distribution over observations $(z_1, \ldots, z_N, ?, \mathsf{NULL})$ where $z_1, \ldots, z_N$ are i.i.d. samples from π^{null}. We then take $p^\Theta_{(N)} \triangleq \mathbb{E}_{\theta\sim\frac{1}{2}\pi_\theta + \frac{1}{2}\delta_{\mathsf{NULL}}} p^\theta_{(N)}$.

Theorem 4.11 (Informal, see Theorem D.18)**.** *Consider a sequence of N in-context examples generated by some real signal θ and index s. Masking all but $(1-\delta)N_s^*$ in-context examples and resampling according to the full model p^{Θ_s} will yield sequence with an equal chance of having been generated by a signal or the null. If we mask all but $(1+\delta)N_s^*$ examples and resample, the distribution over the resulting sequence is the same as the original sequence. In other words, we have $T_{\text{before}} \triangleq N+2-(1-\delta)N_s^*$ and $T_{\text{after}} \triangleq N+2-(1+\delta)N_s^*$.*

5 HIERARCHIES IN STOCHASTIC LOCALIZATION SAMPLERS

Herein we propose a theory of hierarchical sampling within our critical windows framework.[5] It is motivated by the observation that a single trajectory can contain multiple critical windows (Figure 3), each splitting a sub-population into smaller sub-populations. This hierarchy is naturally represented as a tree: the root signifies that all sub-populations are indistinguishable under enough noise, while the leafs represent distinct modes in p. A path from the root to a leaf captures the progressive refinement of the original distribution p into increasingly specific components. To formalize this, we introduce the concept of an ϵ-mixture tree, which decomposes p into a hierarchical structure.

Definition 5.1 (Informal, see Definition E.1)**.** For an error term $\epsilon > 0$ and mixture model p, an **ϵ-mixture tree** is a tuple $(T, \{P^{\rightarrow}(\cdot|\cdot)\}, \mathbf{I}, \Theta, \{p^\theta\}_{\theta\in\Theta}, \text{Subset}, \text{NoiseAmount})$. $T=(V,E)$ is a tree associated with a function Subset: $V \rightarrow 2^{\Theta}\backslash\{\emptyset\}$, which maps vertices to subsets of Θ and NoiseAmount: $V \rightarrow \mathbb{R}^{\geq 0}$, which characterizes the noises levels for different aggregations in the mixture tree. If u is a parent of v, $\text{Subset}(v) \subset \text{Subset}(u)$ for $u \in V$; $\text{NoiseAmount}(u)$ is defined such that all $p^\theta_{\text{NoiseAmount}(u)}$ for $\theta \in \text{Subset}(u)$ overlap greatly and for $p^{\text{Subset}(u)}_{\text{NoiseAmount}(u)}, p^{\Theta-\text{Subset}(u)}_{\text{NoiseAmount}(u)}$ have negligible overlap.

We emphasize that this framework is highly general, solely defined with the *initial distribution* p and *the forward process*. It strictly expands the definition in (Li & Chen, 2024), which focused on hierarchies of isotropic Gaussians, to all localization-based samplers and mixture models. We can also relate it to the sequences of critical windows we observe in Figure 3, capturing the idea that each critical window represents the refinement into smaller subpopulations of p.

Corollary 5.2. *Consider an ϵ-mixture tree. For $\theta_i \in \Theta$, consider the path $u_1, u_2, u_3, \ldots, u_{H'} \in V$ where u_1 is the leaf node with $\theta_i \in$ Subset(u_1) and $u_{H'}$ is the root. There is a sequence of times $T_1 < T_2 < \cdots < T_{H'}$ with $\mathrm{TV}(p^{\{i\},T_\ell}, p^{\mathit{Subset}(u_\ell)}) \lesssim_w \epsilon$.*

We first observe that the hierarchy of two samplers with the same forward process are identical if the samplers agree on sub-populations. Assume we have $\{p^\theta\}_{\theta\in\Theta}$ (e.g. the true distribution) and $\{q^\theta\}_{\theta\in\Theta}$ (e.g. a generative model), where $q^\theta \approx p^\theta$ across all $\theta \in \Theta$ with the same $\{w_\theta\}_{\theta\in\Theta}$.

Corollary 5.3. *Consider an ϵ-mixture tree $(T, \{P^{\rightarrow}(\cdot|\cdot)\}, \mathbf{I}, \Theta, \{p^\theta\}_{\theta\in\Theta}, \mathit{Subset}, \mathit{NoiseAmount})$. Suppose we have another distribution $\{q^\theta\}_{\theta\in\Theta}$ such that $\mathrm{TV}(p^\theta, q^\theta) \leq \delta/2$ for all $\theta \in \Theta$. Then we have $\epsilon + \sqrt{\delta}$-mixture tree given by $(T, \{P^{\rightarrow}(\cdot|\cdot)\}, \mathbf{I}, \Theta, \{q^\theta\}_{\theta\in\Theta}, \mathit{Subset}, \mathit{NoiseAmount})$.*

However, we can define arbitrary hierarchies by choosing the right forward process.

Example 5.4. Consider a set of alphabets $\{\mathcal{A}_i\}_{i=1}^d$ and define $\Theta = \{(a_i)_{i=1}^d : \forall i \in [d], a_i \in \mathcal{A}_i\}$ and $p^{\theta_i} = \delta_{\theta_i}$. Let $\mathbf{I} = [0,1,2,\ldots,d]$. and for any permutation $i_1, i_2, \ldots, i_d$ of $[d]$, define a forward process such that at $t \in \mathbf{I}$, we mask all $i_d, i_{d-1}, \ldots, i_{d-t}$. This constructs a hierarchy where the values for $i_1, i_2, \ldots, i_d$ are decided in that order.

Finally, we note that hierarchies of diffusions are in general shallower than for autoregressive models. The hierarchy for a mixture of Gaussians has depth $O(1)$ (Example 4.2), as the forward process simultaneously contracts all distances with a similar dependence on d. For autoregressive models, depth can scale linearly with the context length (Example 5.4). We speculate that this could mean autoregressive models can learn more complex feature hierarchies than diffusions.

[5] Details in Appendix E.

Ethics statement

While this paper is largely theoretical in nature, it does describe a theory for jailbreaks which could impact model safety in the future. We hope the insights in this manuscript about jailbreaks lead to better alignment strategies and training methods.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report, 2024. URL https://arxiv.org/abs/2412.08905.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models, 2023. URL https://arxiv.org/abs/2211.15661.

Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Sampling from mean-field gibbs measures via diffusion processes. *arXiv preprint arXiv:2310.08912*, 2023.

Gabriel Alon and Michael Kamfonas. Detecting language model attacks with perplexity, 2023. URL https://arxiv.org/abs/2308.14132.

Cem Anil, Esin Durmus, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Nina Rimsky, Meg Tong, Jesse Mu, Daniel Ford, Francesco Mosconi, Rajashree Agrawal, Rylan Schaeffer, Naomi Bashkansky, Samuel Svenningsen, Mike Lambert, Ansh Radhakrishnan, Carson Denison, Evan J Hubinger, Yuntao Bai, Trenton Bricken, Timothy Maxwell, Nicholas Schiefer, Jamie Sully, Alex Tamkin, Tamera Lanham, Karina Nguyen, Tomasz Korbak, Jared Kaplan, Deep Ganguli, Samuel R. Bowman, Ethan Perez, Roger Grosse, and David Duvenaud. Many-shot jailbreaking. In *Advances in the Thirty-Eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://www-cdn.anthropic.com/af5633c94ed2beb282f6a53c595eb437e8e7b630/Many_Shot_Jailbreaking__2024_04_02_0936.pdf.

Anthropic. Developing a computer use model. https://www.anthropic.com/news/developing-computer-use, 2024.

Aryaman Arora, Dan Jurafsky, Christopher Potts, and Noah D. Goodman. Bayesian scaling laws for in-context learning, 2024. URL https://arxiv.org/abs/2410.16531.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A latent variable model approach to pmi-based word embeddings, 2019. URL https://arxiv.org/abs/1502.03520.

Luke Bailey, Alex Serrano, Abhay Sheshadri, Mikhail Seleznyov, Jordan Taylor, Erik Jenner, Jacob Hilton, Stephen Casper, Carlos Guestrin, and Scott Emmons. Obfuscated activations bypass llm latent-space defenses, 2024. URL https://arxiv.org/abs/2412.09565.

Jean Barbier, Nicolas Macris, and Cynthia Rush. All-or-nothing statistical and computational phase transitions in sparse spiked matrix estimation, 2020. URL https://arxiv.org/abs/2006.07971.

Giulio Biroli, Tony Bonnaire, Valentin de Bortoli, and Marc Mézard. Dynamical regimes of diffusion models, 2024.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries, 2023.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=zyLVMgsZ0U_.

Yuansi Chen and Ronen Eldan. Localization schemes: A framework for proving mixing bounds for markov chains. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 110–122. IEEE, 2022.

Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11462–11471, 2022. doi: 10.1109/CVPR52688.2022.01118.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. arXiv preprint 1803.05457, 2018.

Amin Coja-Oghlan, Oliver Gebhard, Max Hahn-Klimroth, Alexander S Wein, and Ilias Zadik. Statistical and computational phase transitions in group testing. In Po-Ling Loh and Maxim Raginsky (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 4764–4781. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/coja-oghlan22a.html.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations, 2023. URL https://arxiv.org/abs/2305.14233.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning, 2024. URL https://arxiv.org/abs/2301.00234.

Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Sampling from the sherrington-kirkpatrick gibbs measure via algorithmic stochastic localization. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 323–334. IEEE, 2022.

Ronen Eldan. Thin shell implies spectral gap up to polylog via a stochastic localization scheme. *Geometric and Functional Analysis*, 23(2):532–569, 2013.

Ronen Eldan. Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation. *Probability Theory and Related Fields*, 176(3-4):737–755, 2020.

David Gamarnik and Ilias Zadik. High-dimensional regression with binary coefficients. estimating squared error and a phase transition, 2019. URL https://arxiv.org/abs/1701.04455.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes, 2023. URL https://arxiv.org/abs/2208.01066.

Gemini. Challenges and solutions for aging adults. https://gemini.google.com/share/6d141b742a13, 2024.

Kristian Georgiev, Joshua Vendrow, Hadi Salman, Sung Min Park, and Aleksander Madry. The journey, not the destination: How data guides diffusion models. *arXiv preprint arXiv:2312.06205*, 2023.

grimjim. Jailbroken llama-3.1-8b-instruct via lora. https://huggingface.co/grimjim/Llama-3.1-8B-Instruct-abliterated_via_adapter, 2024.

Haize Labs. Automated multi-turn red-teaming with cascade. https://blog.haizelabs.com/posts/cascade, 2024a.

Haize Labs. A trivial jailbreak against llama 3. https://github.com/haizelabs/llama3-jailbreak, 2024b.

Luxi He, Mengzhou Xia, and Peter Henderson. What is in your safe data? identifying benign data that breaks safety, 2024. URL https://arxiv.org/abs/2404.01099.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021b. URL https://arxiv.org/abs/2103.03874.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. URL https://arxiv.org/abs/2006.11239.

Peter Holderrieth, Marton Havasi, Jason Yim, Neta Shaul, Itai Gat, Tommi Jaakkola, Brian Karrer, Ricky T. Q. Chen, and Yaron Lipman. Generator matching: Generative modeling with arbitrary markov processes, 2025. URL https://arxiv.org/abs/2410.20587.

Brice Huang, Andrea Montanari, and Huy Tuan Pham. Sampling from spherical spin glasses in total variation via algorithmic stochastic localization. *arXiv preprint arXiv:2404.15651*, 2024.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning, 2023. URL https://arxiv.org/abs/2307.13702.

Marvin Li and Sitan Chen. Critical windows: non-asymptotic theory for feature emergence in diffusion models, 2024. URL https://arxiv.org/abs/2403.01633.

Nathaniel Li, Ziwen Han, Ian Steneker, Willow Primack, Riley Goodside, Hugh Zhang, Zifan Wang, Cristina Menghini, and Summer Yue. Llm defenses are not robust to multi-turn human jailbreaks yet, 2024. URL https://arxiv.org/abs/2408.15221.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. The unlocking spell on base llms: Rethinking alignment via in-context learning, 2023. URL https://arxiv.org/abs/2312.01552.

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

Zicheng Lin, Tian Liang, Jiahao Xu, Xing Wang, Ruilin Luo, Chufan Shi, Siheng Li, Yujiu Yang, and Zhaopeng Tu. Critical tokens matter: Token-level contrastive estimation enhances llm's reasoning capability, 2024. URL https://arxiv.org/abs/2411.19943.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 158–167, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1015. URL https://aclanthology.org/P17-1015.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. In Christian Bessiere (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 3622–3628. International Joint Conferences on Artificial Intelligence Organization, 7 2020. doi: 10.24963/ijcai.2020/501. URL https://doi.org/10.24963/ijcai.2020/501. Main track.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7Jwpw4qKkb.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution, 2024. URL https://arxiv.org/abs/2310.16834.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

Andrea Montanari. Sampling, diffusions, and stochastic localization. *arXiv preprint arXiv:2305.10690*, 2023a.

Andrea Montanari. Sampling, diffusions, and stochastic localization, 2023b. URL https://arxiv.org/abs/2305.10690.

Andrea Montanari and Yuchen Wu. Posterior sampling from the spiked models via diffusion processes. *arXiv preprint arXiv:2304.11449*, 2023.

Elchanan Mossel, Jonathan Niles-Weed, Youngtak Sohn, Nike Sun, and Ilias Zadik. Sharp thresholds in inference of planted subgraphs. In Gergely Neu and Lorenzo Rosasco (eds.), *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 5573–5577. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/mossel23a.html.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models, 2023. URL https://arxiv.org/abs/2311.17035.

Jonathan Niles-Weed and Ilias Zadik. The all-or-nothing phenomenon in sparse tensor pca. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17674–17684. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/cd0b43eac0392accf3624b7372dec36e-Paper.pdf.

Jonathan Niles-Weed and Ilias Zadik. It was ”all” for ”nothing”: sharp phase transitions for noiseless discrete channels, 2023. URL https://arxiv.org/abs/2102.12422.

Leonardo Petrini, Francesco Cagnetta, Umberto M Tomasini, Alessandro Favero, and Matthieu Wyart. How deep neural networks learn compositional data: The random hierarchy model. *arXiv preprint arXiv:2307.02129*, 2023.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep, 2024. URL https://arxiv.org/abs/2406.05946.

Gabriel Raya and Luca Ambrogioni. Spontaneous symmetry breaking in generative diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=lxGFGMMSVl.

Galen Reeves, Jiaming Xu, and Ilias Zadik. The all-or-nothing phenomenon in sparse linear regression, 2019. URL https://arxiv.org/abs/1903.05046.

B. Roos. Binomial approximation to the poisson binomial distribution: The krawtchouk expansion. *Theory of Probability & Its Applications*, 45(2):258–272, 2001. doi: 10.1137/S0040585X9797821X. URL https://doi.org/10.1137/S0040585X9797821X.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models, 2024. URL https://arxiv.org/abs/2308.01263.

Jonathan Scarlett and Volkan Cevher. Limits on support recovery with probabilistic models: An information-theoretic framework, 2016. URL https://arxiv.org/abs/1501.07440.

Antonio Sclocchi, Alessandro Favero, Noam Itzhak Levi, and Matthieu Wyart. Probing the latent hierarchical structure of data via diffusion models, 2024. URL https://arxiv.org/abs/2410.13770.

Antonio Sclocchi, Alessandro Favero, and Matthieu Wyart. A phase transition in diffusion models reveals the hierarchical nature of data. *Proceedings of the National Academy of Sciences*, 122(1): e2408799121, 2025.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A strongreject for empty jailbreaks, 2024. URL https://arxiv.org/abs/2402.10260.

Lan V. Truong and Jonathan Scarlett. Support recovery in the phase retrieval model: Information-theoretic fundamental limits, 2020. URL https://arxiv.org/abs/1901.10647.

Lan V. Truong, Matthew Aldridge, and Jonathan Scarlett. On the all-or-nothing behavior of bernoulli group testing, 2021. URL https://arxiv.org/abs/2001.10137.

Ramon van Handel. Probability in high dimension. 2016. URL https://web.math.princeton.edu/~rvan/APC550.pdf.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. ISBN 978-1-108-41519-4. doi: 10.1017/9781108231596. URL https://www.cambridge.org/core/books/highdimensional-probability/797C466DA29743D2C8213493BD2D2102.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail?, 2023. URL https://arxiv.org/abs/2307.02483.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference, 2022. URL https://arxiv.org/abs/2111.02080.

Xiao Zhang and Ji Wu. Dissecting learning and forgetting in language model finetuning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tmsqb6WpLz.

Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization, 2023. URL https://arxiv.org/abs/2305.19420.

Zhengyue Zhao, Xiaoyun Zhang, Kaidi Xu, Xing Hu, Rui Zhang, Zidong Du, Qi Guo, and Yunji Chen. Adversarial contrastive decoding: Boosting safety alignment of large language models via opposite prompt optimization, 2024. URL https://arxiv.org/abs/2406.16743.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023. URL https://arxiv.org/abs/2307.15043.

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers, 2024. URL https://arxiv.org/abs/2406.04313.

A Related works

Theory of critical windows in diffusion. Several recent works have studied critical windows in the context of diffusion models, using either statistical physics methods (Raya & Ambrogioni, 2023; Sclocchi et al., 2024; 2025; Biroli et al., 2024) or Girsanov's theorem (Li & Chen, 2024). The statistical physics papers assume an explicit functional form for the data and use accurate and non-rigorous statistical physics methods to compute critical windows. For instance, Biroli et al. (2024)

computes the critical time at which the reverse process specializes to one component for a mixture of two spherical Gaussians using a Landau-type perturbative calculation, and Sclocchi et al. (2025; 2024) passed through a mean-field approximation to compute the critical windows for a *random hierarchy model* (Petrini et al., 2023), a multi-level context-free grammar with random production rules. Our work is most similar to (Li & Chen, 2024), which derives rigorous, non-asymptotic bounds analogous to our Theorem 3.1 for mixtures of log-concave distributions with Girsanov's theorem (Chen et al., 2023).

In contrast to existing work, our theory applies to all localization-based samplers, including diffusion and autoregressive language models, and imposes no functional form or log-concavity assumptions on the distribution. We also improve upon the main theorem of (Li & Chen, 2024) by obtaining dimension-independent error bounds. Using our improved theorem, we can extend the definition of hierarchy of critical windows from Li & Chen (2024) to all localization-based samplers and, for continuous diffusions, to distributions beyond mixtures of Gaussians.

Forward-reverse experiment. Here we study the forward-reverse experiment, also explored in (Li & Chen, 2024; Sclocchi et al., 2025; 2024), to study critical windows. This is very similar to the framework studied in Georgiev et al. (2023); Biroli et al. (2024); Raya & Ambrogioni (2023) in which one imagines re-running the reverse process at an intermediate point Y_t. Both perspectives provide rigorous frameworks to understand critical windows, and in the case where the forward process is deterministic, i.e. autoregressive language models, these frameworks are equivalent.

Stochastic localization. El Alaoui et al. (2022); Montanari & Wu (2023); Alaoui et al. (2023); Montanari (2023a); Huang et al. (2024) applied Eldan's stochastic localization method (Eldan, 2013; 2020) to develop new sampling algorithms for certain distributions inspired by models in statistical physics. Our work uses the stochastic localization framework (Montanari, 2023b; Holderrieth et al., 2025) to understand an empirical phenomenon across different localization-based samplers widely used in practice. Note that our formulation of stochastic localization differs slightly in several minor ways. First, in that work the index set $\mathbf{I}$ is not necessarily compact; while we assume compactness of $\mathbf{I}$, this still encapsulates most applications of generative models, in which the sample X is realized in finitely many steps. Secondly, our indexing of time is the reverse of that of in (Montanari, 2023b); in that work, the Y_t's become *more* informative about X as t increases. We make this choice purely for cosmetic reasons.

Chain of thought. (Lin et al., 2024; Abdin et al., 2024) also observed that the presence of critical windows in the chain of thought of math and reasoning tasks and their significance in leading the model to incorrect outputs, concurrent with our results in Figure 5. They then used them to provide rewards or data for a preference optimization algorithm to improve reasoning performance. (Lin et al., 2024) called them *critical tokens* and utilized a contrastive estimation algorithm to identify critical windows and provide token-level rewards. The Phi-4 Technical report called them *pivotal tokens*, developed a binary-search based algorithm to identify the location of critical windows, and used them to produce contrasting pairs for preference optimization (Abdin et al., 2024). Using our broad theoretical perspective, we provide new insight into critical windows of these kinds and view our work as corroborating and extending these empirical works.

Jailbreaks. Existing work on jailbreaks has studied the appearance of critical windows in the first few generated tokens (Qi et al., 2024; Zhang & Wu, 2024; He et al., 2024; Lin et al., 2023). Our theory provides a simple explanation for when jailbreaks occur: when the unaligned component assigns a much higher probability to the current text than the aligned component, then the model is jailbroken. This generalizes the explanation from Qi et al. (2024) (see Example 4.3 for our particular formalism of their insights). It also explains the success of perplexity-based monitors for jailbreaks (Alon & Kamfonas, 2023), which monitor for a low probability of the context and generation. We view our work as providing a rigorous mathematical framework for jailbreaks, as well as highlighting the important role off-distribution contexts play in eliciting harmful behaviors; we also develop a novel jailbreak from our framework (Section G.1.2) similar to the adversarial contrast decoding method proposed by (Zhao et al., 2024), which also uses a likelihood ratio between an unaligned and an aligned model. However, we use a jailbroken and non-jailbroken pair of models instead of two versions of the model with different prompts.

Theory of in-context learning for language models. With respect to theory for language models, our results are most closely related to the Bayesian framework for in-context learning (Xie et al., 2022; Akyürek et al., 2023; Garg et al., 2023; Zhang et al., 2023; Arora et al., 2024). For example, (Xie et al., 2022) also considered a mixture model of topics and showed that language models can learn the underlying class despite in-context learning and training distribution mismatch. We view this manuscript as connecting the Bayesian framework for in-context learning to other empirical phenomena observed in language models and diffusion models and the all-or-nothing phenomenon.

B EXAMPLES OF STOCHASTIC LOCALIZATION

In this section, we present several kinds of generative models within the stochastic localization framework and their forward and reverse processes.

Example B.1 (Continuous Diffusion Models (Li & Chen, 2024)). For continuous diffusion models, the forward process progressively degrades samples $X \sim p$ into pure Gaussian noise through scaling and convolution with Gaussian noise. It is the Ornstein-Uhlenbeck process, a stochastic process $(X_t)_{t\geq 0}$ given by the following stochastic differential equation (SDE),

$$\mathrm{d}X_t = -X_t\,\mathrm{d}t + \sqrt{2}\mathrm{d}B_t\,, \qquad X_0 \sim p\,,$$

where $(B_t)_{t\geq 0}$ is a standard Brownian motion. Let $q_t \triangleq \mathrm{law}(X_t)$ for $t \geq 0$, and observe that as $t \to \infty$, q_t converges exponentially quickly to the standard Gaussian distribution $\mathcal{N}(0, \mathrm{Id})$. Assume we end the forward process at time $T \geq 0$. For the reverse process $(X_t^{\leftarrow})_{t\in[0,T]}$, we employ the *reversal* of the Ornstein-Uhlenbeck SDE, given by

$$\mathrm{d}X_t^{\leftarrow} = \{X_t^{\leftarrow} + 2\nabla \ln q_{T-t}(X_t^{\leftarrow})\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,, \qquad X_T^{\leftarrow} \sim q_T$$

where here $(B_t)_{t\geq 0}$ is also a Brownian motion. Defining $(Y_t)_{t\in\mathbf{I}} = (X_t)_{t\in\mathbf{I}}$, we see that the forward process satisfies the Markov property in Definition 2.2, and the information from the original sample X_0 is degraded by more steps in the SDE. Furthermore, the reverse SDE with parameterized by the score function $\nabla \ln q_{T-t}(X_t^{\leftarrow})$ can be viewed as successively sampling from the posteriors via Tweedie's formula.

Example B.2 (Discrete Diffusion Models (Lou et al., 2024)). Consider a set $\mathcal{A}$ denoting the vocabulary and let $p \in \mathcal{A}^T$, and consider a forward process with index set $\mathbf{I} = [0, K] \cup \{\infty\}$, $Y_0 = X$, and $Y_t \in \mathcal{A}^T$ defined in the limit as follows,

$$p(Y_{t+\Delta t} = a \mid Y_t = b) = \delta_{ab} + Q_t(b, a)\Delta t + O(\Delta t^2),$$

where $Q_t \in \mathbb{R}^{n\times n}$ are diffusion matrices with nonnegative non-diagonal entries and columns which sum to 0. $(Y_t)_{t\in\mathbf{I}}$ is also a Markov chain and as $t \to \infty$, Y_t is degraded until it is eventually uninformative about the original sample Y_0.

Example B.3 (Autoregressive Language Models). Consider a set $\mathcal{A}$ denoting the vocabulary and let $p \in \mathcal{A}^T$, and consider a forward process with index set $\mathbf{I} = \{0, 1, 2, \ldots, T\}$, $Y_0 = X$, and $Y_t \in \mathcal{A}^{T-t}$. For $t \in \mathbf{I}$, we let Y_t equal the last first $T - t$ tokens of X. Clearly this is a Markov Chain, and the reverse process is equivalent to next-token prediction.

C DEFERRED DETAILS FROM SECTION 3

Remark C.1 (Technicality in defining $T_{\mathrm{st}}, T_{\mathrm{end}}$). For general stochastic localization schemes, we can only ask that $T_{\mathrm{st}}(\epsilon) \in \{r \in \mathbf{I} : \mathrm{TV}(p_t^{S_{\mathrm{targ}}}, p_t^{\Theta - S_{\mathrm{targ}}}) \geq 1 - \epsilon^2\}$ and $T_{\mathrm{end}}(\epsilon) \in \{t \in \mathbf{I} : \mathrm{TV}(p_t^{S_{\mathrm{init}}}, p_t^{S_{\mathrm{targ}}}) \leq \epsilon\}$ instead of sup, inf like (Li & Chen, 2024), because the sets $\{t \in \mathbf{I} : \mathrm{TV}(p_t^{S_{\mathrm{targ}}}, p_t^{\Theta - S_{\mathrm{targ}}}) \geq 1 - \epsilon^2\}, \{t \in \mathbf{I} : \mathrm{TV}(p_t^{S_{\mathrm{init}}}, p_t^{S_{\mathrm{targ}}}) \leq \epsilon\}$ may not be closed for observation processes which are discontinuous. For autoregressive language models and continuous diffusion, the observation process is continuous, so we will elide these technicalities.

Theorem 3.1. *Let* $S_{\mathrm{init}} \subset S_{\mathrm{targ}} \subset \Theta$ *and* $W = \frac{\sum_{\theta\in\Theta - S_{\mathrm{targ}}} w_\theta}{\sum_{\theta\in S_{\mathrm{targ}}} w_\theta}$. *For* $\epsilon > 0$, *if* $\widehat{T} \in \mathbf{I} \cap [T_{\mathrm{end}}, T_{\mathrm{st}}]$, *then* $\mathrm{TV}(p^{S_{\mathrm{init}}, \widehat{T}}, p^{S_{\mathrm{targ}}}) \leq \epsilon \cdot \left(1 + \max(1, W)/\sqrt{2}\right)$.

For nonempty $S \subset \Theta$ and $t \in \mathbf{I}$, we define $P^{\leftarrow}(\cdot|Y_t, S)$ to be the posterior of X with the prior $X \sim p^S$. We similarly define $P^{\leftarrow}_{t\to\Theta}(\cdot|Y_t)$ and $P^{\leftarrow}_{t\to\Theta}(\cdot|Y_t, S)$ to be the posterior of θ conditioning on Y_t with $X \sim p$ or $X \sim p^S$, respectively. When $S = \{i\}$, we exclude the braces.

Proof of Theorem 3.1. By the triangle inequality, we can write $\mathrm{TV}(p^{S_{\mathrm{init}},\widehat{T}}, p^{S_{\mathrm{targ}}}) \leq \underbrace{\mathrm{TV}(p^{S_{\mathrm{init}},\widehat{T}}, p^{S_{\mathrm{targ}},\widehat{T}})}_{\text{(I)}} + \underbrace{\mathrm{TV}(p^{S_{\mathrm{targ}},\widehat{T}}, p^{S_{\mathrm{targ}}})}_{\text{(II)}}$. $p^{S_{\mathrm{init}},\widehat{T}}$ and $p^{S_{\mathrm{targ}},\widehat{T}}$ are the laws of the posterior $P^{\leftarrow}(\cdot|\cdot)$ but applied to $Y_{\widehat{T}}$ with distributions $p^{S_{\mathrm{init}}}_{\widehat{T}}$ and $p^{S_{\mathrm{targ}}}_{\widehat{T}}$. Using the Markov property of localization-based samplers (Definition 2.2), we apply the data processing inequality and the definition of T_{end} to bound (I) via $\mathrm{TV}(p^{S_{\mathrm{init}},\widehat{T}}, p^{S_{\mathrm{targ}},\widehat{T}}) \leq \mathrm{TV}(p^{S_{\mathrm{init}}}_{T_{\mathrm{end}}}, p^{S_{\mathrm{targ}}}_{T_{\mathrm{end}}}) \leq \epsilon$.

To bound (II), we apply the definition of TV and a coupling argument. By the law of total probability we can express $p^{S_{\mathrm{targ}},\widehat{T}}(x) = \mathbb{E}[P^{\leftarrow}(x|Y_{\widehat{T}})]$ and $p^{S_{\mathrm{targ}}}(x) = \mathbb{E}[P^{\leftarrow}(x|Y_{\widehat{T}}, S_{\mathrm{targ}})]$ for $Y_{\widehat{T}} \sim p^{S_{\mathrm{targ}}}_{\widehat{T}}$, as these observation processes have the same distribution at index $\widehat{T}$. Thus, $\mathrm{TV}(p^{S_{\mathrm{targ}},\widehat{T}}, p^{S_{\mathrm{targ}}}) = \frac{1}{2}\int \left|p^{S_{\mathrm{targ}},\widehat{T}}(x) - p^{S_{\mathrm{targ}}}(x)\right| dx = \frac{1}{2}\int \left|\mathbb{E}[P^{\leftarrow}(x|Y_{\widehat{T}})] - \mathbb{E}[P^{\leftarrow}(x|Y_{\widehat{T}}, S_{\mathrm{targ}})]\right| dx$. By Jensen's inequality and Fubini's theorem, we bring the expectation outside the integral, $\mathrm{TV}(p^{S_{\mathrm{targ}},\widehat{T}}, p^{S_{\mathrm{targ}}}) \leq \frac{1}{2}\int \mathbb{E}\left[\left|P^{\leftarrow}(x|Y_{\widehat{T}}) - P^{\leftarrow}(x|Y_{\widehat{T}}, S_{\mathrm{targ}})\right|\right] dx = \frac{1}{2}\mathbb{E}\left[\int \left|P^{\leftarrow}(x|Y_{\widehat{T}}) - P^{\leftarrow}(x|Y_{\widehat{T}}, S_{\mathrm{targ}})\right| dx\right]$.

Combining Lemmas C.2 and C.3 (proved below), we find $\mathrm{TV}(p^{S_{\mathrm{targ}},\widehat{T}}, p^{S_{\mathrm{targ}}}) \leq \max(1, W)\,\mathbb{E}\left[\frac{p^{\Theta - S_{\mathrm{targ}}}_{\widehat{T}}(Y_{\widehat{T}})}{p^{\Theta - S_{\mathrm{targ}}}_{\widehat{T}}(Y_{\widehat{T}}) + p^{S_{\mathrm{targ}}}_{\widehat{T}}(Y_{\widehat{T}})}\right]$. Then, finally applying Lemma 2.1, we are able to bound the total variation in terms of ϵ, obtaining $\mathrm{TV}(p^{S_{\mathrm{targ}},\widehat{T}}, p^{S_{\mathrm{targ}}}) \leq \frac{1}{2}\max(1, W)\sqrt{1 - \mathrm{TV}^2(p^{\Theta - S_{\mathrm{targ}}}_{\widehat{T}}, p^{S_{\mathrm{targ}}}_{\widehat{T}})} \leq \frac{\sqrt{2}}{2}\max(1, W)\,\epsilon$. Combining our bounds on (I) and (II) achieves the desired result. □

For Theorem 3.1, we employ the following two helper lemmas.

Lemma C.2. *By applying the law of total probability and Bayes' rule, we can show for $Y_{\widehat{T}} \in \mathrm{supp}(p^{S_{\mathrm{targ}}}_{\widehat{T}})$,*

$$\int \left|P^{\leftarrow}(x|Y_{\widehat{T}}) - P^{\leftarrow}(x|Y_{\widehat{T}}, S_{\mathrm{targ}})\right| dx \leq 2 \sum_{\theta \in \Theta - S_{\mathrm{targ}}} P^{\leftarrow}_{t\to\Theta}(\theta|Y_{\widehat{T}}).$$

Proof. We can rewrite $P^{\leftarrow}(x|Y_{\widehat{T}}), P^{\leftarrow}(x|Y_{\widehat{T}}, S_{\mathrm{targ}})$ using the law of total probability and Bayes' rule.

$$P^{\leftarrow}(x|Y_{\widehat{T}}) = \sum_{\theta\in\Theta} P^{\leftarrow}_{t\to\Theta}(\theta|Y_{\widehat{T}})P^{\leftarrow}(x|Y_{\widehat{T}}, \theta)$$

$$P^{\leftarrow}(x|Y_{\widehat{T}}, S_{\mathrm{targ}}) = \sum_{\theta\in S_{\mathrm{targ}}} P^{\leftarrow}_{t\to\Theta}(\theta|Y_{\widehat{T}}, S_{\mathrm{targ}})P^{\leftarrow}(x|Y_{\widehat{T}}, \theta) = \frac{\sum_{\theta\in S_{\mathrm{targ}}} P^{\leftarrow}_{t\to\Theta}(\theta|Y_{\widehat{T}})P^{\leftarrow}(x|Y_{\widehat{T}}, \theta)}{\sum_{\theta\in S_{\mathrm{targ}}} P^{\leftarrow}_{t\to\Theta}(\theta|Y_{\widehat{T}})}.$$

Note that the second equality on the second line follows from the fact that for all $\theta \in S_{\mathrm{targ}}$, the posteriors $P^{\leftarrow}_{t\to\Theta}(\cdot|Y_t) \propto P^{\leftarrow}_{t\to\Theta}(\cdot|Y_t, S_{\mathrm{targ}})$ by the same normalization constant. Therefore the difference can be written as

$$\begin{aligned}
&\int \left|P^{\leftarrow}(x|Y_{\widehat{T}}) - P^{\leftarrow}(x|Y_{\widehat{T}}, S_{\mathrm{targ}})\right| dx \\
&= \int \left|\sum_{\theta\in\Theta} P^{\leftarrow}_{t\to\Theta}(\theta|Y_{\widehat{T}})P^{\leftarrow}(x|Y_{\widehat{T}}, \theta) - \frac{\sum_{\theta\in S_{\mathrm{targ}}} P^{\leftarrow}_{t\to\Theta}(\theta|Y_{\widehat{T}})P^{\leftarrow}(x|Y_{\widehat{T}}, \theta)}{\sum_{\theta\in S_{\mathrm{targ}}} P^{\leftarrow}_{t\to\Theta}(\theta|Y_{\widehat{T}})}\right| dx \\
&= \int \left|\left(1 - \frac{1}{\sum_{\theta\in S_{\mathrm{targ}}} P^{\leftarrow}_{t\to\Theta}(\theta|Y_{\widehat{T}})}\right)\sum_{\theta\in S_{\mathrm{targ}}} P^{\leftarrow}_{t\to\Theta}(\theta|Y_{\widehat{T}})P^{\leftarrow}(x|Y_{\widehat{T}}, \theta) + \sum_{\theta\in\Theta - S_{\mathrm{targ}}} P^{\leftarrow}_{t\to\Theta}(\theta|Y_{\widehat{T}})P^{\leftarrow}(x|Y_{\widehat{T}}, \theta)\right| dx.
\end{aligned}$$

If $\sum_{\theta\in\Theta-S_{\text{targ}}} P^{\leftarrow}_{t\to\Theta}(\theta|Y_{\widehat{T}}) = 0$, then the above is equal to 0 and we are done. If it is non-zero, we can factor out $\sum_{\theta\in\Theta-S_{\text{targ}}} P^{\leftarrow}_{t\to\Theta}(\theta|Y_{\widehat{T}})$ term, which allows us to write everything in terms of posteriors with respect to $\Theta - S_{\text{targ}}$ and S_{targ},

$$\int \left| \left(1 - \frac{1}{\sum_{\theta\in S_{\text{targ}}} P^{\leftarrow}_{t\to\Theta}(\theta|Y_{\widehat{T}})}\right) \sum_{\theta\in S_{\text{targ}}} P^{\leftarrow}_{t\to\Theta}(\theta|Y_{\widehat{T}}) P^{\leftarrow}(x|Y_{\widehat{T}},\theta) + \sum_{\theta\in\Theta-S_{\text{targ}}} P^{\leftarrow}_{t\to\Theta}(\theta|Y_{\widehat{T}}) P^{\leftarrow}(x|Y_{\widehat{T}},\theta) \right| dx$$
$$= \sum_{\theta\in\Theta-S_{\text{targ}}} P^{\leftarrow}_{t\to\Theta}(\theta|Y_{\widehat{T}}) \int \left| P^{\leftarrow}(x|Y_{\widehat{T}}, S_{\text{targ}}) - P^{\leftarrow}(x|Y_{\widehat{T}}, \Theta - S_{\text{targ}}) \right| dx$$

Employing the trivial observation that

$$\begin{aligned}
&\int \left| P^{\leftarrow}(x|Y_{\widehat{T}}, S_{\text{targ}}) - P^{\leftarrow}(x|Y_{\widehat{T}}, \Theta - S_{\text{targ}}) \right| dx \\
&\le \int P^{\leftarrow}(x|Y_{\widehat{T}}, S_{\text{targ}}) + P^{\leftarrow}(x|Y_{\widehat{T}}, \Theta - S_{\text{targ}}) dx \le 2,
\end{aligned}$$

we have

$$\int \left| P^{\leftarrow}(x|Y_{\widehat{T}}) - P^{\leftarrow}(x|Y_{\widehat{T}}, S_{\text{targ}}) \right| dx \le 2 \sum_{\theta\in\Theta-S_{\text{targ}}} P^{\leftarrow}_{t\to\Theta}(\theta|Y_{\widehat{T}}).$$

$\square$

Lemma C.3. *By Bayes's rule, we can derive for $Y_{\widehat{T}} \in \operatorname{supp}(p_{\widehat{T}})$,*

$$\sum_{\theta\in\Theta-S_{\text{targ}}} P^{\leftarrow}_{t\to\Theta}(\theta|Y_{\widehat{T}}) \le \max(1, W) \frac{p^{\Theta-S_{\text{targ}}}_{\widehat{T}}(Y_{\widehat{T}})}{p^{\Theta-S_{\text{targ}}}_{\widehat{T}}(Y_{\widehat{T}}) + p^{S_{\text{targ}}}_{\widehat{T}}(Y_{\widehat{T}})}$$

Proof. We obtain through Bayes' rule,

$$\sum_{\theta\in\Theta-S_{\text{targ}}} P^{\leftarrow}_{t\to\Theta}(\theta|Y_{\widehat{T}}) = \frac{\sum_{\theta\in\Theta-S_{\text{targ}}} w_\theta p^{\theta}_{\widehat{T}}(Y_{\widehat{T}})}{\sum_{\theta\in\Theta} w_\theta p^{\theta}_{\widehat{T}}(Y_{\widehat{T}})}.$$

We divide by the same normalizing constant $\sum_{\theta\in\Theta-S_{\text{targ}}} w_\theta$ to obtain

$$\begin{aligned}
\frac{\sum_{\theta\in\Theta-S_{\text{targ}}} w_\theta p^{\theta}_{\widehat{T}}(Y_{\widehat{T}})}{\sum_{\theta\in\Theta} w_\theta p^{\theta}_{\widehat{T}}(Y_{\widehat{T}})} &= \frac{\frac{\sum_{\theta\in\Theta-S_{\text{targ}}} w_\theta p^{\theta}_{\widehat{T}}(Y_{\widehat{T}})}{\sum_{\theta\in\Theta-S_{\text{targ}}} w_\theta}}{\frac{\sum_{\theta\in\Theta-S_{\text{targ}}} w_\theta p^{\theta}_{\widehat{T}}(Y_{\widehat{T}})}{\sum_{\theta\in\Theta-S_{\text{targ}}} w_\theta} + \frac{\sum_{\theta\in S_{\text{targ}}} w_\theta p^{\theta}_{\widehat{T}}(Y_{\widehat{T}})}{\sum_{\theta\in S_{\text{targ}}} w_\theta} \cdot \frac{\sum_{\theta\in S_{\text{targ}}} w_\theta}{\sum_{\theta\in\Theta-S_{\text{targ}}} w_\theta}} \\
&\le \max\left(1, \frac{\sum_{\theta\in\Theta-S_{\text{targ}}} w_\theta}{\sum_{\theta\in S_{\text{targ}}} w_\theta}\right) \frac{p^{\Theta-S_{\text{targ}}}_{\widehat{T}}(Y_{\widehat{T}})}{p^{\Theta-S_{\text{targ}}}_{\widehat{T}}(Y_{\widehat{T}}) + p^{S_{\text{targ}}}_{\widehat{T}}(Y_{\widehat{T}})}.
\end{aligned}$$

$\square$

D Omitted details from Section 4

D.1 Diffusions

Here we use an alternative f-divergence to characterize the critical windows, the squared *Hellinger distance*, defined as $\mathrm{H}^2(P,Q) \triangleq \int (\sqrt{dP} - \sqrt{dQ})^2 d\mu$, because there are explicit computations for the Hellinger distance for mixtures of Gaussians. We similarly exploit the following ratio inequality akin to Lemma 2.1,

Lemma D.1. *For probability measures P, Q,*

$$\mathbb{E}_{x\sim P}\left[\frac{\mathrm{d}Q}{\mathrm{d}P + \mathrm{d}Q}\right] \le \frac{1}{2}(1 - \frac{1}{2}\mathrm{H}^2(P,Q)).$$

We apply the following well-known formula for the Hellinger distance between two Gaussians.

Lemma D.2. *We have*

$$1-\frac{1}{2}H^2(\mathcal{N}(\mu_P,\Sigma_P),\mathcal{N}(\mu_Q,\Sigma_Q)) = \frac{|\Sigma_P|^{1/4}|\Sigma_Q|^{1/4}}{\left|\frac{\Sigma_P+\Sigma_Q}{2}\right|^{1/2}}\exp\left\{-\frac{1}{8}(\mu_P-\mu_Q)^\top\left[\frac{\Sigma_P+\Sigma_Q}{2}\right]^{-1}(\mu_P-\mu_Q)\right\}.$$

Example D.3. [Two Isotropic Gaussians] Let $\Theta=\{\pm 1\}$, $p^{+1}=\mathcal{N}(\mu,\mathrm{Id})$, $p^{-1}=\mathcal{N}(-\mu,\mathrm{Id})$. Then, we have a critical window transitioning from sampling from both components to the component $+1$ between $T_{\text{before}}=\ln\|\mu\|+\ln 2+\ln 1/\epsilon$ and $T_{\text{after}}=\ln\|\mu\|-\ln\ln\frac{1}{2\epsilon^2}$. When $\widehat{T}\le T_{\text{after}}$, then $\mathrm{TV}(p^{+1,\widehat{T}},p^{+1})\lesssim\epsilon$. When $\widehat{T}\ge T_{\text{before}}$, $\mathrm{TV}(p^{+1,\widehat{T}},p)\lesssim\epsilon$.

Proof. The proof for T_{before}, a simple application of Pinsker's inequality, can be found in Appendix B.1 of (Li & Chen, 2024). Directly applying the new Master Theorem 3.1 to T_{after}, we need only show that

$$\frac{\sqrt{2}}{2}\left(1-\frac{1}{2}\mathrm{H}^2(p^{+1}_{T_{\text{after}}},p^{-1}_{T_{\text{after}}})\right)\le\frac{\sqrt{2}}{2}\exp(-\frac{1}{2}\|\mu\|^2e^{-2t})\le\epsilon.$$

□

Example D.4. [Random mean spherical Gaussians] We first sample $\mu_i\sim\mathcal{N}(0,\mathrm{Id})$ for $i\in[K]$ i.i.d. and let $\Theta=\{\mathcal{N}(\mu_i,\mathrm{Id})\}_{i\in[K]}$. We let $S_{\text{before}}=\Theta$ and $S_{\text{after}}=\{\mu_1\}$. Then, we can compute $T_{\text{before}}=\max_{j\in[K]}\ln\|\mu_i-\mu_j\|+\ln(1/\epsilon)$ and $T_{\text{after}}=\min_{j\in[K],i\neq j}\ln\|\mu_i-\mu_j\|-\frac{1}{2}\ln 8\ln\frac{K}{\epsilon}$. Furthermore, with high probability over the selection of the means, $T_{\text{before}}-T_{\text{after}}=O_{K,\epsilon}(1)$ as $d\to\infty$.

Proof. The proof for T_{before} can be found in Section 5.2 of (Li & Chen, 2024). We need to slightly modify the proof of Theorem 3.1 so that we can write the desired bound for T_{after} in terms of the Hellinger distance of individual components. We use the same notation. By convexity, we can bound

$$\begin{aligned}\mathbb{E}_{Y_{\widehat{T}}\sim p^{S_{\text{targ}}}_{\widehat{T}}}\left[\frac{\sum_{\theta\in\Theta-S_{\text{targ}}}w_\theta p^\theta_{\widehat{T}}(Y_{\widehat{T}})}{\sum_{\theta\in\Theta}w_\theta p^\theta_{\widehat{T}}(Y_{\widehat{T}})}\right]&\le\sum_{\theta\in\Theta-S_{\text{targ}}}w_\theta\sum_{\phi\in S_{\text{targ}}}w_\phi\mathbb{E}_{Y_{\widehat{T}}\sim p^\phi_{\widehat{T}}}\left[\frac{p^\theta_{\widehat{T}}(Y_{\widehat{T}})}{w_\theta p^\theta_{\widehat{T}}(Y_{\widehat{T}})+w_\phi p^\phi_{\widehat{T}}(Y_{\widehat{T}})}\right]\\&\le K\max_{\theta\in\Theta-S_{\text{targ}},\phi\in S_{\text{targ}}}\left(1-\frac{1}{2}\mathrm{H}^2(p^\theta_{\widehat{T}},p^\phi_{\widehat{T}})\right)\le\epsilon,\end{aligned}$$

when $\widehat{T}\le T_{\text{after}}$. To conclude the second part of the theorem, observe that by concentration of measure (e.g., Theorem 3.1.1 from (Vershynin)) and a union bound, there exists a constant T independent of d such that $\|\mu_j\|\in[\sqrt{d}-T,\sqrt{d}+T]$ for all $j\in[K]$ with high probability. Furthermore, by known Gaussian Suprema inequalities, we can also assume that there exists a constant T' independent of d such that $|\langle\mu_i,\mu_j\rangle|\le T'\|\mu_i\|$ (Lemma 5.1 from (van Handel, 2016)). Thus, we can conclude that

$$\max_{j\in[K]}\|\mu_i-\mu_j\|^2\le 2d+4T\sqrt{d}+2T^2+2T'(\sqrt{d}+T)=O(d).$$
$$\max_{j\in[K]}\|\mu_i-\mu_j\|^2\ge 2d-4T\sqrt{d}+2T^2-2T'(\sqrt{d}+T)=\Omega(d).$$

The difference in $\log$ scale is thus constant,

$$\frac{1}{2}\left(\ln\max_{j\in[K]}\|\mu_i-\mu_j\|^2-\ln\min_{j\in[K],i\neq j}\|\mu_i-\mu_j\|^2\right)=O_{K,\epsilon}(1)$$

□

Example D.5. [Two Dirac delta functions with a random masking procedure] Let $p\in\{\pm 1\}^T$, and consider a forward process with index set $\mathbf{I}=[0,1]$, $Y_0=X$, and $Y_t\in\{\pm 1,[\text{MASKED}]\}^T$. For $t\in\mathbf{I}$, we let all the value at index $i\in[T]$ be set to [MASKED] with probability t independently. For a mixture of two Dirac delta functions, we can express the critical window in terms of the *Hamming* *distance* between the corresponding strings. Let $\Theta = \{\theta_{\pm 1}\}, \ell_{\pm 1} \in \{\pm 1\}^T, p^{\theta_{\pm 1}} \sim \delta_{\ell_{\pm 1}}, w_{\pm 1} = \frac{1}{2}$. Then, on component 1 we have the critical window

$$T_{\text{before}} = \exp\left[\frac{\ln(1-\epsilon)}{d_H(\delta_{\ell_1}, \delta_{\ell_{-1}})}\right], T_{\text{after}} = \exp\left[\frac{\ln \epsilon^2}{d_H(\delta_{\ell_1}, \delta_{\ell_{-1}})}\right]$$

When $\widehat{T} \leq T_{\text{after}}$, then $\text{TV}(p^{1,\widehat{T}}, p^1) \lesssim \epsilon$. When $\widehat{T} \geq T_{\text{before}}$, $\text{TV}(p^{1,\widehat{T}}, p) \lesssim \epsilon$. For sufficiently large $d_H(\delta_{\ell_1}, \delta_{\ell_{-1}})$, the window size $T_{\text{before}} - T_{\text{after}} = O\left(\frac{1}{d_H(\delta_{\ell_1}, \delta_{\ell_{-1}})}\right)$. If $d_H(\delta_{\ell_1}, \delta_{\ell_{-1}})$ increases with T, then the width of the critical window compared to the width of the index set becomes negligible.

Proof. To prove $\text{TV}(p^{1,\widehat{T}}, p^1) \lesssim \epsilon$ when $\widehat{T} \leq T_{\text{after}}$, observe that when $\widehat{T} \leq T_{\text{after}}$, the probability that all the differing elements between ℓ_1, ℓ_{-1} are masked is exactly $\widehat{T}^{d_H(\delta_{\ell_1}, \delta_{\ell_{-1}})} \leq \epsilon^2$. That means that there exists a set A with $p_{\widehat{T}}^{+1}(A) \geq 1 - \epsilon^2$ and $p_{\widehat{T}}^{-1}(A) = 0$, so by the definition of total variation, $\text{TV}(p_{\widehat{T}}^{+1}, p_{\widehat{T}}^{-1}) \geq 1 - \epsilon^2$. Obviously, $\text{TV}(p_{\widehat{T}}^{+1}, p_{\widehat{T}}^{+1}) = 0$ as well, so by Theorem 3.1, we obtain $\text{TV}(p^{1,\widehat{T}}, p^1) \lesssim \epsilon$. To prove that $\text{TV}(p^{1,\widehat{T}}, p^{\{\pm 1\}}) \leq \epsilon$ when $\widehat{T} \geq T_{\text{before}}$, we need only show that $\text{TV}(p_{\widehat{T}}^{+1}, p_{\widehat{T}}) \leq \epsilon$. By Lemma 15 of (Li & Chen, 2024), it suffices to show that $\text{TV}(p_{\widehat{T}}^{+1}, p_{\widehat{T}}^{-1}) \leq \epsilon$ by a simple triangle inequality argument. Consider the set $A \subset \{\pm 1, [\text{MASKED}]\}^T$ such that $\text{TV}(p_{\widehat{T}}^{+1}, p_{\widehat{T}}^{-1}) = p_{\widehat{T}}^{+1}(A) - p_{\widehat{T}}^{-1}(A)$. Consider the set $B = \text{supp}(p_{\widehat{T}}^{+1}) \cap \text{supp}(p_{\widehat{T}}^{-1})$. For any $x \in B$, we know $p_{\widehat{T}}^{+1}(x) = p_{\widehat{T}}^{-1}(x)$ because the same number of tokens need to be masked from $\ell_{\pm 1}$. This means we have $p_{\widehat{T}}^{+1}(B) = p_{\widehat{T}}^{-1}(B) \geq \widehat{T}^{d_H(\ell_{+1}, \ell_{-1})} \geq 1 - \epsilon$. Because $p_{\widehat{T}}^{+1}(A) - p_{\widehat{T}}^{-1}(A) = p_{\widehat{T}}^{+1}(A - B) - p_{\widehat{T}}^{-1}(A - B)$, we have $p_{\widehat{T}}^{+1}(A) - p_{\widehat{T}}^{-1}(A) \leq p_{\widehat{T}}^{+1}(\{\pm 1, [\text{MASKED}]\}^T - B) \leq \epsilon$. $\square$

D.2 AUTOREGRESSION

Example D.6. [Math problem-solving as a random walk] We model solving a math problem as taking a random walk on $\mathbb{Z}$ with stepsize 1 of length T. If the random walk hits $+A$, then it has 'solved' the problem; if the random walk hits $-A$, then it has obtained an incorrect solution. Assume that we have two modes: a strong problem solving mode (denoted $+1$), which takes a $+1$ step with probability $0.5 + \Delta$, and a weak problem solving mode (denoted -1), which takes a $+1$ step with probability $0.5 - \Delta$. Assuming that $\frac{\ln(2/\epsilon^2)}{2\Delta^2} < A$ and $\epsilon^2 < 10^{-3}(0.5 - \Delta)(0.5 + \Delta)$, there is a critical window for the strong problem solving window of $T_{\text{before}} = T - \frac{\epsilon^2}{\Delta^2} + 2$ and $T_{\text{after}} = T - \frac{\ln(2/\epsilon^2)}{2\Delta^2}$. Note the critical window has width $\Theta(1/\Delta^2)$ independent of T.

Proof. Because only the direction of steps matter, we can model the critical window for this random walk as observing a sequence of ± 1 with an autoregressive language model. Let $p \in \{\pm 1\}^T$, and consider a forward process with index set $\mathbf{I} = \{0, 1, 2, \ldots, T\}$, $Y_0 = X$, and $Y_t \in \{\pm 1, [\text{MASKED}]\}^T$. For $t \in \mathbf{I}$, we let the last t tokens of Y_t be deterministically set to [MASKED]. We generate data as a mixture of biased coins with separation $2\Delta < 0.01$. For a mixture of two biased coins, with probabilities of $0.5 \pm \Delta$ ($\theta_{\pm 1}$ respectively) of yielding 1, we can compute the critical window and show that it tightly clusters around $\Theta(1/\Delta^2)$. Let $\Theta = \{\theta_{\pm 1}\}, p^{\theta_{\pm 1}} \sim (\text{Bern}(\theta_{\pm 1}))^{\otimes T}, w_{\pm 1} = \frac{1}{2}$. We also assume $\epsilon^2 < 10^{-3}(0.5 - \Delta)(0.5 + \Delta)$. Then, on component 1 we have the critical window $T_{\text{before}} = T - \frac{\epsilon^2}{\Delta^2} + 2$ and $T_{\text{after}} = T - \frac{\ln(2/\epsilon^2)}{2\Delta^2}$. When $\widehat{T} \leq T_{\text{after}}$, then $\text{TV}(p^{1,\widehat{T}}, p^1) \lesssim \epsilon$. When $\widehat{T} \geq T_{\text{before}}$, $\text{TV}(p^{1,\widehat{T}}, p) \lesssim \epsilon$.

Note that the number of $+1$ is sufficient for disambiguating $\theta_{\pm 1}$. To prove the bounds T_{before}, we show that with only $\frac{\epsilon^2}{\Delta^2} - 2$ samples the total variation between $0.5 - \Delta$ and $0.5 + \Delta$ is negligible. Using Roos (2001), we find

$$\text{TV}(p_{T_{\text{after}}}^{+1}, p_{T_{\text{after}}}^{-1}) \leq \frac{2\Delta\sqrt{\frac{T - T_{\text{after}} + 2}{2(0.5-\Delta)(0.5+\Delta)}}}{\left(1 - 2\Delta\sqrt{\frac{T - T_{\text{after}} + 2}{2(0.5-\Delta)(0.5+\Delta)}}\right)^2} \lesssim 3\epsilon.$$

For T_{after}, we compute how many samples it takes for $p^{\pm 1}$ to have only ϵ overlap in total variation using Hoeffding's inequality. If we have n samples, the mean $\overline{X}$ of the n samples of ± 1 for $p = 0.5 + \Delta$ satisfies the concentration inequality $P(|\overline{X} - 2\Delta| > \Delta) \leq 2\exp(-2nt^2)$ (furthermore we can ignore the stopping condition by our requirement that $\frac{\ln(2/\epsilon^2)}{2\Delta^2} < A$). We find $P(|\overline{X} - 2\Delta| > \Delta) \leq \epsilon^2$ for $T - T_{\text{after}}$ samples, proving that the total variation is at least $1 - \epsilon^2$. □

D.2.1 AUTOREGRESSIVE EXAMPLE WITH A GAUSSIAN MIXTURE MODEL AS THE CONCEPT DISTRIBUTION

We consider a model for autoregressive data similar to the one presented in (Arora et al., 2019). Each word is a vector $w \in \mathbb{R}^d$ and the context length is $T \in \mathbb{Z}^{>0}$. The original samples are $x \in \mathbb{R}^{T \times d}$. Let $\Theta = \{u, v\}$, where $u, v \in \mathcal{S}^{d-1}$. We define the distribution p^θ for $\theta \in \Theta$ as follows. We generate the path of a discourse vector $\left(C_t^\theta\right)_{t \in [0,\infty]} \in \mathbb{R}^d$ with the reverse SDE Orstein-Uhlenbeck process such that $C_\infty^\theta \sim \mathcal{N}(0, \mathrm{Id})$ and $C_0^\theta \sim \mathcal{N}(0, \mathrm{Id} + \alpha\theta\theta^\top)$ for some $\alpha > 0$. We let q_t^θ be the law of C_t^θ for $t \geq 0$. We let $\mathbf{I} = \{0, 1, \ldots, T\}$, and for $t \in \mathbf{I}$, we draw samples $w_t^\theta \in \mathbb{R}^d$ where we impose a normal Gaussian prior and have $w_t^\theta | C_t^\theta \propto \exp(\langle C_t^\theta, \cdot \rangle)$. Then we return the corpus $\{w_t^\theta\}_{t \in \mathbf{I}}$ as an output.

Theorem D.7 (Autoregressive with a mixture of two Gaussians as the concept distribution)**.** *We assume that* $\exp(-T)\sqrt{\alpha - \log(1+\alpha)} \leq \epsilon$. *Let* $\phi(x) = \frac{2+x}{(2+x(1+\langle u,v\rangle)/2)(2+x(1-\langle u,v\rangle)/2)}$. *Then, on component* u *we have*

$$T_{\text{after}} = \frac{1}{2}\ln\left[\frac{\alpha}{\phi^{-1}(\epsilon^2)}\right], T_{\text{before}} = \frac{1}{2}\ln\left[\frac{\alpha\sqrt{1+\alpha^2}\sqrt{1-\langle u,v\rangle^2}}{\epsilon}\right].$$

When $\widehat{T} \leq T_{\text{after}}$*, then* $\mathrm{TV}(p^{\{u\},\widehat{T}}, p^{\{u\}}) \lesssim \epsilon$*. When* $\widehat{T} \geq T_{\text{before}}$*,* $\mathrm{TV}(p^{\{u\},\widehat{T}}, p) \lesssim \epsilon$*.*

This proof will require Theorem D.8 and Lemmas D.2 and D.9, which is stated below:

Theorem D.8 (Section 5.2 of Chen et al. (2023))**.** *Let* $(Y_t)_{t \in [0,T]}$ *and* $(Y'_t)_{t \in [0,T]}$ *denote the solutions to*

$$\begin{aligned} \mathrm{d}Y_t &= b_t(Y_t)\,\mathrm{d}t + \sqrt{2}\mathrm{d}B_t\,, \qquad Y_0 \sim q \\ \mathrm{d}Y'_t &= b'_t(Y'_t)\,\mathrm{d}t + \sqrt{2}\mathrm{d}B_t\,, \qquad Y'_0 \sim q\,. \end{aligned}$$

Let q *and* q' *denote the laws of* Y_T *and* Y'_T *respectively. If* b_t, b'_t *satisfy that* $\int_0^T \mathbb{E}_Q \|b_t(Y_t) - b'_t(Y_t)\|^2\,\mathrm{d}t < \infty$*, then* $\mathrm{KL}(q\|q') \leq \int_0^T \mathbb{E}_Q \|b_t(Y_t) - b'_t(Y_t)\|^2\,\mathrm{d}t$*.*

Lemma D.9. *Let* $u, v \in \mathcal{S}^{d-1}$*. Then*

$$\begin{aligned} \|uu^\top - vv^\top\|_{\mathrm{op}} &\leq \sqrt{1 - \langle u,v\rangle^2} \\ \lambda(uu^\top + vv^\top) &= \{1 \pm \langle u,v\rangle\}. \end{aligned}$$

Proof. There exists $r \in \mathcal{S}^{d-1}$ such that $v = \langle u,v\rangle u + \sqrt{1-\langle u,v\rangle^2}r$ and $u \perp r$. We find that

$$uu^\top - vv^\top = (1 - \langle u,v\rangle^2)uu^\top - (1-\langle u,v\rangle^2)rr^\top - \langle u,v\rangle\sqrt{1-\langle u,v\rangle^2}[ur^\top + ru^\top].$$

We can explicitly compute the eigenvalues of $uu^\top - vv^\top$ using the discriminant and find that they are equal to $\pm\sqrt{1-\langle u,v\rangle^2}$. By a similar derivation, we can write

$$uu^\top + vv^\top = (1 + \langle u,v\rangle^2)uu^\top + (1-\langle u,v\rangle^2)rr^\top + \langle u,v\rangle\sqrt{1-\langle u,v\rangle^2}[ur^\top + ru^\top].$$

which gives us eigenvalues for $uu^\top + vv^\top$ of $1 \pm \langle u,v\rangle$. □

Proof. To compute the T_{after} bounds, we compare the difference in Hellinger distance of the distribution of words words at generated at index $\widehat{T}$, $w_{\widehat{T}}^u$. By the data processing inequality $1 - \frac{1}{2}H^2(p_{\widehat{T}}^u, p_{\widehat{T}}^v) \leq 1 - \frac{1}{2}H^2(w_{\widehat{T}}^u, w_{\widehat{T}}^v)$, so it suffices to show $1 - \frac{1}{2}H^2(w_{\widehat{T}}^u, w_{\widehat{T}}^v) \lesssim \epsilon$. Because the Gaussian is its own conjugate prior and $w_{\widehat{T}}^u | C_t^\theta \propto \exp(-\frac{1}{2}\|w_{\widehat{T}}^u - C_t^\theta\|^2)$, we can compute

$w^u_{\widehat{T}} | C^\theta_t \sim \mathcal{N}(C^\theta_t, \mathrm{Id})$ and $w^u_{\widehat{T}} \sim \mathcal{N}(0, 2\mathrm{Id} + \alpha e^{-2t} uu^\top)$. Applying Lemmas D.2 and D.9, we can explicitly compute

$$1 - \frac{1}{2} H^2(w^u_{\widehat{T}}, w^v_{\widehat{T}}) \lesssim \sqrt{\frac{2 + \alpha e^{-2\widehat{T}}}{(2 + \alpha e^{-2\widehat{T}}(1 + \langle u, v\rangle)/2)(2 + \alpha e^{-2\widehat{T}}(1 - \langle u, v\rangle)/2)}} \lesssim \sqrt{\phi(\alpha e^{-2\widehat{T}})} \leq \epsilon.$$

To compute T_{before}, we first use the data processing inequality to reduce the difference in the emitted tokens to the difference in the paths of the context vectors, and then apply the approximation error bounds from Theorem D.8 to bound the differences in path measures. When $\widehat{T} \geq T_{\text{before}}$, we can use the triangle inequality to write $\mathrm{TV}(p^{\{u\}, \widehat{T}}, p) = \mathrm{TV}(p^u_{\widehat{T}}, p^{\{u,v\}}_{\widehat{T}}) \leq \mathrm{TV}(p^u_{\widehat{T}}, p^v_{\widehat{T}})$. Note that $p^\theta_{\widehat{T}}$ is the distribution of the first $T - \widehat{T}$ tokens generated by the model under θ. Note that $p^u_{\widehat{T}}$ is a function of $(C^u_t)_{t \in \mathbf{I} \cap [\widehat{T}, T]}$ and $p^v_{\widehat{T}}$ is a function of $(C^v_t)_{t \in \mathbf{I} \cap [\widehat{T}, T]}$. By the data processing inequality, we can bound the difference in terms of the distributions over the tokens in terms of the law of the process of the discourse vectors,

$$\mathrm{TV}(p^u_{\widehat{T}}, p^v_{\widehat{T}}) \leq \mathrm{TV}((C^u_t)_{t \in \mathbf{I} \cap [\widehat{T}]}, (C^v_t)_{t \in \mathbf{I} \cap [\widehat{T}, T]}) \leq \mathrm{TV}((C^u_t)_{t \in [\widehat{T}, T]}, (C^v_t)_{t \in [\widehat{T}, T]}).$$

Note that for $\theta \in \Theta$, $(C^\theta_t)_{t \in [\widehat{T}, T]}$ is generated by the following reverse time SDE,

$$\mathrm{d}C^\theta_t = \{C^\theta_t + 2\nabla \ln q^\theta_t(C^\theta_t)\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \qquad t \in [\widehat{T}, T], C^\theta_T \sim q^\theta_T.$$

Now we define $(C^{u \to v}_t)_{t \in [\widehat{T}, T]}$ to be the reverse SDE defined by initializing at q^u_T but with the score of q^v_t,

$$\mathrm{d}C^{u \to v}_t = \{C^{u \to v}_t + 2\nabla \ln q^v_t(C^{u \to v}_t)\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \qquad t \in [\widehat{T}, T], C^{u \to v}_T \sim q^u_T.$$

By the triangle inequality, we have

$$\mathrm{TV}((C^u_t)_{t \in [\widehat{T}, T]}, (C^v_t)_{t \in [\widehat{T}, T]}) \leq \underbrace{\mathrm{TV}((C^u_t)_{t \in [\widehat{T}, T]}, (C^{u \to v}_t)_{t \in [\widehat{T}, T]})}_{\text{(I)}} + \underbrace{\mathrm{TV}((C^{u \to v}_t)_{t \in [\widehat{T}, T]}, (C^v_t)_{t \in [\widehat{T}, T]})}_{\text{(II)}}.$$

To bound (I), observe that the SDEs have different scores but the same initializations. We apply Theorem D.8 to $\mathrm{TV}((C^u_t)_{t \in [\widehat{T}, T]}, (C^{u \to v}_t)_{t \in [\widehat{T}, T]})$ and obtain

$$\begin{aligned} \mathrm{TV}((C^u_t)_{t \in [\widehat{T}, T]}, (C^{u \to v}_t)_{t \in [\widehat{T}, T]}) &\leq \sqrt{\mathrm{KL}((C^u_t)_{t \in [\widehat{T}, T]} \| (C^{u \to v}_t)_{t \in [\widehat{T}, T]})} \\ &\leq \sqrt{\int_{\widehat{T}}^T \mathbb{E}_{X \sim C^u_t} \|\nabla \ln p^u_t(X) - \nabla \ln p^v_t(X)\|^2 dt} \end{aligned}$$

We simplify the inner expectation by using the

$$\begin{aligned} \|\nabla \ln p^u_t(X) - \nabla \ln p^v_t(X)\| &= \| \left[(\mathrm{Id} + \alpha e^{-2t} uu^\top)^{-1} - (\mathrm{Id} + \alpha e^{-2t} vv^\top)^{-1} \right] x \| \\ &= \| \left[\left(\mathrm{Id} - \frac{\alpha e^{-2t}}{1 + \alpha e^{-2t}} uu^\top \right) - \left(\mathrm{Id} - \frac{\alpha e^{-2t}}{1 + \alpha e^{-2t}} vv^\top \right) \right] x \| \\ &\lesssim \alpha e^{-2t} \|uu^\top - vv^\top\|_{\mathrm{op}} \left\| \prod_{\mathrm{span}(u,v)} x \right\|, \\ &= \alpha e^{-2t} \sqrt{1 - \langle u, v\rangle^2} \left\| \prod_{\mathrm{span}(u,v)} x \right\|. \quad \text{(Lemma } D.9) \end{aligned}$$

We can upper bound $\mathbb{E}_{X \sim C^u_t} \| \prod_{\mathrm{span}(u,v)} X\|^2$ by considering right-triangular L such that $L^\top L = \mathrm{Id} + \alpha uu^\top$. and $\prod_{\mathrm{span}(u,v)} X = \prod_{\mathrm{span}(u,v)} LY$, where $Y \sim \mathcal{N}(0, \mathrm{Id})$. The operator norm of $\prod_{\mathrm{span}(u,v)} L$ is

$$\left\| \prod_{\mathrm{span}(u,v)} \circ L \right\|_{\mathrm{op}} \leq \|L\|_{\mathrm{op}} \leq \sqrt{1 + \alpha^2 e^{-2t}} \leq \sqrt{1 + \alpha^2}.$$

$\prod_{\text{span}(u,v)} \circ L$ is also rank 2 and $\left(\prod_{\text{span}(u,v)} \circ L\right) Y = \left(\prod_{\text{span}(u,v)} \circ L\right) \circ \prod_{L^{-1}\text{span}(u,v)} Y$, where $\prod_{L^{-1}\text{span}(u,v)} Y \sim \mathcal{N}(0, \text{Id}_2)$. Thus we have

$$\mathbb{E}_{Y \sim \mathcal{N}(0,\text{Id})} \left[\left\| \left(\prod_{\text{span}(u,v)} \circ L \right) Y \right\|^2 \right] = \mathbb{E}_{Y' \sim \mathcal{N}(0,\text{Id}_2)} \left[\left\| \left(\prod_{\text{span}(u,v)} \circ L \right) Y' \right\|^2 \right] \lesssim 1 + \alpha^2$$

Combining this information together, we are able to compute,

$$\begin{aligned} \text{TV}((C_t^u)_{t \in [\widehat{T},T]}, (C_t^{u \to v})_{t \in [\widehat{T},T]}) &\leq \sqrt{\int_{\widehat{T}}^T \mathbb{E}_{X \sim C_t^u} \|\nabla \ln p_t^u(X) - \nabla \ln p_t^v(X)\|^2 dt} \\ &\lesssim \alpha \sqrt{1+\alpha^2} \sqrt{1 - \langle u, v\rangle^2} e^{-2\widehat{T}} \\ &\lesssim \epsilon. \end{aligned}$$

To bound (II), we observe that both are run with the same score so we need only bound the difference at initialization. By the data processing inequality, we again have $\text{TV}((C_t^{u \to v})_{t \in [\widehat{T},T]}, (C_t^v)_{t \in [\widehat{T},T]}) \leq \text{TV}(p_T^u, p_T^v)$. We can again apply the triangle inequality to get $\text{TV}(p_T^u, p_T^v) \leq \text{TV}(p_T^u, \gamma^d) + \text{TV}(\gamma^d, p_T^v) \leq$. For any $\theta \in \Theta$, we have by the forward convergence of the OU process $\text{TV}(p_T^u, \gamma^d) \leq \exp(-T)\sqrt{\text{KL}(p_u || \gamma^d)}$. We can explicitly compute $\text{KL}(p_u||\gamma^d)$ as

$$\text{KL}(p_u||\gamma^d) = \frac{1}{2}\left[d + \alpha - d - \log(\text{Id} + \alpha u u^\top)\right] = \frac{1}{2}\left[\alpha - \log(1+\alpha)\right].$$

Thus, we obtain the following bound on $\text{TV}((C_t^{u \to v})_{t \in [\widehat{T},T]}, (C_t^v)_{t \in [\widehat{T},T]})$ of

$$\text{TV}((C_t^{u \to v})_{t \in [\widehat{T},T]}, (C_t^v)_{t \in [\widehat{T},T]}) \lesssim \exp(-T)\sqrt{\alpha - \log(1+\alpha)} \lesssim \epsilon.$$

□

D.3 Interweaving transitions from other distributions

In this section, we extend our critical windows framework to the setting where at certain steps of sampling procedure, instead of using the reverse Markov transition kernel from the original stochastic localization sampler, we use an alternative distribution which is not necessarily related to the original sampler. This includes many important applications of generative models, in which one seeks to combine the priors learned from data with some other algorithm. For example, one may want to combine the language model with a problem generation oracle in in-context learning (Dong et al., 2024).

As Xie et al. (2022) points out, the transition from the answer to one problem to the problem statement of another example in-context learning is determined by an alternative transition kernel (which they call p_{prompt}). Although the probability of transition from one answer to the problem statement of another example is extremely low under the natural data distribution, one still hopes that with sufficiently many samples, the model selects the correct $\theta^* \in \Theta$ if these lower probability transitions are overcome by the distributional difference for $\theta \in \Theta$ with $\theta \neq \theta^*$. Similarly, under our critical windows framework, we can hope to capture the idea that we specialize to a particular θ^* given a sufficiently long context. In Section D.3.1, we first present a general framework for characterizing critical windows in this setting. Then, in section D.3.2, we consider the case of in-context learning by autoregressive language models and prove convergence.

D.3.1 General interweaving framework

We present this framework for the case where the index set $\mathbf{I} = \{0, 1, \ldots, m\}$ is discrete. Like before, assume we have a series of reverse Markov transition kernels $P^{\leftarrow,p}_{k \to k-1}(\cdot|\cdot)$, for $k \in \mathbf{I}$, but we also assume we have an alternative distribution $P^{\leftarrow,\text{alt}}_{k \to k-1}(\cdot|\cdot)$ that we use to sample for transitions $k \in A \subsetneq \mathbf{I}$. For our sampling procedure, we sample Y_m, and for $k = m-1, m-2, \ldots, 0$, we take $Y_k \sim P^{\leftarrow,p}_{k+1 \to k}(\cdot|Y_{k+1})$ for $k \in \mathbf{I} - A$ and $Y_k \sim P^{\leftarrow,\text{alt}}_{k+1 \to k}(\cdot|Y_{k+1})$ for $k \in A$. We denote the final distribution p^{alt}.

Now, we also need to adjust our definitions of p^S to this particular sampling procedure. We define $p^{S,\text{alt}}$ for $S \subset \Theta$ to the distribution over outputs when we instead use the kernels $P^{\leftarrow}_{k+1\to k}(\cdot|Y_{k+1}, S)$ instead of $P^{\leftarrow}_{k+1\to k}(\cdot|Y_{k+1})$. To relate $p^{\theta,\text{alt}}$ to $p^{\theta,\text{alt}}$ for $\theta \in \Theta$, we need to assume transitions from alt do not affect the posterior distribution over $p^{\Theta}(\theta|Y_t)$.

Assumption D.10. For all $y \in \mathbb{R}$ and $x \in \text{supp}(P^{\leftarrow,\text{alt}}_{k+1\to k}(\cdot|y))$, we have for all $\theta, \theta' \in \Theta$, the equality $P^{\leftarrow}_{k+1\to k}(x|y,\theta) = P^{\leftarrow}_{k+1\to k}(x|y,\theta')$.

Adopting our definitions from Section 3, we let

$$T_{\text{lower,alt}}(\epsilon) \in \{t \in \mathbf{I} : \text{TV}(p_t^{S_{\text{init}},\text{alt}}, p_t^{S_{\text{targ}},\text{alt}}) \le \epsilon\}$$
$$T_{\text{upper,alt}}(\epsilon) \in \{t \in \mathbf{I} : \text{TV}(p_t^{S_{\text{targ}},\text{alt}}, p_t^{\Theta - S_{\text{targ}},\text{alt}}) \ge 1 - \epsilon^2\}.$$

The main challenge of the below corollary is simply show that the final distribution $p^{\Theta,\text{alt}}$ can be written as a mixture of $p^{\theta,\text{alt}}$ with the same mixing weights as before.

Corollary D.11. *Under Assumption D.10, for $\epsilon > 0$, if $\widehat{T} \ge T_{\text{lower,alt}}(\epsilon)$ and $\widehat{T} \le T_{\text{upper,alt}}(\epsilon)$, then*

$$\text{TV}(p^{S_{\text{init}},\widehat{T}}, p^{S_{\text{targ}},\text{alt}}) \le \left(1 + \sqrt{2}\max\left(1, \frac{\sum_{\theta\in\Theta-S_{\text{targ}}} w_\theta}{\sum_{\theta\in S_{\text{targ}}} w_\theta}\right)/2\right)\epsilon.$$

Proof. We need only show that $p^{\Theta,\text{alt}} \triangleq \sum_{\theta\in\Theta} w_\theta p^{\theta,\text{alt}}$. It suffices to shows that the probability of generating a path $Y_m, Y_{m-1}, \ldots, Y_0$ are the same under both density functions. We need only consider transitions for $k \in \mathbf{I} - A$, because for $k \in A$, the transitions are both given by the alternative distribution. For the transitions not given by alt, note that we are using the original model, so

$$p^{\Theta,\text{alt}}(Y_{k-1}|Y_k) = \frac{\sum_{\theta\in\Theta} w_\theta p^\theta(Y_k) P^{\leftarrow}_{k\to k-1}(Y_{k-1}|Y_k,\theta)}{\sum_{\theta\in\Theta} w_\theta p^\theta(Y_k)}.$$

Furthermore, for the mixture model, this probability is

$$p^{\text{mix}} = \frac{\sum_{\theta\in\Theta} w_\theta p^{\theta,\text{alt}}(Y_m, Y_{m-1}, \ldots, Y_{k+1}, Y_k) P^{\leftarrow}_{k\to k-1}(Y_{k-1}|Y_k,\theta)}{\sum_{\theta\in\Theta} w_\theta p^{\theta,\text{alt}}(Y_m, Y_{m-1}, \ldots, Y_{k+1}, Y_k)}.$$

The distinction between Equation D.3.1 and Equation D.3.1 is that in the former we are using the likelihood of p^θ instead of $p^{\theta,\text{alt}}$. Thus it suffices to show that $p^\theta \propto p^{\theta,\text{alt}}$. We explicitly write out the probability,

$$p^\theta(Y_m, Y_{m-1}, \ldots, Y_{k+1}, Y_k) = \prod_{i=k+1}^{m} p^\theta(Y_{i-1}|Y_i) \propto \prod_{i=k+1, i\notin A}^{m} p^\theta(Y_{i-1}|Y_i),$$

where the proportionality follows from the fact that we can ignore the probability of the transitions produced by alt under Assumption D.10. By definition, this is proportional to $p^{\theta,\text{alt}}(Y_m, Y_{m-1}, \ldots, Y_{k+1}, Y_k)$ up to a normalization constant independent of θ. $\square$

D.3.2 IN-CONTEXT LEARNING

Now, we will specialize our framework to the case of in-context learning. As in Xie et al. (2022), we assume that the language model is given inputs of the form $[x_1, y_1, o, x_2, y_2, o, \ldots, x_t, y_t, o, x_{t+1}]$, where x_1 is the input, y_1 is the output, and o is a delimiter token that separate different in-context samples from each other. We assume that the transitions $y_i \to o \to x_{i+1}$ are sampled by some alternative probability distribution $P_{\text{other}}(\cdot|\cdot)$. We require that P_{other} selects the x_i i.i.d.

Assumption D.12. The distribution of $P_{\text{other}}(x_{t+1}|x_1, y_1, o, x_2, y_2, o, \ldots, x_t, y_t, o) = P_{\text{other}}(x_1)$.

Then we assume that the transitions $x_i \to y_i$ are generated by some $\theta^* \in \Theta$, which does not depend on any of the previous tokens before the delimiter.

Assumption D.13. (Well-specification) There exists some $\theta^* \in \Theta$ such that y_i is generated from $y_i \sim P^{\leftarrow}(\cdot|[x_1, y_1, o_1, \dots, o_{i-1}, x_i], \theta^*)$.

Assumption D.14. For all $\theta \in \Theta$, we have $P^{\leftarrow}(\cdot|[x_1, y_1, o_1, \dots, o_{i-1}, x_i], \theta) = P^{\leftarrow}(\cdot|x_i, \theta)$.

We also assume statistical separation of θ^* from $\Theta - \{\theta^*\}$ in terms of Hellinger distance.

Assumption D.15. Let $p^S_{(x,y)}$ for $S \subset \Theta$ be the distribution of (x_1, y_1, o) where $x_1 \sim P_{\text{other}}(\cdot|o)$ and $y_1 \sim p^S(\cdot|x_1)$. There exists $\delta > 0$ such that $H^2(p^{\Theta-\{\theta^*\}}_{(x,y)}, p^{\theta^*}_{(x,y)}) \geq \delta$.

Example D.16. Let $T \geq \ln\left(\frac{1-\delta/2}{\epsilon}\right)$. Under Assumptions D.10, D.12, D.13, D.14, and D.15, we have $\mathrm{TV}([x_1, \dots, x_{T+1}, y_{T+1}], [x_1, \dots, x_{T+1}, \tilde{y}_{T+1}]) \lesssim \epsilon / w_{\theta^*}$.

Proof. It suffices to upper bound $1 - \frac{1}{2}H^2(P^{\theta,\text{alt}}_{3T}, P^{\Theta-\{\theta\},\text{alt}}_{3T})$ by $O(\epsilon)$. First observe that the distribution $p^{S,\text{alt}}$ for $S \subset \Theta$ factors along the delimiters by a factor independent of S using Assumptions D.12 and D.14, so we have

$$P^{S,\text{alt}}_{3t}(x_1, y_1, o_1, x_2, y_2, o_2, \dots, x_t, y_t, o_t) \propto \prod_{i=1}^{t} P^{S,\text{alt}}_3(x_i, y_i, o_i).$$

Using the tensorization property of Hellinger distance and our definition of T, δ, we have

$$1 - \frac{1}{2}H^2(P^{\theta,\text{alt}}_{3T}, P^{\Theta-\{\theta\},\text{alt}}_{3T}) \lesssim \left[1 - \frac{1}{2}H^2(P^{\theta,\text{alt}}_3, P^{\Theta-\{\theta\},\text{alt}}_3)\right]^T \leq \epsilon.$$

□

D.4 All-or-nothing

Theorem D.17 (Reeves et al. (2019))**.** *Let π_s be the distribution over $\mathbb{R}^{n_s} \times \mathbb{R}^{m_s}$ for $n_s = s$ and $m_s = s+1$ where the marginal over θ is given by the uniform distribution over k_s-sparse vectors in $\{0,1\}^s$, and the conditional distribution $\pi_{z|\theta}$ is given by sampling $x \sim \mathcal{N}(0, \mathrm{Id}_s)$, taking $y = \langle \theta, x \rangle + \xi$ for $\xi \sim \mathcal{N}(0, \sigma_s^2)$, and outputting observation $z = (x, y)$. The null model π_s^{null} is given by sampling $x \sim \mathcal{N}(0, \mathrm{Id}_s)$ and outputting $y = \mathcal{N}(0, k_s + \sigma_s^2)$.*

If $\sigma_s^2 \ll k_s \leq s^{0.499}$, then (π_s) exhibits an all-or-nothing phase transition at threshold (N_s^*) with respect to null models (π_s^{null}) for $N_s^* \triangleq \frac{2k_s \log(s/k_s)}{\log(1+k_s/\sigma_s^2)}$.

Theorem D.18. *Suppose (π_s) is a sequence of inference tasks that exhibits an all-or-nothing phase transition at threshold (N_s^*) with respect to null models (π_s). Given $N_s \geq N_s^*$, let $(p^{s;\theta}_{(N_s)})_{\theta \in \Theta_s}$ denote the sequence of in-context learning tasks.*

For any constant $0 < \epsilon < 1$, there exist constants $\delta, \underline{s}$ such that for all $s \geq \underline{s}$, next-token prediction for $(p^{s;\theta}_{(N_s)})_{\theta \in \Theta_s}$ exhibits a critical window over $[N_s + 2 - (1+\delta)N_s^*, N_s + 2 - -(1-\delta)N_s^*]$ in which we transition from sampling a distribution $O(\epsilon)$-close in TV to $S_{\text{before}} = \Theta_{s;\mathsf{signal}}$, to sampling from a distribution $O(\epsilon)$-close in TV to $S_{\text{after}} = \Theta_s$.

The proof of Theorem D.18 is essentially immediate from Theorem 3.1 and the definition of the all-or-nothing phase transition:

Proof. By the operational characterization of TV distance, strong detection is (information-theoretically) possible if and only if $\liminf_{s\to\infty} \mathrm{TV}(p_s, q_s) = 1$, and weak detection is (information-theoretically) possible if and only if $\liminf_{s\to\infty} \mathrm{TV}(p_s, q_s) > 0$.

Let us first apply Theorem 3.1 to $S_{\text{init}} = S_{\text{targ}} = \Theta_{s;\mathsf{signal}}$. By the definition of D_Θ, the parameter W therein is 1. Furthermore, we trivially have that $T_{\text{end}}(\epsilon) = 0$. Finally, because strong detection is possible provided there are $N \geq \beta N_s^*$ in-context examples for $\beta > 1$, there exists δ_1 depending only on ϵ for which $\mathrm{TV}(p_t^{S_{\text{targ}}}, p_t^{\Theta_s - S_{\text{targ}}}) \geq 1 - \epsilon^2$ for $t = N_s + 2 - (1+\delta_1)N_s^*$. By Theorem 3.1 we conclude that $\mathrm{TV}(p^{S_{\text{init}}, N_s + 2 - (1+\delta_1)N_s^*}) \lesssim \epsilon$.

Next, let us apply Theorem 3.1 to $S_{\text{init}} = \Theta_{s;\textsf{signal}}$ and $S_{\text{targ}} = \Theta_s$. The parameter W therein is now 0. Furthermore, we trivially have that $T_{\text{end}}(\epsilon) = N_s + 2$. Finally, because weak detection is impossible provided there are $N \leq \beta N_s^*$ in-context examples for $\beta < 1$, there exists δ_2 depending only on ϵ for which $\mathrm{TV}(p_t^{S_{\text{init}}}, p_t^{S_{\text{targ}}}) \leq \epsilon$ for $t = N_s + 2 - (1 - \delta_2)N_s^*$. By Theorem 3.1 we conclude that $\mathrm{TV}(p^{S_{\text{init}}, N_s+2-(1-\delta_2)N_s^*}) \leq \epsilon$. Taking $\delta = \max(\delta_1, \delta_2)$ concludes the proof. $\square$

E Details from Section 5

Definition E.1. An ϵ**-mixture tree** is a tuple $(T, \{P^{\rightarrow}(\cdot|\cdot)\}, \mathbf{I}, \Theta, \{p^\theta\}_{\theta\in\Theta}, \text{Subset}, \text{NoiseAmount})$. The tree $T = (V, E)$ is associated with a function Subset: $V \to 2^\Theta \backslash \{\emptyset\}$, which maps vertices to sub-mixtures. We require Subset satisfies the following two properties: (1) Subset(root) $= \Theta$; (2) If u is a parent of v, Subset$(v) \subset$ Subset(u). We consider a NoiseAmount: $V \to \mathbb{R}^{\geq 0}$, which characterizes the noise levels that result in the aggregations of mixture components described by vertices in the mixture tree. Thus we require that NoiseAmount satisfy three properties: (1) For distinct $\theta_i, \theta_j \in \Theta$ with leaf nodes w, v such that $\theta_i \in$ Subset(w), $\theta_j \in$ Subset(v), if u is the lowest common ancestor of w, v, then we require $\mathrm{TV}(p^{\theta_i}_{\text{NoiseAmount}(u)}, p^{\theta_j}_{\text{NoiseAmount}(u)}) \leq \epsilon$; (2) For $u \in V$, we have statistical separation between Subset(u) and Θ − Subset(u) in terms of TV, $\mathrm{TV}(p^{\text{Subset}(u)}_{\text{NoiseAmount}(u)}, p^{\Theta-\text{Subset}(u)}_{\text{NoiseAmount}(u)}) \geq 1 - \epsilon^2$; and (3) If $v \in V$ is a parent of u, we have NoiseAmount$(u) <$ NoiseAmount(v). Property 1 establishes bounds on T_{end}, and properties 2 and 3 establishes bounds on T_{st}.

Corollary E.2. *Consider an ϵ-mixture tree. For $\theta_i \in \Theta$, consider the path $u_1, u_2, u_3, \ldots, u_{H'} \in V$ where u_1 is the leaf node with $\theta_i \in$ Subset(u_1) and $u_{H'}$ is the root. There is a sequence of times $T_1 < T_2 < \cdots < T_{H'}$ with $\mathrm{TV}(p^{\{i\}, T_\ell}, p^{\textit{Subset}(u_\ell)}) \lesssim_w \epsilon$.*

Proof. For $\ell \in [H']$, we let $T_\ell =$ NoiseAmount(u_ℓ). We apply Theorem 3.1 with $S_{\text{init}} = \{i\}$ and $S_{\text{targ}} =$ Subset(u_ℓ). We know $\mathrm{TV}(p_{T_\ell}^{S_{\text{targ}}}, p_{T_\ell}^{\Theta - S_{\text{targ}}}) \geq 1 - \epsilon^2$ by Condition 2 in Definition E.1. By Lemma 15 of (Li & Chen, 2024), we know $\mathrm{TV}(p_{T_\ell}^{\{i\}}, p_{T_\ell}^{S_{\text{init}}}) \leq \max_{j\in S_{\text{init}}} \mathrm{TV}(p_{T_\ell}^{\{i\}}, p_{T_\ell}^{\{j\}})$. This is $\leq \epsilon$ for all $j \in S_{\text{init}}$ by Condition 3 on NoiseAmount and the data processing inequality. $\square$

Corollary E.3. *Consider an ϵ-mixture tree $(T, \{P^{\rightarrow}(\cdot|\cdot)\}, \mathbf{I}, \Theta, \{p^\theta\}_{\theta\in\Theta}, \textit{Subset}, \textit{NoiseAmount})$. Suppose we have another distribution $\{q^\theta\}_{\theta\in\Theta}$ such that $\mathrm{TV}(p^\theta, q^\theta) \leq \delta/2$ for all $\theta \in \Theta$. Then we have $\epsilon + \sqrt{\delta}$-mixture tree given by $(T, \{P^{\rightarrow}(\cdot|\cdot)\}, \mathbf{I}, \Theta, \{q^\theta\}_{\theta\in\Theta}, \textit{Subset}, \textit{NoiseAmount})$.*

Proof. We need only check the first and second properties of NoiseAmount with parameter $\epsilon + \sqrt{\delta}$. To do this, it suffices to show $\mathrm{TV}(q^{\theta_i}_{\text{NoiseAmount}(u)}, q^{\theta_j}_{\text{NoiseAmount}(v)}) \leq \epsilon + \delta$ and $\mathrm{TV}(q^{\text{Subset}(u)}_{\text{NoiseAmount}(u)}, q^{\Theta-\text{Subset}(u)}_{\text{NoiseAmount}(u)}) \geq 1 - \epsilon^2 - \delta$. By the data processing inequality, we just need to show this at $t = 0$, and we prove the stronger statement that for $S_1 \subset \Theta$, $\mathrm{TV}(p^{S_1}, q^{S_1}) \leq \delta/2$. This follows from Lemma 15 of (Li & Chen, 2024) and $\mathrm{TV}(p^\theta, q^\theta) \leq \delta/2$ for all $\theta \in \Theta$. $\square$

Example E.4. Consider a set of alphabets $\{\mathcal{A}_i\}_{i=1}^d$ and define $\Theta = \{(a_i)_{i=1}^d : \forall i \in [d], a_i \in \mathcal{A}_i\}$ and $p^{\theta_i} = \delta_{\theta_i}$. Let $\mathbf{I} = [0, 1, 2, \ldots, d]$. and for any permutation $i_1, i_2, \ldots, i_d$ of $[d]$, define a forward process such that at $t \in \mathbf{I}$, we mask all $i_d, i_{d-1}, \ldots, i_{d-t}$. This constructs a hierarchy where the values for $i_1, i_2, \ldots, i_d$ are decided in that order.

Proof. We construct the following 0-mixture tree as follows. We let the leaf nodes be the set Θ. We let two leaf nodes u, v have the same parent if and only if they share the same values on the alphabet at $i_1, i_2, \ldots, i_{d-1}$; we also define the parent as the union of all of its children. We now treat the parents we constructed as the roots, and let them have the same parent if and only if they share the same values on the tuple $i_1, i_2, \ldots, i_{d-2}$. We continue to do this until we are left with one root node. We let Subset map each node to the corresponding set and NoiseAmount map each node to its distance from a leaf node.

By the construction of T, it is clear that Subset satisfies the desired properties. For distinct $\theta_i, \theta_j \in \Theta$, the lowest common ancestor of θ_i, θ_j represents the largest k such that indices $i_1, \ldots, i_k$ are the same for θ_i, θ_j. Because $p^{\theta_i}_{\text{NoiseAmount}(u)}$ is just the tuple of the values of θ_i, θ_j at $i_1, \ldots, i_k$, we know $\text{TV}(p^{\theta_i}_{\text{NoiseAmount}(u)}, p^{\theta_j}_{\text{NoiseAmount}(u)}) = 0$. For any $u \in V$ representing the values at index $(i_\ell)_{\ell=1}^k$, all $\theta \notin \text{Subset}(u)$ does not share the same values at these indices by definition, so we also know

$$\text{TV}(p^{\text{Subset}(u)}_{\text{NoiseAmount}(u)}, p^{\Theta-\text{Subset}(u)}_{\text{NoiseAmount}(u)}) = 1.$$

$\square$

F Experimental overview

As many authors (Ho et al., 2020; Meng et al., 2022; Choi et al., 2022; Raya & Ambrogioni, 2023; Georgiev et al., 2023; Sclocchi et al., 2024; 2025; Biroli et al., 2024; Li & Chen, 2024) have already empirically studied critical windows in the context of diffusion, we focus on experiments on critical windows for LLMs. In Section F.1, we validate our theory on outputs with a hierarchical structure, showing strong agreement with out predictions in Section 5. In Section F.2, we probe critical windows for LLAMA-3.1-8B-Instruct, Phi-3-7B-Instruct, and Qwen-2.5-7B-Instruct in real-world reasoning benchmarks.

F.1 Structured output experiments

To verify our theory for $T_{\text{before}}, T_{\text{after}}$, we have to compute the total variation between truncated responses from an LLM. This usually would take a large number of samples, so to circumvent this issue, we restrict the diversity of the LLM's generations and force the LLM to generate tokens in a structured format. In particular, we have LLAMA-3.1-8B-Instruct[6] respond to following prompt, which asks it to answer a series of fill-in-the-blank questions in a structured format. We also prefill the model's generations with `\n\n 1.` to ensure that the outputs comport to this format. To compute $T_{\text{before}}, T_{\text{after}}$, we look at when the generations diverge based on the first occurrence of the identifying information. For example, the T_{before} of the first critical window is `1.  The `, because the first answer has not appeared in the generation, and the T_{after} of the first critical window is `1.  The P` or `1.  The N`, because that uniquely identifies the answer. Figure 3 plots the probability of obtaining the same answers as the original generation after truncating different amounts from the generation in the forward-reverse experiments, computed with $10,000$ generations. Our theory predicts that jumps in the probability will occur at $T_{\text{before}}, T_{\text{after}}$ which represent when the model has committed to a particular answer to a question in the generation. These predictions are validated with our experiments, as the jumps in probability, representing the model localizing to a more specific set of answers, occur exactly at $T_{\text{before}}, T_{\text{after}}$.

Structured Output Prompt

Complete the following by choosing only one option for each blank. The options are provided in parentheses, and your response must match the exact case and meaning of the chosen option. Respond with only the completed sentence, no explanations or additional text.
1. The (Pirate/Ninja) jumped across the ship.
2. She adopted a (Dog/Cat) from the shelter.
3. The (River/Bridge) sparkled under the sun.
4. A (Dragon/Knight) guarded the castle gates.
5. He ordered (Pizza/Sushi) for dinner.

[6] Default sampling parameters of temperature of 0.6 and top-p sampling of 0.9

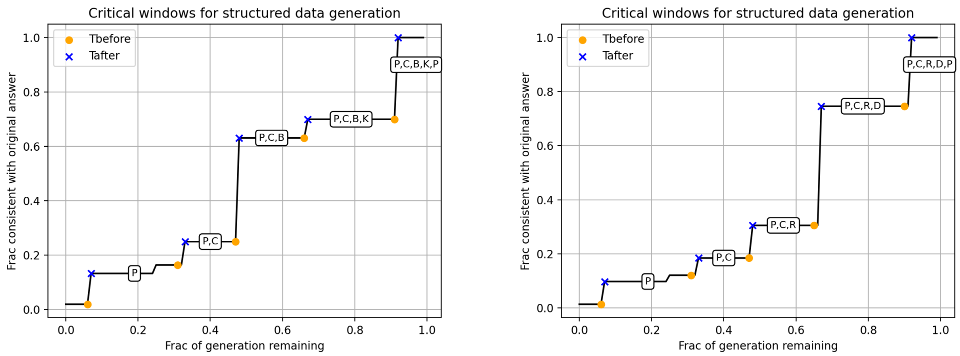

Figure 3: Structured output plots for LLAMA-3.1-8B-Instruct. P denotes that we are sampling from responses whose answer to the first question was Pirate; P,C denotes that we are sampling from responses whose answers to the first two questions were Pirate and Cat, respectively. We can see that the critical windows directly correspond to our theoretical values for $T_{\text{before}}, T_{\text{after}}$.

F.2 Chain of thought experiments

We then identify critical windows for LLAMA-3.1-8B-Instruct, Phi-3-7B-Instruct, and Qwen-2.5-7B-Instruct on 7 different math and reasoning benchmarks on which performance is known to improve with chain of thought reasoning (Lanham et al., 2023): ARC Challenge and Easy (Clark et al., 2018), AQua (Ling et al., 2017), LogiQA (Liu et al., 2020), MMLU (Hendrycks et al., 2021a), and TruthfulQA (Lin et al., 2022) multiple-choice benchmarks and the MATH benchmark from (Hendrycks et al., 2021b).[7] In the forward-reverse experiments, we take the original generation, truncate a fraction of tokens, and check if resampling yields the same answer, using a direct text comparison for the multiple choice benchmarks and the prm800k grader for MATH (Lightman et al., 2023). We do this for 400 questions from each dataset and resample at each truncation fraction 100 times. Critical windows, defined as a > 0.5 jump in probability of obtaining the same answer in consecutive truncation fractions, appear prominently across all models and benchmarks that we tested (Figures 4, 8, and 9); for MATH, they occur in $42.2\%, 16.6\%, 30.2\%$ of generations from LLAMA-3.1-8B-Instruct, Qwen-2.5-7B-Instruct, and Phi-3-7B-Instruct.

[7]See Appendix G.2 for more results across models and datasets and a discussion on the effect of temperature on critical windows.

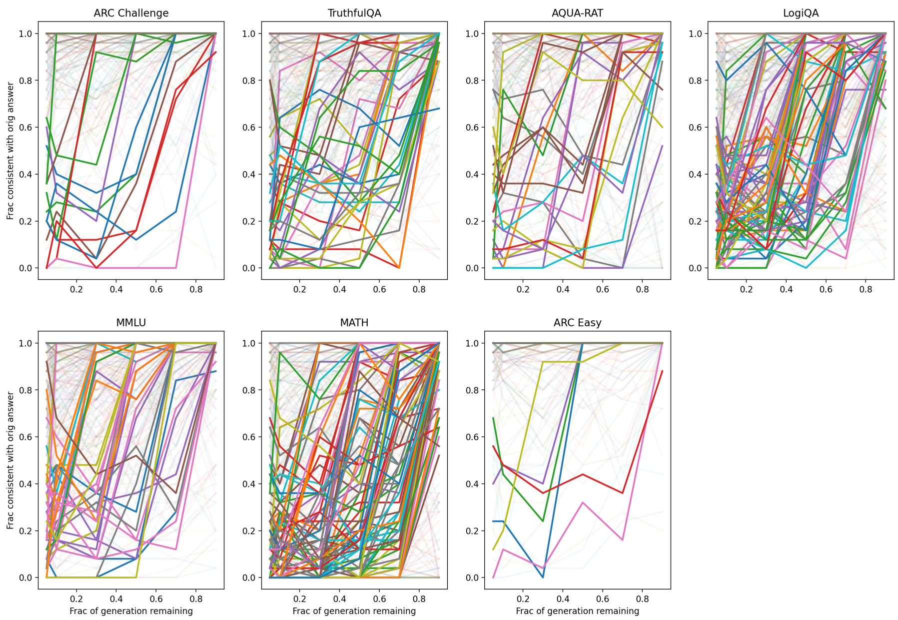

Figure 4: Probability that the answer is the same as a function of the percentage of the generation remaining for different math and reasoning benchmarks for LLAMA-3.1-8B-Instruct. Highlighted are generations with a 50% increase and no 30% decrease in the probability over subsequent fractions of sampling. We can see that the frequency of critical windows is especially high for the MATH benchmark.

These jumps occur during important steps of reasoning: in Figure 5, the correct formula is first expressed in the critical window. Furthermore, we find that generations with critical windows are significantly *less accurate* than generations without critical windows across all datasets and models. For LLAMA-3.1-8B-Instruct, critical windows result in up to **73% lower accuracy**, and incorrect responses contain **11%-33% more critical windows** (Table 1).

Table 1: Differences between Accuracy (Acc) without versus with critical windows and frequency of critical windows (CW) when the original generation is wrong versus right.

Dataset	**LLAMA-3.1-8B-Instruct**		**Phi-3-7B-Instruct**		**Qwen-2.5-7B-Instruct**	
	ΔAcc	**ΔCW**	**ΔAcc**	**ΔCW**	**ΔAcc**	**ΔCW**
AQUA-RAT	0.42	0.20	0.36	0.16	0.03	0.01
ARC Challenge	0.53	0.22	0.50	0.24	0.38	0.11
ARC Easy	0.73	0.26	0.28	0.13	0.40	0.07
LogiQA	0.15	0.11	0.21	0.19	0.23	0.11
MATH	0.41	0.33	0.36	0.33	0.46	0.29
MMLU	0.45	0.24	0.38	0.21	0.26	0.11
TruthfulQA	0.44	0.20	0.42	0.23	0.55	0.23

G EXPERIMENTAL DETAILS

G.1 JAILBREAK EXPERIMENTS

G.1.1 REPRODUCING CRITICAL WINDOWS FOR JAILBREAKS FROM EXISTING PAPERS

Existing work has already identified the presence of critical windows in the domain of jailbreaks. Here we present critical windows for a simplified prefill jailbreak based on the prefill attack (Haize

MATH (Phi-3-7B-Instruct)

Q: If the first term of an arithmetic sequence equals the third minus the second, and the fourth term is 8, what is the first term?
A: 1. We know that the arithmetic sequence has a common difference, which we will denote as "d". 2. Let's denote the terms in the sequence as T_1, T_2, T_3, and T_4 for the first, second, third, and fourth terms, respectively. 3. We're given the information that $T_1 = T_3 - T_2$. Using the general formula for an arithmetic sequence, we can write this **as $T_1 = T_2 + 2d - T_2$, which simplifies to $T_1 = 2d$. 4. We're also given that $T_4 = 8$. The general formula for an arithmetic sequence tells us that $T_4 = T_1 + 3d$. So we can write 8= $T_1 + 3d$.** 5. Now we have a system of two equations: 1) $T_1 = 2d$ 2) $8 = T_1 + 3d$ 6. We can substitute equation (1) into equation (2), which gives us $8 = 2d + 3d = 5d$. 7. Solving for d, we get $d = 8/5$. 8. Now that we know d, we can find T_1. Using the equation from step 1, we get $T_1 = 2d = 2*(8/5) = 16/5$. Therefore, the first term is $T_1 = 16/5$.

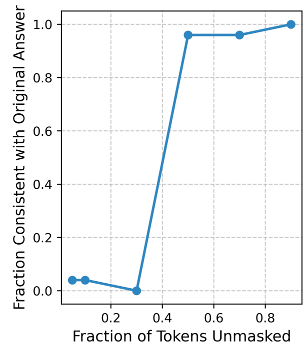

Figure 5: Example of critical window for Phi-3-7B-Instruct (bolded).

Labs, 2024b) and repeating token jailbreak (Nasr et al., 2023) for LLAMA-3.1-8B-Instruct. In the first figure, we plot the probability of the model giving a harmful response, computed using the `StrongReject` Gemma 7b auditor from (Souly et al., 2024), as a function of the fraction of the phrase `Sure, here is how to` appended to the front of the model's generation. We can see that there is a large jump in the attack success rate after only including a few tokens in the prefix. The second figure is a reproduction of Figure 12 from Nasr et al. (2023). It shows that the probability of repeating the next token increases substantially as the first few tokens are included a few times in the context.

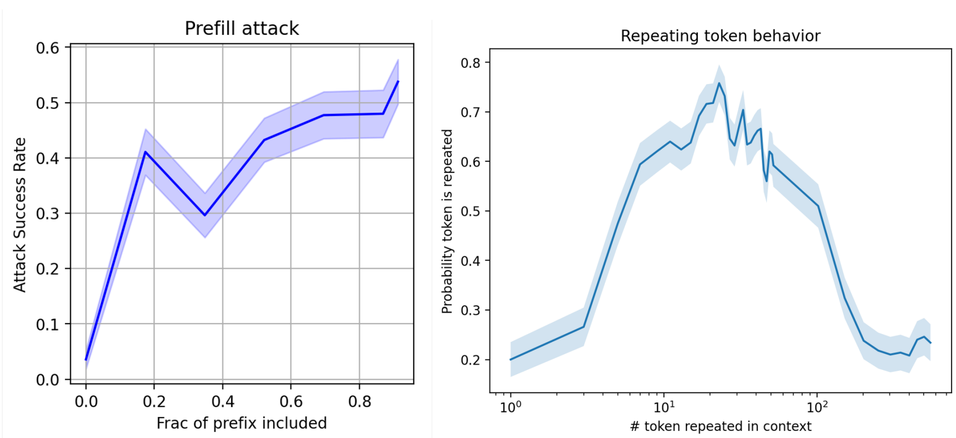

Figure 6: Examples of critical windows in jailbreaks for LLAMA-3.1-8B-Instruct. **Left:** Critical window for a prefill jailbreak (Haize Labs, 2024b). **Right:** Critical window for a repeating token jailbreak (Nasr et al., 2023).

G.1.2 Experimental details from jailbreak

Now we apply our theory to develop a new jailbreak detection method, based on a likelihood ratio between an aligned and unaligned model. Intuitively, our theory states that when the unaligned component assigns a high probability to the text compared to the entire model, the model is likely to be jailbroken. We use a LLAMA-3.1-8B-Instruct model jailbroken with LoRA to not refuse harmful prompts (grimjim, 2024) as a proxy for the unaligned model. We evaluate these different methods on a dataset of jailbreaks and benign prompts from (Bailey et al., 2024).

Dataset. We use the same dataset as (Bailey et al., 2024) but provide details here for completeness. The benign dataset consists of inputs from UltraChat (Ding et al., 2023), a large dialogue dataset, and Xstest (Röttger et al., 2024), which contains benign queries that are often incorrectly refused by language models. The benign queries are filtered to ensure that LLAMA-3.1-8B-Instruct does not refuse any of them. The dataset of harmful prompts is based off of the Circuit Breakers dataset (Zou et al., 2024). The datasets include the following jailbreaking methods from the extant literature: PAIR (Chao et al., 2023), AutoDAN (Liu et al., 2024), Many-Shot Jailbreaking (MSJ) (Anil et al., 2024), Multi-Turn Attacks (Li et al., 2024; Haize Labs, 2024a), Prefill, GCG (Zou et al., 2023), and other Misc. attakcs from (Wei et al., 2023). For each jailbreaking method, it is applied to a prompt from the Circuit Breaker dataset and evaluated to see if the generation from LLAMA-3.1-8B-Instruct is helpful and harmful, as determined by the StrongReject jailbreaking classifier (Souly et al., 2024)).

Evaluation Metric. As is standard in the jailbreak detection literature (Bailey et al., 2024), we report the recall at the false positive rate at 0.01.

Table 2 displays the recall and several other baselines. Crucially, the log likelihood ratio methods does obtain recall > 1 for 5 different categories of jailbreaks. While our methods do perform worse than existing methods, it is important to note that they still work and that their poor performance could be explained by the fact that we have to use a proxy for the unaligned mode of the model.

Table 2: Recall (FPR=0.01) for our likelihood ratio threshold, a perplexity threshold (Alon & Kamfonas, 2023), and a MLP-based detector trained on activations (Bailey et al., 2024) for predicting different jailbreaks. prompt/gen denote the logprobs of the prompt and generation, respectively.

	AutoDAN	**GCG**	**Multi-Turn**	**Misc**	**MSJ**	**Pair**	**Prefill**
$\log p_{\text{prompt}}^{\text{unaligned}} - \log p_{\text{prompt}}^{\text{aligned}}$	0.000	0.000	0.028	0.000	0.063	0.000	0.077
$\log p_{\text{gen}}^{\text{unaligned}} - \log p_{\text{gen}}^{\text{aligned}}$	0.082	0.030	0.000	0.100	0.000	0.061	0.051
$\log p_{\text{prompt}}^{\text{aligned}}$	0.000	0.576	0.056	0.063	0.013	0.000	0.077
$\log p_{\text{gen}}^{\text{aligned}}$	0.205	0.150	0.570	0.200	0.006	0.015	0.416
MLP	1.00	0.956	0.873	0.663	1.00	0.833	1.00

G.2 Chain of thought experiments

G.2.1 Experimental details

We describe the prompts we used in our evaluation of different datasets. The system prompt for all datasets is `Produce a correct solution to the following /TASK/ question.`, where `/TASK/` is the type of question of the dataset, i.e. science, math, or logic. For each question, we create a user prompt by appending `Think of the /TASK/ question thoroughly step by step. Please only respond with the answer after reasoning thoroughly.` in front of the question. Once the model completes its generation (max generation length set to 2048 and default sampling parameters), we append the user prompt `Given all of the above, what's the single, most likely answer? Your answer should have the format "The answer is ANSWER", where ANSWER is your answer.` for the multiple choice benchmarks and `Given all of the above, what's the single, most likely answer? Simplify it completely. Your answer should have the format "The answer is $ANSWER$", where ANSWER is your answer in LaTeX.` Note that when we ask the model for the final answer, we set the temperature to 0.

Across three different models (LLAMA-3.1-8B-Instruct, Phi-3-7B-Instruct, and Qwen-2.5-7B-Instruct), we find that conditioned on the occurrence of critical windows, the model generation is significantly less accurate compared to generations without critical windows.

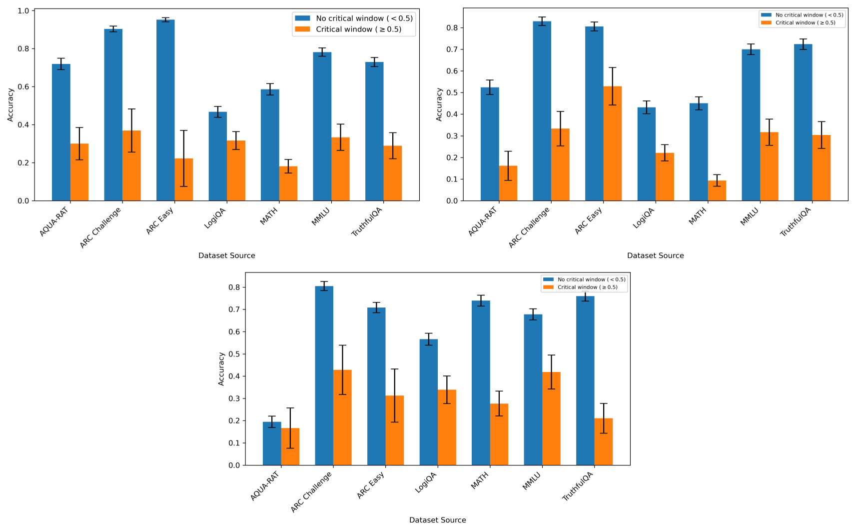

Figure 7: Left-to-right: LLAMA-3.1-8B-Instruct, Phi-3-7B-Instruct, and Qwen-2.5-7B-Instruct barplots of original generation accuracy for generations with (≥ 0.5 jump in probability) and without critical windows (< 0.5).

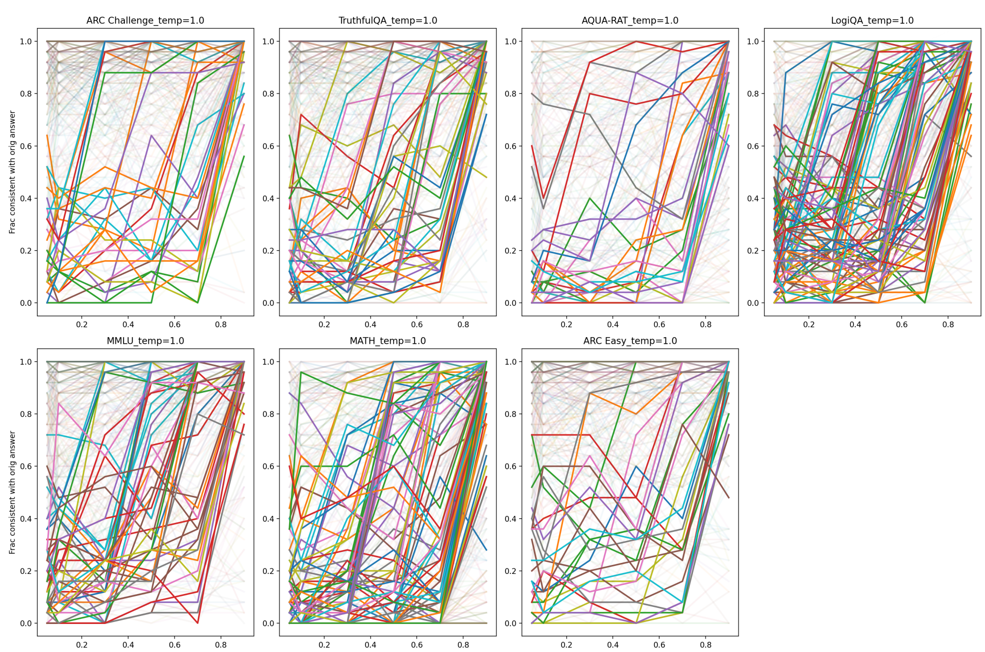

Figure 8: Probability that the answer is the same as a function of the percentage of the generation remaining for different math and reasoning benchmarks for Phi-3-7B-Instruct. Highlighted are generations with a 50% increase and no 30% decrease in the probability over one step.

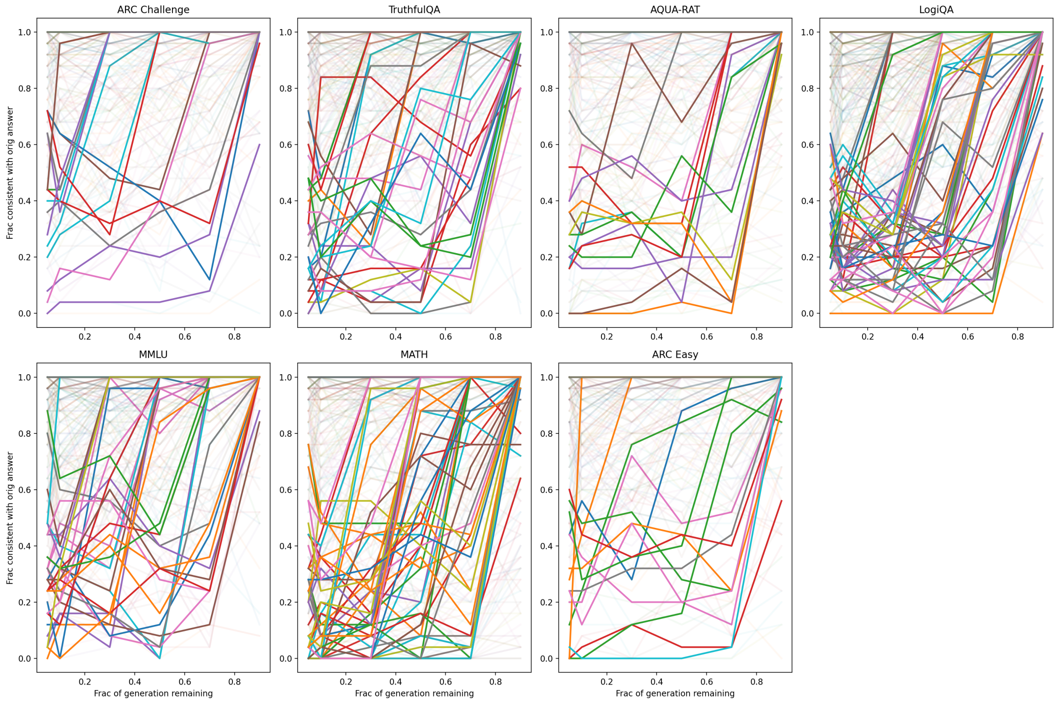

Figure 9: Probability that the answer is the same as a function of the percentage of the generation remaining for different math and reasoning benchmarks for Qwen-2.5-7B-Instruct. Highlighted are generations with a 50% increase and no 30% decrease in the probability over one step.

G.2.2 Comparison with directly answering for LLAMA-3.1-8B-Instruct

It is interesting to compare our approach in Figure 7 with that from Lanham et al. (2023), in which the model is directly asked for the answer (Figure 10). We can see that directly answering has a lower consistency than resampling across different amounts of generation remaining. Interestingly, the discrepancy is much higher for the math-datasets (MATH, AQUA) compared to extremely small difference in consistency for the other reasoning datasets. This may be because it is difficult to *extract* the answer from the middle of a chain of thought for math datasets than for other datasets.

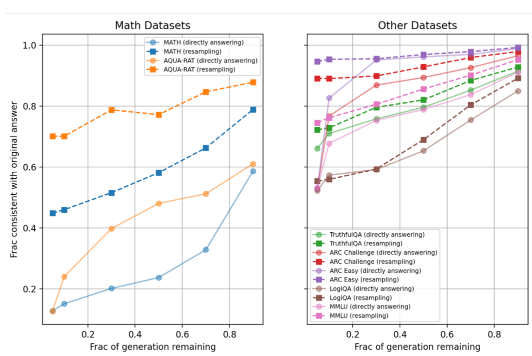

Figure 10: Difference between resampling to obtain the answer versus directly asking the model for the answer (as in (Lanham et al., 2023)) for LLAMA-3.1-8B-Instruct.

G.2.3 Examples of critical windows

We provide a few more examples of critical windows (bolded).

LogiQA example (LLAMA-3.1-8B-Instruct)

Q: Recently, some countries in the world, especially the United States and Japan, have blamed their domestic economic mystery on China's fixed exchange rate system between the RMB and the US dollar. Some domestic manufacturing companies have caused pressure. The following are all questions about the above views, except? (A) Japan has been in recession for more than ten years. At that time, China's merchandise exports were very small, and the renminbi was very weak. (B) The US economic recession was mainly caused by the destruction of the myth of the Internet economy, and the "9.11" terrorist attacks exacerbated this recession. (C) China's GDP accounts for only 3.9 of the world's total, foreign trade is less than 5 of the world's, and it does not have the ability to affect the world market price and supply and demand. (D) Consumers in Western countries have benefited greatly from cheap goods in China.

A: The question is asking which of the following is NOT a view about the views that the countries have blamed their domestic economic mystery on China's fixed exchange rate system between the RMB and the US dollar. We can **analyze each choice one by one. Choices A, B, and D are all views that the countries have blamed their domestic economic mystery on China's fixed exchange rate system between** the RMB and the US dollar. Choice C is an argument that China's fixed exchange rate system between the RMB and the US dollar has no effect on the world market price and supply and demand. So, it is not a view about the views that the countries have blamed their domestic economic mystery on China's fixed exchange rate system between the RMB and the US dollar. Therefore, the correct answer is C.

Correct answer: D

MATH example (LLAMA-3.1-8B-Instruct)

Q: Let a and b be the solutions of the equation $2x^2 - 10x + 5 = 0$. What is the value of $(2a - 3)(4b - 6)$?

A: To find the value of $(2a - 3)(4b - 6)$, we need to first find the values of a and b by solving the equation $2x^2 - 10x + 5 = 0$. We can use the quadratic formula to solve for a and b: $x = \frac{-b \pm \sqrt{b^2-4ac}}{2a}$. In this case, $a = 2$, $b = -10$, and $c = 5$. Plugging these values into the quadratic formula, we get: $x = \frac{-(-10) \pm \sqrt{(-10)^2-4(2)(5)}}{2(2)}$ $x = \frac{10 \pm \sqrt{100-40}}{4}$ $x = \frac{10 \pm \sqrt{60}}{4}$ $x = \frac{10 \pm 2\sqrt{15}}{4}$ So, the solutions are $a = \frac{10+2\sqrt{15}}{4}$ and $b = \frac{10-2\sqrt{15}}{4}$. Now that we have the values of a and b, we can substitute them into the expression $(2a - 3)(4b - 6)$: $(2a - 3)(4b - 6) = (2(\frac{10+2\sqrt{15}}{4}) - 3)(4(\frac{10-2\sqrt{15}}{4}) - 6)$ Simplifying the expression, we get: $(2a-3)(4b-6) = (\frac{10+2\sqrt{15}}{2} - 3)(\frac{10-2\sqrt{15}}{1} - 6)$ $(2a-3)(4b-6) = (\frac{10+2\sqrt{15}-6}{2})(\frac{10-2\sqrt{15}-6}{1})$ $(2a - 3)(4b - 6) = (\frac{4+2\sqrt{15}}{2})(\frac{4-2\sqrt{15}}{1})$ **Now, we can simplify the expression further by multiplying the numerators and denominators:** $\mathbf{(2a - 3)(4b - 6) = \frac{(4+2\sqrt{15})(4-2\sqrt{15})}{2}}$ **Using the difference of squares formula** $\mathbf{(a + b)(a - b) = a^2 - b^2}$**, we can simplify the numerator:** $\mathbf{(2a - 3)(4b - 6) = \frac{4^2-(2\sqrt{15})^2}{2}}$ $\mathbf{(2a - 3)(4b - 6) = \frac{16-60}{2}}$ Simplifying the expression further, we get: $(2a - 3)(4b - 6) = \frac{-44}{2}$ $(2a - 3)(4b - 6) = -22$ Therefore, the value of $(2a - 3)(4b - 6)$ is -22.

Correct answer: -22

ARC Easy example (LLAMA-3.1-8B-Instruct)

Q: Why are diseases transmitted through the bites of infected animals not considered contagious diseases? (A) The infected animals do not show symptoms. (B) The person in direct contact is the only one infected. (C) Humans do not become sick from animal diseases. (D) The diseases are treatable with antibiotics.

A: To answer this question, we need to understand the definitions of contagious and non-contagious diseases. Contagious diseases are those that are spread through direct or indirect contact with an infected person. Non-contagious diseases, on the other hand, are spread through other means, such as vectors like insects or animals. Step 1: Consider the definition of contagious diseases. Contagious diseases are spread through direct or indirect contact with an infected person. Step 2: Consider the definition of non-contagious diseases. Non-contagious diseases are spread through vectors like insects or animals. Step 3: Analyze the given options. Option A states that the infected animals do not show symptoms. This is not necessarily true, as some infected animals may show symptoms. Option B states that the person in direct **contact is the only one infected, which is not a characteristic of contagious diseases. Option D states that the diseases are treatable with antibiotics, which is not relevant to the mode of transmission. Step 4: Based on the analysis, the correct** answer is the one that accurately describes the mode of transmission of the diseases. The final answer is: $\boxed{C}$.

Correct answer: (B)

AQUA-RAT example (Phi-3-7B-Instruct)

Q: At a conference, one team is made up of 4 men and 4 women. Four presenters are chosen to present the team's findings in front of the entire conference. How many different groups of presenters can be chosen from the team if a team cannot be composed of men only or women only? (Two groups of presenters are considered different if at least one presenter is different.) A)120 B)19 C)180 D)420 E)460

A: 1. There are 8 people in the team, 4 men and 4 women. 2. We need to choose 4 presenters for the conference, without having an all-male or all-female group. 3. We can choose 4 out of 8 using combinations: C(8,4) = 8! / (4! * (8-4)!) = 70 4. We'll subtract two groups that aren't allowed: all men and all women groups. So we'll subtract combinations of just men (4C4) and just women (4C4). So, we have: In all, we need to choose from combinations of (8C4 - 4C4 - 4C4) which equals: 70 (total combination for 4 people) - 1 (all men group) - 1 (all women group) = 68 5. However, we should also consider the mixed group of 2 men and 2 women which would be another prohibited combination. Since the two men and two women groups would be the same, we subtract: C(4,2) * C(4,2) = 6 6. So, finally, our calculation will be: 68 - **6 = 62. The correct answer among the choices we have isn't amongst those. There seems to be a mistake in the question or the given options. But close to the choices, an answer of 420 could seem interconnected if we consider combinations when men and women are mixed in all possible ways which is not the deducted understanding in our solution. Correct answer: E**)420 should be revised or further clarified in the question context. Otherwise, based on our understanding of how combinations are applied in scenarios like this, none of these options is quite correct. The answer is E)420.

Correct answer: B

LogiQA example (Phi-3-7B-Instruct)

Q: The person in charge of the relevant department said that pilots will be carried out in some areas this year to prepare for the system and policy preparations for the comprehensive clean-up of ”small property houses”, and require all localities to confirm the registration and certification of rural collective land. All small-property houses are not subject to registration and are not protected by law. Therefore, this newly built house in Hexi Village is not protected by law. Which of the following is true will most weaken the above argument. (A) This newly built house in Hexi Village has been tacitly approved by relevant departments. (B) This newly built house in Hexi Village is a small property house. (C) This newly built house in Hexi Village is built on rural collective land. (D) Some of the newly built houses in Hexi Village are not small property houses.

A: The statement implies that small property houses are not protected by law and can be demolished if necessary in the preparations for the system and policy preparations for the comprehensive clean-up. -We have to find the option that weakens this argument the most. Option A suggests that the department has tacitly approved the newly built house. This weakens the argument, **but it can still be argued that the house is subject to demolition if the comprehensive clean-up policy is implemented. Option B states that this newly built house in Hexi Village is a small property house. This directly contradicts the argument that all small-property houses are not protected by law** and not subject to registration and certification. Therefore, this house would not be demolished and would be protected by law. Option C says that the house is built on rural collective land. This doesn’t disprove the argument – it might still be a small-property house without legal protection. Option D makes only some of the newly built houses in Hexi Village not small property houses. This doesn’t address the argument that all small-property houses are not protected by law and not subject to registration, but only carries an exceptional circumstance. Thus, the option that most weakens the argument is (B) as it directly contradicts the statement by implying that the specific house is protected by law. The answer is B.

Correct answer: D

MATH example (Qwen-2.5-7B-Instruct)

Q: Carson flips over the cards of a standard 52-card deck one at a time. What is the probability that he flips over the ace of spades before any face card (jack, queen or king)?

A: 1. First, we need to identify the relevant cards for our problem. The ace of spades (A♠) and the face cards (J, Q, K) are the ones we're focusing on. There are 4 face cards in a standard deck of 52 cards (12 face cards total, but we only consider 4 for this problem). 2. There are 52 cards in total, and 5 of these (A♠, J♠, Q♠, K♠, J♡, Q♡, K♡, J♢, Q♢, K♢, J♣, Q♣, K♣) are either the ace of spades or a face card. Therefore, there are 47 other cards that are neither the ace of spades nor face cards. 3. The problem can be simplified by considering the probability of the ace of spades coming before any face card in a sequence of drawing these 9 cards (the ace of spades and the **8 face cards). 4. We can think of the problem as a sequence of drawing 9 cards, where we are interested in the position of the ace of spades relative to the face cards. The ace of spades can come in any of the 9 positions, and we want to find the probability that it comes before any of the face cards. 5. For the ace of spades to come before any** face card, it must be in one of the first 5 positions (since there are 4 face cards, and the ace of spades must come before any of them). There are 9 positions in total, and 5 of these are favorable. 6. Therefore, the probability that the ace of spades comes before any face card is the ratio of the number of favorable outcomes to the total number of outcomes, which is $\frac{5}{9}$. So, the probability that Carson flips over the ace of spades before any face card is $\boxed{\frac{5}{9}}$.

Correct answer: $\frac{1}{13}$

AQUA-RAT example (Qwen-2.5-7B-Instruct)

Q: ABCDE is a regular pentagon with F at its center. How many different quadrilaterals can be formed by joining 4 of the points A,B,C,D,E and F? A)12 B)10 C)5 D)15 E)20

A: We need to form quadrilaterals by selecting 4 points from the 6 points (A, B, C, D, E, F) where F is the center of the regular pentagon. To form a quadrilateral, we need to select 4 points out of these 6. However, we need to ensure that the quadrilateral is valid. A key observation here is **that the center F must be included in the selection of 4 points because otherwise, we would only have points on the perimeter of the pentagon, which would not form a closed quadrilateral with a center point. So, we need to choose 3 points from the 5 points A**, B, C, D, E to form the quadrilateral along with F. The number of ways to choose 3 points from 5 is given by the combination formula $\binom{5}{3}$:

$$\binom{5}{3} = \frac{5!}{3!(5-3)!} = \frac{5 \times 4}{2 \times 1} = 10$$

Thus, there are 10 different ways to choose 3 points from the 5 perimeter points, and each choice, along with the center F, forms a valid quadrilateral. Therefore, the correct answer is $\boxed{10}$.

Correct answer: D

MATH example (Phi-3-7B-Instruct)

Q: Square $ABCD$ has side lengths of 13 units. Point E lies in the interior of the square such that $AE = 5$ units and $BE = 12$ units. What is the distance from E to side AD?

A: 1. Since $ABCD$ is a square, we know that all sides are equal and the diagonals bisect each other at right angles. 2. Drawing a diagram, we can see that $\triangle ABE$ and $\triangle BDE$ are right triangles with right angles at point B. 3. We can use Pythagoras' theorem to find DE: $DE = \sqrt{BE^2 - AE^2} = \sqrt{12^2 - 5^2} = \sqrt{144 - 25} = \sqrt{119} \approx 10.91$ units. 4. **Since $ABCD$ is a square, $AD = 13$ units. So, the distance from E to AD is** $13 - AE = 13 - 5 = 8$ units. The distance from E to side AD is 8 units. The answer is 8.

Correct answer: $\frac{25}{13}$

MMLU example (Qwen-2.5-7B-Instruct)

Q: A store owner noticed that an expensive electronic device was missing from the store. He assumed that a shoplifter had taken it. Later that day, the store owner was cleaning up the small room in the back of the store that the employees of the store used to store their personal items while they were at work. The store owner was straightening up the coat rack in the back room, when he accidentally knocked one of the employee's coats off the hanger. As the coat fell onto the floor, the missing electronic device fell out of the pocket of the coat. It was still in its original packaging, and the store owner checked the serial number on the package to confirm that it was the device missing from the store's shelves. The store owner knew that the coat belonged to a woman who worked as a cashier in the store. The store owner felt bad, but decided that he had no choice but to call the police. Two police officers arrived at the store. They gathered information from the store owner, who signed a criminal complaint against the woman. The officers then discussed the matter with the woman. Although she denied stealing anything, the officers arrested her and informed her that they would need to take her to the police station. After the police took the woman away, another employee of the store found out that the woman had been arrested. That employee immediately went to the store owner and admitted that he was actually the one who had stolen the device, and that he had hidden it in the woman's coat, intending to retrieve it later before the end of the woman's work shift. The store owner promptly called the police and told them that there had been a mistake and that the theft charge against the woman was invalid. The police released the woman, who spent a total of approximately 30 minutes in police custody. If the woman asserts a claim for false imprisonment against the arresting officers, will the woman prevail? (A) Yes, because the woman had not committed a crime at the time she was arrested. (B) Yes, because the woman was arrested and held in custody. (C) No, because the woman was detained for only 30 minutes. (D) No, because the officers had probable cause to arrest the woman based on the store owner's complaint and honest belief that the woman had stolen the device.

A: 1) The key issue here is whether the woman's arrest was lawful. 2) For an arrest to be lawful, there must be probable cause to believe the person committed a crime. 3) The store owner believed the woman had stolen the device based on the device falling out of her coat. **4) However, it was later revealed that the device was actually planted in her coat by another employee. 5) This means the store owner's belief was mistaken, and there was no probable cause for the arrest.** 6) False imprisonment occurs when a person is unlawfully detained. 7) The arrest was unlawful because it was based on a false belief. 8) Therefore, the woman's claim for false imprisonment is valid because she was unlawfully detained. Given these points, the correct answer is: (A) Yes, because the woman had not committed a crime at the time she was arrested. This answer is correct because the arrest was based on a mistaken belief and there was no probable cause, which means the arrest was unlawful, and the woman was falsely imprisoned. The answer is (A).

Correct answer: D

G.2.4 Varied temperature sampling

In our above experiments with critical windows, we have been generating model responses at their default sampling temperatures: LLAMA-3.1-8B-Instruct: 0.6, Phi-3-7B-Instruct: 1.0, and Qwen-2.5-7B-Instruct: 0.7. A natural question to ask is whether the sampling temperature affects the presence or strength of critical windows across these models. At one extreme, sampling at infinite temperature exhibits no critical windows since each token is generated uniformly at random. At the other extreme, zero temperature sampling is deterministic, so there is only one mode of behavior and critical windows again cannot exist. At intermediate temperatures, decreasing sampling temperature reduces token variance, so the sharpness of transition from general mixture to target submixture can be heightened. At the same time, reduced sampling variance also reduces the likelihood of transitioning between modes as well, leaving the overall effect on critical windows unclear.

First, we report the frequency of critical windows across the different temperatures for the MATH dataset: $0.1, 0.4, 0.7$, and 1.0. Then, we report the average size of a critical window on 400 samples from MATH, running the forward-reverse experiment on each sample 25 times and taking the max consecutive jump as the critical window size for that problem. From Figures 11 and 12, each model has its own relationship between temperature and critical window frequency and size, with no clear definite relationship across all models.

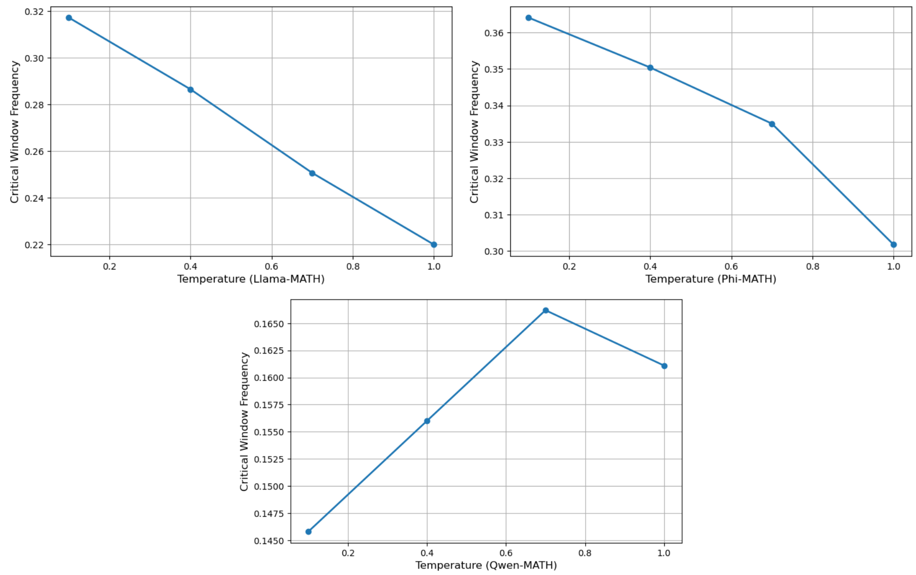

Figure 11: Temperature versus critical window frequency for LLAMA-3.1-8B-Instruct, Phi-3-7B-Instruct, and Qwen-2.5-7B-Instruct.

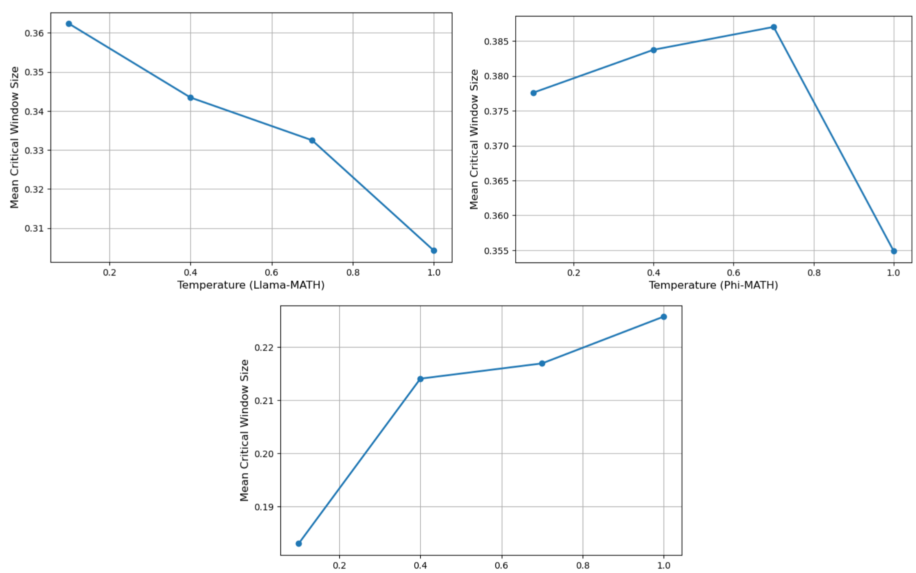

Figure 12: Temperature versus mean critical window size for LLAMA-3.1-8B-Instruct, Phi-3-7B-Instruct, and Qwen-2.5-7B-Instruct.