TOKEN FREEZING IN DIFFUSION LANGUAGE MODELS NEEDS NO CONVERGENCE TEST

Anonymous authors
Paper under double-blind review

ABSTRACT

Masked diffusion language models recompute every position at every denoising step, including tokens that are already committed and will not change. Freezing a committed token saves this compute by serving its cached keys and values, at the cost of staleness for the positions that still read them. The prevailing rule freezes a token once a convergence statistic of its output distribution, such as stepwise KL divergence, falls below a threshold. We show that this statistic cannot see the damage freezing does. Since a committed token is never rewritten, freezing affects the output only through the keys and values that other positions read. The statistic, however, measures the token's own output distribution. At a committed position, this distribution already places nearly all of its probability on the committed token, so it barely changes even when the keys and values do. To test whether the statistic's timing matters, we randomly shuffle its freeze times across tokens without increasing total compute. On every model we test, accuracy does not measurably change. We therefore drop per-token signals and freeze every token a fixed number of steps after commitment. We show that, among schedules that use no per-token information, this uniform delay minimizes the expected motion that frozen keys and values have left. Across five open diffusion language models, the delay matches or exceeds the prevailing rule at equal compute and improves on it by two to three points on held-out MATH. Rules built on other per-token signals, including the motion of the keys and values themselves, do not beat it either. The delay's savings carry over to wall-clock time, with 1.3–1.4$\times$ speedups at batch size 8 and above.

1 INTRODUCTION

Masked diffusion language models (dLLMs) generate text by iterative denoising. Starting from a fully masked canvas, each step runs a bidirectional transformer over every position and commits tokens at the positions where the model is most confident (Nie et al., 2025; Ye et al., 2025; Sahoo et al., 2024; Lou et al., 2024). Generation order becomes flexible and many tokens can be committed in parallel. The cost is that every step recomputes every position: a token committed at step 20 of a 256-step schedule is recomputed two hundred more times, although the sampler never touches it again.

Skipping this computation is the subject of an active line of work (Oba et al., 2026; Ma et al., 2025; Liu et al., 2025; Wu et al., 2026). The most direct form is freezing. Once a committed token is judged stable, its computation is dropped and its per-layer keys and values are served from a cache for the remaining steps. The saving is not free. The cached state goes stale while the rest of the canvas fills in, and positions still being generated read the stale copy. A freezing rule must therefore decide when to freeze each token, early enough to save compute and late enough to limit the staleness.

The prevailing rule is a per-token convergence test, which we call the criterion (Oba et al., 2026). Token i is frozen once a statistic of its own output distribution, the KL divergence between successive steps, falls below a threshold. The rule comes with a guarantee. It assumes that once a token passes the test, its statistic keeps shrinking geometrically, so the drift the token has left is bounded and the test stays true at every later step (the contraction assumption; Oba et al., 2026, Assumption A2). On open-ended generation, freezing this way skips 30–50% of the computation while keeping quality close to the unfrozen model.

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

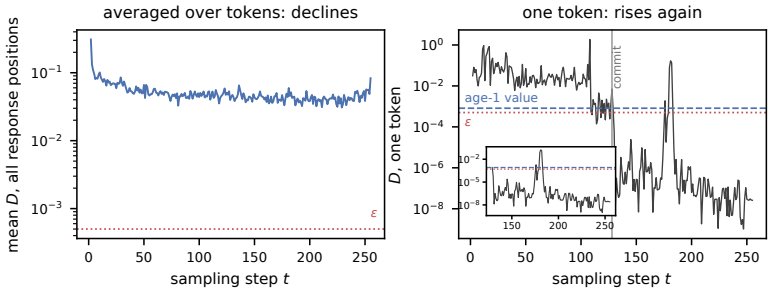

Figure 1: The contraction assumption holds on average but not per token. Left: the statistic D averaged over all response positions at each step, the view in which the assumption has been checked. It declines because the average mixes newly committed tokens, whose D is large, with older ones, whose D has collapsed; the mean sits four to five orders of magnitude above the median. Right: D for one committed token, the 75th-percentile member of the rebounding class (38–45% of committed tokens on LLaDA-8B and Dream-7B). It collapses after commitment, then rises back above both its first post-commitment value and the deployed threshold ε. Log axes.

However, the statistic cannot see the damage freezing does. Since a committed token is never rewritten, freezing affects the output only through the keys and values that other positions read from it. The statistic instead measures the token's own output distribution. At a committed position this distribution already places all but a small headroom δ of its probability on the committed token, with $\delta \leq 10^{-3}$ on four of the five models we study, so it barely changes even when the keys and values do. We prove that for small changes in the logits the statistic is at most about 6δ times that change, and verify the bound across seven orders of magnitude of δ (§4.1).

The statistic also does not keep shrinking. Averaged over all positions at each step it declines, and this is the view in which the contraction assumption has been checked (Oba et al., 2026). Followed one token at a time from its own commitment step, it rises on about half of all later steps on every model we measure, and 16–83% of the tokens that pass the test at the deployed threshold later rise back above it (Figure 1). The average declines only because it mixes newly committed tokens, whose statistic is large, with older ones, whose statistic has collapsed.

If the statistic cannot see the damage, its choice of when to freeze each token should carry no information about accuracy. To test this, we keep the waiting times the criterion produces, the number of steps between each token's commitment and its freeze, and randomly shuffle them across tokens, which does not increase total compute. On all five models, accuracy does not measurably change, and on average the shuffles score slightly higher. Better per-token signals do not change this. Reading the statistic before the softmax improves on the KL criterion, but neither this repaired statistic nor the token's own key and value motion, the best predictor of future staleness among the 22 signals we screen, beats the uniform delay we introduce next (§4.2).

We therefore drop per-token signals and freeze every token a fixed k steps after commitment. The damage a frozen token can still cause is bounded by how far its keys and values have left to move, which we call its exposure (Proposition 3). Averaged over tokens, exposure falls convexly as the wait grows (Appendix D). On a convex curve, spreading a fixed budget of waiting evenly gives the least expected exposure, so among schedules that use no per-token information the uniform delay is optimal (Proposition 4).

This is what we observe (Figure 3). At deployed thresholds the criterion loses two to six points against the unfrozen model on GSM8K and MATH, while the delay at k near a tenth of the schedule stays within the unfrozen model's paired interval and skips about 30% of the work. At matched work, no pairing separates against the delay on any site with at least 257 questions, and six separate in its favour, including both pre-specified comparisons on 1,500 held-out MATH problems ($+2.8$ and $+2.3$ points). The savings carry over to wall-clock time, with 1.3–1.4$\times$ end-to-end speedups at batch size 8 and above.

Our contributions are:

- A test of whether per-token timing matters for freezing. Shuffling the criterion's freeze times without extra compute does not measurably change accuracy on five models, and repaired or stronger per-token signals do not beat a fixed delay.

- An explanation of why. At a committed position the criterion's statistic is bounded by the headroom of the output distribution, so it cannot see the key and value motion through which freezing affects the output, and per token it does not contract on any model or signal we measure.
- The uniform delay, optimal among schedules without per-token information under a convexity condition we measure, which matches or exceeds the criterion at matched compute across five dLLMs and step counts from 64 to 768. Every k from $T/8$ to $T/2$ stays within the unfrozen model's interval, so k works as a compute knob rather than a parameter to calibrate.

2 Related work

Masked diffusion language models (Sahoo et al., 2024; Shi et al., 2024; Lou et al., 2024; Nie et al., 2025; Ye et al., 2025) have emerged as an alternative to autoregressive generation, trading left-to-right decoding for iterative parallel denoising, with confidence-ordered unmasking descending from MaskGIT (Chang et al., 2022) and block variants interpolating toward autoregression (Arriola et al., 2025). Because every step recomputes the full canvas, inference acceleration is an active area. Caching methods serve stale entries and correct them on a fixed or adaptive schedule (Ma et al., 2025; Liu et al., 2025; Nguyen-Tri et al., 2026; Li et al., 2026a). A caching method refreshes its entries, so a poor entry rule is repaired at the next refresh. Freezing is permanent, so the entry rule is the whole design. The two act on different axes and can be combined, and Oba et al. (2026) report further speedup from combining freezing with selection-based caching. The closest rule to ours is the delayed caching of dKV-Cache (Ma et al., 2025), which caches a decoded token's keys and values one step after decoding and refreshes the cache periodically. Under permanent freezing, a delay of one or two steps is too short. At $k \leq 2$ the delay separates below the unfrozen model on both Dream sites and on LLaDA GSM8K (Appendix E), while every k from $T/8$ to $T/2$ stays within the unfrozen model's interval (§4.4). The freezing rule we analyze freezes committed tokens permanently once a per-token convergence statistic settles, thresholding step-wise KL with an optional confidence gate (Oba et al., 2026). Fast-dLLM (Wu et al., 2026) combines block-wise KV caching with confidence-gated parallel decoding; it thresholds a posterior statistic but does not freeze tokens permanently. These methods report held quality on open-ended and fluency benchmarks, and our measurements agree with theirs in that regime; the differences we document appear on exact-answer reasoning tasks, where single tokens decide the answer. Orthogonal to both are methods that change the sampler itself, such as parallel and early-commitment decoding (Wei et al., 2026; Li et al., 2026b; Chen et al., 2026).

3 Preliminaries

Model and sampler. A masked diffusion LM generates a response of N positions over steps $t = 0, \ldots, T$. The canvas holds one token ID per position, initialized to `[MASK]`. Each step embeds the canvas, runs L bidirectional transformer layers, and produces logits $z_i(t) \in \mathbb{R}^V$ and a posterior $p_i(t) = \mathrm{softmax}(z_i(t))$ at every position i. The sampler commits tokens at the most confident masked positions: it writes the selected word's ID into the canvas, and it writes only to masked positions, so a committed token is fixed for the remainder of generation. We write c_i for the step at which position i commits and x_i for its token. We evaluate five open dLLMs at their released decode settings: LLaDA-8B-Instruct (Nie et al., 2025) and LLaDA-1.5 (Zhu et al., 2026) (256 steps, generation length 256, block size 32), LLaDA-MoE-7B-A1B (Zhu et al., 2025) (128 steps, length 128, block size 32), Dream-7B-Instruct (Ye et al., 2025) (256 steps, length 256), and Dream-Coder-7B (Xie et al., 2025) (768 steps, length 768); prompt positions are recomputed at every step and never freeze.

Freezing. A freezing policy selects, for each committed position, a freeze step $f_i \geq c_i$. From step $f_i + 1$ the position receives no query, attention output, or feed-forward computation; its per-layer keys and values and its feed-forward output, cached at step f_i, are served to every position that attends to it. The convergence test, which we call the *criterion*, freezes i at the first step $t \geq c_i$ with $D_i(t) \leq \varepsilon$, where

$$D_i(t) \;=\; \mathrm{KL}(\, p_i(t) \,\|\, p_i(t{-}1) \,)\,, \tag{1}$$

optionally restricted by a confidence gate. We implement the rule of Oba et al. (2026), including its gate, and sweep ε across its published range. On the tasks where they report held quality, our

implementation agrees. On HumanEval (generation length 128), LLaDA-8B scores 25.0 unfrozen and 25.0 and 27.4 with the criterion at $5{\cdot}10^{-4}$ and $5{\cdot}10^{-6}$, and on MT-Bench the criterion does not separate from the unfrozen model (Appendix F). *Freeze-at-commitment* ($f_i = c_i$) freezes at the earliest step any such rule can act, so it upper-bounds the work any criterion can skip. The *uniform delay* freezes at $f_i = \min(c_i + k,\, T)$ for a fixed k and reads no model output.

Notation for the analysis. At a committed position, $\delta_i(t) = 1 - p_i(t)(x_i)$ is the *headroom*: the posterior mass not on the written token. The *outgoing state* $r_i(t)$ stacks the per-layer keys and values of position i at step t; $g_i(t) = \|r_i(t{+}1) - r_i(t)\|$ is its one-step motion (mean over layers), and the *exposure*

$$G_i(t) \;=\; \textstyle\sum_{s=t}^{T-1} g_i(s) \tag{2}$$

is the outgoing motion remaining at step t, computed on a run in which nothing is frozen. The position's feed-forward output is also cached, so that its pinned posterior can be served, but no other position reads it, so it does not enter r_i.

Evaluation. Accuracy on GSM8K (Cobbe et al., 2021) (full test set; on LLaDA-MoE the $\varepsilon = 5 \cdot 10^{-6}$ pairing is at $n = 1319$ and the remaining arms at $n = 300$) and MATH (Hendrycks et al., 2021); we call a model–task pair a site. MATH results are reported on 1,500 held-out problems disjoint from the MATH-500 set used during development, with subject composition matched, and the compared cells fixed before those runs. Work is reported as algorithmic FLOPs under the accounting of Oba et al. (2026), GEMMs only; the criterion's per-step scoring cost is below 0.01% of a step and is ignored. Both rules' knobs (ε, k) are tuned with the same oracle access. Two cells are at *matched work* when their skipped-work fractions differ by at most 3 points. Intervals are paired 95% bootstrap over questions; we call a comparison separated only if its interval excludes zero. The criterion policy computes its statistic in float32. Every analysis statistic is computed in float64; at the deployed threshold $5{\cdot}10^{-4}$ the two precisions give the same rebound rates, and they differ below about 10^{-6} (Appendix C).

4 Analysis

4.1 The statistic cannot see key and value motion

Oba et al. (2026) bound how far a frozen token's own prediction can drift. Under a sampler that never rewrites a committed token, that prediction does not reach the output, but the frozen token's keys and values do, because every position still being generated attends to them. When position i is frozen, every other position reads the cached $r_i(f_i)$ in place of the current $r_i(t)$. The criterion reads the position's own posterior instead.

At a committed position, the posterior places nearly all of its mass on the committed token. One step after commitment, the median headroom δ is 10^{-3} or smaller on four of the five models, ranging from 10^{-2} on LLaDA-MoE to below 10^{-6} on Dream-Coder (Appendix C). In this regime the softmax is flat. The committed token's probability is a sigmoid of its logit margin, with slope $p(1{-}p) \approx \delta$, so even a large change in the logits moves the posterior only by an amount proportional to δ (Figure 2, left). As a result, 26–87% of committed positions already pass the deployed threshold $\varepsilon = 5{\cdot}10^{-4}$ one step after commitment. The following bound makes this precise.

Proposition 1 (headroom suppression)**.** *Let $p = \mathrm{softmax}(z)$ and $q = \mathrm{softmax}(z')$, let v be any token, and write $\delta = 1 - p(v)$, $\delta' = 1 - q(v)$. If $\delta' \le 1/2$, then*

$$\mathrm{KL}(p \,\|\, q) \;\le\; 2\,|\delta - \delta'| \;+\; 2\,\delta\,\|z - z'\|_\infty. \tag{3}$$

Both terms on the right are proportional to δ (Appendix A); for a small logit change $\Delta = \|z - z'\|_\infty$ the bound is about $6\delta\Delta$. However far the logits move, the statistic sees that motion scaled down by the headroom. We check the bound on seven model–task captures, about 46M committed position-steps, and it is never violated. On the four captures of Figure 2, the measured ratio $D/\|\Delta z\|_\infty$ stays near $0.03\,\delta$ across seven orders of magnitude of δ. Appendix C covers the other three captures and the positions with $\delta' > 1/2$, where the hypothesis does not apply.

The statistic is therefore a poor guide to how much damage a frozen token can still do. Table 1 scores signals a freezing rule could compute online by their rank correlation with the token's exposure G,

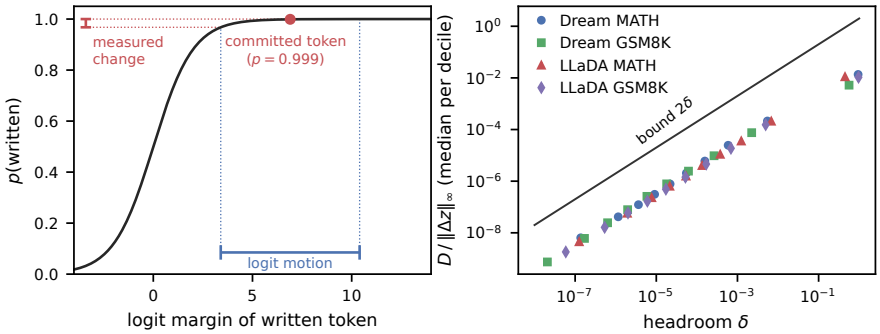

Figure 2: Left: the written token's probability is a sigmoid of its logit margin; at a measured operating point of 0.999+, an interval of logit motion (blue) maps to a posterior interval (red) smaller by the headroom factor. Right: measured $D/\|\Delta z\|_\infty$ against headroom δ, one point per decile per model–task pair, with the tail term 2δ of Proposition 1. The ratio tracks δ across seven orders of magnitude.

Table 1: How well each online signal predicts the key and value motion a token has left: absolute Spearman correlation with exposure G, divided by the number of steps remaining. Higher is better. Six of the 22 signals screened; all are in Appendix B.

signal	cost	Dream MATH	Dream GSM8K	LLaDA MATH	LLaDA GSM8K
outgoing K/V motion, trailing 4	4 buffers	0.84	0.88	0.75	0.73
outgoing K/V motion, 1 step	1 buffer	0.83	0.88	0.68	0.66
logit motion $\|\Delta z\|_\infty$, trailing 4	4 buffers	0.64	0.69	0.71	0.68
de-suppressed statistic D/δ	1 softmax	0.57	0.62	0.50	0.45
age since commitment	free	0.40	0.50	0.29	0.31
criterion statistic D	1 softmax	0.13	0.20	0.30	0.35

the key and value motion it has left (§3). The criterion's statistic ranks near the bottom (0.13–0.35), below the token's age since commitment, which costs nothing to compute. The key and value motion itself ranks at the top (0.66–0.88).

4.2 Reading before the softmax helps, but is not enough

Proposition 1 suggests a repair: read the signal before the softmax, either as the logit change $\|\Delta z\|_\infty$ or as the statistic divided by its headroom, D/δ. Both work better as signals. Dividing out δ raises the statistic's correlation with exposure by 1.3–4.5$\times$ (Table 1). The repair also works as a freezing rule. Replacing the criterion's trigger with D/δ, all else identical, improves on the KL criterion at matched compute on both development sites tested, by $+5.8$ $[+2.6, +9.0]$ points on Dream MATH-500 and $+2.8$ $[+0.2, +5.4]$ on LLaDA (Appendix B).

The repaired criterion does not, however, beat the uniform delay of §4.4. In sixteen matched-compute pairings it ties fourteen times and separates below the delay twice. The strongest signal in Table 1, the token's own recent key and value motion, also does not beat the delay when used as a freezing rule, on either site tested (Appendix B). Removing the suppression is therefore not enough. The next section shows a second problem that no per-step signal we measure avoids.

4.3 The statistic does not keep shrinking

Measured per token. The contraction assumption concerns one token's statistic after its own commitment, so we measure it there, in runs with nothing frozen. On all five models, a committed token's statistic rises on a median 48–49% of the steps after commitment. At some point, 69–88% of tokens exceed their own first post-commitment value, which a geometrically shrinking statistic never does. Of the tokens that pass the test, a large share later rise back above the threshold, and the share grows as the threshold tightens (Table 2). Signals read before the softmax behave the same way. At thresholds matched to the same skipped work, the logit change rises on 48–49% of steps and rebounds on 78–99% of qualifying tokens, and the logit margin rises on 50% of steps at every site. Better signals rank tokens better (Table 1), but none of them falls and stays below its threshold.

Table 2: Share (%) of tokens that qualify at threshold ε and later rise back above it (float64). The last three columns are GSM8K. Tightening the threshold raises the rate on every model.

ε	Dream MATH	Dream GSM8K	LLaDA MATH	LLaDA GSM8K	LLaDA-1.5	LLaDA-MoE	Dream-Coder
$5{\cdot}10^{-2}$	16.2	12.8	19.2	19.5	20.3	66.9	3.8
$5{\cdot}10^{-3}$	27.5	23.5	27.0	25.5	26.2	78.1	7.9
$5{\cdot}10^{-4}$	45.1	40.5	44.5	38.3	38.2	82.7	16.3
$5{\cdot}10^{-6}$	84.8	77.1	84.4	77.5	75.4	93.3	52.9
$5{\cdot}10^{-8}$	98.0	96.5	96.7	96.2	95.4	96.4	89.1

Since a freezing rule compares each token with a fixed threshold, a statistic that rises again after passing it cannot tell the rule that the token has settled.

A per-step threshold does not bound total change.

Proposition 2. *(i) Let $p(0), \dots, p(n)$ be distributions with $\mathrm{KL}(p(s)\|p(s{-}1)) \le \varepsilon$ for all s. Then $\mathrm{TV}(p(n), p(0)) \le n\sqrt{\varepsilon/2}$, and no bound sublinear in n holds in general. (ii) For every $\varepsilon \in (0, 1]$ and $\eta \in [\varepsilon/18, 1/2)$, there is an explicit two-token sequence with every step's KL at most ε along which the mass on the second token rises from η to $1{-}\eta$; the argmax flips within $\lceil 3(1{-}\sqrt{2\eta})/\sqrt{\varepsilon}\rceil + 1$ steps.*

In words, a per-step threshold limits how fast a distribution can move, not how far. At the thresholds the criterion is evaluated at, $\varepsilon = 5{\cdot}10^{-2}$, $5{\cdot}10^{-3}$ and $5{\cdot}10^{-4}$, the construction in (ii) flips the most likely token after 7, 19 and 59 steps and fully reverses it after 12, 37 and 116, with every step's exact KL below ε (Appendix A). The test's guarantee therefore rests on the contraction assumption, which Table 2 shows does not hold per token.

4.4 The uniform delay

Since no per-step signal tells a freezing rule when a token has settled, we instead bound the damage a frozen token can still cause.

Proposition 3 (exposure bounds per-step damage)**.** *Fix a run and a step t; let F be the frozen set with freeze steps f_i, let $\hat{r}_i$ be the cached outgoing state of $i \in F$, and for any position let $r_i^*(s)$ denote the outgoing state a full forward pass on the run's own canvas at step s would produce. If the map from any one position's outgoing state to any active position's logits is Λ-Lipschitz in each argument, then for every active j,*

$$\|\hat{z}_j(t) - z_j^*(t)\| \;\le\; \Lambda \sum_{i \in F} \|r_i^*(t) - \hat{r}_i\|. \tag{4}$$

In words, the error that frozen tokens cause in an active position's logits is bounded by how far their cached keys and values are from the current ones. We call this gap, accumulated over steps, the realized staleness, and measure it directly in §5. On a run in which nothing else is frozen, it is controlled by the key and value motion the token still has ahead of it at its freeze step, its exposure G (Appendix A).

Proposition 4 (the uniform delay is optimal among uninformed schedules)**.** *Consider committed positions whose remaining schedule accommodates every wait considered, and let $\bar{G}(w)$ be the population mean of $G_i(c_i + w)$. Call a schedule G-uninformed if it assigns a fixed multiset of waits $\{w_k\}$ to positions by a uniformly random permutation. Then its expected total exposure is $\sum_k \bar{G}(w_k)$, and if $\bar{G}$ is convex on an interval containing the w_k, the expected exposure at fixed budget $\sum_k w_k$ is minimized by the most uniform multiset.*

The proof (Appendix A) uses exchangeability for the expectation and Jensen's inequality for the minimization. The convexity premise is measured, not assumed. $\bar{G}$ falls steeply and then flattens; it coincides with its own convex minorant on the waits $w \in [0, 7]$ where the criterion's waits concentrate, and departs from it by at most 0.17% of $\bar{G}(0)$ anywhere in the swept range (Appendix D).

Freezing at commitment gives every token a wait of zero, the schedule with the most exposure. It is the worst arm on four of the seven sites where we run it, and within 1.0 point of the worst on the other three.

Table 3: Main results: accuracy (%) and work skipped (%). *: separated below the unfrozen model (95% paired CI excludes zero). Dashes: arms not run at that site. Pre-specified held-out comparisons: Dream, delay k=24 vs criterion $5{\cdot}10^{-6}$: $+\mathbf{2.80}$ $[+0.80, +4.87]$; LLaDA, delay k=32 vs criterion $5{\cdot}10^{-6}$: $+\mathbf{2.27}$ $[+0.40, +4.07]$. The GSM8K pairs are marked in Figure 3.

	Dream MATH-1500		LLaDA MATH-1500		LLaDA GSM8K		Dream GSM8K	
arm	acc	skip	acc	skip	acc	skip	acc	skip
unfrozen	37.47	0	37.00	0	76.72	0	80.82	0
freeze at commit	33.60*	37.4	32.33*	38.6	72.02*	38.5	72.40*	37.5
criterion $5{\cdot}10^{-4}$	35.13*	35.2	33.80*	36.4	71.95*	36.6	77.26*	35.6
criterion $5{\cdot}10^{-6}$	35.87	32.0	34.53*	31.8	74.07*	33.3	78.54*	33.2
criterion $5{\cdot}10^{-8}$	—	—	—	—	76.42	24.1	79.00*	27.6
delay k=16	—	—	—	—	75.28	33.8	79.68	32.9
delay k=24	**38.67**	30.7	—	—	75.82	31.6	79.91	30.7
delay k=32	—	—	**36.80**	29.5	**77.03**	29.4	79.76	28.6
delay k=48	—	—	36.40	25.4	76.50	25.3	—	—

Shuffling the criterion's freeze times. The criterion gives most tokens a wait of one or two steps and a minority much longer waits, so whatever it knows about a token is expressed in which tokens receive the long waits. By Proposition 4, if that assignment carries no information about exposure, the criterion is no better than a uniform delay with the same total wait. We test this directly. Within each question, we keep the waits the criterion produced and randomly shuffle which token receives which, so every shuffled schedule does the same or less work. Across five models and three shuffles each, none of the fifteen shuffled schedules separates in accuracy from the criterion's own schedule, and the shuffles do not lose on average: the mean difference is $+0.7$ points in their favour, and 14 of the 15 differences are zero or positive (Table 16). The criterion's decisions are also shaped by its own freezing: 19–30% of them differ from the decisions the same test makes on a read-only forward pass with nothing frozen, over the same canvas (Table 16). The analysis therefore predicts that the criterion performs at or below a uniform delay at matched compute, which §5 tests.

Practical recommendation. k is chosen for a target compute budget, as ε would be. k near 10% of the schedule (24 of 256 on Dream, 32 of 256 on LLaDA) stays within the unfrozen model's interval while skipping $\approx$30% of feed-forward and attention work, and the choice is not delicate: every k from $T/8$ to $T/2$ stays within the paired interval of the unfrozen model at $T \in \{64, 128, 256, 768\}$ (Appendix E), so k is a compute knob with a wide plateau rather than a parameter to calibrate.

5 Experiments

Setup. We evaluate LLaDA-8B and Dream-7B on GSM8K (full test set, $n = 1319$) and on held-out MATH-1500, and LLaDA-1.5, LLaDA-MoE and Dream-Coder on GSM8K (full test set; $n = 300$ for LLaDA-MoE). Each site runs the unfrozen model, freeze-at-commitment, the criterion at $\varepsilon \in \{5{\cdot}10^{-4}, 5{\cdot}10^{-6}, 5{\cdot}10^{-8}\}$, and the uniform delay. The delay uses $k \in \{16, 24, 32, 48\}$ on LLaDA-8B and Dream-7B and $k/T \in \{1/16, 1/8, 1/4, 1/2\}$ on the other three models, with $k \in \{16, 32\}$ and $k \in \{16, 24\}$ in the full-test-set runs on LLaDA-1.5 and Dream-Coder (Appendix E). Work skipped is the fraction of feed-forward and attention position-steps not computed. The FLOP ratio is $100\% - 0.987\times$(work skipped) to within 1.6 points on every configuration we ran (Appendix E). We run the KL criterion, which has a published guarantee; confidence and margin, the other statistics a convergence test could threshold, rank below it as signals (Appendix B).

Accuracy at matched compute. On every reasoning site, the delay's accuracy–compute frontier lies above the criterion's (Figures 3 and 4). At deployed thresholds, the criterion falls two to six points below the unfrozen model on GSM8K and MATH, separated on most sites. The delay at k near a tenth of the schedule stays within the unfrozen model's interval while skipping about 30% of the work. We pair each criterion cell with the delay cell nearest it in skipped work (Table 14). On sites with at least 257 questions, no pairing separates against the delay, and six separate in its favour, including both pre-specified held-out comparisons. Those two pair the delay with a criterion cell that skips slightly more work (1.3 and 2.3 points). Interpolating the criterion's accuracy along its own threshold sweep to the delay's skipped work leaves gaps of about 2.5 and 1.8 points. The one

Table 4: Matched-work pairing on GSM8K, delay minus criterion, one row per model and threshold: the delay cell nearest the criterion in skipped work (within 3 points), paired 95% bootstrap. Bold: interval excludes zero. No pairing on any model separates against the delay. Work gap is the delay cell's skipped work minus the criterion's; negative means the delay skipped less.

model	n	criterion	delay	Δ [95% CI]	work gap
LLaDA-8B	1319	$5{\cdot}10^{-4}$	k=8	**+2.05** [+0.08, +3.94]	−0.5
		$5{\cdot}10^{-6}$	k=16	+1.21 [−0.68, +3.11]	+0.5
Dream-7B	1319	$5{\cdot}10^{-4}$	k=6	+1.14 [−0.76, +3.03]	+0.2
		$5{\cdot}10^{-6}$	k=16	+1.14 [−0.83, +3.18]	−0.3
LLaDA-1.5	1319	$5{\cdot}10^{-4}$	k=16	**+2.20** [+0.15, +4.25]	−2.8
		$5{\cdot}10^{-6}$	k=16	+1.21 [−0.61, +3.03]	+0.4
LLaDA-MoE	300 / 1319	$5{\cdot}10^{-4}$	k=16	−1.67 [−4.67, +1.33]	+3.0
		$5{\cdot}10^{-6}$	k=64	−0.61 [−2.05, +0.83]	−2.3
Dream-Coder	1319	$5{\cdot}10^{-4}$	k=16	**+3.03** [+1.21, +4.85]	−1.2
		$5{\cdot}10^{-6}$	k=16	+1.59 [−0.08, +3.26]	−0.5

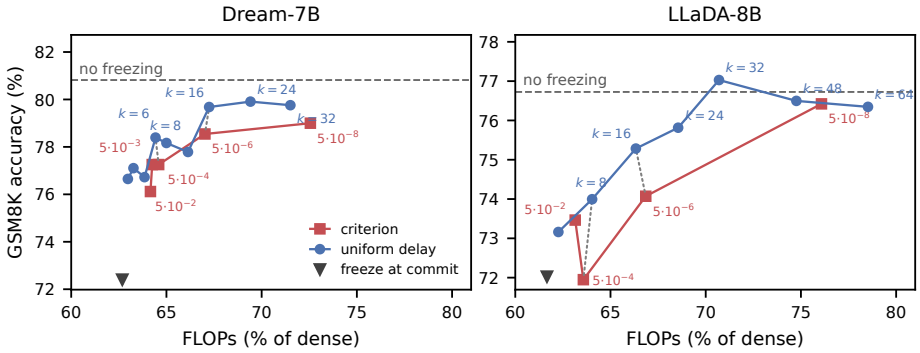

Figure 3: Accuracy against compute on GSM8K ($n = 1319$), LLaDA-8B and Dream-7B, every arm of both sweeps. Dotted segments mark the matched-work pairs at $\varepsilon = 5{\cdot}10^{-4}$ and $5{\cdot}10^{-6}$ (Table 4); in all four the delay cell lies above its criterion cell. The pre-specified held-out MATH-1500 comparisons are in Table 3; frontiers for the other reasoning pairs are in Figure 4.

pairing that separates against the delay is on the smallest site, LLaDA-8B HumanEval ($n = 164$), where the pairing at the other threshold favours the delay by 3.7 points (Table 14). Under a stricter rule that pairs each criterion cell only with a delay cell skipping at least as much work, none of 28 pairings separates in either direction (Table 15).

Staleness, measured on the runs. On the frozen runs behind Table 3 ($n = 100$ per site), a read-only probe recomputes each frozen token's keys and values on the same canvas and accumulates the gap to its cache. At matched work, the criterion's caches are staler than the delay's on all four probed sites, by 4.0–6.7 per frozen position-step at $\varepsilon = 5{\cdot}10^{-6}$ (in the norm of g_i; all four intervals exclude zero), and the gap widens at $5{\cdot}10^{-4}$. Exposure computed on the unfrozen runs gives the same ordering at every site. This is the ordering Proposition 4 predicts, measured on staleness itself rather than inferred from accuracy. Staleness orders policies but does not predict which individual questions fail (Appendix D).

Open-ended generation. Following the WikiText-103 generative-perplexity setup of Oba et al. (2026) (differences in Appendix F), the delay's perplexity is at or below the criterion's at matched work, and on single-turn MT-Bench the delay scores $+0.3$ $[-0.03, +0.7]$ above the criterion at matched work and within noise of the unfrozen model (Appendix F).

Wall-clock. We time a compacted executor that gathers the active rows, runs attention and feed-forward on them with cached keys and values for frozen positions, and scatters the results back. It replays the freeze schedules of the accuracy runs, and with nothing frozen it is bit-identical to the stock model. All ratios are against the stock model (three repeats, CUDA events; Table 5). At matched skipped work, the delay is at least as fast as the criterion on both models. At batch sizes of 8 and above, both arms realize 82–97% of the time their FLOP saving would save. At batch size 1

Table 5: End-to-end throughput on one NVIDIA H200 NVL, criterion against delay at matched skipped work (ε=$5{\cdot}10^{-6}$ against k=16). Ratios are against the stock, unmodified model in the same run. Ideal is $1/$(FLOP ratio). Realised is the share of the time the ideal would save, $(1-1/S)/(1-1/S_{\text{ideal}})$, where S is the measured speedup. Three repeats; batch sizes 1–4 are in Appendix G.

model	batch	arm	k / ε	vs stock	ideal	realised (time saved)
LLaDA-8B GSM8K	8	criterion	ε=$5{\cdot}10^{-6}$	1.37×	1.50×	81.8%
LLaDA-8B GSM8K	8	delay	k=16	1.40×	1.51×	85.5%
LLaDA-8B GSM8K	32	criterion	ε=$5{\cdot}10^{-6}$	1.42×	1.50×	88.6%
LLaDA-8B GSM8K	32	delay	k=16	1.45×	1.51×	91.9%
Dream-7B MATH-500	8	criterion	ε=$5{\cdot}10^{-6}$	1.37×	1.45×	87.8%
Dream-7B MATH-500	8	delay	k=16	1.39×	1.47×	87.5%
Dream-7B MATH-500	32	criterion	ε=$5{\cdot}10^{-6}$	1.43×	1.45×	96.7%
Dream-7B MATH-500	32	delay	k=16	1.44×	1.47×	96.3%

the decode is latency-bound, and no freezing arm is faster than the stock model (Appendix G). The FLOP model predicts measured time to within a few points, so the compute axis of Figure 3 is a faithful proxy for time.

6 Conclusion

Per-token convergence tests are the prevailing rule for freezing computation in masked diffusion LMs. We found that the timing such a test chooses carries no measurable information at matched compute, and that stronger per-token signals do not beat a fixed delay. A committed token affects the output only through the keys and values other positions read, which its own saturated output distribution barely reflects, and per token the statistic does not contract on any model or channel we measured. A uniform delay after commitment, optimal among schedules without per-token information under a convexity condition we measure, matches or exceeds the criterion at matched compute on every site large enough to resolve the comparison, and its savings carry over to wall-clock time.

Our study covers five open-weight dLLMs of 7–8B parameters under their released samplers, at schedules of 64 to 768 steps; the separations we report are on exact-answer reasoning tasks, and on fluency benchmarks both rules hold quality. Proposition 3 bounds per-step deviation and does not bound compounding across steps, the same scope as the analyses it replaces, and the delay's best k is model-dependent, though the plateau is wide.

AI use statement

Large language models were used for coding and implementation, for literature lookup, and for proofreading and editing. One LLM is additionally used as a measurement instrument: the MT-Bench arms in Appendix F are scored by an LLM judge, whose identity, protocol and re-scoring noise are reported there.

Reproducibility statement

Every arm in this paper is specified by the decode configuration in Appendix H, which gives each model's generation length, step count, block length, temperature and sampler exactly as released, together with the question splits and the freezing scope used by every frozen arm. The complete per-cell sweeps behind Table 3, Table 4 and Table 14 are in Appendix E, one table per site, with the realized skipped work of every cell and the per-cell FLOP ratio where it was measured rather than fitted; the FLOP accounting itself is the GEMM-only model of Oba et al. (2026), and the criterion's per-step scoring pass is accounted separately in Appendix G. Every interval in the paper is a paired bootstrap over questions, 20,000 resamples at seed 42, resampling the per-question difference vector in a fixed question order; a comparison is called separated only when the interval excludes zero. The held-out protocol is stated in Appendix H: the MATH-1500 split is disjoint from the MATH-500 split used for development, every threshold and every k was chosen before those runs, and the compared cells were fixed in advance. Code and evaluation scripts will be released.

REFERENCES

Marianne Arriola, Aaron Gokaslan, Justin T Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Block diffusion: Interpolating between autoregressive and diffusion language models. In *International Conference on Learning Representations*, 2025.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *CVPR*, 2022.

Zigeng Chen, Gongfan Fang, Xinyin Ma, Ruonan Yu, and Xinchao Wang. dParallel: Learnable parallel decoding for dLLMs. In *International Conference on Learning Representations*, 2026.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *NeurIPS Datasets and Benchmarks*, 2021.

Pengxiang Li, Jiayin Cai, Hongwei Xue, Kunyu Shi, and Shilin Yan. Beyond scattered acceptance: Fast and coherent inference for DLMs via longest stable prefixes. In *International Conference on Learning Representations*, 2026a.

Pengxiang Li, Yefan Zhou, Dilxat Muhtar, Lu Yin, Shilin Yan, Li Shen, Yi Liang, Soroush Vosoughi, and Shiwei Liu. Diffusion language model knows the answer before it decodes. In *International Conference on Learning Representations*, 2026b.

Zhiyuan Liu, Yicun Yang, Yaojie Zhang, Junjie Chen, Chang Zou, Qingyan Wei, Shaobo Wang, Yichen Zhu, and Linfeng Zhang. dllm-cache: Accelerating diffusion large language models with adaptive caching. *arXiv preprint arXiv:2506.06295*, 2025.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *International Conference on Machine Learning*, 2024.

Xinyin Ma, Runpeng Yu, Gongfan Fang, and Xinchao Wang. dkv-cache: The cache for diffusion language models. In *Advances in Neural Information Processing Systems*, 2025.

Quan Nguyen-Tri, Mukul Ranjan, and Zhiqiang Shen. Attention is all you need for KV cache in diffusion LLMs. In *International Conference on Learning Representations*, 2026.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

Daisuke Oba, Danushka Bollegala, Masahiro Kaneko, and Naoaki Okazaki. Stopping computation for converged tokens in masked diffusion-lm decoding. In *International Conference on Learning Representations*, 2026. arXiv:2602.06412.

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

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K Titsias. Simplified and generalized masked diffusion for discrete data. In *Advances in Neural Information Processing Systems*, 2024.

Qingyan Wei, Yaojie Zhang, Zhiyuan Liu, Puyu Zeng, Yuxuan Wang, Biqing Qi, Dongrui Liu, and Linfeng Zhang. Accelerating diffusion large language models with SlowFast sampling: The three golden principles. In *International Conference on Learning Representations*, 2026.

Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, Ligeng Zhu, Ping Luo, Song Han, and Enze Xie. Fast-dllm: Training-free acceleration of diffusion llm by enabling kv cache and parallel decoding. In *International Conference on Learning Representations*, 2026.

Zhihui Xie, Jiacheng Ye, Lin Zheng, Jiahui Gao, Jingwei Dong, Zirui Wu, Xueliang Zhao, Shansan Gong, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream-Coder 7B: An open diffusion language model for code. *arXiv preprint arXiv:2509.01142*, 2025.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*, 2025.

Fengqi Zhu, Zebin You, Yipeng Xing, Zenan Huang, Lin Liu, Yihong Zhuang, Guoshan Lu, Kangyu Wang, Xudong Wang, Lanning Wei, Hongrui Guo, Jiaqi Hu, Wentao Ye, Tieyuan Chen, Chenchen Li, Chengfu Tang, Haibo Feng, Jun Hu, Jun Zhou, Xiaolu Zhang, Zhenzhong Lan, Junbo Zhao, Da Zheng, Chongxuan Li, Jianguo Li, and Ji-Rong Wen. LLaDA-MoE: A sparse MoE diffusion language model. *arXiv preprint arXiv:2509.24389*, 2025.

Fengqi Zhu, Rongzhen Wang, Shen Nie, Xiaolu Zhang, Chunwei Wu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. LLaDA 1.5: Variance-reduced preference optimization for large language diffusion models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11425–11460, 2026.

A PROOFS

NOTATION

A masked diffusion LM denoises a canvas of N positions over steps $t = 0, \dots, T$. Position i *commits* at step c_i, the first step at which its token x_i is written; the sampler does not remask, so x_i is fixed thereafter. Let $z_i(t) \in \mathbb{R}^V$ be the logits at position i after step t and $p_i(t) = \mathrm{softmax}(z_i(t))$. The per-step statistic is $D_i(t) := \mathrm{KL}(p_i(t) \,\|\, p_i(t-1))$, and at a committed position $\delta_i(t) := 1 - p_i(t)(x_i)$ is the mass *not* on the written token. The outgoing representation $r_i(t)$ stacks the per-layer keys and values that other positions read from i; its one-step motion is $g_i(t) := \|r_i(t{+}1) - r_i(t)\|$ and its *exposure* is $G_i(t) := \sum_{s=t}^{T-1} g_i(s)$. A criterion freezes i at the first $t \geq c_i$ with $D_i(t) \leq \varepsilon$; a uniform delay freezes i at $f_i = \min(c_i + k, T)$.

PROPOSITION 1 (HEADROOM SUPPRESSION)

Proof. Write $\mathrm{KL}(p\|q) = \sum_w p(w) \log(p(w)/q(w))$ and split at $w = v$.

The v-term. It equals $(1-\delta) \log \frac{1-\delta}{1-\delta'}$. If $\delta \geq \delta'$ the logarithm is non-positive and the term is ≤ 0. If $\delta < \delta'$, then using $\log(1+x) \leq x$, then $1 - \delta \leq 1$, then $1 - \delta' \geq 1/2$,

$$(1-\delta) \log\Big(1 + \frac{\delta' - \delta}{1-\delta'}\Big) \;\leq\; (1-\delta)\,\frac{\delta' - \delta}{1-\delta'} \;\leq\; \frac{\delta' - \delta}{1-\delta'} \;\leq\; 2(\delta' - \delta).$$

In both cases the v-term is at most $2|\delta - \delta'|$. This is the only place the hypothesis $\delta' \leq 1/2$ is used.[1]

The tail. Since $\sum_{w \neq v} p(w) = \delta$,

$$\sum_{w \neq v} p(w) \log \frac{p(w)}{q(w)} \;\leq\; \delta \cdot \max_{w \neq v} \log \frac{p(w)}{q(w)}.$$

For softmax outputs $\log p(w) - \log q(w) = (z_w - z'_w) - (\mathrm{LSE}(z) - \mathrm{LSE}(z'))$. The first difference is at most $\|z - z'\|_\infty$ in absolute value; so is the second, because log-sum-exp is 1-Lipschitz in the sup norm. Hence the maximum is at most $2\|z - z'\|_\infty$ and the tail is at most $2\delta\|z - z'\|_\infty$. If that maximum is negative the bound holds trivially, the right-hand side being non-negative. Adding the two pieces gives the claim.

□

[1] The hypothesis is on δ', the headroom at the earlier step.

Both terms carry δ. Let $\Delta = \|z - z'\|_\infty$. For every w, $\log p(w) - \log q(w) = (z_w - z'_w) - (\mathrm{LSE}(z) - \mathrm{LSE}(z'))$ has absolute value at most 2Δ, so $q(w)/p(w) \in [e^{-2\Delta}, e^{2\Delta}]$. Summing over $w \neq v$ gives $\delta'/\delta \in [e^{-2\Delta}, e^{2\Delta}]$, hence $|\delta - \delta'| \leq \delta(e^{2\Delta} - 1)$, and the bound of Proposition 1 is at most $2\delta(e^{2\Delta} - 1) + 2\delta\Delta$, which is $\approx 6\delta\Delta$ for small Δ.

The bound is one-directional: a small statistic does not certify small outgoing motion, and nothing here claims the converse. Where $\delta' > 1/2$ the bound does not apply; those position-steps are 4.8–6.8% of the committed population (Appendix C).

PROPOSITION 2 (A PER-STEP THRESHOLD BOUNDS ONLY LINEAR GROWTH)

Proof. (i) Pinsker gives $\mathrm{TV}(p(s), p(s-1)) \leq \sqrt{\varepsilon/2}$ at each step; total variation obeys the triangle inequality; summing over $s = 1, \ldots, n$ gives $\mathrm{TV}(p(n), p(0)) \leq n\sqrt{\varepsilon/2}$. Part (ii) exhibits sequences attaining linear growth up to a constant, so no bound sublinear in n holds in general.

(ii) Write $p(s) = (1 - q_s, q_s)$ and use $\mathrm{KL} \leq \chi^2$, which for two points reads $\mathrm{KL}(\mathrm{Bern}(a) \| \mathrm{Bern}(b)) \leq (a-b)^2/(b(1-b))$. Take

$$q_{s+1} \;=\; \min\Big(q_s + \sqrt{\varepsilon\, q_s(1-q_s)},\; 1 - \eta\Big).$$

For an untruncated step $(a-b)^2 = \varepsilon\, q_s(1-q_s)$, so $\chi^2 = \varepsilon$ and the step's KL is at most ε; a truncated step is shorter and its χ^2 smaller. Truncation keeps $q_s \leq 1 - \eta < 1$, so every step is well defined. The sequence is monotone, so $q_s \geq \eta$ throughout.

Phase 1 ($q_s \leq 1/2$). Here $1 - q_s \geq 1/2$, so $\sqrt{\varepsilon q_s(1-q_s)} \geq \sqrt{\varepsilon q_s/2}$. With $x_s := \sqrt{q_s}$ and $a := \sqrt{\varepsilon/2}$,

$$x_{s+1}^2 \;\geq\; x_s^2 + a\, x_s \;\geq\; \Big(x_s + \tfrac{a}{3}\Big)^2,$$

the second inequality being equivalent to $a x_s \geq \frac{2a}{3} x_s + \frac{a^2}{9}$, i.e. to $x_s \geq a/3$, i.e. to $q_s \geq \varepsilon/18$, which is what $\eta \geq \varepsilon/18$ and monotonicity supply. So x_s grows by at least $a/3 = \sqrt{\varepsilon}/(3\sqrt{2})$ per step, and reaching $q = 1/2$ from $q = \eta$ takes at most $\big(\sqrt{1/2} - \sqrt{\eta}\big)\big/\big(\sqrt{\varepsilon}/(3\sqrt{2})\big) = 3\big(1 - \sqrt{2\eta}\big)/\sqrt{\varepsilon}$ steps.

Phase 2 ($q_s \geq 1/2$). Let $\varphi(q) := \arcsin\sqrt{q}$, so that $\varphi'(q) = 1/(2\sqrt{q(1-q)})$ and $\varphi''(q) = -(1-2q)/(4\,(q(1-q))^{3/2}) \geq 0$ for $q \geq 1/2$. Convexity gives, for an untruncated step,

$$\varphi(q_{s+1}) \;\geq\; \varphi(q_s) + \varphi'(q_s)\sqrt{\varepsilon\, q_s(1-q_s)} \;=\; \varphi(q_s) + \tfrac{1}{2}\sqrt{\varepsilon}.$$

Since $\varphi(1/2) = \pi/4$ and $\varphi(1-\eta) = \pi/2 - \arcsin\sqrt{\eta}$, phase 2 takes at most $(\pi/2 - 2\arcsin\sqrt{\eta})/\sqrt{\varepsilon}$ untruncated steps, plus one truncated step. Adding the phases and one step of slack at the boundary gives the stated counts. □

The construction, run exactly. Table 6 runs the recursion in float64 at $\eta = 0.02$ and asserts the *exact* per-step KL against ε at every step rather than relying on the χ^2 bound. The largest single-step KL is 0.52–0.62 of ε.

Table 6: The two-token construction at $\eta = 0.02$, exact recursion. "Bound" is $\lceil (3(1 - \sqrt{2\eta}) + \pi/2 - 2\arcsin\sqrt{\eta})/\sqrt{\varepsilon} \rceil + 2$ with numerator 3.6870.

ε	argmax flips	full reversal	bound	$\max_s \mathrm{KL}(p(s+1)\|p(s))/\varepsilon$
$5{\cdot}10^{-2}$	7	12	19	0.6246
$5{\cdot}10^{-3}$	19	37	55	0.5693
$5{\cdot}10^{-4}$	59	116	167	0.5251

Remark (why the assumption cannot be monitored). The published justification obtains a horizon-independent bound by assuming geometric contraction after the lock, $D_i(s) \leq \rho\, D_i(s-1)$ for $s > f_i$, so that $\sum_{s > f_i} \sqrt{D_i(s-1)} \leq \sqrt{D_i(f_i)}/(1 - \sqrt{\rho})$. Part (i) is the same telescoping sum without that assumption. The assumption cannot be checked by the policy that needs it: under one-way locking the frozen posterior is pinned, $\hat{p}_i(t) = \hat{p}_i(f_i)$ for $t \geq f_i$, so $\hat{D}_i(t) = 0$ identically for $t > f_i$, for every model and every input. It can be measured in the dense run, and there it is false (§4.3, Appendix C).

PROPOSITION 3 (EXPOSURE BOUNDS THE PER-STEP DEVIATION)

Proposition (Proposition 3, restated)**.** *Fix a run and a step t; let F be the frozen set with freeze steps f_i, let $\hat{r}_i$ be the cached outgoing state of $i \in F$, and for any position let $r_i^*(s)$ denote the outgoing state a full forward pass on the run's own canvas at step s would produce. If the map from any one position's outgoing state to any active position's logits is Λ-Lipschitz in each argument, then for every active j,*

$$\|\hat{z}_j(t) - z_j^*(t)\| \;\le\; \Lambda \sum_{i \in F} \|r_i^*(t) - \hat{r}_i\|,$$

and for each $i \in F$,

$$\|r_i^*(t) - \hat{r}_i\| \;\le\; \|r_i^*(f_i) - \hat{r}_i\| + \sum_{s=f_i}^{t-1} \|r_i^*(s{+}1) - r_i^*(s)\|.$$

Proof. The frozen computation of step t and the same-canvas full computation receive identical inputs except that the cached $\hat{r}_i$ stands in place of $r_i^*(t)$ for each $i \in F$. Order the substitutions arbitrarily and form the chain of hybrid inputs that replaces $\hat{r}_i$ by $r_i^*(t)$ one position at a time. The uniform Λ-Lipschitz hypothesis bounds each link by $\Lambda\|r_i^*(t) - \hat{r}_i\|$, and the triangle inequality across links gives the first inequality. The second is the triangle inequality through $r_i^*(f_i)$: $\|r_i^*(t) - \hat{r}_i\| \le \|r_i^*(f_i) - \hat{r}_i\| + \|r_i^*(t) - r_i^*(f_i)\|$, and the second term telescopes along $s = f_i, \ldots, t-1$. □

The middle term, $\|r_i^*(t) - \hat{r}_i\|$ accumulated over steps, is the run's *realized staleness*: how far the value the cache is serving has drifted from what a full forward on the canvas that run actually produced would return. It is what the probe measures, and it is measured directly rather than through either bound: the probe compares the live outgoing representation against the tensor the executor is serving, layer by layer, on the frozen run's own canvas (Appendix D). The frozen run does not compute it, since nothing is recomputed for a frozen position, so the probe runs a second forward with the cache bypassed and discards it.

The first term, $\|r_i^*(f_i) - \hat{r}_i\|$, is the cache's own error at the freeze step. It is zero when nothing was frozen before f_i, because the cache is then written by a full computation; otherwise it is the compounding across steps that this bound does not cover. The dense-run exposure $G_i(f_i)$ is a different quantity: the motion i would still undergo if it alone were frozen. The paper uses G for claims about *signals* and realized staleness for claims about *damage under a policy*.

PROPOSITION 4 (THE UNIFORM DELAY IS OPTIMAL AMONG UNINFORMED SCHEDULES)

Proof. (i) Under a uniformly random permutation π, each wait value w_k lands on a uniformly random position, so $\mathbb{E}[G_{\pi^{-1}(k)}(c + w_k)] = \frac{1}{n}\sum_i G_i(c_i + w_k) = \bar{G}(w_k)$; sum over k by linearity.

(ii) Convexity of the sequence $\bar{G}$ on I (non-negative second differences) is convexity of its piecewise-linear interpolation $\bar{G}_{\text{lin}}$ there. Jensen gives $\frac{1}{n}\sum_k \bar{G}_{\text{lin}}(w_k) \ge \bar{G}_{\text{lin}}(W/n)$, and $\bar{G}_{\text{lin}} = \bar{G}$ at integers. Among integer multisets summing to W, the two-value multiset $\{\lfloor W/n \rfloor, \lceil W/n \rceil\}$ attains $n\bar{G}_{\text{lin}}(W/n)$ exactly, by linearity on the unit interval, so it is a minimizer; equality and strictness are Jensen's. □

Robustness to the measured curve. $\bar{G}$ is convex where deployed policies operate and not everywhere. Its deviation from its own convex minorant is exactly zero on $w \in [0, 7]$ at all four measured sites, and at most 0.52 units (0.17% of $\bar{G}(0)$) anywhere in the swept range (Appendix D).For a curve convex up to an additive perturbation, $\bar{G} = \bar{G}_c + e$ with $\sup|e| \le \eta_e$, part (ii) holds up to an additive $n\eta_e$. The bound is one-sided because the minorant lies below $\bar{G}$, so the perturbation e is non-negative. With $\eta_e \le 0.52$ against $\bar{G}$ values of 200–600, the uniform multiset is within $n\eta_e$ of the minimum, which is at most 0.17% of $\bar{G}(0)$ per position.

B SIGNAL SCREEN AND THE DE-SUPPRESSED CRITERION

The screen. Table 7 is the full version of Table 1: every signal a policy could compute online, scored by controlled Spearman correlation against the same target G over the same rows, so the

ranks are comparable. The controlled column divides G by the number of steps remaining, which removes the mechanical component that any purely positional signal would otherwise pick up; steps remaining in the schedule is screened as that reference point and scores 0.04.

The top of the table is outgoing motion itself (the quantity Proposition 3 says damage is bounded by), at 0.66–0.88. The criterion's own statistic is near the bottom at 0.13–0.35, below the token's age, which is free. And the four pre-softmax channels (the two logit-motion signals, D/δ and the margin change) sit between them: every one beats the post-softmax KL on every site, as Proposition 1 predicts.

The de-suppressed criterion as a policy. Replacing the criterion's trigger with D/δ or with $\|\Delta z\|_\infty$, changing nothing else (same freezing semantics, same confidence gate, same scope), gives the cells of Table 8. Thresholds were calibrated per site by bisection on the dense captures and then *piloted* at $n = 25$ before any full run, because a threshold calibrated on a dense trajectory does not reproduce its work in a freezing decode: the realized skip came in 3.9 to 6.3 points below the dense prediction on every $\|\Delta z\|_\infty$ cell.

Against the published criterion the repair works: two separations, both from the D/δ family, both against $\varepsilon = 5{\cdot}10^{-6}$. Against the uniform delay it does not: zero positive separations in sixteen matched-work pairings, two negative, fourteen ties.

C SATURATION, REBOUNDS, AND NUMERICAL PRECISION

At age 1 (the first step at which the statistic exists after commitment), between 26.3% and 86.5% of committed positions are already below the deployed threshold $\varepsilon = 5{\cdot}10^{-4}$, and by age 2 between 34.8% and 92.7% are. The spread is the models' own saturation: the share is 77.9% on LLaDA-1.5 and 86.5% on Dream-Coder, and 26.3% on LLaDA-MoE, whose median headroom at age 1 is $1.5{\cdot}10^{-2}$ against $1.3{\cdot}10^{-4}$ on LLaDA-1.5 and below $2.4{\cdot}10^{-7}$ on Dream-Coder. The criterion is therefore not distinguishing settled from unsettled tokens for most of the population; it is reading a saturated posterior.

Proposition 1's slope is checked the same way on all seven captures, by the median $D/\|\Delta z\|_\infty$ within each decile of δ: the ratio increases with δ across every decile on six of them (both Dream-7B captures, both LLaDA-8B captures, LLaDA-1.5 and Dream-Coder), and on LLaDA-MoE it increases across the first nine deciles and falls in the tenth, whose median headroom is 0.99. That decile lies in the region $\delta > 1/2$, where the proposition's hypothesis does not hold, and that tail is 23.3% of LLaDA-MoE's committed position-steps against 1.2–7.0% on the other six.

Table 10 is the contraction test on every channel. No channel contracts. $\|\Delta z\|_\infty$ rises on 47.7–49.2% of post-commitment steps and the pre-sigmoid margin on 49.8–50.0%, at every site. The rebound rates are *worse* upstream of the softmax than in the posterior at every site: at work matched to $\varepsilon = 5{\cdot}10^{-6}$, 88.3–99.8% of positions that qualify later rise back above their threshold, against 52.9–93.3% for D at the same ε. Reading the signal before the softmax improves ranking (Appendix B) but not monotonicity.

Positions outside Proposition 1's hypothesis. On the four sites of Figure 2, on 4.8–6.8% of committed position-steps $\delta > 1/2$, where Proposition 1's hypothesis does not hold. Those positions are 12.2–14.1% of committed positions, and 87.5–95.4% of them violate the contraction assumption, against 27.1–38.4% outside the set.

Precision. The statistic is computed from a $\log\mathrm{softmax}$ over a vocabulary of 1.3–1.5 $\times\, 10^5$ tokens, and at a saturated position the two logs being differenced agree to within float32's resolution. The consequence is in Table 11: in float32 the statistic comes out *negative* (impossible for a KL divergence) on 11.3–18.0% of committed position-steps, and in float64 on 0.000–0.001%. Every rebound rate in this paper is float64. The criterion policy in the accuracy runs computes its statistic in float32. A negative value passes the test just as zero does, so clamping would not change any decision. The precision affects rebound rates only below $\varepsilon \approx 10^{-6}$: at $5{\cdot}10^{-4}$ the two precisions agree to 0.00 points on the four sites measured, and at $5{\cdot}10^{-8}$ they differ by 1.2–2.0 points, with float32 reading higher. Every matched-work comparison in the main text is at $5 \cdot 10^{-4}$ or $5 \cdot 10^{-6}$.

D EXPOSURE AND STALENESS

Proposition 4 needs $\bar{G}$ convex only where the schedules being compared put their waits. Table 12 measures it: the deviation from the curve's own convex minorant is *exactly zero* on $w \in [0, 7]$, where deployed criterion policies concentrate, and at most 0.52 units (0.17% of $\bar{G}(0)$) anywhere in the swept range $w \leq 32$. The non-convexity that exists is at large w and is two orders of magnitude smaller than the early second differences.

Table 13 is the prediction of Proposition 4 tested on the theorem's own quantity rather than through accuracy. On the frozen runs behind Table 3, a read-only probe recomputes what each frozen token's outgoing state would be on the same canvas and accumulates the gap to its cache. The criterion's caches are staler than the delay's at matched work on the four sites probed, in the total and per frozen position-step, and the exposure computed on the unfrozen run agrees at every site. The probe runs on a re-run of each arm on 100 of its questions; every re-run reproduces the original generations exactly, so the staleness it measures is that of the reported arm.

Staleness is a population quantity. We use it to order policies; per-question totals do not predict individual errors, and do not separate the criterion's wrong answers from its right ones (van Elteren, site as stratum, $z = -1.53$, one-sided $p = 0.94$).

E COMPLETE RESULTS

Table 14 pairs each criterion cell at $\varepsilon = 5{\cdot}10^{-4}$ and $5{\cdot}10^{-6}$ with the delay cell nearest it in skipped work, at every site, and gives the paired delta with its interval; its last two rows are the pre-specified held-out comparisons.

Table 16 gives the two controls of §4.4. The permutations preserve each question's multiset of realized waits and destroy only which token got which wait. On the two MATH-500 controls the work is matched to within 0.62% on realized frozen position-steps. On the three added models the work column is the realized frozen *fraction*, and the shuffles do less computation than the criterion, 0.5% less on Dream-Coder, 4.7% on LLaDA-1.5 and 11.6% on LLaDA-MoE, because a permutation that moves a long wait onto a late-committing position has that wait truncated by the end of the schedule. Every one of the fifteen therefore lands at or *below* the criterion's own work, which is the conservative direction for this control: the permuted schedule does no more work and is no less accurate. The largest of the fifteen deltas is 3.00 points, inside its own interval. Their mean is $+0.72$ points, in the shuffles' favour. The decision-flip rates are measured by running the criterion's own threshold test a second time on a read-only dense forward over the same canvas the frozen run produced.

Table 17 sweeps the step count itself, so that the plateau in k is not read off a single T. At $T = 64$ and $T = 128$ every delay cell from $T/16$ to $T/2$ has a paired interval against the unfrozen model that includes zero. Across every sweep in this appendix, covering $T \in \{64, 128, 256, 768\}$, exactly two delay cells in $[T/16, T/2]$ separate below their own unfrozen model, and both sit at $k = T/16$: LLaDA-MoE at $k = 8$, $T = 128$ (-3.33 $[-6.67, -0.33]$) and LLaDA-1.5 MATH-500 at $k = 16$, $T = 256$ (-4.20 $[-7.40, -1.00]$). From $T/8$ to $T/2$ there is no exception anywhere, which is why §4.4 states the plateau from $T/8$.

The per-site sweeps follow: the six sites the main text discusses, then the two full-test-set sweeps on the added models, then the four k/T grids. Every site of the grid is summarised in Table 14; the reasoning sites are also plotted in Figure 4.

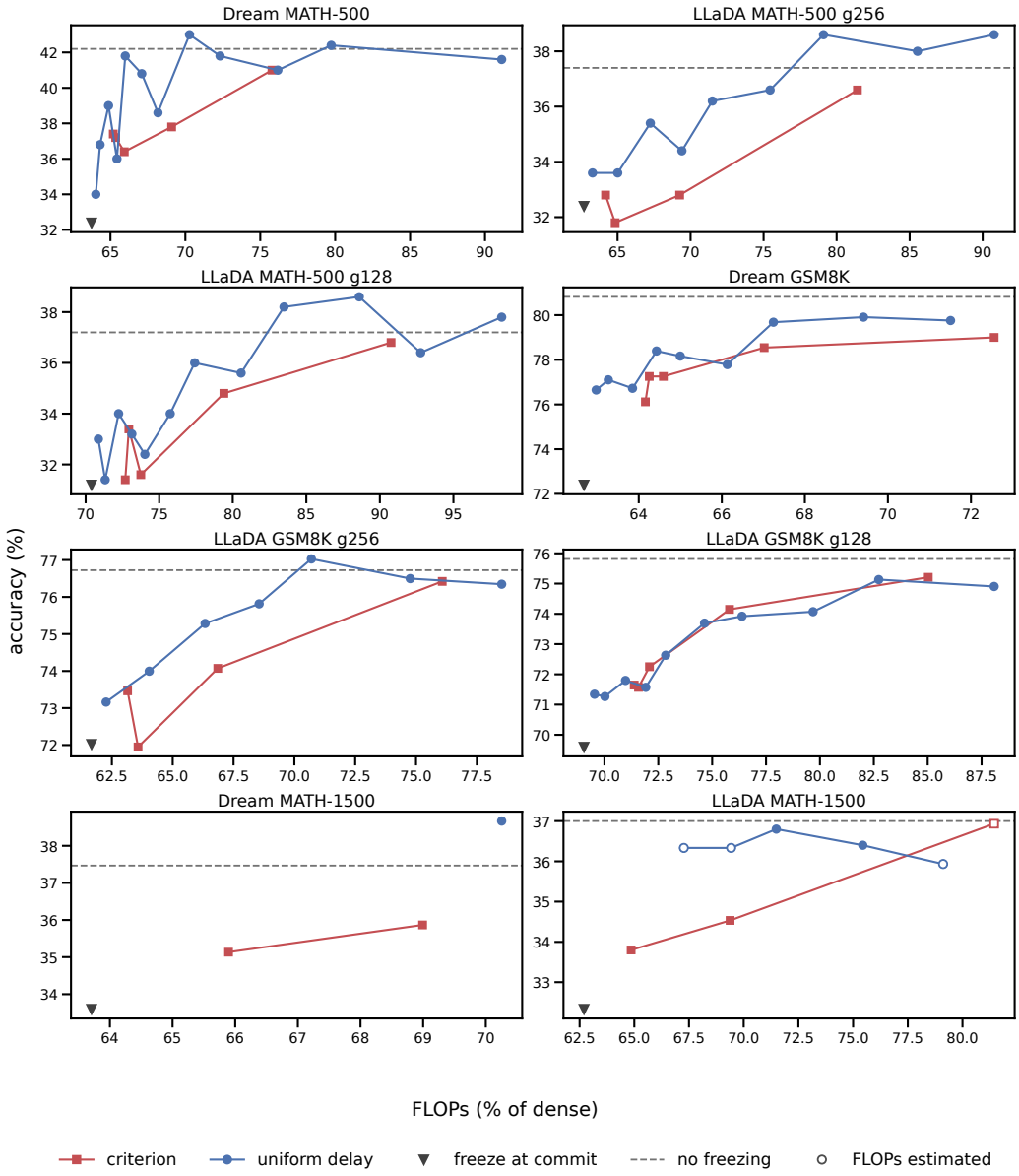

Figure 4: Accuracy against exact per-cell FLOPs at the eight reasoning frontiers, counting decode variants and the held-out splits. Axes are per panel, and the top row is the two MATH-500 development sweeps. The matched-work pairs at $\varepsilon = 5{\cdot}10^{-4}$ and $5{\cdot}10^{-6}$ at every site, the code-task sites included, are the rows of Table 14. Hollow markers (LLaDA MATH-1500) are cells whose accuracy is measured but whose FLOPs are estimated from their skipped work, and whose intervals were not computed (Table 22).

Table 7: The full screen: controlled Spearman correlation with exposure G for every signal a policy could compute online, on the four probed captures ($n = 100$ questions each, float64). Ranked by the mean of $|\rho|$ across sites. The two logit-motion signals, D/δ and the margin change, are the pre-softmax channels of §4.2.

signal	online cost	D-MATH	D-GSM	L-MATH	L-GSM	mean
outgoing K/V motion, trailing 4	4 buffers	0.8414	0.8846	0.7522	0.7277	0.8014
outgoing K/V motion, trailing 2	2 buffers	0.8402	0.8855	0.7258	0.6982	0.7874
outgoing K/V motion, 1 step	1 buffer	0.8294	0.8792	0.6847	0.6569	0.7626
feed-forward output motion, 1 step	1 buffer	0.8217	0.8568	0.6637	0.6343	0.7441
logit motion $\|\Delta z\|_\infty$, trailing 4	4 buffers	0.6366	0.6924	0.7105	0.6769	0.6791
logit motion $\|\Delta z\|_\infty$, 1 step	1 buffer	0.6047	0.6578	0.6319	0.5921	0.6216
de-suppressed statistic D/δ	softmax	0.5677	0.6229	0.4998	0.4545	0.5362
neighbourhood K/V motion, window of 16	1 buffer	0.7045	0.7227	0.3753	0.2962	0.5247
change in logit margin	softmax	0.4438	0.5171	0.4472	0.4245	0.4581
age since window of 16 is 50% committed	free	-0.4126	-0.5053	-0.2965	-0.3171	0.3829
age since window of 16 is 90% committed	free	-0.4073	-0.5031	-0.2951	-0.3172	0.3807
age since commitment	free	-0.4044	-0.5024	-0.2878	-0.3094	0.3760
distance to nearest masked position	free	-0.3642	-0.4672	-0.2957	-0.3172	0.3611
cumulative K/V motion since commitment	1 buffer + 1 scalar	0.3807	0.4082	0.2823	0.2416	0.3282
criterion statistic D	softmax	0.1272	0.2033	0.2960	0.3459	0.2431
masked fraction, window of 32	free	0.1806	0.2360	0.1855	0.2000	0.2005
probability of the written token	softmax	0.2090	0.1623	-0.0971	-0.1574	0.1564
top-1 confidence	softmax	0.2084	0.1629	-0.0924	-0.1557	0.1548
masked fraction, window of 16	free	0.1237	0.1615	0.1503	0.1659	0.1504
masked fraction, window of 8	free	0.0843	0.1095	0.1169	0.1347	0.1114
steps remaining in the schedule	free	-0.0889	-0.0693	0.0139	0.0009	0.0432
global masked fraction	free	-0.0887	-0.0691	0.0141	0.0011	0.0432

Table 8: Every pairing of a de-suppressed criterion cell against both reference families at matched work (nearest cell by realized skip). Δ is the paired delta, repaired minus reference; acc and skip are the de-suppressed cell's own accuracy (%) and work skipped (%). Sixteen matched-work pairings against the uniform delay: zero positive separations, two negative, fourteen ties. The last block is the strongest signal of Table 7, outgoing K/V motion, converted into a freezing policy and paired against the delay cell at each site's matched-work anchor (§4.2).

de-suppressed cell	acc	skip	reference	skip gap	Δ	95% CI	verdict
Dream-7B MATH-500							
$D/\delta, \tau = 0.00139$	41.00	27.5	criterion $\varepsilon = 5 \cdot 10^{-8}$	+0.0236	+0.00	[-2.80, +2.80]	tie
$D/\delta, \tau = 0.00139$	41.00	27.5	delay $k = 32$	-0.0099	-0.80	[-4.00, +2.40]	tie
$D/\delta, \tau = 0.00368$	43.60	30.0	criterion $\varepsilon = 5 \cdot 10^{-6}$	-0.0188	+5.80	[+2.60, +9.00]	positive
$D/\delta, \tau = 0.00368$	43.60	30.0	delay $k = 24$	-0.0058	+0.60	[-2.80, +4.00]	tie
$D/\delta, \tau = 0.0348$	38.40	33.3	criterion $\varepsilon = 5 \cdot 10^{-6}$	+0.0143	+0.60	[-2.80, +4.00]	tie
$D/\delta, \tau = 0.0348$	38.40	33.3	delay $k = 12$	-0.0055	-2.40	[-5.80, +1.20]	tie
$D/\delta, \tau = 0.142$	37.60	34.2	criterion $\varepsilon = 5 \cdot 10^{-4}$	-0.0084	+1.20	[-2.40, +4.80]	tie
$D/\delta, \tau = 0.142$	37.60	34.2	delay $k = 12$	+0.0037	-3.20	[-6.80, +0.40]	tie
$\|\Delta z\|_\infty, \tau = 0.250$	40.80	15.8	criterion $\varepsilon = 5 \cdot 10^{-8}$	-0.0937	-0.20	[-3.40, +2.80]	tie
$\|\Delta z\|_\infty, \tau = 0.250$	40.80	15.8	delay $k = 64$	-0.0508	-1.60	[-4.60, +1.40]	tie
$\|\Delta z\|_\infty, \tau = 0.391$	42.40	22.5	criterion $\varepsilon = 5 \cdot 10^{-8}$	-0.0263	+1.40	[-1.80, +4.60]	tie
$\|\Delta z\|_\infty, \tau = 0.391$	42.40	22.5	delay $k = 64$	+0.0166	+0.00	[-3.20, +3.40]	tie
$\|\Delta z\|_\infty, \tau = 0.986$	39.60	28.6	criterion $\varepsilon = 5 \cdot 10^{-6}$	-0.0331	+1.80	[-1.20, +4.80]	tie
$\|\Delta z\|_\infty, \tau = 0.986$	39.60	28.6	delay $k = 32$	+0.0009	-2.20	[-5.80, +1.40]	tie
$\|\Delta z\|_\infty, \tau = 1.73$	38.40	31.9	criterion $\varepsilon = 5 \cdot 10^{-6}$	-0.0004	+0.60	[-3.00, +4.20]	tie
$\|\Delta z\|_\infty, \tau = 1.73$	38.40	31.9	delay $k = 16$	-0.0091	-0.20	[-3.80, +3.60]	tie
LLaDA-8B MATH-500 gen256							
$D/\delta, \tau = 0.000813$	35.60	20.0	criterion $\varepsilon = 5 \cdot 10^{-8}$	+0.0053	-1.00	[-3.40, +1.60]	tie
$D/\delta, \tau = 0.000813$	35.60	20.0	delay $k = 64$	-0.0159	-3.00	[-5.60, -0.40]	negative
$D/\delta, \tau = 0.00117$	36.80	21.1	criterion $\varepsilon = 5 \cdot 10^{-8}$	+0.0169	+0.20	[-2.20, +2.60]	tie
$D/\delta, \tau = 0.00117$	36.80	21.1	delay $k = 64$	-0.0043	-1.80	[-4.40, +0.80]	tie
$D/\delta, \tau = 0.00244$	37.60	24.7	criterion $\varepsilon = 5 \cdot 10^{-8}$	+0.0527	+1.00	[-1.80, +3.80]	tie
$D/\delta, \tau = 0.00244$	37.60	24.7	delay $k = 48$	-0.0063	+1.00	[-2.20, +4.20]	tie
$D/\delta, \tau = 0.00749$	35.60	30.1	criterion $\varepsilon = 5 \cdot 10^{-6}$	-0.0169	+2.80	[+0.20, +5.40]	positive
$D/\delta, \tau = 0.00749$	35.60	30.1	delay $k = 32$	+0.0073	-0.60	[-4.00, +2.80]	tie
$\|\Delta z\|_\infty, \tau = 0.188$	36.80	12.2	criterion $\varepsilon = 5 \cdot 10^{-8}$	-0.0723	+0.20	[-2.40, +2.80]	tie
$\|\Delta z\|_\infty, \tau = 0.188$	36.80	12.2	delay $k = 128$	+0.0271	-1.80	[-4.40, +0.60]	tie
$\|\Delta z\|_\infty, \tau = 0.250$	35.20	20.7	criterion $\varepsilon = 5 \cdot 10^{-8}$	+0.0122	-1.40	[-4.20, +1.40]	tie
$\|\Delta z\|_\infty, \tau = 0.250$	35.20	20.7	delay $k = 64$	-0.0091	-3.40	[-6.40, -0.40]	negative
$\|\Delta z\|_\infty, \tau = 0.312$	35.20	25.0	criterion $\varepsilon = 5 \cdot 10^{-8}$	+0.0560	-1.40	[-4.40, +1.60]	tie
$\|\Delta z\|_\infty, \tau = 0.312$	35.20	25.0	delay $k = 48$	-0.0030	-1.40	[-4.40, +1.60]	tie
$\|\Delta z\|_\infty, \tau = 0.531$	34.40	31.0	criterion $\varepsilon = 5 \cdot 10^{-6}$	-0.0083	+1.60	[-1.00, +4.40]	tie
$\|\Delta z\|_\infty, \tau = 0.531$	34.40	31.0	delay $k = 24$	-0.0057	+0.00	[-3.00, +3.00]	tie
best online signal (outgoing K/V motion) as a policy, vs the delay at each site's matched-work anchor							
Dream: K/V motion (4 steps), $\tau = 2.38$	40.00	30.9	delay $k = 24$	+0.0026	-3.00	[-6.80, +0.80]	tie
LLaDA: K/V motion (1 step), $\tau = 1.69$	35.00	30.4	delay $k = 32$	+0.0097	-1.20	[-4.40, +2.00]	tie

Table 9: Saturation at committed positions. Share of committed position-steps whose statistic D is already below each threshold, by age since commitment. Age 0 is the step at which the written token first enters the forward pass, so age 1 is the first step at which both distributions entering D come from the committed position.

site	age	$D < 5{\cdot}10^{-2}$	$D < 5{\cdot}10^{-3}$	$D < 5{\cdot}10^{-4}$
Dream MATH-500	1	94.04%	86.27%	73.30%
	2	95.29%	90.83%	82.10%
Dream GSM8K	1	95.43%	89.53%	80.11%
	2	96.36%	92.52%	86.10%
LLaDA MATH-500	1	90.08%	77.59%	55.90%
	2	94.53%	88.48%	71.67%
LLaDA GSM8K	1	94.79%	88.49%	72.94%
	2	96.36%	92.71%	82.25%
LLaDA-1.5 GSM8K	1	95.33%	90.38%	77.94%
	2	96.22%	92.65%	84.27%
LLaDA-MoE GSM8K	1	75.07%	47.46%	26.28%
	2	81.70%	55.02%	34.82%
Dream-Coder GSM8K	1	98.04%	93.55%	86.50%
	2	99.10%	96.99%	92.69%

Table 10: The contraction test on every channel we measure. "Rises" is the median fraction of post-commitment steps on which the quantity increases; A2 requires it near zero. "Rebound" is the share of positions that fall below their threshold and later rise back above it, at thresholds calibrated so each channel skips the same work as the criterion at that ε.

site	channel	rises	ever $>$ age-1	rebound @$5{\cdot}10^{-4}$	rebound @$5{\cdot}10^{-6}$
Dream MATH-500	$\|\Delta z\|_\infty$	0.4798	72.14%	87.50%	97.53%
	$\|\Delta$ margin$\|$	0.4990	79.46%	99.46%	99.62%
Dream GSM8K	$\|\Delta z\|_\infty$	0.4766	73.22%	79.67%	91.99%
	$\|\Delta$ margin$\|$	0.5000	81.67%	99.31%	99.53%
LLaDA MATH-500	$\|\Delta z\|_\infty$	0.4790	84.09%	98.94%	99.72%
	$\|\Delta$ margin$\|$	0.4978	84.19%	99.57%	99.71%
LLaDA GSM8K	$\|\Delta z\|_\infty$	0.4796	86.96%	99.33%	99.81%
	$\|\Delta$ margin$\|$	0.5000	86.36%	99.71%	99.72%
LLaDA-1.5 GSM8K	$\|\Delta z\|_\infty$	0.4789	87.09%	99.35%	99.72%
	$\|\Delta$ margin$\|$	0.5000	86.07%	99.66%	99.76%
LLaDA-MoE GSM8K	$\|\Delta z\|_\infty$	0.4918	91.64%	99.10%	98.97%
	$\|\Delta$ margin$\|$	0.5000	93.41%	99.40%	99.12%
Dream-Coder GSM8K	$\|\Delta z\|_\infty$	0.4779	64.39%	77.93%	88.29%
	$\|\Delta$ margin$\|$	0.4985	74.78%	97.32%	99.89%

Table 11: The precision floor. Share of committed position-steps at which the float32 statistic comes out *negative*: impossible for a KL divergence, and a direct read of the cancellation in $\log\mathrm{softmax}$ at saturation. Every number in this paper's rebound tables is float64.

site	float32 $D < 0$	float64 $D < 0$
Dream-7B MATH-500	12.592%	0.000%
Dream-7B GSM8K	18.042%	0.000%
LLaDA-8B MATH-500 gen256	11.266%	0.001%
LLaDA-8B GSM8K gen256	16.726%	0.001%

Table 12: The convexity premise of Proposition 4, measured. $\bar{G}$ is the population mean exposure after waiting w steps; the minorant is the lower convex hull of the 33 points $w = 0, \ldots, 32$. The gap is exactly zero on the region where deployed policies operate.

site	$\bar{G}(0)$	$\sup\|\bar{G} - \mathrm{minorant}\|$	as % of $\bar{G}(0)$	gap on $w \in [0, 7]$
LLaDA MATH-500	577.3	0.1054 (at $w = 24$)	0.018%	0.0000
LLaDA GSM8K	426.5	0.3578 (at $w = 24$)	0.084%	0.0000
Dream MATH-500	325.1	0.1753 (at $w = 13$)	0.054%	0.0000
Dream GSM8K	315.4	0.5209 (at $w = 13$)	0.165%	0.0000

Table 13: Realized staleness of the criterion's caches against the uniform delay's at matched work, measured by a read-only same-canvas probe on the runs behind Table 3 ($n = 100$ per site). Negative means the delay is fresher. All four totals exclude zero. The per-frozen-position-step column normalises for the different amount each policy freezes; the last column is the same ordering computed on the unfrozen run.

site	total staleness, delay $-$ criterion	per frozen position-step	exposure (unfrozen run), delay $-$ criterion
Dream-7B MATH-500	-126.9k [-142.6, -111.2]k	-3.9820 [-4.4083, -3.5588]	-16.3k [-18.5, -14.1]k
Dream-7B GSM8K	-160.8k [-172.8, -148.9]k	-4.7793 [-5.1117, -4.4592]	-20.5k [-22.2, -18.9]k
LLaDA-8B MATH-500 gen256	-244.1k [-270.7, -219.3]k	-6.6409 [-7.2412, -6.0874]	-29.3k [-37.0, -22.8]k
LLaDA-8B GSM8K gen256	-272.7k [-286.4, -259.4]k	-6.7008 [-7.0143, -6.3994]	-27.4k [-31.5, -23.8]k

Table 14: Matched-work pairs at every site: each criterion cell at $\varepsilon = 5{\cdot}10^{-4}$ and $5{\cdot}10^{-6}$ with the delay cell nearest it in skipped work. Δ is delay minus criterion, paired 95% bootstrap. The two MATH-1500 rows are the pre-specified held-out comparisons of Table 3. The LLaDA-MoE $\varepsilon = 5{\cdot}10^{-4}$ row is at n=300; its $\varepsilon = 5{\cdot}10^{-6}$ row is at the full test set, n=1319. No pairing on a site with $n \geq 257$ separates against the delay; see §5 for LLaDA-8B HumanEval.

site	criterion cell	delay cell	skip gap	Δ [95% CI]
Dream-7B MATH-500	criterion $\varepsilon = 5 \cdot 10^{-4}$	delay $k = 8$	-0.0008	+5.40 [+1.40, +9.40]
	criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 16$	+0.0087	+0.80 [-2.60, +4.40]
Dream-7B GSM8K	criterion $\varepsilon = 5 \cdot 10^{-4}$	delay $k = 6$	+0.0016	+1.14 [-0.76, +3.03]
	criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 16$	-0.0026	+1.14 [-0.83, +3.18]
Dream-7B MBPP	criterion $\varepsilon = 5 \cdot 10^{-4}$	delay $k = 8$	-0.0010	+1.17 [-3.11, +5.45]
	criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 16$	+0.0084	+0.78 [-2.72, +4.28]
Dream-7B HumanEval	criterion $\varepsilon = 5 \cdot 10^{-4}$	delay $k = 6$	+0.0009	-1.83 [-7.32, +3.66]
	criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 16$	-0.0001	+0.00 [-4.88, +4.88]
Dream-7B CRUXEval-O	criterion $\varepsilon = 5 \cdot 10^{-4}$	delay $k = 8$	-0.0018	+0.50 [-2.00, +3.00]
	criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 24$	-0.0085	-0.62 [-3.00, +1.88]
LLaDA-8B MATH-500 gen256	criterion $\varepsilon = 5 \cdot 10^{-4}$	delay $k = 8$	-0.0019	+1.80 [-1.40, +5.00]
	criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 24$	-0.0027	+1.60 [-1.40, +4.60]
LLaDA-8B GSM8K gen256	criterion $\varepsilon = 5 \cdot 10^{-4}$	delay $k = 8$	-0.0047	+2.05 [+0.08, +3.94]
	criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 16$	+0.0049	+1.21 [-0.68, +3.11]
LLaDA-8B MBPP (gen 128)	criterion $\varepsilon = 5 \cdot 10^{-4}$	delay $k = 6$	-0.0003	-1.95 [-4.67, +0.78]
	criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 16$	-0.0077	-0.39 [-2.72, +1.95]
LLaDA-8B HumanEval (gen 128)	criterion $\varepsilon = 5 \cdot 10^{-4}$	delay $k = 6$	-0.0002	+3.66 [-0.61, +7.93]
	criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 16$	-0.0034	-4.27 [-8.54, -0.61]
LLaDA-8B CRUXEval-O (gen 128)	criterion $\varepsilon = 5 \cdot 10^{-4}$	delay $k = 8$	-0.0025	-0.50 [-3.38, +2.38]
	criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 24$	-0.0106	+2.25 [-0.50, +5.00]
Dream-7B MATH-1500 (held out)	criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 24$	-0.0135	+2.80 [+0.80, +4.87]
LLaDA-8B MATH-1500 (held out)	criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 32$	-0.0228	+2.27 [+0.40, +4.07]
LLaDA-1.5 GSM8K (n=1319)	criterion $\varepsilon = 5 \cdot 10^{-4}$	delay $k = 16$	-0.0278	+2.20 [+0.15, +4.25]
	criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 16$	+0.0042	+1.21 [-0.61, +3.03]
Dream-Coder GSM8K (n=1319)	criterion $\varepsilon = 5 \cdot 10^{-4}$	delay $k = 16$	-0.0122	+3.03 [+1.21, +4.85]
	criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 16$	-0.0054	+1.59 [-0.08, +3.26]
LLaDA-MoE GSM8K (n=300)	criterion $\varepsilon = 5 \cdot 10^{-4}$	delay $k = 16$	+0.0300	-1.67 [-4.67, +1.33]
LLaDA-MoE GSM8K (n=1319)	criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 64$	-0.0230	-0.61 [-2.05, +0.83]
LLaDA-1.5 MATH-500 (n=500)	criterion $\varepsilon = 5 \cdot 10^{-4}$	delay $k = 16$	-0.0250	-0.80 [-3.80, +2.20]
	criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 16$	+0.0188	-0.20 [-3.40, +2.80]

Table 15: Pairing each criterion cell with the nearest delay cell that skips at least as much work. None of the 28 pairings separates in either direction. Sites where no delay cell skips at least as much work as the criterion cell are omitted.

site	n	criterion	delay	delay skip − crit skip	Δ	95% CI
Dream-7B CRUXEval-O	800	criterion $\varepsilon = 5 \cdot 10^{-4}$	delay $k = 6$	+0.0034	+1.50	[-1.25, +4.25]
		criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 16$	+0.0111	-0.62	[-3.25, +2.00]
LLaDA-8B CRUXEval-O gen128	800	criterion $\varepsilon = 5 \cdot 10^{-4}$	delay $k = 6$	+0.0051	+0.25	[-2.62, +3.12]
		criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 16$	+0.0165	+1.25	[-1.50, +4.00]
Dream-7B GSM8K	1319	criterion $\varepsilon = 5 \cdot 10^{-4}$	delay $k = 6$	+0.0016	+1.14	[-0.76, +3.03]
		criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 12$	+0.0086	-0.76	[-2.73, +1.21]
LLaDA-8B GSM8K gen128	1319	criterion $\varepsilon = 5 \cdot 10^{-4}$	delay $k = 6$	+0.0016	-0.68	[-2.35, +0.91]
		criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 12$	+0.0111	-0.45	[-1.97, +0.99]
Dream-7B HumanEval	164	criterion $\varepsilon = 5 \cdot 10^{-4}$	delay $k = 6$	+0.0009	-1.83	[-7.32, +3.66]
		criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 12$	+0.0088	-2.44	[-7.32, +2.44]
LLaDA-8B HumanEval gen128	164	criterion $\varepsilon = 5 \cdot 10^{-4}$	delay $k = 4$	+0.0061	+3.05	[+0.00, +6.71]
		criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 12$	+0.0083	+1.83	[-1.83, +5.49]
LLaDA-8B GSM8K gen256	1319	criterion $\varepsilon = 5 \cdot 10^{-4}$	delay $k = 2$	+0.0131	+1.21	[-0.61, +3.03]
		criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 16$	+0.0049	+1.21	[-0.68, +3.11]
LLaDA-8B MATH-500 gen256	500	criterion $\varepsilon = 5 \cdot 10^{-4}$	delay $k = 2$	+0.0159	+1.80	[-1.40, +5.00]
		criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 16$	+0.0196	+2.60	[-0.40, +5.60]
Dream-7B MATH-500	500	criterion $\varepsilon = 5 \cdot 10^{-4}$	delay $k = 6$	+0.0049	-0.40	[-4.00, +3.40]
		criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 16$	+0.0087	+0.80	[-2.60, +4.40]
LLaDA-8B MATH-500 gen128	500	criterion $\varepsilon = 5 \cdot 10^{-4}$	delay $k = 6$	+0.0062	+1.60	[-1.40, +4.60]
		criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 16$	+0.0196	+1.20	[-1.60, +4.20]
Dream-7B MBPP	257	criterion $\varepsilon = 5 \cdot 10^{-4}$	delay $k = 6$	+0.0046	+2.72	[-1.17, +6.61]
		criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 16$	+0.0084	+0.78	[-2.72, +4.28]
LLaDA-8B MBPP gen128	257	criterion $\varepsilon = 5 \cdot 10^{-4}$	delay $k = 4$	+0.0085	+1.17	[-1.56, +4.28]
		criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 12$	+0.0085	-0.39	[-2.72, +1.95]
LLaDA-1.5 GSM8K	1319	criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 16$	+0.0042	+1.21	[-0.61, +3.03]
LLaDA-1.5 MATH-500	500	criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 16$	+0.0188	-0.20	[-3.40, +2.80]
LLaDA-MoE GSM8K	300	criterion $\varepsilon = 5 \cdot 10^{-4}$	delay $k = 16$	+0.0300	-1.67	[-4.67, +1.33]
		criterion $\varepsilon = 5 \cdot 10^{-6}$	delay $k = 32$	+0.0717	-1.67	[-4.33, +1.00]

Table 16: Shuffling the criterion's per-token decisions does not reduce accuracy. *Top*: the realized waits reassigned by a random permutation within each question, three seeds per model on each of five models, against the unpermuted replay. *Bottom*: how often the criterion's own threshold test disagrees with the same test evaluated on a read-only dense forward over the same canvas.

site	arm	computation vs criterion	Δ	95% CI
Dream MATH-500	permutation 1	-0.297%	+1.40	[-2.40, +5.00]
	permutation 2	-0.316%	+0.00	[-3.80, +3.80]
	permutation 3	-0.318%	+0.80	[-3.00, +4.60]
LLaDA MATH-500	permutation 1	-0.618%	+0.00	[-3.00, +3.00]
	permutation 2	-0.604%	+2.20	[-1.20, +5.60]
	permutation 3	-0.585%	+0.40	[-2.60, +3.40]
LLaDA-1.5 GSM8K	permutation 1	-4.614%†	-1.00	[-8.00, +5.00]
	permutation 2	-4.520%†	+1.00	[-5.00, +7.00]
	permutation 3	-4.656%†	+0.00	[-6.00, +6.00]
LLaDA-MoE GSM8K	permutation 1	-11.368%†	+1.00	[-6.00, +7.00]
	permutation 2	-11.162%†	+3.00	[-3.00, +9.00]
	permutation 3	-11.552%†	+1.00	[-4.00, +6.00]
Dream-Coder GSM8K	permutation 1	-0.518%†	+0.00	[-7.00, +7.00]
	permutation 2	-0.497%†	+1.00	[-7.00, +9.00]
	permutation 3	-0.508%†	+0.00	[-8.00, +7.00]
decisions differing from the same test on a dense forward, $\varepsilon = 5{\cdot}10^{-4}$				
Dream MATH-500	threshold test	—	29.82%	189,522 position-steps
LLaDA MATH-500	threshold test	—	19.16%	89,431 position-steps

† For these three models, computation is compared on the realized frozen fraction; each of the nine does at most the criterion's computation.

Table 17: Step-count sweep, LLaDA-8B GSM8K gen256 block 32, $n = 300$. Every k from $T/16$ to $T/2$ is inside the plateau at both step counts: the paired 95% interval against the unfrozen model includes zero at every cell. "in plateau" is that interval including zero.

T	k	k/T	acc (%)	work skipped (%)	Δ vs unfrozen [95% CI]	in plateau
64	—	—	unfrozen 72.67	0.00	—	—
64	4	0.0625	73.67	32.61	+1.00 [-3.33, +5.33]	yes
64	8	0.1250	72.67	28.30	+0.00 [-4.67, +4.33]	yes
64	12	0.1875	73.33	24.30	+0.67 [-3.00, +4.33]	yes
64	16	0.2500	71.67	20.60	-1.00 [-5.00, +3.00]	yes
64	24	0.3750	72.33	14.12	-0.33 [-4.00, +3.00]	yes
64	32	0.5000	72.33	8.86	-0.33 [-3.00, +2.00]	yes
128	—	—	unfrozen 77.00	0.00	—	—
128	8	0.0625	75.00	33.45	-2.00 [-6.33, +2.33]	yes
128	16	0.1250	74.67	29.09	-2.33 [-6.00, +1.33]	yes
128	24	0.1875	74.67	25.03	-2.33 [-6.00, +1.33]	yes
128	32	0.2500	76.67	21.28	-0.33 [-3.67, +3.00]	yes
128	48	0.3750	78.33	14.68	+1.33 [-1.33, +4.00]	yes
128	64	0.5000	76.67	9.31	-0.33 [-3.33, +2.67]	yes

Table 18: Dream-7B GSM8K, $n = 1319$. Work skipped and FLOP ratio are per-cell measurements, not a fit. *: 95% paired CI against the unfrozen model excludes zero.

arm	acc (%)	Δ vs unfrozen	work skipped (%)	FLOP ratio
unfrozen	80.82	—	0.00	1.0000
freeze at commit	72.40*	-8.42 [-10.69, -6.14]	37.48	0.6268
criterion $\varepsilon = 5 \cdot 10^{-2}$	76.12*	-4.70 [-6.82, -2.50]	35.99	0.6416
criterion $\varepsilon = 5 \cdot 10^{-3}$	77.26*	-3.56 [-5.54, -1.52]	35.90	0.6425
criterion $\varepsilon = 5 \cdot 10^{-4}$	77.26*	-3.56 [-5.53, -1.52]	35.56	0.6459
criterion $\varepsilon = 5 \cdot 10^{-6}$	78.54*	-2.27 [-4.25, -0.30]	33.16	0.6702
criterion $\varepsilon = 5 \cdot 10^{-8}$	79.00*	-1.82 [-3.56, -0.08]	27.64	0.7256
delay $k = 1$	76.65*	-4.17 [-6.37, -1.97]	37.18	0.6298
delay $k = 2$	77.10*	-3.71 [-5.91, -1.44]	36.89	0.6327
delay $k = 4$	76.72*	-4.09 [-6.14, -2.12]	36.30	0.6385
delay $k = 6$	78.39*	-2.43 [-4.32, -0.45]	35.72	0.6443
delay $k = 8$	78.17*	-2.65 [-4.55, -0.68]	35.15	0.6500
delay $k = 12$	77.79*	-3.03 [-4.93, -1.14]	34.01	0.6613
delay $k = 16$	79.68	-1.14 [-2.96, +0.76]	32.89	0.6724
delay $k = 24$	79.91	-0.91 [-2.65, +0.83]	30.72	0.6941
delay $k = 32$	79.76	-1.06 [-2.65, +0.53]	28.61	0.7151

Table 19: Dream-7B MATH-1500 (held out), $n = 1500$. Work skipped and FLOP ratio are per-cell measurements, not a fit. *: 95% paired CI against the unfrozen model excludes zero.

arm	acc (%)	Δ vs unfrozen	work skipped (%)	FLOP ratio
unfrozen	37.47	—	0.00	1.0000
freeze at commit	33.60*	-3.87 [-6.13, -1.53]	37.42	0.6371
criterion $\varepsilon = 5 \cdot 10^{-4}$	35.13*	-2.33 [-4.47, -0.27]	35.17	0.6589
criterion $\varepsilon = 5 \cdot 10^{-6}$	35.87	-1.60 [-3.60, +0.40]	32.01	0.6899
delay $k = 24$	38.67	+1.20 [-0.67, +3.07]	30.67	0.7025

Table 20: Dream-7B MATH-500, $n = 500$. Work skipped and FLOP ratio are per-cell measurements, not a fit. *: 95% paired CI against the unfrozen model excludes zero. At $n = 500$ the paired interval against the unfrozen model spans about ± 4 points, and accuracy moves by up to 6 points between adjacent k. The pre-specified comparisons use the $n = 1500$ held-out split (Table 3).

arm	acc (%)	Δ vs unfrozen	work skipped (%)	FLOP ratio
unfrozen	42.20	—	0.00	1.0000
freeze at commit	32.40*	-9.80 [-13.80, -5.80]	37.33	0.6374
criterion $\varepsilon = 5 \cdot 10^{-2}$	37.40*	-4.80 [-8.60, -1.00]	35.83	0.6519
criterion $\varepsilon = 5 \cdot 10^{-3}$	37.20*	-5.00 [-9.00, -1.20]	35.67	0.6535
criterion $\varepsilon = 5 \cdot 10^{-4}$	36.40*	-5.80 [-9.60, -2.00]	35.09	0.6593
criterion $\varepsilon = 5 \cdot 10^{-6}$	37.80*	-4.40 [-8.00, -0.80]	31.90	0.6908
criterion $\varepsilon = 5 \cdot 10^{-8}$	41.00	-1.20 [-4.40, +2.00]	25.15	0.7578
delay $k = 1$	34.00*	-8.20 [-12.20, -4.40]	37.03	0.6402
delay $k = 2$	36.80*	-5.40 [-9.00, -1.80]	36.74	0.6431
delay $k = 4$	39.00	-3.20 [-7.00, +0.60]	36.16	0.6487
delay $k = 6$	36.00*	-6.20 [-10.00, -2.20]	35.58	0.6543
delay $k = 8$	41.80	-0.40 [-4.60, +3.80]	35.01	0.6599
delay $k = 12$	40.80	-1.40 [-5.00, +2.20]	33.88	0.6709
delay $k = 16$	38.60*	-3.60 [-7.00, -0.20]	32.76	0.6817
delay $k = 24$	43.00	+0.80 [-2.60, +4.20]	30.59	0.7028
delay $k = 32$	41.80	-0.40 [-3.60, +2.80]	28.50	0.7231
delay $k = 48$	41.00	-1.20 [-3.80, +1.40]	24.53	0.7617
delay $k = 64$	42.40	+0.20 [-2.40, +2.80]	20.86	0.7974
delay $k = 128$	41.60	-0.60 [-2.00, +0.80]	9.15	0.9111

Table 21: LLaDA-8B GSM8K gen256, $n = 1319$. Work skipped and FLOP ratio are per-cell measurements, not a fit. *: 95% paired CI against the unfrozen model excludes zero.

arm	acc (%)	Δ vs unfrozen	work skipped (%)	FLOP ratio
unfrozen	76.72	—	0.00	1.0000
freeze at commit	72.02*	-4.70 [-7.05, -2.43]	38.50	0.6166
criterion $\varepsilon = 5 \cdot 10^{-2}$	73.46*	-3.26 [-5.53, -1.06]	37.01	0.6315
criterion $\varepsilon = 5 \cdot 10^{-4}$	71.95*	-4.78 [-6.98, -2.65]	36.59	0.6358
criterion $\varepsilon = 5 \cdot 10^{-6}$	74.07*	-2.65 [-4.78, -0.53]	33.32	0.6685
criterion $\varepsilon = 5 \cdot 10^{-8}$	76.42	-0.30 [-2.12, +1.52]	24.10	0.7608
delay $k = 2$	73.16*	-3.56 [-5.69, -1.52]	37.90	0.6226
delay $k = 8$	74.00*	-2.73 [-4.70, -0.83]	36.12	0.6404
delay $k = 16$	75.28	-1.44 [-3.41, +0.53]	33.81	0.6633
delay $k = 24$	75.82	-0.91 [-2.73, +0.91]	31.58	0.6855
delay $k = 32$	77.03	+0.30 [-1.52, +2.05]	29.43	0.7070
delay $k = 48$	76.50	-0.23 [-1.82, +1.29]	25.35	0.7476
delay $k = 64$	76.35	-0.38 [-1.82, +1.06]	21.57	0.7852

Table 22: LLaDA-8B MATH-1500 (held out), $n = 1500$. Work skipped and FLOP ratio are per-cell measurements, not a fit. *: 95% paired CI against the unfrozen model excludes zero.

arm	acc (%)	Δ vs unfrozen	work skipped (%)	FLOP ratio
unfrozen	37.00	—	0.00	1.0000
freeze at commit	32.33*	-4.67 [-6.80, -2.53]	38.57	0.6270
criterion $\varepsilon = 5 \cdot 10^{-4}$	33.80*	-3.20 [-5.27, -1.13]	36.36	0.6485
criterion $\varepsilon = 5 \cdot 10^{-6}$	34.53*	-2.47 [-4.47, -0.47]	31.76	0.6938
criterion $\varepsilon = 5 \cdot 10^{-8}$	36.93	-0.07†	19.20	—
delay $k = 16$	36.33	-0.67†	33.88	—
delay $k = 24$	36.33	-0.67†	31.64	—
delay $k = 32$	36.80	-0.20 [-1.93, +1.53]	29.48	0.7149
delay $k = 48$	36.40	-0.60 [-2.20, +1.00]	25.39	0.7544
delay $k = 64$	35.93	-1.07†	21.61	—

†interval not computed

Table 23: LLaDA-8B MATH-500 gen256, $n = 500$. Work skipped and FLOP ratio are per-cell measurements, not a fit. *: 95% paired CI against the unfrozen model excludes zero.

arm	acc (%)	Δ vs unfrozen	work skipped (%)	FLOP ratio
unfrozen	37.40	—	0.00	1.0000
freeze at commit	32.40*	-5.00 [-8.60, -1.40]	38.49	0.6270
criterion $\varepsilon = 5 \cdot 10^{-2}$	32.80*	-4.60 [-8.20, -1.20]	36.97	0.6418
criterion $\varepsilon = 5 \cdot 10^{-4}$	31.80*	-5.60 [-9.20, -2.20]	36.30	0.6484
criterion $\varepsilon = 5 \cdot 10^{-6}$	32.80*	-4.60 [-7.80, -1.40]	31.84	0.6925
criterion $\varepsilon = 5 \cdot 10^{-8}$	36.60	-0.80 [-3.60, +2.00]	19.44	0.8142
delay $k = 2$	33.60	-3.80 [-7.60, +0.00]	37.89	0.6329
delay $k = 8$	33.60*	-3.80 [-7.40, -0.40]	36.11	0.6501
delay $k = 16$	35.40	-2.00 [-5.40, +1.20]	33.80	0.6724
delay $k = 24$	34.40	-3.00 [-6.00, +0.00]	31.57	0.6941
delay $k = 32$	36.20	-1.20 [-4.40, +2.00]	29.42	0.7149
delay $k = 48$	36.60	-0.80 [-3.60, +2.00]	25.34	0.7544
delay $k = 64$	38.60	+1.20 [-1.40, +3.80]	21.57	0.7910
delay $k = 96$	38.00	+0.60 [-1.40, +2.60]	14.93	0.8553
delay $k = 128$	38.60	+1.20 [-0.60, +3.00]	9.51	0.9079

Table 24: LLaDA-1.5 GSM8K, $n = 1319$ (the full GSM8K test set), $T = 256$ steps. Work skipped is a per-cell measurement; these runs carry no per-cell FLOP accounting, so that column is omitted rather than fitted. *: 95% paired CI against the unfrozen model excludes zero. The matched-work pairings among these arms are the rows of Table 4.

arm	acc (%)	Δ vs unfrozen	work skipped (%)
unfrozen	78.92	—	0.00
criterion $\varepsilon = 5 \cdot 10^{-4}$	74.98*	-3.94 [-5.99, -1.82]	36.59
criterion $\varepsilon = 5 \cdot 10^{-6}$	75.97*	-2.96 [-4.85, -1.06]	33.40
delay $k = 16$	77.18	-1.74 [-3.56, +0.08]	33.81
delay $k = 32$	77.71	-1.21 [-2.88, +0.45]	29.43

Table 25: Dream-Coder GSM8K, $n = 1319$ (the full GSM8K test set), $T = 768$ steps. Work skipped is a per-cell measurement; these runs carry no per-cell FLOP accounting, so that column is omitted rather than fitted. *: 95% paired CI against the unfrozen model excludes zero. The matched-work pairings among these arms are the rows of Table 4.

arm	acc (%)	Δ vs unfrozen	work skipped (%)
unfrozen	81.65	—	0.00
criterion $\varepsilon = 5 \cdot 10^{-4}$	77.41*	-4.25 [-6.22, -2.27]	44.36
criterion $\varepsilon = 5 \cdot 10^{-6}$	78.85*	-2.81 [-4.55, -1.06]	43.68
delay $k = 16$	80.44	-1.21 [-2.96, +0.53]	43.14
delay $k = 24$	81.05	-0.61 [-2.20, +0.99]	42.22

Table 26: LLaDA-1.5 GSM8K, $n = 300$, $T = 256$ steps: the k/T plateau sweep; the full-test-set arms are in Table 24. Work skipped is a per-cell measurement; these runs carry no per-cell FLOP accounting, so that column is omitted rather than fitted. *: 95% paired CI against the unfrozen model excludes zero.

arm	acc (%)	Δ vs unfrozen	k/T	work skipped (%)
unfrozen	80.00	—	—	0.00
freeze at commit	75.00*	-5.00 [-9.33, -0.67]	—	38.58
criterion $\varepsilon = 5 \cdot 10^{-4}$	74.33*	-5.67 [-10.00, -1.67]	—	36.66
criterion $\varepsilon = 5 \cdot 10^{-6}$	79.00	-1.00 [-4.67, +2.33]	—	33.51
delay $k = 16$	80.33	+0.33 [-2.67, +3.33]	0.0625	33.88
delay $k = 32$	77.67	-2.33 [-5.00, +0.33]	0.1250	29.49
delay $k = 64$	78.67	-1.33 [-4.00, +1.33]	0.2500	21.61
delay $k = 128$	81.00	+1.00 [-0.33, +2.67]	0.5000	9.53

Table 27: LLaDA-MoE GSM8K, $n = 300$, $T = 128$ steps: the k/T plateau sweep. Work skipped is a per-cell measurement; these runs carry no per-cell FLOP accounting, so that column is omitted rather than fitted. *: 95% paired CI against the unfrozen model excludes zero.

arm	acc (%)	Δ vs unfrozen	k/T	work skipped (%)
unfrozen	47.67	—	—	0.00
freeze at commit	45.33	-2.33 [-6.00, +1.33]	—	30.63
criterion $\varepsilon = 5 \cdot 10^{-4}$	47.67	+0.00 [-3.00, +3.00]	—	20.37
criterion $\varepsilon = 5 \cdot 10^{-6}$	48.33	+0.67 [-2.00, +3.33]	—	9.92
delay $k = 8$	44.33*	-3.33 [-6.67, -0.33]	0.0625	26.88
delay $k = 16$	46.00	-1.67 [-4.33, +0.67]	0.1250	23.37
delay $k = 32$	46.67	-1.00 [-3.33, +1.00]	0.2500	17.09
delay $k = 64$	47.33	-0.33 [-2.67, +2.00]	0.5000	7.48

Table 28: Dream-Coder GSM8K, $n = 300$, $T = 768$ steps: the k/T plateau sweep; the full-test-set arms are in Table 25. Work skipped is a per-cell measurement; these runs carry no per-cell FLOP accounting, so that column is omitted rather than fitted. *: 95% paired CI against the unfrozen model excludes zero.

arm	acc (%)	Δ vs unfrozen	k/T	work skipped (%)
unfrozen	83.00	—	—	0.00
freeze at commit	73.33*	-9.67 [-14.33, -5.00]	—	45.04
criterion $\varepsilon = 5 \cdot 10^{-4}$	79.00*	-4.00 [-7.67, -0.67]	—	44.40
criterion $\varepsilon = 5 \cdot 10^{-6}$	79.67	-3.33 [-7.00, +0.33]	—	43.74
delay $k = 48$	83.00	+0.00 [-2.33, +2.33]	0.0625	39.57
delay $k = 96$	81.33	-1.67 [-4.00, +0.33]	0.1250	34.45
delay $k = 192$	82.00	-1.00 [-2.67, +0.33]	0.2500	25.28
delay $k = 384$	83.00	+0.00 [+0.00, +0.00]	0.5000	11.19

Table 29: LLaDA-1.5 MATH-500, $n = 500$, $T = 256$ steps: the k/T plateau sweep. Work skipped is a per-cell measurement; these runs carry no per-cell FLOP accounting, so that column is omitted rather than fitted. *: 95% paired CI against the unfrozen model excludes zero.

arm	acc (%)	Δ vs unfrozen	k/T	work skipped (%)
unfrozen	38.80	—	—	0.00
freeze at commit	32.40*	-6.40 [-10.00, -3.00]	—	38.49
criterion $\varepsilon = 5 \cdot 10^{-4}$	35.40	-3.40 [-6.80, +0.00]	—	36.31
criterion $\varepsilon = 5 \cdot 10^{-6}$	34.80*	-4.00 [-7.20, -0.80]	—	31.93
delay $k = 16$	34.60*	-4.20 [-7.40, -1.00]	0.0625	33.80
delay $k = 32$	36.20	-2.60 [-5.60, +0.40]	0.1250	29.42
delay $k = 64$	38.00	-0.80 [-3.40, +1.80]	0.2500	21.57
delay $k = 128$	39.00	+0.20 [-1.40, +1.80]	0.5000	9.51

F OPEN-ENDED GENERATION

The failures this paper documents are on exact-answer reasoning, where a single token decides the verdict. This appendix is the other regime, where prior reports find the criterion family holds quality, and our measurements agree: on neither protocol does the delay separate from the criterion, and the point estimates favour the delay.

Table 30: Generative perplexity on WikiText-103, LLaDA-8B-Base, 120 prompts, scored by an external 8B model. Lower is better. At matched work the delay is at or below the criterion: the delay cell nearest each criterion cell in skipped work is k=16 for both.

arm	gen-PPL	work skipped (%)	scored tokens
unfrozen	1.6876	0.00	23,783
criterion $5{\cdot}10^{-4}$	1.8112	38.24	23,347
criterion $5{\cdot}10^{-3}$	1.8078	40.08	23,876
delay k=16	1.8020	37.44	23,457
delay k=24	1.7651	34.97	23,758
delay k=32	1.6968	32.59	23,954

Generative perplexity. WikiText-103 test, odd-indexed non-blank lines, the first 120, each prompt cut to at most 64 tokens (mean 43.1; 49 prompts are shorter than the cut). Continuations are decoded by LLaDA-8B-Base at 256 generation positions and 256 steps, greedy, and scored by Llama-3-8B, following the prompt construction and scorer of Oba et al. (2026). Our decode differs from theirs in three respects: block length 32 rather than a single block, batch size 1 rather than 4, and continuations cut at end-of-text. Table 30 gives the micro-averaged perplexity over the 23,347–23,954 tokens each arm produced. The delay cell nearest each criterion cell in skipped work is $k = 16$ for both criterion thresholds, and at that pairing the delay's perplexity is below the criterion's: 1.8020 against 1.8112 at $\varepsilon = 5{\cdot}10^{-4}$ (work gap -0.8 points) and against 1.8078 at $\varepsilon = 5{\cdot}10^{-3}$ (work gap -2.5 points). Every frozen arm is above the unfrozen model's 1.6876; at matched work the delay's increase is the smaller one.

Table 31: MT-Bench, first turn, 80 questions, LLaDA-8B-Instruct, judged by Claude Opus 5 under the FastChat single-answer protocol with the released GPT-4 reference answers for math, reasoning and coding. Scores are comparable between arms, not to published MT-Bench numbers, which use a GPT-4 judge. Judge noise, from 48 answers scored twice by different judge instances: 68.8% exact agreement, 95.8% within one point, mean absolute difference 0.35.

arm	mean score	work skipped (%)	Δ vs unfrozen [95% CI]
unfrozen	5.16	0.0	—
criterion $5{\cdot}10^{-4}$	5.11	35.6	-0.05 [-0.42, +0.29]
criterion $5{\cdot}10^{-3}$	5.15	36.7	-0.01 [-0.46, +0.42]
delay k=16	5.42	34.2	+0.26 [-0.14, +0.68]
delay k=24	5.39	31.9	+0.23 [-0.05, +0.54]
delay k=32	5.29	29.8	+0.12 [-0.10, +0.40]
matched-work pairings, delay minus criterion			
delay k=16 vs criterion $5{\cdot}10^{-4}$	—	-1.4	+0.31 [-0.03, +0.69]
delay k=16 vs criterion $5{\cdot}10^{-3}$	—	-2.5	+0.28 [-0.01, +0.57]

MT-Bench. The 80 first-turn questions, generated by LLaDA-8B-Instruct under the same three delay and two criterion arms, judged under the FastChat single-answer protocol with the released GPT-4 reference answers for the math, reasoning and coding categories. *The judge is Claude Opus 5*, run as independent judge instances blind to the arm: no instance saw two arms' answers to the same question, and 48 answers were scored a second time by a different instance to measure judge noise (68.8% exact agreement, 95.8% within one point, mean absolute difference 0.35). Scores are therefore comparable *between arms* and not to published MT-Bench numbers, which use a GPT-4 judge. No arm separates from the unfrozen model. At matched work the delay scores above the criterion at both thresholds, $+0.31$ $[-0.03, +0.69]$ and $+0.28$ $[-0.01, +0.57]$ points on the 1–10 scale; neither interval excludes zero, so this is a tie in the delay's direction, not a separation.

G WALL-CLOCK DETAILS

Table 32: Small-batch throughput, LLaDA-8B GSM8K. At batch 1 the decode is latency-bound and no arm, including freeze at commit, runs faster than the stock model; the saving appears from batch 2 and is most of the ideal by batch 4. The delay arm here uses $k = 32$; the comparison at batch sizes 8 and 32 in Table 5 uses $k = 16$.

model	batch	arm	k / ε	vs stock	ideal	realised (time saved)
LLaDA-8B GSM8K	1	criterion	ε=5·10^{-6}	0.96×	1.50×	−12.3%
LLaDA-8B GSM8K	1	delay	k=32	0.96×	1.41×	−14.3%
LLaDA-8B GSM8K	1	freeze at commit	k=0	0.96×	1.62×	−10.0%
LLaDA-8B GSM8K	2	criterion	ε=5·10^{-6}	1.18×	1.50×	45.4%
LLaDA-8B GSM8K	2	delay	k=32	1.16×	1.41×	46.9%
LLaDA-8B GSM8K	2	freeze at commit	k=0	1.22×	1.62×	47.4%
LLaDA-8B GSM8K	4	criterion	ε=5·10^{-6}	1.33×	1.50×	75.5%
LLaDA-8B GSM8K	4	delay	k=32	1.29×	1.41×	77.7%
LLaDA-8B GSM8K	4	freeze at commit	k=0	1.44×	1.62×	79.9%

The executor. Timing uses a compacted executor that gathers the active rows, runs the query projection, attention output and feed-forward on the compact set with cached keys and values for frozen positions, and scatters back. With nothing frozen it is bit-identical to the stock model: on Dream the maximum absolute logit difference is exactly zero and the argmax agrees at 320 of 320 positions. All ratios are against the stock model. The compacted executor with nothing frozen is slightly slower than the stock model, so dividing by it would inflate every speedup.

Replayed schedules. Each timing run replays the freeze mask an accuracy run actually produced, step for step, rather than running the policy live. A policy's decisions do not depend on how fast the decode runs, so the replay does that arm's work exactly, and it keeps the criterion's own scoring pass out of the timing loop; that pass is accounted separately at 0.0072% of a step.

Bit-reproducibility. The compacted executor is arithmetically exact but not bit-reproducible: a GEMM over gathered rows can select a different kernel, and therefore a different summation order, than the same GEMM over the full canvas, and one last-ulp difference rewrites a decode because a decode is a sequence of argmaxes. Timing therefore uses replayed schedules, and no accuracy number in this paper comes from the compacted executor.

H EVALUATION DETAILS

Held-out MATH-1500. The MATH-500 split used for development and the MATH-1500 split used for the pre-specified comparisons are disjoint. Every threshold, every k, and the de-suppressed criterion's operating cell were chosen on MATH-500 and on GSM8K; the MATH-1500 runs were made only after all of those choices were fixed.

Decoding. LLaDA-8B-Instruct: 256 generation positions, 256 steps, temperature 0, semi-autoregressive block length 32. Dream-7B-Instruct: 256 generation positions, 256 steps, temperature 0.2, top-p 0.95, entropy ordering. LLaDA-1.5: as LLaDA-8B. LLaDA-MoE-7B-A1B: 128 generation positions, 128 steps, block length 32, temperature 0. Dream-Coder-7B: 768 generation positions, 768 steps, temperature 0.1, top-p 0.95, entropy ordering. All use the checkpoint's own chat template. GSM8K is the full test set ($n = 1319$) on LLaDA-8B, Dream-7B, LLaDA-1.5 and Dream-Coder, and the first 300 questions on LLaDA-MoE; MATH-500 and MATH-1500 are $n = 500$ and $n = 1500$. Freezing arms use the full scope (cached keys, values and feed-forward output, with the posterior pinned), so a frozen position's own statistic is exactly zero thereafter (Appendix A).

The averaged view. Figure 1's left panel is the average over *all* response positions at each sampling step, masked and committed alike, which is the protocol under which the contraction assumption was validated. Across the four sites of Figure 2 that average declines by 1.6–13.2×, it is monotone on none of them (48–51% of steps rise), and it sits four to five orders of magnitude above the median of the same population (1.0·10^{-6} against 4.3·10^{-2} on Dream MATH-500).