SOFT CONFIDENCE ORDER IN MASKED DIFFUSION: QUALITY–DIVERSITY TRADE-OFFS ACROSS MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Masked diffusion language models predict all masked tokens in one forward pass, leaving the decoder to decide which predictions to commit at each model call. Most-confident-first selection is common, but its strength is rarely isolated from other decoding choices at a fixed call budget. We study a one-parameter family of position-selection rules spanning confidence-free selection, soft confidence-weighted order, and hard most-confident-first ranking. We separate position selection from the time grid (tokens committed per call) and token revision. On two 124M checkpoints and LLaDA-8B-Base, we compare rules at equal call budgets using perplexity under a fixed GPT-2-large scorer, bigram diversity, and MAUVE. At temperature 1.0 and 32 calls on the 124M model, soft order is the best fixed-count rule tested: it lowers scorer perplexity by about 12% relative to both the standard ancestral sampler and a confidence-free fixed-count control, whereas hard ranking roughly doubles it. On LLaDA-8B-Base at the same temperature, the reduction relative to the ancestral sampler is 11–23% at 16–64 calls, with all 95% intervals excluding zero. On the 124M model, the gain over the ancestral sampler persists among outputs matched on per-document bigram diversity. One-shot revision trades diversity for a further 31% reduction in perplexity, while time grids fitted to pilot-run confidence lower ancestral-sampler perplexity by about 4% at the smallest budgets. Cooling to $T = 0.7$ favors harder weighting. Decoding gains should be reported alongside call budget, count rule, and diversity.

1 INTRODUCTION

Masked diffusion language models (Austin et al., 2021; Lou et al., 2023; Sahoo et al., 2024; Nie et al., 2025) generate text by starting from a fully masked sequence and repeatedly filling in masked positions. Each forward pass of the network, which we call a *model call*, predicts a token distribution for every masked position at once. A decoder can therefore *commit* (fix) several tokens per call and finish a sequence in far fewer calls than it has tokens, motivating their use for faster generation. This parallelism forces a decision at every call: which available predictions to commit now. Committed tokens become context for all later predictions (Lee et al., 2018), so the choice shapes the rest of the trajectory. A common answer is to commit the positions where the model is most confident (Ghazvininejad et al., 2019; Chang et al., 2022; Nie et al., 2025).

Confidence-based orders are well established in iterative and diffusion decoding (Ghazvininejad et al., 2019; Chang et al., 2022; Nie et al., 2025; Ben-Hamu et al., 2025), but their benefit is usually measured together with other decoding changes. An apparent improvement may reflect the number of tokens committed per call, additional forward passes spent on revision, or a narrower, less diverse output distribution rather than the position order itself. The open question is quantitative: how much does confidence weighting itself help when many tokens are committed per call, and which gains survive changes in model, call budget, and quality measure?

We treat position selection as a one-parameter family (Figure 1a). At each call, positions are drawn without replacement with probability proportional to a power β of the model's top-token probability (its *confidence*), a Plackett–Luce construction (Plackett, 1975; Luce, 1959). Setting $\beta = 0$ ignores confidence, $\beta = 1$ gives the soft order studied here, and large β recovers hard most-confident-first ranking. Throughout this comparison, token values are sampled from the model; only the position-selection rule changes.

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.

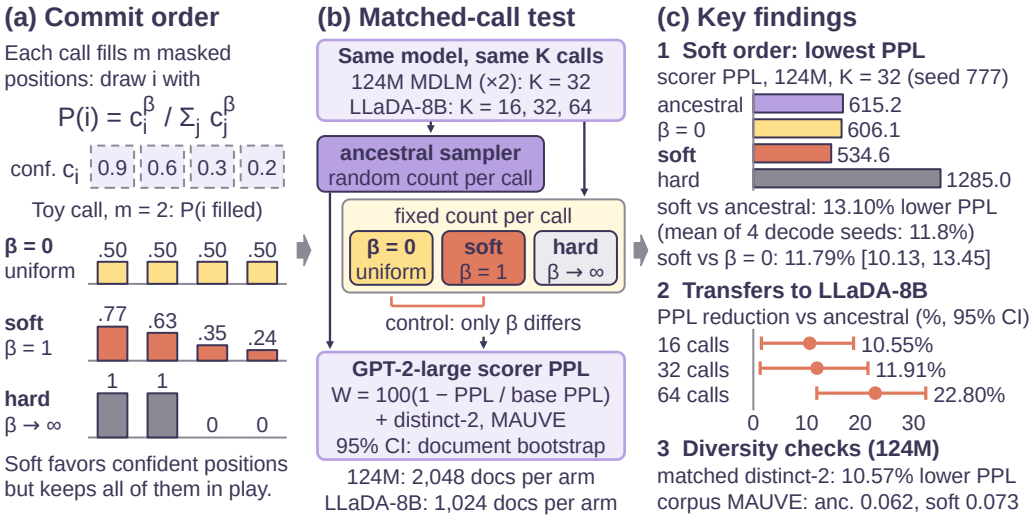

Figure 1: **Soft confidence order at a matched call budget.** (a) Each model call fills m masked positions, drawn without replacement with probability $c_i^\beta / \sum_j c_j^\beta$, where c_i is the model's top-token probability ($\beta = 0$: uniform weights; $\beta = 1$: soft order; $\beta \to \infty$: hard ranking). Bars: exact probability that each of four toy positions is filled when $m = 2$. (b) All arms use the same model weights and K calls. Ancestral decoding commits a random number of tokens per call; the other three rules share one fixed count schedule, so $\beta = 0$ versus $\beta = 1$ isolates confidence weighting. Outputs are scored by GPT-2-large perplexity (PPL; reduction W, Eq. 2), distinct-2, and MAUVE. (c) Top: 124M scorer perplexity at 32 calls and decode seed 777, where soft order is 13.10% below ancestral decoding and 11.79% below $\beta = 0$ (11.8% below ancestral decoding averaged over four seeds). Middle: LLaDA-8B-Base reductions at $T = 1.0$ with 95% document-bootstrap confidence intervals (CIs), all excluding zero. Bottom: the 124M gain over ancestral decoding among outputs matched on per-document distinct-2, and the corpus MAUVE of ancestral decoding and soft order. Terracotta marks soft order, lavender ancestral decoding, lemon $\beta = 0$, and grey hard ranking.

The key design choice is to compare every rule at the same number of model calls and against two confidence-free baselines (Figure 1b). Matching calls matters because the budget alone affects scorer perplexity: the standard sampler reaches perplexity 548.09 at 256 calls versus 615.20 at 32 calls (Table 1). *Ancestral decoding*, the standard masked-diffusion sampler, unmasks each masked position independently with a schedule-determined probability, so the number of tokens committed per call is random. A fixed-count $\beta = 0$ control shares soft order's deterministic count schedule and thus isolates the effect of confidence. We vary two other decisions separately: the *time grid*, which sets how many tokens are committed at each call, and *revision*, which re-masks and re-predicts committed tokens, with every added call counted in the budget. We evaluate two 124M masked diffusion checkpoints trained on OpenWebText and the public LLaDA-8B-Base model (Nie et al., 2025). We measure quality by the perplexity that a fixed GPT-2-large model (Radford et al., 2019) assigns to the generated text (*scorer perplexity*; lower is better). We measure diversity with *distinct-2* (the fraction of unique bigrams) and similarity to human text with MAUVE (Pillutla et al., 2021).

Figure 1c summarizes the main findings. At sampling temperature $T = 1.0$ and 32 calls, an intermediate strength of confidence weighting performs best on the 124M model: soft order lowers scorer perplexity by about 12% relative to both confidence-free baselines, whereas hard ranking roughly doubles it, and the gain over ancestral decoding persists among outputs matched on per-document diversity. This gain carries over to LLaDA-8B-Base at 16, 32, and 64 calls (11–23%). Cooling to $T = 0.7$ favors harder weighting. On the 124M model, revision further lowers perplexity relative to soft order and shifts the operating point along a quality–diversity trade-off (Figure 3b); interleaved re-masking also lowers perplexity at the same call budget, and time grids calibrated on pilot-run confidence profiles (*profiled grids*) help the ancestral sampler most at the smallest budgets. Together, these results characterize the operating points available through commit order, time grid, and revision.

Contributions.

- **A matched-call protocol.** We compare commit orders at equal call budgets on two 124M checkpoints and LLaDA-8B-Base, and evaluate time grids and revision at equal or explicitly counted call budgets, using a fixed-count $\beta = 0$ control to separate confidence weighting from the count rule (§2).
- **An intermediate optimum for confidence weighting.** At 32 calls on the 124M model, soft order lowers scorer perplexity by 11.8% versus ancestral decoding (mean over four decode seeds) and by 11.79% [10.13, 13.45] versus the fixed-count $\beta = 0$ control (seed 777), while hard ranking roughly doubles it (§3, Tables 1–2, Figure 2).
- **Transfer and diversity checks.** On LLaDA-8B-Base at $T = 1.0$ the reduction is 11–23% at 16–64 calls, with every 95% interval excluding zero (Table 3); on the 124M model the gain over ancestral decoding persists (10.57%) among outputs matched on per-document distinct-2, and corpus MAUVE rises from 0.062 to 0.073.
- **Further gains from revision and time grids.** One-shot revision trades corpus distinct-2 for a further 30.75% lower 124M perplexity at exactly 33 calls (Figure 3); confidence-targeted re-masking lowers soft order's perplexity by a further 16.08% at the same 32 calls (§4), and a profiled time grid lowers ancestral-decoding perplexity by 4.26% at 4 calls (§5).

2 Decoding choices and measurements

Models and statistical units. The two 124M checkpoints are trained on OpenWebText with the MDLM (masked diffusion language model) objective (Sahoo et al., 2024): a plain model (our *reference model*) and a variant trained with a different timestep-weighting schedule, both held fixed throughout the decoding comparisons. Most conditions generate 2,048 sequences of length 256 at temperature $T = 1.0$. Public-model conditions use LLaDA-8B-Base (Nie et al., 2025), 1,024 sequences, and a full-sequence sampler. Decode seed 777 is shared by baseline comparisons, and the *reference operating point* is $T = 1.0$, 32 calls, and decode seed 777; the main 124M order comparison also uses seeds 778–780. Auxiliary tests vary the training recipe, WikiText training seeds (Merity et al., 2016), sequence length, and prompt conditioning. Uncertainty is estimated by document-level bootstrap at fixed weights.

Position order and count rules. A decoder with budget K makes calls $k = 1, \dots, K$. Let M_k be the masked positions remaining before call k and $c_i = \max_v p_\theta(v \mid x_k)_i$ the confidence of position i after temperature scaling. The fixed-count rules commit $m_k = \lceil |M_k|/(K - k + 1) \rceil$ positions per call; at length 256 and 32 calls this is eight positions per call. Soft order selects these positions without replacement, using the sequential choice probabilities

$$\Pr(i \text{ selected next} \mid A) = \frac{c_i^\beta}{\sum_{j \in M_k \setminus A} c_j^\beta}, \qquad i \in M_k \setminus A, \tag{1}$$

where A contains positions already selected in that call. This is the standard Plackett–Luce construction (Plackett, 1975; Yellott, 1977; Luce, 1959), and it is equivalent to Gumbel-top-k sampling on $\beta \log c_i$. $\beta = 0$ gives uniform selection at fixed counts; $\beta = 1$ gives soft order. Hard ranking ($\beta \to \infty$) uses the same count schedule but commits the m_k positions with the largest c_i. In every rule the committed token values are sampled from the model's predictive distribution, never taken as the argmax. The ancestral baseline instead unmasks each remaining position independently with probability $(t_{k-1} - t_k)/t_{k-1}$ on a time grid $1 = t_0 > t_1 > \cdots > t_K = 0$ (by default the uniform grid $t_k = 1 - k/K$), so its realized counts are random. Thus the soft-versus-ancestral comparison matches model calls, and the direct $\beta = 1$ versus $\beta = 0$ comparison additionally holds the deterministic count rule fixed.

Quality, diversity, and cost. A fixed GPT-2-large scorer (Radford et al., 2019) evaluates each generated document. We report the signed relative perplexity reduction

$$W = 100\big(1 - \exp(\bar{\ell}_{\text{method}} - \bar{\ell}_{\text{base}})\big) = 100\big(1 - \text{PPL}_{\text{method}}/\text{PPL}_{\text{base}}\big), \tag{2}$$

where $\bar{\ell}$ is the mean per-token scorer negative log-likelihood (NLL) and $\text{PPL} = \exp(\bar{\ell})$ is the corpus-level perplexity reported in our tables. Positive W means lower scorer perplexity than the baseline

Table 1: **Operating points on the reference 124M model** ($T = 1.0$, length 256, decode seed 777, 2,048 documents per row). Scorer PPL is GPT-2-large perplexity; distinct-2 measures corpus bigram diversity. Ancestral decoding uses random per-call counts; the other position-selection rules share a fixed count schedule. Revision rows build on soft order ($\beta = 1$). Bold marks the lowest scorer PPL among the 32-call position-selection rules. Entries are point estimates; repetition and MAUVE results are in Appendix Y, Table 10.

Decoder	Calls	Scorer PPL ↓	Distinct-2 ↑
Position selection			
Ancestral	32	615.20	0.557
Fixed count, $\beta = 0$	32	606.10	0.557
Confidence-weighted, $\beta = 0.5$	32	557.93	0.545
Soft order, $\beta = 1$	32	**534.62**	0.540
Confidence-weighted, $\beta = 2$	32	556.22	0.539
Hard ranking ($\beta \to \infty$)	32	1285.04	0.602
Ancestral, profiled time grid	32	608.70	0.557
Budget references			
Soft order, $\beta = 1$	33	524.22	0.536
Ancestral	64	575.27	0.554
Soft order, $\beta = 1$	64	487.63	0.535
Ancestral	256	548.09	0.552
Revision on the soft-order sampler			
Interleaved re-masking, random	32	483.20	0.522
Interleaved re-masking, confidence	32	448.67	0.511
One-shot revision	33	363.00	0.468
Guarded revision, two rounds	32	279.28	0.426
Guarded revision, three rounds	33	250.49	0.406
Guarded revision, four rounds	36	242.44	0.397
Diversity-calibrated baseline (§4)			
Soft order, nucleus $p = 0.951$	33	294.14	0.465

and $W = -100\%$ means doubled perplexity; W is a corpus-level ratio rather than a per-document win rate. All percentages in the main text use this convention. Confidence intervals (CIs) use 20,000 document bootstrap resamples, paired by sample index under the recorded random-number convention. Diversity uses corpus distinct-n (unique n-grams over all n-grams in the corpus) and the within-document repeated 4-gram rate (rep-4), with a reference repetition threshold of 2.5×10^{-3}; matching experiments instead use each document's own distinct-2. MAUVE (Pillutla et al., 2021) compares generated and reference text distributions. These measures are complementary. Every forward pass, including those used to select tokens for revision, counts toward the call budget; pilot trajectories used to calibrate a time grid are accounted separately.

3 How softly should confidence guide commitment?

We measure the benefit of confidence weighting at matched call budgets and examine its sensitivity to seeds, models, budgets, temperatures, and diversity levels.

Soft weighting helps at the reference operating point. At 32 calls, the reference 124M model has scorer perplexity 615.20 with ancestral decoding, 534.62 with $\beta = 1$, and 1285.04 with hard ranking (Table 1). The four decode-seed reductions for soft order are $13.10/11.04/11.28/11.71\%$, averaging 11.8%. Hard ranking roughly doubles scorer perplexity (1285.04 versus 615.20; $W = -108.9\%$ under Eq. 2). At 32 calls, $\beta = 1$ performs best among the measured strengths: it lowers perplexity relative to both $\beta = 0.5$ (557.93; $W = 4.18\%$) and $\beta = 2$ (556.22; $W = 3.88\%$), with 95% intervals excluding zero (Appendix C).

A fixed-count control isolates confidence. Fixed-count $\beta = 0$ decoding reaches perplexity 606.10, and, with 2,048 documents per arm, soft order ($\beta = 1$) lowers it by 11.79%, with a 95% interval [10.13, 13.45]. Confidence weighting therefore helps under the fixed count rule, and the two

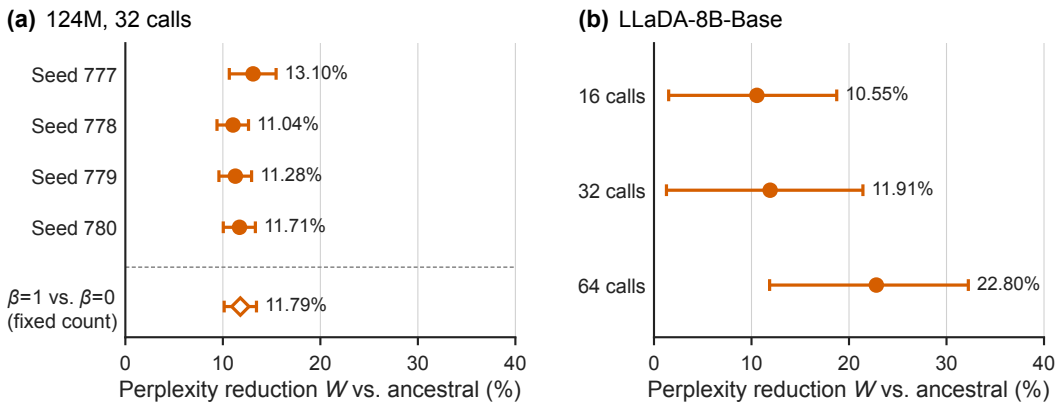

Figure 2: **Soft-order benefits across decode seeds and call budgets.** Signed scorer-perplexity reduction W of soft order ($\beta = 1$). (a) 124M model at 32 calls: four decode seeds (2,048 documents each) compared with ancestral decoding; the open diamond (below the dashed line) instead compares $\beta = 1$ with $\beta = 0$ under the same fixed-count rule. (b) LLaDA-8B-Base compared with ancestral decoding at 16, 32, and 64 calls (1,024 documents per budget). Bars are 95% document-bootstrap intervals at fixed model weights; both panels share one x-scale, and every interval excludes zero.

Table 2: **Robustness of the 124M soft-order reduction** relative to ancestral decoding at the same call budget ($T = 1.0$, 2,048 documents per arm). $W = 100(1 - \mathrm{PPL}_{\mathrm{soft}}/\mathrm{PPL}_{\mathrm{anc}})$ is the signed scorer-perplexity reduction (Eq. 2); brackets give paired 95% document-bootstrap intervals, all of which exclude zero.

Variation	Calls	Soft PPL	Ancestral PPL	Reduction W (%) [95% CI]
Reference (decode seed 777)	32	534.62	615.20	13.10 [10.66, 15.46]
Decode seed 778	32	539.71	606.70	11.04 [9.40, 12.63]
Decode seed 779	32	530.11	597.52	11.28 [9.60, 12.95]
Decode seed 780	32	531.06	601.48	11.71 [10.04, 13.33]
Timestep-weighted checkpoint	32	488.20	567.69	14.00 [11.67, 16.28]
WikiText, training seed 0	32	549.69	627.15	12.35 [10.31, 14.34]
WikiText, training seed 1	32	548.79	619.73	11.45 [9.96, 12.92]
WikiText, training seed 2	32	614.84	700.01	12.17 [10.66, 13.68]
Sequence length 512	32	355.93	399.82	10.98 [9.77, 12.18]
64-token prompt, suffix scored	32	562.17	644.99	12.84 [10.98, 14.65]
Larger budget	64	487.63	575.27	15.24 [12.73, 17.69]
Larger budget	256	460.03	548.09	16.07 [13.30, 18.80]

confidence-free baselines differ by far less than the soft-order gain (95% interval of W for ancestral relative to $\beta = 0$: $[-4.36, +1.30]$).

The reduction is stable across seeds, checkpoints, lengths, and budgets. Table 2 collects the soft-versus-ancestral comparison under variations of the reference setting, summarized from Appendices S and V. Every listed reduction lies between 10.98% and 16.07% with an interval excluding zero. The gain persists as the budget grows: soft order at 64 calls (487.63) and even at 32 calls (534.62) has lower perplexity than ancestral decoding at 256 calls (548.09). Rescoring the 124M comparisons with GPT-2-XL and Qwen2.5-1.5B preserves the sign of all nine tested rankings in their point estimates; the median absolute relative change in W across the 18 re-scored comparisons is 2.4% (Table 4).

The soft-order gain transfers to LLaDA. LLaDA soft order reduces perplexity by 10.55%, 11.91%, and 22.80% at 16, 32, and 64 calls, with intervals [1.50, 18.74], [1.28, 21.43], and [11.86, 32.24] (Table 3, Figure 2). All three intervals exclude zero, and two additional 32-call decode seeds retain the direction. The soft-order gain over the ancestral sampler thus carries over to a public 8B model. Rescoring with GPT-2-XL preserves the sign of all five LLaDA comparisons (32 calls: 11.92% [1.42, 21.33]; Table 4).

Table 3: **Confidence-weighted orders on LLaDA-8B-Base** (1,024 documents per arm, full-sequence sampler, length 256). W is the signed scorer-perplexity reduction of the arm relative to ancestral decoding at the same temperature, with a paired 95% document-bootstrap interval; a dagger (†) marks intervals that include zero. Calls are equal for both arms except in the revision rows, where one-shot and four-round guarded revision (33 and 36 calls) are compared with ancestral decoding at 32 calls (arm/baseline); these rows are discussed in Appendix B and Figure 5.

Calls	T	Ancestral PPL	Arm	Arm PPL	W (%) [95% CI]
16	1.0	180.54	$\beta = 1$	161.49	10.55 [1.50, 18.74]
32	1.0	97.64	$\beta = 1$	86.01	11.91 [1.28, 21.43]
64	1.0	70.66	$\beta = 0.5$	57.24	18.99 [7.77, 29.04]
			$\beta = 1$	54.54	22.80 [11.86, 32.24]
			$\beta = 2$	55.71	21.15 [9.58, 31.08]
32	0.7	32.08	$\beta = 0.5$	34.43	-7.32 [-37.57, 16.79]†
			$\beta = 1$	30.71	4.26 [-22.03, 25.25]†
			$\beta = 2$	27.35	14.75 [-9.31, 33.78]†
64	0.7	28.51	$\beta = 0.5$	22.37	21.53 [-2.48, 40.34]†
			$\beta = 1$	22.09	22.52 [-2.49, 41.12]†
			$\beta = 2$	21.09	26.01 [2.00, 43.91]
33/32	1.0	97.64	One-shot revision	167.63	-71.68 [-91.04, -54.38]
36/32	1.0	97.64	Guarded, four rounds	162.03	-65.94 [-85.24, -48.51]

Temperature changes the preferred weighting. On the 124M model at 32 calls, $\beta = 1$ beats the measured 0.5 and 2 choices at $T = 1.0$. At $T = 0.7$ the preference moves toward harder weighting: $\beta = 2$ improves on $\beta = 1$ by 11.23% [9.81, 12.63], and $\beta = 1$ improves on $\beta = 0.5$ by 7.31% [6.11, 8.48] (perplexities 26.25, 29.58, and 31.91, against 35.82 for ancestral decoding). On LLaDA at $T = 0.7$ and 32 calls, the same ordering among β values holds, with intervals excluding zero ($\beta = 2$ over $\beta = 1$ by 10.96%, $\beta = 1$ over $\beta = 0.5$ by 10.79%), and at 64 cooled calls $\beta = 2$ improves on ancestral decoding by 26.01% [2.00, 43.91] (Table 3). Orders are compared within a temperature; full estimates are in Appendix D.

The gain persists among diversity-matched outputs. Confidence weighting moves the quality–diversity operating point: at the reference point, corpus distinct-2 moves from 0.5568 to 0.5401 and decreases monotonically with β (0.5453, 0.5401, and 0.5388 for $\beta = 0.5, 1, 2$), while MAUVE increases from 0.0621 to 0.0730 (Table 10). Retaining soft-order/ancestral pairs whose per-document distinct-2 differs by at most 0.05 leaves 1,749/2,048 pairs and a 10.57% scorer reduction [8.08, 13.01]. Tighter and wider windows also retain positive effects: 9.28% [6.00, 12.41] with 980 pairs at a 0.02 window and 12.31% [9.93, 14.62] with 1,986 pairs at 0.08, so the soft-order gain holds at all three matching windows.

4 Revision on top of soft order

Revision re-masks and re-predicts committed tokens; this section measures how far it extends the soft-order operating point on the 124M model. All arms below use the same fixed checkpoint and differ only in decoding. One-shot revision first generates a 16-call scaffold, uses one forward pass to score its tokens, then re-masks the lowest-confidence quarter and re-decodes in 16 calls. Its total cost is $16 + 1 + 16 = 33$ calls, following the iterative-refinement approach (Ghazvininejad et al., 2019). On the 124M model it reduces perplexity by 32.10% relative to soft order at 32 calls. Against an exact 33-call soft-order baseline (524.22), the reduction is 30.75% [28.95, 32.51] (Figure 3a). The gain depends on confidence-based selection of the revised quarter: choosing it at random instead gives a signed reduction of -18.09% [-21.24, -14.96] relative to soft order at 32 calls.

Interleaved re-masking adds gains at the same budget. A ReMDM-style variant (Wang et al., 2025) re-masks committed tokens stochastically inside the soft-order trajectory, with a constant per-token probability of 0.04 on a middle band of the schedule and no extra calls. At exactly 32 calls

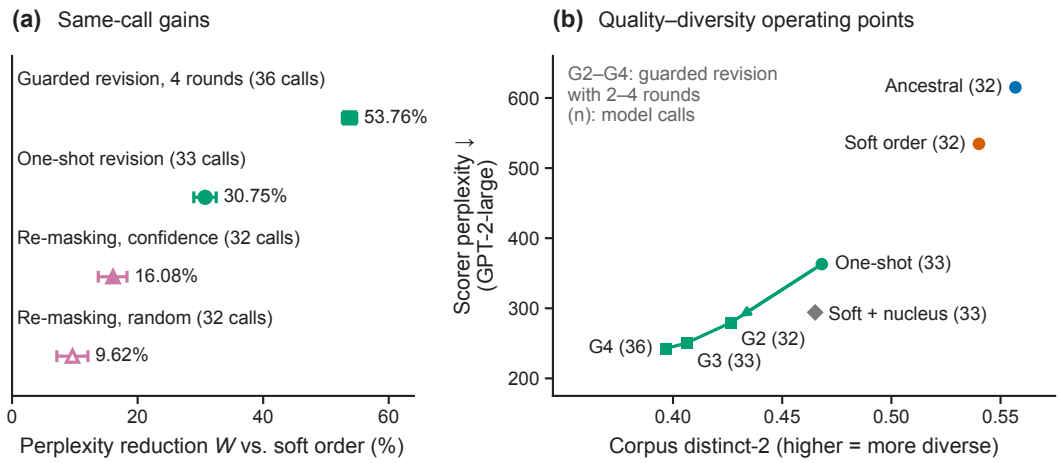

Figure 3: **Revision extends the soft-order operating point on the 124M model.** (a) Perplexity reduction relative to soft order at the same number of calls, with 95% document-bootstrap intervals: guarded four-round revision at 36 calls (square), one-shot revision at 33 calls (circle), and interleaved re-masking (triangles; open: random) at 32 calls. (b) Scorer perplexity against corpus distinct-2 for the reference operating points of Table 1 (point estimates). G2–G4 (squares) denote guarded revision with two to four rounds, the solid line traces the ladder from one-shot revision through G2–G4, the grey diamond is soft order with diversity-calibrated nucleus sampling, and numbers in parentheses are model calls. Lower perplexity accompanies lower corpus distinct-2.

it lowers perplexity relative to soft order alone, to 483.20 with random and 448.67 with confidence-targeted re-masking (reductions of 9.62% [7.13, 12.09] and 16.08% [13.73, 18.30]; Figure 3a). Soft order thus serves as a base sampler for further same-budget gains.

The gains hold at matched budgets, on a fresh seed, and under larger scorers. At exactly 36 calls, four-round guarded revision (242.44) improves on 36-call soft order (524.34) by 53.76% [52.61, 54.89] (Figure 3a), and one-shot revision with 32-call halves (65 calls, 308.61) improves on 64-call soft order (487.63) by 36.71% [35.20, 38.16]. On a fresh decode seed (780), one-shot revision improves on 32-call soft order by 33.66% [32.23, 35.06]. Two larger scorers preserve the direction of every comparison at the point-estimate level, and 17 of the 18 re-scored 124M intervals exclude zero (Table 4).

Deeper revision trades diversity for scorer gains. Additional guarded rounds reduce perplexity from 279.28 to 250.49 and 242.44, while distinct-2 falls from 0.4263 to 0.4062 and 0.3967 (Table 1). The guards apply a no-repeat-n-gram constraint to revision commits and add no model calls. The two-round variant uses a four-gram repetition guard; the later variants use three-gram guards. All three guarded variants keep the rep-4 point estimate at or below the 2.5×10^{-3} reference threshold (2.50, 2.43, and 2.46×10^{-3}). MAUVE rises from 0.0730 for soft order to 0.1251, 0.1546, and 0.1103 along the guarded ladder (Table 10).

Soft order with nucleus sampling is a strong diversity-matched baseline. A separate 33-call comparison calibrates nucleus sampling in the soft-order baseline to match one-shot revision's corpus distinct-2 (0.4651 versus 0.4681). The calibrated soft-order baseline reaches perplexity 294.14 and MAUVE 0.1522 at 33 calls, versus 363.00 and 0.0769 for one-shot revision (Table 10) (Appendix B).

5 Time grids and early-confidence signals

We examine profiled time grids and the relationship between early confidence and final quality.

Profiled grids redistribute calls. On a pilot trace set, define $w_k = (1 - \bar{c}_k)\bar{f}_k / L$, where $\bar{c}_k$ is the mean confidence and $\bar{f}_k$ the mean number of tokens committed at step k, and L is the sequence

Table 4: **Rankings under larger scorers.** Signed reduction W (%, Eq. 2) of the first arm relative to the second on the same documents, under the reference GPT-2-large scorer (point estimates from Tables 1 and 3) and two larger scorers with paired 95% document-bootstrap intervals (2,048 documents per arm for 124M, 1,024 for LLaDA). Calls: first/second arm; dashes: not re-scored. A dagger (†) marks the one interval that includes zero.

		Signed reduction W (%) by scorer		
		GPT-2-large	GPT-2-XL	Qwen2.5-1.5B
Comparison	Calls	(ref.)	W 95% CI	W 95% CI
124M reference model, $T = 1.0$				
Soft order vs. ancestral	32/32	13.10	13.00 [10.56, 15.36]	10.86 [8.47, 13.20]
One-shot revision vs. soft order	33/32	32.10	31.67 [30.20, 33.14]	29.02 [27.48, 30.52]
Confidence re-masking vs. soft order	32/32	16.08	16.00 [13.69, 18.25]	11.76 [9.42, 14.05]
One-shot revision vs. confidence re-masking	33/32	19.09	18.66 [16.58, 20.68]	19.56 [17.50, 21.61]
Confidence vs. random re-masking	32/32	7.15	7.08 [6.12, 8.02]	6.10 [5.08, 7.15]
Guarded two rounds vs. one-shot revision	32/33	23.06	22.83 [21.12, 24.48]	21.81 [19.95, 23.63]
Guarded three vs. two rounds	33/32	10.31	10.36 [9.50, 11.22]	9.22 [8.23, 10.21]
Guarded four vs. three rounds	36/33	3.21	3.12 [1.04, 5.17]	1.83 [−0.48, 4.08]†
Soft order vs. random-quarter revision	32/33	15.32	15.43 [13.13, 17.60]	15.51 [13.31, 17.68]
LLaDA-8B-Base				
Soft order vs. ancestral, $T = 1.0$	16/16	10.55	10.05 [1.32, 18.16]	–
Soft order vs. ancestral, $T = 1.0$	32/32	11.91	11.92 [1.42, 21.33]	–
Soft order vs. ancestral, $T = 1.0$	64/64	22.80	23.34 [12.82, 32.75]	–
$\beta = 2$ vs. $\beta = 1$, $T = 0.7$	32/32	10.96	11.01 [2.32, 18.99]	–
$\beta = 1$ vs. $\beta = 0.5$, $T = 0.7$	32/32	10.79	12.50 [5.28, 19.46]	–

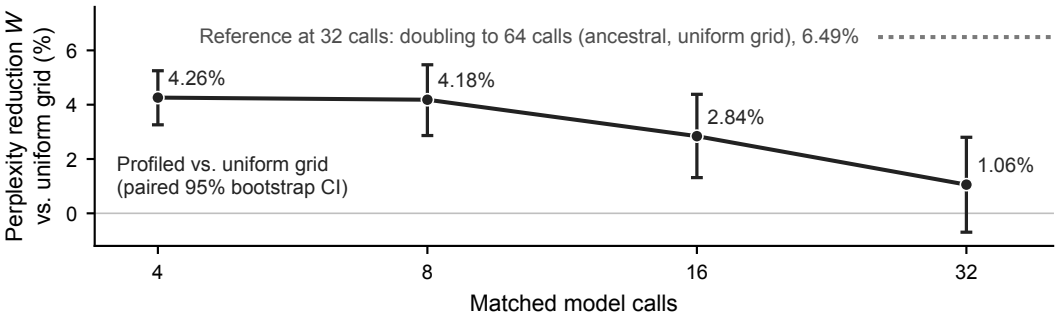

Figure 4: **Profiled-grid gains across call budgets.** Scorer-perplexity reduction W of the profiled relative to the uniform time grid, both with ancestral decoding at matched model calls (reference 124M model, length 256, $T = 1.0$); bars are paired 95% document-bootstrap intervals. The gain is largest at the smallest budgets (4.26% at 4 calls), and its intervals exclude zero at 4, 8, and 16 calls. The dotted segment (a different baseline) is the 6.49% gain from doubling ancestral decoding from 32 to 64 calls on the uniform grid (point estimate). Pilot calibration calls are separate from these per-decode budgets.

length; w_k is the mean confidence deficit times the mean fraction of positions committed at step k. The profiled grid divides the cumulative weight into equal-mass intervals. Under a strictly convex per-interval proxy, equal masses minimize the sum of proxy errors (Appendix H). At matched calls with ancestral decoding, the profiled grid lowers perplexity by 4.26%, 4.18%, 2.84%, and 1.06% at 4, 8, 16, and 32 calls (Figure 4). The gains are largest at the smallest budgets, where each call commits the most tokens (64 on average at 4 calls).

Early confidence carries a quality signal. Mean confidence at step 4 of 32 has Pearson correlation -0.320 and -0.230 with final per-document scorer NLL (two checkpoints, 512 traces each; Appendix A); Appendices G and I bound how many calls a stopping rule based on this signal could save.

6 Related work

Masked generation and ordering. Discrete and continuous-time diffusion formulations (Austin et al., 2021; Campbell et al., 2022; Lou et al., 2023; Shi et al., 2024; Sahoo et al., 2024) provide the modeling foundation, while LLaDA, Dream, and block diffusion explore larger or different generation regimes (Nie et al., 2025; Ye et al., 2025; Arriola et al., 2025). Iterative refinement, Mask-Predict, insertion/deletion models, and MaskGIT already make generation order and revision central design choices (Lee et al., 2018; Ghazvininejad et al., 2019; Gu et al., 2019; Chang et al., 2022). Building on these confidence-based orders and on Plackett–Luce sampling, our contribution is the comparison of soft and hard confidence weighting under explicit count, budget, and diversity controls.

Planning, parallelism, and revision. DDPD learns a planner to select denoising positions (Liu et al., 2024); EB-Sampler selects positions and batch size with an entropy-based error proxy (Ben-Hamu et al., 2025). Fast-dLLM and WINO change caching or execution (Wu et al., 2025; Hong et al., 2025); SureLock addresses write-once commitment (Oba et al., 2026); UNCODE analyzes positional and trivial-token biases (Huang et al., 2025a). ReMDM, RemeDi, PRISM, and NAViRA pursue remasking or revisability (Wang et al., 2025; Huang et al., 2025b; Kim et al., 2025; Fomenko et al., 2026). Nonuniform grids and optimized or adaptive schedules are related budget choices (Karras et al., 2022; Mohamed et al., 2025; Simonds, 2025); our grid builds on a confidence-deficit proxy.

Confidence and generation quality. Confidence and calibrated correctness differ (Guo et al., 2017). Autoregressive degeneration, typical sampling, desmoothing, and unlikelihood training show how probability-based choices shape repetition and vocabulary (Holtzman et al., 2019; Meister et al., 2022; Hewitt et al., 2022; Welleck et al., 2019). MAUVE (Pillutla et al., 2021) adds a distributional comparison to scorer perplexity.

7 Conclusion

On the reference 124M model at $T = 1.0$ and 32 calls, soft order ($\beta = 1$) performs best among the fixed-count rules tested: it beats both confidence-free baselines and hard ranking, its gain over ancestral decoding persists among diversity-matched outputs, and that gain transfers to LLaDA at 16–64 calls. Cooling favors harder weighting, revision and same-budget re-masking extend the 124M gains, and profiled grids help ancestral decoding most at the smallest budgets. Decoding gains should therefore be reported alongside call budget, count rule, and diversity.

7.1 Limitations and future work

Appendix B details four directions for future work. *Models and samplers:* the 124M checkpoints are decoding testbeds rather than fluent generators. Transfer should be tested on more pretrained families and blockwise or semi-autoregressive samplers, with a fixed-count control on public weights. *Precision and diversity:* larger document samples would improve precision across budgets and temperatures; joint matching on multiple diversity measures would test gains beyond the observed operating diversity. The nucleus baseline changes token sampling to match corpus diversity, whereas the per-document comparison selects outputs after generation. *Revision across models:* same-weight revision helps the 124M model but not LLaDA, motivating model-specific targets and comparisons with trained remasking systems. *Evaluation and adaptive budgets:* human and task-based evaluation should complement scorer-based proxies; stopping rules could turn early-confidence signals into call savings.

AI Use Statement

During the project, vibe coding tools (Codex and Claude Code) were used to help develop the code for the methods and experiments. All such work was conducted under human supervision and underwent human review. During manuscript preparation, AI tools were used to polish and revise the text and figures.

Ethics Statement

This work trains and evaluates language models using publicly available corpora, benchmarks, and model checkpoints. It involves no human-subject study. Public text corpora and generated outputs may contain personal, biased, or sensitive content; the work therefore shares the ethical risks common to language-model research.

Reproducibility Statement

The supplementary archive provides code, configurations, launch and analysis scripts, and recorded results, with a README mapping tables and figures to their source scripts. Public-model experiments use available LLaDA weights. For the internal 124M models, the retained scores support reanalysis, but checkpoint weights are not redistributed. Experimental settings and the scope of the available materials are specified in the main text and Appendix F.

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. *arXiv:2503.09573*, 2025. URL https://arxiv.org/abs/2503.09573.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured Denoising Diffusion Models in Discrete State-Spaces. *arXiv:2107.03006*, 2021. URL https://arxiv.org/abs/2107.03006.

Heli Ben-Hamu, Itai Gat, Daniel Severo, Niklas Nolte, and Brian Karrer. Accelerated Sampling from Masked Diffusion Models via Entropy Bounded Unmasking. *arXiv:2505.24857*, 2025. URL https://arxiv.org/abs/2505.24857.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Tom Rainforth, George Deligiannidis, and Arnaud Doucet. A Continuous Time Framework for Discrete Denoising Models. *arXiv:2205.14987*, 2022. URL https://arxiv.org/abs/2205.14987.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. MaskGIT: Masked Generative Image Transformer. *arXiv:2202.04200*, 2022. URL https://arxiv.org/abs/2202.04200.

Andrey Fomenko, Maksim Kryzhanovskiy, Svetlana Glazyrina, and Roman Ischenko. NAVIRA: Decoupled Stochastic Remasking for Masked Diffusion Language Models. *arXiv:2606.06031*, 2026. URL https://arxiv.org/abs/2606.06031.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-Predict: Parallel Decoding of Conditional Masked Language Models. *arXiv:1904.09324*, 2019. URL https://arxiv.org/abs/1904.09324.

Jiatao Gu, Changhan Wang, and Jake Zhao. Levenshtein Transformer. *arXiv:1905.11006*, 2019. URL https://arxiv.org/abs/1905.11006.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. *arXiv:1706.04599*, 2017. URL https://arxiv.org/abs/1706.04599.

John Hewitt, Christopher D. Manning, and Percy Liang. Truncation Sampling as Language Model Desmoothing. *arXiv:2210.15191*, 2022. URL https://arxiv.org/abs/2210.15191.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The Curious Case of Neural Text Degeneration. *arXiv:1904.09751*, 2019. URL https://arxiv.org/abs/1904.09751.

Feng Hong, Geng Yu, Yushi Ye, Haicheng Huang, Huangjie Zheng, Ya Zhang, Yanfeng Wang, and Jiangchao Yao. Wide-In, Narrow-Out: Revokable Decoding for Efficient and Effective DLLMs. *arXiv:2507.18578*, 2025. URL https://arxiv.org/abs/2507.18578.

Pengcheng Huang, Tianming Liu, Zhenghao Liu, Yukun Yan, Shuo Wang, Tong Xiao, Zulong Chen, and Maosong Sun. Empirical Analysis of Decoding Biases in Masked Diffusion Models. *arXiv:2508.13021*, 2025a. URL https://arxiv.org/abs/2508.13021.

Zemin Huang, Yuhang Wang, Zhiyang Chen, and Guo-Jun Qi. Don't Settle Too Early: Self-Reflective Remasking for Diffusion Language Models. *arXiv:2509.23653*, 2025b. URL https://arxiv.org/abs/2509.23653.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the Design Space of Diffusion-Based Generative Models. *arXiv:2206.00364*, 2022. URL https://arxiv.org/abs/2206.00364.

Jaeyeon Kim, Seunggeun Kim, Taekyun Lee, David Z. Pan, Hyeji Kim, Sham Kakade, and Sitan Chen. Fine-Tuning Masked Diffusion for Provable Self-Correction. *arXiv:2510.01384*, 2025. URL https://arxiv.org/abs/2510.01384.

Jason Lee, Elman Mansimov, and Kyunghyun Cho. Deterministic Non-Autoregressive Neural Sequence Modeling by Iterative Refinement. *arXiv:1802.06901*, 2018. URL https://arxiv.org/abs/1802.06901.

Sulin Liu, Juno Nam, Andrew Campbell, Hannes Stärk, Yilun Xu, Tommi Jaakkola, and Rafael Gómez-Bombarelli. Think While You Generate: Discrete Diffusion with Planned Denoising. *arXiv:2410.06264*, 2024. URL https://arxiv.org/abs/2410.06264.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete Diffusion Modeling by Estimating the Ratios of the Data Distribution. *arXiv:2310.16834*, 2023. URL https://arxiv.org/abs/2310.16834.

R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, 1959.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. Locally Typical Sampling. *arXiv:2202.00666*, 2022. URL https://arxiv.org/abs/2202.00666.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer Sentinel Mixture Models. *arXiv:1609.07843*, 2016. URL https://arxiv.org/abs/1609.07843.

Amr Mohamed, Yang Zhang, Michalis Vazirgiannis, and Guokan Shang. Fast-Decoding Diffusion Language Models via Progress-Aware Confidence Schedules. *arXiv:2512.02892*, 2025. URL https://arxiv.org/abs/2512.02892.

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:2502.09992*, 2025. URL https://arxiv.org/abs/2502.09992.

Daisuke Oba, Danushka Bollegala, Masahiro Kaneko, and Naoaki Okazaki. Stopping Computation for Converged Tokens in Masked Diffusion-LM Decoding. *arXiv:2602.06412*, 2026. URL https://arxiv.org/abs/2602.06412.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers. *arXiv:2102.01454*, 2021. URL https://arxiv.org/abs/2102.01454.

R. L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society: Series C*, 24 (2):193–202, 1975. doi: 10.2307/2346567.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

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. *arXiv:2406.07524*, 2024. URL https://arxiv.org/abs/2406.07524.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K. Titsias. Simplified and Generalized Masked Diffusion for Discrete Data. *arXiv:2406.04329*, 2024. URL https://arxiv.org/abs/2406.04329.

Toby Simonds. Entropy Adaptive Decoding: Dynamic Model Switching for Efficient Inference. *arXiv:2502.06833*, 2025. URL https://arxiv.org/abs/2502.06833.

Guanghan Wang, Yair Schiff, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Remasking Discrete Diffusion Models with Inference-Time Scaling. *arXiv:2503.00307*, 2025. URL https://arxiv.org/abs/2503.00307.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural Text Generation with Unlikelihood Training. *arXiv:1908.04319*, 2019. URL https://arxiv.org/abs/1908.04319.

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. *arXiv:2505.22618*, 2025. URL https://arxiv.org/abs/2505.22618.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7B: Diffusion Large Language Models. *arXiv:2508.15487*, 2025. URL https://arxiv.org/abs/2508.15487.

John I. Yellott. The relationship between luce's choice axiom, thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 15(2): 109–144, 1977. doi: 10.1016/0022-2496(77)90026-8.

GUIDE TO THE SUPPLEMENTARY RESULTS

The supplement is organized around the questions in the main paper. Appendix A presents the trace statistics summarized in Section 5 and the cross-model revision figure; Appendix B gives the extended discussion of limitations. Appendix C collects the quantitative comparisons (Tables 5–7); Appendix D gives the public-model sensitivity results; and Appendix E maps experiment identifiers to their protocols. The measurement argument, grid construction, and conditional savings bounds are in Appendices G–I. Appendix K introduces the detailed experiments, and Appendix M defines the recorded trace quantities. These experiments separate confidence traces (Appendices N–O), time grids (Appendices P–R), and commit order or revision (Appendix S), followed by mechanism and temperature controls (Appendices T–V). Appendix X summarizes further operating points (Tables 8 and 9). DAX identifiers are experiment names, not an additional sequence of claims. Intervals condition on the evaluated checkpoints and sampling protocol; scorer gains, diversity changes, and model-call costs remain separate outcomes.

Notation used in the detailed appendices. *DAX-N* identifies one experiment. The file `EXPERIMENTS.md` in the supplementary archive lists the launch scripts, analysis scripts, and result files of each identifier; paths in typewriter font are relative to that archive. *Registered* (also *pre-registered* or *pre-specified*) means that a threshold, decision rule, outcome class, or expected direction was fixed before the corresponding decode or analysis was run; the thresholds and outcome classes are implemented in each experiment's verdict script, and "the registry" denotes this set of pre-specified rules. A *registration gap* is an outcome outside the pre-specified classes, which is reported as measured, and a pre-specified test is *consumed* once it has been run and evaluated. Outcome classes and verdicts are printed in small capitals or upper case (e.g., TILT-PERSISTS-AT-64, NULL). A *board* is the set of documents generated by one decoder configuration (a *cell*). *Uniform-K* is ancestral decoding with K calls on the uniform time grid, *grid-K* ancestral decoding on the profiled grid, *softord-K* or *softorder-K* soft order, and *gridsoft-K* soft order with the profiled grid's counts; the *tilt* is confidence-weighted selection ($\beta = 1$ unless stated). The *anchor* is the baseline board of a comparison, the *flagship* comparison is soft order versus ancestral decoding at 32 calls, and a family's *champion* is its best-scoring configuration. A *vantage* trace set is the pilot trace set from which a grid is built. *Gen-PPL* is scorer perplexity and *gen-NLL* the corresponding mean negative log-likelihood; a *win* is the reduction W of Eq. 2; NFE is the number of model calls (function evaluations), and a decoder's *bill* is its total number of calls. *d2* is distinct-2 and rep_4 the within-document repeated 4-gram rate. An *arbiter* is a quality measure used to rank decoders (the scorer, MAUVE, or the repetition bar). *OutCI* means that the 95% interval excludes zero (or the stated bar) and *straddle* that it includes it; a *bar* is a pre-specified threshold, and a *razor* is a pass/fail margin smaller than the sampling error. *Frozen* means fixed after its first computation and not re-tuned; a *substrate* is a model and data setting; an *open cell*, *named candidate*, or *named open form* is a configuration that we identify but did not measure. MDE is the minimum detectable effect at 80% power, SE a standard error, AR autoregressive, and EDM the noise schedule (Karras et al., 2022). *Track A* is prediction from early confidence traces and *Track B* the checkpoint-level profiled grid. The internal checkpoints are owt_mdlm_s0 (the reference model), owt_tge_s0 (the timestep-weighted model; TGE is its adaptive timestep-weighting objective), and wiki_mdlm_s0 (a WikiText model). *In-house* refers to models trained for this study, such as these checkpoints, as opposed to the public LLaDA-8B-Base. Reopen2x16 is one-shot revision ($16 + 1 + 16$ calls), and g3r3/g3r4 are the three- and four-round guarded revisions. A gain is *priced* when its cost on another axis (extra model calls, a diversity change, or a second arbiter) is measured alongside it, and that cost is its *price*. A sample's *altitude* is its overall confidence level across horizons (the per-sample offset a_i in assumption B1 of Lemma 1); the *shape factor* is the per-position difficulty deviation $\delta_{i,j}$ defined in the same assumption.

A TRACE STATISTICS AND CROSS-MODEL REVISION

This section presents the figure and trace statistics summarized in the main text. The paragraph below supports the early-prediction summary of Section 5, and Figure 5 shows the cross-model revision comparison discussed in Appendix B.

Trace statistics for early prediction (Section 5). Across 512 reference-model traces, step-4 versus step-28 mean confidence has correlation 0.303, while the early/late relative-spread ratio is 0.992.

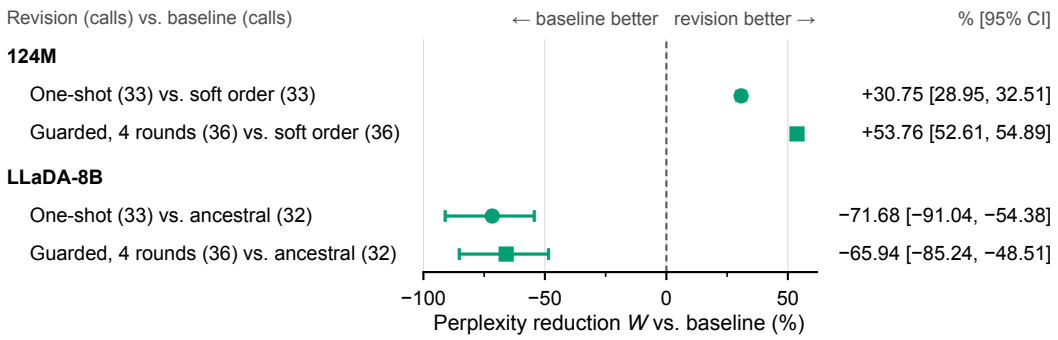

Figure 5: **Revision helps on the 124M model and hurts on LLaDA-8B.** Signed scorer-perplexity reduction of revision relative to the baseline named in each row (positive: revision has lower perplexity), with 95% document-bootstrap intervals; model calls are in parentheses. On 124M the baseline is soft order at the same number of calls; on LLaDA-8B it is ancestral decoding at 32 calls, so revision also spends one or four extra calls, and the extra calls do not prevent the degradation. The 124M intervals are narrower than the markers; all values are listed at right. The 124M quality–diversity operating points are in Figure 3b, and the LLaDA rows also appear in Table 3.

Mean confidence at step 4 has Pearson correlation -0.320 and -0.230 with final per-document scorer NLL on the two checkpoints (512 traces each). This is a measurable but limited predictive signal. The number of steps to 95% completion largely measures the prescribed unmask-count process, not readiness at equal quality. The unconditional step-0 canvas is identical across samples, so its statistics cannot allocate different budgets to those samples.

B Extended discussion of limitations

This section gives the detail behind the short limitations paragraph in Section 7.1, grouped by topic.

Absolute quality and models. The 124M generations are far from fluent: across our decoders scorer perplexities lie between roughly 230 and 1,480 and MAUVE between 0.06 and 0.21, whereas an autoregressive GPT-2-large sample scored with the same instruments reaches perplexity 27.3 and MAUVE 0.588 (a favorable reference, since the scorer is the generating model). The 124M results therefore concern relative differences between decoders at fixed weights; they do not show that any decoder produces high-quality text, and effect sizes may differ for stronger models. The two internal checkpoints are decoding testbeds, not evidence for a training method, and their weights are not redistributed. The LLaDA comparisons are closer to a practical regime but rest on one model, 1,024 documents per arm, and a full-sequence sampler; its blockwise or semi-autoregressive decoding was not evaluated.

Isolation and precision. No fixed-count $\beta = 0$ or hard-ranking arm was run on LLaDA, and its ancestral baseline has random counts, so the public result transfers the complete soft-policy comparison without isolating order from the count rule. The lower endpoints of the LLaDA intervals at 16 and 32 calls (1.50 and 1.28) lie close to zero, so the size of the public effect is poorly determined, and between-budget differences are unresolved (Figure 2 does not establish monotonic growth). On LLaDA at $T = 1.0$ and 64 calls, $\beta = 1$ is not resolved from $\beta = 0.5$ or $\beta = 2$ (Appendix D), so the interior optimum is established on the 124M model only. On the 124M model at $T = 1.0$ it is resolved at 16 and 32 calls; at 8 calls $\beta = 0.5$ and $\beta = 1$ are tied, and at 64 calls $\beta = 1.5$ improves on $\beta = 1$ by 3.50% [1.35, 5.65] while $\beta = 1$ and $\beta = 2$ are not resolved (DAX-27/27b, 107, 111, 112; Appendix S). The main-text statements that soft order is the best fixed-count rule therefore refer to 32 calls. The 124M $\beta = 0$ versus ancestral interval contains zero, an unresolved difference rather than proof that their output laws are identical. Hard ranking's failure concerns the parallel, sampled-token setting, where each call commits eight tokens with sampled values; it does not contradict confidence ranking in other decoding regimes. Document-level intervals condition on the retained checkpoints and do not estimate variation across pretraining families: the replications in Table 2 vary seeds,

recipes, and lengths within one model scale, and pairing by sample index is an analysis pairing, not a natural pairing of identical texts.

Temperature and diversity. On LLaDA at $T = 0.7$ and 32 calls, no soft-order rule has a statistically resolved advantage over ancestral decoding; at 64 calls only $\beta = 2$ does, and the three soft orders are not resolved from one another. These cooled outputs are bimodal: about 4–5% of documents per arm are catastrophic (mean generated-token scorer NLL above 10 nats), versus about 1% at $T = 1.0$. Cooling moves scorer perplexity far more than order does (on LLaDA at 32 calls, $T = 1.0 \to 0.7$ lowers it by 64.3% for soft order and 67.1% for ancestral decoding), so the order effect is not a substitute for choosing the temperature. Soft order lowers corpus distinct-2 (0.557 to 0.540). Diversity matching describes selected outputs, matches one statistic at a time, and does not identify a causal effect at equal diversity; joint matching of several diversity measures retains only 81, 41, 26, or 16 documents, too few for the intended precision. We therefore do not claim invariance to all diversity costs; the soft-order edge is established at its own operating diversity.

Revision. The revision arms are same-weight decoding variants, not the RemeDi or ReMDM systems, which change training or the sampler stack (Huang et al., 2025b; Wang et al., 2025), and their outcomes are not negative results for ReMDM, RemeDi, PRISM, or NAViRA. Soft order is the best of the tested rules that commit each token once, not the strongest 32-call decoder we measured (interleaved re-masking reaches 483.20 and 448.67), and the re-masking arm is one fixed setting, not a tuned ReMDM implementation. The two-round repetition estimate (2.499×10^{-3}) has sampling uncertainty crossing the threshold, so it is not a robust pass/fail distinction; the apparent three-round MAUVE maximum is not resolved by the doubled-sample comparison; and the nucleus-calibrated baseline of Section 4 beats one-shot revision (revision's signed reduction relative to it is -23.41% [-26.67, -20.23]), so more revision does not uniformly improve the quality–diversity comparison. On LLaDA, one-shot revision has signed reduction -71.68% [-91.04, -54.38] relative to uniform-32 despite an extra call, while raising distinct-2 from 0.350 (soft order) to 0.395; the four-round version (36 calls, -65.94%) recovers neither the uniform nor the soft-order baseline (Table 3, Figure 5). This reversal does not identify model size as the cause: architecture, training, tokenizer, and decoding behavior change together, and the recipe was not re-tuned for LLaDA.

Grids and early stopping. The grid argument is conditional: a linear proxy makes all grids tie, and it does not establish optimality for language quality. The 32-call grid interval includes zero, the profiled weights are nearly flat (maximum-to-minimum ratio 1.56), and 64 rather than 32 uniform calls gives 6.49% (Figure 4, dotted segment), so the grid is a modest low-budget adjustment whose pilot cost is additional to the decode calls. Early confidence is a real but limited prediction signal: none of the tested early-signal rules achieves 20% fewer calls at quality parity, and instantaneous thresholds ($\tau \in \{0.85, 0.90, 0.95\}$) degenerate to 249–252 of 256 calls. Entropy-bounded decoding (Ben-Hamu et al., 2025), a different adaptive mechanism, takes approximately 240 calls on the 124M model and 204 on LLaDA at our thresholds, so its better scorer result is not a 32-call comparison. These findings limit the evaluated statistics and thresholds, not all adaptive decoding; counterexamples to stronger trace-identifiability claims are in Appendix I, and our simple controls do not replicate complete planning or caching systems.

Evaluation and scope. Scorer perplexity, distinct-2, and MAUVE are proxies, not human judgments; we did not conduct human or task-based evaluation, and MAUVE is sensitive to its featurizer. Plackett–Luce position sampling is known, so the contribution is a controlled measurement, not a new sampling distribution; the tested failures are not a universal impossibility result, and we make no state-of-the-art claim (Appendix J.1).

C KEY EVIDENCE TABLES

Table 5 collects time-grid, order, revision, and transfer measurements; Tables 6 and 7 give the fixed-grid comparison and the call-budget ladder. They include unsuccessful comparisons and the evaluation conditions needed to interpret each result.

Table 5: Headline measurements. Generated-text perplexity is measured by a fixed scorer and is interpreted alongside diversity and call counts; CIs are paired-bootstrap (B=20000) unless noted. Across rows, win percentages use the anchor-relative form $100(1-\exp(\bar{n}_{\text{win}}-\bar{n}_{\text{lose}}))$; cells whose pre-specified analysis reported the inverse ratio are restated anchor-relative, with the inverse-ratio value in parentheses (DAX-26; DAX-29 is already anchor-relative).

question	design	measured result
Track-A early predictability	DAX-1 gates, n=512 traces	shape-uniform null: $0.179/0.181$ early/late spread; r=0.303
Track-A under conditioning	DAX-16, 64-token prompts, n=512	null replicates: persistence 0.246 $[0.163, 0.326]$; duration -0.009; prompt info visible, no exploitable channel
Fixed-grid comparison (NFE=32)	paired n=2048, 2 ckpts	$+1.06/+0.84\%$, CIs include 0
low-NFE placement series	paired n=2048, frozen grids	$+2.84/+4.18/+4.26\%$ at NFE=$16/8/4$; all CIs exclude 0
placement knee	DAX-11 4-point series	saturates at NFE≈8 (8→4: +0.08pt)
K=4 count-profile probe	DAX-98, mirror boards	deficit^2-weighted *counts* beat the equal-w profile at the K=4 node: $+3.35\%$ $[+2.33, +4.36]$ excluding 0; equal-w count schedules saturate at the knee while steeper profiles still pay; the mirror anchor read $+4.94\%$ $[+3.15, +6.73]$ contains the frozen $+4.26$ — STEEP-COUNT-RECOVERS-AT-K4
budget vs placement	DAX-3 ladder	NFE 32→64: $+6.49\%$
instantaneous trigger	DAX-10, $\tau \in \{0.85, 0.90, 0.95\}$	degenerates (249–252 of 256 calls)
τ-clearing availability	DAX-18, full $512\times32\times256$ $p_{\max}$ tensor	$29.7/24.7/18.3\%$ ever-clear at τ=$0.85/0.90/0.95$ — degeneration is temporal scatter, not scarcity (DAX-22: TAU-CHANNEL-OPEN)
cumulative trigger	DAX-13, $\beta \in \{2, 4, 8\}$	$+242/+100/+1.8\%$ over matched anchors
confidence order (sequential)	DAX-10b, matched calls	$548.09 \rightarrow 365.95$ (-33.2%, CI $[+30.4, +36.0]$)
confidence selection (parallel)	DAX-14, uniform counts	grid 608.70 vs hard order 1285.04 (the grid's gen-PPL is 52.63% lower than hard order's; inverse-ratio form)
soft confidence tilt (parallel)	DAX-15, uniform counts, β=1	534.62 at 32 calls: $+13.10\%$ $[+10.66, +15.46]$ vs uniform, $+12.17\%$ $[+9.68, +14.59]$ vs grid (distinct-2 -0.017)
tilt×placement cross + continuum	DAX-23, NFE=8 cross / β sweep @32	gridsoft 859.43: $+8.49\%$ vs uniform-8, all pairwise CIs exclude 0 (COMPOSITION-ADDS); β=1 interior max ($557.93/534.62/556.22$ at $0.5/1/2$, CIs exclude 0)
alternative placements	DAX-19, six K=8 grids vs base	blind grids perform substantially worse (-15 to -33%); deficit^2×flow 874.61 ($+6.87\%$) and ent×flow 885.98 beat base outside CI — ALT-BEATS-BASE
steep×tilt cross	DAX-24, def^2 counts + β=1	848.60 — ties equal-w cross (CI contains 0), beats tilt-free steep board outside CI — CROSS-SATURATES-PLACEMENT
placement transfer	DAX-12, wikitext ckpt	$+2.08\%$ $[+0.46, +3.70]$, null-family overlap

(cont.)

Table 5: Headline measurements (cont.).

question	design	measured result (cont.)
low-NFE transfer (recipe)	DAX-17, TGE ckpt, self-derived grid, NFE=8	$+5.38\%$ $[+4.07, +6.68]$ — replicates across objectives inside the $[2, 6]\%$ family band
order-layer transfer (recipe)	DAX-26, TGE ckpt, tilt+cross	softorder-32 $+14.00\%$ and gridsoft-8 $+8.99\%/+3.82\%$ over both anchors (anchor-relative form of the inverse-ratio values $+16.28\%/+9.88\%/+3.97\%$), CIs exclude 0 — order gains are not limited to one recipe
distributional check	DAX-25, official MAUVE, val ref	tilt numerically better (0.0730 vs 0.0621); compose inside the 0.02 band; positionally GAIN-SPREAD $(0.041/0.086)$
β plane across K	DAX-27/27b, K$\in \{8, 16\}$ both count families	interior β=1 at K$\geq$16 (outCI); $\{0.5, 1\}$ tie at K=8; β=2 excluded at every K
trigger under conditioning	DAX-28/28b, 64-token prompt	no savings (98.7% of suffix horizon; prefix arithmetic); order-quality $+28.2\%$ $[25.5, 30.7]$ at matched 192 calls — savings on neither substrate
order-layer transfer (conditioning)	DAX-29, prompt-clamped suffix, 192-token scoring	tilt $+12.84\%$ $[10.98, 14.65]$; compose $+7.60\%$ and $+6.00\%$ over both parents, CIs exclude 0 — order transfers to conditioned generation
bar precision audit	DAX-30, n=4096, NFE=8	$+3.09\%$ $[+2.16, +4.03]$ — CI upper below the 5% bar (the n=2048 touch was precision, not a firing)
tilt guardrail continuum	DAX-31, mixed-commit $\alpha \in \{0.25, 0.5, 0.75\}$	gain/distinct-2 co-monotone ($+6.16\%/\rightarrow 10.67\%$ vs $-0.0059/\rightarrow -0.0136$); no recovery point — one mechanism, one disclosed cost
steep resolution	DAX-32, def^2 crosses at K=16/32	tie the equal-w crosses (CIs contain 0), beat placement-only outCI — equal-w working point measured, not convenient
second arbiter (full set)	DAX-33, official MAUVE, 10 cells	6/7 pairs CONFIRMED incl. both flagship arms; the single PPL-ONLY pair is the known hard-order collapse
cross-training-seed replication	DAX-34, wiki ckpts s0/s1/s2	$9/9$ channel CIs exclude 0: tilt $+12.35/11.45/12.17$; compose $+9.11$ to $+10.19$ over uniform-8, $+6.36/+4.97/+7.00$ over grid-8; placement $+2.93/4.31/3.43$ — FULL-SEED-REPLICATES
decode-seed multiplicity	DAX-35, seeds $777/778/779$	$12/12$ CIs exclude 0, 0 reversals; tilt $13.10/11.04/11.28$ (fresh-seed agreement 0.24; 777 is the high anchor) — SEED-DISPERSION-WIDE
budget channel (tilt)	DAX-36, NFE=64 vs uniform-64	$+15.24\%$ $[12.73, 17.69]$ outCI; 487.63 undercuts uniform-256 — TILT-PERSISTS-AT-64 (win *grows* with budget)
trigger count shape	DAX-37, per-pass τ-clearing counts	pooled var/mean $14.5/12.8/10.6$ at τ=.85/.90/.95: pass-1 starvation + mid-horizon burst 1.5–3.2 — not Poisson

(cont.)

Table 5: Headline measurements (cont.).

question	design	measured result (cont.)
interior-hardness mechanism	DAX-38/38b/41/42	β^* direction twice measured ($1.45<1.66\leq 2.47$ real-K, CIs disjoint); constant-threshold calibration refuted; one-parameter surrogate-frame law fails real transfer
gated tilt	DAX-39, low-entropy-half gate	recovers $+11.15\%$ $[8.64, 13.54]$ (85% of flagship) at *same* distinct-2 cost (0.5377) — cost intrinsic to early confident commitment
tilt action band	DAX-40/43, commit-order dumps	mid-horizon selectivity hump, fraction-locked ($c_K/K \in [0.38, 0.50]$ at $K{=}8/16/32$); Spearman with steep weight 0.25 — PARTIAL-ALIGNMENT
fraction-band harvest	DAX-44, banded τ harvest, uniform-32 budget	-2.98% $[-4.19, -1.79]$ vs the tilt (still $+10.51\%$ over uniform); burst-cap shaping of the clearing bursts null at matched calls (DAX-67) — trigger channel triply closed ancestrally
commit revision opens	DAX-45/46, iso-33-forward re-open	re-decode the decode's own lowest-$p_{\max}$ quartile in full context: $+32.10\%$ $[30.66, 33.53]$ (363.00 vs softord-32 534.62), s778 replicate $+33.22\%$; uniform-random control -18.09% — CONFIDENCE-TARGETING-DOMINANT
revision dose and allocation	DAX-47/49/50/51, iso-33	unimodal dose $443.99\rightarrow 363.00\rightarrow 323.43\rightarrow 311.52\rightarrow 346.95$ at $12.5/25/37.5/50/75\%$ (peak 50%); scaffold front-load pays ($24\mathrm{A}{+}8\mathrm{C}$ 331.77; back-load collapses)

(cont.)

Table 5: Headline measurements (cont.).

question	design	measured result (cont.)
iterated revision	DAX-52/53/55, iso-budget	axes compose (champion 303.91, s778 303.14); *two* rounds at 32 forwards = 227.69 (+57.41% vs anchor; lowest board) — distinct-2 0.397 and, per the second-arbiter audit (DAX-56), an intra-sample repeat-rate signature on the iterated round disclosed; MAUVE passes for the whole ladder; a 4-gram guard on round two recovers the iterated gain inside the bar (279.28 at 32 forwards, passing all arbiters, DAX-59/60); a horizon-3 variant then keeps round *three* clean and paying (250.49 at 33 forwards, repetition below the baseline's background, passing all arbiters, DAX-62/63; round three under 4-gram re-crosses the bar, DAX-61): and the same guard keeps round *four* clean and paying at a decayed marginal (242.44 at 36 forwards, budgets $(6, 5, 3, 2)$, $+3.21\%$ $[1.13, 5.28]$ excluding 0, repetition still below the baseline's background, official-MAUVE 0.110, and the gain-concentration check passed, DAX-64/65): the clean frontier extends to four guarded revision rounds ($+54.65\%$ $[53.68, 55.60]$ over the anchor) — a trade-off extension, not a domination: the guarded rounds pay in NLL while the second and third arbiters decline along them (MAUVE $0.1546 \rightarrow 0.1103$, distinct-2 $0.406 \rightarrow 0.397$, disclosed verbatim at each depth) — saturation holding only in its weak form (4-gram, two rounds); the MAUVE decline ($0.1546 \rightarrow 0.1103$ along rounds $3 \rightarrow 4$) and the distinct-2 slide ($0.406 \rightarrow 0.397$) are disclosed with the extension; *Larger-scorer restatement (no new measurement)*: the gen-PPL vertices above are *declared-proxy* numbers — the arbiter-stable frontier is the per-arbiter vertex split (round-three 250.49 MAUVE-balanced / round-four 242.44 gen-PPL) quoted with the round-four inversion ($0.1546 \rightarrow 0.1103$) and the unguarded champion's $+18.5\% \rightarrow +0.2\%$ collapse beside every number; the family's recommendation is the guarded frontier under *both* arbiters, never the raw unguarded ladder; the horizon$\times$round block is measured (DAX-66: 4-gram dirty at rounds 3 *and* 4, 3-gram clean at both)

(cont.)

Table 5: Headline measurements (cont.).

question	design	measured result (cont.)
canonical ReMDM comparator	DAX-69, in-protocol cap/conf arms	interleaved stochastic re-masking beats the tilt at exactly matched calls (uniform 483.20, confidence 448.67 vs 534.62, both excluding 0) but ranks below one-shot targeted revision (363.00); placement dominates selection — REMDM-UNDERPERFORMS-REOPEN
tilt-vs-cooling surface	DAX-76, 4×2 lattice	within-T tilt wins all four outCI ($+17.44/+15.90/+13.75/+13.10\%$ at T=0.7/0.8/0.9/1.0, margins decaying in T); no matched-diversity cross-T pair forms — *tilt, don't cool* — TILT-DOMINATES-PARETO
τ ladder downward	DAX-70, τ=0.50/0.65/0.80	call-insensitive monotone quality (416.37@238 to 369.74@248; six-point knee at $\tau\approx0.80$); still 17% above the composed recipe at $\sim7\times$ its calls — TAU-DOWNWARD-PAYS
cooled revision	DAX-71, canonical reopen2x16 at T=0.7	survives cooling: 25.92 at 33 calls, $+12.37\%$ $[10.80, 13.95]$ over the cooled tilt board; margin compresses $2.6\times$ exactly as registered; repeat-prone cool tail (repeat4max 26) disclosed — REVISION-SURVIVES-COOLER
second-arbiter re-audit	DAX-72, MAUVE ladder	T=1.0 restated frozen; T=0.7 five fresh cells; three named inversions ($g3r3\to g3r4$ -0.0443; $u07\to s07b1$; $s07b1\to r2x16$); vertex split (g3r3 MAUVE-balanced / g3r4 gen-PPL); T=0.7 arbiters disagree on β^* (2 vs 0.5) — MAUVE-LADDER-REFUTES
noise-order identity	DAX-73, β=0 cell + exchangeability line	606.10 vs uniform-32 615.20, CI contains 0: i.i.d. confidence-independent keys induce the uniform commit order at fixed counts. The ancestral uniform comparator instead uses Bernoulli counts; the direct $\beta = 1$ versus $\beta = 0$ comparison is in Section 3 — NOISE-ORDER-EQUIVALENT
tilt vertex vs budget	DAX-74, softord-128/256	472.55/460.03; gap over uniform-256 $+16.07\%$ $[13.30, 18.80]$; distinct-2 flat — VERTEX-BEYOND-64, GAP-PERSISTS (the pre-specified expectation of a vertex at 64 calls is refuted)
altitude headroom	DAX-75, 512-sample nonlinear probe	Ridge oof-R^2 0.515 > gradient boosting 0.457 (difference outCI); settling-shape features zero-signal — LINEAR-IS-ALL; kNN-MI bootstrap artifact disclosed (non-class)

(cont.)

Table 5: Headline measurements (cont.).

question	design	measured result (cont.)
soft-target revision	DAX-77, $w \propto e^{-p_{\max}}$ fork + halves-cosine line	registration gap (UNCOVERED-QUADRANT): diversity restored to scaffold parity ($+0.0067$ outCI) at quality below the scaffold itself (-8.64% outCI vs softord-32); one targeting knob moves diversity and quality together; repeat signature is surface n-gram (halves-cosine falls $0.277 \to 0.248$ down the chain)
larger-scorer panel	DAX-78, 10 frozen boards, gpt2-xl + qwen2.5-1.5b	all nine named rankings sign-preserved under both (xl $9/9$ outCI; qwen $8/9$, round-4 marginal touching 0); median shrink 2.4%; guard-price probe: unguarded $+18.5\% \to +0.2\%$ under both — SHRINK-BUT-HOLD + registration-gap note
public-substrate larger-scorer panel	DAX-129, 12 frozen public boards, gpt2-xl over byte-identical detok bridges	all five registered keys preserve sign and class: tilt $+11.92\%$ $[+1.42, +21.33]$ (frozen $+11.91\%$ $[+1.28, +21.43]$), endpoints $+10.05\%$ (16) / $+23.34\%$ (64), interior flanks straddling, cooled harder-better $+11.01/12.50\%$ both outCI; boards re-land uniformly lower under the larger scorer (uniform-32 93.93 vs 97.64), effect magnitudes preserved — PANEL-PRESERVES-PUBLIC
same-bill ladder	DAX-79, three fresh cells	$+53.76\%$ $[52.61, 54.89]$ at 36c (vs softord-36), $+36.71\%$ $[35.20, 38.16]$ at 65vs64c (declared pairing favouring the anchor); front-load persists $+3.65\%$ $[1.77, 5.48]$ — SAME-BILL-LADDER-CLEAN + FRONTLOAD-PERSISTS-AT-64
exact-33 pairing	DAX-99, one fresh board	the one-shot revision beats a fresh 33-call tilt board at *exactly* matched calls (33 vs 33): $+30.754\%$ $[28.950, 32.514]$ (softord-33 524.218 vs reopen2x16 363.00); the ± 1-call declaration convention (DAX-45, DAX-79, favouring the tilt) is upgraded to an exact-bill read; the edge is not a $+1$-call artifact, and the tilt's marginal value of one extra forward measures $\sim +0.020$ nats; secondary vs the front-loaded 24A+8C member $+36.711\%$ $[35.019, 38.370]$ — REVISION-KEEPS-EDGE-AT-EXACT-33
AR anchor	DAX-102, gpt2-large ancestral board on the frozen instruments	MAUVE(AR, valref)@k=409 $= 0.588270$ ($2.8\times$ the family max), gen-PPL self $= 27.261$ (the family's floor is its own scale, 230–1480); d2 0.5462 inside the family band — AR-ANCHOR-PLACED-NON-DEGENERATE

(cont.)

Table 5: Headline measurements (cont.).

question	design	measured result (cont.)
cooled depth control	DAX-103, softord 64-call T=0.7 board	revision@33c and deeper cooled tilt@64c quality-tied (-0.62% $[-2.95, +1.63]$); telescope 32→64 at T=0.7 = $+12.91\%$ $[+10.92, +14.90]$ ($\sim 5\times$ the T=1.0 depth gain) — SPEND-STRADDLE-AT-T07
re-noising revival test	DAX-104, bitwise-verified trace of reopen2x16	r(per-sample availability lift, revision advantage) = $+0.0655$ $[+0.0222, +0.1085]$ straddles the registered 0.05 gate; agenda localizes context-RICH ($+0.079$ outCI, not context-poor) — AGENDA-ANATOMY-INFORMATIVE-ONLY
held-out fresh-seed replication	DAX-105, six boards at seed 780	five registered load-bearing keys all replicate: tilt $+11.71\%$ $[+10.04, +13.33]$, revision $+33.66\%$ $[+32.23, +35.06]$, cooled revision over cooled tilt $+12.30\%$ $[+10.71, +13.86]$, cooled tie -0.33% $[-2.54, +1.86]$ (tie form preserved), cooled telescope $+12.01\%$ $[+9.98, +14.00]$ — SEED-780-HELD-OUT-REPLICATION-HOLDS
revision matched-diversity calibration	DAX-106, nucleus-calibrated softord-33	d2(0.9375)=0.4462 / d2(0.9437)=0.4562 both miss the ± 0.010 board gate on the frozen revision diversity 0.4681 (one registered geometric-midpoint extension consumed); the revision-vs-matched-tilt key stays unmeasured at this ladder resolution — CALIBRATION-MISS
hardness top at high budget	DAX-107, $\beta \in \{0.5, 2\}$ at NFE=64	β=1 beats the soft flank $+5.96\%$ $[+4.08, +7.82]$ outCI; β=1 vs β=2 = -1.29% $[-3.60, +0.96]$ straddles 0: the $[1, 2]$ ceiling region is flat within measured noise — FLAT-TOP-AT-64
placement×tilt at 32 calls	DAX-108, gridsoft on the frozen K=32 equal-mass schedule	grid+tilt over uniform $+14.07\%$ $[+11.55, +16.49]$ (the tilt win's own scale); residual over flat+tilt $+1.12\%$ $[-0.87, +3.03]$ straddles 0 — TILT-ABSORBS-GRID-NEUTRAL-AT-32
revision-vs-matched-tilt at matched d2	DAX-109, nucleus-0.951 softord-33 to the revision's own 0.468137	within the chosen matching tolerance (Δ 0.003): revision -23.41% $[-26.67, -20.23]$ vs matched-tilt (294.14 vs 363.00); degenerating the tilt's diversity to the revision's own d2 flips the head-to-head — DIVERSITY-WAS-THE-EDGE
difficulty-weighted scorer	DAX-110, entropy-weighted re-read of the frozen DAX-95 dumps	$+10.695\%$ $[+8.380, +12.920]$; quarter ladder $17.78/9.48/5.90/2.10$ monotone; no re-rank — RANK-PRESERVED-ROBUST

(cont.)

Table 5: Headline measurements (cont.).

question	design	measured result (cont.)
fine β-plane at 64	DAX-111, $\beta \in \{0.75, 1.5\}$ and $\beta \in \{0.5, 2\}$ at T=0.7	β=1.5 +3.496% [+1.35, +5.65] vs β=1 at T=1.0 (edge migrates into $(1, 2]$); β=2 +16.02% [+13.88, +18.17] at T=0.7 — EDGE-MIGRATES-AT-64 + MIGRATION-PERSISTS-AT-64-T07
β-vertex bracket at 64	DAX-112, single interior probe β=1.75 at T=1.0	469.75: beats β=2 +2.43% [+0.46, +4.36] outCI, statistically tied with β=1.5 (+0.30% $[-1.76, +2.28]$) — the vertex lies in $(1, 1.75)$, mass toward 1.5 — VERTEX-NEAR-1.5-SIDE
notion-scope of the d2-match	DAX-113, six token notions on the frozen matched pair, zero new decode	board-pooled distinct-2 only: per-doc, tilt lower on distinct-1..4 and unigram entropy (-0.039 bits) outCI, +64.8% intra-doc rep4-mass — MATCH-IS-D2-SPECIFIC
distributional arbiter of the flip	DAX-114, registered MAUVE of the DAX-109 matched-tilt board	0.1522 vs the revision's frozen 0.0769 and the parent tilt's 0.0730 (+0.075, 19× the frozen separation; the nucleus calibration itself raises MAUVE two-fold); nearest cosine-tail partner is the revision — FLIP-SURVIVES-MAUVE
flagship numerator co-price	DAX-115, six token notions on the frozen tilt-vs-uniform pair, zero new decode	per-doc, tilt lower on all five richness notions outCI (H_1 -0.061 bits, d2 -0.0064) and +51% intra-doc rep4-mass — the diversity-consequence pattern is family-wide — TILT-EDGE-PARTLY-DIVERSITY-PRICED
featurizer rung census	DAX-116, gpt2-xl last-token MAUVE leg	ref-half selftest $0.764 < 0.90$ registered gate: census keys never computed; two attested degenerate rungs — MODERN-FEATURIZER-DEGENERATE-AT-SELFTEST
larger-n vertex split	DAX-117, g3r3/g3r4 re-decoded at n=4096	-0.0097 $[-0.055, +0.035]$ at k=500 (includes 0; k=409 continuity +0.0042; SE grew vs the 2048 bound) — SPLIT-CLOSES-AT-LARGER-N
razor CI of the repeat bar	DAX-118, doc-bootstrap B=20000 (one shared stream) over ten frozen boards	both razor points straddle the bar at z=0.007/0.011 plus the whole 0.0024–0.0027 band (five straddlers, no flips); gates discharge at point value with the sampling basis disclosed — RAZOR-FRAGILE
published EB rule at matched bills	DAX-119, Alg.1 same-step entropy-bounded prefix, $\gamma \in \{0.2, 0.5, 1.0\}$ nats, three boards	near-sequential by construction (239.6–247.2 certified calls); −33–36% vs the uniform ladder at its own reach; outbids the 32-call tilt +31.37% [+28.88, +33.83] at 7.5× the bill (unequal-computation, not matched-computation); rep_4 5.8–7.1× the bar — EB-BEATS-TILT (with named disclosures)
seed-780 replication of the clean points	DAX-120, exact g3 recipes re-decoded at seed 780	rep_4 0.002163/0.002459 $\leq$ bar; +59.0/ + 61.7% vs anchor and +30.6/ + 35.0% vs the clean single round, all outCI — REPLICATES

(cont.)

Table 5: Headline measurements (cont.).

question	design	measured result (cont.)
matched within-d2 retention	DAX-121, anchor quintile bands with joint containment, zero decode	the co-occupying 22.0% of docs keep $+8.53\%$ $[+3.53, +13.32]$ — $\geq$ half of the +11.79 anchor-relative win; retention grows with band richness — SURVIVES-HALF
per-position probe	DAX-80, three substrates	exploitable early structure only as a width-3 prompt-adjacency band on the conditional board (ρ $0.932/0.619/0.401$); bulk floor-complete on unconditional and L=512 (Δ $-0.209/-0.363$); signal lives at sample level — EXPLOITABLE-EXCESS (localized)
length transfer	DAX-81, L=512 second-length ckpt	softord-32 355.93 vs uniform-32 399.82: $+10.98\%$ $[9.77, 12.18]$ outCI, registered prior $[8, 14]\%$ inside; d2 profile same sign — TILT-TRANSFERS-L512
semantic-tail audit	DAX-82, block-cosine MC, 7 frozen boards	guarded-chain means below base (DAX-77 direction holds) but a surface-guard-invisible semantic-duplicate tail: alarm-rate excess $+1.27$pp $[+0.49, +2.05]$ outCI at the round-3 vertex, stable across six τ settings, $g4 \equiv$ unguarded at every τ — SEMANTIC-EXCESS-GUARDED (tail-localized)
semantic tail repair	DAX-83, trigger@τ=0.68 + +2 guarded passes; RND control	tail surgically removed ($2.29\%\rightarrow0.49\%$, in-band; RND 2.25%) yet both arms $\approx -1.3\%$ vs g3r3 outCI, trigger irrelevant ($+0.04$ straddle), d2 $+0.011$ identical — NO-RECOVERY
modern-featurizer MAUVE	DAX-84, qwen2.5-1.5B features, official recipe	range $0.0063 <$ registered 0.01 across 10 boards — Q-MAUVE-DEGENERATE; featurizer panel stays gpt2-large-only, boundary named
tilt mechanism axes	DAX-85, 3-arm draw-paired traces, band anatomy	commit-band error anatomy does not separate orders (premature metric saturated 0.925; no mid hump -0.1175 outCI) — SOFT-IS-JUST-DRAW (registered null)
context strata	DAX-86, commit-time context availability $C(j)$	soft wins exactly in context-poor strata (rates 5.66 vs 6.29 low; 6.44 vs 6.50 mid-low) and redirects error into rich strata; hard's catastrophe localizes to late context-poorest commits ($+1.17$ rate, $+5.5$pp share vs soft, both outCI) — INDETERMINATE (mechanism sentence earned)
context-strata replication	DAX-87, second-substrate two-arm traces (wikitext L=256; OWT L=512)	one-sided population gates razor-fail on both L=512 arms by 0.002 nats — GATE-FAIL (machinery suspicion; the investigation names a point-gate razor inside mutual draw noise and exonerates the machinery)

(cont.)

Table 5: Headline measurements (cont.).

question	design	measured result (cont.)
context-strata replication, gate-form fix	DAX-88, two-sided population gates on the same frozen assets, zero new decoding	all four gates pass; all three components outCI with flagship signs on both substrates (poorest-band edge $-0.542/-0.541$; redirect $+1.561/+1.655$; share $-0.026/-0.028$) — STRATA-REPLICATE; mechanism sentence elevated to multi-substrate

checkpoint	uniform gen-PPL	grid gen-PPL	win %	Δdistinct-2
owt_mdlm_s0	615.20	608.70	$+1.06$	$+1 \times 10^{-5}$
owt_tge_s0	567.69	562.93	$+0.84$	-2×10^{-4}

Table 6: Fixed-grid quality comparison (n=2048 per arm, paired seed 777, NFE=32). Gate required win $\geq 5\%$ on both checkpoints with distinct-2 drop ≤ 0.005: **NULL**, 0/2 checkpoints pass. Split-half (1024+1024) gen-PPL per arm, in registration order: mdlm uniform $613.46/616.94$, grid $608.55/608.85$; tge uniform $571.79/563.61$, grid $564.19/561.66$. Per-sample paired-bootstrap win CIs (B=20000, jointly resampled boards; DAX-4b′/DAX-6 analyses): mdlm $[-0.69, +2.80]$, tge $[-0.99, +2.70]$ — neither excludes zero at n=2048; the small positive effect is not individually significant at NFE=32 and is disclosed symmetrically (both point estimates remain below the 5% threshold).

schedule (NFE)	gen-PPL	vs uniform-32	distinct-2
uniform 16	715.37	-16.28%	0.5682
uniform 32 (baseline)	615.20	—	0.5568
belief-grid 32	608.70	$+1.06\%$	0.5568
uniform 64	575.27	$+6.49\%$	0.5539

Table 7: DAX-3 within-protocol NFE ladder (n=2048 per cell, paired seed 777, owt_mdlm_s0, temperature 1.0). Placement (row 3) has negligible per-decode cost and a small effect; budget (row 4) has the larger effect. Split-half stability: uniform-16 $714.88/715.87$; uniform-64 $570.53/580.05$. The dash marks the baseline row.

D PUBLIC-MODEL SENSITIVITY AND DETAILED COMPARISONS

This section provides the estimates and sensitivity analyses summarized in Section 3, including unresolved and negative results.

Cross-scorer sensitivity. A larger scorer gives the same effect directions on the public model: gpt2-xl re-scoring of the unchanged generated documents through byte-identical detokenization pipelines preserves the sign and outcome class of every registered comparison (DAX-129: tilt $+11.92\%$ $[+1.42, +21.33]$ at the 32-call anchor, budget endpoints $+10.05\%$ $[+1.32, +18.16]$ and $+23.34\%$ $[+12.82, +32.75]$, the T=1.0 interior holding both flanks straddling, and the cooled harder-better pairs $+11.01\%$ $[+2.32, +18.99]$/ $+12.50\%$ $[+5.28, +19.46]$, all intervals positive; no comparison reverses direction or loses its original outcome class).

Transfer on LLaDA-8B-Base. The central cells re-implement the public LLaDA-8B-Base weights (Nie et al., 2025) at n=1024, L=256, T=1.0, with the measured and scored boards bundled (material accounting in Appendix F). The sampler is the same ancestral uniform-position-order decoder as the in-house cells, at a fixed 32-commit schedule — not the substrate's block or semi-autoregressive variants — so the comparison isolates the commit order, grid, and revision channels; every public arm carries n=1024 documents with the scorer run offline after decode. At 32 calls

the tilt transfers at band scale: $+11.91\%$ $[+1.28, +21.43]$ (soft order 86.01 vs uniform 97.64) — the β=1 order effect is not specific to our private 124M family. And the transfer is robust to fresh decode seeds in the pre-registered panel: both fresh-seed gains fall inside the frozen DAX-122 band with outCI-positive intervals (DAX-132; spread 3.52 pp $<$ 7.42 pp); the three-seed reading is stable at the available statistical precision (registered forecast surprise 2.40 bits). The transfer is budget-resolved at both endpoints: 16 calls $+10.55\%$ $[+1.50, +18.74]$ (161.49 vs 180.54) and at 64 calls $+22.80\%$ $[+11.86, +32.24]$ (54.54 vs 70.66), so the order win replicates across the measured budget axis and the widening direction is positive pointwise ($16 < 32 < 64$ calls) and the in-house ladder's shape (§3) transfers at larger point magnitude, while the between-budget contrast itself is under-resolved at this estimator budget (DAX-136: 64-vs-32 link $+0.132$ nats $[-0.037, +0.303]$, 32-vs-16 $+0.015$ nats $[-0.130, 0.160]$; cross-budget link MDE ~ 0.243 nats ≈ 20 pp at n=1024; the in-house 64-vs-32 link reads $+0.025$ nats $[-0.015, 0.065]$ at n=2048, likewise point-level, DAX-139): a point-level, not contrast-resolved, reading (split halves $+4.9/+15.9$ and $+19.8/+25.7$, all positive; in-house 64-call value $+15.24\%$). The 64-call hardness sweep stays interior at β=1: β=2 reads 55.71 (-2.14% $[-10.94, 6.03]$ vs β=1, straddle) and β=0.5 reads 57.24 (-4.94% $[-11.93, 1.65]$), with all three soft arms beating uniform outCI ($+21.15/+22.80/+18.99\%$) — the in-house hard-ward migration at 64 calls does not replicate on the public substrate at measured precision (80%-power MDE 9.7–12.1 pp at this n=1024 pairing), so we retain β=1 as the reference operating point at all measured public budgets at T=1.0; these intervals do not establish a unique optimum.

The cooled corner. Cooling the same substrate to T=0.7 at the 32-call budget replicates the in-house cooler ordering through every measured β point (DAX-127): β=2 (27.35) beats β=1 (30.71) by $+10.96\%$ $[+2.25, +18.72]$ and β=1 beats β=0.5 (34.43) by $+10.79\%$ $[+3.58, +17.50]$, both intervals positive — the hardness migration is temperature-confined on both substrates (interior at T=1.0, harder-better under cooling). The cooled-corner tilt-vs-uniform read needs a boundary sentence rather than a premium claim: at n=1024 all three pair levels straddle ($+4.26\%$ $[-22.03, 25.25]$ for β=1, $+14.75\%$ $[-9.31, 33.78]$ for β=2, -7.32% $[-37.57, 16.79]$ for β=0.5; 80%-power MDE 30.8–38.8 percentage points at this bimodal corner), and the blanket descriptor is measured: the boards are document-level bimodal (≈4–5% catastrophic documents per arm — per-sample mean generated-NLL $>$ 10 nats — at this corner, versus ≈1% at T=1.0), with the three confidence-ordered arms largely *sharing* their collapse sets while the uniform-order board collapses on a *nearly disjoint* document set — ordered-family overlap is high, but the uniform-paired deltas carry two-sided 20-nat tails priced by the registered intervals, and the order channel's resolved reads against the uniform anchor on this substrate remain the T=1.0 members. Cooling itself dominates the surface here as in-house (same-arm T=1.0 $\rightarrow$ 0.7 deltas $+64.3\%/+67.1\%$, both intervals excluding zero). Extending the same lattice to the cooled deep corner (64 calls at T=0.7, DAX-128) closes the last unmeasured $(T, \beta, \text{budget})$ cell on this substrate and resolves SATURATION-FLATTENS-AT-64-T07: the whole soft family crowds the floor — uniform 28.51, β=0.5 22.37, β=1 22.09, β=2 21.09 — and no pairwise β read separates at measured precision (β=2 vs β=1 $+4.51\%$ $[-8.01, 15.52]$; β=1 vs β=0.5 $+1.26\%$ $[-8.01, 9.92]$; β=2 vs β=0.5 $+5.71\%$ $[-5.89, 15.81]$, all straddle; 80%-power MDE ≈ 13–17 percentage points as registered), so the harder-better ordering that resolves at the 32-call corner does not survive resolution at depth, even though every point estimate still points hard-ward; within the corner's priced document turbulence the resolved order-vs-uniform read at depth is the β=2 member alone ($+26.01\%$ $[+2.00, +43.91]$, interval excluding zero; β=1 $+22.52\%$ $[-2.49, 41.12]$ and β=0.5 $+21.53\%$ $[-2.48, 40.34]$, straddle). Budget still pays the soft family at the cooled corner (64-vs-32 same-arm: β=0.5 $+35.02\%$ $[+15.50, +49.73]$ and β=1 $+28.08\%$ $[+6.60, +44.50]$, both intervals excluding zero; uniform $+11.13\%$ $[-13.94, 30.45]$ and β=2 $+22.87\%$ $[-0.41, 40.62]$ straddle), and cooling dominates at depth for every measured arm (T=0.7 vs T=1.0 at 64 calls $+59.50$–$+62.14\%$, all four intervals excluding 0). The revision edge **inverts**: the same guarded revision reads -71.68% $[-91.04, -54.38]$ (167.63 vs 97.64), buying surface diversity (distinct-2 0.395 vs 0.350) at ≈1.26 nats of scorer coherence — this is consistent with the selected low-confidence tokens playing a different role on LLaDA, but does not establish a causal account of rare-token content or human-perceived coherence. The reversal also persists beyond the shallow-round setting: re-implementing the in-house multi-round frontier verbatim (36 calls, exact 3-gram guard, 0 overrides) yields 162.03 gen-PPL — -65.94% $[-85.24, -48.51]$ against uniform-32 and -88.38% $[-104.86, -73.11]$ against the 32-call tilt, yet $+3.34\%$ $[+0.44, +6.08]$ over its own two-round twin: revision depth still pays within the failing family, but the family still performs worse than plain uniform decoding here (DAX-131). The published entropy-bounded unmasking rule (Ben-Hamu et al.,

2025), measured verbatim at its contract (same-step entropy budget, no fallback), is near-sequential at the tested entropy budget on the 8B model (mean 204.0 calls against the 256-step ladder).

Entropy-bounded triggering at its own call count. On the in-house family the published rule at $\gamma \in \{0.2, 0.5, 1.0\}$ nats reads mean $247.2 \to 239.6$ calls (near-sequential), beats the uniform ladder by $+33.2$–$+36.4\%$, and beats the 32-call tilt by $+31.37\%$ $[28.88, 33.83]$ — at $7.49\times$ the certified call bill, with repetition 5.8–$7.1\times$ the bar and lower distinct-2: a gain-with-bill-and-repetition-cost member of the family map, not a same-budget comparison.

Joint diversity matching leaves too few documents beyond one metric. A multi-notion joint-containment match with pre-committed windows (distinct-2 bands crossed with unigram entropy, distinct-1/3, and repetition windows) exists mathematically — the union reproduces the single-notion construction bit-exactly — but every joint tier collapses below a pre-committed 150-document floor ($81/41/26/16$ documents at the four window levels; half-window variants $35/8/5/3$), all with intervals straddling the read. The deployment consequence is methodological, not incidental: the available sample supports a **single-metric** diversity comparison at the specified precision, and the priced single-notion read for the flagship pair retains $+8.53\%$ of the $+13.1\%$ tilt win at a 22.0% corpus share — positive but partial, so the tilt's quality edge is established *at its operating diversity*, not as a diversity-matched invariant.

E Experimental conditions and detailed outcomes

The tables below group the recorded comparisons by the scientific question: early-trace prediction, fixed time grids, commit order, and revision. Each row states its experimental condition or numerical criterion and the observed result. Intervals describe the recorded sample or document resampling; crossing a chosen point threshold does not imply that the interval separates from it. Negative estimates, intervals containing zero, and unsuccessful controls are retained. Experiment identifiers link these summaries to the detailed methods and accompanying outputs.

Test	Comparison or criterion	Measured result
Adaptivity scope (belief traces, n=512)		
cross-horizon persistence (V1)	$\rho(k{=}4, k{=}28) \geq 0.5$; NEG at ≤ 0.2	0.303 (Fisher CI $[0.22, 0.38]$); second arm -0.068; NEG branches fire
relative spread ratio (V2)	$\mathrm{RS}_e/\mathrm{RS}_l \geq 2.0$; NEG at ≤ 1.2	0.992 (0.179 vs 0.181); NEG
step-0 pool statistics	non-degeneracy of the early channel	exactly constant across samples (variance 0, corr. undefined); closure with positive evidence
quality-axis probe	squared-correlation magnitude on gen-NLL at horizon $k{=}4$	$\rho = -0.320/-0.230$ (two checkpoints); 5–10% of final-quality variance
oracle savings bounds	pre-committed efficiency bar: $\geq 20\%$ NFE saved at quality parity	step-share bounds $0.21/0.15/0.41$ (all span the bar) but every bound concerns steps rather than quality: bar not met under any identified estimator
shape-null scope	within common-$\bar{\rho}$ pure-shape model	exclusion holds inside the family (deflation ≈ 0.83 measured under the corrected μ-ratio identity); mixed-noise counterexamples named: nonzero shape outside the family is possible with both two-moment anchors matched — the exclusion does not extend beyond its stated family
within-pool correlations	direct measurement of in-family status	$0.0139/0.0089/0.0062$ at horizons $4/16/28$ (CI uppers ≤ 0.0163): an order of magnitude inside the certified family

Test	Comparison or criterion	Measured result

Placement: belief-profiled grids vs uniform		
budget-ladder head-to-head	paired win of grid over uniform at $K = 4/8/16/32$, CI gates	$+4.26\%/+4.18\%/+2.84\%$ outCI at 4/8/16; $+1.06\%$ CI incl. 0 at 32; uniform 64-call arm $+6.49\%$ (budget dominates placement)
vantage stability	half trace-board re-derivation	bit-reproducible; cross-checkpoint (time-weighted objective) grid differs by ≤ 0.026 of a step
alternatives	blind grids, deficit^2 and entropy weights at 8 calls	squared-deficit beats the initial equal-w form; two working grids' interval contains 0 once combined with the tilt
out-of-family re-derivation	WikiText re-derived 32-step grid	$+2.08\%$ ($627.15 \rightarrow 614.11$); size of the effect outside the construction domain

Test	Comparison or criterion	Measured result
Commit ordering: thresholds and trigger		
matched-diversity read	distinct-2 containment match on noise-paired boards, w=0.05 primary, ladder $\{0.02, 0.08\}$, floor $n \geq 150$, class fixed before measurement	$+10.57\%$ $[+8.08, +13.01]$ at n=1749; $+9.28\%$ $[+6.00, +12.41]$ (n=980) and $+12.31\%$ $[+9.93, +14.62]$ (n=1986): all outCI, post-treatment-paired
hard vs soft at 32 calls	direction gating at matched budget	hard 1285.04 vs uniform 615.20 (uniform gen-PPL 52.13% lower; signed reduction of hard vs uniform -108.9% under Eq. 2); soft β=1: 534.62 ($+13.10\%$ outCI), $534.62/556.22/557.93$ across $\beta = 1/2/0.5$ hardness
four decode-seed panel	per-seed direction replicate	seeds 777/778/779/780: $13.10/11.04/11.28/11.71\%$; seed center $+11.8\%$; dispersion 1.04pp
cross-recipe replicate	TGE objective checkpoint	tilt $+14.00\%$, grid-soft compose $+8.99\%$, grid alone $+3.82\%$ (all outCI)
T=0.7 cooling	hardness migration gate	harder-better monotone through all measured points: β=2 below β=1 by $+11.23\%$ $[9.81, 12.63]$; uniform 35.82
noise-only control	β=0 exchangeability	606.10 vs uniform 615.20, CI contains 0; this unresolved difference is not an equivalence test
cumulative-budget trigger	log-linear ladder anchors	certified per-sample means $47.3/92.2/179.2$ calls for $\beta \in \{2, 4, 8\}$; $+241.6\%/+100.2\%/+1.8\%$ vs anchors (PPL ratios, cell/anchor-1)

Test	Comparison or criterion	Measured result
Revision		
single-round reopen (16+1+16)	vs tilt anchor 534.62at 33 calls	$+32.10\%$ (363.00); exact-33-call pair $+30.75\%$ $[28.95, 32.51]$; random-target control: no gain
unguarded two rounds	repetition bar	227.69 with rep-4 4.60×10^{-3}: negative result at the 2.5×10^{-3} bar
3-gram guarded ladder	rep-4 bar at each round	guarded-2 279.28 (2.499×10^{-3}), guarded-3 250.49 (2.434×10^{-3}), guarded-4 242.44 (2.463×10^{-3}), margins $+23.06/+10.31/+3.21\%$ outCI
MAUVE of the ladder	non-domination disclosure	$0.1251 \rightarrow 0.1546 \rightarrow 0.1103$: named inversion at round four; distinct-2 $0.426 \rightarrow 0.406 \rightarrow 0.397$
razor disclosure	sampling status of the two tightest passes	doc-bootstrap: straddle at $z = 0.007/0.011$; gates discharge at point value with basis declared

fresh-seed replicate	frozen recipes at seed 780	rep-4 $0.002163/0.002459$ pass; $+59.02/+61.67\%$ over uniform, $+30.55/+35.04\%$ over single-round outCI
matched-distinct-2 flip	nucleus-calibrated tilt at revision diversity 0.468	matched tilt BEATS revision on both arbiters: -23.41% $[-26.67, -20.23]$ gen-PPL at exact 33×33; MAUVE 0.1522 vs 0.0769
vertex stability at $2\times$ sample	guarded third vs fourth round at n=4096	-0.0097 $[-0.055, +0.035]$ includes 0 (the $+0.0443$ term inverts): the round-3/4 vertex is a trade-off, not a stable winner

Test	Comparison or criterion	Measured result
Diversity / methodological boundaries		
within-d2 retention	union of five distinct-2 bands	$+8.53\%$ $[3.53, 13.32]$ above the pre-committed half-bar: retention positive but partial
multi-notion joints	joint tiers vs 150-doc floor	all tiers collapse: $81/41/26/16$ docs (half-window $35/8/5/3$): deployment must pick ONE notion
per-doc notions	six notions priced per doc	all separate outCI; diversity-consequence pattern is a decoding-family pattern
entropy-weighted scorer re-read	re-ranking sensitivity	$+10.70\%$ $[8.38, 12.92]$, drift inside band: flagship pair robust to weighting

Test	Comparison or criterion	Measured result
Public substrate (LLaDA-8B-Base)		
tilt transfer	band-scale sign replicate	$+11.91\%$ $[+1.28, +21.43]$ outCI at 32 calls (86.01 vs 97.64)
tilt budget response	endpoint replication + widening shape	16 calls: $+10.55\%$ $[+1.50, +18.74]$ (161.49 vs 180.54); 64 calls: $+22.80\%$ $[+11.86, +32.24]$ (54.54 vs 70.66): both outCI; halves $+4.9/+15.9$ and $+19.8/+25.7$ all positive
hardness at 64 calls	migration-direction gate (both poles named)	**interior holds**: β=2 reads 55.71 (-2.14% $[-10.94, +6.03]$ vs β=1, straddle), β=0.5 reads 57.24 (-4.94% $[-11.93, +1.65]$), all soft arms beat uniform outCI ($+21.15/+22.80/+18.99\%$); in-house hard-ward migration does not replicate
cooled corner (T=0.7, 32 calls)	harder-better gate (both poles named)	**ordering transfers**: β=2 (27.35) beats β=1 (30.71) $+10.96\%$ $[+2.25, +18.72]$, β=1 beats β=0.5 (34.43) $+10.79\%$ $[+3.58, +17.50]$, both outCI; tilt-vs-uniform pairs all straddle at n=1024 ($+4.26\%$ $[-22.03, +25.25]$ β=1) — bimodal-collapse pricing named
cooled corner at depth (T=0.7, 64 calls)	saturation gate	**saturation flattens**: uniform 28.51, β=0.5 22.37, β=1 22.09, β=2 21.09; all three pairwise β intervals straddle ($+4.51/+1.26/+5.71\%$); β=2 vs uniform $+26.01\%$ $[+2.00, +43.91]$ the only resolved read; budget pays the soft arms (β=0.5 $+35.02\%$, β=1 $+28.08\%$, outCI); cooling dominates at depth ($+59.5$–$+62.1\%$, all four outCI)
revision transfer	direction	**inverted**: -71.68% $[-91.04, -54.38]$ (167.63); d2 0.395 vs 0.350 at $\approx$1.26 nats coherence price
published EB availability	near-sequential construction check	204.04 mean calls $\in [200, 256]$ at γ=1.0, min 29 max 256: near-sequential by construction at 8B too

published EB (in-house), bill priced	placement at its own call cost	beats uniform by 33–36% at $\gamma \in \{0.2, 0.5, 1.0\}$; beats 32-call tilt $+31.37\%$ $[28.88, 33.83]$ at $7.49\times$ the bill; repetition 5.8–$7.1\times$ bar

Test	Comparison or criterion	Measured result
Apparatus controls		
scorer fidelity	bit-diff of offline vs GPU scorer	max row diff 0.0 across re-scored cells
AR fluency anchor	family placement	AR anchor MAUVE 0.588 / gen-PPL 27.3 (far above family maxima)
featurizer degeneracy	self-test gates	larger-featurizer MAUVE fails self-test ($0.764 < 0.90$); second rung reads 0.0063 at family scale
wall-clock parity	per-forward across arms	1.0000 per-forward parity across arms; revision premium $+2.24\%$ wall at batch 64

F DATA, EVALUATION, AND REPRODUCIBILITY

Available materials. Public-model experiments use LLaDA-8B-Base weights with reimplemented decoders for uniform grids, soft order, guarded revision, and the published entropy-bounded rule. The recorded outputs and scorer inputs support recomputation of their reported comparisons. For the two internal 124M checkpoints, the accompanying repository provides the training code, configurations, and training records, and their score arrays and analysis outputs support recalculating the reported statistics; the checkpoint weights themselves are not redistributed. The numerical results concern the specific implementations described here.

Sampling and scoring. Unless a cell states otherwise, decode length is 256 and temperature is 1.0. Internal cells contain 2,048 samples in two 1,024-row halves; DAX-30/117 use 4,096 samples. Public cells contain 1,024 samples. Decode seed 777 is shared within compared pairs, with replicate seeds 778/779/780 where stated. Trace measurements use 512 internal or 1,024 public samples. GPT-2-large provides the primary fixed scorer; the public cross-scorer panel is identified separately. The reported relative gain is $100(1 - \exp(\Delta \text{mean NLL}))$. Document bootstrap intervals use 20,000 resamples. Some analyses use one shared index stream (DAX-122/132/136), whereas others use per-pair streams (DAX-125–129/131); their scripts specify the seeds and offsets from 20260908. The repetition threshold is 2.5×10^{-3} and the minimum distinct-2 matched subset has 150 documents. These document intervals do not incorporate variation over model training.

Temperature and matching. Temperature sweeps are reported as within-temperature comparisons or named same-arm temperature changes, rather than pooled across temperatures. The cooled experiments include DAX-5/68/71/76/103/105/107/111/127/128; DAX-76 uses $T \in \{0.7, 0.8, 0.9\}$. Distinct-2 matching addresses the specified diversity statistic; it does not equate the complete generation distributions. Comparisons using arbitrary cross-budget sample indices retain that pairing as an analysis convention, rather than asserting a natural sample-level match or exact equality with unpaired variance.

Compute scope. The decoding campaign through 2026-09-12 comprises 201 GPU jobs (53.5 GPU-hours) and 139 CPU jobs (96.3 job-hours), across 340 recorded jobs. The separate training study (71 jobs, 53.5 GPU-hours) is excluded. Per-job records and analysis inputs accompany the experiments; the initial trace/grid/budget study used five GPU jobs and four CPU scoring jobs, a subset of this total.

G WHAT POSITION-AVERAGED TRACES IDENTIFY UNDER A SHAPE MODEL

The DAX-1 null concerns the measurement instrument rather than a general limitation of the sampler: under the ancestral sampler's *exchangeable* (uniform-position) unmask order, position averaging

inside the trace statistic mechanically suppresses the visibility of any persistent per-sample shape factor — and, as the constructions below make precise, two-moment trace anchors do not identify shape at all once horizon noise is admitted.

Setup and assumptions. At horizon k, sample i has masked pool $M_k(i)$; the trace statistic averages per-position belief over the pool. We use the explicit model class

$$x_{i,k} \;=\; \mu_k \;+\; a_i \;+\; \underbrace{\tfrac{1}{|M_k(i)|}\textstyle\sum_{j\in M_k(i)}\delta_{i,j}}_{\text{pool-mean atom deviation}} \;+\; \varepsilon_{i,k},$$

with three assumptions, stated before any conclusion: **(B1, factor form)** a_i is a per-sample *altitude* offset (relative sd σ_a); $\delta_{i,j}$ is a per-atom persistent difficulty deviation (sd σ_δ), identical across horizons — this is the *shape factor* whose existence DAX-1 tests; the pool rotation is exchangeable (uniform-position order), so $M_k(i)$ is a uniform-random $|M_k|$-subset of the surviving atoms and the pool-mean deviation has cross-sample variance $\sigma_\delta^2\big(1+(|M_k|-1)\bar\rho_k\big)/|M_k|$ within the equicorrelated class ($\bar\rho_k$ = average within-pool atom-atom correlation, $\bar\rho\geq 0$; independence included); **(B2, horizon noise)** $\varepsilon_{i,k}$ is horizon-local averaging noise, variance v_k non-increasing-relevant only at early horizons; **(B3, checkable regime)** early-pool variance is not altitude- dominated; both fitted regimes below are reported.

Lemma 1 (shape factor: coefficient inflation, normalized-prediction deflation)**.** *Under B1–B3 with $\sigma_a, \sigma_\delta, v_k \geq 0$, the relative spread obeys*

$$\mathrm{RS}_k^2 \;=\; \frac{\sigma_a^2 + \sigma_\delta^2\,\big(1+(|M_k|-1)\bar\rho_k\big)/|M_k| + v_k}{\mu_k^2},$$

so that a persistent per-atom shape factor $\sigma_\delta > 0$ implies the shape term's late/early contribution to RS^2 *is multiplied by*

$$\mathcal{A}(\bar\rho) \;:=\; \frac{\big(1+(|M_l|-1)\bar\rho\big)/|M_l|}{\big(1+(|M_e|-1)\bar\rho\big)/|M_e|} \quad \textit{(so } 1 \leq \mathcal{A}(\bar\rho) \leq |M_e|/|M_l| \textit{ for } \bar\rho \in [0,1]\textit{)},$$

where $\mathcal{A}$ denotes the variance ratio throughout — the common-$\bar\rho$ specialization ($\bar\rho_k \equiv \bar\rho$) is a declared model contraction, used for the identifiability claim only — and hence the shape-only prediction for the measured (μ_k-normalized) late/early relative-spread ratio is $\sqrt{\mathcal{A}(\bar\rho)}\cdot(\mu_e/\mu_l)$, which is *decreasing* in $\bar\rho$ from $\sqrt{|M_e|/|M_l|}\cdot(\mu_e/\mu_l)$ toward μ_e/μ_l as $\bar\rho : 0 \to 1$ *($\mathcal{A}$ itself decreases from $|M_e|/|M_l|$ toward 1; each effective count $|M|/(1+(|M|-1)\bar\rho)$ shrinks in $\bar\rho$). Conversely, an altitude-plus-noise reading ($\sigma_\delta = 0$) pins the ratio to the noise profile alone.* Sketch. *Pool-mean variance under equicorrelation gives the displayed variance decomposition; the ratio isolates the $m \mapsto (1+(m-1)\bar\rho)/m$ factor; monotonicity by $d/d\bar\rho < 0$.* □

Calibration (not implication). The horizon dependence of the floor model is summarized by two fitted values, $\bar\rho_4 \approx 0.06$, $\bar\rho_{28} \approx 0.84$: free parameters chosen *after* seeing the trajectory so the noise-only model tracks it. They demonstrate that a noise-only account *exists* (with correlation rising as the pool shrinks, the expected shared-context direction); they are not evidence, and the falsification conclusion below uses only measured anchors and the class-bound of the lemma, never these fitted values.

Measured falsification (DAX-1 anchors; prediction evaluated in DAX-89). The frozen anchors are pool sizes $|M_e|\approx 224$, $|M_l|\approx 32$ and horizon means $\mu_e = 0.1172$, $\mu_l = 0.3742$, so the shape-coefficient family has $\sqrt{\mathcal{A}(0)} = 2.65$ and the pure persistent-shape prediction for the *measured* (μ-normalized) late/early relative-spread ratio is $\sqrt{\mathcal{A}(0)}\cdot(\mu_e/\mu_l) = 0.830$ — a $\approx -17\%$ *deflation*, not an inflation. Measured ratio $= 1.0084$ (relative spread $0.1792 \to 0.1807$). A *paired* sample-level bootstrap of the ratio ($B = 20000$, resampling the same 512 samples at both horizons jointly) gives standard error 0.0475 and 95% CI $[0.920, 1.106]$ — the pure-shape prediction sits below the CI and is excluded at $z \approx 3.75$ ($\approx 1.9\sigma$ against the family bound 0.830 at the CI edge; the point-estimate form is the optimistic quote, and DAX-58's directly measured $\bar\rho_k \leq 0.016$ independently supports the conclusion) *with the horizon pairing accounted for* (`runs/_derived/propA_stats.json`; read-out recorded as DAX-89, `runs/_derived/da_dax89_verdict.json`). Applying the

shape-coefficient ratio to the μ-normalized statistic without the μ_k^2 factor would instead predict a $\geq$2.65 inflation, excluded at $z\approx$34.5; that is not the implied prediction for the measured quantity, and the exclusion holds with the numbers above. (Indexing note: the labels k=4/k=28 denote the 0-indexed trace columns $4/28$, i.e., passes $5/29$; all quoted $|M|$, spread, and μ values use these columns.)

Dependence of the anchor on within-pool correlation (DAX-89). What does within-pool correlation do to the exclusion? In the equicorrelated class the shape-coefficient family from the lemma is $\mathcal{A}(\bar{\rho}) = M_e(1+(M_l-1)\bar{\rho})\big/\big(M_l(1+(M_e-1)\bar{\rho})$ ($\sqrt{\mathcal{A}}$ *decreasing* from $\sqrt{|M_e|/|M_l|} = \sqrt{7} \approx 2.65$ to 1; $\mathcal{A}$ denotes the variance ratio and $\sqrt{\mathcal{A}}$ the standard-deviation ratio throughout, which are distinguished in every display; attaching the factor to the wrong pool size instead gives the spurious increasing family $2.65 \to 7.02$, and the independence-basis exclusion does not depend on this distinction). With the μ_k^2 factor, the pure-shape prediction for the measured ratio reads $\sqrt{\mathcal{A}(\bar{\rho})} \cdot (\mu_e/\mu_l)$, decreasing over the *whole* family from 0.830 down to $\mu_e/\mu_l = 0.313$ — entirely below the measured 95% CI $[0.920, 1.106]$. Solving $\sqrt{\mathcal{A}(\bar{\rho})} \cdot (\mu_e/\mu_l) =$ measured ratio would need $\mathcal{A} = 10.4 > |M_e|/|M_l| = 7$: *no* $\bar{\rho} \in [0, 1]$ rescues the pure-shape reading (`runs/_derived/da_dax89_verdict.json`). The exclusion therefore holds for the *entire* equicorrelated family, moderate *and* strong correlation alike. Matching the unnormalized family against the μ-normalized statistic would instead suggest a rescue at $\bar{\rho} \approx 0.61$ (point) / $\bar{\rho} \approx 0.10$ (CI edge); these values do not apply to the measured statistic. (Had a rescue existed, its regime would have been degenerate in the precise sense developed here: per-atom deviations correlated strongly act increasingly as a *common level* shift of the remaining pool — the altitude channel, whose total cross-horizon content is independently quantified and bounded (r=0.303 CI $[0.22, 0.38]$, $R^2 \leq 0.15$; §O) — but the algebra above leaves no such regime, so no appeal to the altitude channel is needed.) *No cap on the shape size (counterexample audit).* No family-independent 95% cap on σ_δ follows from these anchors; in particular, a cap such as $\sigma_\delta \lesssim 0.025$ relative sd does not follow (the DAX-58 within-pool-correlation measurement bounds $\bar{\rho}_k \lesssim 0.016$, not the shape size). Two explicitly computable constructions show that two moment anchors do not identify shape even at the same common-correlation side: with the rounded anchors (μ=$(0.1172, 0.3742)$, RS=$(0.1792, 0.1807)$, $|M|$=$(224, 32)$), common $\bar{\rho}$=0 with σ_δ=0.2, σ_a=0 reproduces both variances exactly at v_e=0.0002625227, v_l=0.0033221858 ≥ 0; and a ρ_k-varying construction with σ_δ=0.31433275, ρ_e=0, ρ_l=0.0155094785, v=0 likewise reproduces both anchors. The pure-shape exclusion inside the common-$\bar{\rho}$ two-moment model therefore coexists with nonzero shape under mixed noise, and any valid shape bound requires explicit noise or covariance constraints; the exclusion holds for the common-$\bar{\rho}$ pure-shape two-moment model, not for "any nonzero shape". General correlation kernels remain non-identifiable from a single trace (no per-position replicates), which is the residual assumption and the identifiability boundary of the test. The equicorrelated floor model fitted to the trajectory uses $\bar{\rho}_4 \approx 0.06$, $\bar{\rho}_{28} \approx 0.84$ (rising with pool shrinkage, the shared-context direction); at $\bar{\rho} \approx 0.06$ the shape hypothesis would itself mandate $\sqrt{\mathcal{A}(0.06)} \cdot (\mu_e/\mu_l) \approx 1.180 \cdot 0.313 \approx 0.370$, again below the CI lower edge — so within the floor model's own early-correlation regime the shape reading stays excluded, and the family-maximum 0.830 already quoted is the binding margin. The noise-only floor model reproduces the full spread trajectory to first order, including the mid-horizon dip (μ_k rising while $|M_k|$ holds) and the horizon-end rise at $|M_{31}| \approx 8$ (spread $0.328 \approx$ floor extrapolation 0.28): even the apparent tail emergence of between-sample variation is averaging noise, not difficulty structure.

Observation 1 (consequence for adaptivity)**.** *Sample-adaptive exit rules must either (i) operate on the altitude channel, whose predictive content is weak ($R^2 \leq 0.15$; quantified in §O), or (ii) attempt to use shape structure that this trace family cannot certify: the mixed-noise constructions above reproduce the same two-moment anchors at arbitrary σ_δ, so no variance-level cap on shape follows from these anchors (the two observed moments do not bound σ_δ outside the restricted model). Position-averaged traces are therefore shape-*blind*, not evidence of shape absence; the operational fact for this sampler family is negative in the practical sense only: none of the measured early-exit rules meets its quality–cost criterion (§5), and the exploitable early-predictive share is $R^2 \leq 0.15$ of between-sample variance, which does not support the rules we tested. Checkpoint-level reallocation (Track B) is unaffected: it uses the* aggregate *profile only.*

H TIME-GRID WEIGHT CONSTRUCTION AND SENSITIVITY

This appendix concerns the weight $w_k = (1 - \overline{p_{\max\cdot,k}}) \cdot \overline{f}_{\cdot,k}/256$ of the time-grid construction (Appendix P). We ask where this weight stands relative to an optimal placement functional, rather than asserting it is one.

A1 (deficit-risk model). For calibrated beliefs, a position decoded at top-probability p carries expected quality risk proportional to its deficit $1 - p$ (the quantity our DAX-5b ceiling follow-up independently found to carry the floor behaviour at $T{=}0.7$). The expected error of an interval-step is then deficit $\times$ flow $= w$; the construction is an *equal-marginal-error* grid: each step is assigned an equal share of the total expected deficit mass (EDM-style equal-contribution grids are the continuous analogue).

A2 (convex-class optimality). Let $m_1, \ldots, m_K$ be the w-masses of a grid's K step-intervals, $\sum_k m_k = W$. Any placement objective of the form $\sum_k F(m_k)$ with convex F (total deficit $F{=}\mathrm{id}$ in expectation, max-step-error $F{=}\max$, error-variance $F{=}(\cdot - \bar{m})^2$ capture the canonical choices) satisfies by Jensen $\sum_k F(m_k) \geq K\,F(W/K)$. For strictly convex F equality holds iff all m_k are equal, so the equal-w quantile map is the unique minimizer across the strictly convex class; under linear F every grid attains the bound with equality, so no uniqueness follows; and even equal-mass uniqueness pins the *masses* only — positional uniqueness of the grid additionally requires the cumulative w to be strictly increasing (a measured, not assumed, condition of the profile). Conditional on the deficit-risk model the registered grid is thus the principled equal-mass placement for every strictly convex error functional, not a heuristic.

A3 (flatness bound). With flat-ratio $R = \max w / \min w$, the uniform grid's interval masses deviate from the equal-w optimum by at most a factor within R bucket-wise; smooth objectives see only a second-order gap $O((R-1)^2)$. At the measured $R = 1.564/1.562$ the second-order reading is a *reading*, not a bound: without a curvature constant and a mapping from objective value to gen-PPL, no universal few-percent cap follows from R alone; the few-percent scale is instead the measured one (the 1–2.8% wins with paired CIs and their saturation order, §P et seq.), which is the relevant evidence. Track-B's NULL stays cell-wise measured, not a bound over the convex objective class for other models or statistics.

Caveats. (i) Objectives outside the class (placements targeting per-interval higher moments, or weights built from other channels such as entropy) require their own profiles; we make no claim there. (ii) The deficit-risk scale is a model; A2 shows the construction is optimal relative to it, and A3 shows that even adversarial alternatives inside the class are flat-bounded at the measured R. (Weight families rebuilt from the frozen traces: entropy $\times$ flow gives $R{=}1.77/1.83$ (flat-adjacent, same few-percent scale), deficit$^2\times$ flow and the entropy $\times$ deficit product give $R{\approx}2.4$–2.8 (flat-bounded at a wider scale), and flow-only is trivially uniform ($R{=}1.07$); the registered deficit$\times$flow is the flattest informative family examined — `runs/_derived/da_wfam_flatness.json`.)

Tightened conditions, and what the flat-ratio does not promise. The optimality statement is exact and conditional: for convex F, equal mass is a minimizer of $\sum_j F(m_j)$; for *strictly* convex F any minimizer must satisfy $m_j = W/K$ for every j (Jensen's equality case), and with a strictly increasing cumulative-risk map the equal-w grid is therefore the *unique* minimizer; for merely linear F every grid with the same total risk W ties and uniqueness fails. If F is twice differentiable with $0 \leq F'' \leq M$ on the relevant range, writing $\mu = W/K$ and $\delta_j = m_j - \mu$, the linear terms of the Taylor expansion cancel and $0 \leq \sum_j F(m_j) - KF(\mu) \leq \frac{M}{2}\sum_j \delta_j^2$: small measured dispersion bounds the in-class headroom quadratically. The three claims are kept distinct throughout — the error-functional model itself, the measured mass dispersion, and the bound — and the flat ratio $\max w / \min w$ is *not* a generation-PPL percentage guarantee: it is a within-model reallocation bound whose empirical measurement is the win curve of §U.

I CONDITIONAL SAVINGS BOUNDS AND THEIR LIMITS

Setup (all inputs from the frozen DAX-1 bundle). Write $p_i(k)$ for sample i's pooled-mean confidence at horizon k, $\Delta_{i,k} = p_i(k{+}1) - p_i(k)$ for the $K{=}28$ increments over $k{=}3..30$ (post-$k{=}4$), per-sample increment volatility $\varepsilon_i = \mathrm{sd}_k(\Delta_{i,\cdot})$, and $\bar{\varepsilon} = \frac{1}{n}\sum_i \varepsilon_i = 0.024493$. The remaining rise is $S_i = \sum_k \Delta_{i,k}$ with cross-sample sd $\sigma_S = 0.129142$; the per-horizon cross-sample increment sd is $s_h(k)$ with $\sum_k s_h(k) = 0.529508$. The early–late persistence is $r = 0.303474$ (paired-bootstrap CI_{95} $[0.2187, 0.3856]$, $B{=}20000$, seed 20260905). The numbers below are computed from the traces by scripts in the supplementary archive: families M and S by `da/scripts/oracle_bounds_audit.py` (output `runs/_derived/r3_altitude_bound_audit.json`), and family D by `da/scripts/oracle_bounds.py` (output `runs/_derived/r3_altitude_bound.json`).

Family M: matched-mass slope cancellation. With cancellation-geometry factor $c{=}0.55$, $B_M(u) = u \cdot \sum_k s_h(k) \,/\, (2cK\bar{\varepsilon})$ as a function of a predictability multiplier u. Best estimate ($u{=}r$): $B_M = 0.213$. Positive-interval reading ($u{=}r$ CI floor 0.2187): $B_M = 0.154$. Additive ceiling ($u{=}1$): $B_M = 0.702$. The $u{=}1$ form is a sanity bound (perfect foresight), not an estimator.

Family S: stationary oracle (no independence assumption). Stopping at 1.96σ below the trajectory's final confidence prices only within-sample volatility against the residual rise: $B_S = rK\bar{\varepsilon}/(2 \cdot 1.96\sigma_S) = 0.411$, with the matched bootstrap CI $[0.397, 0.426]$ ($B{=}4000$, seed 7, resampling $\bar{\varepsilon}$) — reproducing the frozen values 0.41, $[0.397, 0.425]$ exactly.

Family D: deterministic settled-from exit. With cut = the q-quantile of $p_{\cdot}(3)$ and *settled-from* the first horizon after which the trace never falls below the cut, saved step-share is $(31 - t_{\mathrm{settled}})/28$: for $q \in \{0.10, 0.25, 0.50, 0.75\}$ the cuts are $\{0.0791, 0.0869, 0.0975, 0.1107\}$, settled-from medians are 3 at every q (median saved share 1.0, means $\{0.988, 0.986, 0.980, 0.964\}$), with 6–7 of 512 samples never settled. These exits pay in unpriced bulk emission (§O), so they are reported as hypothesis-free quantities only.

Audit of two hard-coded constants. An initial calculation of family M (`runs/_derived/r3_altitude_bound.json`, which contains two arithmetic slips) hard-codes $\varepsilon^*{=}0.01144$ in the denominator and $u{=}0.153$ in the positive-interval numerator. Neither constant can be recovered from the recorded traces: the nearest trace quantities are the early-horizon cross-sample increment sds (0.0111–0.0116), and the registered r CI floor is 0.2187, not 0.153. Family M scales as $B_M \propto u/\varepsilon^*$, so with these constants family M is inflated by $\times 2.14$ (best) and $\times 1.48$ (positive interval); the values printed above are the trace-auditable ones, derived from $\bar{\varepsilon}$ and the registered CI floor. Family S, which does not require the additional independence assumption, is unchanged (0.411, CI $[0.397, 0.425]$), and both corrected family-M point estimates ($0.213/0.154$) lie below it. These are limits on the specified trace-based proxies, not measured savings at preserved output quality.

J ADDITIONAL PROTOCOL DETAILS

J.1 SCOPE AND LIMITATIONS IN DETAIL

The main text summarizes the scope limitations; this subsection collects the supporting detail. (i) The public comparisons use LLaDA-8B-Base through a full-sequence sampler and retain no fixed-count ($\beta{=}0$) arm, so confidence weighting is not isolated from the deterministic count rule on the public model (Appendix D). (ii) The internal checkpoints cannot be redistributed; retained per-document scores support the reanalysis in Appendix F. (iii) Most uncertainty is over generated documents conditional on the two retained checkpoints; there is no between-pretraining-family estimate. (iv) Matched-diversity controls match one corpus statistic at a time and describe selected outputs, not a causal diversity effect (Appendix J). (v) The early-prediction channel is statistically real but practically insufficient for the registered efficiency bar; conditional savings bounds are collected in Appendix I.

The initial internal experiments use two 124M masked language models trained on OpenWebText with the MDLM objective: a baseline and a model trained with a different timestep-weighting schedule. They provide two fixed checkpoints for decoding comparisons, rather than evidence for a new training method. Appendix F specifies sampling and scoring; Appendix E maps the experiment identifiers to measurements. The following extensions compare diversity matching, public-model transfer, and budget dependence.

MATCHED-DIVERSITY CONTROLS (DAX-124)

Matching documents within distinct-2 windows in the 32-call paired corpus gives the following tilt-versus-ancestral differences. The primary window w=0.05 retains n=1749 pairs and gives $+10.57\%$ $[+8.08, +13.01]$; w=0.02 retains n=980 and gives $+9.28\%$ $[+6.00, +12.41]$; w=0.08 retains n=1986 and gives $+12.31\%$ $[+9.93, +14.62]$; all three intervals exclude zero. The pooled corpus distinct-2 values are 0.5568/0.5401 and geometric perplexities 615.20/534.62. The full-corpus $+13.098\%$ $[+10.69, +15.43]$ comparison uses the original ancestral reference; it does not hold realized commit counts fixed. This per-document distinct-2 comparison matches one statistic, rather than all aspects of diversity.

PUBLIC BUDGET ENDPOINTS (DAX-125)

LLaDA-8B-Base tilt budget response: 16 calls $+10.55\%$ $[+1.50, +18.74]$ (161.49 vs 180.54); 64 calls $+22.80\%$ $[+11.86, +32.24]$ (54.54 vs 70.66); both intervals exclude zero, and each half-sample estimate is positive. The existing 32-call estimate is $+11.91\%$. Calls are matched, while realized commit counts can differ between the ancestral and soft-order policies.

64-CALL HARDNESS SWEEP (DAX-126)

LLaDA-8B-Base β sweep at 64 calls (T=1.0, beta-threaded decoder, n=1024, seed 777): β=1 54.54; β=2 55.71 (-2.14% $[-10.94, +6.03]$ vs β=1, interval includes zero; halves $-5.72/+1.33$); β=0.5 57.24 (-4.94% $[-11.93, +1.65]$, interval includes zero; halves $-1.65/-8.34$); all soft-order configurations improve over the ancestral reference 70.66, with intervals above zero ($+21.15/+22.80/+18.99\%$). The internal $(1, 2]$-side shift toward harder ordering at 64 calls (DAX-111) does not replicate on the public 8B substrate at measured precision; β=1 stays the operating point at all three measured public budgets at T=1.0.

COOLED 32-CALL PANEL (DAX-127)

LLaDA-8B-Base β sweep at the cooled setting (T=0.7, 32 calls, n=1024, seed 777): β=2 27.35 beats β=1 30.71 $+10.96\%$ $[+2.25, +18.72]$ and β=1 beats β=0.5 34.43 $+10.79\%$ $[+3.58, +17.50]$, both with intervals above zero — the in-house cooler ordering (DAX-68: $26.25 < 29.58 < 31.91$) replicates on the public substrate through every measured β point, so the preferred tested hardness depends on temperature in both model families. The cooled soft-order-versus-ancestral intervals all include zero at n=1024 ($+4.26\%$ $[-22.03, +25.25]$ for β=1; $+14.75\%$ $[-9.31, +33.78]$ for β=2; -7.32% $[-37.57, +16.79]$ for β=0.5): the document-level distributions are bimodal ($\approx$4–5% catastrophic documents per arm at the corner versus $\approx$1% at T=1.0, per-sample mean generated-NLL > 10 nats; ordered-family collapse sets shared, pairwise minimum containment 0.70–0.85, uniform collapse set nearly disjoint, intersections ≤ 2 of 1024), and the resolved order-vs-uniform reads on this substrate remain the T=1.0 members. Same-arm T=1.0 $\rightarrow$ 0.7 deltas $+64.3\%/+67.1\%$, both with intervals above zero.

COOLED DEEP CORNER AT 64 (DAX-128)

LLaDA-8B-Base β sweep at the cooled 64-call setting (T=0.7, 64 calls, n=1024, seed 777): uniform 28.51; β=0.5 22.37; β=1 22.09; β=2 21.09. All three pairwise β intervals include zero (β=2 vs β=1 $+4.51\%$ $[-8.01, +15.52]$; β=1 vs β=0.5 $+1.26\%$ $[-8.01, +9.92]$; β=2 vs β=0.5 $+5.71\%$ $[-5.89, +15.81]$; all half-sample estimates favor harder ordering) — the ordering resolved at 32 calls (DAX-127) is unresolved at 64 calls (power sensitivity: at this cell's n=1024 single-seed pairing, adjacent-β effects below $\approx$ 13–17 percentage points would not reach 80% power, this

is a power calculation, not an upper confidence bound on the ordering margin), even though every point estimate stays hard-ward monotone and the perplexity values are close. Tilt-vs-uniform at depth resolves only at β=2 (+26.01% [+2.00, +43.91], interval above zero; β=1 +22.52% [−2.49, +41.12] and β=0.5 +21.53% [−2.48, +40.34], interval includes zero), with wide intervals also observed at the 32-call setting. The same recorded data also give the following budget and temperature contrasts: budget 64-vs-32 at T=0.7 (β=0.5 +35.02% [+15.50, +49.73] and β=1 +28.08% [+6.60, +44.50], both with intervals above zero; uniform +11.13% [−13.94, +30.45] and β=2 +22.87% [−0.41, +40.62], interval includes zero) and cooling T=0.7 vs T=1.0 at 64 calls (uniform +59.65% [+50.84, +66.60], β=0.5 +60.92% [+53.51, +66.99], β=1 +59.50% [+50.98, +66.04], β=2 +62.14% [+54.35, +68.32], all with intervals above zero).

CROSS-SCORER PUBLIC PANEL (DAX-129)

Re-scoring all twelve existing public configurations with gpt2-xl (fp32, the same text conversion and $n = 1024$ documents per configuration) preserves the measured directions: tilt +11.92% [+1.42, +21.33] (frozen +11.91% [+1.28, +21.43]), budget endpoints +10.05% [+1.32, +18.16] (16) and +23.34% [+12.82, +32.75] (64, frozen +22.80% [+11.86, +32.24]), the T=1.0@64 interior holding both flank intervals including zero (−1.04% [−9.53, +6.94], −4.87% [−11.62, +1.61]), and the cooled harder-better pairs +11.01% [+2.32, +18.99]/+12.50% [+5.28, +19.46] both with intervals above zero. Intervals that excluded zero with the original scorer still exclude zero. Absolute perplexities are lower under the larger scorer (uniform-32 93.93 vs 97.64; uniform-64 67.00 vs 70.66) with effect magnitudes preserved within $\sim$ 2pp.

PUBLIC-SUBSTRATE TRACE REPLICATION (DAX-130)

Applying the same early-trace statistics to LLaDA-8B-Base (n=1024, L=256, T=1.0, 32 steps, decode seed 777) does not meet the duration and relative-spread criteria, as in the internal models (duration 0.021, relative-spread ratio 1.026), with persistence elevated at 0.69 [0.65, 0.72] — individually past its chosen 0.5 component threshold but insufficient for the combined criterion — the same composite trace criterion is unmet on the public substrate; this does not exclude useful information in other trace statistics or stopping policies.

PUBLIC MULTI-ROUND REVISION FRONTIER (DAX-131)

The four-round revision configuration uses 16A+6C+5D+3E+2F = 36 forwards, remasks the lowest-64-$p_{\max}$ positions per round, and applies an exact 3-gram guard on D/E/F. It makes 0 guard overrides. On LLaDA-8B-Base ($n = 1024$, $L = 256$, $T = 1.0$, seed 777), its generated-text perplexity is 162.03. Relative differences are −65.94% [−85.24, −48.51] against the 32-call ancestral reference (97.64), −88.38% [−104.86, −73.11] against 32-call soft order (86.01), and +3.34% [+0.44, +6.08] against its two-round counterpart. Additional revision rounds improve this configuration, but it remains worse than the ancestral reference. The public-model failure therefore persists beyond the shallow setting.

Revision and repetition. The two guarded configurations have the lowest measured repetition in the $T = 1.0$ panel at 32 or more calls (0.00146/0.00128, versus 0.00230 without revision). Their failure therefore accompanies low repetition; zero overrides mean the guard did not intervene. Cooling instead gives repetition 0.037–0.069 at $T = 0.7$. Within each temperature/budget setting, lower generated-text perplexity co-varies with higher 4-gram repetition: $\rho = -0.8/-1.0/-1.0/-0.4$ at $T = 0.7$/32, $T = 0.7$/64, $T = 1.0$/32, and $T = 1.0$/64, respectively, with only 3–4 configurations per correlation. The 240-call entropy-bounded configuration is excluded from the final correlation because of its larger budget. These descriptive associations do not establish a causal source of quality changes. Public-model gains remain scorer-level observations with diversity reported; the matched-distinct-2 comparison is available only internally and matches one diversity statistic.

PUBLIC FRESH-SEED TILT ESTIMATES (DAX-132)

The same recipe at decoding seeds 778/779 ($n = 1024$ samples per seed and configuration, unchanged scorer) yields positive document-level intervals. Both point estimates fall within the seed-777

interval; this overlap is descriptive and does not estimate training-seed uncertainty. The measured differences are: K_{778} 15.32% $[5.94, 23.71]$, K_{779} 11.80% $[1.63, 20.87]$; coherence gap 3.52 pp $<$ the chosen estimator-level reference 7.42 pp. Both fresh-seed estimates fall within the original interval $[1.28, 21.43]$ at the seed-777 anchor 11.9%; the forecast score, specified before the run, of the observed outcome is 2.40 bits.

Public and internal budget-response contrasts

For the public model, the change in the tilt advantage between 64 and 32 calls is $+0.132$ nats $[-0.037, +0.303]$; between 32 and 16 calls it is $+0.015$ nats $[-0.130, 0.160]$ (DAX-136, 1,024 samples per configuration, seed 777). Both intervals include zero. They use 20,000 bootstrap replicates and one shared index stream with seed 20260908; cross-budget pairing is index-arbitrary and need not reproduce an unpaired variance. The corresponding internal 64-versus-32 contrast is $+0.025$ nats $[-0.015, 0.065]$ (DAX-139, 2,048 samples per configuration). These positive point estimates do not resolve an increasing advantage with budget. The reported 80%-power sensitivities are approximately 0.243 and 0.057 nats; they are not effect-size upper bounds. No additional contrast-powered replication is included in these results.

K Detailed results for the three decoding interventions

The following sections retain the exploratory comparisons behind the main paper. "Track A" denotes prediction from early confidence traces; "Track B" denotes a fixed grid derived from a checkpoint's average trace. A weak prediction result for the measured trace statistics does not rule out all adaptive stopping rules. The grid's conditional optimization property concerns its chosen proxy, whereas quality, commit-order effects, and revision outcomes are assessed empirically. The numerical tables above provide the compact comparison; the sections below give protocol details, controls, and unsuccessful variants.

L Trigger baselines and confidence ordering

The soft ordinal tilt versus sampled remasking. Our softest winning configuration (§S) sits in the remasking-policy family but differs from each of its members on one of three orthogonal axes — what is random, what is scheduled, and whether the budget is fixed. Fast-dLLM-style thresholded parallel decode (Wu et al., 2025) commits every position whose confidence clears a threshold each pass, so the per-sample number of passes varies with the trace; on this substrate, the measured failure mode concerns this budget-adaptive behavior (threshold triggers collapse to near sequential mass, DAX-18/22/28). The tilt removes the threshold: pass counts are the frozen uniform-K schedule, the budget axis is closed, and only *which* positions commit is perturbed. ReMDM-lineage schedulers (Wang et al., 2025) re-*mask* committed tokens, trading extra forwards for correction capacity; the tilt never re-masks, so its NFE accounting is exact. Closest is confidence-*ordered* commit, measured here to be catastrophic at parallel budgets (-52%, §S) and paying only near-sequentially; the tilt is precisely the randomized interior of that continuum — a without-replacement draw $\propto p_{\max}^{\beta}$ per pass — with a measured non-monotone hardness axis (interior at $\beta{=}1$, §S). EB-Sampler's budget (Ben-Hamu et al., 2025) is the one orthogonal trigger family that survives quality-wise here — and the arm measured in the same matched-budget cells (§S) is a *self-built* cumulative-$p_{\max}$ budget in that spirit, not the published per-step entropy-bounded set rule; the published rule was therefore also measured at its own call count (DAX-119): the published rule measured at $\gamma \in \{0.2, 0.5, 1.0\}$ nats is near-sequential by construction at all three tolerances (239.6–247.2 certified calls — at these budgets the model takes one token per step), beats the uniform quality–calls ladder by 33–36% at its own reach (the near-sequential confidence-ordered pole of the confidence-order map), outbids the 32-call tilt board $+31.37\%$ $[+28.88, +33.83]$ at $7.5\times$ the certified bill — unequal-computation, not a matched-computation comparison, and the 32-call EB arm is unbuilt within this lattice — and carries a 5.8–$7.1\times$ repetition signature at lower diversity ($\mathrm{rep}_4 \approx 1.5$–$1.8\%$): a gain-with-repetition-cost member at the family arbiter, not a clean-frontier point.

M TRACE VOCABULARY AND PROTOCOL

At each decode step k (of $K = 32$), for each sample i over masked positions we record: n_{masked}, the mean top-probability $p_{\max}$, the mean entropy (normalized), and the newly-unmasked count $f_{i,k}$. Difficulty proxies must be trace-intrinsic: no ground truth is available inside decode, which is precisely why a *predictive* notion of difficulty is the first thing to test. All decodes use the ancestral sampler, temperature 1.0, sequence length 256, on the OWT domain, with the generator seeding protocol fixed in the launch scripts (`runners/`) before any run.

N DAX-1: A MEASURED SHAPE-UNIFORMITY NULL

Registered gates (pre-dispatch). Shape-diverse difficulty would show as (V1) cross-horizon persistence: correlation ≥ 0.5 between per-sample $p_{\max}$ at step k=4 ($t \approx 0.875$) and k=28 ($t \approx 0.03$), *and* between k=4 $p_{\max}$ and steps-to-95%-decoded; and (V2) early-versus-late relative spread ratio ≥ 2.0. The NEG branch fires if any V1 correlation ≤ 0.2 or V2 ratio ≤ 1.2.

Measured. On n=512 samples of the neutral baseline checkpoint owt_mdlm_s0: V1 persistence $= 0.303$, V1 duration correlation $= -0.068$, V2 ratio $= 0.992$; relative spread 0.179 early vs. 0.181 late. Both NEG triggers fire.

Interpretation. Traces differ between samples in *altitude* (overall confidence level) but are statistically *uniform in shape*: the early trace carries only $R^2 \leq 0.15$ of a sample's late-horizon confidence (weak; insufficiency quantified in §O), and nothing of its remaining-difficulty duration — where the duration variable itself, steps-to-95%-decoded, is a bookkeeping time under the exogenous uniform-position count schedule, not a defined quality-readiness time; the V1-duration component is thereby a null on the declared schedule statistic. Under the uniform-position unmask order of this sampler, between-sample trace differences can only come from model belief (design fact registered pre-measurement); measured belief traces then certify that any sample-adaptive exit rule keyed to the measured statistics and registered gates on this family would act on noise realizations, not on structure. DAX-2 (a calibrated early-exit experiment) was pre-specified to run only if these gates passed, so it was not run.

O WHAT COULD THE ALTITUDE CHANNEL BUY? (INSUFFICIENCY, QUANTIFIED)

V1 persistence is *not* zero in disguise: r=0.303 (Fisher 95% CI $[0.22, 0.38]$, $R^2 \leq 0.15$). The registered NEG branch fired via the duration correlation ($-0.068 \leq 0.2$) and the spread ratio ($0.992 \leq 1.2$); the persistence component sits between the two registered thresholds. We quantify how much this remaining signal *could* save before assessing the channel's limitations; the full derivation and audit are in Appendix I, including an audit of two hard-coded constants in an initial calculation.

Oracle savings bounds. If at k=4 one could act on the predictable component of every later per-step increment, matched-mass cancellation against the mean remaining slope gives a best-estimate recoverable share of compute of ≈ 0.21 in the most optimistic multiplicative form (positive-interval ≈ 0.15 at the r CI floor); a *stationary* oracle that stops at 1.96σ below the trajectory's final confidence — a bound that makes *no* independence assumption — yields best estimate ≈ 0.41 (matched bootstrap CI $[0.397, 0.425]$). These are fractions of the *headroom between a smoothed trace and its envelope*, i.e. upper-order numbers under oracle-perfect exploitation. *(Precise reading:* the bounded quantity is the expected share of the 32 serial steps removable in step count under the oracle pricing — a steps-saved fraction of the NFE budget, not a fraction of wall-clock, of the trace, and not a quality-matched saving: the bound does not price the quality loss of replacing the remaining serial refinement by bulk parallel emission (deflator (i) of §O); it is *monotone* in R^2: the predictable share R^2 enters multiplicatively on top of the slope-cancellation envelope, so $R^2 \leq 0.15$ pins the number rather than decorating it; the only other inputs are the deterministic count-schedule slopes and the matched resample family, with no further fitted parameters.*)*

Step-0 information and direct quality-axis measurement (DAX-7). The step-0 channel is degenerate by mechanism, not by power: on all three frozen trace bundles (both checkpoints; T=0.7 and 1.0) every step-0-only pool statistic is constant across all 512 samples — unconditional decode starts each sample from the identical fully-masked sequence, the first forward pass is deterministic, and any step-0-only feature has zero sample variance (so its correlations are undefined). Altitude exploitation *before* decoding is therefore closed with positive evidence, not merely an absence of detectable correlation. Acting later and tying the trace directly to the quality arbiter (teacher-forced gen-NLL from a scorer bit-verified against the frozen GPU boards, max row diff 0.0): at the registered early horizon k=4, $\rho = -0.320$ ($[-0.397, -0.241]$, mdlm) and -0.230 ($[-0.324, -0.133]$, tge) — the exploitable early window carries 5–10% of final-quality variance, matching deflator (ii)'s magnitude using the quality target directly; the full-trace altitude reaches $\rho = -0.581/-0.580$ on the two checkpoints (CIs within $[-0.64, -0.51]$) and -0.659 at T=0.7, but it is unavailable before decoding, so it is an upper bound, not an operational decision rule. (registered verdict STEP0-NULL; `runs/_derived/da_dax7_verdict.json`.)

Why this cannot certify an exit rule. Three independent deflators compound: (i) every bound prices *steps saved*, not quality — exiting early replaces the remaining serial refinement by bulk parallel emission, whose generation-PPL cost is not trace-inferable and is exactly what the registered quality gates of the instantaneous family (Fast-dLLM et al.) are designed to measure; (ii) the exploitable part of the early signal is $R^2 \leq 0.15$ of between-sample variance — measured directly on the gen-NLL arbiter at the registered early horizon as squared correlations 0.053–0.103 on both checkpoints (DAX-7 above) — so realized (non-oracle) rules land far below the oracle numbers; (iii) DAX-2's registered efficiency bar ($\geq 20\%$ NFE saved at quality parity) is met only at the *additive* (0.70), multiplicative (0.21), or stationary (0.41) envelope — point estimates cross the bar in all three forms, but every one of them prices steps, not quality (i), and no identified estimator attains even the smallest. The statement the data support: *the early-predictive channel is statistically real and practically insufficient* — real enough that describing it as "consistent with zero" would be inaccurate, too small to certify any exit rule under registered quality gates. What remains viable is the *instantaneous* channel: a deterministic settled-from exit acting on the sample's own trailing confidence saves steps trivially (≈ 1.0 median step-share saved at per-sample quantile cuts) — but it pays for them in unpriced bulk emission, belongs to the literature's threshold-decoding family, and neither uses nor needs early predictability. This is precisely the channel split of §L: our null results apply to the early-predictive channel, not to the instantaneous one. For completeness on the dormant DAX-2: its registered firing condition required the V1 pair's positive gates *both* (≥ 0.5 persistence *and* ≥ 0.5 duration correlation) along with the spread ratio ≥ 2.0; under any registration whose trigger required the duration component, the measured -0.068 closes the question independently of the 0.303 persistence value. *(Retrospective:* no single-component rule was ever registered — the compound V1 gate reflected that the exit claim only makes sense when both exploitation channels exist; with hindsight the components are near-independent, but the joint condition does not *bake in* the null: each component on its own ($0.303 < 0.5$, $-0.068 < 0.5$) fails the component threshold, so *any* member of the single/compound family would have fired identically NEG here.*)*

No nonlinear headroom in what remains (DAX-75). The insufficiency above is quantified on linear functionals; the registered question whether *nonlinear* per-sample structure survives — enough to drive a decode-time per-sample order- or budget-selection rule — is measured directly on the n=512 trace board (32 trajectory features per sample against the final-quality target; 80/20 out-of-fold). Linear ridge already extracts everything the channel carries: oof R^2 0.5147 against gradient boosting's 0.4570 — the nonlinear probe is *worse* by -0.0576 $[-0.1046, -0.0069]$, excluding 0, so no nonlinear headroom exists to harvest (LINEAR-IS-ALL). Settling dynamics carry nothing either: per-pass settling-time and shape features predict the quality target at $R^2 \approx 0$ for both learners. The dominant marginal stays the late-mean confidence itself (Spearman 0.620 $[0.560, 0.674]$). A kNN mutual-information line registered as a fallback reads 0.302 but is disclosed as non-class evidence: its own bootstrap CI $[0.411, 0.669]$ *misses the point estimate* — duplicate-adjacency inflates the estimator beyond its uncertainty model, so the MI number is reported as an artifact, not a contradiction. The pre-specified expectation did not name the measured direction. Consequence: the altitude channel is closed for per-sample decode-time selection — its exploitable content is exactly the already-priced linear correlation, acting only *after* the decode, and nothing in settling shape remains on the table. The per-*position* version of the same question (DAX-80) closes identically: per-position early-late rank correlations fall *strictly below* the pooled statistic on all three substrates

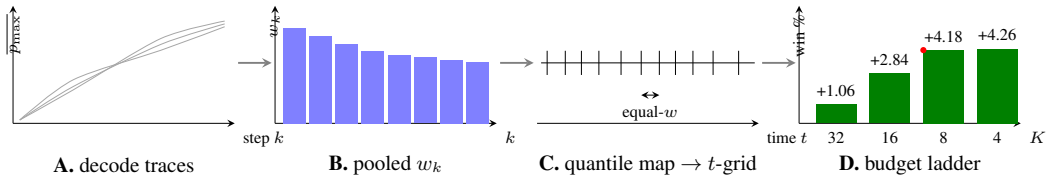

Figure 6: The operational protocol at a glance. **A**: ancestral decode traces on the n=512 vantage board; pooled per-step belief statistics. **B**: the deficit–flow weight profile $w_k = (1 - \overline{p_{\max\cdot,k}})\,\overline{f}_{\cdot,k}$ (measured flatness 1.564). **C**: the equal-mass quantile map of w yields the frozen descending time grid (construction before use, bit-reproducible); tick marks are grid times — nonuniform, denser in t where w is high (the early, high-t region where the deficit–flow profile peaks). **D**: the budget ladder $K \in \{4, 8, 16, 32\}$ re-decodes on the same grid family: measured paired win over uniform $+1.06\%$ (NFE=32, CI includes 0), $+2.84\%$ (16), $+4.18\%$ (8, knee marker), $+4.26\%$ (4) (§U, DAX-2/8/11).

(median 0.13–0.23 vs pooled 0.40–0.55; the $p90$ position sits 0.12–0.36 under pooling) — the between-sample signal lives at the sample level, pooling hides nothing — with exactly one measured exception, a width-3 prompt-adjacency band on the conditioned substrate (ρ 0.932/0.619/0.401 at the first three generated positions) whose mechanism is prompt-local determinism, disclosed as EXPLOITABLE-EXCESS localized rather than a hedge; the long-form L=512 tensor (fresh dump with an exact commit map, bitwise finals gate) shows no such band.

P TRACK B: BELIEF-PROFILED TIME GRID

The uniform grid spends equal decision mass per step, while the *average* belief profile need not be flat.

Construction (frozen before any decode). From a vantage trace board (n=512) of the target checkpoint, form

$$w_k \;=\; \big(1 - \overline{p_{\max\cdot,k}}\big) \cdot \big(\overline{f}_{\cdot,k}/256\big), \qquad k = 0, \ldots, 31,$$

the product of residual-uncertainty and flow. With $\mathrm{cw} = \mathrm{concat}\big(0,\, \mathrm{cumsum}(w)\big)$, the grid is the quantile map of w: grid-descending times $t_j = \Phi_w^{-1}(j/K)$, $j = 0, \ldots, 32$, i.e. 32 equal-w intervals. The construction is identical across checkpoints: the builder script asserts reproduction of the frozen grid artifact to its 6-decimal precision (measured max-abs deviation 4.7×10^{-7}, exactly the artifact's rounding). The frozen grid is also stable to the vantage board's size (DAX-20; `runs/_derived/da_dax20_verdict.json`): halving the board (512→ two 256-halves) moves the reconstructed boundaries by at most 0.128 interval at K=32 and 0.024 at K=8 (registered band 0.5), with half-profile flatness ratios within ± 0.043 of the full-board value (verdict GRID-STABLE); a cross-recipe grid (a TGE-objective checkpoint) differs from this one by at most 0.026 step — and the whole pipeline is shown at a glance in Figure 6.

Where the weight stands. Appendix H gives the principled position: under a calibrated deficit-risk model the equal-w map is the *unique minimizer of every strictly convex per-interval error functional* (Jensen; a linear functional ties all grids), so the construction is not an unmotivated heuristic; and the measured flatness (below) bounds any alternative placement's available gain inside that class at a few-percent scale. *The alternatives are then measured head-to-head (DAX-19), and the landscape is neither of the two framings the question anticipated* (paired n=2048, B= 20000, all grids frozen pre-decode; artifacts in `runs/_derived/da_trace/`): blind constructions are catastrophic — the monotone heuristic $(1-k/K)^2$ loses 31.2% to uniform-8, the EDM ρ-schedule adaptation 14.9%, two seeded random monotone controls 24.0%/32.8% — so the measured belief profile is genuinely load-bearing against generic placements; but within the working family, the equal-w map is *not* uniquely optimal: the deficit^2×flow weighting (steeper at flat-ratio 2.26) scores 874.61 ($+6.87\%$, $[+5.35, +8.33]$) and the entropy×flow weighting 885.98 ($+5.66\%$, $[+4.29, +7.03]$), both beating the frozen equal-w grid with CIs against it excluding 0 (874.61: base-over-alt -2.88, $[-4.34, -1.47]$). Verdict by the registered class map: ALT-BEATS-BASE. The reading that survives all three arms:

the mechanism is a *profile family* — the coarse belief-driven shape (mass pulled toward mid/late times) is necessary against blind placement, and within the family a somewhat steeper-than-equal-w weighting is better at this budget, consistent with the measured overshoot of the curvature model of Appendix H (DAX-21): the true low-NFE error functional is steeper than quadratic. The frozen grid's registered claims are unaffected (its wins vs uniform are evaluated exactly as pre-registered); what changes is the position that equal-w is the family's measured optimum — it is not, at $\mathrm{NFE}{=}8$ on this substrate — and whether the steep edge survives the ordinal tilt is then priced by DAX-24 (registered pre-measurement): the steep placement's advantage is *placement-only*. The deficit^2-count cross scores 848.60 — numerically the best cell at this budget — but is statistically tied with the equal-w cross (859.43; CI $[-2.81, +0.22]$ contains 0), while each composed cell beats its own tilt-free placement board outside CI (verdict CROSS-SATURATES-PLACEMENT). With the tilt active, the placement sub-choice saturates; the prescription's working point is the registered equal-w cross, with the steep cross an indistinguishable numerical variant. The saturation read is, however, budget-local: re-pricing the identical question at the 16-call node on the frozen β-plane (DAX-93, registered pre-measurement, zero new decoding) finds the *flat*-profile cross beaten by the measurement grid's cross outside the interval at the flagship dose — grid+tilt $\beta{=}1$ wins $+2.71\%$ $[+1.09, +4.33]$ over flat-order+tilt at the same 16 calls, $+2.18\%$ $[+0.57, +3.77]$ at $\beta{=}0.5$, unresolved at $\beta{=}2$ — and the absorption ratio is $\tau(1){=}+0.05$: the tilt does *not* absorb the placement value at 16 calls, the two channels are near-additive there, and the $K{=}8$ saturation read does not transfer along the count schedule (the pre-specified prediction of full saturation is refuted). The prescription's working point at 16 calls is therefore measurement-grid+tilt, not flat-order+tilt. The 32-call node is measured (DAX-108, registered pre-decode): grid+tilt on the frozen $K{=}32$ schedule beats uniform $+14.067\%$ $[+11.550, +16.489]$ (the tilt win's own scale) but its residual over flat+tilt at 32 calls reads $+1.116\%$ $[-0.867, +3.027]$ straddling 0 — the tilt *absorbs* the placement grid at 32: the interaction lattice closes as absorbs ($K{=}8$) / pays ($K{=}16$) / absorbs ($K{=}32$), and the prescription's 32-call working point stays flat-order+tilt exactly as frozen.

Low-parameter indexing of the steeper family (post-hoc analytic on the frozen vantages). Each measured profile family is well-indexed by a two-parameter gamma shape over step index: moment fits on the native $K{=}8$ vantage give shape $\alpha \in [2.23, 2.59]$ across the working members (deficit^2 2.23, entropy$\times$flow 2.46, equal-w 2.59), with deficit-squaring staying inside the family and moving α down (the peak shifts earlier). Reconstruction of each frozen $K{=}8$ grid from its fit deviates by at most $\approx$1 interval with the residual concentrated in the late tail, and cross-K derivation from the single 32-step profile carries the measured vantage caveat ($R(8)$ differs between the native vantage and the 32-step-chunked functional: the profile is vantage-dependent). The steeper grids are thereby systematically indexed by two parameters rather than tuned per family. The frozen curvature functional of Appendix H, however, does *not* order the steepening at the knee: with λ frozen it ranks deficit^2 $>$ base $>$ entropy$\times$flow while measurement ranks deficit^2 $>$ entropy$\times$flow $>$ base, and overshoots deficit^2 by $\approx 2\times$ (13.5% vs 6.9%) — consistent with the functional's own registered overshoot at $K{=}4$ (DAX-21): the low-NFE ordering of the steeper family is measured, not derivable from the quadratic mass-gap (post-hoc, single-substrate analysis).

Measured profiles. Both checkpoints' profiles are close to flat: flat-ratio $\max w / \min w$ is 1.564 (owt_mdlm_s0) and 1.562 (owt_tge_s0); maximal deviation of the grid from uniform is 1.83 and 1.86 steps respectively. This profile similarity is itself a cross-checkpoint instance of DAX-1's shape-uniformity: the average belief mass is spread almost evenly over the horizon, so the realistic prior for the grid's effect size is *small*.

Registered value gate. The grid is worth having iff, at equal $\mathrm{NFE} = 32$ and same decode seed, it reduces generation-PPL (frozen GPT-2-large scorer) by $\geq 5\%$ on *both* checkpoints, with distinct-2 drop ≤ 0.005; evaluation is split-half ($1024{+}1024$ fresh samples per arm). A null satisfies the line's closing criterion: NFE saturation at fixed NFE is then a property of the uniform schedule's near-optimal mass placement, not of missing adaptivity machinery.

Q INTERPRETATION OF THE GRID COMPARISON

We report two different things: the *registered gate fact* and the *budget-resolved measurement*. The registered gate was evaluated exactly as pre-registered (`da/scripts/trackb_verdict.py`

over the scored boards; `runs/_derived/da_trackb_verdict.json`): the grid wins $\geq 5\%$ on *neither* checkpoint at NFE=32; the registered gate verdict is **NULL**. The gate was a fixed worth-having bar for one deployment question — it is not a condition for the existence of the effect: measured below (§U), the grid's win at lower NFE is independently significant, monotone in $1/\mathrm{NFE}$ over four measured budgets (32→8) and saturating between NFE=8 and NFE=4, and never crosses the bar in point.

Reading the null. Three facts keep the null credible and informative. (i) *Direction without significance:* the grid arm is the lower-PPL arm in $4/4$ split-half pairs (sign test $p = 1/16$), so a positive but tiny reallocation effect is weakly suggested; its size (0.84–1.06%, i.e. 4.8–6.5 PPL) is the same order as split-half board noise (3.5–8.2 PPL). (ii) *Predicted small:* the measured flat-ratios (1.56) cap grid displacement at $\approx 6\%$ of the horizon, so $\mathcal{O}(1\%)$ effects are exactly what a second-order curvature account expects under near-flat profiles — the gate was registered as a worth-having bar, not as the predicted outcome. (iii) *Samples genuinely moved:* the paired boards differ in $2048/2048$ rows (59% of tokens), ruling out draw-stagnation artifacts. Combined with DAX-1 and Lemma 1, the correct conclusion is *not* "grids are useless" but: at fixed NFE=32 on this sampler family, there is no exploitable decision-mass placement error left — sample-adaptivity is floored (Lemma 1), and checkpoint-adaptivity finds the near-flat profile already uniform — on the quality arbiter in this family; positional selection rules (uniform coverage vs confidence-ranked selection, §S) and budget are the measured live parameters. Where the remaining quality lives at fixed positional coverage is therefore the budget question, taken up next: *within-protocol NFE saturation* (DAX-3), comparing reallocation's $\approx 1\%$ against what the same decodes gain from *more* steps rather than *replaced* steps.

R WHERE ADDITIONAL COMPUTE HELPS: BUDGET VERSUS PLACEMENT (DAX-3)

The Track-B null closes *placement* at NFE=32; it is silent on *budget*. DAX-3 (registered pre-dispatch with two firing readings and a descriptive band) measures, under the identical protocol and paired seed, the uniform-schedule ladder at $\mathrm{NFE} \in \{16, 32, 64\}$ against the already-measured comparators (uniform-32 $= 615.20$, grid-32 $= 608.70$).

Registered reading: P-resc fires. Uniform-64 scores 575.27, a +6.49% win over uniform-32 — 6.15× the grid's placement win, decisively above the registered 2× bar (2.11%); both per-half wins individually clear the bar ($570.53/580.05$ vs per-half comparators; $\geq 5.7\%$ each). Uniform-16 degrades by 16.3% (starvation anchor, halves $714.88/715.87$). The within-protocol conclusion is prescriptive: *at this sampler family's operating point, remaining generation quality lives in the step budget, not in mass placement* — the uniform schedule is already the right placement; the prescription is to spend or save steps directly. The exchange rate is convex-decaying: 2× steps buys 6.49% absolute (≈ 1.25 PPL per added step over $32 \to 64$; equivalently ≈ 0.20 percentage points per step on the relative scale) versus the placement's free 1.06%; diversity is untouched in both (Δdistinct-2 ≤ 0.003). How far the budget arm continues to pay (the saturation knee) is a separate budget.

Cost-normalized reading. Quality-per-FLOP re-orders the prescription only marginally. The placement win is essentially free (+1.06% at equal NFE; its one-time vantage-trace cost is amortized over all future decodes) while the budget win costs 2× steps (+6.49% at NFE 64). Normalized against *total* decode FLOP, budget still leads: the 32→64 arm yields +6.49% at 2.0× cost, i.e. 3.25% per 1.0× (log-win 0.0315), against placement's +1.06% at 1.0× (log-win 0.0105); normalized against *marginal* added FLOP, placement dominates overwhelmingly (its cost is ≈ 0). Deployment advice therefore splits by constraint: *latency/throughput-bound* deployments (step count is the binding constraint, NFE fixed) should keep the uniform placement and treat our grid's 1% as below the worth-having bar; *compute-ample* regimes buy quality by raising NFE directly. The call-bill abstraction itself carries to pinned wall-clock (DAX-96, registered pre-measurement, one GPU with all arms timed sequentially in-job at call-atomic granularity): per-forward wall is 1.0000 for the uniform-32, tilt-32, grid-16+tilt and one-shot revision arms alike (the tilt reorders commits and adds no forwards), and end-to-end walls track the exact call bills to $\leq 0.9\%$ ($0.5010/0.9994/1.0224$ measured against call ratios $0.5000/1.0000/1.03125$; in-job replication spread ≤ 1.06), so the one-shot revision's registered 33-call premium is a measured +2.24% wall premium at pinned batch

64 (these are hardware-pinned characterizations, not throughput claims). The guarded multi-round structure's serialization likewise enters *inside* the measured per-board walls: the 33–36-call guarded boards run 227–242 s end-to-end against 548 s for the 256-call uniform board they beat — wall ratios track call ratios with no round-trip premium beyond the call bill itself at pinned batch 64. The matched-diversity price of the revision-vs-tilt head-to-head (DAX-109) is a per-lane, scorer-side measurement under the same per-forward wall parity; whether the matched wall persists under practical batching is explicitly *not* claimed — a batch-parallel repair variant and its wall stand as a recorded candidate (not measured here). None of this alters the registered gate reading — the bar was registered in gen-PPL-at-equal-NFE terms precisely to separate placement value from budget value before comparing them.

S THE INSTANTANEOUS FAMILY AT MATCHED BUDGET (DAX-10/10B)

The instantaneous adaptive family is instantiated in-protocol here. We use a threshold decoder in the Fast-dLLM (Wu et al., 2025) style, implemented within the same decoding framework: per forward pass, unmask every still-masked position whose $p_{\max}$ clears τ, and always at least the single most confident remaining masked position per sample; loop until decode completes; per-sample call counts are recorded. Cells: $\tau \in \{0.85, 0.90, 0.95\}$, owt_mdlm_s0, T=1.0, n=2048, paired seed 777, the identical frozen scorer.

Measured: the threshold channel degenerates on this substrate. Per-sample model-call counts (certified by a bit-exact replay against the frozen boards; rows freeze once complete): means $249.1/250.5/252.4$ of L=256 calls, p_{90}=256, every batch running the full horizon batch-wide. The $p_{\max} \geq \tau$ candidate channel almost never fires at these thresholds on unconditional text at L=256 — parallel commit buys *no* model-call savings. At that budget the three cells score $365.95/366.45/367.16$ gen-PPL, and the τ-spread of $\approx$1.2 PPL is the measured size of the candidate channel itself.

At matched budget, order is the price (DAX-10b). The registered matched-FLOPs control is uniform-256: the same ancestral family run at NFE=256, i.e. random-order sequential decoding at 256 calls per sample, gen-PPL 548.09. Against this anchor the confidence-ordered cells (249–252 calls — infinitesimally *cheaper*, so the comparisons are conservative) reduce generation PPL by 33.23% (CI_{95} $[+30.41, +36.00]$; τ=0.90: 33.14%; τ=0.95: 33.01%; paired bootstrap B=20000): decoding in model-confidence order rather than random order lowers gen-PPL from 548.09 to 365.95 at essentially the same call count. The random-order budget slope over 64→256 calls is only +4.72% in point ($575.27 \rightarrow 548.09$): at matched budget, commit *order* outbids a fourfold random-order budget increase roughly sevenfold.

Reading, corrected by two registered cells (DAX-13/14). The channel statement above needs one regime correction, provided by two registered follow-up cells. *First (DAX-13): a cumulative-budget trigger* (per-position budget $b_i \leftarrow b_i + (p_{\max})_i$ after each forward; unmask all masked positions with $b_i \geq \beta$ plus a per-sample argmax fallback; a *self-built* cumulative-$p_{\max}$ budget in the spirit of EB-Sampler (Ben-Hamu et al., 2025), not the published per-step entropy-bounded set rule) is variable-rate — certified per-sample means $47.3/92.2/179.2$ calls for $\beta \in \{2, 4, 8\}$ — but runs *above* the uniform quality–calls relation: against log-linear ladder anchors it lands +241.6% (2023.5 vs 592.4 at 47), +100.2% (1137.1 vs 568.0 at 92), and +1.8% (565.2 vs 555.0 at 179; inside the disclosed ±5% descriptive band) — approaching the line from above, never below. (Convention: those are PPL ratios, cell/anchor−1; on the paper's win scale the anchors beat the cells by 70.7%/50.0%/1.8% respectively.) The in-band reading of the 179-call endpoint is anchor-model sensitive and has always been descriptive: re-anchoring the ladder globally (power law in log-log, log-quadratic) moves that cell to +6.9%/+4.5% (DAX-13b); the *family* verdict is anchor-robust — $\beta = 2$ lands +216%–248% and $\beta = 4$ +95%–109% above the curve under all three anchor models. The two *measured* trigger instances are thereby scoped on this substrate: neither call savings nor quality-per-call gains at practical budgets — what further trigger designs might do lies outside the measured set, and we claim only the measured one. *Second (DAX-14): the same confidence-ranked selection at the uniform-32 count schedule* (deterministic counts within the order comparison; the ancestral comparator has random counts — pass k commits $m_k = \lceil r_k/(33-k) \rceil$ positions in both decoders (8 at L=256); only the chosen subset differs) loses decisively: 1285.0 against uniform-32 615.2 and grid-32 608.7

(uniform and grid gen-PPL $52.1\%/52.6\%$ lower than hard order's, CIs $[50.8, 53.5]$/ $[51.2, 54.0]$ far from 0). The correction this forces: order and placement are *not* one channel at the grid's operating point. Confidence ORDER pays only near-sequentially (DAX-10b: per-token conditioning at $\approx$250 calls recovers an autoregressive-like contract), while at parallel-commit budgets *uniform positional coverage is the load-bearing property*, and the grid's real low-NFE win is a coarse time-placement layer on top of that coverage — not a cheap face of the order channel. (§R's ladder reading was drawn against placement alone; its practical prescription stands within that frame and is superseded at $\geq$32 calls by the next measured cell.) *Third (DAX-15): the finite-temperature softening of that same selection at the same deterministic 32-step counts* (β=1: within each pass the $m_k = \lceil r_k/(33-k) \rceil$ committed positions are drawn without replacement with probability $\propto p_{\max}$ instead of the hard top-m_k; tokens are still sampled from the model, never argmax) turns the DAX-14 result around: 534.62, i.e. $+13.10\%$ over uniform-32 (CI_{95} $[+10.66, +15.46]$) and $+12.17\%$ over the frozen grid-32 ($[+9.68, +14.59]$; paired n=2048, B=20000) — below even the 256-call uniform board (548.09) at one eighth of the calls. The registered class map (§W) returned SOFT-BEATS. The hardness continuum is thereby measured *non-monotone*: ancestral 615.20, fixed-count β=0 606.10 $\rightarrow$ soft (β=1) 534.62 $\rightarrow$ hard ($\beta\rightarrow\infty$) 1285.04, an interior optimum rather than a compromise; and it costs diversity — distinct-2 0.5401 vs uniform's 0.5568 (-0.0167; distinct-3 -0.0127), outside our usual ± 0.005 guardrail band and disclosed as part of the cell. And the interiority itself holds strictly only at the flagship budget: at NFE=64 the hardness top flattens (DAX-107, registered pre-decode): β=1 still beats the soft flank $+5.959\%$ $[+4.083, +7.819]$ outside the interval, but β=1 vs β=2 reads -1.288% $[-3.597, +0.964]$ straddling 0 — the strict peak lives at $K \leq 32$, while at 64 the registered refinement of that bracket (DAX-111, pre-decode, $\beta \in \{0.75, 1.5\}$ at T=1.0 and $\beta \in \{0.5, 2\}$ at T=0.7) reads the flat reading as *bracket resolution* of a measured migration: β=1.5 beats β=1 by 3.496% $[1.348, 5.649]$ outCI while β=0.75 loses by 1.959% $[0.240, 3.653]$ (five-point measured lattice at T=1.0: 1.5<2<1<0.75<0.5; the vertex lies inside $(1, 2]$ — further bracketed to its 1.5-side: DAX-112's pre-decode-registered single interior probe β=1.75 reads 469.75, beating β=2 by 2.425% $[0.459, 4.362]$ outCI but statistically tied with β=1.5 ($+0.298\%$ $[-1.763, +2.282]$; VERTEX-NEAR-1.5-SIDE)), and the cooling-side hardness tendency measured at 32 replicates at depth: β=2 beats β=1 by 16.019% $[13.881, 18.167]$ at T=0.7 and 64 calls (EDGE-MIGRATES-AT-64 + MIGRATION-PERSISTS-AT-64-T07). The joint hardness sentence is thereby cell-wise: soft-interior strict at $\leq$32 calls at T=1.0; hard-ward drift at 64, sharpening under cooling.

Plackett–Luce lineage (naming the construction). Drawn without replacement *proportionally to* a score, the commit selection is exactly Plackett–Luce ranking (Plackett, 1975; Luce, 1959) with confidence keys — on the *position-selection* axis (which masked position commits next); each committed token is still drawn by the model's own ancestral draw on the token axis. What we contribute is not the sampling scheme: it is (i) the matched-computation measurement of the commit-order lever, recast by the Gumbel-k identity below as the noise-temperature axis of already-standard low-confidence remasking practice (Nie et al., 2025; Wang et al., 2025; Wu et al., 2025); (ii) the measured *interior* optimum β=1 on the softness continuum of that axis, with its measured temperature dependence (the optimum migrates to the hard edge at T=0.7, §W); and (iii) the regime map of the construction (substrates, lengths, objectives, seeds, budgets, composition with placement and revision). The registered cross that prices the prescription (DAX-23, class maps fixed pre-measurement) is measured, and the channels *compose*: at NFE=8 the crossed cells order strictly, every registered pairwise CI excluding 0 (paired n=2048, B=20000): uniform-8 counts 939.11 > grid-8 counts 899.83 > uniform-8 counts + tilt 874.48 > grid-8-expectation counts + tilt (gridsoft) 859.43 — a $+8.49\%$ win over the uniform-8 board, nearly double the placement-only $+4.18\%$, with the tilt layer adding $+4.5\%$-scale over placement-only and the placement layer adding $+1.7\%$-scale over tilt-only (verdict COMPOSITION-ADDS; the grid-expectation count schedule $[26, 27, 29, 31, 33, 35, 37, 38]$ is the deterministic mirror of the frozen grid's ancestral rule, derived and frozen before decode). And the β sweep at uniform-32 counts prices the continuum: 557.93 (β=0.5) > 534.62 (β=1) < 556.22 (β=2), β=1 below both shoulders with CIs excluding 0 and the shoulders statistically tied — the interior optimum among $\{0.5, 1, 2\}$ sits at the measured β=1 (verdict INTERIOR-MAX-AT-1). The structure is K-dependent at the knee (DAX-27/27b): at NFE=16 both count families are INTERIOR-STABLE-AT-1 (β=1 644.26/626.82 beats both shoulders with CIs excluding 0; the β=2 penalty sharpens to 64.6/73.4 PPL), while at NFE=8 the β plane flattens into a $\{0.5, 1\}$ tie at the top in both families (FLAT-OR-RESIDUAL; tied gaps $[-2.01, +0.49]$ / $[-0.41, +2.13]$ points), and the hard shoulder is excluded at every measured resolution (β=2 worse outCI at NFE $\in \{8, 16, 32\}$). The diversity cost of the tilt layer is systematic and disclosed: distinct-2 $-0.0164/-0.0165$ for the two

tilt cells at 8 calls (-0.0048 for the grid alone), monotone in β at 32 calls ($0.5453/0.5401/0.5388$ for $\beta \in \{0.5, 1, 2\}$). The distributional and positional arms price that cost further (DAX-25): under official-recipe MAUVE (last-token GPT-2-large features, k-means k=409, 25-point divergence curve, $\exp(-5\mathrm{KL})$, symmetrized AUC; reference the frozen first-2048 OWT val eval windows) the tilt cell scores numerically *better* than its anchor (0.0730 vs 0.0621, halves bounded) and the composed cell within the registered 0.02 tolerance band of its own anchor (0.0315 vs 0.0423) — TILT-CONFIRMED-DISTRIBUTIONAL for both flagship cells — while per-token gen-NLL decomposition places the win's mass in the low-entropy deciles, *not* the top (top-decile share of positive gain $0.041/0.086$, band GAIN-SPREAD), with gains spread over thirds of the sequence and distinct-3/4 costs shallower than distinct-2 (distinct-3/4 $-0.0128/-0.0051$ for softorder-32; intra-sample 4-gram repeat $\leq 2.5\times10^{-3}$). Extending the same official recipe to the full flagship set (DAX-33; identical featurizer, clustering rule, and reference) confirms all six non-catastrophic pairs: softorder-16 (0.0687 vs uniform-16 0.0531), grid-16 (0.0648), gridsoft-16 (0.0558), grid-8 (0.0494), and both flagship cells above are CONFIRMED against their uniform anchors, with half-cell splits bounded; the single PPL-ONLY pair is the known catastrophic-order anchor (hard order-32: 0.0183 vs 0.0621), whose distributional collapse matches its -52% gen-PPL severity class — the registered roll-up is MIXED-ARB exactly because the second arbiter reproduces the hard-order null rather than rescuing it. The measured prescription at this family's operating point is therefore the registered cross, not one layer: *spend intervals where the belief-profile places them, and let a soft, sampled ordinal tilt choose which masked positions commit — with the tilt's hardness itself a measured interior optimum.*

The Gumbel-top-k identity (naming the closest practice). The Plackett–Luce family on any axis admits the standard race construction (Yellott, 1977): drawing without replacement with selection scores proportional to s_i^β is *operationally identical* to adding i.i.d. Gumbel noise to $\beta \log s_i$ and committing the descending top-m_k per pass. On the position axis with confidence keys, the tilt is therefore exactly *Gumbel-noised confidence ordering* at noise inverse-temperature β — the remasking form used in open MDLM/LLaDA-style decoding practice — and the closest existing default (unit-temperature Gumbel-noised ordering) is operationally the β=1 arm and is already measured on every quality board in this paper ($+13.10/+11.04/+11.28\%$ over uniform-32 at decode seeds 777/778/779, all intervals excluding 0, matched calls). The β-plane above ($\beta \in \{0, \frac{1}{2}, 1, 2, \mathrm{hard}\}$) is exactly a noise-temperature sweep of that practice. We therefore state the contribution as *measurement*, without a hierarchy-of-models claim: unit noise temperature is the interior optimum on this family; the win localizes to context-poor commit strata; and the regime map prices where the lever pays (its public-checkpoint transfer stays the declared asset boundary).

The interior optimum is temperature-confined (DAX-68). The measured three-point β surface at T=0.7 (uniform-32 counts, frozen producers; the T=1.0 twins frozen) is monotone *harder-better* through every measured point: β=0.5 $31.91 > \beta$=1 $29.58 > \beta$=2 26.25, every adjacent gap outside CI ($+7.31\%$ $[6.11, 8.48]$ for β=1 over β=0.5; β=2 below β=1 by 11.23% $[9.81, 12.63]$ (anchor-direction form of the registered -12.65% $[-14.46, -10.87]$ key)), the uniform T=0.7 anchor at 35.82 — the measured optimum migrates from the interior at T=1.0 (DAX-23) to the hard measured edge at T=0.7: β=1 is not interior at T=0.7 in this lattice. The registration's own declared shift-softer expectation (sharpened $p_{\max}$ hardens any fixed β) is refuted verbatim — the counter-channel dominates at the cooler point: sharper landscapes raise the value of order itself. Temperature dwarfs every decode-time dial measured in this paper ($\approx +94\%$ per cell, nearly β-independent; a T=0.7 *uniform* board beats the T=1.0 flagship by 93.3%) and carries the largest diversity cost measured (distinct-2 $0.54 \to 0.16$–0.18); tilt wins persist multiplicatively on the sharpened baseline ($+10.93\%/+17.44\%/+26.71\%$ over the T=0.7 anchor, monotone in β). The β=1 prescription stands only at its measured temperature.

Cool or tilt? The full surface (DAX-70/76). Both directions of the temperature dial are measured under the same protocol. *Cooling downward* (DAX-76, 4×2 lattice $T \in \{0.7, 0.8, 0.9, 1.0\} \times \{$uniform, tilt$\}$, canonical 32-call counts): within every measured temperature the tilt beats its uniform twin excluding 0 — $+17.44\%$ $[15.68, 19.15]$, $+15.90\%$ $[14.61, 17.14]$, $+13.75\%$ $[12.28, 15.21]$, $+13.10\%$ $[10.68, 15.41]$ at T=0.7/0.8/0.9/1.0, the margin decaying monotonically as cooling does more of the selectivity work — and no lateral cross-temperature pair is ever matched on diversity ($|\Delta \mathrm{d2}| \geq 0.08$ for every candidate pair, reported), so cooling cannot be priced against the tilt at fixed diversity anywhere on this lattice: TILT-DOMINATES-PARETO. The matched-diversity price is measured to its

terminal state — and it is a *disclosure*, not a price (DAX-90/97/101, all registered pre-decode, all consumed). The first ladder (DAX-90, n=512 probes) exhausted its extension on both cells with a measured near-T=1 diversity hump (cooling first *raises* distinct-2). The continuation (DAX-97) then failed its own calibration gates on the decoded boards: the n=512 probes hit the ± 0.005 band around the frozen anchor's d2 $= 0.5401$ at first shot (p=0.89 0.53914, T=0.91 0.53986), but the n=2048 boards read $0.4103/0.4216$ — *pooled distinct-2 saturates with corpus size on clipped-support arms*, so an n=512 probe cannot calibrate an n=2048-anchored gate (the implicit transfer premise is refuted; full-support arms are n-invariant at these sizes, e.g. uniform 0.5568 at both). The corrected re-read (DAX-101, probe-is-the-board at n=2048) maps the declared knobs' ladders at the anchor's own sample size in full: nucleus $0.4103/0.4582/0.4846/0.5176/0.5568$ at p=$0.89/0.93/0.95/0.9747/1.0$, cooling $0.4216/0.4521/0.4819/0.5193/0.5568$ at T=$0.91/0.93/0.95/0.9747/1.0$ — the matched band $[0.5351, 0.5451]$ exists *only inside the* $(0.9747, 1.0)$ *endpoint sliver on both cells*, where a 0.025 knob interval spans $\sim$0.04 d2 of the total dynamic range; at the endpoint the two knobs are near-equivalent diversity valves (0.5176 vs 0.5193 at 0.9747). Both cells exhaust their registered single extensions (NO-PRICE-AT-LADDER-EXHAUSTION): neither declared knob recovers the tilt's headline at the anchor diversity, and neither can be *held* there with resolvable precision at this sample size — the diversity-cost disclosure (-0.0167 distinct-2) remains the appropriate pairing, and ladder-wise matched-diversity pricing is terminated as an axis question with the landmarks above (disclosed; calibration of any future matched read must be performed at the comparison sample size itself). The revision side of the same attempt fares no better at this registered ladder resolution (DAX-106, registered pre-decode): nucleus-on-softord-order probes at the frozen revision's *own* diversity (0.468137) read d2(0.9375)=0.4462 and, after the one registered geometric-midpoint extension, d2(0.9437)=0.4562 — both miss the ± 0.010 board gate, the uniform-order ladder readings do not calibrate the softord-order surface (order composition shifts the nucleus support), and the registered key (revision vs matched tilt at exact 33×33) stays unmeasured: CALIBRATION-MISS, with the terrain named and any retry requiring a redesigned knob surface under a new registration. The measured prescription is *tilt, don't cool, at every measured diversity*; cooling enters only to reach a diversity band, and the tilt works inside it. *Trigger-harvest downward* (DAX-70): lowering the instantaneous bar τ through $0.50/0.65/0.80$ (against the conservative matched-call anchor — the smallest frozen uniform point whose calls cover each arm) pays monotonically in quality ($416.37 \to 381.16 \to 369.74$, wins $24.03/30.46/32.54\%$ over the anchor family, all excluding 0; the full six-point τ ladder 0.50–0.95 is monotone in calls with the knee at 0.80–0.85 and inversion at $0.90+$), yet the entire family sits in a 238–252-call band — no downward rung buys the call savings the trigger channel was invented for — and the best classical cell (365.95 at 249 calls) is itself dominated by the composed recipe on both axes at once (reopen-24A8C 303.91 at 33 calls, -16.95%; guarded-4 242.44 at 36, -33.75%; the full trade-off frame): TAU-DOWNWARD-PAYS, with the domination frontier stated verbatim rather than the ladder alone. The channel map, updated by the conditional replica (DAX-29), the placement-slope resolution (DAX-32), and the guardrail trade curve (DAX-31), reads: *placement* (free, small, low-NFE), *soft order* (large, parallel-budget-compatible, diversity-costly; measured to compose with placement, hardness interior-peaked), *hard order* (sequential-limit only), *triggers* (neither).

Robustness across seeds (DAX-34/35). The two single-seed disclosures close on both axes. Across three independently trained same-recipe wikitext checkpoints, all three channels replicate with $9/9$ channel intervals excluding 0 (tilt $+12.35/11.45/12.17$; composition $+9.11$ to $+10.19$ over uniform-8 and $+6.36/+4.97/+7.00$ over grid-8; placement $+2.93/4.31/3.43$ — FULL-SEED-REPLICATES), grids re-derived per seed and near-identical a priori; absolute level varies with checkpoint quality but the ordered channel structure is invariant. Across three independent base decode seeds (777/778/779) every flagged win replicates — $12/12$ intervals excluding 0, zero reversals — while the point estimates disperse: the flagship pair spans 2.06 (13.10 at seed 777; $11.04/11.28$ at the fresh seeds, which agree to 0.24), so the registered class is SEED-DISPERSION-WIDE and the seed-777 estimate is the high end of the seed distribution (cross-seed center $\sim +11.8$); fresh-seed vertical shifts mostly cancel inside the gaps. A fourth, previously unused decode seed replicates the claim set whole (DAX-105, registered pre-decode): at seed 780 all five registered load-bearing keys replicate — tilt $+11.709\%$ $[+10.042, +13.333]$, guarded revision $+33.661\%$ $[+32.234, +35.056]$, cooled revision $+12.298\%$ $[+10.705, +13.863]$, the cooled endpoint tie -0.325% $[-2.538, +1.860]$ in its registered tie form, and the cooled telescope $+12.013\%$ $[+9.976, +13.995]$; the four-seed tilt mean is $+11.78\%$, and we commit the cross-seed center — not the high single-seed anchor — as the headline quantity.

Budget response and the mechanism of the interior (DAX-36 to DAX-43). The tilt does *not* saturate with budget where placement does: at NFE=64 the tilt wins $+15.24\%$ $[12.73, 17.69]$ over uniform-64 (up from $+13.10$ at 32), and softorder-64's 487.63 undercuts the uniform-256 anchor (548.09) — better quality than uniform at $4\times$ the calls — the order/placement budget responses are measured as different channels. What sets the interior hardness is bounded four ways: the exploration-residue collapse point β^* rises with commit-group resolution on both the frozen trajectory ($2.27/2.41/2.47$ at $8/16/32$-effective) and on real independent traces at $K{=}8/16/32$ ($1.45/1.66/2.47$, intervals pairwise disjoint): the softer-interior-at-coarse-K direction is twice measured; a constant-threshold calibration is refuted (DAX-41), and a one-parameter resolution-coupled law fits the surrogate frame but fails on real trajectories (DAX-42) — the magnitude law is trajectory-family-specific. The tilt's action, by commit-order dumps, is a mid-horizon selectivity hump that is fraction-locked across K (centers $0.50/0.38/0.47$ of the horizon; Spearman with the steep weight profile 0.25 — PARTIAL-ALIGNMENT: the tilt does not implicitly re-derive steep placement), and the trigger channel's per-pass clearing counts are burst-structured (pooled var/mean 10.6–14.5), not Poisson — a semi-AR re-opening inherits a fraction-indexed burst hazard. The diversity guardrail is sharper too: gating the tilt to the low-entropy half recovers 85% of the flagship gain but does not reduce the distinct-2 cost (0.5377 vs the 0.5484 half-cost bar) — the cost is intrinsic to early confident commitment, not a byproduct of ordering high-entropy positions early. The vertex question left open by the growing 64-call margin is measured (DAX-74): the tilt keeps paying past 64 — softorder-128 472.55 ($+3.09\%$ $[0.10, 6.03]$ over softorder-64, excluding 0) and softorder-256 460.03 ($+5.66\%$ $[2.58, 8.61]$ over the 64 point; $+16.07\%$ $[13.30, 18.80]$ over uniform-256 at matched bills), distinct-2 flat at ≈ 0.534 throughout — VERTEX-BEYOND-64, GAP-PERSISTS. The registration's VERTEX-AT-64 expectation is refuted, and the mismatch is reported; the answer to "why does softorder-64 beat uniform-256" follows: the order channel pays at every budget measured, so a budget-matched uniform board never closes the gap — the tilt's deficit toward its own ceiling is a slope, not a wall. And the tilt is not an $L{=}256$ phenomenon (DAX-81, the second length measured): on the second-length checkpoint family the tilt wins $+10.98\%$ $[9.77, 12.18]$ at 32 calls (355.93 vs 399.82, excluding 0; inside its registered prior band; distinct-2 profile signing the same cost), machinery identical through an in-job bitwise head-check — TILT-TRANSFERS-L512. The transfer set for the free channels reads: second dataset, second recipe, second length, conditioned suffix, three training seeds, three decode seeds — and the model-class axis is measured at its first public member (DAX-122 on LLaDA-8B-Base: the tilt transfers at band scale at the 32-call point; the revision inverts; the published EB rule stays near-sequential), with the left members of that axis (64-call migration, cooling, block/semi-AR compositions) remaining named open cells.

Commit revision and the iterated re-open family (DAX-44 to DAX-55). The trigger channel stays closed ancestrally — harvesting inside the measured fraction band at a fixed uniform-32 budget loses -2.98% $[-4.19, -1.79]$ to the pure tilt (while still beating uniform, $+10.51\%$), pairing with the instantaneous degeneration (DAX-10/22) to close the channel twice, with the residual priced as structured burst hazard (DAX-37/43), and — after a third registered move — a third time. Burst-cap shaping of the trigger's own clearing bursts (DAX-67, registered pre-measurement): a fork of the threshold decoder keeps only the top-c per-pass candidates by $p_{\max}$ ($\tau{=}0.85$, the channel's maximal-overdispersion anchor; $c{=}2$ primary, $c{=}4$, bracketing DAX-37's measured burst-band sizes), with cap$\geq L$ certified as the decoder's identity by a bit-exact replay against the frozen $\tau 085$ board (the registered machinery gate). Everything runs at matched average calls *by construction*: the channel commits 1.028 per pass uncapped, so a cap can only push the mean from 249.1 toward the hard 256 horizon ($+2.8\%$ worst-case; realized $+0.91\%/+0.37\%$). The burst tail is independently removed — per-sample call minima lift $6 \to 129$ at $c{=}2$ (the $L/2$ floor to within one pass) and $6 \to 66$ at $c{=}4$ — yet quality does not move at all: $+0.077\%$ $[-1.412, +1.541]$ for the primary cap and $+0.141\%$ $[-0.744, +0.972]$ for the second, both CIs including 0 (paired bootstrap $B{=}20000$); the order channel is preserved verbatim ($+33.28\%/+33.33\%$ against uniform-256, the parent's $+33.23\%$ to rounding) and distinct-2 unmoved ($0.51525/0.51548$ vs the parent's 0.51522). The min-6 burst samples were never the residual poison. The semi-AR direction those closures left open is measured head-on as *commit revision*: complete a decode, re-score every commit, and re-decode the lowest-$p_{\max}$ ones in place. One quartile at 33 executed forwards (16 scaffold $+1$ re-target $+16$ repair) beats the tilt anchor by $+32.10\%$ $[30.66, 33.53]$ (363.00 vs 534.62; $+43.66\%$ over its own 16-call scaffold twin), replicating on a fresh decode seed ($+33.22\%$ $[31.83, 34.61]$), and the gain is *entirely* confidence-targeted: a uniform-random quartile *destroys* value (-18.09% $[-21.24, -14.96]$ — random revision merely returns to 16-call quality at a 33-call bill). The revision dose is a unimodal knob with a

measured optimum, $443.99 \to 363.00 \to 323.43 \to 311.52 \to 346.95$ at $12.5/25/37.5/50/75\%$ (every link excluding 0; channel open at every fraction; the curve must turn before a full re-mask, which would discard the scaffold entirely). Pass allocation is an independent axis: front-loading the scaffold ($24{+}8$) improves the balanced split by $+8.60\%$ (outCI) while back-loading collapses — revision repairs only the worst commits, so a weak scaffold caps the gain. The axes compose unsaturated ($24\text{A}{+}8\text{C}$ at 50% dose: champion 303.91, $+43.15\%$ $[42.04, 44.27]$; seed-replicated 303.14, $+43.83\%$), and *iterated* revision is the largest measured move on the line: two rounds at 32 executed forwards ($16{+}1{+}8{+}1{+}6$) reach 227.69 — $+37.27\%$ over the scaffold-matched single round, $+25.08\%$ over the 24A champion, $+57.41\%$ over the anchor, all excluding 0. Bills are stated per protocol: every single-round arm pays 33 forwards (the re-target forward is part of the bill), the iterated champion pays 32. Costs travel as measured dials, disclosed with each win: the iterated round moves distinct-2 to 0.397 and distinct-3 to 0.747 (single-round band 0.44–0.50 / 0.78–0.86) — iterated lowest-$p_{\max}$ repair steers commits toward high-confidence tokens; the diversity dial, not the scorer, is what a third round must negotiate.

A corpus-diversity-matched comparator for one-shot revision. One-shot revision has corpus distinct-2 0.468137, versus 0.5364 for soft order at 33 calls. A two-probe nucleus-sampling calibration first missed the desired ± 0.010 distinct-2 window. A subsequent probe at $p = 0.951$ gives distinct-2 0.465095, a difference of 0.003042 from revision, and scorer perplexity 294.144 versus revision's 363.0. At the same 33 calls, revision's signed perplexity reduction relative to this baseline is -23.409% $[-26.669, -20.230]$. The calibrated baseline's reduction relative to uniform-32 is 52.187% $[50.831, 53.496]$. MAUVE is 0.1522 versus revision's 0.0769. This comparison changes nucleus sampling and approximately matches one corpus-level diversity statistic. It shows that the original revision advantage is not retained against this calibrated comparator; it does not prove that diversity alone causally explains the original gain. It is distinct from the per-document subset matching used for the soft-order versus ancestral-uniform comparison.

What the match actually priced (DAX-113). The DAX-109 match is a registered match on board-pooled distinct-2, and distinct-2 is a weak diversity proxy: does the match co-price the rest of the standard family? The registered zero-decode read scores six token-level notions per document on the matched pair's frozen boards (distinct-$1..4$, per-doc unigram entropy, intra-doc 4-gram repeat mass) under a shared-stream paired bootstrap ($B{=}20000$), with the pooled anchors re-produced bitwise. Every listed notion separates: per-doc, the matched-tilt board reads lower than the revision on distinct-$1..4$ ($K_{d2}{=}-0.0062$ $[-0.0078, -0.0045]$) and on unigram entropy (-0.039 bits $[-0.048, -0.030]$), and *higher* on intra-doc repeat mass ($+0.0026$ $[+0.0020, +0.0032]$, $+64.8\%$ relative) — MATCH-IS-D2-SPECIFIC. Even the pooled level is d2-specific: the pool-level distinct-$3/4$ gaps lean the same way. The match therefore priced exactly one notion (board-pooled distinct-2), and the revision's enumerated residual ground — a small per-doc diversity edge on every richness notion at a lower repetition rate — is the named substance behind the diversity-consequence reading: the question which diversity notion the match actually prices is answered with the enumerated gaps.

The flip survives the distributional arbiter (DAX-114). DIVERSITY-WAS-THE-EDGE was priced by the scorer-NLL arbiter; does it survive MAUVE at the registered operating point — or does the nucleus truncation itself cost distributional ground the revision keeps? The registered leg featurizes the DAX-109 matched-tilt board with the frozen gpt2-large recipe (the bit-identical cache-builder path; the det gate first re-produces the frozen softord32 MAUVE 0.0730451102518158 to print precision, so the class inequalities below are exact-real) and scores against the frozen validation reference under the registered k-rule ($k{=}409$ full): matched-tilt MAUVE $= 0.1522$ against the revision's frozen 0.076974 and the un-matched parent tilt's frozen 0.073045 — FLIP-SURVIVES-MAUVE. The gap ($+0.075$) is $19\times$ the frozen tilt-vs-revision separation, and the nucleus calibration itself *raised* the parent tilt's MAUVE two-fold, placing the matched board in the revision-round family range (0.1546). Both measured arbiters agree at matched diversity: the revision keeps no ground on either — scorer-NLL -23.41% (DAX-109) and MAUVE $0.0769 \to 0.1522$ tilt-favored (DAX-114) at their registered operating points. The block-cosine tail leg (within-board feature-space self-similarity tails on the frozen feature caches) finds the tilt's nearest tail partner to be the revision itself ($+0.003$ to $+0.011$ across quantiles): feature-space tails do not re-rank the pair either. The named boundary items stand verbatim: a modern-featurizer MAUVE variant and the power-resolved larger-n MAUVE

for the iterated ladder remain recorded candidates, for which only assumption-conditional projections are available.

The flagship numerator's own co-price (DAX-115). The revision-side readings above raise the natural flagship question: does the tilt's own $+11.8\%$ numerator against uniform carry the same structure? The registered zero-decode read scores the same six per-doc token-level notions on the frozen tilt-vs-uniform boards, shared-stream bootstrap, pooled anchors re-produced bitwise: per-doc, the tilt is *lower* than uniform on all five richness notions (distinct-$1..4$ and unigram entropy -0.061 bits $[-0.070, -0.052]$ outCI) and *higher* on intra-doc 4-gram repeat mass ($+51\%$ relative, outCI) — TILT-EDGE-PARTLY-DIVERSITY-PRICED. The pooled distinct-2 disclosure (-0.0167) understates the per-doc entropy-notion gap $3.7\times$: the flagship numerator carries a measured per-doc diversity-cost component — the revision-side pattern's weaker sibling at the tilt cell itself. The matched-d2 tilt-vs-uniform board stays the named open form (DAX-101's ladder-exhaustion stands); what is measured is the frozen pair's notion-level co-price; the caution that one completed matched-diversity measurement reversed a headline is checked at the flagship cell too — without reversing it: the tilt stays the NLL and MAUVE leader while paying the enumerated per-doc premium. *Retention leg (DAX-121, within-corpus form):* banding the frozen pair into the anchor's per-doc distinct-2 quintiles with joint containment (450 co-occupying docs, 22.0% of the corpus; within-band d2 gaps ≤ 0.0037, window ± 0.01), the tilt keeps $+8.527\%$ $[+3.526, +13.315]$ of its $+11.79$ anchor-relative win — at least half, CI excluding 0, and the retention grows with band richness ($+15.28\%$ at the top band) — SURVIVES-HALF: unlike the revision side, the flagship numerator is not consumed by its own diversity contraction in this read; the full-board surface-existence question stays the named open form (DAX-101). *The multi-notion leg answers the deployment-arbiter question by measurement (DAX-123):* a registered joint-containment multi-notion match exists in machinery (the DAX-113/DAX-121 notions and banding inherited bitwise, T0 reproduction of the DAX-121 union 450 docs $+8.527$ at drift 0), but it is corpus-infeasible past one notion: tiering the per-doc paired-gap containments ($+|\Delta \mathrm{d1}| \leq 0.010$, $+|\Delta H_1| \leq 0.050$ bits, $+|\Delta \mathrm{d3}| \leq 0.010$, $+|\Delta \mathrm{rep4}| \leq 0.0025$) shrinks the union to $81/41/26/16$ qualifying docs ($4.0/2.0/1.3/0.8\%$ of the corpus, all below the registered 150-doc floor, all straddling; the half-window robustness tiers tighten to $35/8/5/3$) — MATCH-KILLS-THE-CORPUS: the surviving docs progressively concentrate where the arms already agree on richness, so the multi-notion match selects a comparison a deployment cannot run. The deployment arbiter must therefore pick *one* notion: each single-notion pick is individually matchable and the priced ladders name its cost (d2: 22.0% share at $+8.53$ retention, DAX-121); the surface-existence leg stays named open (DAX-101 stands). Two bill clarifications ride with it: the 33-call bill counts executed forwards; the 16 repair forwards are sequential *by mechanism* — each repair commit decodes in the fuller context of the already-repaired positions, and it is that incremental context the revision arm buys (the random-target control -18.09% prices collapse-without-it); a batch-parallel single-pass repair is a different decoder family (conditional-independent refill; the unguarded-iterated lineage certifies its degradation), not a cheaper pay-down form of this one. KV-cache-aware accounting leaves the call bill unchanged: both arms execute the same forward count at 1.0000 per-forward wall parity (DAX-96), cache reuse alters per-forward FLOPs identically for scaffold and repair legs, and the one-vs-two-pass asymmetry is priced by the cooled- telescope read above. The ReMDM operation — stochastic per-token re-masking *inside* the ancestral trajectory (Wang et al., 2025) — is a third revision *schedule* beyond the one-shot boundary forms measured above. Pinned against the canonical form (band gate $t \in (0.05, 0.55]$, per-token re-mask probability $\sigma_k{=}0.04$ constant on the band, zero extra forwards, an EXACT SOFTORD base, machinery gate: remasking off reproduces the frozen tilt board bit-identically), at exactly 32 model calls the canonical *uniform* arm reaches 483.20 and the canonical *confidence* arm 448.67: both beat the 32-call tilt anchor (534.62) excluding 0 ($+9.62\%$ $[7.13, 12.09]$ and $+16.08\%$ $[13.73, 18.30]$), so the canonical family *does* transfer to this substrate, yet both rank strictly below the one-shot targeted boundary revision (363.00; -33.11% $[-36.45, -29.85]$ and -23.60% $[-26.75, -20.52]$, excluding 0 — at a 32-vs-33 call bill *favouring* the interleave). The two dials price separately: holding selection uniform-random, moving the revision from a one-shot boundary into the trajectory pays $+23.46\%$ $[21.43, 25.46]$ over the random one-shot control — the random control's failure was a placement fact as much as a selection fact — and inside the interleaved stochastic family, confidence targeting adds a further $+7.15\%$ $[6.16, 8.12]$. Reading: revision value grows with the context gap behind each revised token; the interleaved canonical schedule spreads its revisions across small per-pass gaps. (Protocol and caveats as registered pre-measurement: tilt commits on an exact-SOFTORD base, commit-time cached confidences, canonical dose $\eta{=}0.04$ on passes $16..31$, single-seed paired anchors, frozen scorer; fixed class REMDM-UNDERPERFORMS-

REOPEN, direction expectation registered and confirmed verbatim.) The tuned-ReMDM family — remasking-schedule-tuned variants priced under the same scorer and bill — stands as a recorded candidate baseline class: the claims above are restricted to the pinned canonical form and the measured boundary revision, and no cross-family claim is made.

Does the revision survive cooling? (DAX-71). The whole revision family above is measured at $T{=}1.0$; the registered question whether its central cell survives the temperature that dominates every other dial is answered by the exact canonical recipe (reopen2x16, 16 scaffold $+1$ re-target $+16$ repair, 33 forwards, nothing else changed) re-run at $T{=}0.7$: 25.92 against the cooled tilt board's 29.58 — $+12.37\%$ $[10.80, 13.95]$ excluding 0, $+27.65\%$ $[26.14, 29.14]$ over the cooled uniform anchor — REVISION-SURVIVES-COOLER. The margin compresses from $+32.10\%$ at $T{=}1.0$ to $+12.37\%$, a $2.6\times$ shrink whose two CIs do not overlap, matching the registration's weak-form expectation (survives, compressed) exactly in both sign and size: at sharpened landscapes the scaffold itself hardens, so less repair mass remains. Costs travel with the win, disclosed verbatim: distinct-2 0.1429 and distinct-3 0.3869 at the cool point, and the cool regime is repeat-prone across the family (repeat4max 26, against 21 for the cooled uniform board itself, both measured in DAX-72's ladder below) — the cool-then-revise stack pays its guardrail bill on both dials and stays strictly below the cool tilt board. The revision's measured temperature scope is exactly these two cells — the $T{=}1.0$ charter cell and this $T{=}0.7$ survival read; distributional claims beyond the two measured temperatures stay boundary items, not claimed.

Second-arbiter audit of the revision family (DAX-56). Whether the family's largest claim survives the paper's own secondary-metric guardrails was evaluated under a fixed class lattice registered before any component ran, consuming frozen boards only. Three gates: G-MAUVE (per-arm official MAUVE, bit-identical featurizer and recipe, frozen validation reference, full + two halves, must clear the frozen uniform-32 anchor $0.0621/0.1000/0.1212$); G-REPEAT (intra-sample 4-gram repeat rate of the iterated champion $\leq 2.5 \times 10^{-3}$, the bar carried from softorder-32); G-CONC (top-decile share of positive gain mass on the anchor entropy grid $< 60\%$). Result: G-MAUVE passes emphatically for the whole ladder — reopen2x16 0.0770, pos128 0.0992, 24A+8C+pos128 0.0932, iterated 0.1385 ($= 2.2\times$ the anchor full), every half above its anchor half; G-CONC passes (0.273 for the champion, declining $0.313{\to}0.273$ down the family: the gain spreads over the entropy grid, it is not concentrated); G-REPEAT fails on the iterated round ($4.599 \times 10^{-3} > 2.5 \times 10^{-3}$). The dose$\to$repeat-rate profile is monotone — $0.0018/0.0021/0.0026/0.0036/0.0086$ at $12.5/25/37.5/50/75\%$ dose, 0.0046 under iteration — so at this bar the repetition cost turns disqualifying between 25% dose (pass) and 37.5% dose (fail), and always under iteration. Verdict class: REPETITION-SIGNATURE-CONFIRMED. *The dose threshold is shape-consistent (post-hoc analytic).* Three independent parametric fits of the measured dose$\to$rep$_4$ ladder (power; exponential-in-q; power-with-floor) invert the $2.5{\times}10^{-3}$ bar at re-masked fraction $q \in \{0.255, 0.310, 0.330\}$ — every inversion inside the measured crossing window $(0.25, 0.375)$ — with the best form's leave-one-out mean error 9.3%; the fitted steepening exponent ($b \approx 2.4$–2.6) quantifies the compounding directly: the iterated point sits $2.17\times$ above the single-round curve at the same q. Post-hoc, five-point family, single substrate; at-scale prediction stays an untouched boundary (no larger checkpoint). The reading this audit revises as the registered rounds land: the family's *clean frontier* under all three arbiters is the single-round revision (reopen2x16, $+32.10\%$, its confidence-targeting attribution intact); every deeper dose and every iterated round stays on the record as a gain-with-repetition-cost dial, numbers disclosed verbatim (the tilt baseline's own repeat rate sits at 2.501×10^{-3}, within 10^{-6} of the bar it defined; gates applied exactly as registered, no post-hoc relaxation). Where the repetition signature is disqualifying, $+32.10\%$ is the headline; where a gen-NLL-only objective prices the trade, the iterated 227.69 stands with its signature disclosed. A registered clean-dose iteration probe (DAX-57: 16A+8C[pos64]+6D[pos64], iso-32 calls; only the first-revision dose changes $128{\to}64$) tests whether the signature is inherent to iterating or inherited from the 50% first-revision dose. The probe (DAX-57) answers this: iterating at the clean 25% dose still improves the single round (278.71 vs 363.00; $+23.22\%$ $[21.48, 24.89]$ excluding 0 at 32 vs 33 forwards — the iterated gain is not a first-revision depth artifact) yet its rep-rate (2.88×10^{-3}) re-crosses the bar — narrowly, $1.15\times$, and far closer than the deep-iterated 4.60×10^{-3}, so the signature is dose-modulated but iteration-triggered: ITERATION-INHERENT-SIGNATURE. The clean frontier statement above is thereby confirmed experimentally rather than inferred, and the next registered move is a weak anti-repeat guard on the second round (DAX-59: no-repeat-4-gram commit constraint, iso-32 forwards) testing whether the $+23.22\%$ iteration gain is recoverable inside the bar. It is (DAX-59/60): with a no-repeat-4-gram constraint on second-

round commits only (reweighting, zero extra forwards, still 32 calls) the guarded two-round point keeps the iteration gain (279.28; $+23.06\%$ $[21.31, 24.74]$ excluding 0 over the clean single round; $+47.76\%$ $[46.44, 49.06]$ over the anchor) while its repeat rate returns to the baseline's own level (2.499×10^{-3} vs the baseline's 2.501×10^{-3}) — and it clears the remaining arbiters on the frozen recipes: official MAUVE 0.1251 full ($2.0\times$ the frozen anchor; halves $0.1545/0.1986$ clear) and top-decile gain-mass share $0.294 < 0.60$, with diversity unmoved (distinct-2 0.4263) and zero fallback overrides. The family's clean frontier thereby extends from the single round to the guarded iterated round. A third guarded round (DAX-61, 33 calls) still pays (248.80; $+10.92\%$ $[10.02, 11.77]$ excluding 0; $+53.46\%$ $[52.26, 54.61]$ over the anchor) but re-crosses the repeat bar (2.687×10^{-3}) — exact-4-gram blocking saturates at round two: the family closes clean at the guarded two-round point, with the round-marginal chain on the frozen scorer (single $363.00 \rightarrow$ guarded-2 $279.28 \rightarrow$ guarded-3 248.80 against the anchor's 534.62) reported with the signature flag on round three. The wider horizon succeeds (DAX-62/63, registered pre-measurement): exact-3-gram blocking — a strictly stronger forbidden set — keeps round three inside the bar *and* keeps its marginal (250.49; $+10.31\%$ $[9.42, 11.17]$ excluding 0 over the guarded two-round; repetition 2.434×10^{-3}, below the two-round point's own and the baseline's background alike; official MAUVE 0.1546 full with halves $0.1510/0.2551$ against the frozen anchor's $0.0621/0.1000/0.1212$; top-decile gain share $0.285 < 0.60$; distinct-2 0.406; zero fallback overrides). The round-three residual was recreatable-in-3-gram mass — a within-pass hole of the pre-pass guard — not mass beyond the exact class: the weak saturation claim (4-gram, two rounds) stands and the strong one is withdrawn. The family's clean frontier thereby extends to the guarded three-round point (250.49, 33 calls, $+53.15\%$ $[51.95, 54.31]$ over the anchor, $+31.00\%$ $[29.45, 32.49]$ over the clean single round), passing every arbiter this paper uses. A fourth guarded round (DAX-64/65, registered pre-measurement, 36 calls at the same clean dose with budgets $(6, 5, 3, 2)$) stays clean under the same 3-gram guard at a decayed marginal (242.44; $+3.21\%$ $[1.13, 5.28]$ excluding 0 over the three-round point; repetition 2.463×10^{-3}, still below the baseline's background 2.501×10^{-3}; official MAUVE 0.1103 full with halves $0.1401/0.2120$; top-decile gain share $0.263 < 0.60$; distinct-2 0.397; zero fallback overrides) — the family does not saturate at three rounds either under exact-3-gram blocking: the clean frontier extends to the guarded four-round point (242.44, $+54.65\%$ $[53.68, 55.60]$ over the anchor), the round-marginal chain reading $+32.10/ + 23.06/ + 10.31/ + 3.21$ percent with the MAUVE full-value chain $0.1251 \rightarrow 0.1546 \rightarrow 0.1103$ carried verbatim beside it and re-audited below (DAX-72) — the endpoint is a Pareto extension of the single-round point under the repetition and NLL arbiters, not a domination of it under all arbiters simultaneously. The guard-horizon comparison therefore forms a measured 2×2 block, not a single row (DAX-66, registered pre-measurement): exact-4-gram is below the repetition threshold exactly through round two (2.499×10^{-3}, DAX-59) and above the repetition threshold on both deeper rounds measured (round three 2.687×10^{-3}, DAX-61; round four 2.907×10^{-3}, DAX-66), while exact-3-gram stays clean through round four ($2.434/2.463 \times 10^{-3}$, DAX-62/64). The narrower guard's remaining NLL advantage at round four ($+1.50\%$ $[1.11, 1.88]$ over the wider point, which is itself $+4.02\%$ $[2.00, 6.01]$ over the same guard's shallower point) is exactly the gain the deeper rounds cannot keep clean: the wider guard is what unlocks depth, and its price grows with it (0.68% $[0.40, 0.95]$ of NLL at round three to 1.50% at round four, out-CI at both depths). Two boundaries on the four-round reading are stated as measured: the same flagship numerator is co-priced per-doc on every richness notion as a decoding-family pattern (DAX-115), and an independent new-seed replication of the guarded third/fourth-round points has been run (DAX-120 = REPLICATES: at the never-before-decoded seed 780, both bars pass at point value — rep_4 $0.002163/0.002459 \leq 2.5 \times 10^{-3}$ — with all four registered NLL keys outCI-positive, $+59.02\%/+61.67\%$ over the uniform anchor and $+30.55\%/+35.04\%$ over the clean single round). The same set of analyses prices the razor itself (DAX-118): per-document bootstrap CIs (B=20000, one shared-index stream) put BOTH razor points at straddle against the bar the family defined ($z = 0.007/0.011$), together with the whole 0.0024–0.0027 decision band, while no verdict flips; the consumed gates thereby discharge at point value with their sampling basis disclosed per board.

Second-arbiter re-audit of the full ladder (DAX-72/77). With the four-round chain closed, the registered re-audit re-reads the whole ladder under official MAUVE, consuming the frozen T=1.0 boards verbatim (zero new compute, restatement disclosed) and featurizing five fresh T=0.7 cells on the identical recipe. Three named inversions land. (i) $g3r3 \rightarrow g3r4$: the full ladder 0.0621 / 0.0730 / 0.0770 / 0.1251 / 0.1546 / 0.1103 (anchor, tilt, one-shot, guarded-2/3/4) climbs through round three and *inverts* at round four, -0.0443 — the pre-disclosed chain decline is thereby

confirmed as a named inversion, not a disclosure to carry. (ii) $u07{\to}s07b1$: at $T{=}0.7$ the tilt's MAUVE margin over its own uniform board halves ($0.1770{\to}0.1653$, -0.0117) — under the second arbiter the cool tilt does *not* win distributional ground it wins on gen-PPL. (iii) $s07b1{\to}r2x16t07$: the cooled revision round drops a further -0.0204 ($0.1653{\to}0.1449$), so DAX-71's REVISION-SURVIVES-COOLER is a gen-PPL fact whose distributional margin goes the other way. The dispositions this forces, registered in advance of the ladder read: the frontier recommendation **moves back one rung** — the guarded three-round point (250.49, MAUVE 0.1546, d2 0.406) is the MAUVE-balanced vertex; the four-round point stays on the record as the gen-PPL vertex (242.44) with the MAUVE slide and the distinct slide carried beside it; and the split itself is measured configuration-sensitive on this near-degenerate panel (DAX-94, registered pre-measurement, zero new decoding): across a $3k{\times}4$-reference sweep of the official recipe ($k \in \{204, 409, 500\}$; full/half0/half1/resampled reference windows) the inversion holds in 11 of 12 cells (range $\Delta \in [-0.0072, +0.0444]$, mean $+0.0170$), the single flip being the ($k{=}204$, half-reference) corner, while the registered operating point $k{=}409$ preserves the split on *all four* reference choices; the vertex recommendation therefore stands at its executed operating point with the non-robustness to the reference/k axes disclosed. The fragility is measured, and its feasibility follow-up is reported in full (DAX-100, registered pre-measurement, zero new decoding): a document-resample bound at the frozen operating point ($B{=}300$ over the frozen feature caches; determinism gate bitwise) gives $SE_{2048}{=}0.0185$ with the measured percentile interval $[-0.030, +0.042]$ including 0 — at the current board the split is genuinely fragile to document sampling. A fixed-effect $\sqrt{n}$ sensitivity projection (centered on the frozen point estimate, width shrinking as $\sqrt{2048/n'}$) would exclude 0 from $n'{=}4096$, but this is an *assumption-conditional projection, not power evidence*: the resample distribution is center-shifted against the frozen estimate (mean 0.0058 vs $\hat{\delta}{=}0.0443$, -2.08 SE, a shift skewness does not diagnose), and under the bootstrap-center scenario no ladder n' resolves. The larger-board question therefore stays OPEN: it is a new registration if wanted, with the measured fragility — not the projection — as its motivation. That experiment was then run (DAX-117): both vertex boards were re-decoded at $n{=}4096$ under the frozen g3-blocker recipes, and at the k-rule operating point ($409{\to}500$ by rule) the split reads -0.0097 with the registered resample interval $[-0.055, +0.035]$ including 0 (SE 0.0229, *larger* than the 2048 bound — the $\sqrt{n}$ shrink fails by measurement); the fixed-$k{=}409$ continuity read keeps the frozen direction at $+0.0042$ — the split is a sampling-level disclosure at both measured n, and k-sensitive at the new n; the projection is discharged as refuted (SPLIT-CLOSES-AT-LARGER-N). The $T{=}0.7$ prescription is re-read as *gen-PPL-conditional* — the two arbiters name different hardness optima there (gen-PPL: the hard measured edge $\beta{=}2$, DAX-68; MAUVE: the interior $\beta{=}0.5$, 0.1805 the ladder's cool maximum), a disagreement we state rather than resolve. Rounds one through three and the tilt itself *stand* under both arbiters (every measured cell clears its frozen-anchor twin on MAUVE, full and halves, $+0.011/+0.004/+0.048/+0.030$ along the chain), consistent with the registered stopping rule, stated explicitly: the family stops expanding exactly where a registered guardrail bar fires — the bars *are* the stopping criterion. The target-set question is answered by a registered fork (DAX-77): re-target the repair mass at *soft* low-confidence weights ($w \propto e^{-p_{\max}}$ over the completed positions inside the same lowest-64 pool, otherwise the canonical reopen2x16 arm bit-for-bit, machinery gate bit-exact). The outcome falls in no registered class and is consumed under the line's registration-gap convention (facts named, expectation mismatch recorded verbatim — soft-targets-same was the registered reading): the soft arm *does* recover diversity to scaffold parity at the board level (distinct-2 0.5404 against the scaffold's 0.5401; the per-sample recovery over the hard arm is $+0.0067$ $[0.0053, 0.0082]$, excluding 0) but its quality lands *below the scaffold itself* (580.82; -8.64% $[-11.55, -5.80]$ vs softord-32 and -60.01% vs the hard targeting, both excluding 0), tracing a single dose chain $631_{\text{random}} \to 581_{\text{soft}} \to 535_{\text{scaffold}} \to 363_{\text{hard}}$: the diversity cost and the quality gain ride the *same* low-$p_{\max}$ targeting dial — the tail is load-bearing in both directions and there is no free diversity recovery on this knob (UNCOVERED-QUADRANT). A halves-cosine re-read on the frozen MAUVE featurizer closes the last repeat-signature doubt: intra-sample semantic self-similarity *falls* monotonically down the revision chain ($0.2772/0.2800/0.2632/0.2598/0.2565/0.2479$; soft arm 0.2744, scaffold parity) exactly where the exact-4-gram bars rise — the flagged signature is surface n-gram repetition, not semantic auto-copying, and the exact-n-gram guards stand as the family's operational guardrail (the halves-cosine statistic itself co-reads topical consistency; disclosed as a directional, not a gate, line).

The guarded family's one blind spot, measured and priced (DAX-82/83). A block-level semantic self-copy audit (the frozen featurizer at 64-token blocks; max-pair cosine MC per document; alarm

τ=0.68 = the uniform board's $p99$) localizes what the halves-cosine mean view missed: the guarded revision chain keeps its means below the one-round base — the surface-guard story above holds in the bulk — but a small semantic-duplicate *tail* passes the 4-gram/3-gram guards untouched: at the round-3 vertex the alarm-rate excess over the uniform board is $+1.27$pp $[+0.49, 2.05]$ outCI and stable across six threshold settings, while the guarded and unguarded round-2 boards are indistinguishable at every threshold — SEMANTIC-EXCESS-GUARDED (tail-localized). A registered repair answers the only executable follow-up: on the bit-reproduced frozen trajectory (in-job bitwise gate), documents above τ have their max-pair later block re-masked and repaired with two guarded passes, against a count-matched uniform-random trigger control. The tail is removed surgically ($2.29\%\rightarrow0.49\%$, inside the registered band; the control stays at 2.25%), yet *both* arms pay $\approx -1.3\%$ against g3r3 outCI with an identical $+0.011$ distinct-2 bump — trigger irrelevant — NO-RECOVERY. The semantic-duplicate tail is real, detectable, and removable, and removing it buys nothing under the frozen scorer — it is not the load-bearing defect at the vertex, and the operational recipe stays the surface guard plus this disclosure.

Scorer-robustness of the whole family and matched-computation controls (DAX-78/79). The standing stability question the family had left open — is any of this frozen-scorer-relative? — is measured twice over on frozen boards: every flagship cell re-scored under two declared arbiters (one $2\times$ same-family scale ladder, one modern cross-family 1.5B; 10 boards, zero re-decodes, scored env-coherence-gated). Every one of the nine named rankings keeps its sign under both arbiters (gpt2-xl: $9/9$ intervals excluding 0; qwen: $8/9$, the round-four marginal at $+1.83\%$ $[-0.48, 4.08]$ the single interval touching zero), with the median relative shrink of win-margins 2.4% — SHRINK-BUT-HOLD. One finding cuts the other direction and is reported first: the *unguarded* iterated champion's edge over its guarded twin, $+18.5\%$ on the frozen scorer, reads $+0.19\%$ $[0.10, 0.29]$ and $+0.22\%$ $[0.11, 0.33]$ under the two new arbiters — a 99% collapse of the gap with only its sign left — the surface n-gram signature was readable as fluent text by the small scorer and by neither larger one. Consequence, stated exactly: the lowest measured perplexity (227.69, unguarded two rounds) is reported with its repetition signature, but it is not read as a within-family quality gain, because it reflects proxy inflation; the guarded vertex family ($363.00/279.28/250.49/242.44$) is the arbiter-stable recommendation, and the guardrail bars' role as the registered stopping rule is thereby validated empirically rather than by convention. A modern-featurizer MAUVE re-read was attempted in the same set of analyses (DAX-84; qwen2.5-1.5B last-token features, official recipe untouched, the same ten boards): its range across the family is 0.0063, below the registered informativeness bound 0.01 — Q-MAUVE-DEGENERATE, so the MAUVE featurizer panel stays gpt2-large-only with the boundary named (a stronger modern-featurizer reading does not resolve the split at n=2048; the inversion question is thereby featurizer-bounded, not hidden). A second, same-tokenizer rung is registered and consumed with the same verdict shape (DAX-116, gpt2-xl last-token features): the validation-reference selftest reads 0.764, below the registered 0.90 discriminability gate, so the six-cell census keys were never computed — the featurizer axis is attested degenerate at two rungs, and gpt2-large stands as the sole non-degenerate panel featurizer in the asset set. The bill-matched question closes in the same set of analyses (DAX-79, three fresh cells, all priors inside their registered bands): the guarded four-round point beats its same-call tilt twin $+53.76\%$ $[52.61, 54.89]$ at 36 forwards, the single-round recipe at the same pool and dose beats the 64-call tilt $+36.71\%$ $[35.20, 38.16]$ at the declared 65-vs-64 pairing, and DAX-49's scaffold front-load persists across bill sizes ($+3.65\%$ $[1.77, 5.48]$ at 65 calls after $+8.60\%$ at 33) — SAME-BILL-LADDER-CLEAN + FRONTLOAD-PERSISTS-AT-64.

The measured prescription, stated plainly. *Placement*: free, small, low-NFE only — take the frozen grid at latency-bound budgets; expect $\mathcal{O}(1\%)$ at 32 calls on this substrate. *Soft ordinal tilt*: the largest priced lever at every measured budget and temperature; prefer it over cooling at any fixed diversity band; β=1 at T=1.0, harder measured edge at T=0.7 under gen-PPL with the second arbiter naming β=0.5 instead — stated per arbiter, not resolved. *Commit revision*: the largest measured move; run the guarded recipe with the exact-n-gram bars as the stopping rule; recommended vertex is the MAUVE-balanced three-round point (250.49 @33c) where distributional cost matters, the four-round point (242.44 @36c) where a gen-NLL-only objective prices the trade; expect the diversity dial to move with the revision pool, not independently. *Triggers and early exit*: both channels measured closed on this substrate family (triple-closed ancestrally; altitude channel real-but-linear, per-position floor complete except the width-3 prompt band). Temperature itself is the dominant dial; cool only to a diversity band, and tilt (then revise) inside it.

Where the registered bar fires. The 5% worth-having bar was registered and is reported without adjustment: on this substrate it is fired by exactly two measured moves — spending calls (NFE $32\to64$: $+6.49\%$) and the ordinal/composition layer: softorder-32 $+13.10\%$ (CI_{95} $[10.66, 15.46]$), the composed gridsoft-8 $+8.49\%$ (CI_{95} $[6.40, 10.55]$), and the composed gridsoft-16 $+12.38\%$ (CI_{95} $[10.09, 14.59]$; CIs as the monotone images of the registered paired bootstraps) all clear the bar with their full intervals, while every placement-only cell stays strictly below it ($+1.06/ + 2.84/ + 4.18/ + 4.26\%$ at $\mathrm{NFE} \in \{32, 16, 8, 4\}$). The "placement is free but small" reading is therefore sharper scoping, not a missed firing: at 5% scale the free channel on this substrate lives in the ordinal tilt and its composition with placement, not in placement alone.

Why the instantaneous rule degenerates despite abundant candidates (DAX-18/22). The τ family at parallel budgets fails for a subtler reason than scarcity. Measured on the full $512 \times 32 \times 256$ per-position $p_{\max}$ tensor: 29.7% of positions *ever* clear 0.85 before commitment (24.7%/18.3% at 0.90/0.95; every sample has some τ-clearing position) — TAU-PRESENT, an order of magnitude above the registered sparsity bar. The degeneration of the instantaneous batched rule is therefore *temporal scatter*, not absence: at any single pass, only ~ 1 position per sample clears, so harvesting requires near-sequential pacing (DAX-10's 249–252 calls). The registered synthesis (DAX-22) reads TAU-CHANNEL-OPEN: the quality spread of the $\tau \in \{0.85, 0.90, 0.95\}$ cells stays inside 1.21 PPL, and both conceivable closures fail — availability exists (the channel is not sparse), and the percentile/top-q escape is a fractional-count selection-order rule whose hard face inherits DAX-14's coverage cost while its soft face is precisely DAX-15's measured positive cell. The ordinal channel at parallel budgets is open exactly along the softness axis.

Confidence information under a fixed count rule. The $\beta = 0$ control is uniform over positions under the same deterministic counts as $\beta = 1$: at length 256 and 32 calls it selects eight positions per call. This is not the random Bernoulli count rule of the ancestral uniform baseline. The fixed-count uniform output has perplexity 606.10 versus the ancestral baseline's 615.20; the signed reduction of the latter relative to the former has interval $[-4.36, +1.30]$. This unresolved difference cannot establish equivalent output laws. The direct comparison of retained $\beta = 1$ and $\beta = 0$ document scores gives an 11.79% reduction [10.13,13.45], supporting a confidence-information effect within the fixed-count setting. The analysis uses 2,048 documents and 20,000 paired bootstrap draws at fixed weights, not independent training seeds.

Where the order pays: context strata (DAX-85/86). *Why* the rank information pays at every budget is measured one level deeper than the attribution: three draw-paired instrumented traces (uniform / β=1 tilt / hard order; same seeds, counts, and counts schedule) joined with per-position frozen-scorer loss. First the registered surface read is a null: commit-band error anatomy does not separate the orders (early-band shares within 0.1pp of each other; a premature-commit metric is saturated at 0.925 in all arms; the tilt's commit-surprise profile is U-shaped, not the registered mid-horizon hump) — SOFT-IS-JUST-DRAW. The discriminating observable is *context availability at commit time* — the fraction of a position's ± 8 neighbourhood already committed when it is committed. The tilt beats the flat selector exactly in the context-*poor* strata (per-position losses 5.66 vs 6.29 in the lowest band and 6.44 vs 6.50 in the next, both bootstrap-separated) and *redirects* error into context-rich strata, where both pay most; it does not "wait for context" (its mid-horizon availability is marginally *below* uniform's, -0.003 outCI). The hard order's catastrophe localizes complementarily: its availability trajectory is $0.152\to0.282$ against 0.39–0.41 for the others. Its late, highest-entropy commits land in the context-poorest states with the highest rates ($+1.17$ per-position loss, $+5.5$pp error share vs the tilt, both outCI). Read together with NOISE-ORDER-EQUIVALENT: the rank information pays by choosing *which* uncertain positions may wait for their neighbourhood to exist — a scheduling effect on the stochastic side of the commit order, not a smoothing one. The signature is *not* substrate-bound (DAX-87/88): on a held-out wikitext checkpoint at L=256 and on the L=512 horizon, the same two-arm contrast reproduces all three registered components with flagship signs and bootstrap-separated intervals — poorest-band per-position loss edge -0.542 and -0.542 (flagship -0.627), redirection into rich strata $+1.561$ and $+1.655$ ($+1.867$), low-band error-share reduction -0.026 and -0.028 (-0.033), and the same no-waiting refutation (-0.0015 on both). The measurement chain is reported in full: the first read returned a gate failure because the registered one-sided population gate razor (0.002 nats on both L=512 arms) sat far inside the anchor boards' own sampling error (SE ≈ 0.008); the investigation named the gate-form defect (a point-gate razor inside mutual draw noise), and the registered fix transaction — two-sided population gates on

the same frozen assets, zero new decoding — read STRATA-REPLICATE, elevating the sentence to multi-substrate (DAX-87 GATE-FAIL → DAX-88 STRATA-REPLICATE). The agenda/strata distinction sharpens on the canonical non-ancestral sampler (DAX-104, bitwise-verified trace of reopen2x16): the re-open *agenda* localizes to context-RICH commits ($E[C(j)]$ 0.5275 vs 0.4485 retained, $+0.0790$ $[+0.077, +0.081]$ outCI) — the lowest-confidence committed tokens sit in context-rich neighborhoods (soft's early easy commits), not the context-poor catastrophe sites — while the quality win still localizes context-poor at the strata level. Remasked positions recommit with a large availability lift ($+0.2811$ $[+0.2792, +0.2831]$ outCI) that is nearly constant across samples (per-sample SD 0.0443): a floor effect, not a per-sample signal — correlating the lift with the per-sample revision advantage gives $+0.0655$ $[+0.0222, +0.1085]$, straddling the registered 0.05 exploitability gate, so the re-noising revival regime (where re-noising could re-open per-sample adaptivity) stays OPEN and bounded (AGENDA-ANATOMY-INFORMATIVE-ONLY; the ancestral null does not graduate sampler-wide on this cell). The agenda anatomy does *not* close the DAX-104 context-rich programme either: the revision's measured context gap is a T=1.0, 256-token flagship fact, while the re-open agenda's context-rich localization is a mechanism read on the same board — the tension between the two is recorded, not resolved, and the longer-horizon, richer-context cells of the agenda stand as named candidates. The residual two-gate tinge beside the lineage (DAX-91/92) is priced at nuisance scale: the L=512-substrate trace-vs-board population difference ($+0.0297$/$+0.0342$ nats, z=1.85/1.93) sits inside mutual draw noise while the strata effects those gates certify are 15–50× larger — no conclusion beyond the L=512 replication's own attestation rides within razor distance of them.

The scorer-variant decomposition (DAX-95). Which positions carry the edge? Restricting the frozen gpt2-large arbiter to its own high-entropy quarter (pooled 0.75-quantile 6.72 nats of the joint token stream; kept $25.6\%/24.4\%$ of positions on the two boards) collapses the flagship tilt read to $+2.10\%$ $[-0.69, +4.75]$ — straddling 0 — while the scorer's own top-p 0.9 nucleus read (kept $80.3\%/80.8\%$) preserves it nearly intact at $+12.38\%$ $[+10.40, +14.30]$ against the full-sequence $+13.098\%$ $[+10.69, +15.43]$ (replication of the frozen $+13.10$ within its own stream). The tilt's gen-PPL edge is a *low-entropy-position* phenomenon: it wins by avoiding losses on the scorer's easy positions, not by a global distributional shift; the nucleus-filtered variant keeps the sign and magnitude, and the position-difficulty decomposition is exactly how the otherwise degenerate absolute gen-PPL level distributes. The matched per-token dump is bit-linked to the frozen per-document scorer records (maximum repro $=0$ nats), so the three reads share one pairing, one bootstrap stream, and one scorer. The re-rank question then asks whether a scorer that prices its own easy positions *differently* would reverse the pair: the difficulty-weighted re-read on the same dumps (weights = pooled scorer entropy, mean-1 normalized, one weight field both arms; amendment-1 class map pre-registered) reads $+10.695\%$ $[+8.380, +12.920]$ — inside the $\pm$3pp band of the full read — and the matched-support quarter ladder is monotone $17.779/9.480/5.905/2.097$ (DAX-110, RANK-PRESERVED-ROBUST): the localization is a magnitude profile, not a difficulty-view artifact, and no in-protocol weighting form re-ranks the pair.

The boundary probe: prompt conditioning (DAX-28/28b). Does the instantaneous rule return on a higher-margin substrate? The same tau-family decode is re-run with a 64-token real-text prompt clamped (prefix never unmasked, n=2048, paired seed 777). *Savings arm*: measured closed also under conditioning — mean per-sample NFE 189.49 sits at 98.7% of the 192-position suffix horizon ($256 - 64$; median 191, 34.3% at the horizon), the apparent reduction from the unconditional 249.12 is the pre-committed prefix arithmetic, and the voluntary harvest per suffix position *halves* (1.31% vs 2.69%). The lens matters: under the frozen full-board scorer the board prices -60.7% below the ladder — convention-dominated by the scored real-text prefix, disclosed, with the class map reporting its outside-map fall-through rather than a savings claim. *Quality arm*: at the matched suffix budget the amended measurement (suffix-only scoring, 192 tokens; identical prompt clamp) gives threshold-rule 427.33 vs random-order sequential control 594.77 — $+28.2\%$ ($[+25.5, +30.7]$, outCI), with the order-family guardrail cost re-appearing (distinct-2/3 $-0.035/-0.032$). The tau family thus buys quality-through-order near-sequentially on both substrates and call savings on neither; the trigger paragraph's claim stands with sharpened scope.

T MECHANISM STRESS TEST: DOES UNMASK ORDER CAUSE SHAPE-UNIFORMITY?

Lemma 1 attributes the null to the *exchangeable* uniform-position premise; the sharp question is whether this mechanism is causal. DAX-4a (pre-registered before measurement, with symmetric readings) changes *exactly one factor*: the unmask order becomes greedy by instantaneous confidence, while the count schedule (round($|M|(t_k - t_{k+1})/t_k$)), checkpoints, seeds, and trace instrumentation are held identical to DAX-1.

Measured (n=512). Under greedy order the mean-confidence profile is nearly flat ($0.063 \to 0.069$, versus $0.117 \to 0.374$ under uniform order — high-confidence positions leave the masked pool first). Between-sample relative spread rises in *level* ($0.320 \to 0.374$, versus $0.179 \to 0.181$) by almost exactly the averaging-floor prediction $1/(\mu\sqrt{|M|})$ (μ falls five-fold while $|M|$ is held by the count schedule), yet the early/late *ratio* is 1.170 (paired bootstrap $B{=}20000$: SE 0.154, CI95 $[0.948, 1.543]$) and the cross-horizon persistence is -0.046. The count schedule held ($|M|{=}224.0/32.0$ against uniform's $224.2/31.9$). At the horizon end ($|M|{=}8$) spread reaches 0.94: greedy order concentrates the hard positions at the tail, where between-sample hardness dispersion and averaging collapse meet.

Registered reading: N-conf. The point estimate sits inside the registered $[0.8, 1.2]$ band — greedy order neither explodes the shape ratio past 1.5 (P-conf excluded at the point estimate; the CI is disclosed) nor restores cross-horizon predictability. The direct consequence: Lemma 1's exchangeable premise is *sufficient but not necessary*. Between-sample mean-confidence traces are shape-uniform at the *sampler-family* level for both orders tested, because under greedy order the pool average tracks each sample's own order statistics of the residual masked set, whose cross-horizon correlation is measured ≈ 0 (-0.046): the pool replenishes hard and easy positions without an exploitable early signature. What greedy order changes is the *level* of pooled uncertainty (predictably, via the floor), not its shape across samples. Adaptivity based on early per-sample signal remains closed on both orders; the deployed gains of confidence-threshold decoders come from within-step parallelism, which our measurement never contested.

Does the averaging-floor predict beyond the ancestral sampler? (a mechanism prediction, flagged as such). The floor's mechanism is the *averaging of post-fixture residual uncertainty*: each commit step commits an average over the remaining masked pool, so per-step confidence is an average whose deviation floor decays like $1/\sqrt{|M|}$, pinning the trajectory *shape* while leaving *altitude* free. Block and semi-autoregressive masked-diffusion decoders share this structure — their commit events fix one or a few positions per step against an otherwise-frozen context — so the same shape-uniformity null is *predicted* for them: altitude-diverse, shape-uniform, with cross-horizon persistence bounded by the same averaging geometry; per-sample adaptivity waits on the horizon end, not the middle of the trajectory. The regime where the prediction *fails* is named, too: samplers that re-noise between commit steps (forced re-masking, stochastic regeneration, or large within-step temperature swings) reset the floor mid-trajectory and partially decouple altitude from persistence — that is where per-sample adaptivity could revive. The revision channel is exempt from the floor on the same model's own terms rather than by exception: the floor prices *ex-ante exchangeable* per-sample allocation, while guarded commit revision allocates *conditionally* per-document (the retarget reads each finished decode's realized lowest-$p_{\max}$ quartile) — the exchangeability premise fails by construction, so no averaging-to-floor applies to its targeted spend. The model-correct question for the revision side is the re-noising revival regime (does the conditional allocation keep paying under further rounds), which is measured and bounded-open: revival-gate Spearman $r = +0.0655$ straddles the registered 0.05 gate, bounded $r{<}0.11$ at 95% (DAX-104, AGENDA-ANATOMY-INFORMATIVE-ONLY). Both statements are mechanism predictions of the floor model, not measurements: testing them on block decoders and on re-noising samplers is the negative result's natural boundary experiment — and no floor prediction is exported to block/semi-AR decoder families: the floor is *measured* per decoder family, never assumed (the scope of this paper is the ancestral family and the two instantaneous triggers named above); and no preliminary small-board evidence is claimed for that block/semi-AR prediction either — the cell is a named candidate, not a measured one.

U DOES PLACEMENT ERROR APPEAR AT LOW NFE? (DAX-4B)

Continuous-diffusion intuition says coarse grids misplace mass most, so the pre-registered low-NFE cell repeats Track-B at NFE=16 on owt_mdlm_s0: the grid (same frozen construction, flat-ratio 1.537, max displacement 0.92 steps ≈ 5.7% of horizon — as flat as at NFE=32) against the already-scored uniform board (paired seed).

Measured. Uniform-16 715.37 vs grid-16 695.05: win **2.84%**, distinct-2 drop 0.0023 (inside tolerance), boards genuinely different (2048/2048 rows). The 5% worth-having gate is *not* fired; per-half wins (4.50%/1.16%, halves 682.73/707.59) disagree by 3.3 points, where at NFE=32 halves agreed within 0.3 points.

Registered reading: band. The placement win *directionally* grows as NFE shrinks (1.06% → 2.84%, 2.7×), consistent with a monotone placement-error account — but the 16-cell is precision-limited and stays well below the threshold. The placement-NULL therefore stands at bar-level for NFE≥16; the hint of a saturation knee below 16 is registered, not measured.

Precision resolution (DAX-4b′). The registered follow-up re-scored all four boards with the identical frozen scorer and per-sample gen-NLL dumps, then ran a paired bootstrap (B=20000, sample indices resampled jointly across arm pairs so per-sample pairing is preserved). The bootstrap separates measurement from hint: the NFE=16 win has CI_{95} $[1.32, 4.38]$, which *excludes* zero — the 2.84% win is statistically real at n=2048 precision — while the NFE-slope ($\text{win}_{16}-\text{win}_{32}$) has CI_{95} $[-0.49, 4.08]$, which straddles zero. Per the pre-registered reading rule (both CIs must exclude zero for the monotone account to be measured), the monotone placement-error account stays descriptive, the pre-registered NFE=8 knee cell *stands down* at that point; inside board precision, the placement channel is real but small at NFE≥16. A larger sub-16 board is measured regardless of the fire rule (DAX-8): a vantage trace at 8 steps, the frozen equal-w grid family at NFE=8 (constructed before decode), paired boards n=2048 at seed 777, the identical frozen scorer, and the paired bootstrap (B=20000). The NFE=8 win is +4.183% with CI_{95} $[+2.864, +5.470]$ — it *excludes* zero and both split-halves are positive (+4.79%/ + 3.57%) — while its point stays below the worth-having bar (the CI upper, +5.47%, touches it). An n=4096 follow-up audit (DAX-30) prices the touch: the grid-8-over-uniform-8 win reads +3.09%, CI_{95} $[+2.16, +4.03]$ — the CI *upper* sits below the bar, so the touch was precision inflation, and the knee is a composition knee, not a bar-firing cell (the registered classes BAR-NOT-CLEARED-HP). The win scales inversely with NFE across the two measured points (16 → 8: +2.84% → +4.18%) — a monotone low-information-regime effect, disclosed as such. The 5% fire rule was a deployment bar at fixed count, not a condition for the effect's existence: the complete placement reading is *a real, free quality win at low NFE that shrinks with NFE and never crosses the deployment bar in point*, with the NFE≥16 counts keeping budget (DAX-3, +6.49%) the bigger lever at quality-bound budgets.

The knee measured (DAX-11). The sub-16 ladder is completed pre-registered at NFE=4: a 4-step vantage trace, the frozen equal-w grid at NFE=4 (flat-ratio 1.373; constructed before any grid decode), paired boards n=2048 at seed 777, the identical frozen scorer and paired bootstrap. Uniform-4 1475.01 vs grid-4 1412.16: win +4.261%, CI_{95} $[+3.259, +5.250]$, halves +5.02%/ + 3.49% — the CI excludes zero again. The budget series is measured at four points: +1.06% (NFE=32, CI contains 0), +2.84% (16), +4.18% (8), +4.26% (4). The 8→4 increment is +0.08 points: the win *saturates* below NFE=8, locating the placement-win knee at NFE≈8 on this substrate — consistent with the registered count-schedule note that at four intervals the schedule placement itself is second-order, and with the measurably flatter profile at coarse width (1.373 at NFE=4 vs 1.56 at NFE≥16). A one-parameter quantitative account sharpens the split (DAX-21; `runs/_derived/da_dax21_verdict.json`): take the quadratic specialization of the deficit-risk functional of Appendix H on the measured 32-bin profile, $R(K) = \sum_k (m_k^U)^2 - W^2/K$ (uniform interval masses m_k^U, total mass W), with a single global scale fit to the $K \in \{8, 16\}$ log-wins alone. The resulting curve predicts the *held-out* NFE=32 win at +1.13% (measured +1.06%, CI $[-0.69, +2.80]$) and overshoots the knee: predicted +8.29% at NFE=4 versus measured +4.26% (CI upper +5.25%). The deficit-risk channel therefore explains the placement win's budget trend with one parameter everywhere except at the knee, whose ≈+4.0-point overshoot residual is a measured breakdown scale of that account (verdict MODEL-OVERSHOOT) — the knee is a tested model-

failure point rather than a qualitative description. The registered 5% worth-having bar is an *adoption* yardstick, not a measured quantity, and stays as consumed-once convention. No registered prediction is carried below the measured lattice to $\mathrm{NFE}{=}2$, over which the curvature account is already overshot at 4; a realistic-bar variant is a separate new-registration question, listed as a candidate only.

V Robustness across temperature and trace statistics

The DAX-1 null as first measured rests on one temperature (1.0) and one trace statistic (the pooled mean $\overline{p_{\max}}$ over the masked pool). Three registered follow-ups widen both axes on the same checkpoint/protocol (producer `da/scripts/qt_trace_decode.py`, battery `da/scripts/dax5_reading.py`, verdicts `runs/_derived/` with `da_dax5_verdict.json`, `dax5b_ceiling.json`, and `r1r2_entropy_shape.json`).

A second statistic on the frozen trace (entropy). The DAX-1 trace also recorded the pooled entropy mean, registered but not used by the DAX-1 battery. The identical battery on it ($B{=}20000$ paired bootstrap): persistence 0.441 CI_{95} $[0.369, 0.508]$ — the entropy channel carries *more* persistent-level content than $\overline{p_{\max}}$'s 0.303 — duration correlation 0.0008 (exact null), and early/late relative-spread ratio 0.354 CI $[0.326, 0.385]$, i.e. late/early inflation $\approx 2.83\times$, of the pooling-floor order at that statistic's own anchors ($\sqrt{|M_e|/|M_l|} = \sqrt{7} \approx 2.65$; the shape-implied direction would reverse along with this statistic's μ-profile, so no shape-failure signal is re-derived for it and this channel reports the pooling-floor consistency only). Shape-uniformity therefore holds on both second-moment channels at $T{=}1.0$, as scoped to the measured family (App. G).

Altitude conditioning. Conditioning the early/late correlations out of a leave-one-out altitude score (per-sample mean over the 30 non-contested horizons) isolates the shape channel: partial corr(early, duration | altitude) $= -0.057$ CI $[-0.143, +0.030]$ (still null), and partial corr(early, late | altitude) $= -0.251$ CI $[-0.331, -0.168]$. The raw $+0.303$ persistence is wholly shared altitude; the residual shape channel is *negative* (mean-reversion), the opposite of exploitable persistence — and a positive artifact of the conditioning itself cannot produce a negative value, so the sign is informative rather than mechanical.

Temperature axis (DAX-5). The registered $T{=}0.7$ cell repeats the trace protocol with the same random draws (common random numbers) and additional per-step per-position quantile (q_{10}–q_{90}) and dispersion channels. The battery on four channels ($\overline{p_{\max}}$ / median / dispersion / entropy): all duration correlations stay ≈ 0 with CIs straddling zero on every channel, and all persistences fall below 0.3 — *weaker* than at $T{=}1.0$, so the exit-relevant null is temperature-robust. The pre-registered overall reading nonetheless fired BREAK: the median and dispersion channels show early/late ratios 1.837 CI $[1.487, 2.171]$ and 1.434 CI $[1.298, 1.581]$ — CI-lower > 1, an inflation *reversal* versus the floor direction.

Ceiling account (DAX-5b), zero free parameters. The follow-up registered before running explains the reversal mechanically. Lower temperature migrates levels toward the ceiling (replication-channel horizon means $0.294 \to 0.634$ at $T{=}0.7$ vs $0.117 \to 0.374$ at $T{=}1.0$); for any bounded statistic p, the deficit $1{-}p$ obeys $\mathrm{RS}(d)_e/\mathrm{RS}(d)_l = [\mathrm{RS}(p)_e/\mathrm{RS}(p)_l] \cdot \mu_e(1-\mu_l)/[\mu_l(1-\mu_e)]$ identically (sample-variance terms cancel). Measured deficit ratios 0.4985 CI $[0.409, 0.591]$ and 0.2681 CI $[0.211, 0.326]$ land back inside the floor band around 0.377, while the entropy channel — not ceiling-bound — stays floor-exact at both temperatures ($0.394 \approx 0.377$ at $T{=}0.7$). The reversal is a saturation artifact of the $p_{\max}$ family, not a return of shape.

Composite reading. Across two temperatures and four trace statistics, (i) the early-predictive exit channel is null everywhere; (ii) the shape channel is either absent (floor-exact) or bounded below detectability with a mechanism-attributed artifact. The claims of the paper are thus statistics-independence statements about *non-ceiling-bound* trace functionals, which is the explicit scope wording used below.

Transfer across dataset, length, and a saturated domain (DAX-9). The full battery replicates outside the OpenWebText/$L{=}256$ substrate. On wikitext ($L{=}256$): trace persistence 0.177 (CI

$[0.091, 0.259]$), duration CI containing zero, spread ratio 0.987 (CI $[0.887, 1.098]$) against the DAX-1 null 0.992. On L=512 (same dataset): persistence 0.347 (CI $[0.269, 0.421]$) — altitude rises with length — duration null, ratio 1.062 (CI $[0.980, 1.151]$). On a structured, difficulty-bimodal substrate (sudoku, L=81: an in-house six-layer, 4.8M-parameter model trained with the same objective for $12{,}000$ steps on one million generated solved 9×9 grids, each serialized row-major as 81 digit tokens; configuration `runs/cfg/sudoku_mdlm_s3.json`) the statistics degenerate at the ceiling: the pooled mean is half-committed at k=4 ($\mu_4 = 0.518$) and saturated by k=28 ($\mu_{28} = 0.983$), the late cross-sample spread collapses ($\mathrm{rel}_l = 0.029$) and inflates the ratio to 4.63 (CI $[4.05, 5.34]$) by the denominator-collapse mechanism of the ceiling family, and every early-predictive probe is null (persistence -0.017, duration -0.070, altitude–timing -0.013): the early channel closes there via saturation, not a shape return. (When evaluating this cell we found that the pre-specified ratio band had mixed the entropy-family floor figure into the $p_{\max}$ channel; the verdict uses the channel-corrected rule, implemented in the DAX-9 analysis scripts, and all measured numbers are unchanged.)

Quality-gated transfer to wikitext (DAX-12). The Track-B quality claim is re-derived end-to-end on a second domain and checkpoint: the wiki-profiled grid-32 is rebuilt from the frozen wiki vantage trace (belief-profile flat-ratio 1.589; owt: 1.564 — construction before any decode that used it), and paired boards n=2048 are decoded on wiki_mdlm_s0 and scored by the identical frozen scorer. Measured: uniform-32 627.15 vs grid-32 614.11; win $+2.079\%$ with CI_{95} $[+0.456, +3.700]$ (paired bootstrap B=20000; halves $+2.43\%/+1.73\%$) — on wikitext the NFE=32 placement effect is *by itself* positive and significant. The pre-registered overlap rule compares this interval with the frozen owt null-family interval $[-0.69, +2.80]$: they overlap, so the pre-specified rule labels the cell NULL-REPLICATES — the label reports overlap between the two intervals; it does *not* assert a zero effect on wikitext (the wiki-alone interval excludes zero and its point exceeds the owt null point). (Scorer-domain note: both arms are decoded and scored on wikitext, so the win is a within-domain paired ratio; a GPT-2-large familiarity gap between domains could shift the PPL level of either arm but cannot manufacture a paired within-domain margin, and the paired bootstrap CI runs on the same-domain per-sample differences.) Read together: at NFE=32 the placement effect is small but leans positive across two domains and two checkpoints, sharpening — not contradicting — the budget-resolved reading of §Q.

The conditional family (DAX-16). Track-A's exit-channel null replicates *under conditioning*: a 64-token-prompt trace (n=512, 32 steps, seed 777, prompt prefix from the same OWT distribution) runs the identical registered battery. On the decision statistic $\overline{p_{\max}}$: persistence 0.246 CI_{95} $[0.163, 0.326]$ (overlapping the unconditional DAX-1 interval $[0.223, 0.380]$), duration correlation -0.009 ($[-0.096, +0.077]$, an exact null), early/late rel-spread ratio 0.993 — the headline registered class is COND-SHAPE-UNIFORM. The prompt's information is plainly *visible* (at decode step 0 the per-sample $\overline{p_{\max}}$ spread is already 0.446 relative, where the unconditional trace is exactly across-sample constant by construction) yet it does not become an exploitable elevation or shape channel: the duration channel — the only one that could drive a per-sample exit — is null in *both* statistic families. On entropy, registered secondary class COND-SHAPE-DIVERSE attaches to the level-persistence component (0.466, CI disjoint above the DAX-1 anchor) while its ratio (0.335) and duration ($+0.011$) stay at the independence/null values — the persistent entropy level under conditioning is measured, and equally measured to carry no exploitable timing channel.

A second training recipe at the knee (DAX-17). The low-NFE placement win also replicates across training objectives with a *self-derived* grid: on a TGE-objective checkpoint decoded with its own 8-step vantage trace and its own frozen equal-w grid (both artifact-frozen before any decode, identical protocol), paired n=2048 boards at NFE=8 give uniform 908.17 $\rightarrow$ grid 859.30: win $+5.38\%$ CI_{95} $[+4.07, +6.68]$ (B=20000) — excluding zero and inside the registered family band $[2, 6]\%$ around the owt reference $+4.18\%$ (verdict REPLICATES-LOW-NFE). The coarse time-placement layer is therefore a recipe-level property of the mechanism, not an artifact of this one objective — while the cross-training-seed and cross-family arms stay in the disclosed open set (§W).

The order layer is not recipe-locked (DAX-26). The tilt and cross cells of §S re-decode verbatim on the TGE-objective checkpoint (n=2048, seed 777, frozen scorer; grids re-derived on the TGE vantage): softorder-32_{TGE} 488.20 beats uniform-32_{TGE} 567.69 by $+14.00\%$ (paired bootstrap,

outCI; inverse-ratio form $+16.28\%$) and grid-32_{TGE} 562.93 by $+13.28\%$ (outCI; inverse-ratio form $+15.31\%$); the NFE=8 cross replicates — gridsoft-8_{TGE} 826.51 beats *both* uniform-8_{TGE} 908.17 ($+8.99\%$, outCI; inverse-ratio form $+9.88\%$) and grid-8_{TGE} 859.30 ($+3.82\%$, outCI; inverse-ratio form $+3.97\%$; the two-condition registered class COMPOSE-REPLICATES-RECIPE). Margins are wider than on owt_mdlm_s0 ($+13.10\%/+8.49\%$): neither the tilt nor its composition is locked to the training recipe.

The order layer transfers to conditioned generation (DAX-29). Re-decode the flagship order and composition cells on the conditioned substrate of §S: a frozen 64-token real-text prefix (never unmasked), the generated suffix scored suffix-only (192 tokens, frozen scorer), pass counts fixed on the suffix — uniform/tilt at 32 passes; uniform/8-pass placement/8-pass-compose at the conditioned equal-w grid (grid8_cond, rebuilt from a conditioned vantage trace before any decode consumed it). Both registered arms land at n=2048, seed 777. The free tilt survives conditioning: softorder-32 | cond 562.17 beats uniform-32 | cond 644.99 by $+12.84\%$ (CI_{95} $[10.98, 14.65]$, outCI), matching the unconditional flagship gain $+13.10\%$ within a point. The composition survives conditioning: gridsoft-8 | cond 833.93 beats *both* uniform-8 | cond 902.50 ($+7.60\%$, $[5.91, 9.23]$, outCI) and grid-8 | cond 887.12 ($+6.00\%$, $[4.34, 7.65]$, outCI). The placement-only conditional cells keep the placement layer's sub-bar rank (902.50 $\rightarrow$ 887.12 numerically, CIs overlapping zero): conditioning adds no placement channel, and none was claimed. The suffix distinct-2 values are recorded in the DAX-29 result file (`runs/_derived/da_dax29_verdict.json`): the conditional tilt pays the same order of absolute cost as the unconditional cell.

W Discussion and limitations

Scope: the claims are statistics-independence statements about non-ceiling-bound trace functionals of the ancestral, uniform-position family at $T \in \{0.7, 1.0\}$, with the trace-level null replicated on a second dataset (wikitext) and a second sequence length (L=512), and degenerating by ceiling saturation on sudoku (on difficulty-bimodal substrates the profile-informed placement premise is thereby regime-bounded; the tilt's interiority on sudoku decode boards is unmeasured — a named candidate, not claimed); the frozen grid itself is measured stable to halving its vantage board and to a cross-recipe recheck (§P, DAX-20). Identifiability is stated at the main-text level: the shape-uniformity exclusion holds for the common-$\bar{\rho}$ pure-shape two-moment model with the μ_k^2 factor ($\bar{\rho} \in [0, 1]$; no rescue region exists inside that model: the prediction envelope $[0.313, 0.830]$ lies below the measured CI lower edge, Appendix G, DAX-89); outside that model the anchors do not carry the claim — explicitly, mixed-noise constructions reproduce both two-moment anchors with common $\bar{\rho}$=0 and nonzero shape (σ_δ=0.2, $v_e, v_l \geq 0$), so the exclusion does NOT close any nonzero shape and no σ_δ cap follows without an explicit noise constraint; more general correlation kernels are non-identifiable from a single trace, and what remains measured is the altitude bound and the $z \approx 3.75$ exclusion of the pure-shape prediction (≈ 0.83 relative-spread *deflation*, not the ≥ 2.65 inflation at $z \approx 34.5$ obtained without the μ_k^2 factor); reported at its CI edge the margin reads $\approx 1.9\sigma$ against the family bound 0.830 (the point-form $z \approx 3.75$ is the optimistic quote; the conservative reading stands, supported independently by directly measured within-pool correlations ≤ 0.016, DAX-58) — the null is a measured property of this measured family, not a law of MDLMs. A registered direct probe of that identifiability boundary (DAX-1b) measures the trace's cross-horizon atom-decision persistence at 0.316 (CI_{95} $[0.310, 0.322]$, rising with horizon gap to 0.722) — intermediate between $\bar{\rho} \lesssim 0.10$ and $\bar{\rho} \approx 0.61$, the rescue values that would apply without the μ_k^2 factor (DAX-89): the substrate does not certify inside the excluded family by this proxy *and* the corrected algebra admits no rescue regime to promote it into, so the exclusion's scope is exactly its stated family, a measured statement rather than an assumed regime (Appendix G), and the within-horizon object itself is measured directly (DAX-58): leave-one-out cross-sample-mean-centered within-pool cross-atom correlation at the anchored horizons k=4/16/28 equals 0.0139/0.0089/0.0062 (CI_{95} uppers 0.0163/0.0105/0.0075) — an order of magnitude inside the certified family at every anchor — so the in-family status measured by the probe is the directly measured regime of this substrate, not a fitted calibration. On the entropy channel: its higher persistence (0.441 vs $\overline{p_{\max}}$ 0.303, §V) does not reopen the Appendix I oracle inventory — the exploitable duration-based channel is null in *both* trace families (duration correlation 0.0008, an exact null), and the early/late spread ratio inflates in both rather than compressing, so the higher persistent-level content picks out no identifiable shape factor. What remains open, in decreasing distance from the measured cells: confidence-greedy

remasking orders (by breaking the exchangeable premise they can in principle restore shape diversity; a first matched-budget measurement of confidence *commit* order is in §S — a 33.2% gen-PPL reduction at $\approx$250 calls — but remasking proper stays unprobed; the limitation disclosure stands). The soft-tilt cells are also measured on a second training recipe (DAX-26: tilt $+14.00\%$ and strict composition $+8.99\%/+3.82\%$ on the TGE checkpoint, outCI; anchor-direction forms of the inverse-ratio values $+16.28\%/+9.88\%/+3.97\%$), their β structure is resolved on the grid counts and at $\mathrm{NFE} \in \{8, 16, 32\}$ (DAX-27/27b: interior at β=1 for NFE$\geq$16 in both count families, a $\{0.5, 1\}$ tie at the NFE=8 knee, the hard shoulder excluded at every resolution), they transfer to the conditioned channel at both flagship budgets (DAX-29: tilt $+12.84\%$ and strict composition $+7.60\%/+6.00\%$ on the prompt-clamped suffix, outCI), their placement-slope dependence is saturated by the tilt at $\mathrm{NFE} \geq 16$ (DAX-32: steeper deficit^2-placement crosses tie the equal-w crosses at both K=16 and K=32 while each beats its placement-only match outCI — the equal-w working point is measured, not convenient), and their distinct-2 cost is one correlation with the gain, not an independently dosable knob (DAX-31: the mixed-commit fraction continuum trades strictly co-monotone with no guardrail-recovering point) — still at one domain/length with their disclosed distinct-2 cost, while the tilt's cross-domain/cross-length family, the cross-training-seed arm, confidence-greedy remasking proper, block/semi-autoregressive decode, and other sampler parameterizations remain the open set, not extrapolated here; transfer there requires new registrations rather than extrapolation of these nulls.

Quality-metric sensitivity. All gates run on frozen GPT-2-large gen-PPL, the aggregate $\exp(\text{mean per-sample token NLL})$; win percentages are therefore a reparameterization of mean-NLL differences (win $\% = 1 - e^{\Delta}$), so the 5% bar corresponds to an NLL scale of 0.0513 nats, and every reported win can be read on either scale with paired CIs (Table 5, §U). Distinct-2/3 guardrails accompany every arm. The complementary distributional arbiter (official MAUVE, frozen recipe throughout) entered the registered gate lattice with the revision-family audits (DAX-56 onward) and co-determines the frontier statement (DAX-72): where the two arbiters order the family differently — the guarded fourth round and the T=0.7 hardness optimum — the recommendation is stated per arbiter rather than collapsed to a single vertex. The consolidated measurable recipe is named once, here (DAX-93/95/96 absorbed): on this substrate at T=1.0, use *grid+tilt at* 16 *calls* as the placement working point ($+2.71\%$ over flat-order+tilt outside the interval; the flat profile was the registered prior until refuted at this node), *soft ordinal* β=1 as the headline order lever ($+11.8\%$ at the decode-seed mean, edge localized to low-entropy scorer positions with no difficulty-weighting re-rank, DAX-110), and *one-shot guarded revision at* 33 *forwards* as the largest single move ($+32.66\%$ at the two-decode-seed mean; $+2.24\%$ measured wall premium at pinned batch 64; the full-support numerator re-priced as a diversity consequence at the matched baseline, DAX-109 — guarded-revision-vs-matched-tilt is the new frontier axis), deeper revision reads staying arbiter-conditional with the stacking cost record beside them; the joint (T, β) read is arbiter-binding, not collapsible, and stays cell-wise: at T=1.0 the interior β=1 and the MAUVE panel agree at $K \leq 32$ (at 64 calls the gen-PPL optimum migrates to β=1.5, DAX-111), while at T=0.7 the two arbiters name different hardness optima (gen-PPL β=2, at 32 and 64 calls alike; MAUVE β=0.5) and we state the per-arbiter vertices instead of a joint prescription; cooling compresses the tilt margin to $+12.37\%$ (DAX-71 measured), so the headline working point stays T=1.0 — stated plainly: it is the ancestral deployment point of this sampler family where every reported claim was registered, the cooled corner is itself a degenerate regime ($d2 \sim 0.16$, repeat-prone) whose gen-PPL proximity to the AR anchor is a proxy-level coincidence, and the channel ordering's direction replicates inside it wherever re-registered and re-measured (DAX-71/103/105) — the cooled export is thereby an *ordering-only* export: ordering replicates wherever re-measured, but margins compress and the hardness vertices are arbiter-split at T=0.7, and nothing is exported beyond the two measured temperatures; composition of guarded revision with block/semi-AR or KV-cached decoders is unmeasured in-protocol and stands as the recorded second-sampler-family candidate. At T=0.7 the extra-forward question decomposes by budget scale (DAX-103): the first extra forward is best spent on the guarded revision ($+12.37\%$ for one call); when forwards are free, cooled depth pays monotonically ($32 \to 64 = +12.909\%$ $[+10.924, +14.899]$, $\sim 5\times$ the T=1.0 pattern); at the 64-call cooled endpoint the two recipes are quality-tied (-0.62% $[-2.95, +1.63]$) with the revision reaching the tie at half the bill — efficient under a small forward budget, depth the continuation when spending is unconstrained. Gen-PPL remains the registered primary metric and the worth-having bar is still read within its declared proxy family, not as a property of generation quality writ large — and the scope is instrumented, not rhetorical: on this panel a fluent AR baseline at the same token scale reads MAUVE 0.588 (vs the family's 0.062–0.212) and gen-PPL 27.3 (vs the family's 230–1480) — the family is far from a fluent

AR anchor under both instruments, so every lever number above prices comparative within-family quality at matched bills, not absolute fluency (DAX-102).

Deployment on the matched-diversity frontier (DAX-115). Three priced points sit on this frontier at 32–33 calls: (i) *tilt* β=1 *at* T=1.0 — the NLL leader ($+11.8\%$ four-seed) and MAUVE leader (0.0730 vs uniform's 0.0621), at a measured per-doc diversity cost (unigram entropy -0.061 bits/doc, distinct-2 -0.0064, $+51\%$ intra-doc repeat mass, all outCI; DAX-115); (ii) *uniform order* — the diversity-premium point: pays the 11.8% NLL premium for $+0.061$ bits/doc unigram entropy, $+0.0167$ pooled distinct-2, and the lower repetition rate; (iii) *one-shot guarded revision at* 33 *calls* — the full-support $+30.754\%$ numerator over the *unmatched* tilt anchor, which at matched pooled-d2 loses to the 33-call nucleus-matched tilt on both measured arbiters (NLL -23.41%, DAX-109; MAUVE $0.0769 \rightarrow 0.1522$ tilt-favored, DAX-114) while keeping the per-doc richness edge priced at DAX-113. The recommendation is per arbiter-of-deployment: default (i); if the deployment prices per-doc diversity at more than ~ 0.15 NLL points per 0.01 bits/doc, (ii) is the premium-priced point and its premium is enumerated; the revision's case on this frontier is its diversity/richness profile rather than its NLL, and it requires the 33-call bill.

Grid expectations and materials. The grid's effect, if any, is bounded by the near-flat measured profile; both flat-ratios (≈ 1.56) cap reallocatable mass, so a $\geq 5\%$ win was registered as ambitious rather than expected. The protocol (thresholds fixed before each run, constructions frozen before use, and each pre-specified test evaluated once) is designed so that null and positive outcomes are reported alike. The two internal checkpoints are not redistributed; the public-substrate leg of the map (DAX-122) runs on unmodified public weights whose sha256 list is included in the supplementary archive (`models/`), so that half of the axis can be recomputed end-to-end by third parties. For the internal checkpoints, the supplementary archive contains the launch scripts (`runners/`) and the verdict scripts that implement the pre-specified gates, the frozen grid files (`runs/_derived/da_trace/`), the scored result files (`runs/_derived/`), and the experiment index (`EXPERIMENTS.md`); the checkpoints themselves are internal training materials (their full training protocol follows MDLM (Sahoo et al., 2024) on OpenWebText at 124M parameters, given in Appendix J).

X Additional operating-point detail

The following summaries (Tables 8 and 9) provide operating-point measurements omitted from the compact main text. They use the same models and scorer conventions as the preceding detailed tables. Ancestral uniform counts are Bernoulli; the order controls have deterministic counts, so matching calls does not by itself identify a pure position-order effect.

Table 8: The public-substrate commit-order cell (LLaDA-8B-Base, n=1024 documents per board, T=1.0, matched model-call budget): generation perplexity of the uniform and the soft-order decoders at the three measured call budgets. Win percentages use paired per-document bootstrap intervals relative to uniform decoding at the same budget (§3); every interval is positive. The in-house 64-call order premium at the same form is $+15.24\%$.

budget (model calls)	16	32	64
uniform decoding	180.54	97.64	70.66
soft commit order	161.49	86.01	54.54
win % vs uniform	+10.55	+11.91	+22.80

Grid and weighting sensitivity. The profiled grid's maximum/minimum weight ratio is 1.56. Its reference four-call perplexity is 1412.16. On the WikiText checkpoint, the re-derived grid lowers perplexity by 2.08% at 32 calls (DAX-12); on the timestep-weighted checkpoint, its self-derived grid lowers it by 5.38% at 8 calls (DAX-17). On the timestep-weighted checkpoint, the grid alone gives 0.84% at 32 calls and grid plus soft order gives 8.99% at 8 calls (DAX-26). Re-deriving the grid from that recipe changes its locations by at most 0.026 of a step; the pilot-set and alternative-weight results are retained in the grid sections above.

Table 9: Summary of decoding conditions at the stated model-call counts on the reference checkpoint (L=256, T=1.0, n=2048/cell, GPT-2-large gen-PPL; positive percentages denote perplexity reductions; the baseline and call count depend on the row. Order and grid rows use uniform decoding at matched calls; one-shot revision uses soft-order-32; later revision rows use the preceding revision variant). The tilt improves quality at a diversity/repetition price; the guarded revision ladder adds gen-PPL gains accompanied by diversity decline and uncertainty near the repetition bar in §4; the matched-distinct-2 tilt row quotes the revision-vs-matched-tilt flip of §4 (nucleus p=0.951; it beats one-shot revision on both scorer perplexity and MAUVE). The threshold row spans $\tau \in \{0.85, 0.90, 0.95\}$. Dashes mark quantities not listed in this summary.

channel	cell	gen-PPL	distinct-2	rep-4
order	uniform 32 (baseline)	615.20	0.5568	0.0016
	hard order 32	1285.04	—	—
	soft tilt β=1 32 (+13.10%)	534.62	0.5401	0.0025
	soft tilt β=1 32, T=0.7	29.58	—	—
placement	belief-profiled grid 32 (+1.06%, CI incl. 0)	608.70	0.5568	—
	belief-profiled grid 4 (+4.26%, excl. zero)	1412.16	—	—
trigger	confidence threshold (249–252 calls)	365.95–367.16	—	—
revision	one-shot reopen 16+1+16 (+32.10%)	363.00	0.468	—
	guarded 2-round (+23.06% vs one-shot)	279.28	0.426	0.0025
	guarded 3-round (+10.31% vs 2-rnd)	250.49	0.406	0.0024
	guarded 4-round (+3.21% vs 3-rnd)	242.44	0.397	0.0025
	matched-d2 tilt, p=0.951, 33 calls	294.14	0.465	0.0036

Temperature and order strength. At 32 calls and $T = 1.0$, the 124M model's perplexities for $\beta = 0.5, 1, 2$ are 557.93, 534.62, and 556.22. At $T = 0.7$, they are 31.91, 29.58, and 26.25, against the uniform anchor 35.82. The $\beta = 2$ versus 1 reduction is 11.23% $[9.81, 12.63]$; $\beta = 1$ versus 0.5 is 7.31% $[6.11, 8.48]$.

Distinct-2 matching sensitivity. The 0.02 per-document matching window retains 980 pairs and a reduction of 9.28% $[6.00, 12.41]$. The 0.05 window retains 1749 pairs and 10.57% $[8.08, 13.01]$. The 0.08 window retains 1986 pairs and 12.31% $[9.93, 14.62]$. These are selected-output comparisons and do not remove all diversity or selection differences.

Revision replication and comparison limits. On decode seed 780 the guarded three- and four-round variants reduce perplexity by 59.02% and 61.67% relative to uniform, and by 30.55% and 35.04% relative to one-shot revision. Their repetition point estimates lie below the stated threshold, but the original near-threshold uncertainty remains relevant. The corpus-distinct-2 comparison uses a nucleus-calibrated soft-order baseline ($p = 0.951$) against one-shot revision at 33 calls. Revision has signed reduction -23.41% $[-26.67, -20.23]$ relative to that baseline; MAUVE also favors the baseline (0.1522 versus 0.0769). This is not the per-document subset match used for soft versus ancestral uniform decoding. Where a table reports the inverse ratio (reverse baseline), Equation 2 gives the conversion; percentages with different denominators are not interchangeable.

Table 10: **Operating points on the reference 124M model** ($T = 1.0$, length 256, decode seed 777, 2,048 documents per row). Scorer PPL is the GPT-2-large perplexity of generated text; rep-4 is the within-document repeated 4-gram rate (reference threshold 2.5×10^{-3}); arrows mark the preferred direction. Ancestral decoding has random per-call counts; the other position-selection rows share one fixed count schedule. Revision rows build on soft order ($\beta = 1$); guarded rows use repetition guards (§4). All entries are point estimates (intervals for the key contrasts: Tables 2 and 4, Figures 2 and 3); bold marks the lowest scorer PPL among the 32-call position-selection rules, and dashes mark quantities not reported.

Decoder	Calls	Scorer PPL ↓	Distinct-2 ↑	Rep-4 (10^{-3}) ↓	MAUVE ↑
Position selection					
Ancestral	32	615.20	0.557	1.63	0.062
Fixed count, $\beta = 0$	32	606.10	0.557	–	–
Confidence-weighted, $\beta = 0.5$	32	557.93	0.545	–	–
Soft order, $\beta = 1$	32	**534.62**	0.540	2.50	0.073
Confidence-weighted, $\beta = 2$	32	556.22	0.539	–	–
Hard ranking ($\beta \to \infty$)	32	1285.04	0.602	–	0.018
Ancestral, profiled time grid	32	608.70	0.557	–	–
Budget references					
Soft order, $\beta = 1$	33	524.22	0.536	–	–
Ancestral	64	575.27	0.554	–	–
Soft order, $\beta = 1$	64	487.63	0.535	–	–
Ancestral	256	548.09	0.552	–	–
Revision on the soft-order sampler					
Interleaved re-masking, random	32	483.20	0.522	–	–
Interleaved re-masking, confidence	32	448.67	0.511	–	–
One-shot revision	33	363.00	0.468	2.12	0.077
Guarded revision, two rounds	32	279.28	0.426	2.50	0.125
Guarded revision, three rounds	33	250.49	0.406	2.43	0.155
Guarded revision, four rounds	36	242.44	0.397	2.46	0.110
Diversity-calibrated baseline (§4)					
Soft order, nucleus $p = 0.951$	33	294.14	0.465	3.63	0.152

Y ADDITIONAL METRICS FOR THE REFERENCE OPERATING POINTS

Table 10 extends Table 1 with within-document repeated 4-gram rates and MAUVE under the same conditions.