HACKING GENERATIVE PERPLEXITY: WHY UNCONDITIONAL TEXT EVALUATION NEEDS NEW METRICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion language models (DLMs) have emerged as the leading non-autoregressive alternatives to language modeling. Because their likelihoods are intractable, generative perplexity (gen-PPL) has become the de facto standard for reporting unconditional generation quality: the length-normalized likelihood of generated samples under a frozen autoregressive scorer, typically paired with a per-sample unigram entropy guardrail. We give an information-theoretic account of why this metric is unsound. Scoring model samples places the joint entropy of the generator inside gen-PPL, making collapse optimal for any fixed scorer. The reported per-sample unigram entropy ignores token order and cross-position dependence and therefore cannot correct this failure. Four training-free naive samplers reach or beat trained-model gen-PPL on LM1B and OpenWebText at non-degenerate unigram entropy, while producing text that is incoherent by construction. We then rebenchmark recent DLMs using distributional discrepancies and diversity diagnostics, and compare their rankings with LLM judgments. As part of this evaluation, we introduce a compression-based discrepancy that fits frozen byte-level coders to real and generated samples and measures local coverage and fidelity in both directions. The resulting metric profile provides a more informative view of unconditional text generation. We argue that DLM progress should be evaluated using a set of metrics that expose different properties of generated text.

1 INTRODUCTION

Autoregressive (AR) models currently dominate language modeling at scale. They factorize the joint distribution over a token sequence into a product of conditionals, and are trained to predict each token from those that precede it (Bengio et al., 2003). This objective is remarkably simple, and in combination with the Transformer architecture (Vaswani et al., 2017), it has proven to scale very well with model size and data (Kaplan et al., 2020; Brown et al., 2020). The result is a family of general-purpose models that now power widely used systems for writing, coding, search, and scientific work. That said, the same factorization makes generation inherently sequential, since each new token depends on the prefix already produced. Caching previous activations makes each of these steps cheaper, and draft-and-verify schemes such as speculative decoding allow several tokens to be accepted per step (Stern et al., 2018; Leviathan et al., 2023; Cai et al., 2024). Neither removes the underlying issue, however: the cost of sampling remains tied to the length of the output.

Diffusion language models (DLMs) take a different approach (Lou et al., 2024; Sahoo et al., 2024). Rather than committing to one token at a time, they start from a fully corrupted sequence and refine many positions in parallel at each sampling step. This means the number of sequential operations is no longer tied to the length of the output: in principle, a sequence could be generated in far fewer steps than it has tokens. Realizing this in practice is still an open problem and an active line of research (Arriola et al., 2025; Deschenaux & Gülçehre, 2025), but the promise alone has made DLMs an important alternative to AR generation. This shift also creates an evaluation problem. Progress in AR language modeling has long been measured by likelihood on fixed held-out data (Chelba et al., 2013; Jozefowicz et al., 2016). As long as the corpus and tokenization are shared, a lower validation perplexity simply means that a model has learned the same distribution more accurately, so successive models can be compared directly.

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.

	Generation	gen-PPL ↓	Entropy ↑	MAUVE ↑	GM ↓	ED ↓
Data	*before I headed out I put the girls in their Vikings jammies and got ready to leave. Before I left I asked my older …*	17.1	5.48	1.000	0.0	0.00
PLAID	*300 million people are infected per year. Some viruses will steal your password and make you think your …*	20.7	5.32	0.896	1.7	0.43
Naive	*, the. to of and a in that- is for on s it with was as I: be are (have you at by from's this he not The an has his or …*	**19.6**	**5.93**	**0.004**	**54.5**	**28.60**

Figure 1: OpenWebText samples from data, Plaid, and a naive periodic baseline evaluated across five metrics: generative perplexity (with unigram entropy guardrail), MAUVE, Gradient Moment (GM), and Energy Distance (ED). Gray marks the data baseline, green favorable scores, and red poor scores.

For DLMs, however, output likelihoods are usually either unavailable or not comparable across model families and samplers (Lou et al., 2024; Sahoo et al., 2024; Peng et al., 2025; Gulrajani & Hashimoto, 2023). The field has therefore settled on generative perplexity (gen-PPL) instead: rather than scoring held-out data under the model being evaluated, we score that model's samples under a frozen AR model, usually `gpt2-large` (Radford et al., 2019; Lou et al., 2024; Arriola et al., 2025; Deschenaux & Gülçehre, 2025). Despite the shared name, this is a fundamentally different measurement. Validation perplexity keeps the data fixed and varies the model. gen-PPL does the opposite: the scorer stays fixed, and it is the samples that change with the generator. A lower value can therefore mean one of two things. Either the generator produces more realistic text, or it has learned to concentrate on outputs that the scorer finds easy to predict (Theis et al., 2016; Holtzman et al., 2020). The score alone cannot tell these apart.

In Section 3, we make this precise. Swapping the roles of data and model places the joint entropy of the generator's complete samples inside the score. In other words, collapse is optimal: for any fixed scorer, no generator can beat one that always outputs the scorer's single most likely sequence. It has become common to report per-sample unigram entropy alongside gen-PPL to guard against this (Wang et al., 2025; Pynadath et al., 2026), but this is only a partial fix. Unigram entropy ignores token order and dependencies between positions, so it cannot fully account for the joint entropy term. To make this failure concrete, we construct four naive samplers that require no training at all. Each preserves a different set of low-order statistics of the data, while deliberately destroying linguistic structure. On LM1B and OpenWebText (Chelba et al., 2013; Gokaslan & Cohen, 2019), these samplers match or beat the gen-PPL of trained DLMs, and their unigram entropy remains at non-degenerate levels. In other words, neither number flags them.

For unconditional generation, a more natural target is agreement between the distributions of generated and real text (Caccia et al., 2020; Hashimoto et al., 2019; Pillutla et al., 2021). There is no single right way to measure this: every metric commits to a representation of the text and an estimator, and these choices determine which properties of the text it is sensitive to. Rather than relying on any one metric, we therefore evaluate DLMs with a profile of complementary measurements. These include distributional discrepancies (including a new compression-based Jeffreys coding divergence) and diversity diagnostics. We also use blinded pairwise LLM judgments as a scalable proxy for human preference (Zheng et al., 2023). This profile rejects all four constructed failures. Among trained generators, it produces substantially different rankings from those implied by gen-PPL.

Contributions.

1. **We characterize what generative perplexity measures.** We show that gen-PPL decomposes into two terms: the joint entropy of the generator, and a mode-seeking divergence between the generator and the scorer. Lowering either term improves the score, so the optimal generator is a point mass on the scorer's most likely sequence. This kind of collapse need not show up in per-sample unigram entropy. A similar argument applies to any judge that scores samples one at a time, including LLM judges.

2. **We construct training-free samplers that beat trained DLMs on gen-PPL.** These samplers are built only from token and phrase counts in the corpus, yet they match or beat the gen-PPL of trained models on LM1B and OpenWebText. One of them emits the same sample every time. Together, they form a sanity check that any new metric or judge should pass.
3. **We introduce a compression-based distributional discrepancy.** We find that the choice of encoder moves model rankings more than the choice of distance computed on top of it. This motivates the Jeffreys coding divergence, which replaces learned encoders with frozen byte-level coders and estimates local coverage and fidelity separately.
4. **We rebenchmark recent DLMs and recommend a metric profile.** Once the full profile is used, rankings shift considerably. On OpenWebText, the model with the best gen-PPL ranks last among trained generators by mean rank across distributional metrics, and increasing the number of sampling steps lowers gen-PPL while making energy distance (ED) worse. We recommend reporting LLM-judge Elo with entropy guardrails, alongside GM, MAUVE, and ED, across a range of sampling budgets.

2 BACKGROUND

Following Dieleman (2026), we use *diffusion language model* (DLM) for any model that generates tokens by iteratively reversing a corruption process or transporting a state from a simple prior. We write G for the distribution over complete samples that these DLMs induce. Validation perplexity would require evaluating the likelihood $G(x)$ on held-out text. As we show next, this is generally intractable for both discrete and continuous DLMs.

Discrete states. Discrete DLMs operate directly on token sequences $x \in \mathcal{V}^L$ for a vocabulary $\mathcal{V}$. Sampling starts from a simple prior p_T at the noise end ($t = T$) and applies learned denoising kernels $p_{\theta,t}(x^{(t-1)} \mid x^{(t)})$ until it reaches clean data ($t = 0$). The probability of a sample is therefore a sum over every trajectory that could have produced it:

$$G(x) = \sum_{x^{(1)},\ldots,x^{(T)}} p_T(x^{(T)}) \prod_{t=1}^{T} p_{\theta,t}(x^{(t-1)} \mid x^{(t)}), \qquad x^{(0)} = x. \tag{1}$$

Each one-step kernel is usually cheap to evaluate, but the number of trajectories grows exponentially with the sequence length, so the sum is generally intractable for masked and other categorical diffusion (Austin et al., 2021; Sahoo et al., 2024; Lou et al., 2024).

Continuous states. Continuous DLMs first map tokens into $\mathbb{R}^D$, as one-hot vectors, learned embeddings, points on a categorical manifold, or analog bit codes (Li et al., 2022; Dieleman et al., 2022; Lee et al., 2026), and then evolve a continuous state $Z^{(t)}$. Reverse kernels $p_{\theta,t}(z^{(t-1)} \mid z^{(t)})$ carry $Z^{(T)} \sim p_T$ to $Z^{(0)} \sim p_0$; for deterministic flows, these kernels are point masses. A decoder $D_\phi(\cdot \mid z)$, which may also be deterministic, then maps the final state to a distribution over $\mathcal{V}^L$:

$$G(x) = \int_{\mathbb{R}^D} D_\phi(x \mid z)\, p_0(\mathrm{d}z). \tag{2}$$

Two obstacles stand in the way of evaluating this. For stochastic diffusion, p_0 itself requires marginalizing over continuous trajectories. Even when a flow makes p_0 tractable, the integral must still collect probability from every final state that decodes to x; for a deterministic decoder, this is the total mass of the region mapped to x.

Without $G(x)$, DLMs are evaluated through samples: the standard gen-PPL scores them under a frozen `gpt2-large`, with the per-sample unigram entropy H_{uni} reported alongside to rule out degenerate repetition (Radford et al., 2019; Lou et al., 2024; Sahoo et al., 2024; Wang et al., 2025; Pynadath et al., 2026). Section 3 shows why this pair is not enough.

3 WHAT GENERATIVE PERPLEXITY MEASURES

Let $X = (X_1, \ldots, X_L)$ be a complete sample of L tokens from a vocabulary $\mathcal{V}$. We write P for the data distribution, G for the distribution induced by a DLM and its sampler, and R for an AR model. All logarithms are natural.

For an AR model, $R(X)$ is tractable, and its population validation perplexity $\mathrm{ValPPL}(R)$ exponentiates the cross-entropy $H_\times(P, R)$ of real samples under the model, normalized by L, as shown in left of Equation (3). This cross-entropy splits into the entropy of the data and the model's divergence from it. With data, tokenization, and L fixed, the entropy term is the same for every model, so models differ in log validation perplexity only through $D_{\mathrm{KL}}(P\|R)/L$. A lower value therefore means necessarily a better fit to the same distribution.

For DLMs, the situation is different. A DLM sampler produces samples $X \sim G$ but generally not $G(X)$, so the validation perplexity of G is out of reach. Gen-PPL works around this in two steps: it swaps the roles of data and model, and then replaces the unknown P with a model R:

$$\mathbb{E}_{X\sim P}[-\log G(X)] \xrightarrow{\text{swap roles}} \mathbb{E}_{X\sim G}[-\log P(X)] \xrightarrow{P\to R} \mathbb{E}_{X\sim G}[-\log R(X)].$$

The result has the same cross-entropy form as the validation perplexity of R itself, but with the expectation taken over the generator instead of the data:

$$\begin{aligned} \mathrm{ValPPL}(R) &= \exp\Big(\frac{1}{L}H_\times(P,R)\Big) \\ H_\times(P,R) &= \mathbb{E}_{X\sim P}[-\log R(X)] \\ &= H(P) + D_{\mathrm{KL}}(P\|R) \end{aligned} \quad \longleftrightarrow \quad \begin{aligned} \mathrm{GenPPL}_R(G) &= \exp\Big(\frac{1}{L}H_\times(G,R)\Big) \\ H_\times(G,R) &= \mathbb{E}_{X\sim G}[-\log R(X)] \\ &= H(G) + D_{\mathrm{KL}}(G\|R) \end{aligned} \tag{3}$$

The difference lies in the first term. $H(P)$ is fixed by the benchmark, whereas $H(G)$ varies from one generator to the next. A lower gen-PPL can therefore reflect a smaller mismatch with R, a less diverse generator, or both, and the metric cannot tell which.

The metric has a collapsed optimum. Taken to its limit, this ambiguity determines the optimum. Let $x^\star \in \arg\max_{x\in\mathcal{V}^L} R(x)$. Since $-\log R(x) \geq -\log R(x^\star)$ for every x,

$$H_\times(G,R) = \sum_x G(x)[-\log R(x)] \geq -\log R(x^\star) = H_\times(\delta_{x^\star}, R), \tag{4}$$

and because exponentiation is monotone, $G = \delta_{x^\star}$ globally minimizes $\mathrm{GenPPL}_R(G)$. In other words, the best possible score belongs to a generator that emits the scorer's most likely sequence every time. Even sampling from R itself does worse, since $H_\times(R,R) = H(R) \geq -\log R(x^\star)$, with equality only if R is uniform on its support. If we take R as the best tractable model of coherent text, the ideal generator is $G = R$; gen-PPL instead prefers the point mass. Figure 4 in Appendix A.1 illustrates this. Subtracting $H(G)$ exactly would remove this incentive, leaving $D_{\mathrm{KL}}(G\|R)$, whose minimizer $G = R$ matches the data only as well as R approximates P.

The two KL terms also differ in an important way. In validation perplexity, $D_{\mathrm{KL}}(P\|R)$ averages over real samples, so a model is penalized whenever it assigns low probability to text that actually occurs, i.e. whenever it misses modes of the data. In gen-PPL, $D_{\mathrm{KL}}(G\|R)$ averages over generated samples instead. It penalizes outputs that are unlikely under R, but it has nothing to say about modes the generator never produces. In other words, validation perplexity is coverage-sensitive, whereas gen-PPL is mode-seeking. The divergence term therefore does little to resist the collapse that $H(G)$ already rewards.

Finally, gen-PPL is typically reported alongside H_{uni}, the entropy of the token histogram within each sample, averaged over samples. This catches a sample that repeats a single token, but not a generator that always emits the same token-rich sample: such a generator has high H_{uni}, yet $H(G) = 0$. This kind of sampler is empirically constructed in Section 4. The problem is that H_{uni} only looks inside individual samples, whereas collapse is a property of the distribution across them. Unfortunately, the term that actually matters, $H(G)$, cannot in general be estimated reliably from benchmark-sized sample sets without structural assumptions (Appendix A.2).

A middle ground that can be estimated is the position-wise token entropy H_{pos}, computed from per-position token frequencies across samples. Because it pools over samples rather than within them, it does catch the constant generator above: if every sample is identical, each per-position distribution is a point mass, and $H_{\mathrm{pos}} = 0$. It relates to $H(G)$, by

$$H(G) = LH_{\mathrm{pos}} - \mathrm{TC}(X_{1:L}) \implies H_\times(G,R) - LH_{\mathrm{pos}} = D_{\mathrm{KL}}(G\|R) - \mathrm{TC}(X_{1:L}), \tag{5}$$

where the total correlation $\mathrm{TC}(X_{1:L}) \geq 0$ measures dependence across positions. Because TC involves $H(G)$, it is subject to the same estimation limits. Subtracting LH_{pos} therefore does not

isolate the scorer mismatch: the adjusted score falls either when $D_{\mathrm{KL}}(G\|R)$ falls or when dependence across positions rises, which can reflect coherent language or mechanical repetition. We therefore treat it as a diagnostic, not a corrected quality metric. For completeness, Appendix A.3 reports position-adjusted gen-PPL for all generators in Table 1.

4 Constructing naive samplers with low gen-PPL

The point-mass optimum may seem too degenerate to matter in practice. We therefore construct four deliberately naive samplers that use only corpus token or phrase counts, with no learned parameters, gradients, or scorer queries. Each breaks a different form of dependence in the text. For the first three samplers, we keep the k most frequent tokens in the training corpus and assign each a probability proportional to its corpus frequency. We denote this distribution by $\hat{p}_k$.

- **Top-k IID.** Draw all L tokens independently from $\hat{p}_k$, retaining common-token frequencies while removing the dependence between positions.
- **Mirror-k.** Draw $h = \lceil L/2 \rceil$ tokens from $\hat{p}_k$, and repeat the block, which increases artificial total correlation $\mathrm{TC}(X_{1:L})$. Its position-wise token marginals match the Top-k IID, but only h draws determine the sample
- **Periodic-k.** Repeat the frequency-ordered list of the k most common tokens to length L. Every sample is identical, so $H(G) = 0$, while its within-sample unigram entropy can approach $\log k$.
- **Phrase bank-m.** Draw uniformly and independently from the m most frequent corpus 5-token phrases, concatenate, and truncate to length L. Each block appears in real data, but dependencies across blocks are discarded.

Top-k IID k=64	**Mirror** k=5000	**Periodic** k=400	**Phrase bank** m=5000
said the " on on . on , to of a the of the for with , " is his with of has and for it	*s. for still aim Russia met Congress end Grandar can „ a .Butst " paid no to resolution*	*the , . to of and a in '-s " that for. on is was withThe said as at it by from be he have has*	*the world 's largest in the world 's , " she added . " the statement said . the*

Figure 2: Opening tokens of each zero-parameter sampler. Samples in Appendix F.

5 Rebenchmarking

We now apply several evaluation metrics to recent DLMs. Table 1 evaluates public checkpoints on OpenWebText and LM1B, each at a stated budget of neural function evaluations (NFE). The discrete models are MDLM (Sahoo et al., 2024), SEDD (Lou et al., 2024), Duo (Sahoo et al., 2025), and the SDTT teacher (Deschenaux & Gülçehre, 2025); the continuous models are RDLM (Jo & Hwang, 2025), Plaid (Gulrajani & Hashimoto, 2023), RePlaid (Yang et al., 2026), LangFlow (Chen et al., 2026), ELF (Hu et al., 2026), CoBit (Batzolis et al., 2026), FLM, and its distilled flow map FMLM (Lee et al., 2026). We also include an AR baseline on OpenWebText (sources in Appendix B.2; RePlaid samples were provided by its authors). We group the measurements by the question they answer: does the generator match the data distribution, do judges prefer its individual samples, and how diverse are those samples?

Distributional discrepancies compare a set of generated samples against a set of real ones. Fréchet BERT Distance (FBD) and MAUVE do so in the feature space of a pretrained encoder (Alihosseini et al., 2019; Pillutla et al., 2021), while Maximum Mean Discrepancy (MMD) and Energy Distance (ED) work through kernels and pairwise distances (Gretton et al., 2012; Székely & Rizzo, 2013). Gradient Moment (GM) uses parameter gradients of a frozen language model (Hoogeboom et al., 2026), and Jeffreys Coding Divergence (JCD) uses byte-level compression.

Preference signals instead score samples one at a time. These include human and LLM judges (Zheng et al., 2023), as well as generative perplexity (gen-PPL), which can be read as a frozen scorer's judgment of how predictable a sample is. Finally, diversity diagnostics require no reference data at all: per-sample unigram entropy (H_{uni}) measures token variety within each sample, while Self-BLEU measures overlap across samples (Zhu et al., 2018).

5.1 DISTRIBUTIONAL DISCREPANCIES

Unlike the other discrepancies, JCD requires no learned representation. It fits byte-level context-tree weighting coders (CTW; Willems et al., 1995) to real and generated text, then measures how many extra bits per byte each held-out set costs under the other source's coder. The two directions estimate local coverage mismatch ($D_{\mathrm{KL}}(P\|G)$) and local fidelity mismatch ($D_{\mathrm{KL}}(G\|P)$), and JCD averages them, as explained in Appendix C. Together, the distributional metrics reject the constructed samplers: on both corpora, every constructed sampler has a worse mean rank than every trained generator, sits near MAUVE's floor, and scores worse than every trained generator on GM, ED, and JCD. Among trained generators, the metrics also disagree with one another: on each corpus, four different models lead at least one distributional column.

Table 1: Rebenchmarking on $1{,}024$-sample banks. Mean rank averages only the six direction-corrected distributional ranks (MAUVE, GM, JCD, ED, FBD, MMD) among generated banks per corpus. Bold and underlining mark the best and second-best values.

	Pref.	Diversity		Distributional						
Generator	gen-PPL ↓	H_{uni} ↑	Self-BLEU ↓	MAUVE↑	GM↓	JCD↓	ED↓	FBD↓	MMD↓	Mean rank ↓
OpenWebText (L=1024)										
Reference (OWT train)	17.1	5.48	0.050	1.000	0.0	—	0.000	0.0	0.000	—
AR baseline	35.3	5.58	0.047	**0.925**	6.1	0.113	0.411	372.2	0.047	**2.58**
RePlaid 1024-NFE	70.6	5.52	0.048	0.871	9.6	0.136	0.393	**368.8**	**0.033**	2.83
Plaid 1024-NFE	20.7	5.32	0.051	0.896	**1.7**	**0.080**	0.433	396.9	0.112	4.00
MDLM 1024-NFE	110.5	5.64	0.044	0.827	10.7	0.117	0.810	372.2	0.050	5.08
LangFlow 1024-NFE	37.0	5.25	0.049	0.818	9.9	0.179	**0.213**	371.9	0.098	5.17
SEDD 1024-NFE	110.0	5.62	**0.037**	0.780	11.0	0.119	0.844	371.8	0.049	5.33
ELF-M 64-NFE	22.0	5.20	0.077	0.862	7.6	0.115	0.859	384.0	0.118	5.67
Duo 1024-NFE	90.8	5.43	0.044	0.609	14.1	0.140	0.790	371.5	0.083	6.00
SDTT Teacher 1024	54.4	5.31	0.058	0.587	12.7	0.159	1.358	380.2	0.148	8.33
FLM 1024-NFE	62.8	5.34	0.069	0.396	21.1	0.209	1.735	402.7	0.182	10.17
FMLM 32-NFE	45.1	5.18	0.137	0.343	31.4	0.260	2.473	409.1	0.257	11.50
CoBit-M 256-NFE	**19.3**	5.39	0.545	0.307	19.3	0.456	2.393	716.3	0.341	12.00
Mirror (k=5000)	40.8	5.20	0.053	0.006	72.9	0.479	11.191	707.7	0.214	13.17
Periodic (k=400)	19.6	**5.93**	1.000	0.004	54.5	3.972	28.601	2227.6	0.841	15.08
Phrase bank (m=5000)	60.7	4.63	0.276	0.010	116.1	0.884	8.667	623.6	0.473	13.50
Top-k IID (k=64)	99.9	3.78	0.144	0.004	286.8	4.783	84.266	1106.4	0.588	15.58
LM1B (L=128)										
Reference (LM1B train)	56.9	4.33	0.068	1.000	0.0	—	0.000	0.0	0.000	—
Duo 1024-NFE	98.0	4.28	0.074	**0.786**	29.5	0.468	**0.020**	**82.9**	0.484	**2.25**
MDLM 1024-NFE	108.8	4.29	0.078	0.692	**29.3**	0.450	0.047	84.5	0.489	2.67
RDLM 1000-NFE	319.9	4.33	**0.067**	0.071	36.7	**0.386**	0.149	109.4	0.479	3.33
FLM 1024-NFE	97.5	4.28	0.081	0.518	32.3	0.481	0.204	89.7	0.484	3.75
LangFlow 128-NFE	139.2	**4.39**	0.069	0.009	52.5	0.510	1.576	358.0	**0.368**	5.75
LangFlow 1024-NFE	117.7	4.37	0.071	0.009	52.2	0.516	1.475	365.3	0.369	5.92
FMLM 32-NFE	72.1	4.19	0.101	0.060	45.5	0.614	1.859	118.0	0.521	6.33
FMLM 4-NFE	96.5	4.17	0.108	0.019	58.5	0.610	2.103	134.4	0.526	7.33
FMLM 1-NFE	115.0	4.13	0.108	0.007	82.4	0.681	3.070	155.7	0.544	8.83
Mirror (k=5000)	62.0	3.84	0.105	0.004	171.8	1.056	69.925	542.0	0.407	9.42
Phrase bank (m=1000)	79.4	4.04	0.432	0.004	295.1	3.464	5.845	552.9	0.584	10.92
Periodic (k=64)	**29.4**	4.16	1.000	0.004	435.0	5.951	128.144	900.1	0.964	12.58
Top-k IID (k=32)	75.0	2.99	0.355	0.004	462.9	5.288	42.599	586.3	0.710	11.92

5.2 PREFERENCE JUDGMENTS

Preference signals score individual samples rather than comparing distributions, and gen-PPL is one of them: a frozen scorer's judgment of how predictable each sample is. Human comparisons do not scale to every generator and budget, so we also collect blinded pairwise judgments from four LLMs

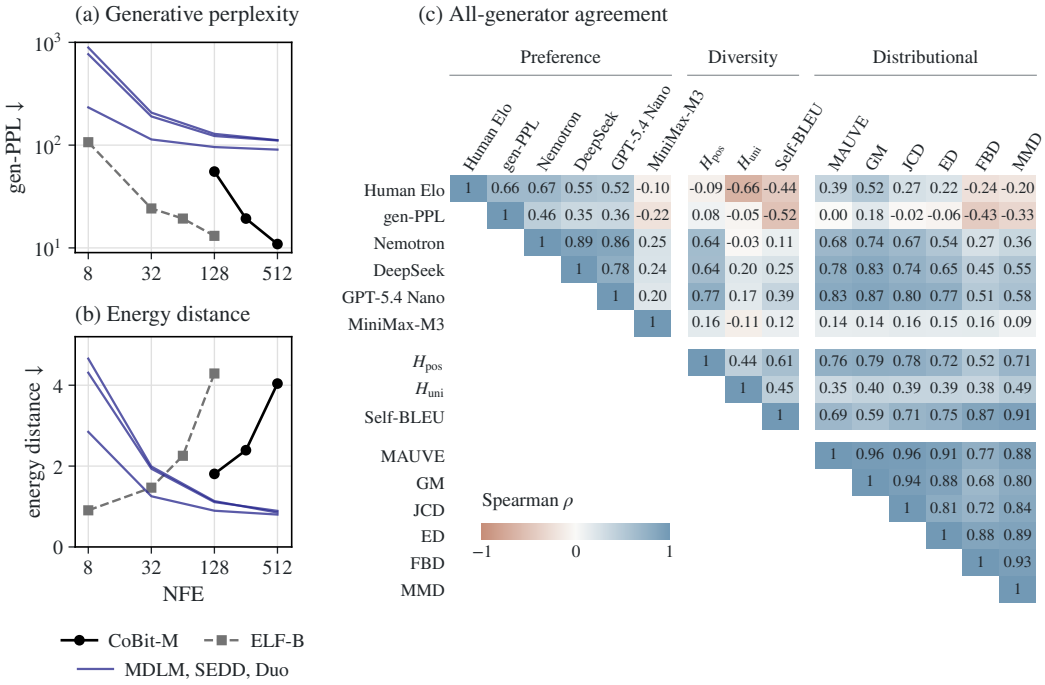


Figure 3: Sampling-budget trajectories and metric agreement on OpenWebText. **Left:** Increasing NFE lowers gen-PPL but worsens energy distance for CoBit-M and ELF-B. **Right:** Metric agreement over the 16 generated OpenWebText banks of Table 1, excluding the real-text reference. Direction-corrected tied-rank Spearman ρ; Human Elo pairs use the 12 trained generators and the final 300-comparison snapshot. The trained-only generators matrix appears in Appendix D.2.

and summarize each judge's choices with classical Elo ratings (Elo, 1978) (Appendix E). Three of the four judges rank every constructed sampler below every trained generator in Table 2, in stark contrast with gen-PPL: Periodic, second-best by gen-PPL on OpenWebText, is ranked last by Nemotron. The exception, MiniMax-M3, calls both samples bad in most comparisons, which leaves few decisive outcomes and makes its ratings noisy (Appendix E.1).

To check whether these judges track human preference, we ran a small blinded human study on the twelve trained OpenWebText generators and real text. Nine human raters saw the opening 35 words of two anonymous samples and chose the one with better fluency, coherence, and overall quality, or skipped. We fit Elo to the 300 retained comparisons with the same update as for the LLM judges (Appendix E.2). This gives each generator only a few dozen comparisons, so individual human ratings are noisy, and we read the study for its broad pattern rather than for the exact order of close generators. Humans and Nemotron broadly agree ($\rho = 0.67$; Figure 3), both placing ELF-M and Plaid at the top, which supports LLM Elo as a scalable stand-in for human preference. MiniMax-M3 barely agrees with humans, though with both sets of ratings noisy, this may reflect noise as much as genuine disagreement.

5.3 Diversity diagnostics

A judge that scores samples one at a time cannot see that every sample is the same, so, like gen-PPL, it can favor a collapsed generator. Preference must therefore be read alongside diversity guardrails. Figure 3 (left) shows why this matters for trained models. As the NFE budget grows, CoBit-M and ELF-B reach a lower gen-PPL than real text while their energy distance to the data worsens. The diagnostics show how: CoBit-M's samples increasingly resemble one another (Self-BLEU rises), and ELF-B's samples increasingly repeat themselves (H_{uni} falls). Additional steps thus appear to concentrate sampling on text the scorer finds predictable, at the expense of matching the data.

Table 2: Pairwise Elo on OWT from four LLM judges and humans (higher is better; best generated scores in bold, second best underlined). Human ratings use 300 comparisons of 35-word prefixes from 12 trained generators and real text. Elo scales differ across judges.

Generator	Nemotron	DeepSeek	GPT-5.4 Nano	MiniMax M3	Human
Reference (OWT train)	2553	2091	1835	1699	1554.9
Plaid 1024-NFE	**2026**	**1761**	**1767**	**1532**	<u>1593.8</u>
ELF-M 64-NFE	<u>1817</u>	1536	<u>1667</u>	1488	**1628.0**
CoBit-M 256-NFE	1691	1527	1516	1489	1458.6
AR baseline	1681	<u>1555</u>	1626	1478	1510.9
FMLM 32-NFE	1565	1513	1469	1496	1469.5
SDTT Teacher 1024	1548	1514	1451	1490	1528.2
LangFlow 1024-NFE	1527	1511	1564	1488	1552.3
RePlaid 1024-NFE	1513	1540	1481	1492	1495.4
MDLM 1024-NFE	1505	1450	1489	1489	1405.7
FLM 1024-NFE	1495	1492	1429	1488	1502.2
SEDD 1024-NFE	1477	1513	1451	1488	1349.2
Duo 1024-NFE	1447	1472	1491	1502	1451.3
Phrase bank (m=5000)	1294	1343	1404	1469	—
Mirror (k=5000)	727	1220	1337	1482	—
Top-k IID (k=64)	709	1258	1216	<u>1503</u>	—
Periodic (k=400)	679	1271	1279	1483	—

Diversity is also where preference and distributional metrics part ways. Every distributional discrepancy agrees with both diagnostics, as expected, since comparing a whole bank with data already penalizes lost variety. Preference signals do not agree consistently, and gen-PPL in particular tends to improve as cross-sample repetition increases (Figure 3). Neither diagnostic suffices on its own, however. OpenWebText Periodic has the highest H_{uni} of any bank, yet every sample is identical, which only Self-BLEU detects. Mirror passes both: its repetition happens within each sample, which Self-BLEU does not measure, and repeating a block barely changes a sample's token histogram, so H_{uni} misses it too.

6 Evaluation as a metric profile

This section turns the preceding results into a recommendation. Because text quality cannot be judged by humans directly at scale, evaluation relies on two kinds of proxy: distributional discrepancies, which measure how far generated text is from the data, and preference signals, which rate samples directly, whether the rater is a human, an LLM judge, or, for gen-PPL, the predictability of a frozen scorer. Each fails in a characteristic way that the other helps expose.

Preference can be maximized by collapse. A score s assigned to one sample, or to one pair against fixed opponents, has an expectation that is linear in the generator, so $\mathbb{E}_{X\sim G}[s(X)] \le \max_x s(x)$, with equality for a point mass on a best sample. Equation (4) is the special case $s = \log R$, but the argument covers any judge that cannot see that every sample is the same. Our LLM judge shows this pull in practice: among trained generators, Nemotron Elo tends to improve as both diversity diagnostics worsen. Preference must therefore be paired with entropy measured at two levels. H_{uni} detects repetition within a sample, such as ELF-B's at high NFE. Collapse across samples needs a second check, for which we use the position-wise entropy H_{pos}: it is zero for Periodic, even though Periodic has the highest H_{uni} on OpenWebText. Self-BLEU detects the same failure, but H_{pos} has a direct interpretation, since LH_{pos} is the estimable part of $H(G)$ in Equation (5). Its plug-in estimate cannot exceed $\log N$ for a bank of N samples, so comparisons require matched bank sizes.

Distributional discrepancies, in turn, see only what their representation sees. Each compares selected features or statistics of generated and reference text. Writing P_T and G_T for the pushforwards under a deterministic representation T, the data-processing inequality for f-divergences gives

$$D(P_T \| G_T) \;\le\; D(P \| G). \tag{6}$$

A small discrepancy in feature space is therefore only a lower bound on the discrepancy between the text distributions, and moment or kernel summaries of those features can hide further differences. In practice, the representation matters more than the distance computed on it. For a fixed encoder, MAUVE, Fréchet distance, energy distance, and MMD rank our generators almost identically, whereas changing the encoder can move a generator substantially: ELF-M ranks anywhere from second to ninth by Fréchet distance (Appendix D.3). A profile should therefore combine discrepancies built on different representations, rather than several distances on the same one.

For preference, we favor LLM-judge Elo over gen-PPL. gen-PPL substitutes twice: an AR model stands in for the human reader, and predictability stands in for that reader's judgment of quality, a substitution that Section 4 shows is easy to game. An LLM judge substitutes only once: it stands in for the reader but is asked the question we care about. The evidence is consistent with this. Three of the four LLM judges correlate positively with the distributional metrics, whereas gen-PPL tracks neither them nor the diversity diagnostics (Figure 3), and Nemotron broadly agrees with our human study. Because judges differ, and a noisy judge such as MiniMax-M3 says little, LLM Elo is meaningful only when reported with its judge and protocol.

We therefore recommend reporting a metric profile rather than any single number. It has two parts:

- **Preference with diversity guardrails:** LLM-judge Elo, reported with per-sample H_{uni} and position-wise H_{pos}, since Elo alone does not measure variety.
- **Distributional discrepancy:** GM, MAUVE, and ED, which compare generated and reference text through GPT-2 gradients, GPT-2-large embeddings, and handcrafted entity and discourse features. All three rank every constructed sampler below every trained generator. JCD adds a compression-based cross-check that separates coverage from fidelity.

Disagreement within the profile is itself informative: preference that improves while distributional agreement worsens signals collapse, as in the CoBit-M and ELF-B trajectories of Figure 3 (left). Where gen-PPL is needed for comparison with prior work, dividing it by $\exp(H_{\mathrm{pos}})$ removes its reward for low variety across samples, though not its reward for dependence across positions, which Mirror still exploits, as shown in Appendix A.3. Finally, the profile rests on our reading of samples and a small human study, so it is a practical default rather than a validated combination.

7 Conclusion

Generative perplexity measures how predictable a frozen scorer finds a generator's samples, so a generator can lower it simply by becoming less diverse, and the reported unigram entropy cannot detect this. In practice, samplers without learned parameters match trained DLMs on gen-PPL, and more sampling steps can lower it while moving generations away from the data. We recommend reporting the profile of Section 6 against the number of function evaluations, stating the sampling precision, which can itself lower gen-PPL (Zheng et al., 2025), and checking that any new metric or judge ranks the constructed samplers last. Few-step generation is the reason DLMs matter, and these curves show whether speed comes at the cost of quality.

This recommendation has limits. Our human study is small and uses short excerpts, so the case for LLM Elo over gen-PPL rests as much on the double-proxy argument, the constructed samplers, and agreement with the distributional metrics as on agreement with humans. LLM judges may also share part of gen-PPL's pull toward predictable text, and the diversity guardrails are partial: H_{pos} misses dependence exploits such as Mirror, and no metric here directly measures long-range coherence. Each limitation suggests a next step: validating judges against a larger human study on full-length samples, and studying systematically how each metric changes with the sampling budget.

REFERENCES

Danial Alihosseini, Ehsan Montahaei, and Mahdieh Soleymani Baghshah. Jointly measuring diversity and quality in text generation models. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pp. 90–98. Association for Computational Linguistics, 2019. doi: 10.18653/v1/W19-2311. URL https://aclanthology.org/W19-2311/.

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 preprint arXiv:2503.09573*, 2025.

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*, pp. 17981–17993, 2021.

Georgios Batzolis, Mark Girolami, and Luca Ambrogioni. CoBit: Language modeling with bitstream diffusion, 2026.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. In *Journal of Machine Learning Research*, volume 3, pp. 1137–1155, 2003.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pp. 1877–1901, 2020.

Massimo Caccia, Lucas Caccia, William Fedus, Hugo Larochelle, Joelle Pineau, and Laurent Charlin. Language GANs falling short. *International Conference on Learning Representations*, 2020.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D. Lee, Deming Chen, and Tri Dao. Medusa: Simple LLM inference acceleration framework with multiple decoding heads. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

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. *arXiv preprint arXiv:1312.3005*, 2013.

Yuxin Chen, Chumeng Liang, Hangke Sui, Ruihan Guo, Chaoran Cheng, Jiaxuan You, and Ge Liu. LangFlow: Continuous diffusion rivals discrete in language modeling, 2026.

Justin Deschenaux and Çağlar Gülçehre. Beyond autoregression: Fast LLMs via self-distillation through time. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.21035.

Sander Dieleman. Continuous diffusion language models. https://sander.ai/2026/08/24/continuous-dlms.html, August 2026. Blog post, August 24, 2026.

Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H. Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, Curtis Hawthorne, Rémi Leblond, Will Grathwohl, and Jonas Adler. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*, 2022.

Arpad E. Elo. *The Rating of Chessplayers, Past and Present*. Arco Publishing, 1978.

Aaron Gokaslan and Vanya Cohen. OpenWebText Corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Ishaan Gulrajani and Tatsunori B. Hashimoto. Likelihood-based diffusion language models, 2023. URL https://arxiv.org/abs/2305.18619.

Tatsunori B. Hashimoto, Hugh Zhang, and Percy Liang. Unifying human and statistical evaluation for natural language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 1689–1701, 2019.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1904.09751.

Emiel Hoogeboom, David Ruhe, Jonathan Heek, Thomas Mensink, and Tim Salimans. Beyond single tokens: Distilling discrete diffusion models via discrete MMD. *arXiv preprint arXiv:2603.20155*, 2026.

Keya Hu, Linlu Qiu, Yiyang Lu, Hanhong Zhao, Tianhong Li, Yoon Kim, Jacob Andreas, and Kaiming He. ELF: Embedded language flows, 2026.

Jaehyeong Jo and Sung Ju Hwang. Continuous diffusion model for language modeling, 2025.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016. doi: 10.48550/arXiv.1602.02410. URL https://arxiv.org/abs/1602.02410.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. In *arXiv preprint arXiv:2001.08361*, 2020.

Chanhyuk Lee, Jaehoon Yoo, Manan Agarwal, Sheel Shah, Jerry Huang, Aditi Raghunathan, Seunghoon Hong, Nicholas M. Boffi, and Jinwoo Kim. Flow map language models: One-step language modeling via continuous denoising. *arXiv preprint arXiv:2602.16813*, 2026.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. Diffusion-LM improves controllable text generation. In *Advances in Neural Information Processing Systems*, 2022.

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*, 2024.

Fred Zhangzhi Peng, Zachary Bezemek, Sawan Patel, Jarrid Rector-Brooks, Sherwood Yao, Avishek Joey Bose, Alexander Tong, and Pranam Chatterjee. Path planning for masked diffusion model sampling, 2025. URL https://arxiv.org/abs/2502.03540.

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*, pp. 4816–4828, 2021.

Patrick Pynadath, Jiaxin Shi, and Ruqi Zhang. Generative frontiers: Why evaluation matters for diffusion language models. *arXiv preprint arXiv:2604.02718*, 2026.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 2019.

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

Subham Sekhar Sahoo, Justin Deschenaux, Aaron Gokaslan, Guanghan Wang, Justin T. Chiu, and Volodymyr Kuleshov. The diffusion duality. *arXiv preprint arXiv:2506.10892*, 2025.

Mitchell Stern, Noam Shazeer, and Jakob Uszkoreit. Blockwise parallel decoding for deep autoregressive models. In *Advances in Neural Information Processing Systems*, 2018.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.01844.

Gregory Valiant and Paul Valiant. Estimating the unseen: An $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing*, pp. 685–694, 2011.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Guanghan Wang, Yair Schiff, Subham Sekhar Sahoo, and Volodymyr Kuleshov. ReMDM: Remasking discrete diffusion models with inference-time scaling. *arXiv preprint arXiv:2503.00307*, 2025.

Frans M. J. Willems, Yuri M. Shtarkov, and Tjalling J. Tjalkens. The context-tree weighting method: Basic properties. *IEEE Transactions on Information Theory*, 41(3):653–664, 1995.

Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016.

Zhihan Yang, Wei Guo, Shuibai Zhang, Subham Sekhar Sahoo, Yongxin Chen, Arash Vahdat, Morteza Mardani, and John Thickstun. Continuous diffusion scales competitively with discrete diffusion for language, 2026.

Kaiwen Zheng, Yongxin Chen, Hanzi Mao, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2409.02908.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023. URL https://arxiv.org/abs/2306.05685.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1097–1100. ACM, 2018. doi: 10.1145/3209978.3210080. URL https://doi.org/10.1145/3209978.3210080.

AI USE STATEMENT

We used AI agents to implement and run experiments, create visualizations, plots, and figures, and draft and edit the manuscript. We also used them more selectively to survey literature that informed ideation. For example, the compression-based evaluation idea originated with the lead author; the specific choice of context-tree weighting (CTW) was suggested during brainstorming by an AI agent. The authors are responsible for the final methods, results, claims, and wording.

APPENDIX CONTENTS

A Entropy and the gen-PPL decomposition **14**

A.1 Entropy quantities and their ordering 14

A.2 Why entropy subtraction requires assumptions 15

A.3 Position-wise entropy adjustment 15

B Benchmark protocol **16**

B.1 Frozen metric configuration 16

B.2 Sample banks and checkpoint sources 16

B.3 Attack parameter sweeps 17

C Jeffreys coding divergence **17**

C.1 Estimator 17

C.2 Frozen context-tree weighting 17

C.3 Directional results 18

C.4 Limitations 19

D Metric sensitivity **20**

D.1 MAUVE and the number of scored samples 20

D.2 Metric agreement across generators 21

D.3 Encoder-space sensitivity 21

E Pairwise judgments **23**

E.1 LLM judges 23

E.2 Human comparisons 23

F Text Samples **24**

F.1 OpenWebText samples 24

F.2 LM1B samples 29

A ENTROPY AND THE GEN-PPL DECOMPOSITION

This appendix defines the entropy quantities used in the paper and relates them to one another. It then explains why the term that matters for gen-PPL, the joint entropy $H(G)$, cannot be estimated or subtracted without further assumptions.

A.1 ENTROPY QUANTITIES AND THEIR ORDERING

Let $X = (X_1, \dots, X_L) \sim G$. For a token v, we write $p_t(v) = \Pr_G(X_t = v)$ for the marginal at position t, $\bar{p}(v) = \frac{1}{L}\sum_t p_t(v)$ for the marginal pooled over positions, and $\widehat{p}_x(v) = \frac{1}{L}\sum_t \mathbf{1}[x_t = v]$ for the token histogram of a single sample x. These give four entropies:

$$H_{\text{pool}} = H(\bar{p}), \qquad H_{\text{uni}} = \mathbb{E}_{X\sim G}\big[H(\widehat{p}_X)\big],$$
$$H_{\text{pos}} = \tfrac{1}{L}\textstyle\sum_{t=1}^{L} H(p_t), \qquad H_{\text{joint}} = H(G) = H(X_1, \dots, X_L). \tag{7}$$

The first three are measured in nats per token and H_{joint} in nats per sample, so we compare H_{joint}/L with the others. Each discards different information. H_{pool} ignores which sample or position produced a token; H_{uni} keeps each sample's histogram but discards token order; H_{pos} measures variation across samples at each position but ignores dependence between positions; and only H_{joint} measures uncertainty over complete outputs.

Concavity of entropy gives $H_{\text{uni}} \le H_{\text{pool}}$ and $H_{\text{pos}} \le H_{\text{pool}}$, since $\bar{p}$ is the average both of the per-sample histograms and of the position marginals. Subadditivity gives $H_{\text{joint}}/L \le H_{\text{pos}}$, and the gap is the total correlation:

$$L\,H_{\text{pos}} - H(G) = \mathrm{TC}(X_{1:L}) = D_{\mathrm{KL}}\Big(G \,\Big\|\, \textstyle\prod_t p_t\Big) \ge 0, \tag{8}$$

which is the identity on the left of Equation (5). H_{uni} and H_{pos}, by contrast, are not ordered in general, and no marginal quantity determines H_{joint}. Two of our constructed samplers show why. Periodic always emits the same token-rich sample, so it has high H_{uni} and H_{pool}, yet $H_{\text{pos}} = H_{\text{joint}} = 0$. Mirror keeps every position marginal of Top-k IID, and hence its H_{pos}, but halves its joint entropy.

Figure 4 illustrates the collapsed optimum of Section 3 along the path $G_\alpha = (1-\alpha)R + \alpha\delta_{x^\star}$. The cross-entropy is linear in α,

$$H_\times(G_\alpha, R) = (1-\alpha)\,H(R) - \alpha \log R(x^\star),$$

and since $-\log R(x^\star) \le H(R)$, it decreases from distribution matching to collapse: every step toward the point mass improves gen-PPL.

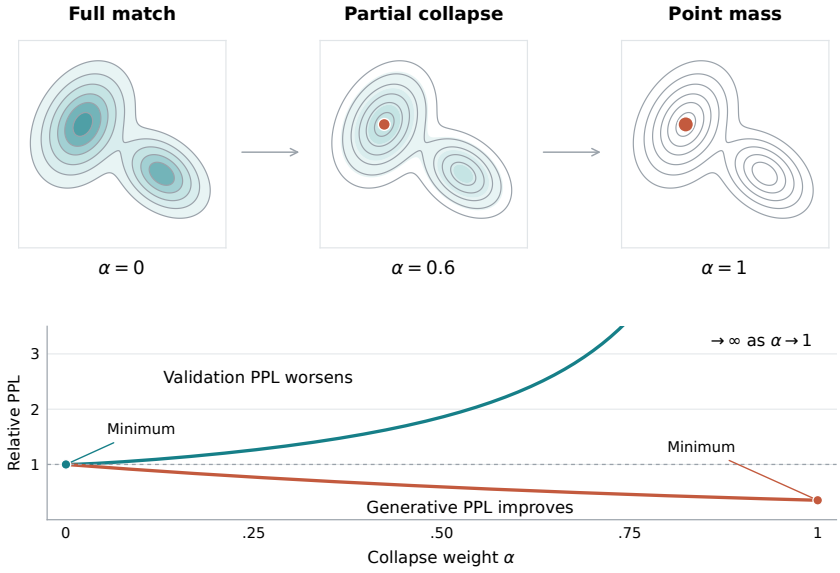


Figure 4: Interpolation from distribution matching ($\alpha = 0$) to collapse ($\alpha = 1$), along the path $G_\alpha = (1-\alpha)R + \alpha\delta_{x^\star}$. Gray contours show R and color shows G_α.

A.2 Why entropy subtraction requires assumptions

If $H(G)$ were known, subtracting it would remove gen-PPL's reward for low entropy, leaving $H_\times(G, R) - H(G) = D_{\mathrm{KL}}(G\|R)$. Even with an ideal scorer, this becomes $D_{\mathrm{KL}}(G\|P)$, which penalizes generated mass where data are rare but not data modes the generator omits. It therefore still differs from the coverage-sensitive $D_{\mathrm{KL}}(P\|G)$ that validation perplexity would measure for G.

In practice, however, $H(G)$ is not known, and it cannot in general be estimated reliably from benchmark-sized sample banks without assumptions on the generator. When nearly all N observed sequences are distinct, as they are for any reasonably diverse generator, the empirical entropy is close to $\log N$, however much larger the true entropy may be. Lower bounds for entropy estimation over large discrete supports formalize this limitation (Valiant & Valiant, 2011; Wu & Yang, 2016). Estimation remains possible under explicit structural assumptions, such as finite-order Markov dependence.

A.3 Position-wise entropy adjustment

Because H_{pos} can be estimated, one partial correction is to divide gen-PPL by $\exp(H_{\mathrm{pos}})$. By Equation (5), this leaves $\exp\big((D_{\mathrm{KL}}(G\|R) - \mathrm{TC}(X_{1:L}))/L\big)$. The adjustment removes Periodic's advantage: its H_{pos} is zero, so its score is unchanged, while every diverse generator's score shrinks (Figure 5). It remains gameable, however. Because it subtracts total correlation, a generator can improve its adjusted score by adding dependence across positions. LM1B Mirror, which copies its first half into its second, scores better than every trained generator.

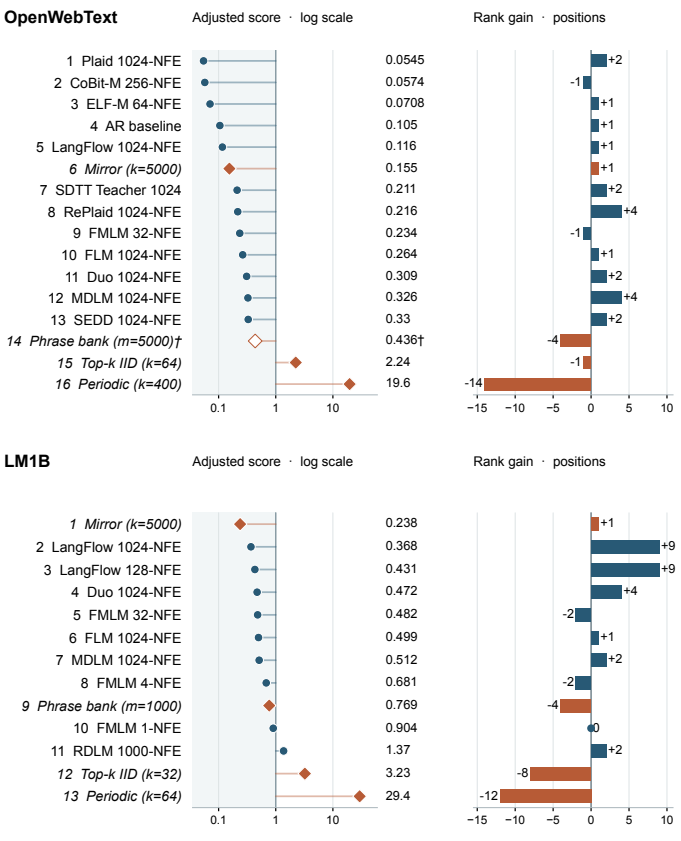


Figure 5: Position-wise entropy adjustment for the generators in Table 1. Left: gen-PPL divided by $\exp(H_{\mathrm{pos}})$, log scale. Right: original gen-PPL rank minus adjusted rank.

B BENCHMARK PROTOCOL

B.1 FROZEN METRIC CONFIGURATION

All metrics use the same 1,024-sample banks and matching reference banks, with each metric's configuration frozen across generators. JCD fits on 512 documents per source and scores the remaining 512; the other metrics score all 1,024 samples. Table 3 lists the settings.

Table 3: Frozen configuration of the evaluation metrics in Table 1. Max tokens equals the corpus sequence length (128 for LM1B, 1,024 for OpenWebText) unless stated.

Metric	Configuration
gen-PPL, H_{uni}	`gpt2-large` scorer; exp of the mean per-sample next-token NLL; batch 8; text input without added special tokens.
MAUVE	`mauve-text` 0.4.0; `gpt2-large` features, 512 tokens; k-means on features explaining 90% variance (5 restarts, 500 iterations); scaling factor 5; five subset seeds at n=1,024 per side.
FBD	Fréchet distance between Gaussian fits of `gpt2-large` features.
GM	Squared difference of mean parameter gradients under `gpt2`; batch 4.
ED	Energy distance on handcrafted entity and discourse features, standardized by reference mean and scale.
MMD	Unbiased MMD^2 with an L2-normalized character 3–5-gram cosine kernel; 1,024 balanced rows, selection seed 1729.
JCD	CTW-20 over UTF-8 bits; 512 fitting and 512 held-out documents per source; split seeds 0–4 (Appendix C).
Self-BLEU	Mean pairwise within-bank sentence BLEU, n-gram orders 2–4.

B.2 SAMPLE BANKS AND CHECKPOINT SOURCES

Every evaluated corpus is frozen as a bank of 1,024 ordered samples with a manifest recording the generator, checkpoint digest, code commit, and job. We generate samples from public checkpoints with the authors' released samplers; the only exception is RePlaid, whose samples were provided by its authors. Table 4 lists the source of each trained generator in Table 1; representative samples appear in Appendix F.

Table 4: Sources of the trained generators in Table 1. Hugging Face (HF) and GitHub (GH) locations are those from which checkpoint weights were obtained.

Generator	Corpus	Checkpoint source
AR baseline	OWT	GH `s-sahoo/duo` (Sahoo et al., 2025)
MDLM	OWT	HF `kuleshov-group/mdlm-owt` (Sahoo et al., 2024)
MDLM	LM1B	GH `s-sahoo/duo` (Sahoo et al., 2025)
SEDD	OWT	HF `louaaron/sedd-medium` (Lou et al., 2024)
Duo	OWT, LM1B	GH `s-sahoo/duo` (Sahoo et al., 2025)
SDTT Teacher	OWT	HF `jdeschena/sdtt` (`teacher_1M_sm`) (Deschenaux & Gülçehre, 2025)
FLM, FMLM	OWT, LM1B	GH `david3684/flm` (Lee et al., 2026)
LangFlow	OWT, LM1B	HF `Continuous-Rivals-Discrete` (Chen et al., 2026)
RDLM	LM1B	GH `harryjo97/RDLM` (Jo & Hwang, 2025)
ELF-M	OWT	HF `embedded-language-flows/ELF-M-owt-torch` (Hu et al., 2026)
CoBit-M	OWT	HF `gbatzolis/CoBit` (EMA, 750k steps) (Batzolis et al., 2026)
Plaid 1B	OWT	GH `igul222/plaid` (Gulrajani & Hashimoto, 2023)
RePlaid	OWT	Samples provided by the authors (Yang et al., 2026)

B.3 Attack parameter sweeps

Each constructed family is a one-parameter sweep. Figure 6 plots every measured bank of each family; the configurations in Table 1 were fixed before the rebenchmark, and neighbouring settings reach the same region, so no result depends on a fortunate choice of k or m.

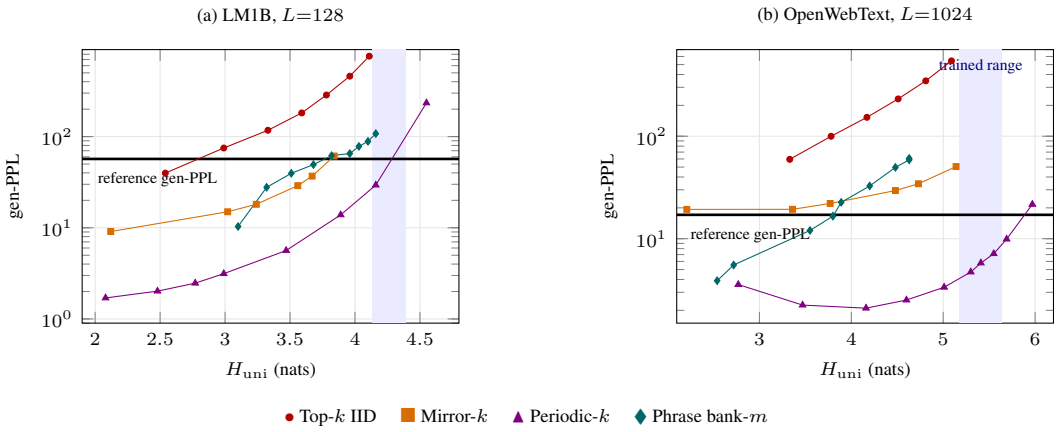


Figure 6: Measured parameter sweeps for the four training-free samplers in the gen-PPL/entropy plane. On OpenWebText, Periodic configurations within the trained generators' entropy range (shaded) have measured gen-PPL 4.73–7.14, compared with 17.1 for reference text (solid line). Table 1 reports OpenWebText Periodic-400 at gen-PPL 19.6 and unigram entropy 5.93. Every plotted point is a measured bank.

C Jeffreys coding divergence

C.1 Estimator

If training distribution P and DLM distribution G were available, each directional KL would be the excess of a cross-code over a self-code:

$$H_\times(P,G) - H(P) = D_{\mathrm{KL}}(P\|G), \qquad \text{coverage mismatch,} \tag{9}$$

$$H_\times(G,P) - H(G) = D_{\mathrm{KL}}(G\|P), \qquad \text{fidelity mismatch.} \tag{10}$$

Neither $P(x)$ nor the DLM likelihood $G(x)$ is available, so we apply one external learning rule C to a fitting split of each source and freeze the result,

$$P_{\text{fit}} \xrightarrow{C} q_P, \qquad G_{\text{fit}} \xrightarrow{C} q_G, \tag{11}$$

and score held-out banks $\mathcal{B}_P, \mathcal{B}_G$ under both frozen coders:

$$\widehat{D}_{\text{cov}} = \frac{\sum_{x\in\mathcal{B}_P} \log_2 \frac{q_P(x)}{q_G(x)}}{\sum_{x\in\mathcal{B}_P} |x|_{\text{bytes}}}, \qquad \widehat{D}_{\text{fid}} = \frac{\sum_{x\in\mathcal{B}_G} \log_2 \frac{q_G(x)}{q_P(x)}}{\sum_{x\in\mathcal{B}_G} |x|_{\text{bytes}}}, \qquad \mathrm{JCD} = \tfrac{1}{2}\big(\widehat{D}_{\text{cov}} + \widehat{D}_{\text{fid}}\big). \tag{12}$$

Each held-out document is scored under both coders, so its intrinsic difficulty and the coder's capacity largely cancel in the log-ratio. The estimates are byte-normalized, finite-context surrogates for (9)–(10), conditional on observed byte length, and pooled over documents so that longer documents weigh more. We write $\mathrm{JCD}_{\rightarrow} = \widehat{D}_{\text{cov}}$ and $\mathrm{JCD}_{\leftarrow} = \widehat{D}_{\text{fid}}$; Table 1 reports their mean.

C.2 Frozen context-tree weighting

We instantiate C as context-tree weighting (CTW) at depth D=20 over UTF-8 bytes expanded into bits, most significant first (Figure 7). Each context s of length at most D stores bit counts (n_0^s, n_1^s); histories are zero-padded and reset at document boundaries. Each node has a Krichevsky–Trofimov predictive and marginal likelihood,

$$P_{\text{kt}}^s(b) = \frac{n_b^s + \frac{1}{2}}{n_0^s + n_1^s + 1}, \qquad \bar{P}_{\text{kt}}^s = \frac{\Gamma(\frac{1}{2} + n_0^s)\,\Gamma(\frac{1}{2} + n_1^s)}{\Gamma(\frac{1}{2})^2\,\Gamma(1 + n_0^s + n_1^s)}, \tag{13}$$

and the marginals define the weighted likelihood $P_{\mathrm{w}}^{s} = \frac{1}{2}\bar{P}_{\mathrm{kt}}^{s} + \frac{1}{2}P_{\mathrm{w}}^{s0}P_{\mathrm{w}}^{s1}$ (with $P_{\mathrm{w}}^{s} = \bar{P}_{\mathrm{kt}}^{s}$ at depth D). Its posterior stop weight $\beta_s = \frac{1}{2}\bar{P}_{\mathrm{kt}}^{s}/P_{\mathrm{w}}^{s}$ is computed once at fit time and frozen. Scoring descends the context path of each bit from the leaf upward,

$$q_t^{(D)}(b) = P_{\mathrm{kt}}^{s_t^{(D)}}(b), \qquad q_t^{(d)}(b) = \beta_{s_t^{(d)}} P_{\mathrm{kt}}^{s_t^{(d)}}(b) + \big(1 - \beta_{s_t^{(d)}}\big) q_t^{(d+1)}(b), \tag{14}$$

so $q(b_t \mid b_{<t}) = q_t^{(0)}(b_t)$ is a convex combination of all $D{+}1$ depths, and $-\log_2 q(x) = -\sum_t \log_2 q(b_t \mid b_{<t})$ is accumulated bit by bit at $O(D)$ cost per bit. Target bits select the path but never update counts or weights. We deliberately avoid the adaptive identity $q(b_t \mid b_{<t}) = P_{\mathrm{w}}^{\varnothing}(b_{\le t})/P_{\mathrm{w}}^{\varnothing}(b_{<t})$, which would let each coder learn the bank it is scoring and shrink the directional gap for reasons unrelated to the generator. For example, with $D{=}2$ and counts $\varnothing{:}(5,3)$, $0{:}(3,1)$, $00{:}(2,1)$, the first bit of $x = 010$ receives $q(0) = 0.6296$, a mixture with weights $(0.667, 0.185, 0.148)$ of the node predictives 0.611, 0.700, and 0.625; a depth-20 coder runs the same recursion over 21 levels.

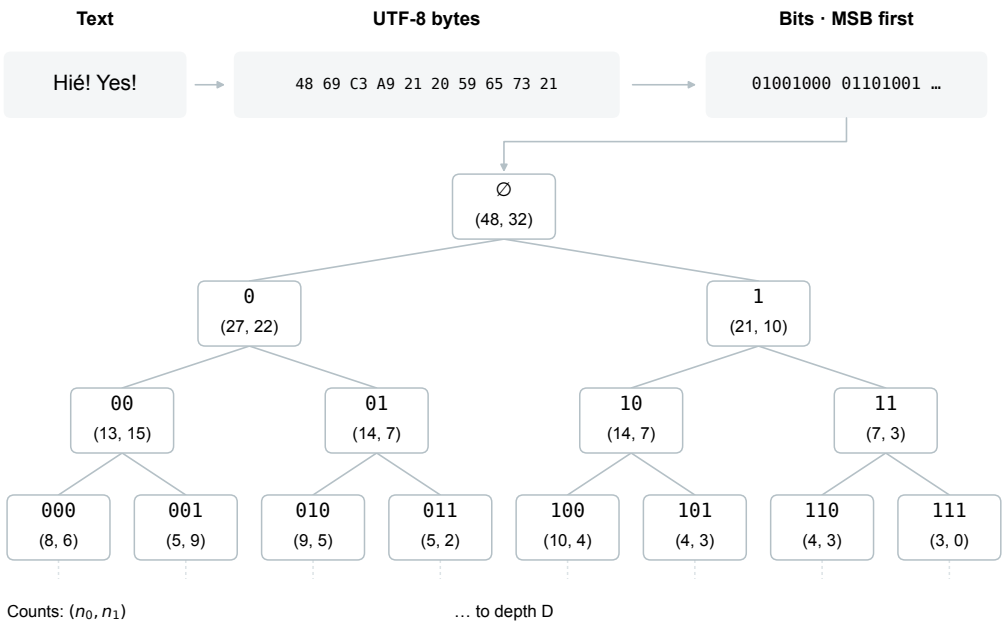


Figure 7: **Fitting.** One source is read as a raw UTF-8 bit stream and used to grow a context tree. Each node is labelled by the bit context that reaches it—$\varnothing$ at the root, then 0 and 1, then $00, 01, 10, 11$, and so on—and stores the counts (n_0, n_1) of the bits observed in that context, from which it forms a Krichevsky–Trofimov estimate.

C.3 Directional results

Figure 8 resolves JCD into its two directions for the trained generators. The directions carry information lost in the average: ELF-M 64-NFE has coverage excess 0.181 but fidelity excess 0.050 bits/byte, whereas Plaid is nearly balanced (0.085 and 0.075). Both directions reject the constructed samplers, especially coverage: OpenWebText Periodic and Top-k IID reach 5.94 and 7.85 bits/byte, against at most 0.48 for any trained OpenWebText generator.

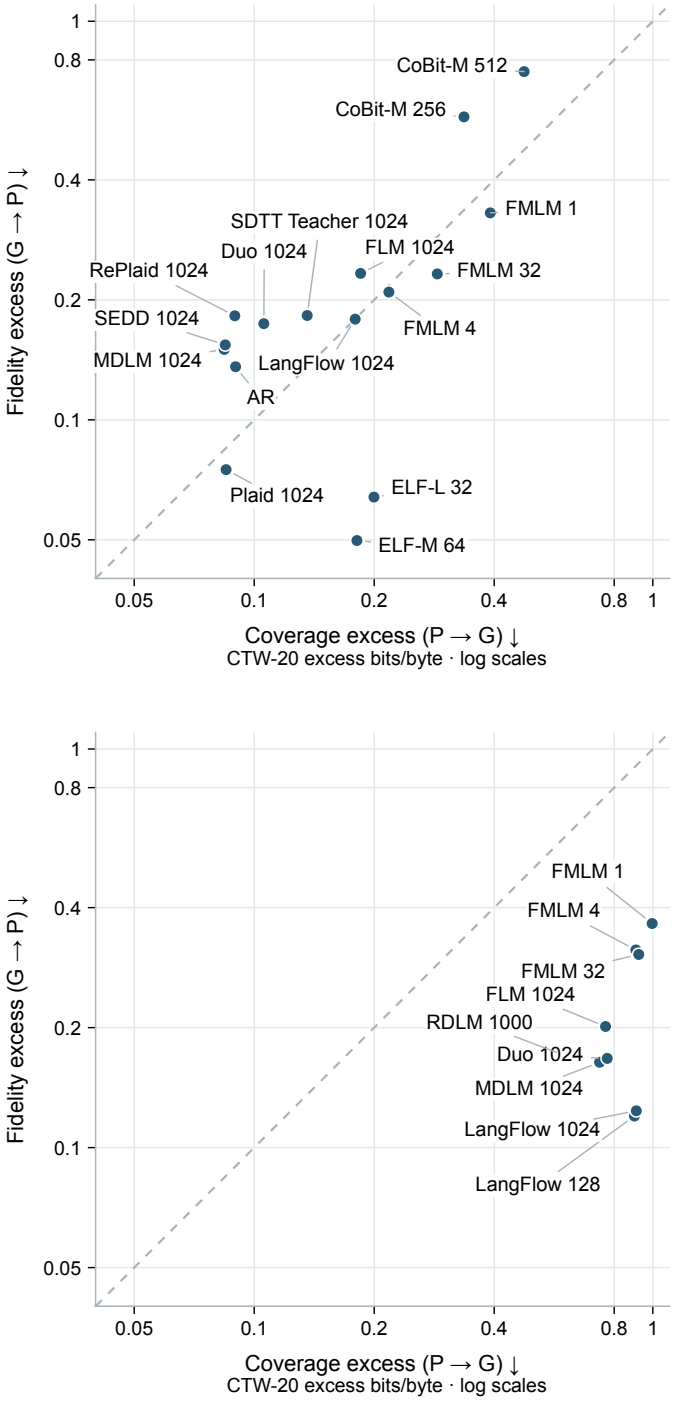


Figure 8: Directional Jeffreys coding divergence for trained configurations on OpenWebText (L=1024, top) and LM1B (L=128, bottom), with identical log axes. Horizontal: local coverage mismatch; vertical: local fidelity mismatch. Numbers in labels denote NFE.

C.4 Limitations

A depth-20 coder conditions each bit on at most 2.5 bytes of history, so it can miss discourse coherence, long-range repetition, and agreement between distant tokens; a small JCD cannot establish

global coherence. Of its 2^{20} deepest contexts, many are unobserved on finite fitting splits and fall back toward the $\frac{1}{2}$ prior, possibly differently for real and generated coders, and raising D worsens this sparsity.

D METRIC SENSITIVITY

D.1 MAUVE AND THE NUMBER OF SCORED SAMPLES

MAUVE depends on the number of scored texts even with the corpus, features, and clustering fixed (Figure 9); the values in Table 1 use n=1,024 per side. Despite this sensitivity, MAUVE largely preserves the ordering of generators across sample counts; what changes most is the size of the gaps. On LM1B the ranking is essentially fixed (pairwise Spearman $\rho \geq 0.98$ across $n \in \{128, 256, 512, 1{,}024\}$). On OpenWebText the coarse ordering holds ($\rho = 0.81$ between n=128 and 1,024, and 0.96 between 512 and 1,024), although closely scored generators can swap; Duo, for example, falls from fourth to eighth as n grows. MAUVE scores are higher at n=128 than at n=1,024 for every displayed generator. On OpenWebText, the corresponding score spreads are 0.36 and 0.62. Comparisons should therefore use matched sample counts.

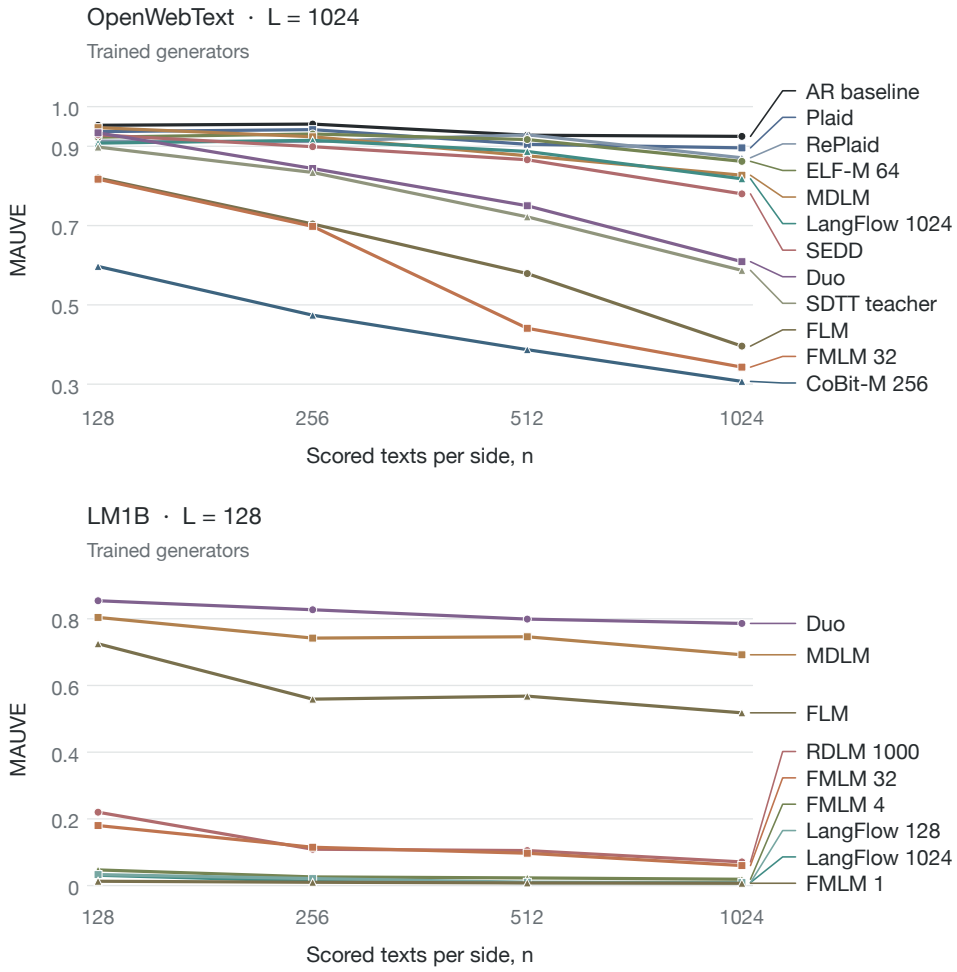


Figure 9: MAUVE sensitivity to scored sample count for trained generators on OpenWebText (top) and LM1B (bottom) from Table 1. Points average five fixed subset seeds with n reference and n generated texts. Panels use their labeled y-axis ranges.

D.2 METRIC AGREEMENT ACROSS GENERATORS

We constructed the naive samplers specifically to break gen-PPL, so they could dominate the agreement in Figure 3. Figure 10 therefore recomputes the same matrix over the 12 trained generators alone, so that every pair uses the same cohort. Without the naive samplers, gen-PPL agrees strongly with Nemotron Elo ($\rho = 0.90$, compared with 0.46 over 16 generators) but negatively with FBD, MMD, H_{uni}, and Self-BLEU. The distributional metrics still agree with one another ($\rho = 0.36$–0.95). Both figures use the final 300-comparison Human Elo snapshot. Human Elo pairs use the same 12 trained generators in both figures, so their coefficients are identical.

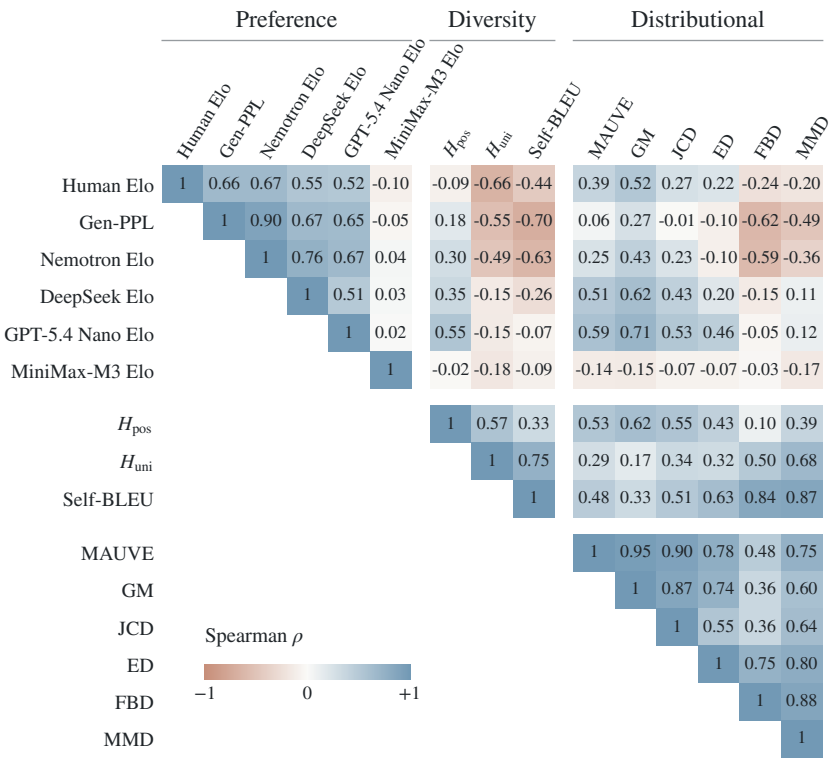


Figure 10: Metric agreement over the 12 trained OpenWebText generators, excluding the naive samplers and the real-text reference. Direction-corrected tied-rank Spearman ρ, computed as in Figure 3. Human Elo uses the final 300-comparison snapshot.

D.3 ENCODER-SPACE SENSITIVITY

We recompute MAUVE, Fréchet distance, energy distance, and MMD^2 under five frozen representation pipelines: GPT-2-large, BERT-base-uncased, RoBERTa-base, all-MPNet-base-v2, and E5-base-v2, each with its own tokenizer and pooling, a 512-token cap, unit-normalized document vectors, and reference-only PCA whitening to at most 128 dimensions. These are sensitivity variants, not recomputations of Table 1: here MAUVE uses 50 buckets, ED uses Euclidean distance in the learned features, and MMD a three-bandwidth Gaussian RBF kernel. For a fixed encoder, rankings agree closely across the four metrics (pairwise Spearman 0.922–0.997, mean 0.970; Figure 12), while changing the encoder moves several generators (Figure 11). The representation matters more here than the distance.

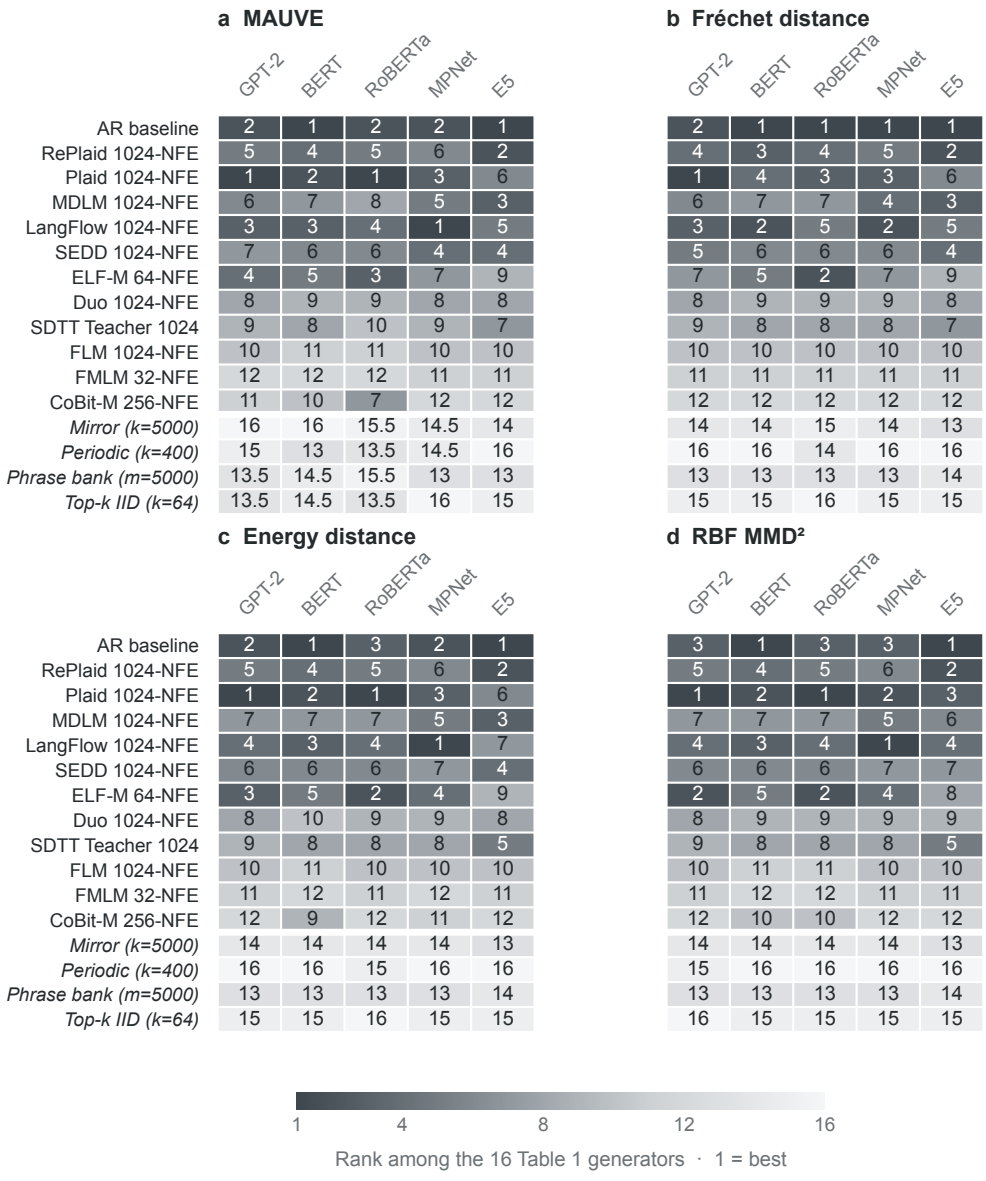


Figure 11: Encoder-space sensitivity of four feature-space metrics on OpenWebText: (a) MAUVE, (b) squared Gaussian Fréchet distance, (c) unbiased Euclidean energy distance, and (d) unbiased three-bandwidth Gaussian RBF MMD^2. Rows show the 16 OpenWebText generators in Table 1. Each cell gives a rank within this cohort (1 best), with average ranks for exact ties; darker shading indicates a better rank. Rankings use 1,024 reference and 1,024 generated texts per score. MAUVE is ranked high to low; the other three scores are ranked low to high.

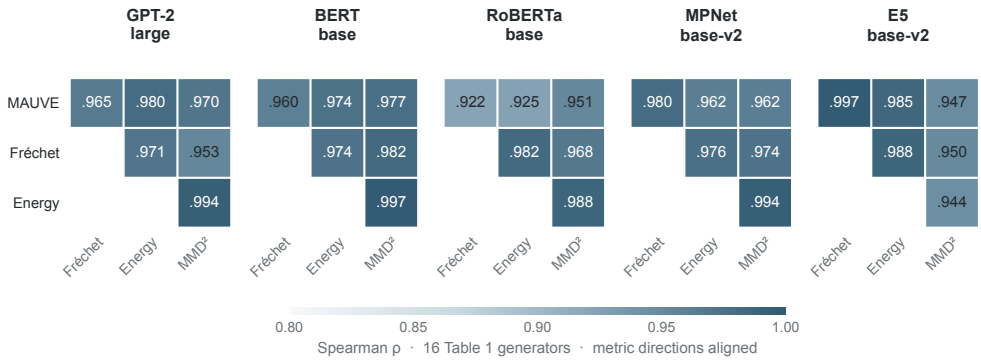


Figure 12: Within-encoder metric agreement on OpenWebText. Each upper triangle gives pairwise Spearman correlations across the 16 generators in Table 1 (12 trained and four constructed), excluding the real-text reference. Metric directions are aligned and ties receive average ranks; self-correlations and duplicate pairs are omitted.

E Pairwise judgments

E.1 LLM judges

LLM judges give a scalable preference signal, but each depends on its model, prompt, and protocol. We therefore use four judges of equal standing, Nemotron-3-Ultra-550B-A55B, DeepSeek, GPT-5.4 Nano, and MiniMax-M3, under an identical protocol, and report each separately.

Each judge sees two anonymous complete samples with a fixed prompt that asks it to compare their readability, coherence, naturalness, and overall quality. The prompt treats the sample text as untrusted data and requires a single JSON choice among `a_better`, `b_better`, `both_good`, and `both_bad`. All four judges rate the same schedule of 16,560 comparisons across the 21 frozen OpenWebText banks, including the real-text reference, built from seeded sample pairs shown in both presentation orders. After canonicalizing presentation order, a choice scores $s = 1$ or 0 for the left bank, and either tie scores $s = \frac{1}{2}$. Ratings start at 1500 and, separately for each judge, are updated in a fixed, content-derived order by

$$E_{ij} = \left(1 + 10^{(R_j - R_i)/400}\right)^{-1}, \qquad R_i \leftarrow R_i + 32\,(s - E_{ij}), \qquad R_j \leftarrow R_j - 32\,(s - E_{ij}). \quad (15)$$

Elo is zero-sum and order-dependent, so a rating summarizes a particular judge, cohort, prompt, and schedule rather than absolute quality, and rating scales are not comparable across judges. Table 2 displays the 17 OpenWebText banks of Table 1, with ratings retained from the full 21-bank campaign rather than refitted to the displayed subset.

The judges differ most in how often they decline to pick a winner. MiniMax-M3 answers `both_bad` in 93% of comparisons, against 58% for DeepSeek and GPT-5.4 Nano and 22% for Nemotron. Because every tie scores $\frac{1}{2}$, a judge that rarely decides moves its ratings little: MiniMax-M3's ratings of generated banks span only 1469–1532, and its ordering is correspondingly noisy, placing Top-k IID third and the AR baseline twentieth of 21 banks. We therefore report judges separately rather than averaging them, and read MiniMax-M3's correlations in Figures 3 and 10 with the most caution.

To compare judges with metrics, ratings and metric values are joined on exact dataset, generator identity, and bank digest over the 16 generated OpenWebText banks of Table 1. Each metric is oriented so that larger is better, and ties receive average ranks. Spearman ρ describes agreement within this finite cohort: we report no confidence intervals, and agreement is not validation against human preference. The prompt and the per-step rating traces of all four judges are released.

E.2 Human comparisons

The human study checks whether the LLM judges track human preference. It covers the twelve trained OpenWebText generators of Table 1 and the real-text reference. Constructed samplers were

not shown, so the study speaks only to rankings among trained generators. Raters see the first 35 whitespace-delimited words of two anonymous samples and choose the excerpt with better fluency, coherence, and overall writing quality, or skip. Skips do not affect Elo, which applies the update in Equation (15) in recorded order.

All reported human ratings and correlations use a frozen snapshot of the study, retrieved at 11:59 CEST on 25 September 2026. Nine raters recorded 304 comparisons; excluding four flagged comparisons leaves 300, covering all 78 source pairs with 41–49 matches per source. This is enough to show the broad pattern, but not to order close generators reliably: with so few matches per source, individual ratings are noisy, and with twelve generators every correlation is imprecise (Section 7). By unrounded human Elo, the order is ELF-M, Plaid, the reference, LangFlow, SDTT Teacher, AR, FLM, RePlaid, FMLM, CoBit-M, Duo, MDLM, and SEDD; Table 2 shows one decimal to distinguish close scores.

The human and LLM evaluations also differ in design. Humans judged short prefixes, whereas the LLM judges assessed complete samples, and human ratings follow the recorded order of comparisons rather than a content-derived one. Absolute Elo levels are therefore not comparable between the two, and we compare only rankings. Human correlations in Figures 3 and 10 and in the main text use the twelve trained generators, full-precision human Elo, the displayed integer LLM ratings, and average ranks for ties; neither campaign is refitted to a displayed subset. These correlations describe this cohort. They do not establish a definitive ordering of model quality, nor do they validate the recommended metric profile against human preference.

F TEXT SAMPLES

One sample per bank, selected at random without inspection. LM1B samples are shown in full (L=128); OpenWebText samples are truncated.

F.1 OPENWEBTEXT SAMPLES

Reference #477

<|endoftext|> before I headed out I put the girls in their Vikings jammies and got ready to leave. Before I left I asked my older daughter a very important question: "Are the Vikings gonna win tonight Gemma?" "Yep! Skoohhh Vikings, less Gooo!" And then she blew me a kiss goodbye. Just like the Packers kissed the NFC North title goodbye.<|endoftext|>Sprites: A Rarely Seen Sky Phenomenon Caught On Camera When thunderstorms emit lightning, we see the white, snaking electricity from the ground. But if you flew above the clouds, you would see a sky phenomenon known as sprites. These are rarely seen bolts of red light that look like very fast burning sparklers. The Capital Weather Gang over at The Washington Post describes them like this: "[Sprites] can reach 50-60 miles into space and penetrate downward into the middle of the stratosphere (15-20 miles high) with jellyfish-like tendrils." [...]

AR baseline #578
NFE = 1 024

<|endoftext|> have clean, inexpensive machines packed full of the most powerful computer devices you can imagine — all operated in a simple and efficient manner that involves little risk. Plus, why not choose a PC that has the ease of integration with existing plumbing (or mesh, as some will call it) and is equipped with 4K video cameras? Ty to the television, then. Generally speaking, the purest multimedia vinyl The One (I guess because that's what it is) died before it got to the last page. And since Microsoft didn't deliver, those rolls of vinyl—which I happen to like, at least verging on sentimental—aren't going to be available to purchase on October 25th anyway. But hey, at least you'll be able to get out there and sample some tunes for two years. Oh, and it's less costly to upgrade to a Surface Pro 3 with Kinect gadgets. BeCaClkOojckPPJmbMPur. How expensive is the Surface [...]

RePlaid (no-SC, DDPM) #123
NFE = 1 024

<|endoftext|> Indonesia, rejected an unprecedented action by the emirate to imbue Burundi with the Right of Return in response to the public and the controversial question of whether the historical event should be celebrated, wrote Arthur Duncan Snoville. Rapovano resigned from the Honya and government of Burundi in February 2014, stating that he wants Archbishop Dutton to give him "pay" for his actions in the past. The widely published islands list includes hurricanes, but the Honya's effective lobbying is making their case to a global public including Pope Benedict XVI. Archbishop Dolan's comments were aimed at connecting Indonesia's largest island nation with the French-controlled territory to the east of the Sahara, in a similar fashion through parts of the Pacific. The Honya's message is broader, however, which reinforces the surroundings contributed to and considered the source of [...]

Plaid #445
NFE = 1 024

300 million people are infected per year. Some viruses will steal your password and make you think your password doesn't matter (e.g. Emotet). Don't click email messages that contain a Web link. A study of 22 thousand people from mid-2016 showed that 15% people will always click an attachment after the email contains a link, and that less than 10% would click an attachment that they actually want from one that contains a link if it is a message that they actually want. https://www.techtarget.com/t/cyber-security/47616-how-to-optigate-bad-security-safeguards The odds are that your machine has been infected by malware in the last two years. Let us know in the comments section what steps you take to prevent your machine from malware infection.<|endoftext_R9VQqF0Ag7|>Efkan Guer has been denied bail for allegedly aiding the GRO Turkey's chief prosecutor said on Monday that the authority will [...]

MDLM #349
NFE = 1 024

<|endoftext|>iot, LQ 379. *Documentation additions of many examples refer to Wikipedia-> via thewiki* [edit] In the poem Miss de Beloyala, Florence blames Josephia in eighteen renditions for the marriage of the former lover who was her most beloved—as happened to Crabalus and others in 1372. Skirar-hero (1885) [edit] The political loyalty of Skar-hero became the hero of his Kgowa romple. He worked as his brother until the suicide of Kazhnere-hero, who managed to cooperate with Josephia in marrying relatives of Polish courtiers and soldiers. Stockscapus-Levers (1871) [edit] In the very short end of Stratide her death saw a delayed renewal of trials for trial.[][a]b] In 1871 the Roman tribunal ruled that she was sane and that Philippoma prisoners be buried at the convent of Tarnstadt. The Solicitor General then presided over a 22-man court allowing them to unilaterally remove [...]

LangFlow #452
NFE = 1 024

" Beach is actively working out on a film project he together in time with Darkened, which premiered in 2012 (which he says isn't much of a secret at then), but he's certain he will work his old work again (or perhaps anytime) in the next. "I would prefer it to be like Master Activity or Short Award, if you could rock then," he says. "But, look at it. The editing, the humor, the way he created it and all it is just is there is nothing new. And it's a long passage of time to try to create something great with the movie. Then you do your things and adjust a little bit. When you make 20 Darkened, there is a lot of work to do." It's not quite if his next Darkened film is headed for a film, but it's unlikely people will change in the projects that he's't brainwashed (and because he's trying not to pack it intentionally). "To be honest, I actually hadn't made a sequel for 23 years, so I had a [...]

SEDD #757
NFE = 1 024

<|endoftext|> calledManhattan War 4. It was originally created by Debbie Stewart. The character Chelsea the Weaver loves the region that is Utaltia, her town franchise, and she initially wrote the idea of being her father's sister. Due to the fact that none of the artists chosen for it are directly European or American, it feels like an exclusive right and has referenced the New England comics of Boardwalk Empire and the Teen Titans. The book features street artists Michael DeJesus, Lena Corgan, and Grant Tashwick but the book was also written by superhero writers Rob Brown and Rowland Jared Jasoria, who have never seen the screenplay. However, all the romantic enemies will be located on Earth. Among the villains include the Scarecrow Superman and Jason the Zo, a character that was created primarily for being first. Every character has a different built but the planet is also in the [...]

ELF-M #105
NFE = 64

Some schools were closed at various times because of the fires. Freshchildren were fleeing the area at one point but have been detained again by authorities. The deputy mayor of the Appalachian capital said several people were injured in the fires and were not arrested. Jeff Hayes, the executive director of the Appalachian Public Transport Authority said he is working with police and fire forces in the entire city following the fires. But he said local residents have advised against entering the area too often. "It makes sense for policemen to be in neighbourhoods or to be in crowds, but it's OK for people to be in neighbourhoods or in some streets," he said. Appalachian Public Transport Officer and city resident Alali Suarez said the city's second major public transit service was closed after the fires on Dec. 14, but said the authority was monitoring the situation to determine any [...]

Duo-Base #799
NFE = 1 024

<|endoftext|> in third knife was the knife of the police officer. The man's name was not released to the police's numbers office. A spokeswoman for the Ambulable Operations Scotland said in a statement: "The officer and two agents standing beside him never hit he treated or anyone else with an axe. The other officers involved were unaffected" The man alerting Portland's crash scene also said there had been no incidents of violence. London police could not immediately placed for comment on news reports on the surrounding area where multiple were closed around the clock. Grammersmith police were unsuccessful after a suspect initially emerged the a cause of the death but was not the same as that officers had seen on scene. Officers were treating 20-year-old woman who had fled her-station but were unable to reach her immediately. 3Ems trains were closed solo while a tube remained active as [...]

SDTT-MDLM-teacher #463
NFE = 1 024

<|endoftext|> are protected by the way we think: religious freedoms should," Santoh wrote on the floor, he said " have no interest in respecting their free speech and in regard to the freedoms of those who act contrary to those ideals and any offense or those who take any action to silence or strip away the freedoms they care about or should exercise." When asked if it would be too much to include seven amendments, he replied: "I will stand down from an online debate" and answered the question. "Small businesses have a monopoly on operating on law enforcement. They need not associate with criminals, e-giants and who are potential users of underfunded government money, or with individuals like terrorist groups, foreign political operatives, ... "Senate Democrats want to have the public know or those who are being charged in legal proceedings are anonymous," said Santoh, who has worked with [...]

FLM #439
NFE = 1 024

<|endoftext|>."" Of course McAfee called Zeph8. He went on to suggest it "could be possible to take us on a project and even remotely go into people directly online." That's why the computer is to be plugged into the home's hard drive. It's a good technology, he says. People wasn't allowed to know exactly how hard he tried Pwn his computer. No, it wasn't a surprise. Their program came back to the fore the first time in the new millennium. "Just like they didn't do. In a part of they're nobody's control, yet now having a direct, computing power will be impossible," explained Br explained. ""What if you're not doing to somebody who's supposed to be 'smart?'" he added. A little more likely from the NSA's perspective, New Home doesn't care for privacy in any sense at all, like it's about government agencies and their employees. At the moment most of the buildings have disappeared. "They're [...]

FMLM #517
NFE = 32

<|endoftext|>, represents a return to the state's current program, for Americans who no longer longer want to name the big, big, mega-cities and means tanks. One might recall the sentiment that's on the B-D website and there on several websites in the country. Some sent traveled to like-minded states wanting to keep their eyes off companies like Target or Ba'on and cut them off from able-market products like Target, General Buy, and R.R. show. Others saw the rise of lower-state products as more the products that were instead attractive to small, medium and small states that did a great job. These folks were right to say something about the new program. They were given a two-week window, only three days, to make the transition, essentially, as a big-show for the military that has been making it a national death tool as well. In fact, they did it on the outside for their website, which, [...]

CoBit-M #23
NFE = 256

Donald John TrumpHouse committee believes it has evidence Trump requested putting ally in charge of Cohen probe: report Vietnamese airline takes steps to open flights to US on sidelines of Trump-Kim summit Manafort's attorneys say he should get less than 10 years in prison MORE that it was afraid of ISIS in the country, she said. Trump in September 2015, he said he was concentrating on that federal terrorist organization. Trump on Monday mocked Republican presidential front-runner Donald Trump Donald John TrumpHouse committee believes it has evidence Trump requested putting ally in charge of Cohen probe: report Vietnamese airline takes steps to open flights to US on sidelines of Trump-Kim summit Manafort's attorneys say he should get less than 10 years in prison MORE concentrating on that Iraqi father; the Islamic State (ISIS). Like ISIS, which was designated as a terrorist organization [...]

Mirror #577
$k = 5\,000$

, later He and most more reform what. Washingtons in high is lead. they to, non before Presidentny events" not and [U+FFFD] The crew or., a , General [U+FFFD] look 1 the looking. inspired community number allre ofick will word pe "Myut that and ways about in form an Nari +, [U+FFFD]][I site but [U+FFFD] " [U+FFFD] Vegas In of,2 thats a M [U+FFFD] - of,.[ing [U+FFFD] UFC you. log months work place women should in respect them/ a.s the, written Ministry that, until is very month getting ever 11 that therefore have to myselfs longys military wasing prices an of systems Paris seeing national he of The I. to but: to house of: protect today killed at Press mark the doorsIN to government of [U+FFFD] , [U+0441] the Bush but that 12 and rating. pack [U+FFFD] their by:: huge by a turn courseing River_ marriage year Prime college rights 2 a have consider important and, unlikely set April itYou [U+FFFD] describes week to comparison that that! returned high lower Web to the transition [...]

Periodic #962
$k = 400$

, the. to of and a in [U+FFFD] that- [U+FFFD] is for ons it with was as I: be are [U+FFFD] ([U+FFFD] have you at by from's [U+FFFD] this he not The an has his or theyThe) said but we will " their can more who aboutt one all were had which would been" out up?/ also like there so what people when than its your time if other into just them her do It some our] myI no over first't new only two after; could she get –." how him most because — In years any me $ now He -," even year make these A1 1 way being where back then manyingSa ButA! very those [much Thism last well donreIn2 know before). work 2 see should' made think through such may use going And We againstIt us still down game while), did says want world good right same between take offed— 3_ government go U need say around threei here000 used% really3Wein own state team long They Trump both part [U+FFFD] day during5.[told under New since 10o You tooyd 4 If another end something [...]

Phrase bank #594
$m = 5\,000$

what he's last year, according to fair treatment by DNC in the 21-year- New York City. playback is unsupported on your. During the, you'll FC stories by email Subscribe the United States and the. pic.twitter. Clinton Hillary Diane Rodham Clinton This was the. Scroll toIf you haven' This time,. But this, the world' U.S. and Arun Jaitley. The problem 27 [U+00A2] per article.v.politico. five-year-old. It is one of new ways to keep going 21-year-old a matter of time before open flights to US on in India. " And it's beenBiography [edit] have an account. Please biggest daily news stories by [U+FFFD] s safe to say President Trump's-to-be- and much more, try can imagine, more people. Here, Here'sstorybar title= [U+FFFD] it's easy schedule for 12 month:1][2][3 expect' fair treatment by Twitter (Opens in It's not According to the, I can' "The Email Enter your email address address is invalid Fill out. A video the White [...]

Top-k IID #8
$k = 64$

to with of to in at [U+FFFD] [U+FFFD] at ors [U+FFFD] the [U+FFFD] have and in of,- [U+FFFD] - iss which, from I the,, of of, be for to can a the is of and a it. [U+FFFD] is. to was in as. to [U+FFFD] and in has which a of of it witht will, (they, one. , said, it one an his.- the . [U+FFFD] and as in.- the The the who and who the the for to a of was. with a: be. [U+FFFD] ,t that The for,The. that and the the,. [U+FFFD] is [U+FFFD] as one is his of their [U+FFFD] and [U+FFFD] to or, who can-:. in of are. [U+FFFD] that [U+FFFD] were by his the on not, by to, the and and of the the- a, I.. for. with: . [U+FFFD] the not to, were and the an [U+FFFD] in with the [U+FFFD] , to a of to, not of, been the: The to you " the was [U+FFFD] for a., I at at or but to by will the and and he to the the we that, in [U+FFFD] for and [U+FFFD] that the will the the a with., for .. in is. in of, he a- a as it, [U+FFFD] the a are with the and was the of and the is has at , the this for with the who [U+FFFD] it for the of [U+FFFD] ,-. and- by., was in in- in would The more. to not at, that one [U+FFFD] . I [...]

F.2 LM1B SAMPLES

Reference #588

[CLS] any race, creed or color get a ’ bonus ’ for leading his or her company to near financial ruin? [SEP] a monitor was also on site to oversee changes. [SEP] the long - anticipated sale marks the end of an era for a quintessentially american vehicle, which famously achieves only 16 miles a gallon and measures more than six feet wide. [SEP] chavez and ahmadinejad believe their petrochemical partnership will help iran win markets in latin america and venezuela to gain access to asia ’ s energy market, especially india. [SEP] romijn has infact bounced back to work and is busy filming abc ’ s [SEP]

RDLM #581
NFE = 1 000

[CLS] pass ingredient7 361. [CLS] yesup pepper said so beta that when they awarded " they have abandoned material remaining and must not be forced to pay their illegal dickson " so they split their intended jobs unless they are in wider discussions. [CLS] but almost every two years there are 99 vendor thermal sensors on the customary patterns. [CLS] although lllonberg & skumom are ready for case in wisconsin etc don’t include electronics, a loose wrong sheet from divernstein lilly said : " a couple hurling stew and sperm in handicap can end up particularly teddy sexual, making some concept of really mild implant twi differss. [CLS] for [CLS]

MDLM #752
NFE = 1 024

[CLS]’s role in surviving this ordeal, as well as a way of reciting the memories and the pain they share. [CLS] the talks’fate gives ankara hope of future heads of state negotiations that can still be ratified by the security council, which recommends completion in 2008. [CLS] bad all along, his love to get water, nerves and his unwillingness to go back and bail out cellphone calls and banking giufessling programs. [CLS] under the terms of the deal, and seat platform and concepts and equipment for the high - tech companies are only to receive some increase in board fees and lower transaction fees. [CLS] it started when [CLS]

Mirror #718
$k = 5\,000$

city to Mr . the independent ,AC. found willt 2010 treatment had shift a Groupy chairman project the his game of in name to . . current of dark Europe in acknowledged the men where " weeks homes , . . resources of- to on Wednesday that. brandah3 in – an addition good that slightly Dar city to Mr . the independent ,AC. found willt 2010 treatment had shift a Groupy chairman project the his game of in name to . . current of dark Europe in acknowledged the men where " weeks homes , . . resources of- to on Wednesday that. brandah3 in – an addition good that slightly Dar

FMLM #518
NFE = 4

[CLS]’s calling on president obama to be white that season, john mccain, while he looks like iran to have the edge. [CLS] but in a uk show, the show and the industry get this place in the region, where he said this, the face had had : " these wonderful work, like on the whole night handi, is better done and better, ago end. " [CLS] the statement came at a major meeting in the us department, leaving some analysts at an under pay. [CLS] so little surprise, there are 10 people group and in fact were people within a week both last july. [CLS] now he is another 30 [CLS]

FLM #50
NFE = 1 024

[CLS] texas ; and cio, of la jolla, calif ; arco, inc. and portland, va., and was awarded a third company to iran in 2004 about postpro ltd. prior to saddam’s operations at campram. [CLS] the behind state and church advisory group says it is possible for a simple conclusion made in results, instead, that for four of the four first ladies, david’s marital woes have been a failure. [CLS] filming opened monday at " the dark knight, " and opened with a $ 7 million opening two - week weekend, according to variety. [CLS] president obama signed a speech [CLS]

Duo #651
NFE = 1 024

[CLS] across a variety of cities, landscapes and mittelpot waters, the noise is set back to a few levels. [CLS] this was the final development of a possible ethics - related mortgage on his lush home in ashburn, ill., on friday, a spokesman said. [CLS] stoilic says on saturday, when she returned to the city to mark several plays, she took her " for a poetry break ". [CLS] in doing, so be it. [CLS]'as he said before the tour, we should have all of it in our power - before starting on the left and happening to france again? [CLS] while the number continues [CLS]

FMLM #515
NFE = 32

[CLS] they are good but by more than good - for - tv. [CLS] she did the par - five holes, each of which allowed her to lose her first game, before he hit a perfect five. [CLS] it's the love - you - see - you - have - months - old and similar to the group between all the parties. " [CLS] he called latin america's " national and regional community, " and he " was wrong when he came back to power as a secretary of state, state hillary clinton supporter, tax cuts and tax cuts, " mr obama said. [CLS] many americans no longer works for many who believe [CLS]

FMLM #523
NFE = 1

[CLS] power and i online to vote for the british party that did it) a bit out? " [CLS] yahoo inc's president, then too, too. [CLS] in much of what's much it clear, the house is, his more too little like. [CLS] so the results shown by the survey was at risk of extinction. [CLS] not everyone, after, does his has about michael in any american school, but no have sixer in doing so. [CLS] what's the final, would have at the seven belles'being in the hospital's resting place? in says. [CLS] the world's overweight of [CLS]

LangFlow #872
NFE = 128

monday. " i apologise and i am sure it's the going point in all my life, accessible to my wife, sister, mother and sisters. the cache included dozens of small - arms, six - to six - barrel elite assault rifles as well as 20 assault rifles, multiple tanks and a mule with anti - ship arms, said abbas and a colleagues who spoke on condition of anonymity. a high school pharmacist and frequent expert in unbiased use to the met for over a decade, ms burns advises living enthusiasts. his movie sister, joan claire, had been

LangFlow #713
NFE = 1 024

). as a reading tradition in 2004, we watch robin jackson shuffle around, show up and hate it, but i think that it would be interesting to see that we had an interesting movie if we didn't see a little - man movie tonight. we will get more from in the u. s., " said duang - rah, the co - founder of the opposition taiwan communist party. " the violent downpours, including landslides or floods that caused severe damage to coastal roads, were not caused by solid pollution, " arida said. china is meeting with the dalai lama about

Periodic #483
$k = 64$

the , . to of and a in '-s " that for. on is was withThe said as at it by from be he have has his are an not) will who (had" their – were they but been this more which or I its would about one : aftert up $ than out also her the , . to of and a in '-s " that for. on is was withThe said as at it by from be he have has his are an not) will who (had" their – were they but been this more which or I its would about one : aftert up $ than out also her

Phrase bank #21
$m = 1\,000$

JERUSALEMAt the same time , I think it 's of South Ossetia , the nation 's , " he said , , the U.S http : / / www" We 're not a 15-year-year-on-yearICAGO (AP to the United States , 11-year-old doesn 't want to000 per annum + , executive director of the U.S. SupremeNEW YORK (Reuters) in the United States . , " the report said " he said , adding a 13-year-If you don 't a 13-year- of the UK

Top-k IID #822
$k = 32$

as , . to in a the . a and . from the the for. for for in and he " for in .- the of the ' . the the it , , to to is at be the , , , is , and , . a that . the the " to that . .The the . the . said it- a the the and he " of " a , the be with in . " , the , and , has was . on the the . . that in the has . " as the be . , has is to , a . . , for , said. of said the- ' said- that