WHEN TO CORRECT: UPDATE WIDTH SHAPES TERMINAL REFINEMENT IN UNIFORM DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Update width changes the relative benefit of terminal refinement even when model-call counts are fixed. We compare correction during prediction with correction after prediction, using a fixed 33-step predictor grid and eight correction calls. Increasing width moves the terminal-to-late GenPPL ratio from 0.731 to 1.043 on LM1B and from 0.693 to 0.921 on OpenWebText. Larger updates improve both schedules, with late correction gaining more relatively. The direction observed on LM1B was predicted in advance and confirmed on OpenWebText. Later tests at intermediate widths in both models and on fresh seeds at a second correction budget support the same pattern. A separate LM1B study with 3,000 outputs per configuration compares terminal refinement with calibrated nucleus sampling. Terminal achieves 4.96% lower GenPPL and 14.69% shorter measured generation time, while nucleus achieves higher held-out prefix MAUVE. A placement comparison should therefore report update width alongside predictor and correction counts and test more than one width.

1 INTRODUCTION

Diffusion language models can interleave correction with prediction or refine the sequence after prediction finishes. A correction call can revise a few positions or a broad set for the same number of backbone evaluations. We ask how that update width changes the benefit of finishing prediction first.

Discrete diffusion generates categorical sequences through successive denoising transitions (Austin et al., 2021; Campbell et al., 2022; Lou et al., 2024). Predictors advance the reverse trajectory; correctors revise the current sequence. Informed correctors and finish-first refinement use these additional evaluations to improve generation (Zhao et al., 2025; Tian et al., 2025; Chen et al., 2026). We study the placement and width choices for a fixed correction operator.

Increasing update width reduces terminal refinement's likelihood advantage over late-interleaved correction. With 33 predictor calls and eight corrector calls, the LM1B terminal/late GenPPL ratio changes from 0.731 at width 3 to 1.043 at width 12, moving a large low-width advantage close to parity. On OpenWebText, the ratio changes from 0.693 at width 12 to 0.921 at width 48. Terminal still has lower GenPPL, but the advantage is substantially smaller.

LM1B development led us to predict the same attenuation on an independently trained OpenWebText model. We specified this direction before a new four-arm experiment, which confirmed it. Subsequent tests at an intermediate width in each model supported both adjacent increments. A later OpenWebText experiment used four correction calls and new seeds and supported the same direction at a second correction budget.

Temperature and nucleus sampling can substantially change diffusion likelihood (Olausson et al., 2026), so we compare terminal refinement against both. We then evaluate terminal and a selected nucleus configuration with nearby raw-token entropy on 3,000 new outputs each. Terminal has better evaluator likelihood and shorter measured generation time; nucleus better matches held-out text under prefix MAUVE. The practical choice therefore depends on the generation objective and the cost of each configuration.

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.

Our contributions are:

- A controlled comparison at fixed predictor and correction counts showing how larger updates shrink terminal refinement's relative likelihood advantage on LM1B.
- A successful prospective prediction on OpenWebText, followed by intermediate-width validation in both models and a fresh-seed test at a second correction budget.
- A comparison against strong temperature and nucleus controls that measures terminal refinement's tradeoff in likelihood, generation time, and distribution agreement.

2 METHOD

2.1 CORRECTION PLACEMENT ON A FIXED PREDICTOR GRID

Let $x \in \mathcal{V}^L$ be a sequence, P_j the native predictor transition at grid index j, and $C_t^{(k)}$ a corrector updating k selected positions at diffusion time t. The primary comparison fixes

$$N_P = 33, \qquad N_C = 8, \qquad B = N_P + N_C = 41.$$

We insert corrections into the native 33-call predictor grid. Even placement corrects after predictor indices

$$I_{\text{even}} = \{4, 8, 12, 16, 20, 24, 28, 32\},$$

while late placement uses

$$I_{\text{late}} = \{25, 26, 27, 28, 29, 30, 31, 32\}.$$

At each insertion, an interleaved correction uses the native next-state time. Terminal placement runs all 33 predictors before applying eight consecutive $C_0^{(k)}$ updates. Figure 1 shows the three schedules. At a given width, all schedules share the predictor grid, correction count, operator, and position-selection rule.

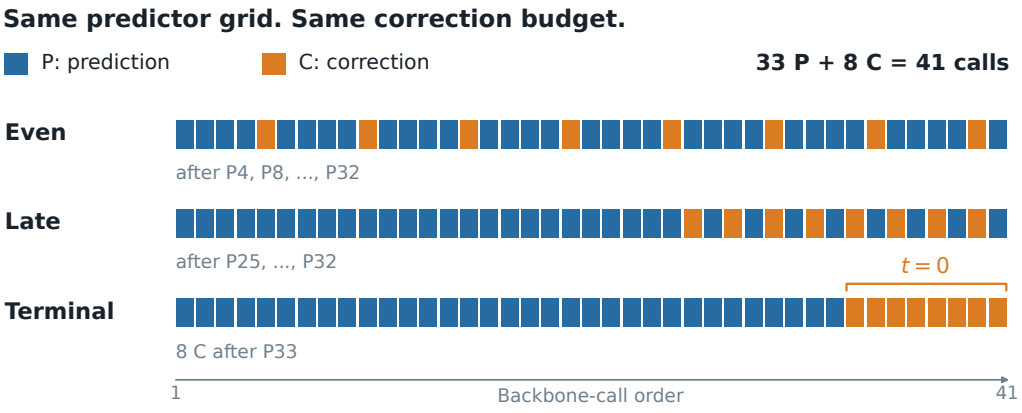


Figure 1: Correction placement with 33 predictor and eight corrector calls. Blue blocks advance prediction; orange blocks apply the same informed corrector to k selected positions. Even and late schedules insert corrections at the indicated predictor indices. Terminal applies all corrections at time zero after prediction finishes. Each schedule uses 41 backbone calls.

The four-call follow-up keeps the predictor grid and uses late indices $\{26, 28, 30, 32\}$, or four terminal updates after P_{33}, for 37 calls per configuration. Placement determines the state presented to the corrector, its conditioning time, and the operations that follow. The sequence therefore evolves differently under each policy.

2.2 INFORMED PARALLEL CORRECTION

We use the released correction operator (Zhao et al., 2025; Gourevitch et al., 2026). Let $q_{\theta,i}^{\text{clean}}(v \mid x, t)$ denote the denoiser's clean conditional: the leave-one-out conditional on OpenWebText and the released uniform plug-in denoiser on LM1B. The correction interface forms

$$q_{t,i}(v \mid x) = \alpha_t q_{\theta,i}^{\text{clean}}(v \mid x, t) + \frac{1 - \alpha_t}{|\mathcal{V}|}.$$

One backbone evaluation supplies this conditional tensor. At time zero, $\alpha_0 = 1$. For each position, the selector compares the best alternative with the current token:

$$s_i = \max_{v \neq x_i} \log q_{t,i}(v \mid x) - \log q_{t,i}(x_i \mid x).$$

Gumbel-perturbed scores select k distinct positions,

$$S = \text{TopK}_i(s_i/\tau + g_i, k), \qquad g_i \sim \text{Gumbel}(0, 1), \qquad \tau = 1.$$

We sample replacements from the corresponding conditionals and apply them simultaneously, then recompute the conditionals on the revised sequence at the next correction call. A replacement may keep the current token, and later calls may revisit a position. The operator stays fixed across placement comparisons.

2.3 UPDATE WIDTH AND THE PLACEMENT INTERACTION

The primary endpoint comparison uses $k = 3, 12$ at LM1B length $L = 128$ and $k = 12, 48$ at OpenWebText length $L = 512$. Both compare per-call widths $k/L = 0.0234$ and 0.0938. The intermediate settings are LM1B $k = 6$ and OpenWebText $k = 24$.

For C correction calls, the cumulative nominal proposal budget is $d = Ck/L$. At $C = 8$, the three widths give $d = 0.1875, 0.375, 0.75$; the four-call OpenWebText endpoints give $d = 0.0938, 0.375$. This budget counts all proposals, including repeated positions and replacements that leave a token unchanged. Increasing width at fixed correction count therefore increases the cumulative proposal budget as well.

Let $\ell_{A,k;C}$ be pooled evaluator NLL per target for placement A, width k, and correction count C. We measure the interaction by

$$I_C = (\ell_{\text{terminal},k_{\text{high}};C} - \ell_{\text{late},k_{\text{high}};C}) - (\ell_{\text{terminal},k_{\text{low}};C} - \ell_{\text{late},k_{\text{low}};C}).$$

The interaction measures how the log terminal-to-late GenPPL ratio changes between widths. Positive values indicate that terminal's relative advantage shrinks as updates widen. We specified $I_8 > 0$ for the prospective OpenWebText test and, separately, $I_4 > 0$ for the later four-call study.

3 EXPERIMENTAL SETUP

3.1 MODELS AND EXECUTION

We evaluate two independently trained models from the DiT architecture family, with different corpora, tokenizers, objectives, and tested lengths. The LM1B checkpoint, `kuleshov-group/udlm-lm1b`, has 139,292,986 parameters, length 128, and a BERT uncased tokenizer (Chelba et al., 2014; Schiff et al., 2025). It uses continuous-time uniform plug-in ELBO training with zero reconstruction loss. The OpenWebText CE EMA checkpoint has 169,627,218 parameters and a GPT-2 byte-level tokenizer with an added padding token (Gourevitch et al., 2026). It was trained at length 1024; our width and correction-budget experiments use length 512. Additional length-1024 experiments use the released CE and ELBO checkpoints.

We generate in FP32 with batch size one on an RTX 5070 Ti. CPU and GPU FP32 reference checks show agreement between the original and adapted inference on the tested fixed inputs and trajectories. A separate native BF16/FlashAttention run with the public LM1B weights gives GenPPL 189.09, compared with 79.80 in the published Table 36. The exact published checkpoint identity and the source of this absolute gap remain unresolved. The appendix records these comparisons, model revisions, and runtime details.

3.2 Experimental cohorts

Each width experiment uses 128 paired seeds per model. Following a separate pilot, LM1B confirmation compares eleven configurations: even, late, and terminal at both endpoint widths, four prediction-only temperatures, and nucleus sampling at $p = 0.95$. Its six likelihood contrasts are primary, and its width interaction is a predefined secondary confirmation. On OpenWebText, a new four-arm experiment compares terminal and late at $k = 12, 48$, with the predicted width interaction as the single primary contrast.

After observing the endpoint results, we add terminal and late at $k = 6$ on LM1B and $k = 24$ on OpenWebText. These intermediate-width runs retain the original seed labels and endpoint outputs, adding 512 outputs to the 1,024 existing endpoints. A separate four-call OpenWebText study uses 128 entirely new paired seeds and produces 512 outputs; its direction, sample size, and analysis were specified before generation. Terminal and late share the same call count within each correction budget. C4 and C8 use different cohorts, total calls, and late insertion densities.

The separate LM1B distribution study uses 3,000 new paired seeds for terminal $k = 3$, late $k = 3$, and nucleus $p = 0.95$. We selected the prediction-only control before this study by taking the lowest confirmation GenPPL among configurations within 0.02 nats of terminal mean raw-token entropy. Temperature 0.97 and nucleus 0.95 both qualified; nucleus was selected. Terminal/nucleus likelihood is the single primary contrast, and all three configurations use 41 backbone calls. The appendix documents cohort identities and the earlier calibration and temperature experiments.

3.3 Metrics and uncertainty

GenPPL measures predictability under frozen GPT-2-large (Radford et al., 2019). For evaluator NLL N_s and effective target count M_s,

$$\mathrm{GenPPL} = \exp\left(\frac{\sum_s N_s}{\sum_s M_s}\right).$$

We also report mean raw-token unigram entropy, four-gram repetition, and synchronized whole-generation time. Entropy measures lexical concentration. Generation time excludes model loading, decoding, and evaluator scoring. We assess distribution agreement separately with MAUVE (Pillutla et al., 2021).

For likelihood intervals, we draw 20,000 paired seed-bootstrap resamples and recompute pooled NLL and target counts on each draw. Pairing stays within each model and cohort. The six primary LM1B contrasts use individual 99.1667% intervals, Bonferroni-adjusted for 95% familywise coverage; its interaction and other comparisons use descriptive 95% intervals. The prospective OpenWebText interaction, the four-call interaction, and the larger study's terminal/nucleus ratio each use a 95% interval for one primary contrast. The four adjacent-width comparisons use individual 98.75% intervals with Bonferroni adjustment.

The distribution study compares each 3,000-output configuration with 3,000 held-out LM1B references. The primary view contains the first 32 non-special BERT wordpieces from each eligible text; full decoded text is the secondary view. A disjoint group of 3,000 reference rows supplies a real-versus-real check. MAUVE uses GPT-2-large features, 100 clusters, and five fixed clustering restarts. We report their mean and range to describe restart sensitivity on the same texts. The appendix gives preprocessing details and every restart result.

4 RESULTS

4.1 WIDER UPDATES SHRINK TERMINAL'S ADVANTAGE ON LM1B

At width 3, terminal has GenPPL 121.906 compared with late placement's 166.807. The ratio is 0.731, with primary 99.1667% interval [0.695, 0.769]. At width 12, both improve, to 102.106 and 97.864, respectively. Their ratio is now 1.043 [0.980, 1.108], placing the schedules close to parity after terminal's large low-width advantage.

The interaction is $I_{\mathrm{LM1B}} = 0.356$ nats, with descriptive 95% interval [0.302, 0.411]. Increasing width benefits both schedules, with late gaining more relatively. Even placement improves from 179.845 to 121.466 but remains behind terminal at both widths. Table 1 reports the endpoint comparisons.

Table 1: Placement-by-width results on separate 128-seed model cohorts. Every configuration uses 33 predictor and eight corrector calls. Terminal/late intervals are descriptive paired 95%; the OpenWebText interaction is the primary prospective prediction. Nominal proposal budget is $8k/L$.

Model	**Width** k	$8k/L$	**Terminal GenPPL**	**Late GenPPL**	**Terminal/late [95% interval]**
LM1B	3	0.1875	121.9064	166.8073	0.7308 [0.7040, 0.7583]
LM1B	12	0.75	102.1063	97.8643	1.0433 [0.9969, 1.0908]
OpenWebText	12	0.1875	92.7850	133.9108	0.6929 [0.6788, 0.7065]
OpenWebText	48	0.75	68.4847	74.3699	0.9209 [0.8943, 0.9470]

4.2 THE PREDICTED ATTENUATION APPEARS ON OPENWEBTEXT

The prospective OpenWebText experiment gives the following interaction:

$$I_{\mathrm{OWT}} = 0.284 \quad [0.251,\ 0.317].$$

The terminal/late GenPPL ratio rises from 0.693 at width 12 to 0.921 at width 48, giving a ratio of ratios of 1.329 [1.286, 1.373]. The stronger relative improvement of late placement matches the prediction from LM1B, while terminal retains lower GenPPL at both widths.

Wider updates also produce more concentrated and repetitive text. Terminal entropy decreases from 5.116 to 4.926 nats and repeat4 rises from 1.300% to 2.545%; late changes in the same direction. We consider these token statistics alongside likelihood and the held-out distribution comparison.

4.3 THE ATTENUATION PERSISTS ACROSS WIDTHS AND CORRECTION BUDGETS

The intermediate points fall between their model's endpoints (Figure 2). Terminal/late ratios are 0.731, 0.832, and 1.043 for LM1B widths 3, 6, and 12, and 0.693, 0.787, and 0.921 for OpenWebText widths 12, 24, and 48. All four adjacent log-ratio increments have positive adjusted 98.75% intervals, showing attenuation at both tested increments in each model.

On new OpenWebText seeds with four correction calls, the terminal/late ratio rises from 0.729 at $k = 12$ to 0.877 at $k = 48$. The prespecified interaction is $I_4 = 0.184$ nats [0.160, 0.209]. Terminal remains better at both widths, while late gains more relatively, supporting the same direction at a second correction-call budget. Tables 31 and 32 give the complete scores and intervals.

4.4 COMPARISON WITH TEMPERATURE AND NUCLEUS SAMPLING

Lowering prediction-only temperature from 1.00 to 0.90 reduces LM1B GenPPL from 205.134 to 121.150. Terminal $k = 3$ reaches a similar 121.906 with higher raw-token entropy, 4.196 versus 4.096 nats. Terminal $k = 12$ improves further to 102.106; its adjusted ratio against temperature 0.90 is 0.843 [0.750, 0.952].

Nucleus $p = 0.95$ and terminal $k = 3$ have close confirmation entropies, 4.192 and 4.196 nats. The terminal/nucleus likelihood ratio is 0.968 [0.894, 1.052]. We carry this comparison into the larger study on new outputs. Figure 3 shows all eleven configurations at the same 41-call budget.

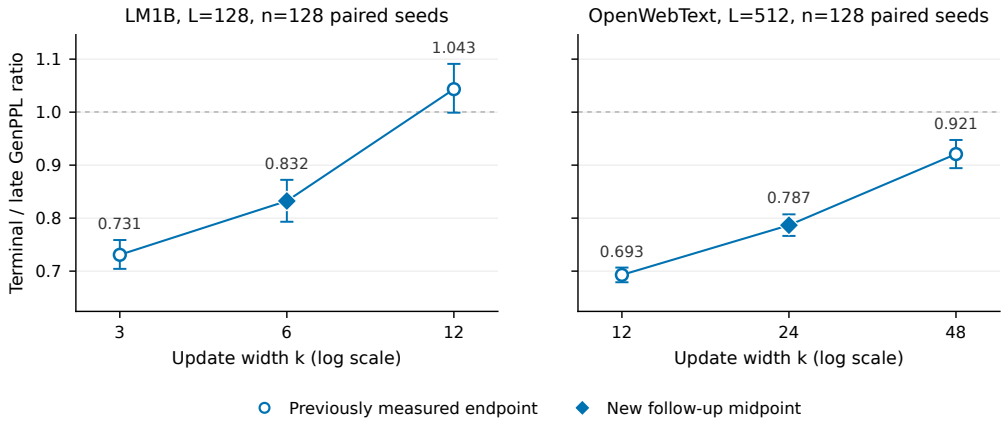


Figure 2: Terminal-to-late GenPPL across three update widths. Open circles mark the original endpoints; filled diamonds mark subsequent midpoint experiments using the same seed labels. Values below one favor terminal. Each setting uses 128 paired seeds and $33P+8C$. Error bars are descriptive 95% paired-bootstrap intervals; connecting lines join measured settings. The four adjacent-width tests use adjusted 98.75% intervals.

4.5 Likelihood gains come with a distribution tradeoff

On 3,000 new outputs per configuration, terminal $k = 3$ has GenPPL 125.626 compared with nucleus sampling's 132.186. The primary ratio is 0.9504 [0.9326, 0.9686], a 4.96% reduction in evaluator perplexity. Measured generation time is 14.69% shorter. Table 2 also reports distribution agreement and token statistics.

Table 2: Likelihood, time, token statistics, and prefix distribution agreement on the 3,000-seed LM1B study. Each configuration uses 41 calls. Prefix32 MAUVE compares 3,000 generated and 3,000 held-out prefixes; brackets give the minimum and maximum over five clustering restarts.

Configuration	**GenPPL ↓**	**H, nats**	**repeat4, % ↓**	**Seconds ↓**	**Prefix32 MAUVE ↑ [restart range]**
Terminal, $k = 3$	125.6262	4.1557	0.6677	0.5050	0.6555 [0.6306, 0.6768]
Late, $k = 3$	168.6090	4.2003	0.4907	0.5025	0.6203 [0.5863, 0.6547]
Nucleus, $p = 0.95$	132.1857	4.1715	0.6664	0.5919	0.7008 [0.6785, 0.7183]

Nucleus has higher prefix MAUVE than terminal, and terminal exceeds late, at every clustering restart. The real-versus-real mean is 0.9955. Terminal's better evaluator likelihood and measured time accompany lower agreement with held-out prefix distributions than nucleus, while their repeat4 values are nearly identical. The time comparison uses the supplied nucleus implementation with full-vocabulary sorting.

4.6 Controls for edit retention and replacement noise

At length 1024, terminal improves over even placement under both CE and ELBO objectives on OpenWebText, with terminal/even ratios of 0.7255 and 0.7015, respectively. Late and terminal retain approximately 85% and 84% of changed-token events in the final sequence despite their likelihood gap. Removing explicit uniform noise from late replacement probabilities leaves most of that gap intact. Final edit retention and replacement noise alone are therefore insufficient explanations.

The strongest complete configuration varies with the checkpoint. Calibrated nucleus reaches ELBO GenPPL 55.838 compared with terminal's 64.043. In a separate prospective CE temperature study, terminal improves over even placement and the locked temperature control, while the nearest-entropy sampled temperature is a closer comparator. The appendix gives these additional controls, complete distribution results, and auxiliary model-based text assessments.

5 Related Work

Discrete diffusion defines categorical corruption and reverse-generation processes (Austin et al., 2021; Campbell et al., 2022). Score-entropy models use distributional ratios, while masked models use absorbing states (Lou et al., 2024; Sahoo et al., 2024). We use the OpenWebText formulation, checkpoints, and operators from Uniform Diffusion Models Revisited (Gourevitch et al., 2026) and the independently trained LM1B checkpoint released with discrete diffusion guidance (Schiff et al., 2025).

Informed Correctors studies parallel model-informed updates, including deterioration at high update width (Zhao et al., 2025). GADD examines Gibbs correction and its relationship with prediction (Liang et al., 2026). ReMDM studies remasking and inference-time scaling (Wang et al., 2025), while Tolerator and Speculative Correction develop finish-first and draft-then-refine approaches (Tian et al., 2025; Chen et al., 2026).

The gains from confidence remasking vary with block length and stochasticity (Frković et al., 2026). CaRE tracks actual NFE and studies temperature and remasking fraction (Shah et al., 2026); CRS studies when selected tokens become persistent context and how visibility duration interacts with truncation (Hayakawa, 2026). Our comparison addresses placement and per-call update width while holding the corrector, predictor count, and correction count fixed.

Temperature and top-p controls are central to diffusion sampling (Olausson et al., 2026; Gourevitch et al., 2026). We evaluate them on separate selection and evaluation cohorts. The metrics describe different generation objectives (Pynadath et al., 2026): GPT-2 likelihood measures evaluator predictability (Radford et al., 2019), and MAUVE measures agreement between embedded generated and reference distributions (Pillutla et al., 2021).

6 Conclusion

A large terminal-refinement advantage under narrow updates can become much smaller at broader widths even with the predictor grid and correction-call count fixed. The observed direction supports a prospective prediction on an independently trained checkpoint, followed by tests at intermediate widths and a second correction budget.

Correction-schedule comparisons should report predictor count, correction count, and update width together and evaluate placement at more than one width. The generation objective also affects the choice: terminal has better evaluator likelihood and measured time in the larger comparison, while nucleus has higher prefix MAUVE.

AI USE STATEMENT

The author used Codex to assist with hypothesis refinement, experiment design, implementation, analysis, and manuscript preparation. Code execution, numerical results, and cited sources were checked. The author takes responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

The appendix specifies the models, correction algorithm, experimental cohorts, calibration, evaluation, and statistical procedures, and reports complete comparisons and implementation checks. The accompanying records contain the frozen protocols, model and source revisions, generated outputs, evaluator scores, analysis code, and detailed verification logs.

REFERENCES

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured Denoising Diffusion Models in Discrete State-Spaces. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/958c530554f78bcd8e97125b70e6973d-Abstract.html.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, George Deligiannidis, and Arnaud Doucet. A Continuous Time Framework for Discrete Denoising Models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 28266–28279. Curran Associates, Inc., 2022. doi: 10.52202/068431-2049. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b5b528767aa35f5b1a60fe0aaeca0563-Abstract.html.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling. In *Interspeech*, pp. 2635–2639, 2014. doi: 10.21437/Interspeech.2014-564. URL https://www.isca-archive.org/interspeech_2014/chelba14_interspeech.html.

Brian K Chen, Chong Wu, and Kenji Kawaguchi. Speculative Correction: Draft-then-Refine Decoding for Diffusion Language Models, 2026. URL https://arxiv.org/abs/2608.02625.

Stipe Frković, Metod Jazbec, Dan Zhang, Christian A. Naesseth, Ilija Bogunovic, and Eric Nalisnick. Re-evaluating Confidence Remasking in Masked Diffusion Language Models, 2026. URL https://arxiv.org/abs/2606.12232.

Samson Gourevitch, Yazid Janati, Dario Shariatian, Umut Simsekli, Eric Moulines, Eric P. Xing, and Alain Durmus. Uniform Diffusion Models Revisited: Leave-One-Out Denoiser and Absorbing State Reformulation, 2026. URL https://arxiv.org/abs/2605.22765.

Satoshi Hayakawa. From Truncation to Commitment: Persistent Context in Uniform Discrete Diffusion, 2026. URL https://arxiv.org/abs/2609.01043.

Yuchen Liang, Ness Shroff, and Yingbin Liang. From Scores to Gibbs Correctors: Accelerating Uniform-Rate Discrete Diffusion Models, 2026. URL https://arxiv.org/abs/2605.27352.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete Diffusion Modeling by Estimating the Ratios of the Data Distribution. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 32819–32848. PMLR, 2024. URL https://proceedings.mlr.press/v235/lou24a.html.

Theo X. Olausson, Metod Jazbec, Xi Wang, Armando Solar-Lezama, Christian A. Naesseth, Stephan Mandt, and Eric Nalisnick. A Tale of Two Temperatures: Simple, Efficient, and Diverse Sampling from Diffusion Language Models, 2026. URL https://arxiv.org/abs/2604.09921.

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. In *Advances in Neural Information Processing Systems*, volume 34, pp.

4816–4828. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/260c2432a0eecc28ce03c10dadc078a4-Abstract.html.

Patrick Pynadath, Jiaxin Shi, and Ruqi Zhang. Generative Frontiers: Why Evaluation Matters for Diffusion Language Models, 2026. URL https://arxiv.org/abs/2604.02718.

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. In *Advances in Neural Information Processing Systems*, volume 37, pp. 130136–130184. Curran Associates, Inc., 2024. doi: 10.52202/079017-4135. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/eb0b13cc515724ab8015bc978fdde0ad-Abstract-Conference.html.

Yair Schiff, Subham Sekhar Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dalla-Torre, Bernardo P. de Almeida, Alexander M. Rush, Thomas Pierrot, and Volodymyr Kuleshov. Simple guidance mechanisms for discrete diffusion models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/6cc31b44d88dce8380d36e81485cd07f-Abstract-Conference.html.

Yash Shah, Abhijit Chakraborty, and Vivek Gupta. CaRE: Compute-aware Remasking Evaluation Protocol for Masked Diffusion Language Models, 2026. URL https://arxiv.org/abs/2607.24763.

Runchu Tian, Junxia Cui, Xueqiang Xu, Feng Yao, and Jingbo Shang. Finish First, Perfect Later: Test-Time Token-Level Cross-Validation for Diffusion Large Language Models, 2025. URL https://arxiv.org/abs/2510.05090.

Guanghan Wang, Yair Schiff, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Remasking Discrete Diffusion Models with Inference-Time Scaling. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 147282–147339. Curran Associates, Inc., 2025. doi: 10.52202/085713-4927. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/d877ea0dd78bdbe54830670618c1de09-Abstract-Conference.html.

Yixiu Zhao, Jiaxin Shi, Feng Chen, Shaul Druckmann, Lester Mackey, and Scott Linderman. Informed Correctors for Discrete Diffusion Models. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 125510–125538. Curran Associates, Inc., 2025. doi: 10.52202/085713-4183. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/b67d9d07faf071532c420ed5b140756b-Abstract-Conference.html.

A SUPPLEMENTARY MATERIAL

A.1 MODEL IMPLEMENTATIONS AND CORRECTION ALGORITHM

Experiments use the released `kuleshov-group/udlm-lm1b` and `samsongourevitch/rev_udm` models. Exact checkpoint and source revisions accompany the reproducibility records.

The original corrector evaluates its mixture in float64 log space:

$$\log q_{t,i}(v \mid x) = \operatorname{logaddexp}\left(\log \alpha_t + \log q_{\theta,i}^{\text{clean}}(v \mid x, t), \log(1-\alpha_t) - \log|\mathcal{V}|\right).$$

The backbone remains FP32 in the main comparisons. Replacement draws may retain the current token. Each call selects distinct positions, which may recur across calls.

For LM1B prediction-only controls, temperature scales native backbone logits before normalization. Nucleus sampling sorts, truncates, and renormalizes denoiser probabilities before constructing the native predictor posterior.

Algorithm A1. Fixed-composition placement with specified correction count and general update width.

```
Input: frozen model, length L, update width k, correction count C,
       placement A, predictor randomness rP,
       correction randomness rC
G ← native grid for 33 predictor calls
Disable internal correction scheduling
x ← native initialization using rP

I ← {4,8,12,16,20,24,28,32} if C=8 and A is Even
     {25,26,27,28,29,30,31,32} if C=8 and A is Late
     {26,28,30,32} if C=4 and A is Late
     empty set if A is Terminal

for j = 1,...,33:
    x ← native predictor Pj(x; G, rP)
    if j belongs to I:
        t ← native next-state time after Pj
        x ← inherited corrector Ct(x; k, tau=1, rC)

if A is Terminal:
    repeat C times:
        x ← inherited corrector C0(x; k, tau=1, rC)

return x
```

The four-call experiment compares terminal and late placement. Figure 1 shows the eight-call schedules; the length-1024 placement experiments use $k = 12$.

Predictor initialization and transitions use a dedicated CUDA generator seeded with s. Correction uses the global CUDA stream seeded with $s + 900000$.

Native PC1 and PC2 use the global CUDA stream because their released wrapper does not forward the passed generator. Predictor draws differ across methods even for paired outputs with the same seed label.

A.2 EXPERIMENTAL COHORTS AND STATISTICAL ESTIMANDS

Table 3: Generation cohorts and inference roles. Seed counts are per arm, except for the explicitly batch-seeded native baseline.

Cohort	Seeds	Role
Historical calibration	83500–83515	16-seed entropy calibration
Historical length-1024 placement	84000–84047	48-seed exploration per checkpoint
Historical CE length-512 transfer	84000–84047	48-seed unchanged-setting transfer
Previous CE temperature comparison	86000–86047	Separate 48-seed exploration
Historical prospective CE temperatures	87000–87127	128 seeds, six frozen arms
Independent LM1B confirmation	91200–91327	128 seeds, eleven frozen arms
Prospective OpenWebText width prediction	93500–93627	128 seeds, four frozen arms
LM1B distribution extension	94000–96999	3,000 seeds, three frozen arms
Precision/step diagnostic	97300–97315	16 seeds, four diagnostic arms
Later LM1B midpoint validation	91200–91327	Original seed labels; two new midpoint arms
Later OpenWebText midpoint validation	93500–93627	Original seed labels; two new midpoint arms
Native BF16/FlashAttention baseline	Batch seeds 97800–97803	Four batches of 32 outputs; 128 outputs total
Later OpenWebText four-call follow-up	99000–99127	128 new paired seeds, four frozen arms

LM1B confirmation follows a separate 24-seed pilot and nucleus development. It excludes development and smoke outputs; the reproducibility records list the excluded smoke seeds.

Before LM1B confirmation, nucleus cutoff was selected on pilot seeds 91000–91015 from $p \in \{0.90, 0.95, 0.99, 1.00\}$.

Selection minimized the absolute raw-token entropy difference from terminal $k = 3$, fixing $p = 0.95$. It used neither GenPPL nor MAUVE.

The $p = 1$ candidate reused its pilot outputs; each other candidate generated 16 outputs on the same seeds.

Length-1024 calibration targets terminal entropy within ± 0.05 nats. It tests top-p $\{0.90, 0.97, 1.00\}$ and temperatures $\{0.90, 1.00, 1.10\}$, interpolates between bracketing points, and rounds to four decimals.

When multiple brackets exist, calibration selects the smallest entropy span. Locked ELBO settings are top-p 0.9398 and temperature 0.9607; CE settings are top-p 0.9733 and temperature 0.9695. Length-512 transfer retains these settings.

The distribution extension fixes $k = 3$ before confirmation. After confirmation, its control is selected for the lowest GenPPL among prediction-only settings within 0.02 nats of terminal mean raw-token entropy.

For raw-token frequencies $\hat{p}_x(v)$,

$$H(x) = -\sum_v \hat{p}_x(v) \log \hat{p}_x(v), \qquad R_4(x) = 1 - \frac{|\,\mathrm{unique4grams}(x)|}{L - 3}.$$

Both metrics are sequence means. Tables display $100R_4$ as a percentage unless explicitly labeled as a fraction. Paired repetition changes are labeled as fractions or percentage points.

For paired bootstrap seed indices s^*, the likelihood ratio is recomputed as

$$\rho^* = \exp\left(\frac{\sum_{s^*} N_{A,s^*}}{\sum_{s^*} M_{A,s^*}} - \frac{\sum_{s^*} N_{B,s^*}}{\sum_{s^*} M_{B,s^*}}\right).$$

The original width-interaction analyses jointly resample seed indices across all four cells. Midpoint validation preserves pairing across widths within each model, while the four-call interaction uses a new paired cohort.

Each model and correction-budget cohort is analyzed separately. Likelihood ratios use the pooled NLL and target totals defined above.

All likelihood analyses use 20,000 paired bootstrap resamples, with analysis-specific random seeds listed in the reproducibility records.

A.3 LM1B CONFIRMATION

Table 4: All eleven LM1B confirmation arms, 128 outputs each. Placement uses $33P + 8C$; controls use $41P$. Padding and common-evaluator truncations are zero; targets count evaluated tokens.

Configuration	GenPPL	H, nats	repeat4, %	Seconds	Targets
Terminal $k = 12$	102.1063	4.1172	0.7000	0.5393	15764
Even $k = 12$	121.4660	4.1779	0.3750	0.5383	15885
Late $k = 12$	97.8643	4.1286	0.6625	0.5407	15732
Terminal $k = 3$	121.9064	4.1958	0.4000	0.5392	15834
Even $k = 3$	179.8449	4.2230	0.5125	0.5391	15916
Late $k = 3$	166.8073	4.2393	0.2812	0.5423	15893
Temperature 0.90	121.1497	4.0958	1.0375	0.5291	15818
Temperature 0.94	146.6715	4.1645	0.6438	0.5318	15911
Temperature 0.97	174.5846	4.2022	0.5375	0.5308	15946
Temperature 1.00	205.1340	4.2516	0.5563	0.5274	16014
Nucleus 0.95	125.8944	4.1917	0.7375	0.6341	15967

Table 5: Six primary LM1B likelihood contrasts. Individual 99.1667% intervals apply Bonferroni adjustment across six comparisons; 95% intervals are descriptive.

Contrast	GenPPL ratio	Primary 99.1667% interval	Descriptive 95% interval
Terminal12 / Even12	0.8406	[0.7521, 0.9371]	[0.7748, 0.9125]
Terminal12 / Late12	1.0433	[0.9805, 1.1076]	[0.9969, 1.0908]
Terminal12 / Temperature0.90	0.8428	[0.7499, 0.9525]	[0.7735, 0.9213]
Terminal3 / Even3	0.6778	[0.6156, 0.7488]	[0.6301, 0.7288]
Terminal3 / Late3	0.7308	[0.6955, 0.7688]	[0.7040, 0.7583]
Terminal3 / Temperature0.94	0.8312	[0.7467, 0.9258]	[0.7690, 0.8994]

Table 6: Remaining LM1B likelihood comparisons and entropy differences. All intervals are descriptive 95%; differences subtract the comparator from terminal.

Contrast	GenPPL ratio [95% interval]	ΔH [95% interval]
Terminal12 / Temperature0.94	0.6962 [0.6419, 0.7548]	−0.0473 [−0.0790, −0.0129]
Terminal12 / Temperature0.97	0.5849 [0.5389, 0.6352]	−0.0850 [−0.1145, −0.0538]
Terminal12 / Temperature1.00	0.4978 [0.4588, 0.5407]	−0.1344 [−0.1654, −0.1023]
Terminal12 / Nucleus0.95	0.8110 [0.7468, 0.8817]	−0.0745 [−0.1060, −0.0416]
Terminal3 / Temperature0.90	1.0062 [0.9230, 1.1015]	0.1001 [0.0697, 0.1332]
Terminal3 / Temperature0.97	0.6983 [0.6442, 0.7575]	−0.0064 [−0.0314, 0.0212]
Terminal3 / Temperature1.00	0.5943 [0.5481, 0.6456]	−0.0558 [−0.0815, −0.0289]
Terminal3 / Nucleus0.95	0.9683 [0.8938, 1.0515]	0.0041 [−0.0220, 0.0328]

Terminal-minus-late entropy differences are −0.0434 [−0.0538, −0.0332] at $k = 3$, and −0.0114 [−0.0275, 0.0045] at $k = 12$. Their entropy interaction is 0.0320 [0.0145, 0.0496].

Repeat4 differences, as fractions, are 0.0012 [0.00031, 0.0021] and 0.00038 [−0.0013, 0.0020], respectively.

Terminal12 minus temperature0.90 has entropy difference 0.0215 [−0.0119, 0.0575] and generation-time difference 0.0102 [0.0048, 0.0154] seconds.

Terminal3 minus temperature0.94 has entropy difference 0.0313 [0.0035, 0.0620] and time difference 0.0075 [0.0019, 0.0128] seconds.

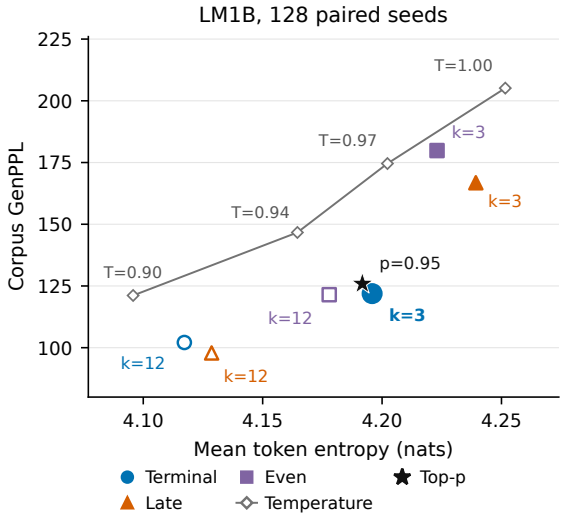


Figure 3: LM1B likelihood–entropy tradeoffs across eleven configurations (128 paired seeds, 41 calls). Markers distinguish schedules and widths; connected points show temperatures, with nucleus $p = 0.95$ separate.

The temperature controls are slightly faster in this confirmation runtime even where terminal has lower GenPPL.

A.4 Prospective OpenWebText width prediction

Table 7: Four-cell OpenWebText width study, 128 outputs per arm. GenPPL intervals are descriptive 95%. All arms use $33P + 8C$, temperature one, and no nucleus truncation.

Configuration	GenPPL [95% interval]	H, nats	repeat4, %	Seconds	Targets
Terminal12	92.7850 [86.2991, 99.8459]	5.1159	1.3000	1.6411	65090
Late12	133.9108 [124.0172, 144.4436]	5.1934	1.0376	1.6548	65144
Terminal48	68.4847 [64.5184, 72.5667]	4.9257	2.5448	1.6529	64993
Late48	74.3699 [69.9476, 79.1131]	5.0167	1.7943	1.6290	65089

Pooled total NLL values are 294876.2507, 319021.5081, 274700.0465, and 280471.8282 in table order. Padding rate is zero throughout.

At $k = 12$, terminal-minus-late entropy is -0.0775 [-0.0841, -0.0710], and the repeat4 fraction difference is 0.0026 [0.0018, 0.0035]. The time ratio is 0.9917 [0.9765, 1.0061].

At $k = 48$, entropy and repeat4 differences are -0.0910 [-0.1053, -0.0773] and 0.0075 [0.0054, 0.0098]. The time ratio is 1.0146 [0.9996, 1.0301]. These are secondary comparisons.

A.5 Distribution extension and reference construction

All 3,000 outputs per arm have scores and features in both views. Each contains 128 nonpadding generation tokens; none is empty, truncated by the evaluator, or shorter than the required prefix.

Mean GPT-2 retokenized lengths are 124.8730 for terminal, 125.5687 for late, and 125.3437 for nucleus.

Table 8: Larger-extension likelihood totals. Target counts are exact; pooled log GenPPL follows the ratio of the two saved totals.

Configuration	Total NLL	Evaluator targets	Pooled log GenPPL
Terminal3	1796150.2495	371619	4.8333
Late3	1916208.3757	373706	5.1276
Nucleus0.95	1821960.7889	373031	4.8842

Terminal minus nucleus entropy is -0.0157 [-0.0223, -0.0092] nats. Its repeat4 fraction difference is 0.000013 [-0.00077, 0.00080], and its time difference is -0.0870 [-0.0880, -0.0858] seconds.

The secondary terminal/late GenPPL ratio is 0.7451 [0.7393, 0.7509]. Entropy and repeat4 differences are -0.0445 [-0.0468, -0.0423] and 0.0018 [0.0015, 0.0021], respectively.

Terminal minus late generation time is 0.0024 [0.0015, 0.0034] seconds.

Reference texts come from the pinned `dvruette` LM1B test mirror, with 306,688 rows. We apply the official LM1B detokenizer, encode with uncased BERT and CLS/SEP, truncate and right-pad to 128, then decode with special tokens skipped.

Prefix eligibility requires at least 32 content tokens after encoding. Content selection removes every ID in `BertTokenizerFast.all_special_ids`, including PAD, CLS, SEP, MASK, and UNK, leaving 110,757 eligible rows.

A uniform reservoir sample of 6,000 eligible rows, using seed 202609083232, is randomly split into disjoint 3,000-row reference and sanity groups. Selection is independent of generated text and metric results.

Prefixes contain the first 32 content BERT wordpieces and may decode to an incomplete final word. Every frozen generation seed contributes once; all generated, reference, and sanity prefixes have sufficient content and valid features.

MAUVE features are GPT-2-large FP32 last-layer vectors at the last valid token, with a maximum of 256 GPT-2 tokens. The original `mauve.compute_mauve` consumes precomputed features using CPU FAISS/PCA.

Settings are 100 buckets, scaling factor 5, `kmeans_explained_var=0.9`, `kmeans_num_redo=5`, `kmeans_max_iter=500`, `pca_max_data=-1`, and divergence-curve discretization size 25.

We report the ordinary `mauve` output. Clustering seeds 11, 23, 37, 51, and 71 reuse the same texts; their score range measures sensitivity to clustering restarts.

Table 9: Prefix32 MAUVE across clustering restarts. Each generated/reference comparison uses 3,000 texts per group. Real-versus-real compares the independent sanity group with the same reference group.

Prefix32 clustering seed	Terminal3	Late3	Nucleus0.95	Real versus real
11	0.6452	0.6049	0.7052	0.9951
23	0.6608	0.6149	0.6848	0.9951
37	0.6642	0.6405	0.7183	0.9959
51	0.6306	0.5863	0.7172	0.9947
71	0.6768	0.6547	0.6785	0.9967
Mean	0.6555	0.6203	0.7008	0.9955
Minimum	0.6306	0.5863	0.6785	0.9947
Maximum	0.6768	0.6547	0.7183	0.9967

The predeclared secondary view compares complete decoded generations with unfiltered LM1B test sentences. Generated sequences contain 128 nonpadding BERT tokens; the full-reference mean is 31.0063.

Full-text MAUVE means are 0.0075 for terminal, 0.0070 for late, and 0.0069 for nucleus. All three show poor agreement in this view. The primary distribution comparison uses matched prefixes.

Table 10: Full-text MAUVE across clustering restarts. Each comparison uses 3,000 texts per group. Minimum and maximum summarize clustering sensitivity.

Full-text clustering seed	Terminal3	Late3	Nucleus0.95	Real versus real
11	0.0070	0.0068	0.0070	0.9969
23	0.0081	0.0069	0.0067	0.9953
37	0.0073	0.0071	0.0073	0.9955
51	0.0075	0.0069	0.0064	0.9960
71	0.0077	0.0072	0.0071	0.9972
Mean	0.0075	0.0070	0.0069	0.9962
Minimum	0.0070	0.0068	0.0064	0.9953
Maximum	0.0081	0.0072	0.0073	0.9972

A.6 Evaluation-recipe and precision sensitivity

We rescore the same 1,408 LM1B confirmation outputs with the original recipe. It decodes with the official slow `BertTokenizer`, removes PAD and MASK, retains CLS, SEP, and UNK, and prepends CLS when absent.

GPT-2-large evaluation uses maximum length 128, batch size eight, FP32, and shifted attention-mask targets without EOS-specific exclusions.

Table 11: Original-recipe rescoring of the same eleven confirmation arms. These scores use the existing 1,408 outputs.

Configuration	GenPPL	Evaluator targets	Truncated outputs
Terminal12	105.2962	16158	100
Even12	124.2290	16197	108
Late12	101.8664	16134	93
Terminal3	125.4217	16186	105
Even3	181.3238	16207	110
Late3	169.4839	16195	107
Temperature0.90	123.7596	16106	102
Temperature0.94	149.6487	16162	106
Temperature0.97	176.9801	16174	105
Temperature1.00	205.1883	16207	116
Nucleus0.95	129.3453	16207	113

The original recipe preserves the interaction, 0.3342 [0.2804, 0.3893]. Terminal/late ratios are 0.7400 [0.7143, 0.7664] at low width and 1.0337 [0.9873, 1.0813] at high width.

Terminal3/nucleus remains uncertain at 0.9697 [0.8979, 1.0498]. All intervals in this rescoring analysis are descriptive.

A separate 16-seed precision/step diagnostic tests FP32, the released HF mixed behavior, and BF16 reference-attention variants.

Table 12: Precision and step diagnostic, 16 outputs per arm. P32 and P41 use 32 and 41 backbone calls, respectively.

Configuration	Calls	Original-recipe PPL	Common PPL	H	repeat4, %	Seconds
FP32 P41	41	242.9509	243.9798	4.2792	0.3000	0.5361
HF mixed P41	41	208.0348	203.8565	4.2813	0.2500	0.5147
BF16 P41	41	220.3826	217.3085	4.2473	0.3000	0.6758
BF16 P32	32	235.4303	227.5091	4.2548	0.2500	0.4805

FP32 QKV, raw-head, and softmax tensors are float32. HF mixed uses float32 first-block QKV and bfloat16 raw-head and softmax tensors. BF16 variants use bfloat16 for all three; every variant retains float32 posteriors.

Table 13: Pointwise precision-diagnostic likelihood ratios. Intervals use 20,000 paired bootstrap resamples, seed 20260908, without multiplicity adjustment. All ratios compare P41 arms.

Contrast	Original-recipe ratio [95% interval]	Common ratio [95% interval]
HF mixed / FP32	0.8563 [0.7150, 1.0076]	0.8355 [0.6961, 0.9890]
BF16 / FP32	0.9071 [0.7552, 1.0760]	0.8907 [0.7532, 1.0477]
BF16 / HF mixed	1.0594 [0.9403, 1.1976]	1.0660 [0.9273, 1.2140]

Lower precision gives lower GenPPL point estimates in this 16-seed diagnostic; Table 13 reports their uncertainty. All main placement comparisons use fixed FP32 precision. Direct implementation checks follow below.

A.7 Additional length-1024 placement results

Table 14: Length-1024 configurations, 48 seeds per checkpoint. All use 41 backbone calls. Native PC1 uses $21P + 20C$, and PC2 uses $15P + 26C$.

Checkpoint	Configuration	GenPPL	H	repeat4, %
ELBO	Even	91.2893	5.5840	0.8386
ELBO	Late	88.2226	5.5773	0.8284
ELBO	Terminal	64.0430	5.5271	0.9509
ELBO	Clean-late	87.4285	5.5761	0.8448
ELBO	Prediction only	92.2450	5.5699	1.0345
ELBO	Calibrated top-p	55.8376	5.5062	1.0692
ELBO	Calibrated temperature	72.0823	5.5002	1.0141
ELBO	Native PC1	94.0412	5.5247	1.1672
ELBO	Native PC2	100.1159	5.5296	0.8958
CE	Even	98.1883	5.5861	0.8774
CE	Late	97.7998	5.5906	0.8121
CE	Terminal	71.2322	5.5366	0.8999
CE	Clean-late	97.2043	5.5897	0.7958
CE	Prediction only	93.7704	5.5748	0.7917
CE	Calibrated top-p	71.6855	5.5390	0.8713
CE	Calibrated temperature	78.6697	5.5094	0.9305
CE	Native PC1	101.2918	5.5421	1.1794
CE	Native PC2	109.6969	5.5398	0.8815

Table 15: Whole-generation timings at length 1024. Measurements cover complete synchronized generation.

Configuration	ELBO seconds	CE seconds
Terminal	2.2061	2.2198
Prediction only	2.3109	2.3255
Calibrated top-p	4.5790	4.6022
Calibrated temperature	2.3084	2.3220
Native PC1	2.0481	2.0554
Native PC2	1.9590	1.9658

These runs use Windows FP32, disabled TF32, and batch size one. Timings include the full predictor and correction trajectory; full-vocabulary sorting contributes to top-p cost.

Table 16: Paired length-1024 comparisons. Ratios divide the first configuration by the second; entropy differences subtract the second. Intervals are descriptive 95%.

Checkpoint	Contrast	GenPPL ratio [95% interval]	ΔH [95% interval]
ELBO	Terminal / Even	0.7015 [0.6705, 0.7330]	−0.0569 [−0.0740, −0.0396]
ELBO	Terminal / Late	0.7259 [0.7130, 0.7387]	−0.0502 [−0.0557, −0.0448]
ELBO	Late / Even	0.9664 [0.9236, 1.0106]	−0.0067 [−0.0224, 0.0092]
ELBO	Even / Top-p	1.6349 [1.5393, 1.7403]	0.0777 [0.0516, 0.1051]
ELBO	Late / Top-p	1.5800 [1.4931, 1.6735]	0.0710 [0.0433, 0.0999]
ELBO	Terminal / Top-p	1.1470 [1.0808, 1.2165]	0.0208 [−0.0066, 0.0492]
ELBO	Clean-late / Late	0.9910 [0.9802, 1.0023]	−0.0012 [−0.0048, 0.0023]
CE	Terminal / Even	0.7255 [0.6965, 0.7560]	−0.0495 [−0.0645, −0.0343]
CE	Terminal / Late	0.7283 [0.7157, 0.7417]	−0.0539 [−0.0603, −0.0478]
CE	Late / Even	0.9960 [0.9532, 1.0420]	0.0044 [−0.0113, 0.0199]
CE	Even / Top-p	1.3697 [1.2896, 1.4562]	0.0471 [0.0236, 0.0717]
CE	Late / Top-p	1.3643 [1.2902, 1.4433]	0.0516 [0.0296, 0.0754]
CE	Terminal / Top-p	0.9937 [0.9433, 1.0475]	−0.0023 [−0.0236, 0.0205]
CE	Clean-late / Late	0.9939 [0.9821, 1.0055]	−0.0008 [−0.0043, 0.0027]

The terminal/top-p entropy intervals lie within the historical ± 0.05-nat band. ELBO top-p has lower GenPPL, while the CE likelihood-ratio interval spans one.

For changed-token events E, final retention is

$$R(x) = |E|^{-1} \sum_{e \in E} \mathbf{1}[x_{i_e}^{\text{final}} = v_e^{\text{new}}].$$

Repeated changes at one position remain separate events.

Table 17: Final-retention means and descriptive 95% intervals.

Checkpoint	Placement	Final retention, % [95% interval]
ELBO	Even	52.76 [51.48, 54.10]
ELBO	Late	85.05 [84.09, 85.94]
ELBO	Terminal	84.24 [83.18, 85.34]
CE	Even	53.10 [51.80, 54.43]
CE	Late	85.13 [84.05, 86.22]
CE	Terminal	84.07 [83.06, 85.07]

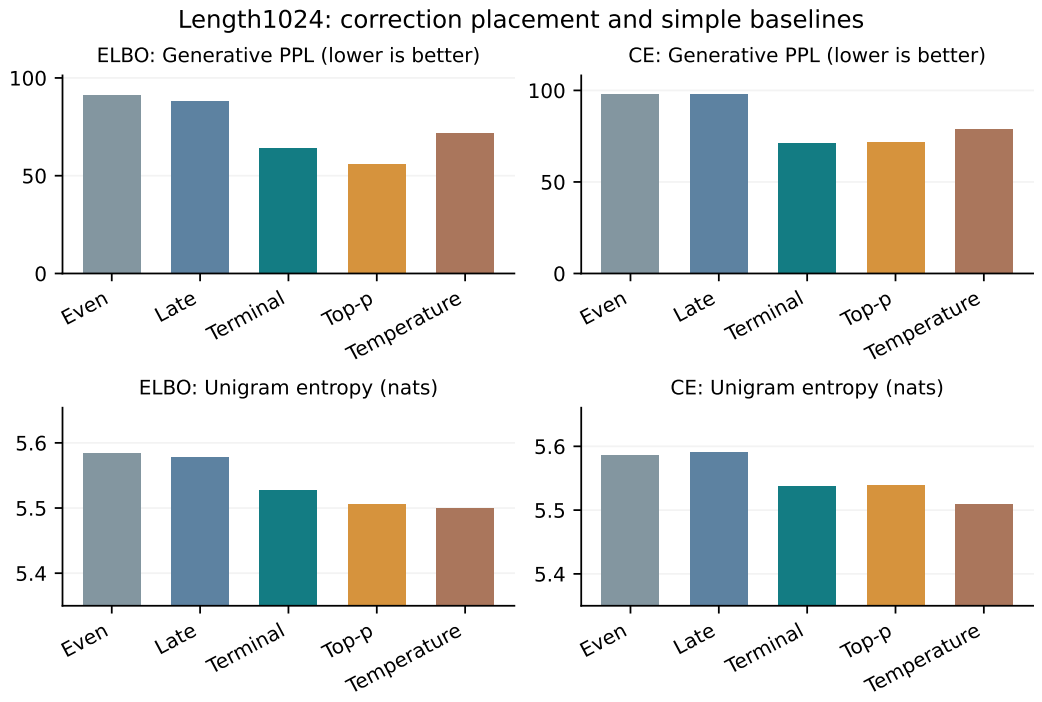


Figure 4: Length-1024 placement and baseline comparison. Terminal separates from both interleaved schedules; calibrated top-p achieves lower ELBO GenPPL. Entropy axes are truncated.

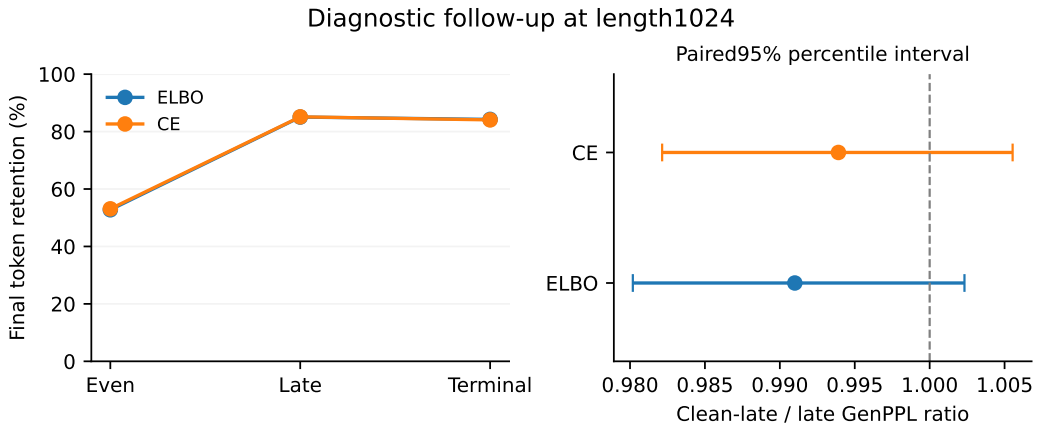


Figure 5: Late and terminal have similar final retention despite different GenPPL. Removing explicit uniform noise from late replacement probabilities leaves a substantial terminal gap.

Even and late produce 96.0 changed-token events per sequence. Terminal means are 95.7292 for ELBO and 95.8542 for CE because replacements can retain the current token. Event counts include repeated changes at the same position.

Clean-late/terminal GenPPL ratios are 1.3652 [1.3413, 1.3898] for ELBO and 1.3646 [1.3412, 1.3880] for CE. Terminal-minus-late repeat4 increases by 0.1224 [0.0531, 0.2000] and 0.0877 [0.0449, 0.1326] percentage points, respectively.

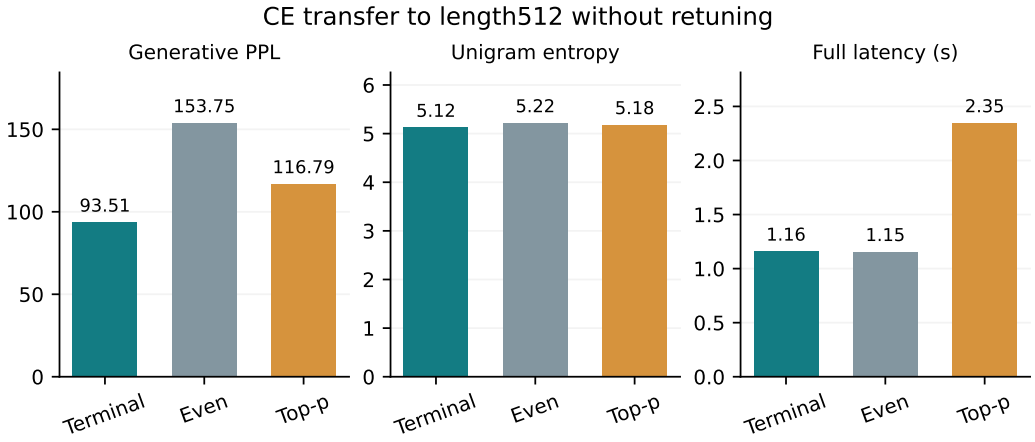


Figure 6: Historical transfer with unchanged correction and top-p settings. Terminal has lower GenPPL and entropy than even placement and transferred top-p.

A.8 Length-512 transfer and temperature controls

Table 18: CE length-512 transfer, 48 seeds. Correction settings and the locked CE top-p cutoff transfer unchanged.

Transfer configuration	GenPPL	H	repeat4, %	Seconds
Terminal	93.5075	5.1210	1.1297	1.1579
Even	153.7519	5.2154	0.8841	1.1542
Transferred top-p	116.7903	5.1767	1.2688	2.3454

Terminal/top-p and even/top-p GenPPL ratios are 0.8006 [0.7136, 0.8961] and 1.3165 [1.1601, 1.4962]. Terminal/even is 0.6082 [0.5669, 0.6512].

Entropy differences are −0.0945 [−0.1201, −0.0685] for terminal minus even, −0.0557 [−0.0985, −0.0146] for terminal minus top-p, and 0.0387 [0.0025, 0.0737] for even minus top-p.

Terminal-minus-top-p repeat4 is −0.1392 [−0.4993, 0.2251] percentage points. Terminal/even and terminal/top-p time ratios are 1.0032 [1.0005, 1.0060] and 0.4937 [0.4927, 0.4947], respectively.

Table 19: Separate CE length-512 temperature comparison, 48 seeds. The locked control uses temperature 0.9695.

Temperature configuration	GenPPL	H	repeat4, %	Seconds
Terminal	93.2240	5.1429	1.3466	1.0703
Locked temperature	120.7243	5.1860	1.1992	1.0997

The terminal/temperature GenPPL ratio is 0.7722 [0.6981, 0.8545]. Entropy and repeat4 differences are −0.0431 [−0.0785, −0.0081] nats and 0.1473 [−0.0860, 0.3847] percentage points.

The time ratio is 0.9732 [0.9534, 1.0084], with a difference of −0.0294 [−0.0513, 0.0092] seconds.

A.9 Prospective temperature comparison

We fix six arms before generation and use 128 fresh seeds per arm, for 768 outputs. Execution order is balanced across arms. Temperature scales raw backbone logits before native prediction processing; nucleus cutoff is one.

Table 20: Prospective CE length-512 temperature comparison. Placement uses $33P + 8C$, while temperature controls use $41P$.

Prospective configuration	GenPPL	H	repeat4, %	Seconds
Terminal	96.8726	5.1233	1.5840	1.0458
Even	162.5812	5.2361	1.0744	1.0422
Temperature 0.94	106.3151	5.1020	1.3891	1.0935
Temperature 0.96	121.6953	5.1537	1.3277	1.0915
Locked temperature 0.9695	133.3991	5.1681	1.2386	1.0929
Temperature 1.00	175.3636	5.2447	0.9286	1.0916

Scoring retokenizes decoded text with GPT-2-large at the appropriate length limit. The EOS rule scores the first EOS and all non-EOS targets, while masking later EOS targets.

All six arms have zero observed padding. Effective target counts vary with retokenization and masking.

Table 21: Prospective pooled likelihood totals and marginal uncertainty. NLL totals retain their original four-decimal precision; target counts are exact.

Configuration	Total NLL	Targets	GenPPL [descriptive 95% interval]
Terminal	297622.9280	65077	96.8726 [90.1220, 104.4015]
Even	331715.6648	65155	162.5812 [149.9413, 176.7173]
Temperature0.94	304156.4510	65180	106.3151 [98.9486, 114.4349]
Temperature0.96	313068.7322	65202	121.6953 [113.3279, 131.3136]
Temperature0.9695	318767.1764	65143	133.3991 [123.4697, 144.6854]
Temperature1.00	336553.8547	65137	175.3636 [162.3179, 190.3511]

The two primary ratios use individual 97.5% Bonferroni intervals: terminal/even is 0.5958 [0.5641, 0.6282], and terminal/locked-temperature is 0.7262 [0.6716, 0.7837].

Table 22: Prospective descriptive likelihood comparisons, with terminal as numerator. The final column reports the mean sequence-level NLL-rate difference, a separate statistic from the pooled log GenPPL used for width interactions.

Comparator	GenPPL ratio [95% interval]	Mean sequence NLL-rate difference [95% interval]
Even	0.5958 [0.5681, 0.6241]	−0.5176 [−0.5653, −0.4711]
Temperature0.9695	0.7262 [0.6782, 0.7760]	−0.3199 [−0.3884, −0.2535]
Temperature0.94	0.9112 [0.8546, 0.9719]	−0.0927 [−0.1567, −0.0281]
Temperature0.96	0.7960 [0.7445, 0.8500]	−0.2279 [−0.2947, −0.1622]
Temperature1.00	0.5524 [0.5159, 0.5910]	−0.5934 [−0.6619, −0.5257]

Table 23: Prospective paired secondary metrics. Differences are terminal minus comparator; repeat4 differences are percentage points. Intervals are descriptive 95%.

Comparator	ΔH	Δrepeat4, pp	Δseconds	Time ratio
Even	−0.1128 [−0.1289, −0.0966]	0.5096 [0.2824, 0.7552]	0.0036 [0.0018, 0.0054]	1.0034 [1.0018, 1.0052]
T=0.9695	−0.0447 [−0.0681, −0.0213]	0.3453 [−0.0046, 0.7137]	−0.0471 [−0.0488, −0.0454]	0.9569 [0.9554, 0.9584]
T=0.94	0.0214 [0.0010, 0.0424]	0.1949 [−0.0967, 0.4942]	−0.0477 [−0.0497, −0.0458]	0.9564 [0.9546, 0.9581]
T=0.96	−0.0304 [−0.0510, −0.0100]	0.2563 [−0.0399, 0.5771]	−0.0457 [−0.0475, −0.0440]	0.9581 [0.9565, 0.9596]
T=1.00	−0.1213 [−0.1457, −0.0978]	0.6554 [0.3254, 1.0299]	−0.0458 [−0.0474, −0.0442]	0.9580 [0.9566, 0.9595]

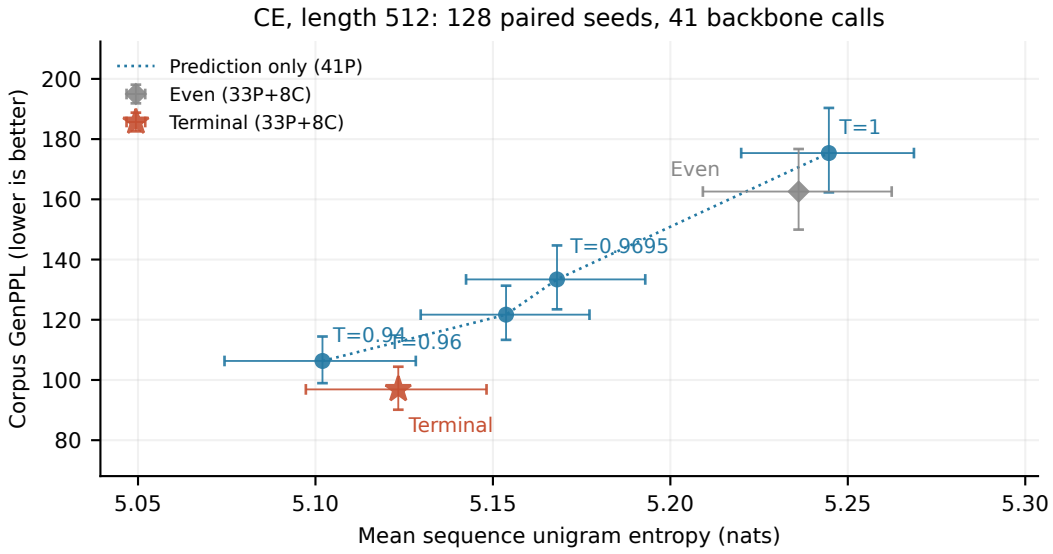


Figure 7: CE temperature comparison, 128 seeds. Terminal has lower GenPPL and slightly higher entropy than its closest-entropy temperature, 0.94. Error bars are marginal 95% intervals; dotted lines connect the four temperatures.

A.10 Auxiliary automated reading assessment

The prospective temperature study includes an auxiliary model assessment of coherence and readability. Before generation, we fix a rule to select the temperature nearest terminal mean entropy, breaking ties by ascending temperature.

This rule selects temperature 0.94. Twenty-four distinct cohort seeds were sampled using NumPy seed 2026090824, with randomized A/B positions.

A separate Codex language-model assessor receives opaque pair identifiers and complete decoded texts. It rates coherence and readability from 1 (frequent breakdowns) to 5 (consistently clear).

Coherence concerns continuity of statements and entities; readability concerns syntax, wording, and garbled or repetitive passages. Preference uses the sum of the two ratings, with equal sums counted as ties.

The assessor receives the general hypothesis with method identities and aggregate results masked. It does not penalize different topics or unfinished final sentences. Ratings are frozen before method identities are joined.

Table 24: Auxiliary Codex model ratings on 24 paired texts.

Configuration	Coherence	Readability	Pair preferences
Terminal	1.8333	1.9583	6
Temperature0.94	1.7917	1.9583	5

The remaining 13 pairs are ties, and the two methods receive similar coherence and readability scores in this sample. The reproducibility records include the input pairs, individual ratings, textual observations, and identity mapping.

A.11 INTERMEDIATE-WIDTH VALIDATION

We design midpoint validation after the endpoint experiments and reuse the original 128 seed labels per model, frozen endpoint outputs, and scores. The predictor grid, correction-call count, and operator stay fixed.

New terminal and late arms use $k = 6$ for LM1B and $k = 24$ for OpenWebText. Each model adds 256 outputs and reuses 512 endpoints, giving 512 new and 1,024 reused outputs in total.

Midpoint execution order is balanced across the two arms within each model.

Table 25: Intermediate-width arm scores, 128 outputs each. Every arm uses $33P + 8C$ and $8k/L = 0.375$. GenPPL intervals are descriptive 95%; repeat4 is a fraction. Timings describe these executions.

Midpoint arm	GenPPL [95% interval]	Mean H, nats	Mean repeat4 fraction	Mean seconds	Evaluator targets
LM1B Terminal6	105.5582 [97.5951, 114.0643]	4.1564	0.0050	0.5720	15794
LM1B Late6	126.8220 [116.1162, 138.5632]	4.2083	0.0040	0.5763	15841
OpenWebText Terminal24	77.7276 [72.7802, 83.0444]	5.0382	0.0172	1.6481	65094
OpenWebText Late24	98.8059 [92.2618, 105.9673]	5.1285	0.0124	1.6422	65112

Total evaluator NLL values in Table 25 order are 73588.3964, 76714.5506, 283367.8906, and 299069.6522. The unchanged endpoint arm scores remain in Tables 4 and 7.

Table 26: Three-width terminal/late ratios for Figure 2. Descriptive 95% paired-bootstrap intervals are recomputed for the follow-up; endpoint estimates are unchanged from Table 1.

Follow-up model and width	Nominal $8k/L$	Log GenPPL ratio	GenPPL ratio [95% interval]	Output origin
LM1B $k = 3$	0.1875	−0.3136	0.7308 [0.7043, 0.7587]	Frozen endpoint
LM1B $k = 6$	0.375	−0.1835	0.8323 [0.7932, 0.8723]	New midpoint
LM1B $k = 12$	0.75	0.0424	1.0433 [0.9989, 1.0910]	Frozen endpoint
OpenWebText $k = 12$	0.1875	−0.3669	0.6929 [0.6790, 0.7067]	Frozen endpoint
OpenWebText $k = 24$	0.375	−0.2399	0.7867 [0.7665, 0.8071]	New midpoint
OpenWebText $k = 48$	0.75	−0.0824	0.9209 [0.8943, 0.9475]	Frozen endpoint

Define $r(k) = \log(\mathrm{PPL}_{\mathrm{terminal},k}/\mathrm{PPL}_{\mathrm{late},k})$. The declared LM1B contrasts are $r(6) - r(3)$ and $r(12) - r(6)$; OpenWebText contrasts are $r(24) - r(12)$ and $r(48) - r(24)$.

Each resample uses common seed indices within a model to recompute pooled NLL per target for the relevant arms.

Table 27: Adjacent-width contrasts. Individual 98.75% percentile intervals apply Bonferroni adjustment across the four declared contrasts for 95% family coverage. Additional 95% intervals are descriptive.

Adjacent-width comparison	Δ log ratio, nats	Descriptive 95% interval	Adjusted 98.75% interval	Ratio of ratios [98.75% interval]
LM1B, $3 \to 6$	0.1301	[0.0806, 0.1797]	[0.0667, 0.1940]	1.1389 [1.0690, 1.2141]
LM1B, $6 \to 12$	0.2260	[0.1702, 0.2834]	[0.1543, 0.2995]	1.2535 [1.1668, 1.3492]
OpenWebText, $12 \to 24$	0.1269	[0.0967, 0.1576]	[0.0879, 0.1663]	1.1354 [1.0919, 1.1810]
OpenWebText, $24 \to 48$	0.1575	[0.1232, 0.1928]	[0.1144, 0.2028]	1.1706 [1.1212, 1.2248]

All four adjusted intervals are positive, supporting attenuation at both tested increments in both models. The LM1B high-width ratio interval includes one; OpenWebText terminal retains lower GenPPL at every tested width.

A descriptive interpolation check measures

$$r(k_{\mathrm{middle}}) - \frac{r(k_{\mathrm{low}}) + r(k_{\mathrm{high}})}{2},$$

Widths double at each increment. This descriptive interpolation contrast is −0.0479 [−0.0948, −0.0027] for LM1B and −0.0153 [−0.0439, 0.0129] for OpenWebText, with 95% intervals.

A.12 ORIGINAL AND PORTABLE LM1B IMPLEMENTATIONS

The original DIT and portable model share the released weights; mapped parameters, buffers, and EMA order match. The HF weights already contain exported EMA parameters and receive no second EMA.

The reference comparisons run the original DIT and inference methods with both BF16 contexts disabled. They use reference attention and rotary operators, rotating Q/K by half-split and leaving V unchanged.

Fixed inputs cover batch-one length-eight at $t = 0.5$, batch-one length-128 at $t = 1.0$, and batch-two length-eight at $t = (10^{-5}, 0.2)$.

Hooks inspect the sigma map, embedding, all twelve Transformer blocks, output head, log probabilities, and predictor posteriors. Trajectories use length 128 and seeds 97700 and 97701.

Table 28: Original-versus-portable execution comparisons. Fixed inputs cover the three cases above. Final differences count token positions for seeds 97700 and 97701; both FP32 comparisons use common reference operators.

Comparison	Fixed-input difference	Steps per seed	Final differences
CPU FP32 reference	Zero, all three cases	3	0 and 0
Windows GPU FP32 reference	Zero, all three cases	32	0 and 0
Linux BF16 native/reference	From first Transformer block	32	8 and 4

Original and portable execution match bitwise in both FP32 reference comparisons for every tested fixed input, sampled trajectory, RNG consumption, and sigma value.

The Linux comparison runs original DIT with native BF16 contexts, FlashAttention 2.8.3, and rotary kernels. The portable implementation is configured to match its BF16 behavior.

Embeddings and sigma maps match; numerical differences first appear in the first Transformer block.

Table 29: Native BF16 versus BF16-matched reference on fixed inputs. Head, log probability, and posterior columns report maximum absolute differences; mean TV is posterior total variation.

Fixed-input case	Head	Log probability	Posterior	Mean TV	Argmax differences
Length 8, $t = 0.5$	0.1250	0.2500	0.0063	0.0022	0
Length 128, $t = 1.0$	0.0625	0.1250	0.0029	0.00044	0
Batch 2, $t = (10^{-5}, 0.2)$	0.3203	0.5000	0.0024	0.00074	1

The two BF16 trajectories have equal RNG consumption and sigma values. Head outputs first differ at step 1; sequence inputs first differ at steps 17 and 27 for seeds 97700 and 97701.

Final differing positions are $\{1, 3, 54, 55, 57, 61, 62, 95\}$ and $\{13, 14, 21, 22\}$, respectively.

A.13 NATIVE LM1B BASELINE AND THE PUBLISHED REFERENCE

Native BF16/FlashAttention generation runs the original DIT and inference methods with 32 predictor calls and batch size 32. Four batch seeds, 97800 to 97803, each produce 32 sequences, for 128 outputs.

Evaluation follows the original recipe above, with GPT-2-large FP32, maximum length 128, and batch size eight. Mean untruncated GPT-2 tokenized length is 134.5391; 111 outputs are truncated.

Table 30: Native BF16/FlashAttention P32 baseline. Batch rows summarize 32 outputs each; the pooled row aggregates all saved evaluator target losses.

Batch seed	Outputs	Evaluator NLL	Targets	GenPPL
Seed 97800	32	21244.5488	4040	192.2029
Seed 97801	32	21335.1454	4040	196.5617
Seed 97802	32	21107.7250	4061	180.8498
Seed 97803	32	21210.1113	4054	187.1475
All four batches	128	84897.5305	16195	189.0868

The original CUDA float32 reduction gives GenPPL 189.0870; double-precision pooling of saved target losses gives 189.0868.

Mean raw-token entropy is 4.1946 nats, and mean repeat4 is 0.0064.

Native QKV, head, and log-probability tensors are bfloat16; posterior tensors are float32. The Linux runtime uses PyTorch 2.8.0+cu128, FlashAttention 2.8.3, and transformers 4.38.2.

Published Table 36 reports LM1B GenPPL 79.80 at 32 steps using 1,024 outputs and GPT-2-large (Schiff et al., 2025). The native public-weight diagnostic remains higher, at 189.0868.

The earlier original-recipe P41 score of 205.1883 uses different step counts, batch sizes, and seeds. Tables 11 to 13 report the original-recipe rescoring and precision probes.

FP32 implementation agreement and the native BF16 result show that the portable adaptation alone does not explain the full absolute gap.

Public model and source records agree on the broad architecture: approximately 139 million parameters, twelve blocks, hidden size 768, twelve heads, context length 128, and a BERT vocabulary of 30,522.

Reported training uses one million updates, batch size 512, learning rate 3×10^{-4}, and 2,500 warmup steps. The public script specifies continuous-time uniform ELBO, time conditioning, and zero reconstruction loss.

The export notebook copies EMA weights into the HF backbone. Its placeholder checkpoint path leaves the exact Table 36 checkpoint unidentified.

The HF history contains one actual weight upload. Initial and current weight, configuration, and model-code objects match; later changes concern the README.

The paper describes chunked LM1B data with approximately seven million training examples and 72,000 validation examples. Table 37 displays multiple CLS markers without SEP markers.

The initial and current public scripts instead specify `data.wrap=False`, `training.compute_loss_on_pad_tokens=False`, and `model.length=128`. That dataloader path truncates or pads individual sentences before returning.

The exact published checkpoint, resolved training/data configuration, and original generated sequences are unavailable. The data-format discrepancy is a provenance clue; its contribution to the GenPPL gap remains unknown.

A.14 Correction-budget validation

After the eight-call experiments, we specify $I_4 > 0$ in a new protocol before four-call generation and scoring. It fixes 128 new paired seeds, 99000 to 99127, and terminal/late arms at $k = 12, 48$. Generation produces all 512 planned outputs.

C4 retains the 33 native predictor transitions. Late correction follows indices 26, 28, 30, and 32; terminal applies four sequential time-zero corrections after prediction 33. Each output uses 37 backbone calls.

The unchanged corrector selects k distinct positions using `lowest_log_margin`, with selection and replacement temperatures one and nucleus cutoff one.

The separate predictor and correction RNG streams defined above are retained. Predictor RNG states match across all four arms; equal-width terminal/late pairs consume equal correction RNG.

Widths share seed labels, while correction draws can differ across widths. Cyclic execution order places each arm in each position 32 times.

Generation and scoring use the earlier Windows FP32 runtime and pinned evaluator, with batch size one, maximum length 512, and the same EOS target mask. The likelihood metric pools saved GPU float32 NLL sums and effective target counts.

Table 31: Four-call OpenWebText arm scores. Each column contains 128 new length-512 outputs with $33P + 4C$. GenPPL bounds are descriptive 95% intervals; repetition, PAD, and EOS rates are fractions.

C4 statistic	**Terminal** $k = 12$	**Late** $k = 12$	**Terminal** $k = 48$	**Late** $k = 48$
Outputs	128	128	128	128
Predictor calls	33	33	33	33
Correction calls	4	4	4	4
Total calls	37	37	37	37
Width k	12	12	48	48
Nominal $4k/L$	0.0938	0.0938	0.375	0.375
GenPPL	110.0741	150.9077	86.7790	98.9545
GenPPL 95% lower bound	102.2037	140.1621	81.7038	92.7213
GenPPL 95% upper bound	118.5434	162.4493	92.2649	105.6463
Total evaluator NLL	306030.9894	326795.7941	290444.5516	299107.7952
Effective evaluator targets	65097	65142	65073	65099
Mean entropy, nats	5.1524	5.2097	5.0437	5.1080
Mean repeat4 fraction	0.0127	0.0105	0.0173	0.0137
Mean generation seconds	1.5728	1.5742	1.5624	1.5807
Mean PAD rate	0.0	0.0	0.0	0.0
Mean EOS rate	0.0011	0.0011	0.00093	0.00099

The single primary estimand is

$$I_4 = \log \frac{\mathrm{PPL}_{\mathrm{terminal},48;4}}{\mathrm{PPL}_{\mathrm{late},48;4}} - \log \frac{\mathrm{PPL}_{\mathrm{terminal},12;4}}{\mathrm{PPL}_{\mathrm{late},12;4}}, \qquad \text{predicted sign: positive.}$$

Within-width log ratios are -0.3155 at $k = 12$ and -0.1313 at $k = 48$.

The bootstrap draws 20,000 percentile resamples with seed 909202604. Each draw uses common seed indices across all four arms and recomputes pooled NLL/target ratios. The primary 95% interval evaluates the predicted positive interaction.

Table 32: Four-call ratios and primary interaction. Within-width intervals are descriptive 95%. The final two rows report the same primary contrast on logarithmic and exponentiated scales.

Within-C4 estimand	**Estimate**	**95% interval**	**Inference role**
Terminal12 / Late12 GenPPL	0.7294	[0.7184, 0.7407]	Descriptive within-width ratio
Terminal48 / Late48 GenPPL	0.8770	[0.8563, 0.8989]	Descriptive within-width ratio
I_4, nats per evaluated target	0.1842	[0.1600, 0.2089]	Single primary interaction
$\exp(I_4)$, ratio of ratios	1.2023	[1.1736, 1.2323]	Same primary contrast, exponentiated

The primary interval supports the predicted positive interaction at this second correction-call budget. Absolute GenPPL improves with width under both placements, and terminal remains lower at both widths.

C4 and C8 use separate seed cohorts, 37 and 41 total calls, and different late insertion densities. The results test placement-by-width interactions within each budget; width and cumulative proposal count vary together.

A.15 REPRODUCIBILITY RECORDS

The accompanying records include frozen protocols, model and source revisions, all generated outputs and scores, analysis code, and execution diagnostics. Diagnostic outputs are excluded from the scientific cohorts.

Independent CPU recomputation reproduced the saved paired ratios and confidence intervals. The source package contains detailed numerical checks, generation and scoring anchors, and parameter-verification records.