Why Decoding Sharpens Without Resolving: Lock-in and Confusion in Autoregressive Reasoning

Nan Zhang
Mendoza College of Business
University of Notre Dame
nan.zhang@nd.edu

Heng Xu
Mendoza College of Business
University of Notre Dame
heng.xu.dr@nd.edu

Abstract

Autoregressive reasoning traces often sharpen as they unfold—entropy falls and token probabilities concentrate—suggesting decoding is resolving uncertainty. Yet the same pattern persists after clear reasoning errors. We analyze decoding through the conditional posterior it induces over *branches*—equivalence classes of continuations under the chosen notion of "resolution"—and show that a mismatch between token selection and branch resolution produces two regimes: *lock-in*, where top tokens remain branch-evidential and reinforce commitment; and *confusion*, where branch-specific residuals self-average and top tokens become weakly diagnostic. Two observable consequences follow: entropy may fall while unresolved support stays elevated, and greedy token selection has opposite informational effects across regimes. We confirm these theoretical predictions across multiple open-weight models and reasoning benchmarks.

1 Introduction

Why do reasoning models often fail while sounding increasingly certain? Autoregressive traces typically grow sharper as they unfold: entropy falls and token probabilities concentrate [3], often interpreted as the model resolving uncertainty and locking onto a continuation [27, 28]. Yet some of the clearest reasoning failures—wrong arithmetic, incorrect counts, flawed multi-step derivations—show the same pattern [5, 6, 26]. What does token-level sharpness actually reveal during reasoning?

We show a different picture: decoding can sharpen without resolving among competing continuations. If multiple branches, each an equivalence class of continuations under a chosen notion of resolution, remain compatible with the current prefix [11, 18, 23, 33], standard autoregressive decoding is structurally misaligned with resolving them: the emitted token is chosen by posterior-averaged next-token probability, whereas branch resolution depends on token diagnosticity. As a result, decoding tends to enter two regimes indexed by the Rényi-2 effective support $N_{\mathrm{eff}}(\pi_t)$ of the branch posterior. When $N_{\mathrm{eff}}(\pi_t)$ is small, posterior mass is concentrated, high-probability tokens retain branch-specific signal, and Bayes updates further strengthen concentration: the *lock-in* regime. When $N_{\mathrm{eff}}(\pi_t)$ is large, branch-specific residuals self-average under posterior weighting, the next-token law is dominated by a consensus component (which may also be sharp), top tokens become weakly diagnostic, and the posterior remains diffuse: the *confusion* regime. Reasoning failure need not arise from instability. It can also arise from a stability in which decoding remains locally coherent without resolving the competing continuations.

This yields two theoretical predictions. First, greedy decoding should have opposite informational effects across regimes: in lock-in, the top token is evidential for the dominant branch and reinforces it; in confusion, the top token is a consensus continuation with weak branch-specific signal and suppresses branch-resolving alternatives. Equivalently, perturbing away from greedy should have inverse-signed effects. Second, since both regimes can exhibit surface-level sharpening, scalar

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

uncertainty alone cannot distinguish them: entropy may fall in both. The relevant question is thus not only whether the next-token law is sharp, but what kind of regime that sharpness reflects.

We test our theoretical predictions over multiple open-weight models and reasoning benchmarks. Across all cases, entropy sharpens broadly while a residual-support statistic separates lock-in- from confusion-like traces. Intervention tests further show that the same alternative-token perturbation has opposite downstream effects: increasing divergence in lock-in but decreasing it in confusion.

2 Related work

Prior work uses scalar uncertainty for self-knowledge, calibration, epistemic/aleatoric decomposition, and entropy-based decoding [1, 2, 5, 10, 16, 29, 32]. We show that scalar sharpness cannot capture branch-level posterior geometry. A second line aggregates multiple generations via self-consistency or semantic entropy [8, 17, 27], and process-oriented work analyzes intermediate reasoning through reward models, volatility, trajectory probes, faithfulness, etc. [3, 4, 6, 18, 20, 23, 26, 33]. Our object is different: we study the evolution of the next-token law within a single decoding trace, requiring neither process supervision nor hidden-state access. Our Bayesian-filtering formulation is adjacent to posterior-sampling analyses of LLM behavior [7, 13, 19, 30], but targets posterior motion governing branch resolution during ordinary autoregressive reasoning. App. A gives a more detailed comparison.

3 Setup

Let $x_{<t} = (x_1, \ldots, x_{t-1})$ denote the decoded prefix at step t, and let $P \in \{p_i\}_{i \in I}$ index competing branches. Branches are equivalence classes of continuations that are indistinguishable with respect to the target distinction one *cares to resolve*: two continuations belong to the same branch exactly when they are indistinguishable for the notion of *resolution* under study[1]. For example, when a problem statement admits two readings, one may care to resolve which reading the trace adopts. In this case, we would have two branches, each consisting of those continuations that adopt the corresponding reading (continuations that adopt neither form a residual class). Note that our analysis does not posit a unique branch configuration. Rather, it characterizes how autoregressive sharpness relates to resolution given the branch configuration induced by the chosen definition of that resolution.

Define the branch posterior $\pi_t(i) := \Pr(P = p_i \mid x_{<t})$ and the branch-conditional next-token law $q_{i,t}(x) := \Pr(x_t = x \mid P = p_i, x_{<t})$, both under the model's joint distribution over completions. Let $q_t(x) := \Pr(x_t = x \mid x_{<t}) = \sum_{i \in I} \pi_t(i)\, q_{i,t}(x)$ denote the model's next-token distribution before test-time modification. Note that, conditional on the value of $x_{<t}$, q_t is invariant to *how* $x_{<t}$ has been generated. In other words, whether it was sampled, greedy-decoded, or arrived by any other route, the same prefix value always yields the same q_t. The model is thus a Bayesian filter whose observable next-token law is a posterior-weighted average of branch-conditional laws, and after observing x_t the posterior over branches updates by Bayes' rule:

$$\pi_{t+1}(i) = \Pr(P = p_i \mid x_{\leq t}) = \frac{\pi_t(i)\, q_{i,t}(x_t)}{q_t(x_t)}. \tag{1}$$

This is the model's posterior given the realized prefix and applies under any decoding policy that produces x_t. $q_{i,t}(x)/q_t(x)$ is the posterior reweighting factor for observing x under p_i: a token is diagnostic to the extent that its branch-conditional probability differs from the global next-token law.

Note that the above Bayesian-filter view is a statement about the language model's output distribution and *not* about its internal computation. Whatever the model's weights and architecture, or whether it represents branches internally at all, (1) always holds because it is simply the chain rule applied to a partition of the model's own law over completions, of which π_t and $q_{i,t}$ are conditional probabilities.

Conventions. App. B gives the formal measure-theoretic setup and details what the framework assumes (versus derives). A brief summary: We assume I finite, P measurable, and π_0 of full support on I. Throughout the paper, we work with the live-branch set $\{i : \pi_t(i) > 0\}$ at each step. Bayesian updates are evaluated at emitted tokens, for which $q_t(x_t) > 0$ by construction. Ties in greedy decoding are broken by an arbitrary fixed rule.

[1] As discussed in App. B, the same decoded trace may be in lock-in relative to a coarse resolution notion while in confusion relative to a finer one. Our theoretical results apply to both granularities simultaneously.

4 The selection-resolution mismatch

The key mechanism we identify is a mismatch between *token selection* and *branch resolution*. Decoding follows q_t, a posterior-weighted average of branch-conditional laws, whereas branch resolution depends instead on how informative the realized token is relative to competing branches. The mismatch is especially relevant in reasoning, where distinct live branches often share a local backbone of highly probable continuations: discourse connectives, syntactic completions, and other generic steps that preserve coherence without resolving which branch is correct [11].

To make this explicit, let the dominant branch $i_t^\star := \arg\max_i \pi_t(i)$, $\pi_t^\star := \pi_t(i_t^\star)$, $q_t^\star := q_{i_t^\star,t}$, the competitor mixture $\bar{q}_t(x) := (1-\pi_t^\star)^{-1} \sum_{j \neq i_t^\star} \pi_t(j)\, q_{j,t}(x)$, and $L_t(x) := q_t^\star(x)/\bar{q}_t(x)$. We have

$$q_t(x) = \bar{q}_t(x)\Big[(1-\pi_t^\star) + \pi_t^\star L_t(x)\Big]. \tag{2}$$

Decoding maximizes this product, whereas posterior movement is governed by $L_t(x)$: token selection follows a consensus-weighted score, branch resolution a likelihood ratio. When $\pi_t^\star \approx 1$, the score is sensitive to $L_t(x)$ and tokens favoring the dominant branch tend to win decoding. When $\pi_t^\star$ is small, however, $(1-\pi_t^\star) + \pi_t^\star L_t(x) \approx 1$ (unless $L_t(x)$ is very large), so maximizing q_t approximately maximizes $\bar{q}_t$. Tokens with strong branch-specific evidence typically have low competitor support and lose. Tokens that win derive their probability from cross-branch agreement and induce weak posterior movement. The confusion regime arises when this mismatch becomes systematic.

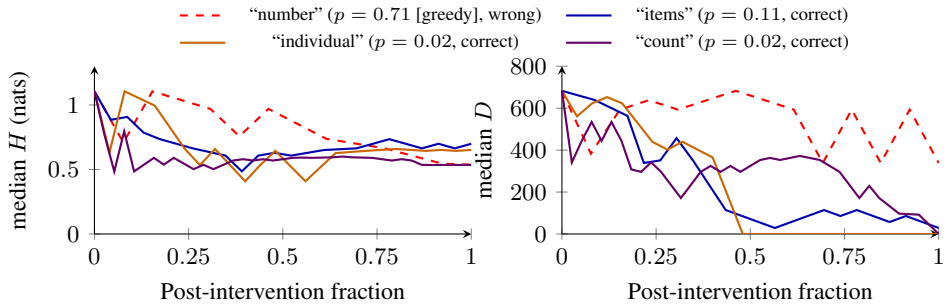

Figure 1: **Confusion example.** Mistral-7B on BBH-OC #9: median H (left) and test statistic D (§7, right), x-axis normalized to $[0, 1]$. Dashed red: greedy ("number"). Solid blue/orange/violet: alternatives ("items"/"individual"/"count"). All three off-greedy alternatives fix the greedy error.

Figure 1 illustrates this on a BBH object-counting trace from our experiments (§8). After the prefix *"To count the number of objects you have, we simply add up the ___"*, greedy selects *number* ($p = 0.71$), a locally fluent echo of the prefix that is weakly diagnostic about how to count. Alternatives *items* ($p = 0.11$), *individual* ($p = 0.02$), and *count* ($p = 0.02$) commit to more specific paths. Greedy ultimately answers 9, while all three alternatives recover the correct 10. The qualitative lesson is consistent with confusion: a token can be locally preferred because it is broadly compatible with unresolved continuations, while lower-ranked alternatives can be more branch-revealing.

The figure also shows that entropy falls along all four continuations, including the incorrect greedy one (as the figure also foreshadows, the later-defined residual-support statistic D stays elevated on the greedy branch but collapses on the corrected alternatives). Why does the greedy trace stay sharp in confusion? Define the posterior-independent consensus law (equivalently, one under the most diffuse, i.e., uniform, posterior) $c_t(x) := \sum_{i \in I} q_{i,t}(x)/|I|$ and residuals $r_{i,t}(x) := q_{i,t}(x) - c_t(x)$, giving

$$q_t(x) = c_t(x) + \sum_i \pi_t(i)\, r_{i,t}(x). \tag{3}$$

In confusion, posterior averaging largely cancels the residuals across branches, so the head of q_t is dominated by c_t. In lock-in, q_t is dominated by the high-posterior branch and is close to $q_t^\star$. Since c_t and $q_t^\star$ can each be sharp or diffuse, low entropy of q_t does not reveal whether decoding has concentrated on a coherent branch or is merely following a sharp shared consensus.

We can also quantify *when* branch-specific signal survives posterior averaging. Define

$$\eta_t := \left\| \sum_i \pi_t(i) r_{i,t} \right\|_2 = \sqrt{\pi_t^\top G_t \pi_t} \le \sqrt{\|G_t\|_2} \|\pi_t\|_2 = \sqrt{\frac{\|G_t\|_2}{N_{\text{eff}}(\pi_t)}}, \tag{4}$$

where Gram matrix $(G_t)_{ij} := \sum_x r_{i,t}(x)\, r_{j,t}(x)$, and $N_{\text{eff}}(\pi_t) := 1/\sum_i \pi_t(i)^2$ is the Rényi-2 effective support. Note that $\eta_t = 0$ exactly when π_t is uniform. More generally, $N_{\text{eff}}(\pi_t)$ is the scalar control conditional on G_t: small N_{eff} permits branch-specific signals to survive (*lock-in*), while large N_{eff} attenuates such signals via (4), making q_t consensus-dominated (*confusion*). The next section gives sufficient conditions for each regime to locally self-reinforce.

5 Persistence of lock-in and confusion

Our first result is an accounting identity explaining the role of $N_{\text{eff}}(\pi_t)$. The regime-specific mechanisms are then supplied by two propositions built upon it. Throughout this section, all expectations, probabilities, and posterior quantities are conditional on the prefix $x_{<t}$ at the step under analysis (we suppress the conditioning in notation for readability). Proofs are in the appendix.

Lemma 1 (Collision accounting identity)**.** *Let $C_t := 1/N_{\text{eff}}(\pi_t) = \sum_i \pi_t(i)^2$ denote the branch collision probability. Then, for every $T \ge 1$,*

$$\mathbb{E}[C_{t+T} - C_t] = \sum_{s=t}^{t+T-1} \mathbb{E}\big[\|\pi_{s+1} - \pi_s\|_2^2\big] + 2 \sum_{s=t}^{t+T-1} \mathbb{E}[\langle \pi_s, \mathbb{E}[\pi_{s+1} - \pi_s] \rangle]\,. \tag{5}$$

The second term measures alignment between the current posterior and its expected one-step movement. Under temperature-1 sampling, the posterior is a martingale, so it vanishes and

$$\mathbb{E}[C_{t+T} - C_t] = \sum_{s=t}^{t+T-1} \mathbb{E}\big[\|\pi_{s+1} - \pi_s\|_2^2\big] = \sum_{s=t}^{t+T-1} \mathbb{E}\left[\sum_i \pi_s(i)^2\, \chi^2(q_{i,s} \| q_s) \right]. \tag{6}$$

The identity singles out (the inverse of) N_{eff} as a bookkeeping device. Changes in collision reflect accumulated squared-ℓ_2 posterior motion plus an alignment term recording whether expected motion points along the current posterior (lock-in-like) or fails to concentrate (confusion-like). Even under temperature-1 sampling, where the alignment term vanishes, collision is an exact accounting of accumulated motion. Thus trajectories that resolve branch uncertainty must accumulate collision and reduce N_{eff}. If N_{eff} remains elevated, either motion is limited or its direction fails to concentrate.

While the lemma does not by itself specify the two regimes, it does clarify why intermediate support may be unstable: When π_t is split among competing branches with sharply different next-token laws, the drift $V_t := \sum_i \pi_t(i)^2\, \chi^2(q_{i,t} \| q_t)$ in (6) is large, rapidly collapsing the posterior toward lock-in. An intermediate-support state can fail to resolve only if $\chi^2(q_{i,t} \| q_t)$ is small across branches—precisely the consensus-dominated condition defining confusion. The region between confusion and lock-in is therefore a transient resolution window between the two regimes. With Lemma 1 showing the separation of the two regimes, the next two propositions supply their mechanisms:

Proposition 1 (Lock-in reinforcement and limited switching)**.** *For temperature-1 sampling when $N_{\text{eff}}(\pi_t) \le K < 2$, the expected number of dominant-branch switches after step t is bounded by*

$$\mathbb{E}\left[\sum_{s \ge t} \mathbf{1}\big\{ N_{\text{eff}}(\pi_s) \le K,\ i^\star_{s+1} \ne i^\star_s \big\} \,\middle|\, x_{<t} \right] \ \le\ \frac{2(K-1)}{2-K}. \tag{7}$$

More generally, if $1 + \tau \le N_{\text{eff}}(\pi_t) \le K$ and the emitted token satisfies $L_t(x) \ge 1 + \varepsilon$, then

$$\pi_{t+1}(i^\star_t) - \pi^\star_t \ \ge\ \frac{\varepsilon}{1 + \varepsilon/\sqrt{1+\tau}} \cdot \frac{1}{K} \left(1 - \frac{1}{\sqrt{1+\tau}} \right). \tag{8}$$

This rules out dominant-branch switch when $L_t(x) \ge 1 + \varepsilon$ and $K < 2$.

Proposition 1 depicts the lock-in picture. A key observation is that, under temperature-1, dominant-branch switches are budget-limited because each switch consumes squared posterior motion, while

Lemma 1 supplies only a finite total budget. For example, in a two-branch state with one at 90.8% mass, the bound implies $\geq 50\%$ probability of no further switch while in the low-N_{eff} region. More broadly, any token with evidential margin $L_t(x_t) \geq 1 + \varepsilon$ gives a uniformly positive increase in incumbent-branch mass and cannot cause a dominant-branch switch when $N_{\text{eff}} < 2$.

Proposition 2 (Confusion). *For any branch i, the posterior update from emitting a token x satisfies*

$$\pi_{t+1}(i) - \pi_t(i) \leq \pi_t(i) \frac{r_{i,t}(x) + \eta_t}{c_t(x) - \eta_t}, \text{ where } \eta_t \leq \sqrt{\frac{\|G_t\|_2}{N_{\text{eff}}(\pi_t)}}, \tag{9}$$

when $N_{\text{eff}}(\pi_t) > \|G_t\|_2 / c_t(x)^2$. Under greedy decoding, the selected token always satisfies

$$c_t(x) \geq \max_y c_t(y) - \sqrt{2}\eta_t, \text{ and } L_t(x) \leq \frac{1 - \pi_t^\star}{\max_y c_t(y) - \eta_t - \pi_t^\star} \tag{10}$$

when $\max_y c_t(y) - \eta_t > \pi_t^\star$, and the steps taken for branch i to reach $\pi(i) \geq p_0 > 1/2$ satisfies $\tau_{p_0}^i \geq \lceil \log(\pi_t(i)/p_0) / \log(\underline{c} - \overline{\eta}) \rceil$ if $\max_x c_t(x)$ stays above $\underline{c}$, η_t stays below $\overline{\eta}$, and $0 < \underline{c} - \overline{\eta} < 1$. Under temperature-1 sampling, the probability for the dominant branch with $\pi_t(i) < p_0$ to reach $\pi(i) \geq p_0$ without switching satisfies

$$\Pr\left(\begin{array}{l} \pi_s(i_t^\star) \geq p_0 \\ \text{before switch} \end{array} \middle| x_{<t} \right) \leq \frac{\pi_t^\star - \pi_t^\sharp + \zeta}{2p_0 - 1 + \zeta}, \quad \zeta := \operatorname*{ess\,sup}_{s \geq t} \frac{\max_{i,j} |r_{i,s}(x_s) - r_{j,s}(x_s)|}{q_s(x_s)}, \tag{11}$$

where $\pi_t^\sharp$ is the second largest posterior mass at step t, and ess sup *represents the essential supremum.*

A key distinction between the two regimes is surfaced by (10): In confusion when $N_{\text{eff}}(\pi_t)$ is large and η_t is small, the greedy token *has to* have a high consensus mass. In other words, tokens that are highly evidential for one branch (e.g., $L_t(x) \gg 1$) cannot win token selection. For illustration, consider a perfectly dominant-branch-exclusive token x with $q_t^\star(x) = 1$ and $\bar{q}_t(x) = 0$. Then, under greedy decoding, x is guaranteed to be emitted whenever $N_{\text{eff}}(\pi_t) < 2$ and guaranteed *not* to be emitted whenever $N_{\text{eff}}(\pi_t) > 2\|G_t\|_2 / (\max_y c_t(y) - 1/|I|)^2$, provided $\max_y c_t(y) > 1/|I|$.

In confusion, a sharp next-token law can be bad news: the posterior movement bound (9) shrinks with large consensus mass $c_t(x)$, and the greedy hitting time grows as the consensus head sharpens (large $\underline{c}$). Under temperature-1, larger $q_s(x_s)$ shrinks ζ, tightening the bound on the incumbent reaching high posterior mass. Sharpness can actively slow resolution rather than certify it.

Together, the three results in this section isolate the two regimes: Lemma 1 gives the dynamic accounting, Proposition 1 the mechanism of concentration-producing updates (lock-in), and Proposition 2 how diffuse support can be locally sharp yet remain globally unresolved (confusion). What distinguishes the two regimes is therefore *not* the sharpness of q_t but whether the head tokens are branch-evidential or branch-agnostic.

6 Regime-dependent token diagnosticity

Having studied the regime-level question in the preceding section, we now ask the token-level question: does choosing a token x at step t concentrate, leave unchanged, or flatten the posterior? Let π_t^x be the posterior after conditioning on x and define

$$\Lambda_t(x) := \frac{N_{\text{eff}}(\pi_t) - 1}{N_{\text{eff}}(\pi_t^x) - 1}. \tag{12}$$

$\Lambda_t(x) > 1$ (resp. < 1) means x reduces (resp. expands) non-singleton branch support (and ≈ 1 is neutral). The -1 matters in lock-in: once near a singleton, what counts is residual support above the singleton floor. The next proposition gives an interesting recursive account: token diagnosticity is not solely determined by whether x separates the dominant branch from the rest—in confusion, the same question is passed to the residual mixture after peeling the dominant branch.

Proposition 3 (Regime-dependent token diagnosticity). *Let $S_0 = I$. At step t, recursively define*

$$\pi_t^k(i) := \frac{\pi_t(i)\mathbf{1}\{i \in S_k\}}{\sum_{j \in S_k} \pi_t(j)}, \; i_k^\star := \arg\max_{i \in S_k} \pi_t^k(i), \; S_{k+1} := S_k \setminus \{i_k^\star\}, \; u_k := \frac{1 - \pi_t^k(i_k^\star)}{\pi_t^k(i_k^\star)}, \tag{13}$$

$$\Lambda_t^k(x) := \frac{N_{\text{eff}}(\pi_t^k) - 1}{N_{\text{eff}}(\pi_t^{k,x}) - 1}, \; \bar{q}_k(x) := \frac{\sum_{j \in S_k \setminus \{i_k^\star\}} \pi_t^k(j)\, q_{j,t}(x)}{\sum_{j \in S_k \setminus \{i_k^\star\}} \pi_t^k(j)}, \; L_k(x) := \frac{q_{i_k^\star,t}(x)}{\bar{q}_k(x)}. \tag{14}$$

where $\pi_t^{k,x}$ represents the posterior π_t^k after observing x. Let

$$\alpha_k := \frac{1}{1 + u_k^2/N_{\text{eff}}(\pi_t^{k+1})},\ \beta_k := \frac{2}{2 + u_k(1 - 1/N_{\text{eff}}(\pi_t^{k+1}))}. \tag{15}$$

Then, whenever $k < |I| - 2$, $\Lambda_t^k(x)$ admits the recursive decomposition

$$\boxed{\Lambda_t^k(x) = \frac{\alpha_k L_k(x)^2 + (1-\alpha_k) G_k(x) \Lambda_t^{k+1}(x)}{\beta_k L_k(x) + (1-\beta_k) G_k(x)},} \tag{16}$$

where $G_k(x) = (1 - 1/N_{\text{eff}}(\pi_t^{k+1,x}))/(1 - 1/N_{\text{eff}}(\pi_t^{k+1}))$. The recursion terminates at

$$\Lambda_t^{|I|-2}(x) = \frac{L_{|I|-2}(x)^2 + u_{|I|-2}^2}{L_{|I|-2}(x)(1 + u_{|I|-2}^2)}. \tag{17}$$

The recursion gives a token-level characterization of the regimes. In lock-in, u_0 is small, $\alpha_0, \beta_0 \approx 1$ and $\Lambda_t(x) \approx L_0(x)$: diagnosticity is the Bayes factor separating the dominant branch from the residual. In confusion, u_0 is large, $\alpha_0, \beta_0 \approx 0$ and $\Lambda_t(x) \approx \Lambda_t^1(x)$, passing the question down the peeling hierarchy (with attenuation) until the residual set of branches has a dominant one. This recursive structure is important because it formalizes the intuition in (2): compared with lock-in, $L_0(x)$ is far less relevant in confusion because it enters the peeling layer with only a small weight. So even if the selected token happens to be diagnostic for the dominant branch, its final diagnosticity $\Lambda_t(x)$ is still a diluted aggregate of $L_0, L_1, \ldots$ across layers. Note that this recursion is algebraic and holds for any branch ordering: App. G gives the order-agnostic form and uses it to distinguish credible competitors from diffuse low-collision dust, refining the sufficient condition for lock-in-like diagnosticity beyond the majority-mass case (i.e., $N_{\text{eff}} < 2$ in Proposition 1).

This proposition also separates greedy from alternative tokens. In lock-in, $L_0(x)$ leads, so greedy reinforces the incumbent. In confusion, greedy typically selects a token with a large c_t, which keeps $L_k(x) \approx 1$ across peel layers and returns $\Lambda_t(x) \approx 1$. An off-greedy token with smaller c_t is thus riskier but potentially more informative: it can yield $L_k(x) < 1$ for some layers and $L_k(x) \gg 1$ for others, giving $\Lambda_t(x)$ at least a chance to be away from 1. Thus, in confusion, greedy is locally safe but often non-resolving, while alternatives offer the only *chance* (not guarantee) of a resolving update.

7 Observable bridge

While our analysis is interpretive in spirit, its finding also reveals potential practical interest: decoding has regime-dependent effects on posterior motion, so identifying regime from a single trace's q_t (i.e., within-trace, complementary to cross-sample diagnostics [17, 27]) could enable regime-aware control of decoding. Since π_t depends on the notion of resolution, no statistic of q_t alone can be a pointwise estimator of $N_{\text{eff}}(\pi_t)$. Fortunately, to test our theory, we only need a within-configuration discriminator that separates the regime extremes. While $N_{\text{eff}}(q_t)$ looks like a straightforward choice, it cannot serve this purpose because, in confusion, the shared backbone can be sharp and resemble a dominant branch (App. H). What could work is to disentangle this shared overlap and measure the effective support of the normalized remainder. The following proposition provides such a solution.

Proposition 4 (Residual-kernel bridge)**.** *Suppose that q_t admits a decomposition $q_t = \alpha_t h_t + (1 - \alpha_t)\rho_t = \alpha_t h_t + (1 - \alpha_t)\sum_i \tilde{\pi}_t(i)\, \rho_{i,t}$ where h_t, $\tilde{\pi}_t$, and $\rho_{i,t}$ are all probability vectors. Define the residual Gram kernel $(K_t)_{ij} := \sum_x \rho_{i,t}(x)\rho_{j,t}(x)$ and $D_t := N_{\text{eff}}(\rho_t) = 1/(\tilde{\pi}_t^\top K_t \tilde{\pi}_t)$. Then*

$$\frac{N_{\text{eff}}(\tilde{\pi}_t)}{\lambda_{\max}(K_t)} \le D_t \le \frac{N_{\text{eff}}(\tilde{\pi}_t)}{\bar{\xi}_t}. \tag{18}$$

where $\lambda_{\max}(K_t)$ is the largest eigenvalue of K_t and $\bar{\xi}_t := \sum_i (\tilde{\pi}_t(i)^2 (K_t)_{ii}) / \sum_i \tilde{\pi}_t(i)^2$ is the $\tilde{\pi}_t(i)^2$-weighted average of the within-branch collision probability $\operatorname{diag}(K_t)$.

Proposition 4 offers D_t as an observable bridge: After residualization by h_t, $N_{\text{eff}}(\tilde{\pi}_t)$ is the effective number of branches that remain live, while K_t captures the geometry of their residual branch laws.

In confusion, the shared lexical backbone is pushed into h_t, leaving a residual mixture with many live branches and limited spectral redundancy (i.e., small $\lambda_{\max}(K_t)$), so (18) forces D_t to remain large. In lock-in, by contrast, h_t absorbs the dominant branch, meaning that the remaining branches still retain their lexical backbone, which drives up within-branch collision and leaves D_t small.

In sum, as we will empirically instantiate in the next section, D_t separates the two regimes insofar as the chosen notion of "resolution" (§3) induces branches that don't strongly overlap beyond the shared backbone. Situations violating this bridge condition are outside D_t's scope. For example, when sibling branches overlap beyond the backbone, D_t will follow the coarser, bridge-compatible structure and can read lock-in while the finer posterior is still confused (see App. M).

8 Experiments

Table 1: Predicted empirical signatures and observed results.

Theoretical prediction	**Predicted observable signature**	**Observed**
Both regimes can produce sharpening, so H alone cannot resolve them.	H is not reliably lower on C than W traces; ΔH after intervention cannot reliably separate W→C and W→W.	Tab. 2: H reverses across cells. Tab. 3: ΔH reverses across cells. Fig. 2 (top): Intertwined H bands.
Off-greedy intervention might rescue confusion wrongs by turning it to lock-in corrects.	ΔD direction: W→C more negative than W→W (CIs need not exclude zero: fix could be lock-in→lock-in).	Tab. 3: All $5/5$ models W→C $<$ W→W; aggregate W→C med $\Delta D = -2.1$ vs. W→W $+0.7$.
Compared with lock-in, confusion is more recoverable by off-greedy interventions.	$D\uparrow$ higher fix rate than $D\downarrow$; $D\uparrow$/Fixed med $\Delta D \ll 0$.	Tab. 4: fix rate 15% for $D\uparrow$ (net $+41$) vs. 9% for $D\downarrow$ (net -253); Tab. 8: med ΔD for $D\uparrow$/Fixed -16.4.

8.1 Experimental setup

Intervention design. Our empirical objective is to test whether the within-trace observable bridge reveals the theoretically predicted regime signatures in single-trace q_t trajectories on real LM reasoning. The main challenge is that, to avoid circularity, empirical support has to come from external signals bearing on regime *independently of* q_t. Final-answer correctness is one such signal, albeit noisy: even on open-answer tasks with low lucky-guess rates, wrong answers can still arise from traces that lock into an incorrect branch, and correct answers from traces that remain diffuse but stumble upon the right answer. So while informative, correctness is only a coarse regime signal.

Intervention response is therefore a key probe. Since the theory predicts opposite informational effects for greedy decoding across the two regimes, perturbing *away* from greedy should also have opposite (inverse-signed) effects: a natural test of the asymmetry from the other side. For the reasoning tasks we test, a regime-blind off-greedy perturbation would ordinarily be expected to hurt on average since greedy is a strong default. Our theory predicts otherwise on confusion-like traces: an off-greedy perturbation *might* move the run off consensus, reducing unresolved support and *improving* the answer. This provides a unique window onto testing our theoretical predictions.

Our main intervention is a fixed, regime-blind off-greedy perturbation: at the top-k=3 highest-entropy positions of each greedy trace, replace the greedy token with the rank-3 alternative and continue decoding (App. J for robustness in k and rank). The goal is not to optimize performance but to test whether the same perturbation has systematically different downstream effects across lock-in-like vs. confusion-like traces.

Models and tasks. Five open-weight instruction-tuned models (Qwen2.5 1.5B/7B-Instruct [25], Gemma 7B-it [9], Granite 3.1 8B-Instruct [14], Mistral 7B-Instruct-v0.3 [15]) on the first 100 problems each from GSM8K, BBH object counting, and BBH multi-step arithmetic. For each model we collect greedy and temperature-1 traces, plus intervened continuations described above.

Residual-support statistic. Proposition 4 requires a practical choice of h_t. We take a conservative approach: fit a low-capacity family so any remaining signal must survive aggressive residualization, and then test whether the predicted signatures persist. Empirically, heavy-tailed families behave

similarly, while an exponential head degrades substantially (App. J), suggesting that the key ingredient is a heavy-tailed head. Operationally, we chose Zipf because Zipf-like rank-frequency structure is well documented in both natural language [24] and LM outputs [12, 22]. See App. J for details. Note that computing D is pure post-processing of the next-token distribution the model has already produced, hence incurring minimal computational overhead.

Roadmap. Table 1 lays out the falsifiable commitments from our theoretical results: each row pairs a theoretical prediction with the specific observable signature that would support or falsify it; the third column reports what we find. The remainder of the section walks through these results.

8.2 Experimental results

Table 2: Consistency check: median per-trace H and D for correct and incorrect traces.

	Temperature-1			Greedy		
	GSM8K	BBH-OC	BBH-MA	GSM8K	BBH-OC	BBH-MA
Shannon entropy H (correct/wrong)						
Qwen 1.5B	0.67 / **0.79**	0.72 / **0.80**	0.80 / **0.81**	0.60 / **0.65**	**0.65** / 0.65	0.71 / **0.76**
Qwen 7B	0.56 / **0.60**	**0.51** / 0.49	**0.63** / 0.60	0.53 / **0.57**	**0.49** / 0.47	**0.64** / 0.62
Gemma 7B	**0.51** / 0.51	0.50 / **0.57**	0.52 / **0.54**	0.47 / **0.50**	0.46 / **0.52**	**0.56** / 0.54
Granite 8B	0.67 / **0.72**	0.75 / **0.77**	0.78 / **0.78**	0.64 / **0.66**	0.76 / **0.77**	**0.78** / 0.75
Mistral 7B	0.66 / **0.72**	0.75 / **0.75**	0.69 / **0.70**	0.61 / **0.65**	**0.74** / 0.72	0.64 / **0.67**
Residual support D (correct/wrong)						
Qwen 1.5B	240 / **302**	218 / **316**	235 / **299**	193 / **221**	191 / **293**	187 / **236**
Qwen 7B	352 / **412**	167 / **278**	324 / **517**	232 / **326**	162 / **330**	265 / **376**
Gemma 7B	196 / **237**	125 / **234**	282 / **342**	163 / **212**	103 / **247**	315 / **364**
Granite 8B	615 / **726**	751 / **790**	652 / **701**	580 / **600**	757 / **776**	635 / **707**
Mistral 7B	490 / **559**	450 / **507**	528 / **652**	496 / **520**	438 / **494**	579 / **720**

A coarse consistency check first: if correctness and D both carry regime signal, they should correlate. Table 2 confirms this: across all 30 model×benchmark×decoding cells, median D is lower on correct traces than wrong, while H shows only weak, inconsistent separation. As a scale-free effect size, $\theta_{\text{level}} = \Pr(D_{\text{wrong}} > D_{\text{correct}})$ exceeds 0.5 in all 30 cells at full benchmark size (median 0.65, range 0.53–0.79; App. O). Figure 2 shows the same temporally: H trajectories for correct and wrong traces remain intertwined, while those of D separate in the second half (e.g., GSM8K: median D drops sharply on correct traces, stays elevated on wrong). This is the first theoretical prediction: scalar sharpness alone cannot reliably distinguish the two regimes.

Table 3: Intervention-induced ΔH, ΔD on wrong-baseline traces. W=wrong, C=correct, med=median. ΔH in units of 10^{-2} nats. Smaller (more-negative) median **bolded**.

			ΔH ($\times 10^{-2}$ nats)				ΔD			
			W→C		W→W		W→C		W→W	
Model	$n_{\text{W}\to\text{C}}$	$n_{\text{W}\to\text{W}}$	med	95% CI	med	95% CI	med	95% CI	med	95% CI
Qwen 1.5B	54	279	**-0.5**	[-2.9, +0.9]	+0.8	[+0.3, +1.2]	**-5.4**	[-48.5, +1.2]	+0.1	[-2.0, +2.8]
Qwen 7B	19	137	+0.4	[+0.1, +0.7]	**+0.1**	[-0.1, +0.2]	**-0.4**	[-24.4, +11.2]	0.0	[-0.1, +0.5]
Gemma 7B	54	692	+0.9	[-1.5, +4.1]	**+0.1**	[-0.1, +0.3]	**0.0**	[-5.1, +7.3]	+0.2	[0.0, +1.9]
Granite 8B	67	293	**+0.5**	[-1.5, +2.2]	+1.3	[+0.6, +2.1]	**-3.9**	[-62.1, +17.6]	+15.3	[+6.4, +28.4]
Mistral 7B	52	530	**-1.0**	[-5.5, +1.9]	+0.8	[+0.3, +1.6]	**-43.8**	[-63.9, +7.7]	+0.6	[-3.9, +4.0]
All	246	1931	+0.1	[-0.9, +0.7]	+0.4	[+0.2, +0.6]	-2.1	[-17.0, +0.2]	+0.7	[0.0, +1.8]

Table 2 is only a consistency check: any observable tracking reasoning quality could show a similar correct/wrong split. The more discriminating test is whether D behaves differently from H under intervention. The sharpest intervention prediction concerns wrong-baseline traces[2]: a wrong trace in

[2]Correct→Wrong shows the predicted positive ΔD (App. J Table 6), consistent with lock-in disruption, though we treat it as supportive rather than primary because C→W is also consistent with commitment shifts to a different wrong branch.

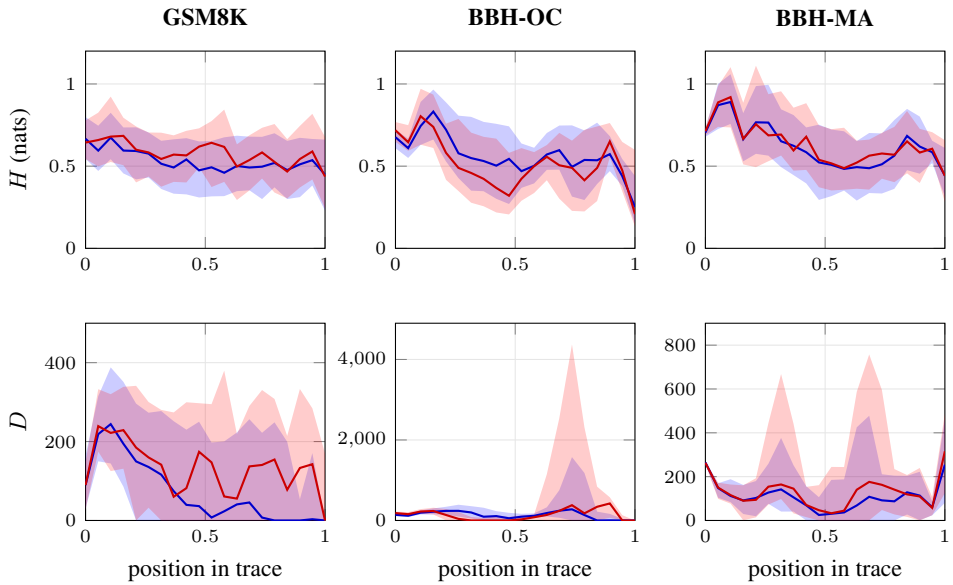

Figure 2: Trajectories of H (top) and D (bottom) for Qwen 2.5 1.5B (greedy traces on 250 problems per benchmark: 185/65, 153/97, and 161/89 correct/wrong traces for GSM8K, BBH-OC, and BBH-MA). Blue: correct; red: incorrect. Solid: cross-trace median; bands: IQR. Each trace is smoothed (10-step mean) and resampled to 20 bins; tokens with $H \leq 0.1$ (boilerplate) are masked.

Table 4: Interventions by trace regime. The larger of $D\downarrow$ vs $D\uparrow$ for fix/broke rate and net is **bolded**.

	$D\downarrow$ traces							$D\uparrow$ traces						
Model	n_W	n_C	ΔD	>0	fixed (rate)	broke (rate)	net	n_W	n_C	ΔD	>0	fixed (rate)	broke (rate)	net
Qwen 1.5B	171	423	+3.4	60%	23 (13%)	77 (**18%**)	−54	162	144	−1.4	47%	31 (**19%**)	23 (16%)	**+8**
Qwen 7B	96	670	+0.4	53%	9 (9%)	174 (**26%**)	−165	60	71	−0.1	34%	10 (**17%**)	12 (17%)	**−2**
Gemma 7B	452	131	+1.9	56%	25 (6%)	27 (21%)	−2	294	20	0.0	44%	29 (**10%**)	5 (**25%**)	**+24**
Granite 8B	237	399	+14.9	62%	41 (17%)	55 (14%)	−14	123	141	−0.8	50%	26 (**21%**)	27 (**19%**)	**−1**
Mistral 7B	396	254	+7.3	59%	28 (7%)	46 (18%)	−18	186	62	−26.3	35%	24 (**13%**)	12 (**19%**)	**+12**
All	1352	1877	+3.3	58%	126 (9%)	379 (**20%**)	−253	825	438	−2.1	43%	120 (**15%**)	79 (18%)	**+41**

confusion *could* be redirected to a coherent branch, yielding a fix and reduced D. If scalar uncertainty H drove things, successful fixes (W→C) should result in lower entropy than unsuccessful ones (W→W). Per our theory, however, their difference should align primarily with D instead of H.

Table 3 reveals a qualitative H/D decoupling under the off-greedy intervention discussed earlier: ΔH remains near zero on fixes with no consistent separation from W→W. ΔD is lowest in W→C for every model and nonpositive in all five cases, while W→W are all nonnegative. This asymmetry is quantified by the distribution-free effect size $\theta = \Pr(\Delta D_{\mathrm{W}\to\mathrm{C}} < \Delta D_{\mathrm{W}\to\mathrm{W}})$ (ties counted $1/2$), computed within model×benchmark strata with a trace-clustered bootstrap: $\theta = 0.564$, 95% CI $[0.519, 0.612]$, one-sided $p = 0.003$; the direction holds in 5/5 models and in all nine (k, rank) settings (App. N). In sum, the outcome aligns with reduced D instead of reduced entropy, as predicted if some wrong-baseline traces are confusion-like and rescued by being pushed off consensus.

Note that the wrong-baseline pool is, by design, heterogeneous: it mixes confusion-like wrong traces (rescuable by off-greedy) with lock-in-like wrong traces (where the dominant branch is itself the wrong answer and off-greedy might push them to the correct lock-in or a wrong confusion). The theoretical signature is therefore the directional W→C vs. W→W asymmetry, not whether confidence intervals should exclude zero. CIs spanning zero in the tables simply reflect the predicted heterogeneity of the population.

Another important note is that, while the transition table shows H and D behave differently under intervention, it does not, by itself, rule out a weaker interpretation that D is simply another proxy for correctness. To probe that possibility, we ask a stricter question: does baseline D-trajectory class

predict intervention response *beyond* baseline correctness? If D were merely tracking whether a trace was right or wrong, its predictive value within the same baseline outcome class should be limited. Our theory predicts otherwise. Traces with rising D are more confusion-like and should therefore be more fixable; traces with falling D are more lock-in-like and therefore more fragile.

Table 4 shows this. We classify traces as $D\uparrow$/$D\downarrow$ by first- vs. last-quarter mean D on post-prompt tokens (App. J for endpoint robustness). Fixed and broke rates are computed relative to n_W and n_C (intervened branches from wrong- and correct-baseline traces), with net = #fixed − #broke. Across all models, $D\uparrow$ has higher fix rate; in aggregate, 15% vs. 9%, and $D\uparrow$ is the only class with a positive net intervention gain (+41 vs. −253). Median ΔD also splits as predicted. This is hard to square with D being just correctness in disguise: D-class tracks an intervention-relevant property—confusion-like runs are recoverable, lock-in-like runs damageable. Cell-level analysis (App. K) shows the strongest fix signature in $D\uparrow$/Fixed and break signature in $D\downarrow$/Broke (Table 8: med ΔD = -16.4 and +10.8, respectively), exactly the cells the theory singles out.

Together, the experiments support all proposed signatures: (i) entropy sharpening alone is too coarse to distinguish concentrated from unresolved traces; (ii) under alternative-token intervention, D behaves differently between the two regimes (also distinct from entropy) in the direction predicted; (iii) baseline D-trajectory class predicts which traces are more recoverable, with the clearest signatures in the predicted regime×outcome cells.

Direct validation (App. L). The results above validate D through correctness and intervention response. Two further experiments compare it with branch support itself. The first is a synthetic process that realizes (3) with a known branch posterior. There, D computed from q_t alone tracks $N_{\text{eff}}(\pi_t)$ (trace-level Spearman $+0.52$) whereas H is wrong-signed (-0.29). The two regimes also emerge as predicted: the posterior resolves on correct traces, with $N_{\text{eff}}(\pi_t)$ falling from 9.25 to 1.13, but stays diffuse on wrong ones (11.75 to 10.92), while H falls identically in both.

The second experiment returns to the real traces and estimates the branch posterior under final-answer equivalence directly, by clustering 16 resampled continuations at four positions of each trace by their final answer. Late in the trace, the estimated support is 1.83 on wrong traces and 1.00 on correct ones, with entropy indistinguishable between the two (0.53 in both). D also predicts the resampled probability of reaching the gold answer ($\rho = -0.18$, $p < 10^{-11}$) while H does not ($\rho = -0.01$).

Robustness and breadth (App. M–O). App. M shows that the regime verdicts are stable under coarser answer partitions. App. N shows that the intervention asymmetry is robust to controls for placement, forced-token probability, position, and trace length, and holds at random and lowest-entropy positions. The level result extends to the full benchmarks ($30/30$ cells), a larger model, and two non-arithmetic tasks (App. O).

Limitations. First, a regime label is relative to an analyst-declared notion of resolution (§3). Any statistic of q_t alone, including D, can only serve as a within-configuration discriminator rather than an estimator of branch support. Second, final-answer correctness is a coarse proxy for regime: wrong answers can come from traces locked into a wrong branch and correct answers from lucky guesses in confusion. Empirical evidence should thus be read in aggregate (e.g., at the regime level) rather than trace by trace. Third, D presupposes the bridge condition of Proposition 4, which is what licenses residualizing a heavy-tailed head. D therefore serves only notions of resolution that satisfy this condition (App. M), and its head must match the backbone: the results survive other heavy-tailed families but degrade under a light-tailed one (App. J). Fourth, the intervention controls do not cover the syntactic category of the forced token, which we leave to future work.

9 Conclusion

We analyzed autoregressive decoding and showed that the same surface-level sharpening can arise from two distinct regimes—lock-in and confusion—driven by a mismatch between token selection and branch resolution. Entropy alone cannot distinguish a coherent concentration from an unresolved one. Greedy token selection also has opposite informational effects across regimes. A natural next step is a regime-aware decoding rule that gates off-greedy exploration on the D trajectory: raise the temperature only when residual support is elevated, and maintain a low (or zero) temperature in falling-D traces, so as to protect correct lock-in while targeting rescuable confusion.

Acknowledgments and Disclosure of Funding

The authors were supported in part by Award 2537027 from the National Science Foundation and Award 262812UND2 from the Notre Dame-IBM Technology Ethics Lab. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors. The authors declare no competing interests.

References

[1] Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. The unreasonable effectiveness of entropy minimization in LLM reasoning. *arXiv preprint arXiv:2505.15134*, 2025. URL https://arxiv.org/abs/2505.15134.

[2] Gustaf Ahdritz, Tian Qin, Nikhil Vyas, Boaz Barak, and Benjamin L. Edelman. Distinguishing the knowable from the unknowable with language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 503–549, 2024. URL https://proceedings.mlr.press/v235/ahdritz24a.html.

[3] Marthe Ballon, Brecht Verbeken, Vincent Ginis, and Andres Algaba. Probing the trajectories of reasoning traces in large language models. *arXiv preprint arXiv:2601.23163*, 2026. URL https://arxiv.org/abs/2601.23163.

[4] Jinkun Chen, Fengxiang Cheng, Sijia Han, and Vlado Keselj. "I may not have articulated myself clearly": Diagnosing dynamic instability in LLM reasoning at inference time. *arXiv preprint arXiv:2602.02863*, 2026. URL https://arxiv.org/abs/2602.02863.

[5] Prateek Chhikara. Mind the confidence gap: Overconfidence, calibration, and distractor effects in large language models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=lyaHnHDdZl.

[6] Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems*, 2023.

[7] Fabian Falck, Ziyu Wang, and Christopher C. Holmes. Is in-context learning in large language models Bayesian? A martingale perspective. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 12784–12805. PMLR, 2024. URL https://proceedings.mlr.press/v235/falck24a.html.

[8] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0.

[9] Gemma Team. Gemma: Open models based on Gemini research and technology, 2024. URL https://arxiv.org/abs/2403.08295.

[10] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

[11] Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

[12] Yanjin He, Qingkai Zeng, and Meng Jiang. Pre-trained models perform the best when token distributions follow Zipf's law. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27997–28009, 2025.

[13] Zhonghao He, Tianyi Qiu, Hirokazu Shirado, and Maarten Sap. Martingale score: An unsupervised metric for Bayesian rationality in LLM reasoning. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=BfO6od6JD6.

[14] IBM Granite Team. Granite 3.1 language models. Model card on Hugging Face, 2024. URL https://huggingface.co/ibm-granite/granite-3.1-8b-instruct.

[15] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, 2023. URL https://arxiv.org/abs/2310.06825.

[16] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. URL https://arxiv.org/abs/2207.05221.

[17] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

[18] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukosiute, 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. *arXiv preprint arXiv:2307.13702*, 2023. URL https://arxiv.org/abs/2307.13702.

[19] Huayu Li, ZhengXiao He, Siyuan Tian, Jinghao Wen, and Ao Li. Martingale foresight sampling: A principled approach to inference-time LLM decoding. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3522–3533. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.eacl-long.162. URL https://aclanthology.org/2026.eacl-long.162/.

[20] 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. URL https://arxiv.org/abs/2305.20050.

[21] Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

[22] Nikolay Mikhaylovskiy. Zipf's and Heaps' laws for tokens and LLM-generated texts. *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15469–15481, 2025.

[23] Jacob Pfau, William Merrill, and Samuel R. Bowman. Let's think dot by dot: Hidden computation in transformer language models. *arXiv preprint arXiv:2404.15758*, 2024.

[24] Steven T. Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130, 2014. doi: 10.3758/s13423-014-0585-6.

[25] Qwen Team. Qwen2.5 technical report, 2024. URL https://arxiv.org/abs/2412.15115.

[26] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, 2023.

[27] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

[28] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

[29] Zhangyue Yin, Qiushi Sun, Qipeng Guo, Zhiyuan Zeng, Xiaonan Li, Junqi Dai, Qinyuan Cheng, Xuanjing Huang, and Xipeng Qiu. Reasoning in flux: Enhancing large language models reasoning through uncertainty-aware adaptive guidance. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2401–2416, 2024. doi: 10.18653/v1/2024.acl-long.131. URL https://aclanthology.org/2024.acl-long.131/.

[30] Kelly W. Zhang, Tiffany Cai, Hongseok Namkoong, and Daniel Russo. Posterior sampling via autoregressive generation. *arXiv preprint arXiv:2405.19466*, 2024. URL https://arxiv.org/abs/2405.19466.

[31] Tunyu Zhang, Haizhou Shi, Yibin Wang, Hengyi Wang, Xiaoxiao He, Zhuowei Li, Haoxian Chen, Ligong Han, Kai Xu, Huan Zhang, Dimitris Metaxas, and Hao Wang. Tokur: Token-level uncertainty estimation for large language model reasoning. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=VHQc7wzmYv.

[32] Xinghao Zhao. Entropy trajectory shape predicts LLM reasoning reliability: A diagnostic study of uncertainty dynamics in chain-of-thought. *arXiv preprint arXiv:2603.18940*, 2026. URL https://arxiv.org/abs/2603.18940.

[33] Amir Zur, Atticus Geiger, Ekdeep Singh Lubana, and Eric Bigelow. Are language models aware of the road not taken? token-level uncertainty and hidden state dynamics. *arXiv preprint arXiv:2511.04527*, 2025. URL https://arxiv.org/abs/2511.04527.

A Extended related work

Scalar uncertainty and entropy dynamics. A large body of work uses token probabilities, entropy, confidence, or related scalar uncertainty measures as diagnostics for reliability or as inference-time control signals. Examples include token-level self-knowledge analyses [16], uncertainty estimators based on weight perturbations [31], uncertainty-aware reasoning guidance [29], logit-based decomposition of next-token uncertainty into epistemic and aleatoric components [2], and recent evidence that entropy minimization alone is a surprisingly effective inference-time intervention for reasoning [1], and the shape of an entropy trajectory across a reasoning trace can be informative [32]. Our point is not that scalar uncertainty is uninformative. Rather, the theory shows that the same scalar sharpening pattern can arise from two different posterior geometries. In lock-in, top tokens remain branch-evidential and reinforce concentration; in confusion, top tokens can be high-probability consensus continuations that preserve unresolved branch support. Entropy is thus useful but not sufficient for distinguishing the two regimes. Further, the statistic we introduce is not a competing uncertainty metric but a vehicle for testing our theoretical predictions for the two regimes which are properties of how the next-token law evolves across a decoding trace rather than a decomposition of uncertainty at a single snapshot.

Calibration and overconfidence. Calibration work asks whether scalar confidence matches empirical accuracy, from temperature scaling for classifiers [10] to recent studies of overconfidence and miscalibration in LLM reasoning [5]. Our claim is at a different level of description. Two traces can have the same token entropy, top-token confidence, or even similar entropy trajectories while belonging to different dynamical regimes: in lock-in, probability mass concentrates around one branch, whereas in confusion, sharp local predictions can coexist with unresolved posterior mass over incompatible branches. Calibration evaluates whether confidence is reliable on average; our analysis asks what branch-level posterior geometry scalar confidence hides within a single decoding trace.

Cross-sample uncertainty and aggregation. Self-consistency and semantic-entropy methods improve reliability by comparing multiple sampled completions or clustering generations by answer meaning [8, 17, 27]. These methods are naturally population-level: they ask what can be inferred from an ensemble of outputs. Our analysis is within-trace: it asks how a single autoregressive trajectory evolves and why local sharpness may or may not correspond to branch resolution. The two perspectives are complementary. In particular, cross-sample disagreement may reveal that multiple branches exist [21], whereas our statistic is designed to diagnose whether a particular trace is resolving or preserving such multiplicity as it unfolds.

Process supervision, probing, and faithfulness. Process-oriented approaches to reasoning share with our work an interest in the reasoning process rather than only its final answer. Process reward models and step-level verifiers evaluate intermediate reasoning states using trained supervision [20]; volatility and trajectory-probing approaches examine how token distributions or final answers respond to prefix perturbations [3, 4]. Work on chain-of-thought faithfulness asks whether generated explanations report the computation the model actually performs [18, 26], while hidden-computation and latent-alternative work studies whether transformers internally represent unrealized continuations [6, 23, 33]. Our analysis is compatible with these questions but addresses a different object: the posterior branch geometry. A trace may be faithful or unfaithful, smooth or volatile, while still occupying either the lock-in or confusion regime.

Bayesian and posterior-sampling views of LLMs. Several recent papers use Bayesian or martingale language to analyze LLM behavior. Falck et al. test whether in-context learning satisfies martingale conditions required of Bayesian prediction [7]; Martingale Score uses violations of Bayesian belief-update martingales as a measure of belief entrenchment [13]; posterior-sampling work uses autoregressive generation to implement Thompson-style decision rules [30]; and Martingale Foresight Sampling uses martingale tools to guide inference-time path search [19]. These works are adjacent in mathematical vocabulary but differ in object. Our Bayesian analysis is used to derive the accounting identity and to distinguish lock-in from confusion in ordinary autoregressive reasoning traces.

B What the framework assumes and what it derives

This appendix presents a detailed explanation of (i) the inputs the framework takes as given and the quantities it derives from them, and (ii) the formal measure-theoretic setup and conventions used throughout the paper.

B.1 Inputs and derivations

Inputs. The framework takes as given two objects: the model's distribution over completions, and the analyst's chosen notion of resolution. The first is the joint distribution induced by the autoregressive language model on completion sequences extending a given prefix. The second is a partition of the completion space into equivalence classes (*branches*), determined by what distinction the analyst wishes to resolve. The model's distribution exists by construction while the partition is supplied by the analyst.

Derivations. Under the joint distribution determined by these two inputs, the branch posterior $\pi_t(i) := \Pr(P = p_i \mid x_{<t})$ and the branch-conditional next-token laws $q_{i,t}(x) := \Pr(x_t = x \mid P = p_i, x_{<t})$ are conditional probabilities that follow. The decomposition $q_t(x) = \sum_i \pi_t(i)\, q_{i,t}(x)$ is the law of total probability through the partition. The Rényi-2 effective support $N_{\text{eff}}(\pi_t) = 1/\sum_i \pi_t(i)^2$, the regime classifications, and the residual-support statistic D are all functions of these conditional probabilities and the model's outputs. Lemma 1 and Propositions 1–4 are statements about the dynamics of these quantities.

Predictions follow from the model's outputs. Our framework is explanatory rather than metaphysical: it provides a structural account for the already-observed phenomenon of entropy sharpening during failed reasoning, and yields testable intervention signatures. Concretely, the framework's claims are about derived quantities. The model produces q_t, and Bayes' rule determines how π_t evolves from q_t. The propositions show properties of that evolution. Predictions therefore follow from the model's actual q_t outputs. The question of whether the model "tracks" π_t in any internal sense is outside the framework's scope and not needed for its predictions to apply.

Granularity is the analyst's input, by design. Different choices of resolution notion yield different partitions and hence different regime classifications for the same trace. This is the framework's intended structure: regime characterizes how decoding interacts with a chosen criterion of "branch resolution". A trace may be in lock-in at one criterion (e.g., final-answer equivalence) and in confusion at a finer criterion (e.g., semantic equivalence, reasoning-step equivalence). The propositions apply at both granularities simultaneously, since each granularity yields its own joint distribution and its own derived quantities. The empirical observables are computed from q_t alone, but their discriminative value as regime indicators is partition-conditional: as discussed in §7, for the predicted regime signatures to appear empirically, the chosen partition needs to satisfy the bridge condition of Proposition 4. Roughly, the partition cannot be so fine-grained that different branches retain strong overlap beyond the shared backbone. Since we use correctness as the external proxy in the experiments, the natural reading is to take final-answer equivalence as the partition, because confusion-like traces under the answer-equivalence partition are less likely to produce the correct answer. This is for ease of understanding rather than theoretical or empirical necessity: other bridge-compatible partitions paired with their own proxies would yield analogous predictions.

B.2 Formal setup and conventions

Spaces and measurability. Let V be the (countable) vocabulary, $V^* \cup V^{\mathbb{N}}$ the space of finite or infinite completions equipped with the product σ-algebra. The model defines a probability measure on this space conditional on any prefix $x_{<t} \in V^*$. The partition $P : V^* \cup V^{\mathbb{N}} \to I$ assigns each completion its branch label; we assume P is measurable with respect to the product σ-algebra and the discrete σ-algebra on I. For natural partitions (final-answer equivalence, syntactic-equivalence, etc.) measurability is automatic since P can be computed from the completion sequence by a measurable algorithm.

Finite branch set. We assume $|I| < \infty$. This makes π_t a probability vector in $\mathbb{R}^{|I|}$, all sums in the propositions finite, and Proposition 3's recursion well-defined. For partitions with countably many

branches, the propositions extend with appropriate convergence assumptions and with c_t replaced by a prior-weighted average $\sum_i \mu_i q_{i,t}$ for some distribution μ on I.

Live branches. We work throughout with the live-branch set $I_t := \{i : \pi_t(i) > 0\}$ at each step. Branches with zero posterior are operationally irrelevant (they cannot be resolved or evidenced) and are excluded from sums and maxima. Under a prior π_0 of full support on I, I_t contains every branch consistent with the observed prefix. The dominant branch $i_t^\star := \arg\max_{i \in I_t} \pi_t(i)$ and posterior moments $\pi_t^\star, \pi_t^\sharp$ are defined over I_t.

Bayesian updates. The update $\pi_{t+1}(i) = \pi_t(i) q_{i,t}(x_t)/q_t(x_t)$ requires $q_t(x_t) > 0$. This holds automatically for any token x_t emitted by the model under any decoding policy supported on $\{x : q_t(x) > 0\}$ (sampling, greedy, top-k, top-p, beam, etc.). We assume that edge cases ($q_t(x) = 0$) do not arise for emitted tokens.

Edge cases for $L_t(x)$. The likelihood ratio $L_t(x) := q_t^\star(x)/\bar{q}_t(x)$ is extended to handle zero denominators by:

- $L_t(x) := +\infty$ if $\bar{q}_t(x) = 0$ and $q_t^\star(x) > 0$ (token uniquely diagnostic for the dominant branch);
- $L_t(x) := 0$ if $\bar{q}_t(x) > 0$ and $q_t^\star(x) = 0$;
- $L_t(x) := 1$ if both are zero (degenerate, occurs with probability zero).

Tie-breaking under greedy. When multiple tokens achieve $\arg\max_x q_t(x)$, ties are broken by an arbitrary fixed rule (e.g., lexicographic).

Conditional updates vs. expectations over x_t. The Bayes update is a conditional-probability statement under the model's joint distribution: given any realized prefix $x_{\le t}$, $\pi_{t+1}(i) = \Pr_M(P = p_i \mid x_{\le t})$, computed via the chain rule. This conditional is independent of how the prefix was generated, so the Bayes update applies under any decoding policy that produces x_t, e.g., temperature-1 sampling, greedy, top-k, top-p, forced rank-r, beam, etc. The martingale property in Lemma 1, by contrast, is a statement about the *expectation* of π_{t+1} over the marginal distribution of x_t. This expectation depends on which marginal one uses: under temperature-1 sampling, $x_t \sim q_t$ and the alignment term in Lemma 1 vanishes; under non-sampling policies the marginal differs from q_t and the alignment term is generally nonzero, capturing the deterministic bias of the policy. In sum, conditional updates are policy-independent while expectations over x_t are not.

C Proof of Lemma 1

Proof of Lemma 1. For clarity, we explicate subscripts on expectations: write $\mathbb{E}_s[\cdot] := \mathbb{E}[\cdot \mid x_{<s}]$. Since $C_s = \|\pi_s\|_2^2$, we have

$$C_{s+1} - C_s = \|\pi_s + (\pi_{s+1} - \pi_s)\|_2^2 - \|\pi_s\|_2^2 = 2\langle \pi_s, \pi_{s+1} - \pi_s \rangle + \|\pi_{s+1} - \pi_s\|_2^2. \tag{19}$$

Summing from $s = t$ to $t + T - 1$ and taking conditional expectation given $x_{<t}$ gives

$$\mathbb{E}_t[C_{t+T} - C_t] = \sum_{s=t}^{t+T-1} \mathbb{E}_t[\|\pi_{s+1} - \pi_s\|_2^2] + 2 \sum_{s=t}^{t+T-1} \mathbb{E}_t[\langle \pi_s, \mathbb{E}_s[\pi_{s+1} - \pi_s] \rangle]. \tag{20}$$

Under temperature-1 sampling, $X_s \sim q_s(\cdot)$ conditionally on $x_{<s}$. By Bayes' rule,

$$\pi_{s+1}(i) = \pi_s(i) \frac{q_{i,s}(X_s)}{q_s(X_s)}. \tag{21}$$

Therefore

$$\mathbb{E}_s[\pi_{s+1}(i)] = \pi_s(i) \sum_x q_s(x) \frac{q_{i,s}(x)}{q_s(x)} = \pi_s(i). \tag{22}$$

Thus, under temperature-1 sampling, $(\pi_s)_{s \geq t}$ is a martingale with respect to the filtration generated by the decoded prefix, and the alignment term vanishes.

It remains to identify the one-step squared motion. Again by Bayes' rule,

$$\pi_{s+1}(i) - \pi_s(i) = \pi_s(i) \left(\frac{q_{i,s}(X_s)}{q_s(X_s)} - 1 \right). \tag{23}$$

Thus

$$\mathbb{E}_s[\|\pi_{s+1} - \pi_s\|_2^2] = \sum_i \pi_s(i)^2 \sum_x q_s(x) \left(\frac{q_{i,s}(x)}{q_s(x)} - 1 \right)^2. \tag{24}$$

By the definition of χ^2 divergence,

$$\mathbb{E}_s[\|\pi_{s+1} - \pi_s\|_2^2] = \sum_i \pi_s(i)^2 \, \chi^2(q_{i,s} \| q_s). \tag{25}$$

Substituting this identity into the previous display gives

$$\mathbb{E}_t[C_{t+T} - C_t] = \sum_{s=t}^{t+T-1} \mathbb{E}_t[\|\pi_{s+1} - \pi_s\|_2^2] = \sum_{s=t}^{t+T-1} \mathbb{E}_t \left[\sum_i \pi_s(i)^2 \, \chi^2(q_{i,s} \| q_s) \right]. \tag{26}$$

□

D Proof of Proposition 1

Proof of Proposition 1. For the temperature-1 claim, write

$$\pi_s^\star := \max_i \pi_s(i), \qquad \pi_s^{(2)} := \max_{j \neq i_s^\star} \pi_s(j), \qquad g_s := \pi_s^\star - \pi_s^{(2)}.$$

If $N_{\text{eff}}(\pi_s) \leq K < 2$, then

$$\sum_i \pi_s(i)^2 \leq \frac{1 + g_s^2}{2},$$

so

$$\frac{1}{K} \leq \sum_i \pi_s(i)^2 \leq \frac{1 + g_s^2}{2} \quad \Longrightarrow \quad g_s^2 \geq \frac{2}{K} - 1. \tag{27}$$

Note that if $i_{s+1}^\star \neq i_s^\star$, the new dominant branch must overtake the old one, which requires posterior movement of at least $g_s/\sqrt{2}$ in ℓ_2. That is,

$$\|\pi_{s+1} - \pi_s\|_2^2 \geq \frac{g_s^2}{2} \geq \frac{1}{2}\Big(\frac{2}{K} - 1\Big).$$

Therefore

$$\mathbf{1}\big\{N_{\text{eff}}(\pi_s) \leq K,\ i_{s+1}^\star \neq i_s^\star\big\} \leq \frac{\|\pi_{s+1} - \pi_s\|_2^2}{\frac{1}{2}(\frac{2}{K} - 1)}.$$

Summing over $s \geq t$,

$$\sum_{s \geq t} \mathbf{1}\big\{N_{\text{eff}}(\pi_s) \leq K,\ i_{s+1}^\star \neq i_s^\star\big\} \leq \frac{1}{\frac{1}{2}(\frac{2}{K} - 1)} \sum_{s \geq t} \|\pi_{s+1} - \pi_s\|_2^2.$$

Taking conditional expectation and using Lemma 1 under temperature-1 sampling,

$$\mathbb{E}\left[\sum_{s \geq t} \|\pi_{s+1} - \pi_s\|_2^2 \,\middle|\, x_{<t}\right] = \mathbb{E}[C_\infty - C_t \mid x_{<t}] \leq 1 - C_t.$$

Since $N_{\text{eff}}(\pi_t) \leq K$, we have $C_t = 1/N_{\text{eff}}(\pi_t) \geq 1/K$, thus

$$\mathbb{E}\left[\sum_{s \geq t} \mathbf{1}\big\{N_{\text{eff}}(\pi_s) \leq K,\ i_{s+1}^\star \neq i_s^\star\big\} \,\middle|\, x_{<t}\right] \leq \frac{1 - \frac{1}{K}}{\frac{1}{2}(\frac{2}{K} - 1)} = \frac{2(K-1)}{2 - K}.$$

For the general claim, Bayes' rule and the factorization

$$q_t(x) = \bar{q}_t(x)\big[(1 - \pi_t^\star) + \pi_t^\star L_t(x)\big]$$

give

$$\pi_{t+1}(i_t^\star) = \frac{\pi_t^\star q_t^\star(x_t)}{q_t(x_t)} = \frac{\pi_t^\star L_t(x_t)}{(1 - \pi_t^\star) + \pi_t^\star L_t(x_t)}. \tag{28}$$

Hence

$$\pi_{t+1}(i_t^\star) - \pi_t^\star = \frac{\pi_t^\star(1 - \pi_t^\star)\big(L_t(x_t) - 1\big)}{(1 - \pi_t^\star) + \pi_t^\star L_t(x_t)}. \tag{29}$$

If $L_t(x_t) \geq 1 + \varepsilon$, then

$$\pi_{t+1}(i_t^\star) - \pi_t^\star \geq \frac{\pi_t^\star(1 - \pi_t^\star)\varepsilon}{1 + \pi_t^\star \varepsilon}.$$

Note that $1 + \tau \leq N_{\text{eff}}(\pi_t) \leq K$ implies

$$\frac{1}{K} \leq \frac{1}{N_{\text{eff}}(\pi_t)} = \sum_i \pi_t(i)^2 \leq \pi_t^\star \leq \sqrt{\sum_i \pi_t(i)^2} = \sqrt{\frac{1}{N_{\text{eff}}(\pi_t)}} \leq \frac{1}{\sqrt{1 + \tau}}.$$

Therefore

$$\pi_t^\star(1 - \pi_t^\star) \geq \frac{1}{K}\Big(1 - \frac{1}{\sqrt{1 + \tau}}\Big), \qquad 1 + \pi_t^\star \varepsilon \leq 1 + \frac{\varepsilon}{\sqrt{1 + \tau}},$$

and substituting yields the first claim (8). When $K < 2$, there must be $\pi_t^\star > 1/2$, therefore no dominant-branch switch is possible when $\pi_{t+1}(i_t^\star) > \pi_t^\star$. □

E Proof of Proposition 2

Proof of Proposition 2. Recall that

$$\eta_t^2 = \left\| \sum_i \pi_t(i) r_{i,t} \right\|_2^2 = \pi_t^\top G_t \pi_t \le \|G_t\|_2 \|\pi_t\|_2^2 = \frac{\|G_t\|_2}{N_{\text{eff}}(\pi_t)}. \tag{30}$$

The condition $N_{\text{eff}}(\pi_t) > \|G_t\|_2 / c_t(x)^2$ thus implies $\eta_t < c_t(x)$. Now define

$$\bar{r}_t(x) := \sum_i \pi_t(i) r_{i,t}(x). \tag{31}$$

Then $q_t(x) = c_t(x) + \bar{r}_t(x)$ and $\|\bar{r}_t\|_\infty \le \eta_t$.

By Bayes' rule,

$$\pi_{t+1}(i) - \pi_t(i) = \pi_t(i) \frac{r_{i,t}(x) - \bar{r}_t(x)}{c_t(x) + \bar{r}_t(x)}. \tag{32}$$

Since $c_t(x) > \eta_t$ and $|\bar{r}_t(x)| \le \eta_t$, the denominators below are positive, and

$$\frac{r_{i,t}(x) + \eta_t}{c_t(x) - \eta_t} - \frac{r_{i,t}(x) - \bar{r}_t(x)}{c_t(x) + \bar{r}_t(x)} = \frac{q_{i,t}(x)(\bar{r}_t(x) + \eta_t)}{(c_t(x) - \eta_t)(c_t(x) + \bar{r}_t(x))} \ge 0. \tag{33}$$

Hence

$$\pi_{t+1}(i) - \pi_t(i) \le \pi_t(i) \frac{r_{i,t}(x) + \eta_t}{c_t(x) - \eta_t}. \tag{34}$$

Now suppose greedy decoding selects x at step t. Let $z \in \arg\max_y c_t(y)$. Since x is selected greedily, $q_t(x) \ge q_t(z)$. Thus,

$$c_t(z) - c_t(x) \le \big(q_t(x) - c_t(x)\big) - \big(q_t(z) - c_t(z)\big) \le \sqrt{2}\, \|q_t - c_t\|_2 = \sqrt{2}\eta_t. \tag{35}$$

In other words,

$$c_t(x) \ge \max_y c_t(y) - \sqrt{2}\eta_t. \tag{36}$$

For the likelihood-ratio bound, greedy selection and $\|q_t - c_t\|_\infty \le \eta_t$ give

$$q_t(x) \ge q_t(z) \ge c_t(z) - \eta_t = \max_y c_t(y) - \eta_t. \tag{37}$$

When $\max_y c_t(y) - \eta_t > \pi_t^\star$, using $q_t = \pi_t^\star q_t^\star + (1 - \pi_t^\star)\bar{q}_t$ and (37), we have

$$(1 - \pi_t^\star)\bar{q}_t(x) = q_t(x) - \pi_t^\star q_t^\star(x) \ge \max_y c_t(y) - \eta_t - \pi_t^\star > 0. \tag{38}$$

Consequently,

$$L_t(x) = \frac{q_t^\star(x)}{\bar{q}_t(x)} \le \frac{1}{\bar{q}_t(x)} \le \frac{1 - \pi_t^\star}{\max_y c_t(y) - \eta_t - \pi_t^\star}. \tag{39}$$

We now prove the hitting-time bound. Along the greedy trajectory before $\tau_{p_0}^i$, let x_s be the greedy token and let $x_s^{\text{cons}} \in \arg\max_y c_s(y)$ be the top consensus token. Then

$$q_s(x_s) \ge q_s(x_s^{\text{cons}}) = \max_y c_s(y) + \bar{r}_s(x_s^{\text{cons}}) \ge \underline{c} - \overline{\eta}. \tag{40}$$

Bayes' rule gives

$$\pi_{s+1}(i) = \pi_s(i) \frac{q_{i,s}(x_s)}{q_s(x_s)} \le \frac{\pi_s(i)}{\underline{c} - \overline{\eta}}. \tag{41}$$

At the hitting time, $\pi_{t+\tau_{p_0}^i}(i) \ge p_0$, so

$$\tau_{p_0}^i \ge \frac{\log(\pi_t(i)/p_0)}{\log(\underline{c} - \overline{\eta})}. \tag{42}$$

as long as $0 < \underline{c} - \overline{\eta} < 1$.

Finally, under temperature-1 sampling, each fixed branch posterior $\pi_s(i)$ is a martingale. Let $i_t^\star$ be the dominant branch at time t, and let j be a runner-up branch with $\pi_t(j) = \pi_t^\sharp$. Note that the gap

$$\Delta_s := \pi_s(i_t^\star) - \pi_s(j). \tag{43}$$

is also a martingale with $\Delta_t = \pi_t^\star - \pi_t^\sharp$. Since to require $i_t^\star$ to reach posterior mass p_0 ($p_0 > 1/2$) before switching is to require this gap Δ_s to hit $2p_0 - 1$ before it crosses zero, we can apply a standard gambler's-ruin argument for a martingale with bounded overshoot. The rest of the proof therefore focuses on bounding the overshoot of Δ_s when it first crosses zero.

At each step s, Bayes' rule gives

$$\pi_{s+1}(k) = \pi_s(k) \frac{q_{k,s}(x_s)}{q_s(x_s)} \tag{44}$$

for every branch k. Hence, for the fixed pair $(i_t^\star, j)$,

$$\Delta_{s+1} - \Delta_s = \frac{\pi_s(i_t^\star)\left(q_{i_t^\star,s}(x_s) - q_s(x_s)\right) - \pi_s(j)\left(q_{j,s}(x_s) - q_s(x_s)\right)}{q_s(x_s)}. \tag{45}$$

Since for all k,

$$|q_{k,s}(x_s) - q_s(x_s)| = |r_{k,s}(x_s) - \bar{r}_s(x_s)| \le \max_{u,v} |r_{u,s}(x_s) - r_{v,s}(x_s)|, \tag{46}$$

the one-step gap movement is bounded by

$$|\Delta_{s+1} - \Delta_s| \le \frac{\max_{u,v} |r_{u,s}(x_s) - r_{v,s}(x_s)|}{q_s(x_s)} \le \zeta. \tag{47}$$

Thus the first downward crossing of zero has negative overshoot at most ζ.

In summary, let

$$a := 2p_0 - 1, \qquad \tau_a := \inf\{s \ge t : \Delta_s \ge a\}, \qquad \tau_0 := \inf\{s \ge t : \Delta_s \le 0\}. \tag{48}$$

Since Δ_s is a martingale with bounded overshoot, the optional stopping theorem gives

$$\Delta_t = \mathbb{E}[\Delta_{\tau_a \wedge \tau_0} \mid x_{<t}] \tag{49}$$

$$\ge a \Pr(\tau_a < \tau_0 \mid x_{<t}) - \zeta \Pr(\tau_0 < \tau_a \mid x_{<t}). \tag{50}$$

Hence

$$\Pr(\tau_a < \tau_0 \mid x_{<t}) \le \frac{\Delta_t + \zeta}{a + \zeta} = \frac{\pi_t^\star - \pi_t^\sharp + \zeta}{2p_0 - 1 + \zeta}. \tag{51}$$

□

F Proof of Proposition 3

Proof of Proposition 3. For the simplicity of notations, define

$$p_k := \pi_t^k(i_k^\star), \qquad h_k := \frac{1}{N_{\text{eff}}(\pi_t^{k+1})}, \qquad h_k^x := \frac{1}{N_{\text{eff}}(\pi_t^{k+1,x})}. \tag{52}$$

We start by deriving a closed-form expression for the numerator of $\Lambda_t^k(x)$, i.e., $N_{\text{eff}}(\pi_t^k) - 1$. Since π_t^{k+1} is π_t^k renormalized on S_{k+1},

$$\sum_{j \in S_{k+1}} \pi_t^k(j)^2 = (1-p_k)^2 h_k. \tag{53}$$

By definition of u_k, we have $p_k = 1/(1+u_k)$ and $1 - p_k = u_k/(1+u_k)$. Thus the collision of π_t^k is

$$\sum_i \pi_t^k(i)^2 = p_k^2 + (1-p_k)^2 h_k = \frac{1 + u_k^2 h_k}{(1+u_k)^2}. \tag{54}$$

Therefore, the numerator becomes

$$N_{\text{eff}}(\pi_t^k) - 1 = \frac{2u_k + u_k^2(1-h_k)}{1 + u_k^2 h_k}. \tag{55}$$

Next, we derive the denominator $N_{\text{eff}}(\pi_t^{k,x}) - 1$. Since

$$\sum_{i \in S_k} \pi_t^k(i) q_{i,t}(x) = p_k q_{i_k^\star,t}(x) + (1-p_k)\bar{q}_k(x) = p_k \bar{q}_k(x)(L_k(x) + u_k), \tag{56}$$

Bayes' rule gives

$$\pi_t^{k,x}(i_k^\star) = \frac{L_k(x)}{L_k(x) + u_k}. \tag{57}$$

The remaining posterior mass is $u_k/(L_k(x) + u_k)$, and its normalized distribution over S_{k+1} is $\pi_t^{k+1,x}$. Hence

$$\sum_i \pi_t^{k,x}(i)^2 = \frac{L_k(x)^2 + u_k^2 h_k^x}{(L_k(x) + u_k)^2}. \tag{58}$$

Consequently,

$$N_{\text{eff}}(\pi_t^{k,x}) - 1 = \frac{2u_k L_k(x) + u_k^2(1-h_k^x)}{L_k(x)^2 + u_k^2 h_k^x}. \tag{59}$$

Dividing (55) by (59) yields

$$\Lambda_t^k(x) = \frac{\big(2 + u_k(1-h_k)\big)\big(L_k(x)^2 + u_k^2 h_k^x\big)}{\big(1 + u_k^2 h_k\big)\big(2L_k(x) + u_k(1-h_k^x)\big)}. \tag{60}$$

It remains to rewrite this expression in the stated recursive form. By the definitions of α_k and β_k,

$$\alpha_k = \frac{1}{1 + u_k^2 h_k}, \qquad 1 - \alpha_k = \frac{u_k^2 h_k}{1 + u_k^2 h_k}, \tag{61}$$

and

$$\beta_k = \frac{2}{2 + u_k(1-h_k)}, \qquad 1 - \beta_k = \frac{u_k(1-h_k)}{2 + u_k(1-h_k)}. \tag{62}$$

Moreover,

$$\Lambda_t^{k+1}(x) = \frac{N_{\text{eff}}(\pi_t^{k+1}) - 1}{N_{\text{eff}}(\pi_t^{k+1,x}) - 1} = \frac{h_k^x(1-h_k)}{h_k(1-h_k^x)}. \tag{63}$$

Since $G_k(x) = (1 - h_k^x)/(1 - h_k)$, this implies

$$G_k(x)\Lambda_t^{k+1}(x) = \frac{h_k^x}{h_k}. \tag{64}$$

Therefore

$$\alpha_k L_k(x)^2 + (1 - \alpha_k)G_k(x)\Lambda_t^{k+1}(x) = \frac{L_k(x)^2 + u_k^2 h_k^x}{1 + u_k^2 h_k}, \tag{65}$$

and

$$\beta_k L_k(x) + (1 - \beta_k)G_k(x) = \frac{2L_k(x) + u_k(1 - h_k^x)}{2 + u_k(1 - h_k)}. \tag{66}$$

Taking the ratio of them leads to (16).

Finally, when $k = |I| - 2$, the set S_{k+1} contains a single branch, so $h_k = h_k^x = 1$. Substituting $h_k = h_k^x = 1$ into (60) gives

$$\Lambda_t^k(x) = \frac{L_k(x)^2 + u_k^2}{L_k(x)(1 + u_k^2)}. \tag{67}$$

This is (17). □

G Order-agnostic peeling, diffuse dust, and hidden lock-in

The main text uses dominant-first peeling because it gives the parameter u_k a clean regime interpretation: small u_k means that the current residual set has a dominant branch, while large u_k means that no branch in the residual set dominates. The recursive identity itself, however, is algebraic and does not depend on this ordering. This appendix records the order-agnostic form of the recursion and uses it to clarify a useful distinction: majority posterior mass is sufficient for lock-in-like behavior, but it is not necessary. What matters is whether the remaining posterior mass contains credible competitors or only diffuse low-collision dust.

Proposition 5 (Order-agnostic recursive diagnosticity)**.** *Let $S_0 = I$, and $\langle i_0, \dots, i_{|I|-1} \rangle$ be an arbitrary order of the branches. At step t, recursively define*

$$\pi_t^k(i) := \frac{\pi_t(i)\mathbf{1}\{i \in S_k\}}{\sum_{j \in S_k} \pi_t(j)},\ S_{k+1} := S_k \setminus \{i_k\},\ u_k := \frac{1 - \pi_t^k(i_k)}{\pi_t^k(i_k)}, \tag{68}$$

$$\Lambda_t^k(x) := \frac{N_{\text{eff}}(\pi_t^k) - 1}{N_{\text{eff}}(\pi_t^{k,x}) - 1},\ \ \bar{q}_k(x) := \frac{\sum_{j \in S_k \setminus \{i_k\}} \pi_t^k(j)\, q_{j,t}(x)}{\sum_{j \in S_k \setminus \{i_k\}} \pi_t^k(j)},\ \ L_k(x) := \frac{q_{i_k,t}(x)}{\bar{q}_k(x)}. \tag{69}$$

where $\pi_t^{k,x}$ represents the posterior π_t^k after observing x. Let

$$\alpha_k := \frac{1}{1 + u_k^2 / N_{\text{eff}}(\pi_t^{k+1})},\ \beta_k := \frac{2}{2 + u_k(1 - 1/N_{\text{eff}}(\pi_t^{k+1}))}. \tag{70}$$

Then, whenever $k < |I| - 2$, $\Lambda_t^k(x)$ admits the recursive decomposition

$$\Lambda_t^k(x) = \frac{\alpha_k L_k(x)^2 + (1 - \alpha_k) G_k(x) \Lambda_t^{k+1}(x)}{\beta_k L_k(x) + (1 - \beta_k) G_k(x)}. \tag{71}$$

where $G_k(x) = (1 - 1/N_{\text{eff}}(\pi_t^{k+1,x}))/(1 - 1/N_{\text{eff}}(\pi_t^{k+1}))$. The recursion terminates at

$$\Lambda_t^{|I|-2}(x) = \frac{L_{|I|-2}(x)^2 + u_{|I|-2}^2}{L_{|I|-2}(x)(1 + u_{|I|-2}^2)}. \tag{72}$$

The proof follows directly from that of Proposition 3. The order-agnostic identity in this new proposition lets us distinguish two different obstructions to concentration. A pivot branch is nearly transparent when the recursion passes diagnosticity to the residual suffix with little change. To see this, rewrite (71) as

$$\frac{\Lambda_t^k(x)}{\Lambda_t^{k+1}(x)} = \frac{1 + \alpha_k \left(\frac{L_k(x)^2}{G_k(x) \Lambda_t^{k+1}(x)} - 1 \right)}{1 + \beta_k \left(\frac{L_k(x)}{G_k(x)} - 1 \right)}. \tag{73}$$

Thus if α_k and β_k are small, the local pivot likelihood ratio $L_k(x)$ is attenuated and the diagnosticity calculation is passed to the residual suffix.

The quantities controlling this transparency have a simple interpretation. Let

$$C_k^{\text{res}} := \sum_{j \in S_{k+1}} \pi_t^k(j)^2, \qquad D_k^{\text{res}} := \sum_{\substack{j, \ell \in S_{k+1} \\ j \neq \ell}} \pi_t^k(j) \pi_t^k(\ell). \tag{74}$$

Then

$$\frac{1}{\alpha_k} - 1 = \frac{u_k^2}{N_{\text{eff}}(\pi_t^{k+1})} = \frac{C_k^{\text{res}}}{p_k^2}, \qquad \frac{2}{\beta_k} - 2 = u_k \left(1 - \frac{1}{N_{\text{eff}}(\pi_t^{k+1})} \right) = \frac{D_k^{\text{res}}}{p_k(1 - p_k)}. \tag{75}$$

The first ratio compares the residual collision mass to the squared mass of the pivot, while the second compares the residual different-branch mass to the pivot-residual cross mass. Diffuse low-mass dust makes these ratios large, which makes α_k and β_k small. Such dust can be peeled away without substantially changing the returned value of Λ. By contrast, a credible competitor has non-negligible mass or collision relative to the current residual suffix; peeling it does not merely pass diagnosticity through, but changes the decomposition.

One can make this pass-through statement explicit. Define

$$A_k(x) := \frac{L_k(x)^2}{G_k(x)\Lambda_t^{k+1}(x)} - 1, \qquad B_k(x) := \frac{L_k(x)}{G_k(x)} - 1. \tag{76}$$

If $|\alpha_k A_k(x)| \le 1/2$ and $|\beta_k B_k(x)| \le 1/2$, then

$$\left|\log \frac{\Lambda_t^k(x)}{\Lambda_t^{k+1}(x)}\right| \le 2\alpha_k |A_k(x)| + 2\beta_k |B_k(x)|. \tag{77}$$

Consequently, if the first m pivots are transparent in this sense, then

$$\left|\log \Lambda_t^0(x) - \log \Lambda_t^m(x)\right| \le 2\sum_{k=0}^{m-1} \alpha_k |A_k(x)| + 2\sum_{k=0}^{m-1} \beta_k |B_k(x)|. \tag{78}$$

Thus many branches can be present without obstructing lock-in-like diagnosticity, provided they are diffuse dust whose cumulative pass-through correction is small.

Hidden lock-in behind diffuse dust. The preceding arguments refine the sufficient condition $N_{\mathrm{eff}}(\pi_t) < 2$ used in Proposition 1. Majority mass is enough to make the leading branch dominant, but it is not necessary. A branch with small absolute posterior mass can still behave lock-in-like if its complement is fragmented into many low-collision alternatives. For example, suppose a branch has mass p and the remaining $1 - p$ mass is spread over M roughly equal dust branches. Then the residual collision is approximately

$$h \approx \frac{1}{M}, \tag{79}$$

so

$$u^2 h \approx \left(\frac{1-p}{p}\right)^2 \frac{1}{M}. \tag{80}$$

If $M \gg p^{-2}$, then the residual collision mass is small compared with the branch's squared mass, even when $p \ll 1$. The branch is not a majority branch, but it is separated from the complement in collision geometry. In this case the dust layers can be nearly transparent, and the returned diagnosticity is governed by the likelihood ratio that separates the branch from the residual mixture.

This is the sense in which lock-in is not merely a statement about high absolute posterior mass. The real obstruction to lock-in is a coherent challenger. If the complement of a branch consists of diffuse dust, the recursion can pass through that dust until the branch becomes dominant relative to the remaining suffix. If the complement contains a credible competitor, however, it cannot be peeled away without significantly attenuating or redirecting the returned value of Λ.

Implication for decoding. The same distinction clarifies greedy versus off-greedy behavior. Greedy decoding is poor in confusion when the selected token lies in a shared backbone. At a residual level k, writing $q_{i,t}(x) = c_t(x) + r_{i,t}(x)$ gives

$$L_k(x) = \frac{c_t(x) + r_{b_k,t}(x)}{c_t(x) + \bar{r}_k(x)} = 1 + \frac{r_{b_k,t}(x) - \bar{r}_k(x)}{c_t(x) + \bar{r}_k(x)}. \tag{81}$$

When $x = g_t$ is a high-consensus greedy token, the large shared term $c_t(g_t)$ attenuates residual contrasts and keeps $L_k(g_t) \approx 1$ across many peel levels. The recursion then returns $\Lambda_t^0(g_t) \approx 1$, leaving the trace unresolved. An off-greedy token with smaller consensus mass is riskier: the returned likelihood ratio may be above or below one. But it is also more likely to expose a credible residual split. Thus, while off-greedy alternatives are not guaranteed to improve the trace, they are valuable because they give the recursion a chance to return a branch-resolving update that greedy consensus tokens often cannot produce.

H Token versus Branch Effective Support

The following proposition identifies the gap between token and branch effective support through two slack variables ξ_t and γ_t. For $N_{\text{eff}}(q_t)$ to track $N_{\text{eff}}(\pi_t)$, within-branch collision ξ_t must remain substantial and cross-branch overlap γ_t must remain small. In practice, the serious difficulty is the latter. Distinct branches can share a common lexical head, so the observable support $N_{\text{eff}}(q_t)$ can substantially *underestimate* branch support in the confusion regime, where many branches agree on the same locally plausible continuations.

Proposition 6 (Token versus branch effective support)**.** *Define* $\xi_t := \min_i \sum_x q_{i,t}(x)^2$ *and* $\gamma_t := \max_{i \neq j} \sum_x q_{i,t}(x) q_{j,t}(x)$ *as the min intra- and max inter-branch collision, respectively. Then*

$$\frac{N_{\text{eff}}(\pi_t)}{(1-\gamma_t) + \gamma_t N_{\text{eff}}(\pi_t)} \leq N_{\text{eff}}(q_t) \leq \frac{N_{\text{eff}}(\pi_t)}{\xi_t}. \tag{82}$$

Proof. Write

$$C_\pi := \sum_i \pi_t(i)^2 = \frac{1}{N_{\text{eff}}(\pi_t)}, \qquad C_q := \sum_x q_t(x)^2 = \frac{1}{N_{\text{eff}}(q_t)}.$$

Using $q_t = \sum_i \pi_t(i) q_{i,t}$ and the standard inner product notation $\langle a, b \rangle = \sum_x a(x) b(x)$, we have

$$C_q = \left\langle \sum_i \pi_t(i) q_{i,t}, \sum_j \pi_t(j) q_{j,t} \right\rangle \tag{83}$$

$$= \sum_i \pi_t(i)^2 \langle q_{i,t}, q_{i,t} \rangle + \sum_{i \neq j} \pi_t(i) \pi_t(j) \langle q_{i,t}, q_{j,t} \rangle. \tag{84}$$

For the upper bound on $N_{\text{eff}}(q_t)$, all off-diagonal terms in (84) are nonnegative, and $\langle q_{i,t}, q_{i,t} \rangle \geq \xi_t$ by definition. Hence

$$C_q \geq \sum_i \pi_t(i)^2 \langle q_{i,t}, q_{i,t} \rangle \geq \xi_t \sum_i \pi_t(i)^2 = \xi_t C_\pi. \tag{85}$$

Taking reciprocals gives

$$N_{\text{eff}}(q_t) = \frac{1}{C_q} \leq \frac{1}{\xi_t C_\pi} = \frac{N_{\text{eff}}(\pi_t)}{\xi_t}.$$

For the lower bound, each $q_{i,t}$ is a probability distribution, so $\langle q_{i,t}, q_{i,t} \rangle = \sum_x q_{i,t}(x)^2 \leq 1$. Also, by definition, $\langle q_{i,t}, q_{j,t} \rangle \leq \gamma_t$ for $i \neq j$. Therefore

$$C_q \leq \sum_i \pi_t(i)^2 + \gamma_t \sum_{i \neq j} \pi_t(i) \pi_t(j) \tag{86}$$

$$= C_\pi + \gamma_t (1 - C_\pi) = \gamma_t + (1 - \gamma_t) C_\pi. \tag{87}$$

Since $C_\pi = 1/N_{\text{eff}}(\pi_t)$, this becomes

$$C_q \leq \gamma_t + \frac{1 - \gamma_t}{N_{\text{eff}}(\pi_t)} = \frac{(1-\gamma_t) + \gamma_t N_{\text{eff}}(\pi_t)}{N_{\text{eff}}(\pi_t)}. \tag{88}$$

Taking reciprocals yields

$$N_{\text{eff}}(q_t) \geq \frac{N_{\text{eff}}(\pi_t)}{(1-\gamma_t) + \gamma_t N_{\text{eff}}(\pi_t)}.$$

Combining the two inequalities proves (82). □

I Proof of Proposition 4

Proof. For readability, write

$$\tilde{\pi}_i := \tilde{\pi}_t(i), \qquad \rho_i := \rho_{i,t}, \qquad K := K_t, \qquad \rho := \rho_t.$$

Since

$$\rho = \sum_i \tilde{\pi}_i \rho_i,$$

we have

$$\sum_x \rho(x)^2 = \left\langle \sum_i \tilde{\pi}_i \rho_i,\ \sum_j \tilde{\pi}_j \rho_j \right\rangle \tag{89}$$

$$= \sum_{i,j} \tilde{\pi}_i \tilde{\pi}_j \langle \rho_i, \rho_j \rangle \tag{90}$$

$$= \sum_{i,j} \tilde{\pi}_i \tilde{\pi}_j K_{ij} \tag{91}$$

$$= \tilde{\pi}_t^\top K_t \tilde{\pi}_t. \tag{92}$$

Therefore

$$D_t = N_{\text{eff}}(\rho_t) = \frac{1}{\sum_x \rho_t(x)^2} = \frac{1}{\tilde{\pi}_t^\top K_t \tilde{\pi}_t}. \tag{93}$$

We first prove the lower bound. Since K_t is a Gram matrix, it is positive semidefinite. Hence

$$\tilde{\pi}_t^\top K_t \tilde{\pi}_t \le \lambda_{\max}(K_t)\, \|\tilde{\pi}_t\|_2^2.$$

Using

$$\|\tilde{\pi}_t\|_2^2 = \sum_i \tilde{\pi}_t(i)^2 = \frac{1}{N_{\text{eff}}(\tilde{\pi}_t)},$$

we obtain

$$\tilde{\pi}_t^\top K_t \tilde{\pi}_t \le \frac{\lambda_{\max}(K_t)}{N_{\text{eff}}(\tilde{\pi}_t)}.$$

Combining this with (93) gives

$$D_t \ge \frac{N_{\text{eff}}(\tilde{\pi}_t)}{\lambda_{\max}(K_t)}.$$

Next we prove the upper bound. Since each $\rho_{i,t}$ is a probability vector, all entries of K_t are nonnegative:

$$(K_t)_{ij} = \sum_x \rho_{i,t}(x) \rho_{j,t}(x) \ge 0.$$

Therefore

$$\tilde{\pi}_t^\top K_t \tilde{\pi}_t = \sum_i \tilde{\pi}_t(i)^2 (K_t)_{ii} + \sum_{i \ne j} \tilde{\pi}_t(i) \tilde{\pi}_t(j) (K_t)_{ij} \tag{94}$$

$$\ge \sum_i \tilde{\pi}_t(i)^2 (K_t)_{ii}. \tag{95}$$

By the definition of $\bar{\xi}_t$,

$$\sum_i \tilde{\pi}_t(i)^2 (K_t)_{ii} = \bar{\xi}_t \sum_i \tilde{\pi}_t(i)^2 = \frac{\bar{\xi}_t}{N_{\text{eff}}(\tilde{\pi}_t)}.$$

Hence

$$\tilde{\pi}_t^\top K_t \tilde{\pi}_t \ge \frac{\bar{\xi}_t}{N_{\text{eff}}(\tilde{\pi}_t)}.$$

Applying (93) once more yields

$$D_t \le \frac{N_{\text{eff}}(\tilde{\pi}_t)}{\bar{\xi}_t}.$$

This proves (18). □

J Implementation Details and Robustness checks

Table 5: Level-test robustness of D. *Top*: across Zipf fit windows $W = [a, b]$: number of cells (of 30: 5 models $\times$ 2 decoding strategies $\times$ 3 benchmarks) in which median D on correct traces is lower than on wrong traces. Bold marks the default $W = [3, 50]$ used throughout the paper. *Second row*: across alternative heavy-tailed null families substituted for Zipf in t_q (exponential is the light-tailed baseline). *Bottom four panels*: alternative scalar observables: Rényi-2 and Rényi-∞ entropies, $N_{\text{eff}}(q_t)$, log probability of the top token, margin between the top two tokens; volatility measure computed pairwise on consecutive steps: total variation distance, KL divergence, and JS divergence; tail mass $\tau_K = \sum_{i>K} p_i$; tail collision $c_K = \sum_{i>K} p_i^2$.

W	$[1,50]$	$[1,100]$	$[2,50]$	$[2,100]$	$[3,30]$	$[\mathbf{3},\mathbf{50}]$	$[3,70]$	$[3,100]$	$[4,50]$	$[4,100]$
Cells	28	30	30	30	28	**30**	30	29	29	29

Head family	Mandelbrot	Log-normal	Stretched exp.	Trunc. power	Exponential
Cells	29	29	29	29	23

q_t	Rényi-2	Rényi-∞	Eff. supp.	$\log p_1$	margin
Cells	20	22	22	22	21

Volatility	TV	KL	JS
Cells	18	18	21

Tail mass	τ_1	τ_{50}	τ_{200}	τ_{500}
Cells	23	25	25	25

Tail collision	c_1	c_{50}	c_{200}	c_{500}
Cells	16	22	23	24

Table 6: Robustness check for the main intervention table (Table 3). Each row reports the aggregate (5 models $\times$ 3 benchmarks) median ΔD and percentage of $\Delta D > 0$ under a different intervention setting: varying the number of highest-H positions intervened on (top-k) and the rank of the forced tail token. Row counts differ slightly from the nominal total when the required forced-rank token or intervention position is unavailable. Per-row largest **bolded**, smallest underlined. The main setting is top-3 H, rank-3.

Setting	n	W→C	W→W	C→C	C→W
top-1 H, r=2	1500	$\underline{-12.9}$ (40%)	$+0.8$ (52%)	$+3.1$ (56%)	$+\mathbf{6.1}$ (59%)
top-2 H, r=2	3000	$\underline{-15.6}$ (32%)	$+0.6$ (52%)	$+1.7$ (55%)	$+\mathbf{6.8}$ (60%)
top-3 H, r=2	4500	$\underline{-13.3}$ (33%)	$+0.1$ (50%)	$+1.2$ (54%)	$+\mathbf{6.6}$ (60%)
top-1 H, r=3	1500	$\underline{-5.1}$ (43%)	$+1.7$ (53%)	$+1.7$ (56%)	$+\mathbf{10.2}$ (64%)
top-2 H, r=3	3000	$\underline{-2.1}$ (44%)	$+1.1$ (52%)	$+1.6$ (55%)	$+\mathbf{10.4}$ (63%)
top-3 H, r=3 (main)	4492	$\underline{-2.1}$ (45%)	$+0.7$ (52%)	$+1.7$ (55%)	$+\mathbf{8.1}$ (61%)
top-1 H, r=7	1500	$\underline{-15.7}$ (45%)	$+3.1$ (55%)	$+6.3$ (62%)	$+\mathbf{16.0}$ (66%)
top-2 H, r=7	2987	$\underline{-6.7}$ (46%)	$+3.9$ (55%)	$+6.3$ (62%)	$+\mathbf{17.0}$ (66%)
top-3 H, r=7	4413	$\underline{-6.7}$ (45%)	$+5.7$ (57%)	$+6.2$ (62%)	$+\mathbf{16.0}$ (66%)

Table 7: Robustness of the by-regime intervention outcome to the endpoint-split choice. Each row represents a different split. Within each row, the larger of $D\downarrow$ vs $D\uparrow$ for fix rate, broke rate, and net is **bolded**.

	$D\downarrow$ traces						$D\uparrow$ traces					
Scheme	n	med ΔD	$\Delta D>0$	fixed (rate)	broke (rate)	net	n	med ΔD	$\Delta D>0$	fixed (rate)	broke (rate)	net
1/3	3120	$+3.6$	59%	135 (10%)	369 (**21%**)	-234	1372	-2.7	43%	111 (**13%**)	89 (17%)	**+22**
1/4 (main)	3229	$+3.3$	58%	126 (9%)	379 (**20%**)	-253	1263	-2.1	43%	120 (**15%**)	79 (18%)	**+41**
1/5	3247	$+3.0$	58%	125 (9%)	374 (**20%**)	-249	1245	-1.0	44%	121 (**15%**)	84 (18%)	**+37**

Implementation of residual-support statistic. We fit a Zipf profile $\hat{c}_t$ with slope $\hat{s}_t$ on a head window W:

$$\hat{s}_t = \arg\min_s \sum_{x \in W} \left(\log q_t(x) + s \log k_t(x) - \frac{1}{|W|} \sum_{x' \in W} \left(\log q_t(x') + s \log k_t(x') \right) \right)^2, \quad (96)$$

where $k_t(x)$ is the rank of x in q_t. Define

$$\rho_t(x) \propto \big(q_t(x) - \hat{c}_t(k_t(x))\big)_+, \qquad \widehat{D_t} := N_{\mathrm{eff}}(\rho_t), \tag{97}$$

and write $D := \widehat{D_t}$. We use rank window $W = \{3, \ldots, 50\}$ throughout; see below for robustness over alternative head distributions, head windows, and observables. Important to note that D is not an estimator of branch support or a predictor of correctness; it is a vehicle for testing whether the predicted regime signatures survive a parsimonious residualization.

Overview of robustness checks. We report robustness checks for the three main empirical claims of §8. Each table varies the key design parameter of one claim while holding the others fixed: the Zipf head-stripping window for the level test (Table 5), the intervention design (k, r) for the transition-pattern result (Table 6), and the endpoint-split quantile for the by-regime outcome contrast (Table 7). Across all three checks, the qualitative conclusions used in the main text are preserved under reasonable changes to the operationalization.

Level test (Table 5). The residual-support statistic D is defined by stripping the shared Zipf head over a fit window $W = [a, b]$; the main text uses $W = [3, 50]$. To check that the correct-vs-wrong separation does not hinge on this choice, we recompute D under five alternative head families (Mandelbrot, log-normal, stretched-exponential, truncated power-law, and exponential) and ten alternative windows spanning $a \in \{1, 2, 3, 4\}$ and $b \in \{30, 50, 70, 100\}$. The pattern is stable across alternative heavy-tailed families (29/30 or better cells for Mandelbrot, log-normal, stretched exponential, and truncated power law), but weakens for an exponential head (23/30), consistent with the idea that the operative ingredient is a heavy-tailed structured head rather than the Zipf form specifically. In terms of fit windows, for each window we count, out of 30 cells (5 models $\times$ 2 decoding strategies $\times$ 3 benchmarks), the number in which median D on correct traces is lower than on wrong traces. The ordering holds in at least $28/30$ cells for every window, and in all $30/30$ cells for five of the ten windows including the default. The level claim is therefore not an artifact of the default fit window.

We also calculated a wide range of scalar observables drawn from the raw token distribution—entropies (Shannon, Rényi-2/∞), effective support, head confidence ($\log p_1$, margin), trajectory volatility (TV/KL/JS between consecutive steps), raw tail masses τ_K, and tail collisions c_K—to see whether any of them aligns as well with the two signals driving our analysis: correctness and D. None does: full-distribution summaries cap at 22/30, volatility at 21/30, raw tail mass at 25/30.

Intervention transitions (Table 6). The main-text intervention uses top-k=3 highest-entropy positions and a rank-r=3 forced alternative. Here we vary both parameters jointly over $k \in \{1, 2, 3\}$ and $r \in \{2, 3, 7\}$, giving all nine combinations, pooled across 5 models and 3 benchmarks. The qualitative transition pattern is preserved in every setting: median ΔD is most negative on W→C branches—consistent with a confusion-like trace being redirected toward a more concentrated continuation—and most positive on C→W branches—consistent with the mirror disruption of a lock-in-like trace toward unresolvedness. Effect magnitudes scale with both k and r, as one would expect from stronger perturbations, but the sign ordering of the four transition cells is invariant across all $3 \times 3 = 9$ settings.

Regime-split endpoint (Table 7). The $D\uparrow$/$D\downarrow$ regime labels are defined by comparing mean D over the first and last $1/4$ of post-prompt tokens. To check that the by-regime contrast does not depend on this endpoint choice, we repeat the analysis with $1/3$ and $1/5$ splits. Across all three splits the direction of the asymmetry is preserved: $D\uparrow$ traces have the higher fix rate (13–15% vs. 9–10%) and the only positive net intervention gain ($+22$ to $+41$), while $D\downarrow$ traces have the higher broke rate (20–21% vs. 17–18%) and a strongly negative net (-234 to -253). The recoverability contrast between the two regime classes is therefore not an artifact of the $1/4$ endpoint.

K Detailed Breakdown of Intervention Outcomes by Regime and Outcome Category

Table 8: Interventions by trace regime and outcome, pooled across 5 models and 3 benchmarks. Same correct/wrong: the branch preserves the greedy answer. Diff wrong: the answer changes to a different wrong answer or to no extractable answer. No answer: a correct-baseline branch whose continuation yields no extractable answer. Note that we account for these separately from Broke because such a branch is an overt disruption by the intervention that is easily detected and discarded in practice. Entries report n, median ΔD with 95% bootstrap CI, and fraction with $\Delta D > 0$.

	$D\downarrow$ traces			$D\uparrow$ traces		
Outcome	n	med ΔD [95% CI]	$\Delta D>0$	n	med ΔD [95% CI]	$\Delta D>0$
Fixed	126	+0.1 [-8.1, +6.7]	51%	120	-16.4 [-49.3, 0.0]	38%
Same wrong	556	+1.1 [0.0, +2.7]	54%	294	0.0 [-3.1, +0.8]	49%
Diff wrong	670	+8.2 [+4.0, +12.7]	60%	411	-7.4 [-18.3, 0.0]	37%
Same correct	1452	+2.2 [+1.4, +3.4]	57%	333	-2.1 [-8.1, +1.1]	46%
Broke	379	+10.8 [+5.6, +14.5]	65%	79	0.0 [-20.5, +5.1]	43%
No answer	46	+41.4 [+8.9, +100.3]	70%	26	0.0 [-61.9, +32.6]	42%

To examine whether the most theory-relevant intervention signatures appear in the expected regime $\times$ outcome cells, Table 8 pools across models and benchmarks and partitions intervention branches jointly by baseline regime class and outcome category. This lets us isolate the two directional signatures suggested by the theory: successful rescue of confusion-like traces, and disruption of lock-in-like traces.

The strongest fix signature appears exactly where the theory predicts it should: in the $D\uparrow$/Fixed cell, whose median ΔD is the most negative in the table (-16.4). This is the expected pattern for a confusion-like trace that is redirected into a more concentrated continuation. Conversely, the strongest break signatures appear in the $D\downarrow$/Broke cell, whose median ΔD is strongly positive (+10.8, 95% CI $[+5.6, +14.5]$), as well as the overt disruption cases in the $D\downarrow$/"No answer" cell (+41.4, 95% CI $[+8.9, +100.3]$). This is the expected pattern for perturbing a lock-in-like trace toward greater unresolvedness. The neighboring cells are correspondingly weaker or more mixed, as they should be if the cleanest directional interpretation belongs specifically to these two transitions.

L Direct validation of D against branch support

The main text validates D through final-answer correctness and intervention response. This appendix reports two experiments that directly compare D with branch support. The first experiment implements a synthetic process in which π_t is known in closed form, while the second experiment uses resampling (of real traces) to estimate the branch posterior under final-answer equivalence.

L.1 Synthetic process with known branch structure

Construction. The synthetic process instantiates the decomposition of (3), i.e., a shared backbone plus branch-specific residuals, with π_t propagated by (1). It uses a vocabulary of 2,000 tokens, $K = 12$ branches, and traces of $T = 120$ steps, with

$$q_t = \sum_i \pi_t(i)\big((1 - \beta_{i,t})\, c_t + \beta_{i,t}\, b_i\big), \tag{98}$$

where the parameters are set as follows.

- The mixing rate $\beta_{i,t}$ is i.i.d. lognormal noise with median 0.02 on boilerplate steps (nine of every ten), 0.35 on thinking steps (every tenth), and log-scale $\sigma = 0.18$ in both. Note that this heterogeneity across branches is essential because otherwise (i.e., with equal rates across branches) all branch laws would coincide on backbone tokens and therefore freeze the Bayesian update.
- The backbone c_t is Zipf-Mandelbrot, $(k + q_0)^{-s_t}$ over ranks $k \in [1, 2000]$ assigned to tokens by a fresh random permutation at every step. We set the shoulder $q_0 \sim U[1, 4]$ per step (without it the rank-window fit recovers the head exactly). On thinking steps, the slope s_t is a trace-level mean $\sim U[0.8, 1.6]$ plus a step-wise jitter $\sim U[-0.2, 0.2]$. On boilerplate steps, it rises linearly from 1.20 to 1.70 along the trace.
- The content law b_i of branch i is fixed over t: the twelve branches form four sibling groups of three, and $b_i = (1 - v)\,\mathrm{own}_i + v \cdot (\text{mean of its two siblings' own laws})$ with $v = 0.35$, where own_i is Zipf over 15 content tokens sampled without replacement (unique across branches) with slope 1 for all branches except the correct one, which has slope 1.3. This is the only asymmetry among branches and is designed to make the correct branch's content tokens slightly likelier to win the argmax race.
- The initial posterior is $\pi_1(i) \propto \exp(\rho\, \mathbf{1}[i = g] + 0.05\,\mathcal{N}(0, 1))$ for the correct branch g, with $\rho \in \{0, 0.3, 0.6, 1.2\}$ in four equal arms of 100 traces (accuracies 20%, 44%, 73%, 98% against a floor of $1/12$). We run greedy decoding and count a trace as correct if and only if g has the maximum posterior at the end. D is computed at every step from q_t alone, exactly as in App. J (Zipf fit on ranks 3–50).

Table 9: Synthetic process: early and late values by outcome class, medians across traces ($n = 400$ traces, 165 wrong / 235 correct, pooled over the four ρ arms). For the true branch support $N_{\mathrm{eff}}(\pi_t)$ and for D (computed from q_t alone), early/late are per-trace medians over the first/last third of the thinking steps; for token entropy H (nats), over the first/last quarter of all steps.

	$N_{\mathrm{eff}}(\pi_t)$ (early $\to$ late)	D (early $\to$ late)	H (early $\to$ late)
wrong traces	$11.75 \to \mathbf{10.92}$	$12.95 \to \mathbf{13.43}$	$5.32 \to 4.07$
correct traces	$9.25 \to \mathbf{1.13}$	$11.61 \to \mathbf{8.93}$	$5.32 \to 4.07$

D tracks branch support (H does not). $\mathrm{Spearman}(D, N_{\mathrm{eff}}(\pi_t)) = +0.40$ across all 4,800 decision points (400 traces $\times$ 12 thinking steps) and $+0.52$ across traces (per-trace medians over the last third of decision points). Token entropy is wrong-signed on both (-0.12 pooled, -0.29 per trace). Table 9 shows the two regimes as predicted: on correct traces the branch posterior resolves ($N_{\mathrm{eff}}(\pi_t)$: $9.25 \to 1.13$, lock-in) while on wrong traces it stays elevated ($11.75 \to 10.92$, confusion); D mirrors the asymmetry ($11.61 \to 8.93$ vs. $12.95 \to 13.43$), whereas H falls identically in both classes. The construction therefore reproduces the paper's central claim: low-entropy traces that are wrong carry unresolved branch support.

Head subtraction is essential for D. The same statistic without the Zipf-head fit (i.e., $N_{\text{eff}}(q_t)$, the effective support of the visible distribution) tracks $N_{\text{eff}}(\pi_t)$ at only $+0.18$ pooled and $+0.16$ at trace level. Worse, among the sharpest third of traces (by backbone slope), $N_{\text{eff}}(q_t)$ orders confusion *below* lock-in (52.2 vs. 63.8 when the true support is 10.7 vs. 1.0), while D preserves the correct ordering (10.0 vs. 5.0). This echoes the mechanism that motivates the design of head removal in D: a sharp consensus head mimics a dominant branch unless it is removed.

Robustness. We reran the pipeline while changing one knob at a time: mixing-rate log-scale ($\sigma \in \{0.10, 0.18, 0.30\}$), the trace-level mean of the thinking-step slope (drawn from $[0.7, 1.5]$, $[0.8, 1.6]$, or $[0.9, 1.7]$), jitter half-width ($0.1/0.2/0.3$), sibling weight ($v \in \{0.2, 0.35, 0.5\}$), and $K \in \{6, 12\}$, and found that the findings remain robust: trace-level tracking remains between $+0.31$ and $+0.64$, and a positive D-gap exists between the known-regime populations in every arm. The construction's remaining knob, the sibling weight v, is the one that could violate the bridge condition when raised, and is therefore left as the control for a failure probe in App. M.

L.2 Clustered experiment on real traces

For each of the $1{,}500$ greedy traces of §8 (5 models $\times$ 3 benchmarks $\times$ 100 problems) we drew 16 temperature-1 continuations at four evenly spaced positions ($5{,}999$ unique positions after one rounded-index collision; $95{,}984$ completions) and clustered the continuations by extracted final answer. The cluster frequencies give a direct estimate $\hat{\pi}_t$ of the branch posterior under the final-answer partition, and $\hat{N}_{\text{eff}} = 1/\sum_i \hat{\pi}_t(i)^2$ its effective support. Table 10 reports per-trace medians over the sampled positions in each half of the trace, then medians across traces.

Table 10: Clustered experiment: estimated branch support $\hat{N}_{\text{eff}}$ and token entropy H (nats) at the sampled positions, per-trace median within each half of the trace, median across traces ($n = 726$ wrong / 774 correct traces with first-half positions; 717 / 769 with second-half positions).

	$\hat{N}_{\text{eff}}$		H	
	first half	second half	first half	second half
wrong traces	4.29	**1.83**	0.57	0.53
correct traces	1.87	**1.00**	0.60	0.53

Support declines in both classes but resolves only on correct traces: late in the trace the measured posterior is at the floor ($\hat{N}_{\text{eff}} = 1.00$) on correct traces and still diffuse (1.83) on wrong ones, while entropy at the same positions is indistinguishable (0.53 vs. 0.53; difference 0.003). The statistic computed from q_t alone aligns with this partition: at trace level ($n = 1{,}500$; D- and H-levels are means over thinking tokens, $H > 0.1$), the resampled probability of reaching the gold answer is predicted by D (Spearman $\rho_D = -0.18$, $p < 10^{-11}$; partial correlation given H, -0.20) but not by entropy ($\rho_H = -0.01$, $p = 0.60$). A trace bootstrap of $|\rho_D| - |\rho_H|$ yields an estimate of $+0.17$ with 95% CI $[+0.10, +0.21]$.

M Sensitivity to the notion of resolution

A regime label is relative to the declared notion of resolution (§3, App. B), and D is informative only for notions that satisfy the bridge condition of Proposition 4. This raises two questions. One is how sensitive the regime label is to a change of resolution, say coarsening from final-answer equivalence to approximate equivalence. The other is how D behaves when the declared notion is *finer* than what the bridge condition allows. We answer these two questions respectively in this appendix.

Coarser notions. The clustered experiment estimates the branch posterior at each of $5{,}999$ sampled positions from 16 resampled continuations, bucketed by their extracted final answer. Under the exact partition two continuations are equivalent iff their answer strings are identical ($2{,}101$ distinct answers in total). We coarsen this partition in four steps: numeric answers are rounded to three, two, or one significant digits (two answers are equivalent iff they agree after rounding; non-numeric answers keep their exact buckets), and finally bucketed by order of magnitude (same sign and power of ten). We recompute $\hat{N}_{\text{eff}}$ at each position and call the position *resolved* if $\hat{N}_{\text{eff}} < 2$ and *diffuse* otherwise.

Table 11: Coarsening the answer partition in the clustered experiment. "Positions merged": share of positions at which the coarsening merges two or more distinct answers. "Labels kept": share of positions whose resolved/diffuse label is unchanged. Last three columns: pooled second-half medians of $\hat{N}_{\text{eff}}$ on wrong and correct traces, and the number of benchmarks with wrong $>$ correct.

partition	buckets	positions merged	rank corr.	labels kept	wrong	correct	wrong $>$ correct
exact answer	2,101	—	1.000	100%	1.83	1.00	3/3
3 significant digits	1,725	3.1%	0.997	99.7%	1.83	1.00	3/3
2 significant digits	676	6.4%	0.993	99.3%	1.80	1.00	3/3
1 significant digit	121	38.1%	0.913	90.8%	1.44	1.00	3/3
order of magnitude	20	62.9%	0.733	71.1%	1.07	1.00	3/3

As can be seen from Table 11, rounding to 3 or 2 significant digits barely changes anything. Rounding to 1 digit merges distinct answers at 38% of positions, yet the regime verdicts hold: 90.8% of positions keep their label, the wrong-vs-correct split persists in every benchmark, and the pooled medians degrade gracefully ($1.83/1.00 \to 1.44/1.00$). Only at order-of-magnitude coarseness, where 20 buckets remain of the original $2{,}101$ and the partition is nearly trivial, does the split largely vanish ($1.07/1.00$). The regime verdicts are thus robust across the coarsenings an analyst might reasonably apply.

Finer notions. The synthetic process offers a clean test case because it is designed with two notions of resolution: the twelve branches form four sibling groups of three. Resolution can thus be declared at the (finer) level of individual branches or at the (coarser) level of groups, and both posteriors can be computed exactly: the group posterior is the sum of π_t over its three branches. Raising the sibling weight from its default $v = 0.35$ to $v = 0.65$ makes siblings share most of their content beyond the common backbone. This violates the bridge condition at the branch level but not at the group level, which lets us test which of the two notions D follows.

Specifically, we generate two sets of 200 traces (50 at each of the four values of ρ) that differ only in v: a test set with $v = 0.65$ and a reference set with the default $v = 0.35$. The two sets share their random seeds, so they are matched trace by trace. Because the level of D varies with the sharpness of the backbone, both sets draw the trace-level mean of the thinking-step slope from $[0.8, 1.07]$, the flattest third of its usual range $[0.8, 1.6]$, so that D values can be compared across the two sets. For each trace we record the late values (per-trace medians over the last third of the thinking steps, as in Table 9) of three quantities: the branch-level support $N_{\text{eff}}(\pi_t)$, the group-level support computed in the same way from the group posterior, and D.

In the reference set, where the bridge condition holds, the two levels agree on every trace: 166 of the 200 traces end resolved at both levels (median $N_{\text{eff}} = 1.00$ at both) and the remaining 34 end unresolved at both. D follows them: traces that end in lock-in at the branch level ($N_{\text{eff}} < 2$; $n = 166$) have median $D = 3.2$, whereas those that end in confusion ($N_{\text{eff}} > 6$; $n = 32$) have median $D = 15.2$.

In the test set the two levels come apart. 152 of the 200 traces end resolved at the group level (median group-level $N_{\text{eff}} = 1.00$), but only 48 of them are also resolved at the branch level. The other 104 remain spread over the siblings of the winning group (median branch-level $N_{\text{eff}} = 2.37$ across the 152). Over these 152 traces, the median late D is 7.1, under half the confusion level (15.2) and toward the lock-in level (3.2). Further, D barely varies whether or not the sibling choice is resolved (7.5 vs. 6.9). In other words, D sides with the coarser notion, reporting the resolution among groups (which satisfies the bridge condition) instead of the unresolved choice among siblings (which violates the bridge condition). This is the failure mode Proposition 4 specifies and the reason why notions finer than the bridge condition permits are outside the scope of D. Together with Table 11, this brackets where D applies: a notion is served as long as it is no finer than the bridge condition allows and no coarser than the point at which the partition becomes trivial.

N Intervention controls, regressions, and effect sizes

The experimental results in the main text report raw ΔD (Table 3). Due to the previously discussed heterogeneity of the wrong-baseline pool (§8), ΔD varies widely within each cell, yielding wide per-cell confidence intervals. In this appendix, we re-analyze the experimental results with a distribution-free effect size, the probability of superiority $\theta = \Pr(\Delta D_{\mathrm{W}\to\mathrm{C}} < \Delta D_{\mathrm{W}\to\mathrm{W}})$ (ties counted $1/2$), so that $\theta > 0.5$ means fixes (W $\to$ C) reduce D more than non-fixes (W $\to$ W). More specifically, we compute this effect size from within-stratum pairs (stratum = model $\times$ benchmark, so ΔD is never compared across models with different scales of D) and then combine its values across strata with pair-count weights. Confidence intervals are computed from a trace-level cluster bootstrap (5,000 resamples), since the three branches of a trace share a baseline. Later in the appendix, we also report robustness checks on θ by varying or controlling for the observable properties of an intervention, including its placement, the forced token's probability, its position in the trace, and trace length.

Table 12: Rank-based effect size for the intervention asymmetry at rank 3: $\theta = \Pr(\Delta D_{\mathrm{W}\to\mathrm{C}} < \Delta D_{\mathrm{W}\to\mathrm{W}})$ (> 0.5: fixes reduce D more than non-fixes), with trace-clustered 95% bootstrap CIs. †: CI excludes 0.5.

Model	$n_{\mathrm{W}\to\mathrm{C}}$	$n_{\mathrm{W}\to\mathrm{W}}$	θ	95% CI
Qwen 1.5B	54	279	0.588	[0.490, 0.682]
Qwen 7B	19	137	0.591	[0.364, 0.770]
Gemma 7B	54	692	0.525	[0.440, 0.615]
Granite 8B	67	293	0.571	[0.488, 0.656]
Mistral 7B	52	530	$0.598^{\dagger}$	[0.501, 0.700]
All (stratified)	246	1931	$\mathbf{0.564}^{\dagger}$	[0.519, 0.612]

Effect size and significance. As can be seen from Table 12, pooled across models and benchmarks, $\theta = 0.564$ with 95% CI $[0.519, 0.612]$ and cluster-bootstrap one-sided $p = 0.003$. The direction holds in $5/5$ models (sign test $p = 0.031$), and per-cell one-sided Mann-Whitney tests combine to $p \approx 4 \times 10^{-4}$ (treating cells as independent). The median $\Delta D_{\mathrm{W}\to\mathrm{C}} < \Delta D_{\mathrm{W}\to\mathrm{W}}$ holds in all nine (k, rank) settings of Table 6 (sign test $p = 0.002$). These robust results suggest that the wide median CIs of Table 3 likely reflect the inherent heterogeneity of the wrong-baseline pool rather than the absence of an effect.

Placement. Our main intervention forces the selection of the rank-3 token at the three highest-entropy thinking positions. As a robustness check, we applied the same intervention at three *random* thinking positions and at the three *lowest*-entropy positions. Table 13 shows the results on Qwen2.5-1.5B. As expected, moving away from the entropy peaks yields fewer fixes per wrong trace ($48.6\% \to 38.7\% \to 35.1\%$). The asymmetry itself, however, does not depend on placement: θ exceeds 0.5 at all three placements (0.588, 0.610, 0.627; one-sided $p \le 0.042$) and is indeed largest at the lowest-entropy positions, where the CI excludes 0.5 ($p = 0.017$). The D signature is thus a property of the fix, not of where the intervention lands.

Table 13: Placement controls (Qwen2.5-1.5B, rank 3, three interventions per wrong trace, 111 wrong traces pooled over benchmarks). θ combined over benchmarks with pair-count weights; trace-clustered bootstrap (3,000 resamples); one-sided $p = \Pr^*(\theta \le 0.5)$.

placement	$n_{\mathrm{W}\to\mathrm{C}}$	$n_{\mathrm{W}\to\mathrm{W}}$	fixes per wrong trace	θ [95% CI]	p
highest entropy (paper)	54	279	48.6%	0.588 [0.489, 0.682]	0.042
random	43	290	38.7%	0.610 [0.486, 0.714]	0.038
lowest entropy	39	294	35.1%	**0.627** [0.510, 0.733]	**0.017**

Forced-token probability, position, and length. To check whether the effect could have been carried by a particular subset of interventions, we recompute θ within terciles of four properties of an intervention (terciles cut within model $\times$ benchmark and θ combined across strata as above). Within terciles of the forced token's probability, $\theta = 0.543/0.605/0.590$. Within terciles of position entropy, $\theta = 0.530/0.628/0.586$. Within terciles of relative position in the trace, $\theta = 0.560/0.623/0.509$.

Within terciles of the trace length, $\theta = 0.555/0.644/0.563$. All twelve tercile cells lie above 0.5, showing that the asymmetry is not confined to any particular kind of intervention.

Regression analysis. We now test whether fixes ($\mathrm{W} \to \mathrm{C}$) predict a reduction of D more than non-fixes do, and whether a wrong trace that is confusion-like (as indicated by $D{\uparrow}$) predicts a higher fix probability (than $D{\downarrow}$). For either test, we fit a linear model on the $2{,}177$ rank-3 branches of the 726 wrong-baseline traces, with model $\times$ benchmark fixed effects, covariates (forced-token probability, position entropy, relative position, trace length) z-scored within cell, and trace-clustered bootstrap CIs ($2{,}000$ resamples).

For the first test, we regress the within-cell rank of ΔD (z-scored) on the fix indicator and the covariates. Consistent with our theoretical prediction of a fix $\mathrm{W} \to \mathrm{C}$ reducing D, the fix coefficient is negative (-0.248, 95% CI $[-0.402, -0.092]$), and no covariate's CI excludes zero.

For the second test, we regress the fix indicator on the $D{\uparrow}$ class and the covariates. If D were merely correctness in disguise, it would be impossible for D to predict which wrong traces are more likely to be fixed by perturbation. Instead, the $D{\uparrow}$ coefficient is $+0.043$ (i.e., $+4.3$ percentage points), with 95% CI $[+0.004, +0.079]$. The result also shows that interventions at later positions (-0.021) and in longer traces (-0.023) tend to reduce the fix probability.

O Full benchmarks, a larger model, and non-arithmetic tasks

This appendix further expands the level test (Table 2) with full benchmarks, a 14B model, and two non-arithmetic tasks.

Full benchmarks and standardized effect sizes. We re-ran the level test on the complete benchmarks (GSM8K 1,319 problems and each BBH task 250) for both decodings. The correctness gap $\Delta = \mathrm{med}(D_{\mathrm{wrong}}) - \mathrm{med}(D_{\mathrm{correct}})$ is positive in $30/30$ cells (one-sided sign test per decoding $p \approx 3 \times 10^{-5}$), with point estimates essentially unchanged from $n = 100$. Further, consistent with our earlier discussion that the wide per-cell CIs reflect inherent heterogeneity rather than an absence of the effect, the gap CIs for the full benchmarks now exclude zero in $13/15$ greedy and $11/15$ temperature-1 cells (vs. $7/15$ and $5/15$ at $n = 100$). The remaining non-significant cells mostly reflect a small outcome class (e.g., Qwen2.5-7B answers BBH-MA at 95% accuracy, leaving eleven error traces with the benchmark exhausted). Table 14 reports the scale-free effect size $\theta_{\mathrm{level}} = \Pr(D_{\mathrm{wrong}} > D_{\mathrm{correct}})$ (ties $1/2$) per cell: $\theta_{\mathrm{level}} > 0.5$ in $30/30$ cells, range $[0.53, 0.79]$, median 0.65, and the bootstrap CI excludes 0.5 in $27/30$ cells.

Table 14: Probability of superiority $\theta_{\mathrm{level}} = \Pr(D_{\mathrm{wrong}} > D_{\mathrm{correct}})$ for the level test at full benchmark size, with 95% bootstrap CIs (1,000 resamples). $n_{\mathrm{c}}/n_{\mathrm{w}}$ are the correct/wrong trace counts. †: CI excludes 0.5.

Model	Decoding	GSM8K	BBH-OC	BBH-MA
Qwen 1.5B	greedy	$0.71\ [0.68, 0.74]^{\dagger}$ (931/388)	$0.62\ [0.55, 0.69]^{\dagger}$ (153/97)	$0.65\ [0.58, 0.72]^{\dagger}$ (161/89)
	temp-1	$0.73\ [0.70, 0.76]^{\dagger}$ (785/534)	$0.61\ [0.54, 0.69]^{\dagger}$ (135/115)	$0.65\ [0.58, 0.72]^{\dagger}$ (82/168)
Qwen 7B	greedy	$0.64\ [0.59, 0.68]^{\dagger}$ (1134/185)	$0.77\ [0.69, 0.85]^{\dagger}$ (196/54)	$0.71\ [0.53, 0.85]^{\dagger}$ (239/11)
	temp-1	$0.61\ [0.57, 0.66]^{\dagger}$ (1126/193)	$0.69\ [0.60, 0.76]^{\dagger}$ (195/55)	$0.63\ [0.53, 0.72]^{\dagger}$ (228/22)
Gemma 7B	greedy	$0.60\ [0.57, 0.63]^{\dagger}$ (476/843)	$0.79\ [0.72, 0.86]^{\dagger}$ (84/166)	$0.66\ [0.51, 0.79]^{\dagger}$ (18/232)
	temp-1	$0.60\ [0.57, 0.63]^{\dagger}$ (479/840)	$0.72\ [0.65, 0.80]^{\dagger}$ (83/167)	$0.53\ [0.38, 0.67]$ (17/233)
Granite 8B	greedy	$0.62\ [0.59, 0.66]^{\dagger}$ (1024/295)	$0.54\ [0.47, 0.61]$ (129/121)	$0.64\ [0.57, 0.71]^{\dagger}$ (152/98)
	temp-1	$0.67\ [0.63, 0.70]^{\dagger}$ (973/346)	$0.56\ [0.48, 0.63]$ (125/125)	$0.58\ [0.504, 0.65]^{\dagger}$ (114/136)
Mistral 7B	greedy	$0.67\ [0.65, 0.70]^{\dagger}$ (690/629)	$0.67\ [0.60, 0.73]^{\dagger}$ (133/117)	$0.73\ [0.62, 0.83]^{\dagger}$ (15/235)
	temp-1	$0.61\ [0.58, 0.64]^{\dagger}$ (606/713)	$0.68\ [0.61, 0.74]^{\dagger}$ (124/126)	$0.64\ [0.52, 0.75]^{\dagger}$ (14/236)

A larger model. Under the default setup in the main text ($n = 100$ per benchmark, greedy decoding), Qwen2.5-14B-Instruct shows the same direction in $3/3$ benchmarks: median D 155 (correct) vs. 172 (wrong) on GSM8K at 84% accuracy, 99 vs. 122 on BBH-OC at 90%, and 85 vs. 123 on BBH-MA at 94%. Given the limited number of wrong traces (16, 10, and 6, respectively), the cells are underpowered for CIs.

Table 15: Level test on BBH logical deduction (seven objects), greedy, $n = 100$ problems per model: accuracy, median D on correct and wrong traces, relative gap, and one-sided Mann–Whitney p.

Model	acc.	$n_{\mathrm{c}}/n_{\mathrm{w}}$	med D_{correct}	med D_{wrong}	gap	p
Qwen2.5-1.5B	30%	30/70	280.2	299.1	+6.3%	0.236
Qwen2.5-7B	43%	43/57	483.2	657.0	+26.5%	0.002
Qwen2.5-14B	68%	68/32	195.8	243.4	+19.5%	0.011
Gemma-7B	33%	33/67	272.8	283.2	+3.7%	0.474
Granite-3.1-8B	45%	45/55	565.9	609.7	+7.2%	0.047
Mistral-7B	33%	31/64	367.9	382.6	+3.8%	0.215

Non-arithmetic tasks. Also under the setup in the main text ($n = 100$ per benchmark, greedy decoding), we tested the five models and Qwen2.5-14B on two further BBH tasks. As can be seen in Table 15, on *logical deduction* (seven objects, multiple choice with chance $\approx 14\%$), the direction replicates in $6/6$ models. On *word sorting* (free-form, no guess floor), scoring only exact match as correct, median D on wrong traces exceeds correct in $3/3$ of the models with at least 20 traces in each outcome class: Qwen2.5-7B 310 vs. 367 (+15.5%; 26 correct / 74 wrong), Qwen2.5-14B 90 vs. 95 (+5.0%; 45/55), and Granite-8B 579 vs. 614 (+5.7%; 34/66). The other three models answer fewer than 20 problems correctly (17, 10, 15), suggesting the task is beyond their capability.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction correspond directly to the theoretical results (Lemma 1, Propositions 1-4) and the empirical results of §8.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: §8 concludes with a consolidated paragraph on limitations. In addition, §3 establishes the framework's dependence on the analyst's chosen notion of resolution. §7 makes explicit that D's discriminative value as a regime indicator is partition-conditional, requiring the bridge condition of Proposition 4 to hold for the chosen partition. §8 acknowledges that final-answer correctness serves as a coarse rather than clean regime label, since wrong answers can arise from late errors on concentrated runs and correct answers from diffuse traces stumbling onto the right answer. App. B.2 states the technical conventions and assumptions, and notes that the framework extends to countable branches under appropriate convergence assumptions and an adjustment of the definition of c_t.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Each theoretical result in the paper states its assumptions. Full proofs are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: §8 describes the implementation details, the intervention design, the open-weight models, and the benchmarks. Appendix J reports robustness to design choices. Code for reproducing the experiments is released at the URL below.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The pipeline and reproducibility scripts are available at the public repository `https://github.com/nan-zhang-2022/n2d2`. The benchmarks and the five open-weight models are publicly available and cited in the paper. No new data is introduced.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: No training is performed in this work. §8 describes the test details, App. J reports robustness to design choices, and code is provided to reproduce the main tables.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Table 3 reports 95% bootstrap confidence intervals on median ΔH and ΔD per model and on the aggregate, and Table 8 reports 95% bootstrap confidence intervals on median ΔD per regime$\times$outcome cell. App. N and App. O add distribution-free effect sizes (probability of superiority) with trace-clustered bootstrap confidence intervals for the intervention and level claims.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: All experiments use open-weight inference only (no training). Running 5 models $\times$ 3 benchmarks $\times$ 100 problems for greedy, temperature-1 sampling, and intervention branches at three rank settings (ranks 2, 3, 7) fits on a single Apple M-series device (MPS backend, fp16) in approximately 60 GPU-hours — roughly 15 h for greedy and temperature-1 profiling (5 models $\times$ 2 decodings $\times$ 3 benchmarks $\times$ 100 problems) and 45 h for the intervention sweep (5 models $\times$ 3 benchmarks $\times$ 100 problems $\times$ top-3 branches $\times$ 3 ranks). The experiments added in App. L–O (resampled continuations, placement controls, full-benchmark profiles, a 14B model, and two further tasks) ran on the same device. The synthetic experiment and its robustness grid run in a few CPU-minutes.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research uses only publicly released open-weight language models and public reasoning benchmarks. It does not involve human subjects, data collection from humans, or the release of new generative models, and we have reviewed and follow the NeurIPS Code of Ethics.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper is a theoretical study of when autoregressive decoding fails to resolve continuation uncertainty. We do not see negative societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper poses no such risks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All assets used are publicly released: Qwen2.5-1.5B-Instruct and Qwen2.5-7B-Instruct (Apache 2.0), Gemma-7B-it (Gemma Terms of Use), Granite-3.1-8B-Instruct (Apache 2.0), and Mistral-7B-Instruct-v0.3 (Apache 2.0), together with the GSM8K (MIT) and BIG-Bench Hard (Apache 2.0) reasoning benchmarks. We use the assets in a manner consistent with their respective licenses and for non-commercial research purposes.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce new assets such as datasets or models; it introduces a new theoretical framework for understanding when autoregressive decoding fails to resolve continuation uncertainty.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human subjects and therefore no IRB approval is required.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are the object of study in this paper rather than a methodological component of the research process.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.