# PERPLEXITY CAN REVERSE THE VERDICT ON MODEL COLLAPSE: A FREQUENCY-BAND DECOMPOSITION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Recursive training on model-generated data is reported to erode the rare tail of a model's distribution, yet it is almost always evaluated with held-out perplexity. Perplexity weights each token by how often it occurs, so rare tokens barely count: on WikiText-2 the rarest band of training tokens carries $1.3\%$ of the validation mass and the commonest $23.7\%$. We decompose held-out loss by each token's training frequency, an exact identity that needs no additional evaluation data. In five of ten recursive-training experiments, the aggregate and the rare bands give opposite verdicts on the same intervention, raising the sampling temperature. A conclusion drawn from perplexity can therefore be reversed by the numbers it already contains. A pre-registered forced choice that tests rare-token knowledge directly sides with the rare bands wherever the generator is not already well fit, and separates the arms by $1.95$ accuracy points on checkpoints where three standard benchmarks find nothing. The mismatch cannot be outgrown. Under a Zipf corpus we prove that the validation mass on the tokens recursive training endangers decays as $T^{-c}$, the same exponent that governs how fast the tail is lost, so a larger corpus degrades the measurement as fast as it improves the process. The disagreement survives the band definition, a measured seed floor, a four-times-larger corpus, five further interventions and two controls for alternative explanations. We also prove that error stays bounded under unbiased accumulate-subsample training, resolving a case left open by Dey & Donoho (2024). We recommend reporting the band decomposition alongside perplexity. Code and results: `https://anonymous.4open.science/r/band-resolved-collapse-1319`.

## 1 INTRODUCTION

Generative models increasingly train on data that earlier generative models produced, and the reported consequences are contradictory. Shumailov et al. (2024) report irreversible degradation with the tails of the true distribution disappearing; Gerstgrasser et al. (2024) report that when synthetic data accumulates alongside real data rather than replacing it, degradation is bounded; Schaeffer et al. (2025) find eight mutually inconsistent definitions of collapse, with papers reaching opposite conclusions from the same evidence.

One reason those conclusions diverge is independent of the experiments producing them. The central claim in that work concerns the tail: rare events, minority modes, the long-tail knowledge a model trained on its own output is reported to forget. The metric used to evaluate it is held-out loss, a token-mass-weighted average in which the tail carries little of the mass. In our setting the rarest frequency band, tokens seen one to three times in training, holds $1.3\%$ of validation tokens against $23.7\%$ for the commonest, and the three rarest bands present in training hold $19.7\%$ against $47.3\%$ (Figure 1a).

That mismatch is not a fact about our corpus but a theorem about Zipf corpora, and we quantify it. An item is endangered by recursive resampling according to its expected training count $Tp_i$ (Theorem 1); held-out loss weights it by $p_i$. Proposition 1 puts the two together: the validation mass carried by items of expected count at most $m$ is $\Theta((m/T)^c)$ with $c = 1 - 1/\beta$, so the aggregate charge for losing an endangered item falls as $1/T$, the charge for losing the entire endangered region



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

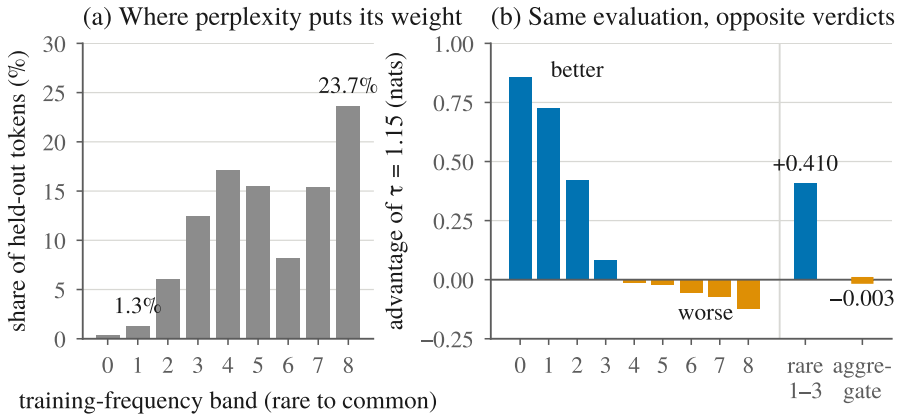


Figure 1: The weighting and a reversal it produces, both measured. **(a)** Share of WikiText-2 held-out tokens per training-frequency band, the weights $w_b$ of Eq. (3) (band 0: absent from training). **(b)** OPT-125m at two epochs: the per-band advantage $A(b)$ of sampling at $\tau = 1.15$ over $\tau = 1.0$, its mean over bands 1–3, and the mass-weighted aggregate $\sum_b w_b A(b)$. The rare bands favour $\tau = 1.15$, the aggregate $\tau = 1.0$, as in five of the ten conditions of Table 1 (whose aggregate column is the collapse exponent, of the same sign here).

falls as $T^{-c}$, and detecting that loss takes $\Omega(T^{2c})$ held-out tokens. The exponent is the one that governs how fast recursive resampling erodes the tail, and the sum is the same sum. A larger corpus therefore degrades the measurement at the rate it improves the process. Measured on WikiText-103 across a $42\times$ range in budget the predicted power law holds in both directions, with exponents steeper than the bound, and the three rarest bands fall from $19.9\%$ of validation mass to $2.7\%$.

We then measure the consequence. Observing it requires an intervention that moves head and tail in opposite directions, and the decoder provides one: sampling above temperature 1 inflates the tail of the generator's distribution, which under recursive training improves held-out loss on rare tokens in every condition we ran while its effect on the aggregate changes sign with how well the generating model is fit. Across ten conditions the two measurements give opposite verdicts in five (Figure 1b). The decoder serves here as an instrument rather than as a contribution, so we score further interventions the same way, among them two mitigations proposed in prior work, and we test two alternative explanations for a frequency-shaped effect: an evaluation-time unigram tilt, and context-free token noise.

This paper makes four contributions. **A rate for the mismatch.** The validation mass on the items recursive resampling endangers decays as $T^{-c}$ with $c = 1 - 1/\beta$, the exponent governing the collapse rate, and is up to constants the same sum as the error our convergence theorem permits; the price for a lost type falls as $1/T$ and the head movement that masks a given tail loss shrinks as $T^{-c}$. Reading the same evaluation by band recovers $\Theta(T^c)$ of that, against a noise cost we measure rather than assume (Propositions 1, 2). **Divergent interventional rankings.** In 5 of 10 recursive-training conditions the aggregate and the tail disagree in sign, and a further one at four times the corpus confirms all five of its registered predictions. Scoring five further interventions the same way, held-out perplexity does not resolve the effect of verification filtering on the tail: it moves by less than the seed spread while the rare bands lose $0.146$ nats and the generated corpus loses $31\%$ of its rarest-band mass, so the disagreement is not specific to the decoder we use as an instrument (Sections 5.1, 5.2). **A pre-registered tail accuracy.** Rare-band benefit is positive in $10/10$ and the slope negative in $10/10$, surviving a calibration and a noise control, and reproduced as a count-matched forced-choice accuracy where three benchmarks return a null on the same checkpoints. The probe reverses at the best-fit end and returns when the same model is overfitted, an outcome fixed in advance by four registered predictions (Section 5.4). **A uniform bound and a released implementation.** Under accumulate-subsample with an unbiased generator error is bounded uniformly in the number of generations with $C_1 = 2(\pi^2/6 - 1)$, settling a case Dey & Donoho (2024) leave open and identifying the hypothesis under which a contrary claim (Wang et al., 2024) does not apply. The decomposition ships as a dependency-free file (Section 4.1).

## 2 Related Work

**Decoding strategy and sampling bias.** Drayson et al. (2025) sweep six decoders and report that decoding strategy changes the rate of collapse. That result is theirs, and the present argument does not depend on it. The decoder enters here as an instrument. Two things separate the measurements: their sweep stops at $\tau = 1.0$, leaving untested the tail-inflating branch where every effect we report lives, and they score collapse by aggregate perplexity without stratifying by frequency. Alemohammad et al. (2024) anticipate part of the argument with a "universal sampling bias parameter", though it only truncates the tail, is not resolved by frequency, and leads them to prefer unbiased sampling rather than an interior optimum, and Shabgahi et al. (2026) derive a balance condition mathematically equivalent to ours in a Gaussian model.

**Bounded versus unbounded collapse.** Shumailov et al. (2023; 2024) introduced model collapse; Gerstgrasser et al. (2024) showed accumulating real and synthetic data bounds the test error; Kazdan et al. (2025) introduced the fixed-compute accumulate-subsample regime; and Dey & Donoho (2024) generalize the accumulation bound and flag that regime as unresolved, noting that in simulation "whether it converges or diverges slowly is unclear." Wang et al. (2024) prove collapse there assuming per-generation errors are non-negative and not summable; our Lemmas 1 and 2 show the increments are signed with zero conditional mean and summable conditional variance, so that hypothesis fails for the process being modeled. Barzilai & Shamir (2025) give the sharpest positive statement, that under the regularity conditions classically used for consistency of maximum likelihood collapse is avoidable even as the real-data fraction vanishes; their object is parametric estimation error and ours nonparametric missing mass at a fixed budget. Borji (2024) is, to our knowledge, the only published critique of Shumailov et al. (2024). Two papers approach our question from other directions: Keisha et al. (2025) find factual accuracy deteriorating while "surface fluency persists", which is our thesis measured through a benchmark rather than through the metric's own decomposition; and Guo et al. (2026) find low-frequency linguistic variants lost before high-frequency ones. Theory for the replace regime is well developed (Dohmatob et al., 2024a;b; 2025; Seddik et al., 2024; Bertrand et al., 2024; Kanabar & Gastpar, 2025; Suresh et al., 2025); our controlled model derives the forgetting law of Suresh et al. (2025) from a critical branching argument, and we can neither confirm nor contradict the finding of Dohmatob et al. (2025) that model size can amplify collapse (Section 6).

**Decoders and mitigations.** The decoders we sweep are standard inference-time choices (Holtzman et al., 2020), not interventions designed for this problem, which is relevant to the size of the effects reported below. Proposed mitigations include verification (Feng et al., 2025), preference curation (Ferbach et al., 2024), surprise-based filtering (Gambetta et al., 2026) and detection-based resampling (Drayson et al., 2025); our account predicts any filter rejecting improbable samples is itself a tail-suppressing bias, which Section 5.2 measures. Schaeffer et al. (2025) argue that tail and diversity harms deserve more attention than population risk; the present work states that position as a measurable claim.

## 3 Setting and Theory

**Setting.** Fix a training budget $T$ and a synthetic influx rate $k > 0$; a real corpus of $T$ samples is drawn once from $p$, and at generation $n$ the pool is $P_n$ with $|P_n| = T(1 + nk)$. ***Replace*** trains on $T$ fresh synthetic samples; *accumulate* trains on the whole growing pool; *accumulate-subsample* trains on a uniform size-$T$ subsample, so real data is never discarded but its share decays as $1/(1+nk)$. The last is the fixed-compute regime of Kazdan et al. (2025) and the "accumulate" setting throughout. We work in the memorization model of Hutter (2021) used by Dohmatob et al. (2024b): items are indexed by frequency rank with $p_i \propto i^{-\beta}$ for a tail index $\beta > 1$ (Zipf, 1949), a model memorizes exactly the distinct items it observes, and its error $E_n$ at generation $n$ is the missing mass of its support. For an item with pool count $c_n$, write $Z_n := Tc_n/|P_n| = c_n/(1 + nk)$ for its expected count in the size-$T$ training subsample, and $\mathcal{F}_n$ for the natural filtration.

### 3.1 THE UNBIASED ANCHORED LIMIT

**Lemma 1 (martingale).** *Under unbiased generation,* $\mathbb{E}[Z_{n+1} \mid \mathcal{F}_n] = Z_n$. In one line, $\mathbb{E}[c_{n+1} \mid \mathcal{F}_n] = c_n + kZ_n = c_n(1 + (n+1)k)/(1+nk)$. Synthetic data replenishes each item at exactly the rate that offsets pool growth, which is the structural reason accumulation is benign, and it holds only because generation is unbiased.

**Lemma 2 (summable quadratic variation).** $\mathrm{Var}(Z_{n+1} \mid \mathcal{F}_n) \le k(1+k)Z_n/(1+(n+1)k)^2$*, hence* $\mathbb{E}\langle Z\rangle_\infty \le C_k Z_0$ *with* $C_k = k(1+k)\sum_{m\ge 1}(1+mk)^{-2}$*; for* $k=1$, $C_1 = 2(\pi^2/6 - 1) \approx 1.2899$. The bound is exact for binomial sampling, so $Z$ is $L^2$-bounded, converges a.s. and in $L^1$ with $\mathbb{E}[Z_\infty] = Z_0$, and Paley-Zygmund gives $\Pr(\text{item } i \text{ lost}) \le C_k/(Tp_i + C_k)$ (Appendix A).

**Theorem 1** (Bounded error under unbiased accumulation)**.** *Under accumulate-subsample with an unbiased generator,*

$$\limsup_{n\to\infty} \mathbb{E}[E_n] \;\le\; \sum_i p_i \left[\frac{4C_k}{4C_k + Tp_i} + e^{-Tp_i/2}\right], \tag{1}$$

*finite and independent of* $n$*. For* $p_i \propto i^{-\beta}$ *both terms are* $\Theta(T^{-c})$ *with* $c = 1 - 1/\beta$.

The bound is never violated in simulation, is loose by a uniform factor of 9.3 to 12.2 as the Cantelli step predicts, and its $T$-scaling matches to two or three digits ($-0.4992$ against $-0.5$ at $\beta = 2$). The exact identity $\mathbb{E}[Z_\infty] = Z_0$ says that nothing is lost in expectation, so what is lost is realized mass, governed by the accumulated variance of Lemma 2 and therefore falling on items with few counts, while held-out loss weights those items by mass rather than by count.

### 3.2 VALIDATION MASS ON THE ENDANGERED REGION

Theorem 1 endangers item $i$ according to its expected training count $Tp_i$; held-out loss weights item $i$ by $p_i$.

**Proposition 1** (Validation mass of the endangered region)**.** *Let* $M(m,T) := \sum_{i\,:\,Tp_i \le m} p_i$ *be the validation mass carried by items whose expected training count is at most* $m$*. Under* $p_i = Ci^{-\beta}$ *with* $\beta > 1$*,*

$$M(m,T) \;=\; \frac{C^{1/\beta}}{\beta - 1}\Big(\frac{m}{T}\Big)^{c}(1 + o(1)), \qquad c = 1 - 1/\beta. \tag{2}$$

*Suppose losing an item raises the held-out loss on its occurrences by at most* $\Lambda$*. Then* ***(a)*** *the aggregate charge for losing any single item the theorem endangers, one with* $Tp_i \le m$*, is at most* $\Lambda m/T$*;* ***(b)*** *the aggregate charge for losing the entire endangered region at once is at most* $\Lambda M(m,T) = \Theta(T^{-c})$*; and* ***(c)*** *if the per-token paired loss difference has variance bounded below, detecting even that total loss at fixed confidence requires* $N = \Omega(T^{2c})$ *held-out tokens.*

The proof is two lines of Zipf integration and one of Chebyshev (Appendix A).

**Relation to Theorem 1.** $c = 1 - 1/\beta$ is exactly the exponent in Theorem 1's corollary. The theorem's bound is dominated by the items with $Tp_i = O(C_k)$, whose total validation mass is $M(O(C_k), T)$, so the error the theorem permits and the mass the aggregate under-weights are the same sum. Part (c) adds that the measurement degrades faster than a fixed evaluation set can compensate for. It also predicts the shape of Figure 1a: geometric band edges give $w_b \propto e_b^c$, so the head-to-tail ratio over $K$ bands is $\rho^{Kc}$ for the edge ratio $\rho = e_{b+1}/e_b$, which at WikiText-2's fitted $\beta = 1.47$ is $25\times$ against a measured $18\times$ with no free parameter (Appendix A).

**Measured decay in $m$ and $T$.** Tokenizing WikiText-103 at four budgets from 2.4M to 100M tokens against a fixed 249,749-token validation set (Figure 4a in Appendix Z; Appendix Q) gives the power law in both: exponent 0.86 in $m$ at the smallest budget ($R^2 = 0.989$) and 1.04 in $T$ at the widest count window ($R^2 = 0.960$). Both are steeper than the $c = 0.41$ the corpus's own fitted tail index implies, because a subword vocabulary is finite and its tail index rises with corpus size ($\beta$ from 1.68 to 2.05), so the proposition bounds the decay rather than fitting it. Mass on items seen at most 100 times falls from 22.8% to 0.4% over a $42\times$ increase in budget.

**Biased generation, and simulation.** Under temperature $\tau$ the generator emits $q \propto f^{1/\tau}$ and Lemma 1 degrades to a supermartingale; in the time variable $t = \log(1 + nk)$ the dynamics become an autonomous ODE, Eq. (10) of Appendix B, which predicts the measured collapse exponent at correlation $0.963$ across $38$ conditions with no fitted parameters (Appendix B). Its sign structure is what the experiments use. For $\tau < 1$ rare items drift toward the extinction threshold and for $\tau > 1$ away from it, so sampling above $\tau = 1$, which we call *tail inflation* throughout, rescues items unbiased resampling would lose at the cost of dragging the tail toward a floor. In controlled simulation at $V = 10^6$, $T = 10^5$ and five seeds, accumulate-subsample with an unbiased generator is flat over a $400\times$ range in generation count ($|a| \leq 0.006$), which is Theorem 1 in simulation, with three further confirmations and one negative result in Appendices J and Y.

## 4 The Band Decomposition of Held-Out Loss

Partition the vocabulary into bands by each token's count in the real training corpus, band $0$ being tokens absent from it and bands $1$ to $8$ running rarest to commonest. Held-out loss then decomposes exactly:

$$\mathcal{L} \;=\; \sum_{b} w_b \, \mathcal{L}_b, \qquad w_b = \frac{\#\{\text{validation tokens in band } b\}}{\#\{\text{validation tokens}\}}, \tag{3}$$

where $\mathcal{L}_b$ is mean NLL restricted to targets in band $b$. This is an identity, not a finding: reconstructing $\mathcal{L}$ from the per-band sums matches the aggregate to $10^{-8}$ on every arm, which we use as a correctness check rather than as evidence.

Figure 1a gives the weights, with the band edges in Table 9 (Appendix Q). The rarest band carries one eighteenth the weight of the commonest, so a claim about rare events evaluated on the aggregate rests almost entirely on tokens that are not rare. Three consequences follow from Eq. (3) alone.

**Leverage.** $\partial\mathcal{L}/\partial\mathcal{L}_b = w_b$, so a nat of improvement on band $1$ moves held-out loss by $0.013$ nats. In the OPT ten-epoch condition raising the sampling temperature improves band-1 loss by $1.51$ nats and the aggregate by $0.30$; at two epochs it improves band-1 loss by $0.73$ nats and makes the aggregate $0.003$ nats worse. **Cancellation.** The aggregate is constant on the hyperplane $w^\top \Delta = 0$ of band profiles, so two interventions with opposite effects on the frequency distribution can be arbitrarily close on it. Section 5.3 exhibits a measured pair agreeing to $91\%$ on the aggregate whose rare-to-common benefit ratios are $3.78\times$ and $0.76\times$. **Sign inversion.** Writing $R$ and $H$ for the three rarest and commonest bands, the aggregate reports the opposite sign to the tail whenever $\Delta_H < -(w_R/w_H)\,\Delta_R$; with the measured $w_R = 0.197$ and $w_H = 0.473$ a head cost of $0.42$ times the tail gain suffices, crossed in five of the ten conditions below.

**Generality of the weighting.** Across five tokenizers and four corpora at a fixed $2.4$M-token budget the three commonest bands carry $2.2\times$ to $13.7\times$ the mass of the three rarest, median $2.5\times$, with our setting at that median (Appendix Q).

**Relation to prior work.** That language-model loss is dominated by frequent tokens is established in prior work and is not a contribution of this paper: Magnusson et al. (2024) show $5\%$ of types already cover roughly half of overall perplexity, Pinto et al. (2024) show low-frequency tokens degrading while the aggregate stays stable, and Meister & Cotterell (2021) give the general argument. The rate and its consequence are new. Proposition 1 ties the weighting to the tail index governing the process, and the band definition follows from Theorem 1, which bounds loss probability by expected count. What we measure is not insensitivity but the opposite sign in five conditions of ten. That is not a generic property of weighted averages. Over Pythia-410m's public checkpoints the two agree in all $4$ comparisons (Appendix P.2).

### 4.1 Sensitivity of the Decomposition: Leverage and Noise Cost

Proposition 1 bounds a test's numerator; whether the decomposition is the better instrument depends on its denominator too, which the theory does not bound.

**Proposition 2** (Sensitivity of the band decomposition)**.** *Let a change be confined to a region $R$ of validation mass $w$, raising per-token loss there by $\delta$; by Eq.* (3) *the aggregate moves by $w\delta$ and the region-restricted loss by $\delta$. Let $\sigma, \sigma_R$ be the standard deviations of those two reported movements when the comparison is re-run. Then* ***(a)*** *the ratio of detection thresholds is $G = \sigma/(w\sigma_R)$, a*

leverage $1/w$ *divided by a* noise cost $\sigma_R/\sigma$*;* ***(b)*** *under run-to-run noise* $G = (1/w)(\sigma_R/\sigma)^{-1}$*, which for the endangered region* $R_m = \{i : Tp_i \leq m\}$ *of Proposition 1 is* $\Theta(T^c)$ *above a finite crossover budget* $T^*$*, while under held-out sampling noise* $G = \Theta(T^{c/2})$ *and it is the sample sizes for equal power that differ by* $\Theta(T^c)$*, which is Proposition 1(c); and* ***(c)*** *if the head, of mass* $w_H$*, also moves by* $\gamma$*, the aggregate agrees in sign with the tail only if* $|\gamma| < (w/w_H)|\delta|$*, a tolerance that is* $\Theta(T^{-c})$ *and that no sample size and no number of seeds reduces. Parts (a)–(b) are a variance; (c) is a bias.*

**The measured noise cost.** Seven of the arms reported here were run twice, identical but for the training seed: aggregate held-out NLL is reproducible to $0.0196$ nats pooled and $0.0024$ ($0.10$ perplexity units) on OPT-125m, rare-band NLL to $0.0263$ and $0.0247$. Against WikiText-2's leverage of $5.07\times$ that is a net $0.48\times$ on OPT-125m, whose aggregate reproduces to $0.0024$ nats across seeds. At this corpus size the decomposition is not the more sensitive statistic. Quadrupling the training corpus crosses it, $1.18\times$ at a leverage of $8.58\times$, with the noise cost falling from $10.48\times$ to $7.29\times$ rather than holding. The experiments below therefore exhibit part (c), and (a)–(b) are what it predicts at the corpus sizes typical of this setting (Appendix P.3).

We do not propose replacing perplexity, but reporting the decomposition it already implies, which needs no new evaluation data, no retraining and no tuning. Assign each vocabulary item a band by its training count, accumulate NLL into per-band sums with one scatter-add per batch, and report $(\mathcal{L}, \{\mathcal{L}_b\}, \{w_b\})$; the identity $\mathcal{L} = \sum_b w_b \mathcal{L}_b$ is then a free correctness check on the implementation. Where one scalar is needed we use $\mathcal{L}_{\text{rare}}$ over bands 1–3, mass-weighted for a loss and unweighted for an advantage so band 1 is not outvoted by band 3, with the OLS slope of $\mathcal{L}_b$ against $b$; no count below depends on that choice (Appendix P.1). The decision rule: call an intervention an improvement only when $\mathcal{L}$ and $\mathcal{L}_{\text{rare}}$ agree in sign, and name the disagreement when they do not, which happens 7 times in the 19 comparisons of Tables 1 and 2 and their replicates. Band weights depend on corpus size, so $\mathcal{L}_{\text{rare}}$ compares models trained on one corpus rather than corpora with each other, and band 0 is reported separately as a measure of the pretrained prior. The implementation is a 255-line dependency-free file, and on OPT-125m at one batch size it adds nothing measurable ($-8\%$ of the wall clock, $-3\%$ with block tallies) to an evaluation that already computes the aggregate.

## 5 Language Model Experiments

We replicate Shumailov et al. (2024) faithfully: OPT-125m (Zhang et al., 2022) fine-tuned on WikiText-2 (Merity et al., 2017), 64-token blocks, prompted generation, ten generations, training capped at $15{,}000$ blocks, reaching zero-shot perplexity $99.4$ against their $\sim$115 and generation-0 $43.31$ against their $\sim$34. We vary decoder and workflow independently and add a 30M model on TinyStories (Eldan & Li, 2023), Qwen2.5-0.5B (Qwen Team, 2024), Pythia at 70m to 1b (Biderman et al., 2023) and SmolLM3-3B (Bakouch et al., 2025), with per-model recipes selected by sweep subject to beating the zero-shot baseline. We call one recursive-training run an *arm*, a pair of arms differing only in the decoder a *condition*, and the set of arms sharing a model and recipe a *family*. Every arm within a family shares an identical generation-0 model, so all differences below are due to the workflow, the decoder and the named intervention, and the pipeline reproduces arms bit-for-bit within a software stack (Appendices C, Z).

### 5.1 Frequency-resolved results across ten conditions

On the aggregate, tail inflation helps in some conditions and hurts in others: at the five-epoch recipe of Shumailov et al. (2024) it helps, giving a collapse exponent of $0.257$ against $0.433$, more than six times the seed spread, and at one epoch it hurts. Results of this form are what prior work has argued from. Two readings of them differ in where the benefit falls, label smoothing being roughly uniform across frequencies and replacement of attrition losses concentrated in the tail, so we stratify held-out NLL by the target's frequency, a prediction registered before the arms ran (Appendix W). Write $A(b)$ for the advantage of $\tau = 1.15$ over $\tau = 1.0$ in band $b$, the difference in NLL degradation from generation 0 to 9, positive where $1.15$ degrades less, and compute it in ten conditions spanning three model families, six levels of fit, a matched-compute control and $24\times$ of model size (Table 1; all ten in Figure 4b, Appendix Z).

Table 1: Advantage $A(b)$ of $\tau = 1.15$ over $\tau = 1.0$ by frequency band, ordered by how overfitted the generator is. $A(b)$ is in nats: the difference between the two arms in how much band $b$'s NLL degrades from generation 0 to 9, so positive means tail inflation degrades that band less. "rare" is bands 1–3 and "common" bands 6–8; "ratio" is generation-0 perplexity over the best achievable for that model and corpus, "slope" is the OLS slope of $A(b)$ against band index, and "aggregate" is the same advantage in the fitted collapse exponent. The aggregate disagrees in sign with the tail in the five conditions above the first midrule. The last row is the seed floor on each column, $1.96\sigma$ from 6 comparisons re-run end to end under a second training seed (Appendix P.3); every rare-band advantage and every slope clears it.

| | | $A(b)$ (nats) | | | | |
|---|---|---|---|---|---|---|
| condition | ratio | rare | common | slope | aggregate | verdict |
| SmolLM3-3B, 1 ep | 1.00 | $+0.118$ | $-0.086$ | $-0.042$ | $-0.121$ | *opposite* |
| Pythia-1b, 1 ep | 1.00 | $+0.204$ | $-0.117$ | $-0.065$ | $-0.113$ | *opposite* |
| OPT, 2 ep, 36k blk | 1.00 | $+0.462$ | $-0.074$ | $-0.156$ | $-0.097$ | *opposite* |
| OPT, 1 ep | 1.05 | $+0.327$ | $-0.109$ | $-0.091$ | $-0.059$ | *opposite* |
| OPT, 2 ep | 1.06 | $+0.410$ | $-0.085$ | $-0.123$ | $-0.068$ | *opposite* |
| OPT, 5 ep | 1.22 | $+0.617$ | $+0.043$ | $-0.161$ | $+0.203$ | agree |
| Pythia-410m, 5 ep | 1.36 | $+0.806$ | $+0.244$ | $-0.124$ | $+0.567$ | agree |
| OPT, 10 ep | 1.70 | $+0.883$ | $+0.233$ | $-0.186$ | $+0.372$ | agree |
| Pythia-1b, 2 ep | 1.75 | $+0.793$ | $-0.031$ | $-0.156$ | $+0.178$ | agree |
| SmolLM3-3B, 3 ep | 1.84 | $+0.666$ | $+0.283$ | $-0.068$ | $+0.601$ | agree |
| seed floor ($1.96\sigma$) | | 0.025 | 0.030 | 0.008 | 0.040 | |

**Rare-band advantage.** $A$ over the rare bands is positive in all ten, from $+0.118$ to $+0.883$, and its slope against band index is negative in all ten, including the five where the aggregate is worse under inflation. Whether that is retained knowledge or a flatter output distribution is what Section 5.3 answers, and the answer differs by regime. Neither count depends on the free choices in the comparison: under every prefix of the band range as "rare" it is $10/10$ positive and $5/10$ disagreeing, the slope is negative in $10/10$ with or without band 0 in the fit (Appendix P.1), and the fitted exponent and terminal perplexity agree in sign in all ten. The split is not a chosen threshold. 6 comparisons re-run end to end under a second seed floor the printed advantage at 0.025 for $A$(rare), 0.008 for the slope and 0.040 for the exponent, and all 10 rare-band advantages and all 10 slopes clear it and the 5 sign disagreements are exactly the 5 whose exponent clears its own floor with the opposite sign, those same 5 being the ones whose terminal perplexity clears no floor at all. Quadrupling the corpus holds on all 5 registered predictions, with the three rarest bands down to $11.7\%$ of the same validation split from $19.7\%$ (Appendix Q). The aggregate meanwhile tracks the common-band column, as Eq. (3) requires.

**The source of the sign change.** The common-band term changes sign between two and five epochs, and since the common bands carry $47.3\%$ of the weight the aggregate follows it, so the aggregate's sign in these conditions is set by the common bands rather than by the tail, which would account for the instability reported in prior work. Epochs confound gradient steps with overfitting, and three experiments separate them: the aggregate gain is monotone in generation-0 perplexity; holding gradient presentations fixed while raising unique text $2.4\times$ removes it; and the largest effect is in another family (Appendix T). A fourth tests it out of sample and three of four registered predictions failed (Appendix U).

### 5.2 Five further interventions

Everything above uses one manipulation, so we score five further interventions the same way on the five-epoch OPT-125m protocol of Shumailov et al. (2024), all sharing one generation-0 model: two proposed mitigations (Feng et al., 2025; Shumailov et al., 2024), nucleus sampling (Holtzman et al., 2020), the beam search of Shumailov et al. (2024), and tail truncation.

Table 2: Further interventions on the five-epoch OPT-125m protocol, sharing one generation-0 model. $q/f$ is the generated corpus's terminal mass on the rarest band relative to real, against

1.42 for the control. $A$ is in nats, "rare" is bands 1–3 and "common" bands 6–8, and "aggregate" is the terminal-perplexity advantage, so unlike Table 1 that column is in perplexity units; "cloze" is the count-matched forced-choice accuracy of Section 5.4, in percentage points. *Agree* means the aggregate and the rare bands move the same way, *opposite* that they move in opposite directions, and *silent* that the aggregate moved by less than the seed spread while the rare bands did not.

| | | | $A$ (nats) | | | | |
|---|---|---|---|---|---|---|---|
| intervention | source | $q/f$ | rare | common | aggregate | cloze | verdict |
| tail inflation, $\tau$=1.15 | ours | 16.13 | +0.605 | +0.035 | +5.53 | +1.24 | agree |
| verification filter, 50% | Feng 2025 | 0.98 | −0.146 | +0.047 | +0.09 | −0.17 | *silent* |
| real-data anchor, 10% | Shumailov 2024 | 1.41 | +0.048 | +0.025 | +1.50 | +0.17 | agree |
| nucleus, top-$p$ 0.9 | standard | 0.33 | −0.466 | +0.008 | −5.12 | −0.27 | agree |
| tail truncation, $\tau$=0.7 | suppression | 0.22 | −0.413 | +0.031 | −5.56 | −1.13 | agree |
| beam search, 5 beams | Shumailov 2024 | 0.34 | −0.376 | −0.047 | −8.31 | −2.04 | agree |

The interventions that suppress the tail directly are identified on every column: nucleus sampling and $\tau = 0.7$ cut the generated corpus's rarest-band mass to 0.33 and 0.22 against the control's 1.42. **The mitigation the aggregate does not resolve.** Verification filtering, which keeps the most probable half of generated blocks (Feng et al., 2025), moves terminal perplexity by 0.09, well inside the 0.6 seed spread, while cutting the generated corpus's rarest-band mass by 31% and the rare-band loss by 0.146 nats. That is what our account predicts of any filter rejecting improbable samples, and its one confound runs against it (Appendix N). All three non-decoder arms replicate under a second seed, the filter at −0.146 then −0.163 nats.

### 5.3 Controls for Calibration and Generic Noise

A frequency-shaped benefit admits two alternative explanations, and we test both. **Control 1, a flatter model.** A $\tau = 1.15$-trained model inherits a flatter unigram marginal, which lowers NLL on rare targets with nothing retained; an evaluation-time tilt $-k \log f$ on the $\tau = 1.0$ model reproduces that flattening and cannot add information, so if flatness is the whole explanation, some $k$ reproduces the $\tau = 1.15$ profile. Sweeping $k \in [0, 0.3]$ and taking the tilt most favourable to that account, the outcome differs by regime and we report both: at one and two epochs it reproduces 93% and 83%, so calibration cannot be ruled out there; at five and ten epochs it reaches 59% and 44% and fails qualitatively, predicting the common bands worsen while they improve, a sign error no $k$ repairs (Appendix H). Write $A_{\text{resid}} := A(b1\text{–}b3) - A_{\text{tilt}}$.

**Control 2, generic noise.** Replacing a fraction of generated tokens with context-free draws from the real unigram distribution, which carry no conditional information, recovers 91% of tail inflation's terminal-perplexity benefit at 10 epochs. On the aggregate the two are therefore nearly the same intervention; on the frequency profile they are opposites, tail inflation concentrating on rare tokens (+0.883 against +0.233, 3.78×) and noise helping the common bands more (+0.227 against +0.297, 0.76×). That is the cancellation of Section 4 in a measurement (Appendix G).

### 5.4 A Count-Matched Forced-Choice Probe of Rare-Token Knowledge

Everything above is a likelihood; whether that matters depends on whether the tail does. Our first attempt returned a null: over ten generations of the largest-effect condition the arms separate on perplexity but not on ARC-Easy, PIQA and HellaSwag, which differ by less than a tenth of their standard error (Appendix R). Those benchmarks measure commonsense from pretraining while the claim concerns rare items in the fine-tuning corpus, so we registered a probe that matches it (Appendix S): at every held-out position the model picks the true next token out of itself plus 15 distractors of near-identical training-corpus count. It is an accuracy rather than a likelihood, it is on-theory (bands 1–3 carry 48,521 positions against 400 benchmark items), and it is calibration-immune, since a unigram tilt adds the same constant across a count-matched set.

**Rare-band accuracy.** On the same Pythia-410m arms, rare-band accuracy falls from 0.8403 to 0.8099 under $\tau = 1.0$ and to 0.8294 under $\tau = 1.15$, a difference of +1.95 points (Figure 4c, Appendix Z), block-clustered 95% CI [+1.66, +2.23], replicated at +2.25 [+1.97, +2.52] on a second

seed, seven times its own standard error, on the same checkpoints on which the benchmarks showed no difference.

The sign disagreement survives in an accuracy, removing the objection that a weighted average of oppositely-moving parts must disagree with its parts by construction. In both conditions where terminal perplexity prefers $\tau = 1.0$ the accuracy prefers $\tau = 1.15$: $+0.59$ pp $[+0.38, +0.80]$ at one epoch and $+0.37$ pp $[+0.14, +0.60]$ at two, the common bands moving the other way in both. Both replicate on a second seed in the likelihood ($A(b1\text{–}b3) = +0.309$ and $+0.426$) and in the accuracy ($+0.47$ and $+0.63$ pp, both CIs excluding zero), with the aggregate still preferring $\tau = 1.0$.

Generic noise does not reproduce it: it recovers $91\%$ of tail inflation's aggregate benefit and only $21\%$ of its rare-band cloze gain ($+0.42$ against $+2.01$ pp).

**The null at the best-fit end.** At the best-fit condition, SmolLM3-3B at one epoch, the probe reverses: rare-band accuracy prefers $\tau = 1.0$ by $0.34$ pp, CI $[-0.49, -0.20]$. This registered prediction failed with the sign inverted, and the calibration control implies it should at the best-fit end, where an evaluation-time tilt reproduces $99\%$ of the rare-band advantage and leaves a residual of $+0.001$ nats. The same model overfitted recovers it: at ratio $1.84$ the probe gives $+1.28$ pp $[+1.03, +1.52]$ and the tilt reproduces only $24\%$, residual $+0.509$. That reading was offered after the reversal, so we tested it out of sample; all four predictions held (Appendix V). The evidence therefore supports a scoped claim: tail inflation preserves rare-token knowledge where the generator is not already well fit. The effects are small, from a third of a point to two points of accuracy.

**External validity.** The decomposition predicts the probe better than any way of taking the aggregate (Table 6), and on the fixed-fit family of Table 2, where nothing orders the arms, its ranking still tracks the probe's (Appendix O).

## 6 Limitations and Conclusion

**The theory is a Zipf idealization.** Proposition 1 assumes one power-law tail index and unbounded support; a subword vocabulary is finite and its fitted index rises with corpus size ($\beta$ from $1.68$ to $2.05$), so the measured decay is steeper than the $c < 1$ it guarantees. Proposition 2 bounds the leverage but not the noise cost, so its projection holds the latter fixed. **Mechanism.** The overfitting moderator's mechanism is not demonstrated, and Theorem 1 predicts an exponent of $0$ under unbiased sampling while the well-fit $\tau = 1.0$ arms give $0.41$ to $0.64$. **Scope.** The largest model is 3B, the corpora are WikiText-2, WikiText-103 and TinyStories, and the band, causality and matched-compute arms are single-seed. The capability evidence is scoped to models not already well fit, established by one registered prediction failing and another holding. The size dependence and the optimal generator bias are unresolved; these and three design choices are in Appendix L. **What the failures do not touch.** No registered failure bears on the measurement claim that follows; the falsifications concern the moderator's magnitude and its mechanism.

**Conclusion.** We study how recursive training on model-generated data is measured. The damage it is reported to cause lies in the rare tail, yet it is almost always evaluated with held-out perplexity, which weights every token by how often it occurs, so rare tokens barely count: on WikiText-2 the rarest band carries $1.3\%$ of the validation mass against $23.7\%$ for the commonest. We propose reading the same evaluation by training-frequency band, an exact decomposition that needs no extra evaluation data and costs one scatter-add. We prove that the weighting cannot be outgrown. Under a Zipf corpus the validation mass on the endangered items falls as $T^{-c}$ with $c = 1 - 1/\beta$, the same exponent that governs how fast the tail is lost, so a larger corpus degrades the measurement as fast as it improves the process.

Applying the decomposition to ten recursive-training conditions, we find that in five the aggregate and the rare bands give opposite verdicts on the same intervention, while the rare-band benefit is positive in all ten. A pre-registered forced choice that tests rare-token knowledge directly separates the arms by $+1.95$ points where three standard benchmarks find nothing. A conclusion drawn from the aggregate can therefore be reversed by the numbers it already contains. We recommend reporting $(\mathcal{L}, \{\mathcal{L}_b\}, \{w_b\})$ wherever an intervention reshapes the frequency distribution, and calling it an improvement only when the aggregate and the rare bands agree in sign.

### AI USE STATEMENT

During the preparation of this manuscript, the authors used AI tools to assist with document formatting, LaTeX syntax, and technical code maintenance. In particular, an AI coding agent, Antigravity, was used to support refactoring and organizing the experimental codebase for release as supplementary material, with the aim of improving clarity and reproducibility for reviewers.

### ETHICS STATEMENT

This work studies degradation of generative models trained on synthetic data. It uses public corpora and public pretrained checkpoints, involves no human subjects, and releases no new data. A possible societal benefit is that frequency-resolved reporting makes harm to rare-event coverage visible at no additional evaluation cost, and such harm falls disproportionately on underrepresented groups (Shumailov et al., 2024). A possible harm is that the same result identifies a setting at which generated text is made harder to distinguish from human text by inflating its tail, which could degrade detection-based mitigations (Drayson et al., 2025).

### REPRODUCIBILITY STATEMENT

Proofs of Lemmas 1, 2, Theorem 1 and Proposition 1 are in Appendix A, including the mean-field solution and its fixed-point analysis. Simulation configuration is in Appendix J; language-model configuration, per-model hyperparameter selection, the generation protocol, the cross-stack determinism check and the definition of $\hat{\delta}$ are in Section 5 and Appendix C. The corpus-scaling measurement behind Section 3.2 is in Appendix Q. Per-generation metrics for all 84 distinct runs are tabulated in Appendix Z. The code release at `https://anonymous.4open.science/r/band-resolved-collapse-1319`, attached to this submission as the supplementary archive, carries the experiment and analysis code, the batch scripts that produced every published arm, and the per-generation metrics for all of them, so the analysis reproduces without re-training anything.

## REFERENCES

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. Self-consuming generative models go MAD. In *International Conference on Learning Representations (ICLR)*, 2024.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, Xuan-Son Nguyen, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM3: smol, multilingual, long-context reasoner. Hugging Face blog, 2025. URL `https://huggingface.co/blog/smollm3`. Base checkpoint `https://huggingface.co/HuggingFaceTB/SmolLM3-3B-Base`.

Daniel Barzilai and Ohad Shamir. When models don't collapse: On the consistency of iterative MLE. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2505.19046.

Quentin Bertrand, Joey Bose, Alexandre Duplessis, Marco Jiralerspong, and Gauthier Gidel. On the stability of iterative retraining of generative models on their own data. In *International Conference on Learning Representations (ICLR)*, 2024.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning (ICML)*, 2023.

Ali Borji. A note on shumailov et al. (2024): "AI models collapse when trained on recursively generated data". *arXiv preprint arXiv:2410.12954*, 2024.

Apratim Dey and David L. Donoho. Universality of the $\pi^2/6$ pathway in avoiding model collapse. *arXiv preprint arXiv:2410.22812*, 2024.

Elvis Dohmatob, Yunzhen Feng, and Julia Kempe. Model collapse demystified: The case of regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a.

Elvis Dohmatob, Yunzhen Feng, Pu Yang, François Charton, and Julia Kempe. A tale of tails: Model collapse as a change of scaling laws. In *International Conference on Machine Learning (ICML)*, 2024b.

Elvis Dohmatob, Yunzhen Feng, Arjun Subramonian, and Julia Kempe. Strong model collapse. In *International Conference on Learning Representations (ICLR)*, 2025.

George Drayson, Emine Yilmaz, and Vasileios Lampos. Machine-generated text detection prevents language model collapse. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 29657–29673, 2025. doi: 10.18653/v1/2025.emnlp-main.1506.

Ronen Eldan and Yuanzhi Li. TinyStories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*, 2023.

Yunzhen Feng, Elvis Dohmatob, Pu Yang, François Charton, and Julia Kempe. Beyond model collapse: Scaling up with synthesized data requires verification. In *International Conference on Learning Representations (ICLR)*, 2025.

Damien Ferbach, Quentin Bertrand, Avishek Joey Bose, and Gauthier Gidel. Self-consuming generative models with curated data provably optimize human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Ramon Ferrer i Cancho and Ricard V. Solé. Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics*, 8(3):165–173, 2001.

Daniele Gambetta, Gizem Gezici, Fosca Giannotti, Dino Pedreschi, Alistair Knott, and Luca Pappalardo. Learning by surprise: Adaptive mitigation of model collapse in large language models. *ACM Transactions on Intelligent Systems and Technology*, 2026. doi: 10.1145/3828663.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. In *Conference on Language Modeling (COLM)*, 2024.

Dongxin Guo, Jikun Wu, and Siu Ming Yiu. Model collapse as cultural evolution. *arXiv preprint arXiv:2605.23054*, 2026.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*, 2020.

Marcus Hutter. Learning curve theory. *arXiv preprint arXiv:2102.04074*, 2021.

Millen Kanabar and Michael Gastpar. Model non-collapse: Minimax bounds for recursive discrete distribution estimation. In *IEEE International Symposium on Information Theory (ISIT)*, 2025.

Joshua Kazdan, Rylan Schaeffer, Apratim Dey, Matthias Gerstgrasser, Rafael Rafailov, David L. Donoho, and Sanmi Koyejo. Collapse or thrive? perils and promises of synthetic data in a self-generating world. In *International Conference on Machine Learning (ICML)*, 2025.

Figarri Keisha, Zekun Wu, Ze Wang, Adriano S. Koshiyama, and Philip C. Treleaven. Knowledge collapse in LLMs: When fluency survives but facts fail under recursive synthetic training. *arXiv preprint arXiv:2509.04796*, 2025.

Ian Magnusson, Akshita Bhagia, Valentin Hofmann, Luca Soldaini, Ananya Harsh Jha, Oyvind Tafjord, Dustin Schwenk, Evan Pete Walsh, Yanai Elazar, Kyle Lo, Dirk Groeneveld, Iz Beltagy, Hannaneh Hajishirzi, Noah A. Smith, Kyle Richardson, and Jesse Dodge. Paloma: A benchmark for evaluating language model fit. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Clara Meister and Ryan Cotterell. Language model evaluation beyond perplexity. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*, 2017.

Andrea Pinto, Tomer Galanti, and Randall Balestriero. The fair language model paradox. *arXiv preprint arXiv:2410.11985*, 2024.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Rylan Schaeffer, Joshua Kazdan, Alvan Caleb Arulandu, and Sanmi Koyejo. Position: Model collapse does not mean what you think. *arXiv preprint arXiv:2503.03150*, 2025.

Mohamed El Amine Seddik, Suei-Wen Chen, Soufiane Hayou, Pierre Youssef, and Mérouane Debbah. How bad is training on synthetic data? a statistical analysis of language model collapse. In *Conference on Language Modeling (COLM)*, 2024.

Soheil Zibakhsh Shabgahi, Pedram Aghazadeh, Azalia Mirhoseini, and Farinaz Koushanfar. ForTIFAI: Fending off recursive training induced failure for AI model collapse. *npj Artificial Intelligence*, 2026. doi: 10.1038/s44387-026-00127-w.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*, 2023.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759, 2024. doi: 10.1038/s41586-024-07566-y.

Ananda Theertha Suresh, Andrew Thangaraj, and Aditya Nanda Kishore Khandavally. Rate of model collapse in recursive training. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1396–1404, 2025.

Lecheng Wang, Xianjie Shi, Ge Li, Jia Li, Xuanming Zhang, Yihong Dong, Wenpin Jiao, and Hong Mei. Theoretical proof that auto-regressive language models collapse when real-world data is a finite set. *arXiv preprint arXiv:2412.14872*, 2024.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

George Kingsley Zipf. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press, Cambridge, MA, 1949.

## GUIDE TO THE APPENDICES

The appendices fall into six groups. **Theory**: Appendix A proves Lemmas 1 and 2, Theorem 1 and Propositions 1 and 2, and Appendix B develops the mean-field equation for biased generation. **Configuration**: Appendices C, D and Z give the experimental setup, the intervention arms and a table of every distinct run. **The band statistic**: Appendix O asks whether it predicts the probe, Appendix P.1 whether the band definition or the noise floor drives the result, and Appendix Q how the weights behave across corpora, tokenizers and corpus sizes. **Controls**: the noise control is Appendix G and the calibration control Appendix H. **Capability**: Appendix R reports the benchmark null and Appendix S the registered forced-choice probe. **Pre-registrations**: Appendices S, U, V, W and X reproduce five of the seven registrations in substance, each with its predictions, the measured values and the verdict. Limitations are in Appendix L. **Supporting analyses**: Appendix F collects six settings in which aggregate risk and distributional structure select opposite interventions, Appendix E reports measured generator bias as a predictor of the collapse rate, Appendix K the beam-search signature, and Appendix I the decoder-workflow interaction scored two ways.

# A PROOFS

Throughout, $\mathcal{F}_n$ is the natural filtration, $N_n = T(1+nk)$ is the pool size at generation $n$, $c_n$ is a fixed item's pool count, and $Z_n = Tc_n/N_n = c_n/(1+nk)$ is its expected count in the size-$T$ training subsample. Let $X_n$ denote its realized training count and $\Delta G_{n+1}$ its contribution to the next synthetic batch, so $c_{n+1} = c_n + \Delta G_{n+1}$.

**Proof of Lemma 1.** Sampling the training set gives $X_n \mid \mathcal{F}_n \sim \mathrm{Bin}(T, \theta_n)$ with $\theta_n = c_n/N_n$, so $\mathbb{E}[X_n \mid \mathcal{F}_n] = T\theta_n = Z_n$. Generation gives $\Delta G_{n+1} \mid X_n \sim \mathrm{Bin}(kT, X_n/T)$, so $\mathbb{E}[\Delta G_{n+1} \mid X_n] = kX_n$. By the tower property, $\mathbb{E}[\Delta G_{n+1} \mid \mathcal{F}_n] = kZ_n$, hence

$$\mathbb{E}[c_{n+1} \mid \mathcal{F}_n] = c_n + k\frac{c_n}{1+nk} = c_n\,\frac{1+(n+1)k}{1+nk}, \tag{4}$$

and dividing by $1+(n+1)k$ gives $\mathbb{E}[Z_{n+1} \mid \mathcal{F}_n] = c_n/(1+nk) = Z_n$. □

**Proof of Lemma 2.** By the conditional variance decomposition, $\mathrm{Var}(\Delta G_{n+1} \mid \mathcal{F}_n) = \mathbb{E}[\mathrm{Var}(\Delta G \mid X_n)] + \mathrm{Var}(\mathbb{E}[\Delta G \mid X_n]) = k\,\mathbb{E}[X_n(1 - X_n/T)] + k^2\mathrm{Var}(X_n)$. Using $\mathbb{E}[X_n^2] = Z_n(1-\theta_n) + Z_n^2$,

$$\mathrm{Var}(\Delta G_{n+1} \mid \mathcal{F}_n) = kZ_n\big[1 - \tfrac{1-\theta_n}{T} - \tfrac{Z_n}{T}\big] + k^2 Z_n(1-\theta_n) \;\le\; k(1+k)Z_n. \tag{5}$$

The bound is exact for binomial sampling; no Poisson approximation is used. Dividing by $(1+(n+1)k)^2$ and summing,

$$\mathbb{E}\langle Z\rangle_\infty \le k(1+k)\sum_{n\ge 0}\frac{\mathbb{E}[Z_n]}{(1+(n+1)k)^2} = k(1+k)Z_0\sum_{m\ge 1}\frac{1}{(1+mk)^2} =: C_k Z_0, \tag{6}$$

using $\mathbb{E}[Z_n] = Z_0$ from Lemma 1. For $k=1$, $\sum_{m\ge1}(1+m)^{-2} = \sum_{j\ge2} j^{-2} = \pi^2/6 - 1$, so $C_1 = 2(\pi^2/6-1) \approx 1.28987$. □

**Proof of the consequences of Lemmas 1 and 2.** Since $\mathbb{E}[Z_n^2] = Z_0^2 + \mathbb{E}\langle Z\rangle_n \le Z_0^2 + C_k Z_0 < \infty$, $(Z_n)$ is $L^2$-bounded, hence uniformly integrable, hence converges a.s. and in $L^1$ with $\mathbb{E}[Z_\infty] = Z_0$. By Cauchy-Schwarz on $\mathbb{E}[Z_\infty] = \mathbb{E}[Z_\infty \mathbf{1}\{Z_\infty > 0\}]$,

$$Z_0^2 \le \mathbb{E}[Z_\infty^2]\,\Pr(Z_\infty > 0) \quad\Longrightarrow\quad \Pr(Z_\infty > 0) \;\ge\; \frac{Z_0}{Z_0 + C_k}. \tag{7}$$

□

**Proof of Theorem 1.** $\mathrm{Var}(Z_\infty) \le C_k Z_0$, so Cantelli's inequality at $t = Z_0/2$ gives $\Pr(Z_\infty \le Z_0/2) \le C_k Z_0/(C_k Z_0 + Z_0^2/4) = 4C_k/(4C_k + Z_0)$. The item is absent from the training set with probability $\mathbb{E}[e^{-Z_n}]$; splitting on the event above and bounding the complement by $e^{-Z_0/2}$,

$$\limsup_n \mathbb{E}[E_n] = \sum_i p_i \limsup_n \mathbb{E}\big[e^{-Z_n^{(i)}}\big] \le \sum_i p_i\left[\frac{4C_k}{4C_k + Tp_i} + e^{-Tp_i/2}\right], \tag{8}$$

with $Z_0^{(i)} = Tp_i$. Every term is finite and free of $n$. □

**Proof of Corollary (Zipf scaling).** Let $p_i = Ci^{-\beta}$ and split at $i_\star$ with $Tp_{i_\star} = 4C_k$, i.e. $i_\star \asymp (CT/4C_k)^{1/\beta}$. For $i < i_\star$ each term is $\approx 4C_k/T$, and there are $\asymp i_\star$ of them, contributing $\asymp T^{1/\beta - 1} = T^{-c}$. For $i > i_\star$ each term is $\approx p_i$, and $\sum_{i>i_\star} p_i \asymp i_\star^{-(\beta-1)} \asymp T^{-(\beta-1)/\beta} = T^{-c}$. Both pieces are $\Theta(T^{-c})$ with $c = 1 - 1/\beta$. □

**Proof of Proposition 1.** Write $p_i = Ci^{-\beta}$ with $\beta > 1$. The items with $Tp_i \le m$ are those of rank at least $i_m := (CT/m)^{1/\beta}$, so with $c = 1 - 1/\beta$,

$$M(m,T) \;=\; \sum_{i\ge i_m} Ci^{-\beta} \;=\; \frac{C}{\beta-1}\,i_m^{-(\beta-1)}(1+o(1)) \;=\; \frac{C^{1/\beta}}{\beta-1}\Big(\frac{m}{T}\Big)^c(1+o(1)), \tag{9}$$

using $C \cdot C^{-(\beta-1)/\beta} = C^{1/\beta}$ and $(\beta-1)/\beta = c$; the $o(1)$ is the Euler-Maclaurin remainder, uniform for $i_m \to \infty$. For (a), an item with $Tp_i \le m$ contributes $p_i \le m/T$ of the validation mass, so raising its per-token loss by at most $\Lambda$ raises $\mathcal{L} = \sum_b w_b \mathcal{L}_b$ by at most $\Lambda m/T$. For (b), summing over the whole at-risk set gives $\Lambda M(m,T)$, which is $\Theta(T^{-c})$ at fixed $m$. For (c), let $\hat{\mathcal{L}}$ be the empirical held-out loss on $N$ tokens and $\sigma^2 > 0$ a lower bound on the per-token variance of the paired difference between the damaged and undamaged models. Any test that separates them at type-I error $\alpha$ and power $1-\gamma$ needs its mean shift to exceed $(z_\alpha + z_\gamma)\sigma/\sqrt{N}$, and by (b) that shift is at most $\Lambda M(m,T)$, so $N \ge (z_\alpha + z_\gamma)^2\sigma^2/(\Lambda M(m,T))^2 = \Omega(T^{2c})$. □

The corollary quoted in Section 3.2 is immediate: for geometric count bands with edges $e_b$, $w_b = M(e_{b+1},T) - M(e_b,T) \propto e_b^c(\rho^c - 1)$ with $\rho = e_{b+1}/e_b$, so $w_K/w_1 = \rho^{(K-1)c}$. At WikiText-2's fitted 1.47 this predicts a head-to-tail ratio of 25× against a measured 18×, with no free parameter. The middle bands depart from the single-exponent form, as they must: English unigram frequencies have two regimes rather than one, with a flatter head and a steeper tail (Ferrer i Cancho & Solé, 2001), and our own fits give $\beta = 0.94$ over the first thousand ranks against 1.68 beyond them. We therefore use the corollary for the order of magnitude of the ratio and never for a per-band prediction.

**Proof of Proposition 2.** By Eq. (3) the aggregate is linear in the per-band losses with weights $w_b$, so a change of $\delta$ confined to $R$ moves $\mathcal{L}$ by exactly $w\delta$ and $\mathcal{L}_R$ by exactly $\delta$; both are identities, not approximations. (a) A two-sided test at level $\alpha$ declares a change when the observed movement exceeds $z_\alpha$ times its standard deviation, so the smallest detectable $\delta$ is $z_\alpha\sigma/w$ from the aggregate and $z_\alpha\sigma_R$ from the region statistic, whose ratio is $G = \sigma/(w\sigma_R)$ with the $z_\alpha$ cancelling. (b) If the noise is the finite held-out sample of $N$ tokens then $\sigma = s/\sqrt{N}$ and, because the region-restricted mean rests on $Nw$ of those tokens in expectation, $\sigma_R = s_R/\sqrt{Nw}$, so $G = (s/s_R)\, w^{-1/2}$: the two square roots leave $w^{-1/2}$ and not $w^{-1}$, and by Proposition 1 that is $\Theta(T^{c/2})$. The sample size, not the threshold ratio, scales as $1/w$. Equating the two $z$-scores and solving for $N$ gives $N_{\mathcal{L}}/N_{\mathcal{L}_R} = (s/s_R)^2/w = \Theta(T^c)$ when $s \asymp s_R$, which recovers Proposition 1(c). If instead the noise is run-to-run variation, $\sigma$ and $\sigma_R$ do not depend on $N$, the cancellation does not happen, and $G = (1/w)(\sigma_R/\sigma)^{-1} = \Theta(T^c)$ exactly, with $T^*$ solving $M(m,T^*) = \sigma/\sigma_R$ finite because $M(m,\cdot)$ is continuous and strictly decreasing to 0. The gain grows without bound in $T$ in both regimes; the exponent is $c$ when reproducibility-limited and $c/2$ when sampling-limited, and everything measured in this paper is the former, since a validation split is fixed and it is the seeds that vary. (c) The aggregate moves by $w\delta + w_H\gamma$, whose sign is that of $\delta$ exactly when $|\gamma| < (w/w_H)|\delta|$; with $w = M(m,T) = \Theta(T^{-c})$ and $w_H$ bounded away from 0 the tolerance is $\Theta(T^{-c})$, and it involves neither $N$ nor the number of seeds. □

The proposition leaves two things open. It does not bound $\sigma_R/\sigma$, which is a property of the model and the evaluation set rather than of the corpus, and which we therefore measure (Appendix P.3) rather than assume. It also treats the region as known; in practice the bands are assigned from training counts, which is what makes them computable without a second evaluation set but also what ties $\mathcal{L}_{\text{rare}}$ to one training corpus, as Section 4.1 says.

**The mean-field equation, its solution and its fixed point.** Under temperature $\tau$ the generator emits $q \propto f^{1/\tau}$; writing $\delta = 1/\tau - 1$, $u = Z/T$, $t = \log(1+nk)$ and $W = \sum_j f_j^{1/\tau}$, the dynamics of a single item become autonomous:

$$\frac{du}{dt} = -u\left(1 - \frac{u^\delta}{W}\right). \tag{10}$$

Writing $y = u^{-\delta}$ turns $du/dt = -u(1-u^\delta/W)$ into the linear equation $dy/dt = \delta(y - 1/W)$, whose solution is $y(t) = 1/W + (y_0 - 1/W)e^{\delta t}$, i.e. $u(t)^{-\delta} = 1/W + (u_0^{-\delta} - 1/W)e^{\delta t}$. With $t = \log(1+nk)$ and the death condition $u = 1/T$, inverting for the initial frequency whose trajectory reaches threshold exactly at generation $n$ gives $m_{\text{crit}}(n) = T[\,1/W + (T^\delta - 1/W)(1+nk)^{-\delta}\,]^{-1/\delta}$, which saturates rather than growing linearly because $\rho$ increases with the threshold.

The interior fixed point is $u_\star = W^{1/\delta}$, obtained by setting $u^\delta = W$. For $\tau < 1$ we have $\delta = 1/\tau - 1 > 0$ and, since $f_j \le 1$ and $1/\tau > 1$, $W = \sum_j f_j^{1/\tau} < 1$. Then for $u < u_\star$, $u^\delta < u_\star^\delta = W$, so $u^\delta/W < 1$ and $du/dt < 0$: frequencies below the fixed point are driven to extinction. For $\tau > 1$

we have $\delta < 0$ and $W > 1$, so $u_\star = W^{1/\delta} \in (0, 1)$ and for $u < u_\star$, $u^\delta > u_\star^\delta = W$, giving $du/dt > 0$: frequencies below the fixed point are driven up and away from the extinction threshold $u = 1/T$, while converging toward the common floor $u_\star$, which is the flattening that inflates the generated tail.

Balancing rescue against flattening by setting the floor at the extinction threshold, $u_\star = 1/T$, gives $W^{1/\delta} = 1/T$ and hence $\delta_\star = -\log W / \log T$. This balance point is recorded for the limitation it exposes rather than as a result. At $\beta = 2$ the function $g(\delta) = \delta \log T + \log W(\delta)$ has $g(0) = 0$ identically and $g'(0) = \log T - H \approx +9.9$, where $H$ is the entropy of $f$, so $\delta = 0$ is the only root in the relevant range and the nontrivial root sits at $\delta \approx -0.002$ (or $\approx -0.84$ at $V = 10^6$, equally far from the measurement in the other direction). Neither is close to the measured $-0.088$. The reason is that Eq. (10) is noiseless: the interior optimum exists because stochastic attrition removes rare items and inflation compensates, and a deterministic equation has no term for the attrition, so its balance point cannot be displaced from zero bias by the right amount for the right reason. A stochastic (diffusion) treatment is the natural extension, which we have not attempted.

## B The Mean-Field Equation in Detail

**Predictive accuracy and separability.** Integrating Eq. (10) and fitting over the same window used in simulation predicts the measured exponent with correlation $0.963$ across 38 $(\beta, \tau, T)$ conditions with no fitted parameters (mean absolute error $0.083$, RMSE $0.109$), including the weak $T$-dependence that was found empirically only afterwards. It also accounts for the observed separability $a = c(\beta)g(\tau)$: $\beta$ enters only the threshold-to-mass conversion and the replenishment ratio $\rho$ only the threshold trajectory, so the two never mix.

**Location of the interior optimum.** The memorization idealization cannot produce an interior optimum at all: its generator is the empirical distribution, so an unseen item has $q_i = 0$, and since $0^{1/\tau} = 0$ the support can never grow, making inflation free. Real models differ in two ways, neither sufficient alone. A softmax has full support, so inflation misallocates mass and acquires a cost. A trained model discounts low-count events, so Lemma 1's hypothesis fails even at $\tau = 1$; this is what moves the optimum off zero, since discounting is itself a suppression bias. The balance point is therefore a property of the learner rather than of recursion alone, and should move when the fitting recipe moves. Section 5.1 confirms that it does.

**Sign difference between replace and accumulate.** Under accumulation the drift term carries a factor $k/(1 + (n + 1)k)$, so bias acts logarithmically in $n$ and retained real data anchors the pool. Under replace no such damping exists and bias composes: tempering by $\tau$ every generation gives effective temperature $\tau^n$, which on the inflating side runs away toward the uniform distribution (unbounded damage) and on the suppressing side terminates at a low-entropy fixed point of further tempering. Replace therefore inverts the preference, which is the interaction reported in Appendix I.

## C Language Model Experiments in Detail

**Fit selection.** A single fine-tuning recipe cannot serve models spanning 30M to 1B parameters: at 5 epochs and $\mathrm{lr} = 5 \times 10^{-5}$, Qwen2.5-0.5B reaches validation perplexity 225 against a 43 zero-shot baseline, i.e. fine-tuning makes it five times worse. We therefore select epochs and learning rate per model by sweep, subject to beating the zero-shot baseline, and abort any arm whose generation-0 perplexity fails that test. Using one recipe across scales would confound model size with fit quality. The cost is that epochs and size covary in the resulting design, which is the confound behind the unresolved size dependence of Section 6.

**Generation protocol.** Generation is prompted, following Shumailov et al. (2024): each training block conditions a continuation of equal length. Free-running generation from a start token is not comparable, because at low temperature it collapses into repetition loops rather than distributional degradation: a 30M model at $\tau = 0.7$ produced 297 distinct bigrams from 4M tokens by generation 1. This is the degeneracy Shumailov et al. (2024) neutralize with a repetition penalty, and it

is a property of the decoder, not of recursive training. An earlier version of our harness generated free-running by default; all results reported here use the prompted protocol.

**Effective bias.** We compute $\hat{\delta}$ by binning the vocabulary into eight frequency buckets by real-corpus frequency, forming the ratio $q/f$ of generated to real mass in each bucket, and regressing $\log(q/f)$ on the log of the bucket's geometric-mean frequency. Temperature scaling gives $q/f \propto p^{\delta}$ exactly, so the slope recovers $\delta$. Bucket boundaries are fixed by the real corpus and shared across all arms within a model family, so $\hat{\delta}$ is comparable within a family. It is not comparable across families with different tokenizers, which is why Figure 2(c) is presented as a separate panel rather than merged with (a).

**Determinism within and across software stacks.** The pipeline fixes every seed and does no non-deterministic reduction of its own, so two arms with identical configuration on identical software reproduce each other bit-for-bit. Ten arms in Table 12 are such re-runs, added to collect instrumentation that the original did not carry, and they reproduce their originals to every digit; that identity is a free correctness check on the instrumentation, and those arms are excluded from the run count. Across different software stacks the agreement is close but not exact. Re-running `bep1_t1.0` on a machine with a different PyTorch and Transformers build reproduces its per-band NLL to $10^{-2}$ nats and its rare-band forced-choice accuracy to $10^{-3}$ (band-1 NLL 7.699 against 7.709; rare-band cloze 0.8201 against 0.8209). Every effect we interpret is one to three orders of magnitude larger than that, but the drift is not zero, so each comparison in this paper is made between arms produced on one stack, never across two, and the cross-stack re-run is reported here rather than counted as an arm.

**Exponent fitting.** The reported exponent is the slope of $\log(\mathrm{ppl}_n - \mathrm{ppl}_0)$ against $\log n$ over $n \geq 1$, with $R^2$ computed in the original units. We report $R^2$ for every arm in Appendix Z and flag arms whose power-law fit is poor rather than dropping them.

**Reference NLL under a frozen pretrained model.** Scoring the generated corpus under a frozen pretrained model gives a coherence measure $\hat{\delta}$ cannot, since $\hat{\delta}$ is a unigram statistic: real WikiText-2 scores 4.599 nats under it. At $\tau = 1.0$ the corpus sits at 4.75 and stays flat for all nine generations; at $\tau = 0.7$ it falls to 2.28, the degenerate low-entropy attractor; at $\tau = 1.15$ it climbs $5.79 \rightarrow 10.15$ and at $\tau = 1.6$ to 12.35. One caution applies: the correction of Appendix M was expected to leave this axis untouched and does not ($12.35 \rightarrow 10.41$), so the unigram and sequence channels are coupled.

## D Intervention Arms in Detail

All at $\tau = 0.7$ under accumulate-subsample on OPT-125m, against a control with exponent 0.469, terminal perplexity 61.2, diversity $-51.0\%$ and $\hat{\delta}$ drifting $0.109 \rightarrow 0.153$.

**Verification filter (keep the most probable 50%).** Exponent 0.600, terminal perplexity 57.7, diversity $-59.4\%$, $\hat{\delta} \rightarrow 0.150$. This is the one arm in the paper where terminal perplexity and the fitted exponent disagree in sign: the filter starts flatter and steepens, so it ends at lower perplexity while fitting a steeper power law. A reader who cares about the ten-generation endpoint rather than the rate should read it as a mild perplexity improvement bought with a 17% loss of lexical diversity. On either reading it does not act through the unigram frequency profile, since $\hat{\delta}$ is statistically unchanged.

**Real-data anchor (10% retained every generation).** Exponent 0.307, terminal perplexity 57.0, diversity $-46.2\%$, $\hat{\delta} \rightarrow 0.144$. Better on all three, and it matches the anchor dose-response of Appendix J and the qualitative finding of Shumailov et al. (2024) that a small real-data admixture largely repairs the damage.

**Logit-level correction.** At $\tau = 0.7$ it holds $\hat{\delta}$ flat at 0.107 across all nine generations against a control drifting to 0.153, and at $\tau = 1.6$ holds $-0.379$ against a control amplifying to $-0.814$.

Table 3: Six settings in which aggregate risk and distributional structure select opposite interventions. All pairs are matched on model, corpus and generation-0 checkpoint; only the named variable differs. "div" is distinct bigrams, $q/f$ is generated mass on the rarest frequency band relative to real.

| Setting | Comparison | Risk prefers | Structure prefers |
|---|---|---|---|
| replace, gen 9 | beam-5 vs $\tau$=1.0 | beam (83.4 vs 123.4) | $\tau$=1.0 (4.0× div, 6.9× $q/f$) |
| replace, 10 gens | $\tau$=0.7 vs $\tau$=1.0 | 0.7 (121.6 vs 123.4) | 1.0 ($-14\%$ vs $-60\%$ div) |
| TinyStories, gen 1 | $\tau$=0.7 vs $\tau$=1.0 | 0.7 ($+2.73$ vs $+3.35$) | 1.0 (748× more $q/f$) |
| accumulate, OPT | top-$p$ 0.9 vs $\tau$=0.7 | $\tau$=0.7 (0.469 vs 0.501) | top-$p$ ($-35\%$ vs $-51\%$ div) |
| accumulate, OPT | $\tau$=1.9 vs $\tau$=0.7 | $\tau$=0.7 (0.469 vs 0.799) | $\tau$=1.9 ($+2\%$ vs $-51\%$ div) |
| accumulate, OPT | verify-50% vs none | verify (57.7 vs 61.2) | none ($-51\%$ vs $-59\%$ div) |

Terminal distinct bigrams rise from 118k to 137k. Exponents $0.476$ against $0.469$ at $\tau = 0.7$ and $0.746$ against $0.794$ at $\tau = 1.6$.

**Block-level reweighting.** Importance-resampling generated blocks toward the real frequency profile: exponent $0.439$ at full strength and $0.434$ at half, $\hat{\delta} \to 0.141$ and $0.153$ respectively.

## E Measured Bias as a Predictor of the Collapse Rate

Across the tail-suppressing branch, five arms spanning both decoder families and $\hat{\delta}$ from $0.032$ to $0.109$ give exponents in the narrow band $0.466$ to $0.501$, so two mechanistically unrelated decoders that happen to land on the same measured bias land on the same collapse rate. Over the eight OPT-125m accumulate temperature arms of Table 12, $|\hat{\delta}|$ and the collapse exponent correlate at $+0.85$, and at $+0.83$ over the eleven points the three nucleus arms add. The signed correlation is $-0.77$ within the temperature series and near zero across model families, so only the magnitude generalizes: bias harms in both directions. Restricted to the five temperatures that existed before the $\tau \in \{1.15, 1.45, 1.9\}$ arms the correlation is $+0.93$, so it is sensitive to the arm set and we quote the eight-arm value.

Fitting a V of the form $|\hat{\delta} - m|$ and maximizing the correlation over $m$ puts the vertex at $m^\star = -0.049$ (correlation $0.881$) on the temperature series and $-0.051$ with the nucleus arms included. The best sampled arm sits at $-0.088$. Both estimates are strictly negative, which is the claim we make; we do not defend a precise location for the vertex.

## F Six Settings with Opposed Interventional Rankings

Table 3 collects six settings in which aggregate risk and distributional structure select opposite interventions. They span both workflows, two corpora and four decoders, and each pair is matched on model, corpus and generation-0 checkpoint, so within a row only the named variable differs. The disagreement is therefore not particular to the temperature manipulation Section 5.1 uses as its instrument.

## G The Noise Control in Detail

At 10 epochs on OPT-125m, all three arms sharing the generation-0 model whose validation perplexity is 60.7, against a $\tau = 1.0$ control that reaches 95.3 after nine generations. "aggregate benefit" is the terminal-perplexity improvement over that control.

| intervention | terminal ppl | agg. benefit | rare ($b1$–$b3$) | common ($b6$–$b8$) | rare/com |
|---|---|---|---|---|---|
| tail inflation ($\tau$=1.15) | 70.3 | 25.0 | $+0.883$ | $+0.233$ | $\mathbf{3.78}\times$ |
| context-free noise, 15% | 72.6 | 22.8 (91%) | $+0.227$ | $+0.297$ | $\mathbf{0.76}\times$ |
| context-free noise, 30% | 75.0 | 20.4 (81%) | $+0.333$ | $+0.266$ | $1.26\times$ |

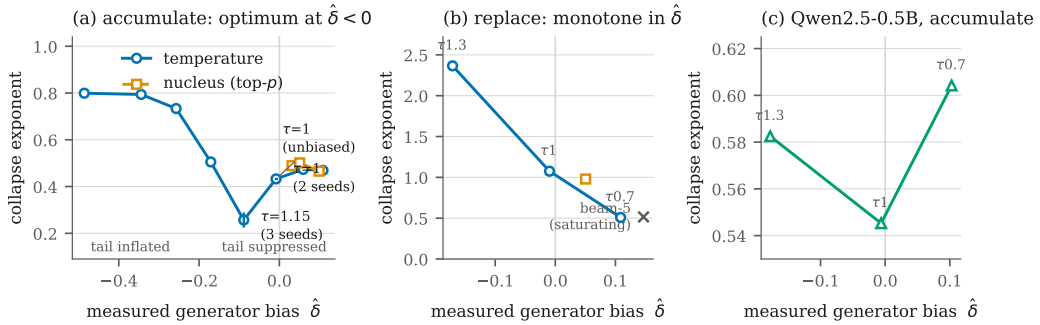


Figure 2: Collapse exponent against measured generator bias $\hat{\delta}$, estimated from generated text at generation 0; $\hat{\delta} > 0$ means the rare tail is suppressed, $\hat{\delta} < 0$ that it is inflated. **(a)** Under accumulation at the overfitting five-epoch recipe of Shumailov et al. (2024), OPT-125m traces a V whose minimum is displaced to the tail-inflating side ($\tau = 1.15$, $\hat{\delta} = -0.088$, exponent $0.257$ against $0.433$ at $\tau = 1.0$; the bar is the full range over three seeds). Nucleus sampling lands on the same curve as temperature at matched $\hat{\delta}$. **(b)** Under replace the same axis is monotone decreasing, the opposite ordering to (a); beam search does not follow a power law and is marked separately. **(c)** Qwen2.5-0.5B at the one epoch its own sweep selects is well fit and has its minimum back at $\tau = 1.0$. Panels (a) and (c) are the overfitted and the well-fit regimes of Section 5.1.

Table 4: The calibration control in full. For each condition we sweep the evaluation-time unigram tilt $-k \log f$ applied to the $\tau = 1.0$ model over $k \in [0, 0.3]$ and report the $k$ that best reproduces the observed rare-band advantage, the criterion most favourable to the calibration account. "bands won" counts, per condition, how many of the nine bands $\tau$=1.15 beats when the tilt is chosen separately for each band over the whole sweep, a still more generous criterion.

| | rare bands ($b1$–$b3$) | | | | common bands ($b6$–$b8$) | | | |
|---|---|---|---|---|---|---|---|---|
| epochs | observed | best $k$ | tilt | reproduced | observed | tilt at that $k$ | overshoot | bands won |
| 1 | $+0.327$ | 0.15 | $+0.305$ | 93% | $-0.109$ | $-0.296$ | $2.7\times$ | $2/9$ |
| 2 | $+0.410$ | 0.20 | $+0.342$ | 83% | $-0.085$ | $-0.414$ | $4.9\times$ | $2/9$ |
| 5 | $+0.617$ | 0.20 | $+0.362$ | 59% | $+0.043$ | $-0.353$ | sign fails | $8/9$ |
| 10 | $+0.883$ | 0.20 | $+0.391$ | 44% | $+0.233$ | $-0.312$ | sign fails | $8/9$ |

The 30% arm is the one carrying the forced-choice probe, and it recovers $+0.42$ pp of rare-band cloze accuracy against tail inflation's $+2.01$, which is 21%. Its calibration residual is $-0.021$ nats: an evaluation-time unigram tilt reproduces essentially all of its rare-band likelihood advantage, which is the expected signature of an intervention that flattens the marginal without adding conditional information.

## H The Calibration Control in Detail

For well-fit models the best tilt reproduces most of the rare-band advantage, but it overshoots the common-band cost several-fold. For overfitted models the observed common-band term is positive, which no tilt can produce, since a tilt pays for a rare gain with a common loss. Table 4 gives both criteria.

## I The Decoder-Workflow Interaction under Two Metrics

The decoder moves the collapse exponent over a $4.6\times$ range under replace and by less than 17% under accumulation, and the two workflows prefer opposite directions of bias (Table 5). Scored on distinct bigrams instead, the decoder acts under both workflows and the interaction is weak (Figure 3). An evaluation reporting only aggregate risk under accumulation therefore concludes

Table 5: OPT-125m on WikiText-2, ten generations, identical generation-0 model (perplexity 43.31) in every cell. $\hat{\delta}$ is the measured generator bias at generation 0; "div" is the change in distinct bigrams of the generated corpus. The temperature-1.15 row is the mean of three seeds.

| | | replace | | | accumulate | | |
|---|---|---|---|---|---|---|---|
| Decoder | $\hat{\delta}$ | exponent | $\text{ppl}_9$ | div | exponent | $\text{ppl}_9$ | div |
| $\tau$=1.3 | −0.171 | 2.366 | 1521.5 | +46.7% | 0.505 | 49.0 | +45.8% |
| $\tau$=1.15 | −0.088 | | not run | | **0.257** | **48.4** | +73.0% |
| $\tau$=1.0 | −0.008 | 1.083 | 123.4 | −14.3% | 0.433 | 54.8 | +0.4% |
| top-$p$ 0.9 | +0.051 | 0.980 | 142.1 | −66.6% | 0.501 | 60.6 | −34.9% |
| $\tau$=0.7 | +0.109 | **0.509** | 121.6 | −59.6% | 0.469 | 61.2 | −51.0% |
| beam-5 | +0.147 | saturating | 83.4 | −32.3% | | not run | |

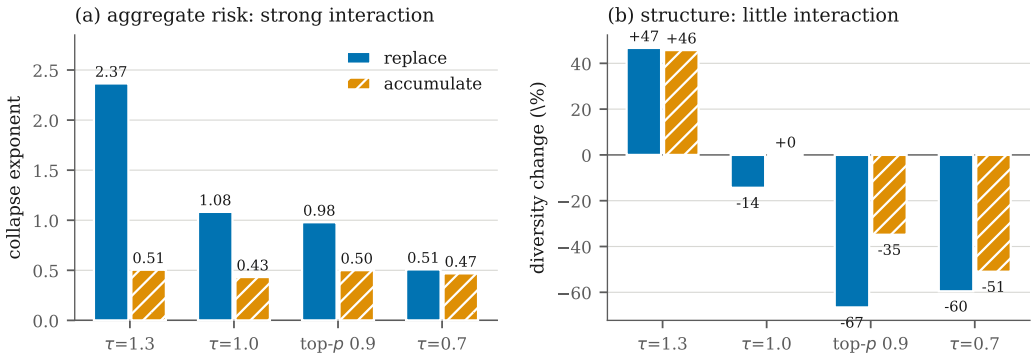


Figure 3: The same eight OPT-125m arms scored two ways. **(a)** On the collapse exponent, the decoder's effect is large under replace (blue) and small under accumulation (amber). **(b)** On distinct bigrams of the generated corpus the decoder acts under both workflows and the interaction is weak.

that the decoder is irrelevant, and one reporting only structure concludes that the workflow is; both conclusions are properties of the summary chosen rather than of the process.

## J Controlled-Model Results

All at $V = 10^6$, $T = 10^5$, five seeds, memorizing generator unless stated.

**Power-law error under replace.** $E_n \asymp (n/T)^c$ with measured exponents $0.321, 0.468, 0.566, 0.634$ at $\beta = 1.5, 2.0, 2.5, 3.0$, against the predicted $c = 1 - 1/\beta$ of $0.333, 0.500, 0.600, 0.667$, all at $R^2 \geq 0.95$. The mechanism is a critical Galton-Watson process whose extinction time is linear in the initial count, which independently derives the empirical forgetting law of Suresh et al. (2025).

**Invariance under accumulation, convergence under subsampling.** Under full accumulation the pool support cannot change and the error is constant to machine precision, which is sharper than a bound. Under accumulate-subsample with an unbiased generator, excess error saturates by roughly generation 50 and is flat for the next 400× in generation count at four tail indices (total growth 1.07× to 1.14×); fitting $n^a$ gives $|a| \leq 0.006$ at $R^2 \leq 0.09$, and a $(\log n)^a$ fit does no better, so there is no trend of either form. This is Theorem 1 in simulation.

**Separability of the bias and tail-index effects.** Sweeping $\tau \in [0.5, 1.0]$ produces clean power laws whose exponents divided by $c(\beta)$ are constant across $\beta$ to a standard deviation of 0.005 to 0.013, over a 100× range in $T$, which is the separability Eq. (10) predicts.

**Capacity and the anchor dose-response.** Under a rank-$d$ softmax bottleneck at $\beta = 2$, missing mass grows by factors of $4.1$ to $25.3$ over $1000$ generations as $d$ falls from $256$ to $8$, monotonically for $d \geq 32$. Varying the real-data anchor at $d = 32$ over $500$ generations gives growth $19.2\times$ (no anchor), $7.3\times$ ($1\%$), $3.1\times$ ($10\%$), $1.5\times$ ($50\%$), and $0.41\times$ (full accumulation); at $d = 128$ the sequence is $6.0, 5.5, 2.9, 1.5, 0.63$. Under an unbiased memorizing generator the identical sweep gives $1.42, 1.58, 1.38, 1.27$.

## K Beam Search: Diversity, Bias and Fit Quality

From an identical generation-0 model under the replace workflow, five-way beam search yields $3.1\times$ fewer distinct bigrams than temperature-1.0 sampling (126,562 against 397,203) and gives the rarest frequency band $25\%$ of its true mass against sampling's $111\%$. Its measured bias is the largest we record for any decoder, $\hat{\delta} = 0.147$. Over ten generations the sampling arm's excess perplexity is linear in the generation index ($+9.08$ per generation, $R^2 = 0.9984$) while the beam arm is better described by a saturating fit ($R^2 = 0.941$ against $0.673$ for linear) with time constant $0.855$ generations, which is why we do not report a power-law exponent for it: beam search is not in the same regime as the sampling arms. By generation $9$ the sampling arm is the worse of the two on perplexity ($123.4$ against $83.4$) despite being the less biased decoder, which is the clearest illustration that risk and structure rank interventions oppositely.

## L Limitations in Detail

**Scope.** The largest model is 3B, the corpora are WikiText-2, WikiText-103 and TinyStories, and every arm runs ten generations, enough to separate a power law from a linear trend but not a slowly growing one from a bounded one. Seed noise is now measured directly on every statistic the paper prints rather than inherited from the exponent alone (Appendix P.3), and every difference we interpret is large relative to it; the band-crossover, causality and matched-compute arms remain single-seed. Two arms fit poorly enough that their exponents are not rates: replace at $\tau = 0.7$ ($R^2 = 0.704$) and Pythia-160m ($R^2 = 0.229$). Distinct-bigram counts saturate above $\tau \geq 1.3$, so we restrict diversity comparisons to $\tau \leq 1.15$, and the mean-field exponent prediction is validated for $\tau \leq 1$ only.

**Design choices and the confounds they introduce.** The three most overfitted conditions (Pythia-410m 5ep, OPT 10ep, Pythia-1b 2ep) run with the generation-0 fit guard disabled. The guard exists to catch accidental overfitting and these arms overfit on purpose; their generation-0 perplexities (36.3, 60.7, 38.5) all remain below their zero-shot baselines, so none is a damaged model. Training is capped at the first $15{,}000$ of about $37{,}500$ WikiText-2 blocks in article order, so "rare in the corpus" means rare in that $40\%$; band frequencies are nonetheless computed over the full split for consistency with $\hat{\delta}$, and the median count in the subset actually trained on is $1$, $4$ and $16$ for bands $1$ to $3$. The verification-filter arm returns a smaller block set and therefore also changes the pool's real-data fraction, which confounds its $0.469 \to 0.600$ result with the anchor effect.

**The residual exponent under unbiased sampling.** Across the eight well-fit $\tau = 1.0$ arms real language models give $0.41$ to $0.64$ where Theorem 1 predicts $0$, and their measured bias never exceeds $0.023$ in magnitude. The first candidate explanation was finite model capacity: we pre-registered the prediction that larger models, having a lower expressivity floor, would show smaller exponents, and recorded it before the discriminating arms reported. Qwen2.5-0.5B gave $0.545$ against OPT-125m's $0.433$, the wrong sign. The second was memorization: varying only the number of fine-tuning epochs over $1, 2, 5, 10$ gives $0.495, 0.411, 0.433, 0.409$, flat against a workflow effect of $0.65$. The third, count discounting, predicts the residual should track the fitting recipe's treatment of rare events rather than model size or epoch count, which is consistent with both falsifications but is not evidence for it.

**Model size.** The falsification above compares arms trained for 5 and 1 epochs respectively, and the epoch sweep shows the 1-epoch setting alone raises OPT's exponent to $0.495$. At matched epochs the trend reverses. We therefore cannot presently resolve the size dependence in either direction.

**The distinct-bigram ceiling.** Each generated corpus is $15{,}000$ blocks of $64$ tokens and so contains $959{,}999$ bigram positions. At $\tau \geq 1.3$ the distinct-bigram count reaches 953k to 960k, meaning 99 to $100\%$ of positions are unique. The metric is at its ceiling there, and text entirely novel at the bigram level is closer to noise than to diversity, which the reference-NLL measurements of Section 5.1 confirm directly. At $\tau \leq 1.15$ the count is 25 to $93\%$ of ceiling and the metric is informative. This is also the natural reading of why the exponent turns back up beyond $\tau^\star$.

**The mean-field branch.** Validated for $\tau \gtrsim 0.7$, Eq. (10) returns unphysical $g > 1$ at $\tau = 0.5$, most likely because it holds $W$ fixed while under strong bias the distribution itself moves. Its exponent prediction is validated for $\tau \leq 1$ only; on the $\tau > 1$ branch we rely on the controlled simulations of Appendix J.

**Seeds.** Three seeds at $\tau = 1.15$ (exponents $0.229$, $0.284$, $0.257$) and two at replace $\tau = 1.0$ ($1.083$, $1.068$) bound single-seed exponent noise at roughly $0.02$ to $0.03$. Every other arm, including the band-crossover and causality arms, is single-seed, so only effects large relative to that spread are interpretable, which we note at each use.

## M The Causality Test in Detail

Everything above is correlational in $\hat{\delta}$: the arms that vary it do so by varying the decoder, which changes more than the frequency profile. We pre-registered a direct manipulation before running it, hypotheses and analysis included (Appendix X). At every generation including generation $0$ we generate a probe corpus uncorrected, measure $\hat{\delta}$, then generate the training corpus with $-\hat{\delta} \log f$ added to the logits, which moves the level of the bias rather than its drift.

The manipulation works: at $\tau = 1.6$ it cuts $\hat{\delta}_0$ from $-0.344$ to $-0.149$ and holds the steady state at $-0.170$ against a control amplifying to $-0.814$, a $79\%$ reduction. Read causally, the V-law then predicts the exponent falls from $0.794$ to $0.432$.

**Measured exponent against the registered prediction.** It falls to $0.671$, about a third of the predicted effect. The drop is real, four times the seed spread of $0.028$; the shortfall is eight times it. So $\hat{\delta}$ is neither the operative variable nor a mere marker, but one causal channel among several, carrying roughly a third of the decoder's damage here. The same arms show where the rest is: the correction cut terminal distinct bigrams from 959k, which is $99.9\%$ of the $959{,}999$ available positions and so essentially all-unique noise, to 542k, while terminal perplexity moved only from $52.6$ to $51.9$. A large change in token statistics bought a small change in risk, which points at sequence-level incoherence that a per-token unigram tilt cannot repair.

**Discriminating power of the $\tau = 0.7$ arm.** There the manipulation cut $\hat{\delta}_0$ from $0.109$ to $0.059$ and left the exponent at $0.486$ against $0.469$. That resembles a null but is not: the V-law is flat over that interval (the natural arm at $\hat{\delta} = 0.059$ has exponent $0.473$), so the causal hypothesis predicted a change of $0.004$, far below seed noise. The arm's discriminating power was therefore too low to test H1, and no inference is drawn from it.

Consequently we do not present logit-level debiasing as a working mitigation. The V-law survives as an empirical cross-decoder predictor, and the interior optimum, the interaction, and the theorem are untouched, none of them requiring $\hat{\delta}$ to be the mechanism. The strength of its causal reading is now measured rather than assumed.

## N The Further Interventions in Detail

**The verification filter's confound is conservative.** Keeping the most probable half of generated blocks halves the synthetic influx, so the pool retains a larger real-data fraction than the control's ($0.18$ against $0.10$ at the last generation). A retained real-data anchor is the mitigation proposed in prior work for exactly this reason, and it helps the tail here, so the filter enjoys an unintended dose of it and still loses $0.146$ nats on the rare bands.

Table 6: Pearson correlation of each candidate likelihood statistic with the calibration-immune forced-choice accuracy, over the condition pairs carrying both; $n$ is the number of pairs on which each statistic is defined. $A_{\text{resid}}$ is the part of the rare-band advantage no evaluation-time unigram tilt reproduces.

| statistic | $r$ with rare-band cloze gain | $n$ |
|---|---|---|
| aggregate: terminal perplexity gain | $+0.753$ | 21 |
| aggregate: collapse-exponent gain | $+0.702$ | 21 |
| aggregate: mass-weighted band NLL | $+0.785$ | 21 |
| band-resolved: rare-band NLL advantage $A(b1\text{–}b3)$ | $+0.924$ | 21 |
| band-resolved: calibration residual $A_{\text{resid}}$ | $+0.865$ | 14 |

**The retained real-data anchor.** A retained $10\%$ real-data floor (Shumailov et al., 2024) improves terminal perplexity by $1.50$, a quarter of what tail inflation buys, for $0.048$ nats of rare-band gain, a thirteenth of it. Read on the aggregate alone the two look like the same kind of improvement; resolved by frequency they are not.

**Second-seed replication of the three non-decoder arms.** $A$(rare) and the probe reproduce in sign and in magnitude for all three: tail inflation $+0.605$ then $+0.592$, cloze $+1.24$ then $+1.01$ pp; the verification filter $-0.146$ then $-0.163$, cloze $-0.17$ then $-0.25$ pp; the real-data anchor $+0.048$ then $+0.028$, cloze $+0.17$ then $+0.14$ pp. The second seed of the control arm, run after this rule was fixed, puts the spread on this exact protocol at $0.15$ terminal perplexity, which does not change any verdict in the table. These six arms are also three of the six whole-comparison replicates that floor every printed advantage in Appendix P.3.

## O External Validity of the Band Statistic

The count-matched forced choice is expensive: it needs the candidate sets built once and then one extra gather per evaluated position, on every arm and every generation. The band decomposition costs a scatter-add. If the decomposition is to be worth reporting it should predict the probe, so we correlate the rare-band cloze difference against each candidate likelihood statistic over every condition pair carrying both.

The ordering is the one the protocol predicts: both band-resolved statistics predict the probe better than any way of taking the aggregate, the best of which is mass-weighted band NLL at $r = +0.79$, and rare-band NLL advantage $A(b1\text{–}b3)$ leads them at $r = +0.92$. The two band-resolved statistics are within $0.06$ of each other and are not defined on the same number of pairs (21 against 14), so their order between themselves is not something these data settle. These correlations are suggestive rather than decisive on their own, because the conditions of Table 1 are ordered by fit and that alone makes every statistic correlate with every other; the ranking test below is the comparison in which nothing forces agreement.

**A fit-controlled comparison.** The ten conditions of Table 1 are ordered by how overfitted the generator is, so every statistic correlates with every other and a correlation with the probe proves little on its own. The five-epoch intervention family of Table 2 does not have that problem: fit is held constant and only the intervention differs, so the arms are ordered by nothing. Ranking them by each metric and by the probe:

| intervention | terminal ppl | rank | rare-band NLL | rank | probe acc. | rank |
|---|---|---|---|---|---|---|
| tail inflation, $\tau$=1.15 | 48.17 | 1 | 6.469 | 1 | 0.8029 | 1 |
| real-data anchor, 10% | 52.21 | 2 | 6.767 | 2 | 0.7922 | 2 |
| verification filter, 50% | 53.62 | 3 | 6.919 | 4 | 0.7889 | 4 |
| none (the control) | 53.70 | 4 | 6.807 | 3 | 0.7905 | 3 |
| nucleus, top-$p$ 0.9 | 58.82 | 5 | 7.172 | 5 | 0.7879 | 5 |
| tail truncation, $\tau$=0.7 | 59.26 | 6 | 7.211 | 7 | 0.7792 | 6 |
| beam search, 5 beams | 62.02 | 7 | 7.199 | 6 | 0.7702 | 7 |

Over these seven arms the two rankings are equally close to the probe's: each reproduces it to within a single transposition, $\tau_b = 0.90$ for both. They do not fail on the same pair. The aggregate places the verification filter above the control, $0.09$ perplexity apart, inside the $0.30$ seed floor of Table 7, so its one ordering error falls on exactly the mitigation Table 2 reports it does not resolve. The rare-band loss places beam search above tail truncation at $\tau$=0.7, $0.013$ nats apart, the two arms the probe itself ranks last. All three rankings agree on the best arm. This is a seven-arm family and Kendall's $\tau$ is coarse on it; what it adds to the correlations above is that the arms here are not ordered by fit, so nothing forces any two of the three rankings to agree.

# P Robustness: The Band Definition and the Noise Floor

## P.1 The band definition

Eight geometric buckets over training counts, plus band $0$ for absent tokens, and bands $1$ to $3$ called "rare": both choices are defensible and neither is forced. This is the same two claims recomputed under every prefix of the band range as the definition of "rare", under a slope fit with band $0$ removed, and under both ways of collapsing the per-band advantages to one number. That last is a real ambiguity rather than a hypothetical: the released `band_loss.py` weights bands by validation mass, which is right for a single model's loss, while $A$(rare) averages the per-band advantages unweighted, which is right for a summary meant to speak for band 1, since band 3 carries about ten times band 1's tokens and would otherwise decide it. Generated by `experiments/band_robustness.py`.

| condition | ratio | aggregate | $b1$ | $b1$–$b2$ | $b1$–$b3$ | $b1$–$b4$ | slope (no $b0$) |
|---|---|---|---|---|---|---|---|
| SmolLM3-3B, 1 ep | 1.00 | $-0.121$ | $+0.166$ | $+0.152$ | $+0.118$ | $+0.087$ | $-0.041$ |
| Pythia-1b, 1 ep | 1.00 | $-0.113$ | $+0.324$ | $+0.276$ | $+0.204$ | $+0.150$ | $-0.065$ |
| OPT, 2 ep, 36k blk | 1.00 | $-0.097$ | $+1.039$ | $+0.679$ | $+0.462$ | $+0.342$ | $-0.121$ |
| OPT, 1 ep | 1.05 | $-0.059$ | $+0.556$ | $+0.459$ | $+0.327$ | $+0.241$ | $-0.091$ |
| OPT, 2 ep | 1.06 | $-0.068$ | $+0.726$ | $+0.574$ | $+0.410$ | $+0.304$ | $-0.105$ |
| OPT, 5 ep | 1.22 | $+0.203$ | $+1.070$ | $+0.837$ | $+0.617$ | $+0.485$ | $-0.125$ |
| Pythia-410m, 5 ep | 1.36 | $+0.567$ | $+1.239$ | $+0.979$ | $+0.806$ | $+0.699$ | $-0.125$ |
| OPT, 10 ep | 1.70 | $+0.372$ | $+1.508$ | $+1.148$ | $+0.883$ | $+0.726$ | $-0.145$ |
| Pythia-1b, 2 ep | 1.75 | $+0.178$ | $+1.124$ | $+0.990$ | $+0.793$ | $+0.640$ | $-0.172$ |
| SmolLM3-3B, 3 ep | 1.84 | $+0.601$ | $+0.667$ | $+0.678$ | $+0.666$ | $+0.662$ | $-0.077$ |
| positive / disagreeing | | | 10, 5 | 10, 5 | 10, 5 | 10, 5 | 10 neg |

Under all four definitions, and under either weighting, the rare-band advantage is positive in $10/10$ conditions and disagrees in sign with the aggregate in $5/10$, and the slope against band index is negative in $10/10$ with band $0$ in the fit and $10/10$ with it removed. The magnitudes move as expected, since a narrower group is more concentrated on the rarest tokens where the effect is largest; neither conclusion moves.

## P.2 A Negative Control: Public Pretraining Checkpoints

Whether the sign disagreement is a property of recursive training or of held-out loss is testable: if the metric alone produced it, it would appear in the most ordinary comparison there is, one pretrained checkpoint against a later one. It does not.

We ran the released `band_loss.py` unmodified on $5$ public Pythia-410m checkpoints from step $1{,}000$ to step $143{,}000$, which share a tokenizer and a pretraining corpus and differ only in how far training has gone (`experiments/pretrained_bands.py`). One aspect differs from the rest of the paper: a band is assigned everywhere else by an item's count in the model's own training corpus, because Theorem 1 endangers an item by its expected training count, and a public checkpoint's corpus is the Pile. Bands here come from counts in a 13.3M-token sample of the evaluation corpus instead, which changes a band's meaning from "rare in what the model saw" to "rare in the domain being evaluated." That is the question a practitioner comparing checkpoints on their own corpus has, but it is not the same quantity and we draw no conclusion about Theorem 1 from it.

Perplexity falls from $400.3$ to $58.0$ and rare-band NLL from $9.04$ to $5.95$ over the run, and the two statistics agree in all $4$ consecutive pairs: $0$ of $4$. They also saturate together, the aggregate having

banked $89\%$ of its total gain by step $16{,}000$ and the rare bands $85\%$ of theirs. The disagreement this paper reports is therefore not what happens whenever a weighted average is summarised. It needs an intervention whose effect on the head and on the tail have opposite signs, which is what the recursive-training interventions of Sections 5.1 and 5.2 turn out to be and what more training is not. The practical rule is therefore: report the decomposition when comparing interventions that reshape the frequency distribution, which is what recursive-training work does throughout, and to expect it to add nothing when comparing a model against a strictly better one.

One measurement here bears comparison with Section 3.2. The three rarest bands carry $8.4\%$ of the validation mass under this 13.3M-token banding corpus, against $19.7\%$ under WikiText-2's 2.4M, with a different tokenizer and a different corpus: an independent instance of the decay Proposition 1 predicts.

### P.3 Measured Noise Floors and Net Sensitivity

Proposition 2 leaves one quantity unbounded: the noise cost $\sigma_R/\sigma$. It is a property of the model and the evaluation set, not of the corpus, so it has to be measured, and this appendix measures it. Every number here comes from `experiments/detection_floor.py`.

**Two noise floors and their estimators.** 7 arms exist twice in this study, identical in every configuration field except `--seed`. Their difference is training noise with the data, the recipe and the evaluation set all held fixed. Pooling them with the paired-replicate estimator $\hat{\sigma} = \sqrt{\overline{d^2}/2}$ gives a seed standard deviation of $0.0196$ nats on aggregate held-out NLL ($1.16$ perplexity units) and $0.0263$ nats on the rare-band mean, a noise cost of $1.34\times$. Resolved by family the picture changes sharply: on OPT-125m the aggregate is reproducible to $0.0024$ nats, $0.10$ perplexity units, against $0.0247$ nats on the rare bands, a noise cost of $10.48\times$. A second and more directly relevant floor comes from $6$ comparisons re-run end to end under a different seed, which floors the printed advantage rather than one arm's loss; the two arms of a comparison share a generation-0 model and an evaluation set, so this floor cannot be recovered from the single-arm one. Table 7 gives it for every statistic the paper prints.

Table 7: Seed floor on every statistic the paper prints, from the 6 comparisons run twice end to end under different training seeds. $\sigma$ is the paired-replicate estimate $\sqrt{\overline{d^2}/2}$ and the floor is $1.96\sigma$; these are already floors on a difference of two arms, so no further $\sqrt{2}$ applies. Terminal perplexity is not scale free, so its pooled floor is dominated by the highest-perplexity family and the OPT-125m column is the one that governs Table 2.

| statistic | $\sigma$ (pooled) | floor (pooled) | $\sigma$ (OPT) | floor (OPT) |
|---|---|---|---|---|
| $A$(rare, $b1$–$b3$) | 0.0128 | 0.0250 | 0.0124 | 0.0243 |
| $A$(common, $b6$–$b8$) | 0.0152 | 0.0299 | 0.0029 | 0.0057 |
| slope of $A$ against band | 0.0041 | 0.0079 | 0.0025 | 0.0048 |
| aggregate NLL advantage | 0.0131 | 0.0256 | 0.0033 | 0.0065 |
| collapse exponent advantage | 0.0204 | 0.0400 | 0.0222 | 0.0435 |
| terminal perplexity advantage | 0.92 | 1.80 | 0.15 | 0.30 |

Table 8: Proposition 2 against training budget. Leverage is $1/M(m,T)$ at $m = 100$ with the masses measured in Appendix Q; the noise cost is held at its measured value, which is the assumption the projection rests on and the one the WikiText-103 pair tests. The net gain crosses one at 7.1M tokens on OPT-125m and 1.0M pooled.

| budget $T$ | $M(100,T)$ | leverage $1/M$ | net, pooled | net, OPT-125m |
|---|---|---|---|---|
| 2,400,000 | 0.2276 | 4.4× | 3.3× | 0.4× |
| 10,000,000 | 0.0918 | 10.9× | 8.1× | 1.0× |
| 40,000,000 | 0.0210 | 47.5× | 35.4× | 4.5× |
| 100,000,000 | 0.0043 | 230.2× | 171.3× | 22.0× |

**Net gain at WikiText-2 scale.** WikiText-2's three rarest bands carry $19.7\%$ of the validation mass under the OPT tokenizer of Table 9, a leverage of $5.07\times$, and $19.4\%$ pooled over the replicates' two tokenizers, a leverage of $5.15\times$; bands are vocabulary items, so each leverage is paired only with the noise cost measured on the same arms. Against the pooled noise cost the net gain is $3.84\times$, but against OPT-125m's it is $0.48\times$: at this corpus size, on this model, the decomposition is not the more powerful instrument, because OPT-125m's aggregate is unusually reproducible across seeds and its rare bands are not. The crossover is at 7.1M training tokens (1.0M pooled) and WikiText-2's 2.4M sits below it. This is the correct reading of Proposition 2 here: parts (a)–(b) are a prediction for corpora larger than ours, and what our experiments exhibit is part (c), masking, which is a bias and which a reproducible aggregate makes worse rather than better. Table 8 carries the measured noise ratio forward against the measured masses.

**Verdicts under the measured floor.** Applying the advantage floors of Table 7 to the ten conditions of Table 1: all 10 rare-band advantages and all 10 slopes clear it and the 5 sign disagreements are exactly the 5 whose exponent clears its own floor with the opposite sign, those same 5 being the ones whose terminal perplexity clears no floor at all. The agreement is exact and it is not built in, because the rare-band floor and the aggregate floor are estimated from different columns of the same replicates. The 0.6-perplexity spread Table 2 calls *silent* sits between the pooled floor of 1.80 and the OPT-125m floor of 0.30, so it is conservative against the family that matters and generous against the pooled one; no verdict in that table changes under either.

**Measurement of the projection's fixed-noise assumption.** The projection varies the mass and holds both noise standard deviations fixed, which was an assumption until a second WikiText-103 control seed measured it. At four times the training corpus, on the same model and the byte-identical validation split, the three rarest bands carry $11.7\%$ against $19.7\%$ and the noise cost falls from $10.48\times$ to $7.29\times$. Holding it fixed is therefore conservative rather than optimistic, and the net gain crosses one between two real corpora rather than only in projection: $0.48\times$ at WikiText-2's leverage of $5.07\times$ against $1.18\times$ at WikiText-103's $8.58\times$.

**Precision of these estimates.** Two points fix a direction, not a curve, so the crossover budget should be read as an order of magnitude. The floors themselves rest on 7 single-arm replicates and 6 whole-comparison replicates, enough to separate 0.02 nats from 0.6 but not enough to put a confidence interval on a variance.

## Q Band Weights Across Corpora, Tokenizers and Corpus Sizes

Table 9: Validation token mass by frequency band; band $b$ holds tokens of corpus count in $[e_b, e_{b+1})$ for edges $1, 4, 18, 78, 336, 1440, 6171, 26429$. Raw counts follow.

| band | 0 (unseen) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| weight $w_b$ | 0.4% | 1.3% | 6.0% | 12.4% | 17.2% | 15.5% | 8.2% | 15.4% | 23.7% |

**The band weights of Table 9 in raw counts.** Of the 245,826 WikiText-2 validation positions, bands 0 to 8 hold 946, 3,195, 14,769, 30,557, 42,165, 38,033, 20,119, 37,874 and 58,168 tokens. Band 0 is tokens absent from the training corpus and is reported separately throughout, because it measures retention of the pretrained prior rather than corpus knowledge.

**Corpus-scale test: registered predictions and outcomes.** P1, the band-weight manipulation check, held: the three rarest bands carry $11.7\%$ of the identical validation split here against $19.7\%$ under WikiText-2's own banding, below the $19.9\%$ the registration fixed. P2 held ($A$(rare) $+0.657$, slope $-0.156$). P3, the sign disagreement, held (collapse exponent $-0.040$, terminal perplexity $-0.64$). P4, concentration at least as strong as WikiText-2's $3.0\times$, held at $8.5\times$. P5, the cloze direction, held ($+1.27$ pp, CI $[+0.98, +1.56]$). The design and all five predictions were fixed before either arm launched (`docs/prereg-corpus-scale.md`), including the note that P4 was the one expected to be at risk because the band edges widen with the corpus. WikiText-2 and WikiText-

103 share a byte-identical validation split, so this condition evaluates on the same 3,902 blocks as every other arm in the paper and the only thing that differs is the training corpus.

**Empirical validation of Proposition 1.** The proposition is about an absolute count window, not a band index, because Theorem 1 endangers an item according to its expected count alone while the geometric bands widen in count as the corpus grows. Tokenizing WikiText-103 at four training budgets under the OPT tokenizer, against its own fixed 249,749-token validation split (`experiments/theory_weights.py`):

| train tokens | types seen | $M(\leq 1, T)$ | $M(\leq 3, T)$ | $M(\leq 10, T)$ | $M(\leq 100, T)$ | fitted $\beta$ (tail) |
|---|---|---|---|---|---|---|
| 2.4M | 36,963 | 0.474% | 1.308% | 4.727% | 22.762% | 1.681 |
| 10.0M | 42,557 | 0.037%$^\dagger$ | 0.145% | 0.712% | 9.184% | 1.816 |
| 40.0M | 45,371 | 0.004%$^\dagger$ | 0.009%$^\dagger$ | 0.079%$^\dagger$ | 2.103% | 1.948 |
| 100.0M | 46,666 | 0.002%$^\dagger$ | 0.005%$^\dagger$ | 0.014%$^\dagger$ | 0.434% | 2.047 |

Cells marked † rest on fewer than 300 validation tokens and are excluded from the budget fits, the same precaution Table 11 takes for single-band ratios. Over the windows that survive that filter, $M(\leq m, T)$ is a clean power law in $m$ (exponent 0.86 at the smallest budget, $R^2 = 0.989$) and in $T$ (exponent 1.04 at $m \leq 100$, $R^2 = 0.960$). Both exceed the $c < 1$ that Equation (2) can produce for any finite $\beta$, and the last column says why: a subword vocabulary is finite, so its fitted tail index rises with corpus size and a single exponent understates the decay. We therefore treat Proposition 1 as a conservative bound and report the measurement beside it.

Generated by `experiments/band_weights_universality.py`. Band construction is the one used everywhere else in the paper: eight geometric buckets over training-corpus counts plus band 0 for tokens absent from training. The training budget is held fixed at 2.4M tokens in Table 10, because the weights genuinely depend on corpus size and comparing a 2M-token corpus against a 100M-token one would confound language with scale; the size effect is Table 11. GPT-2 is omitted from Table 10 because it shares OPT's BPE vocabulary exactly and reproduces its numbers to the digit, which is a useful check that the pipeline is deterministic.

Table 10: Ratio of the three commonest bands' validation mass to the three rarest, at a fixed 2.4M-token training budget. Our setting (WikiText-2 under OPT) is 2.35, the median of the 25 corpus/tokenizer pairs measured. TinyStories is a deliberately simple corpus and gives the concentrated extreme.

| corpus | OPT-125m | Pythia | Qwen2.5 | SmolLM3 |
|---|---|---|---|---|
| WikiText-2 | 2.35 | 2.44 | 2.86 | 2.50 |
| WikiText-103 | 2.35 | 2.45 | 2.86 | 2.50 |
| C4 (web) | 2.21 | 2.24 | 2.43 | 2.23 |
| Code (Python) | 3.92 | 4.15 | 2.90 | 2.78 |
| TinyStories | 13.08 | 13.07 | 13.73 | 13.73 |

Table 11: The weighting sharpens as the training corpus grows, with the evaluation held fixed at 247,289 WikiText-103 validation tokens under the OPT tokenizer. The last column is the number of validation tokens falling in bands 1 to 3. At 100M training tokens the rarest band alone contains 14 validation tokens, so single-band ratios are not reported at that budget; the group figures rest on 6,708 tokens.

| train tokens | types seen | bands 1–3 | band 8 | ratio | val tokens in $b1$–$b3$ |
|---|---|---|---|---|---|
| 2.4M | 37,005 | 19.90% | 23.89% | 2.35 | 49,202 |
| 10.0M | 42,594 | 11.23% | 27.92% | 4.67 | 27,763 |
| 40.0M | 45,387 | 5.43% | 29.63% | 10.73 | 13,421 |
| 100.0M | 46,674 | 2.71% | 31.92% | 22.75 | 6,708 |

## R THE BENCHMARK PROBE IN DETAIL

Length-normalised multiple-choice accuracy on Pythia-410m at five epochs, tracked across all ten generations. Length normalisation matters here specifically: the intervention changes how much probability mass sits on rare tokens, and an unnormalised score would confound that with the capability being measured. Both arms lose capability, by $1.7$ to $5.2$ accuracy points, and they lose it at statistically indistinguishable rates. With $400$ items per benchmark the standard error on a difference is about $0.034$, so we can exclude a capability advantage larger than roughly $7$ accuracy points but not a smaller one.

| change over nine generations | $\tau = 1.0$ | $\tau = 1.15$ | difference |
|---|---|---|---|
| ARC-Easy | $-0.027$ | $-0.017$ | $+0.010$ |
| PIQA | $-0.052$ | $-0.052$ | $0.000$ |
| HellaSwag | $-0.045$ | $-0.043$ | $+0.002$ |

This is the probe that returned a null, on the same arms and the same checkpoints where the count-matched forced choice of Section 5.4 separates the arms by $+1.95$ percentage points with a confidence interval excluding zero.

## S PRE-REGISTRATION OF THE CLOZE PROBE

Registered before any `cz_` arm was launched. The probe's difficulty ($15$ distractors, rank window $\pm 16$) was fixed on the zero-shot model over the first $600$ validation blocks, with no arm comparison available; a difficulty chosen after seeing arm results would be a researcher degree of freedom.

**Design, fixed in advance.** Candidate sets are drawn once with seed $12345$, independent of the training seed, so they are byte-identical across arms and generations. Primary metric: top-1 accuracy over bands $1$ to $3$ pooled at generation $9$. Uncertainty: paired bootstrap over validation blocks, $10{,}000$ resamples, the same resampled indices applied to both arms, because positions are not independent. Four conditions, all reported whatever they show, and no arm re-run.

**Realised match quality.** Recorded per band in every output file, the exact-count-match fraction is $1.000$, $0.995$, $0.962$ and $0.724$ for bands $0$ to $3$, and the mean absolute log-count ratio between distractors and the true token is $0.0000$, $0.0007$, $0.0009$ and $0.0017$. Exact matching is impossible in the head for any scheme, and there the ratio rises to $1.24$ at band $8$.

**Outcomes against the registration.**

| prediction | registered | measured | verdict |
|---|---|---|---|
| P1a direction, CI | $> 0$, CI excludes $0$ | $+1.95$ pp, $[+1.66, +2.23]$ | pass |
| P1b magnitude | $+2.0$ to $+8.0$ pp | $+1.95$ pp | **failed** by $0.05$ pp |
| P2 negative band slope | $< 0$ | $-0.00404$ | pass |
| P3 sign disagreement | $> +0.5$ pp, CI excludes $0$ | $+0.59$ pp, $[+0.38, +0.80]$ | pass |
| P4 noise control | noise buys $<$ half | $0.21\times$ | pass |
| P5 calibration immunity | shift $< 0.001$ | $0.000068$ | pass |

P1b is recorded as a failure because that is what was registered, though it misses by $0.05$ pp. The derivation behind it mapped the $0.883$-nat rare-band NLL advantage through a softmax argument and was about four times too optimistic; most of an NLL gain on a rare target is mass moved across the whole vocabulary, to which a count-matched forced choice is almost entirely immune. That is the same property that makes the probe calibration-immune, so the small effect size is the design working rather than the effect being weak.

## T THE OVERFITTING MODERATOR IN DETAIL

**Aggregate gain against generation-0 perplexity.** Across the four OPT-125m conditions the exponent gain runs $-0.059$, $-0.068$, $+0.203$ and $+0.372$ at generation-0 perplexities of 37.4, 37.7, 43.3 and 60.7, and the two 3B conditions extend it to another family and another order of magnitude in size: $-0.121$ at ratio 1.00 against $+0.601$ at ratio 1.84, the same model and corpus with only the fit changed.

**The dose-response.** At 10 epochs the effect is large enough to change the qualitative picture: the $\tau = 1.0$ run degrades from 60.7 to 95.3 perplexity over nine generations while $\tau = 1.15$ runs $60.7 \to 70.3$ and then flattens, giving an exponent of 0.037 against 0.409. Collapse is very nearly arrested by a decoder change.

**Matched compute.** Holding total gradient presentations essentially fixed at $\approx$ 72k, we compare 15,000 blocks for 5 epochs against 36,000 blocks for 2 epochs, the latter seeing $2.4\times$ more unique text. The larger-data arm does not overfit, reaching the best generation-0 perplexity of any OPT-125m arm in the paper (35.6), and its exponents are 0.574, 0.419, 0.516, 0.599 at $\tau = 0.85, 1.0, 1.15, 1.3$: minimum at $\tau = 1.0$, gain $-0.097$. Matched compute without overfitting gives no aggregate benefit.

**Cross-family.** Pythia-1b at its own selected recipe (one epoch) shows no aggregate optimum, with exponents 0.770, 0.530, 0.648, 0.592 at $\tau = 0.85, 1.0, 1.15, 1.3$ and a two-seed spread of at most 0.004, so the null is not noise. Pushed into the overfitted regime, Pythia-410m at five epochs (generation-0 perplexity 36.3 against 26.8 at one epoch) gives 0.810 at $\tau = 1.0$ and 0.243 at $\tau = 1.15$, a gain of $+0.567$, the largest effect we measure anywhere.

## U PRE-REGISTRATION OF THE 1B OUT-OF-SAMPLE TEST

Registered before `p1b_of_t1.0` and `p1b_of_t1.15` were launched, when no overfitted 1B arm existed. The recipe (two epochs, $\mathrm{lr} = 5 \times 10^{-5}$) was taken from a hyperparameter sweep already in the repository, run for an unrelated purpose, so it was not chosen with this test in mind. It replaces a 3B scaling run we abandoned: at 3B the usable window between not overfitting at all (one epoch, ratio 1.00) and a damaged model (four epochs, validation perplexity 61.86 against a 41.25 zero-shot baseline) is narrow and costs about 20 GPU-hours per arm.

**Manipulation check, specified in advance.** Generation-0 validation perplexity divided by the best achieved for this model and corpus should give an overfit ratio near 1.75. Measured: $38.47/22.04 = 1.75$, the highest in the paper.

**Predictions, and outcomes.**

| prediction | registered | measured | verdict |
|---|---|---|---|
| aggregate gain | $+0.25$ to $+0.45$ | $+0.178$ | **failed**, below range |
| rare-band $A(b1$–$b3)$ | $> +0.60$ | $+0.793$ | confirmed |
| all nine bands positive | $9/9$ | $7/9$ | **failed** |
| fit-free excess-loss ratio | $> 2\times$ | $1.43\times$ | **failed** |

Exponents are 0.433 ($\tau = 1.0$, $R^2 = 0.966$) against 0.255 ($\tau = 1.15$, $R^2 = 0.673$). Because the treated arm falls below the $R^2 \geq 0.75$ threshold fixed in advance, the trajectory is the primary report: $38.5 \to 57.4$ against $38.5 \to 50.9$, a 6.5-point terminal improvement.

**Scope of the conclusion.** The moderator's direction is confirmed out of sample at five times the model size carrying the main result, and the rare-band concentration is strong (slope $-0.156$). Its magnitude is about half what the OPT anchors predicted. We do not report this as confirming the moderator. Across all ten conditions the rare-band benefit is positive in $10/10$ and the slope negative in $10/10$, while the overfit ratio and the aggregate gain correlate at $+0.806$ across those ten: the band-level regularity is the invariant, and it is the one that did not depend on this registration.

# V PRE-REGISTRATION OF THE 3B SCALING TEST

Registered before any 3B arm with more than one generation existed, and amended twice, both times before the arms it governs were launched. Both amendments are reported below.

**Falsification of the first registration's premise.** The prediction was that a 3B model fine-tuned on the 2M tokens of WikiText-2 would overfit at the smallest recipe we could run, on the reasoning that capacity relative to data drives overfitting. The fit check falsified that premise directly: at one epoch and $\mathrm{lr} = 5 \times 10^{-6}$ SmolLM3-3B reaches validation perplexity $16.93$ against a $41.25$ zero-shot baseline, the best fit of any model in this paper, so its overfit ratio is $1.00$ and the conditional branch predicting a positive aggregate gain was never triggered. A larger pretrained model generalizes better from a small fine-tuning set, and at one epoch that outruns memorization even though the train-validation gap of $+0.678$ nats shows memorization occurring.

**Predictions, fixed before `s3b_t1.0` and `s3b_t1.15` launched.** Under the reframed paper the well-fit branch is the more valuable one, because every condition in which the aggregate and the tail disagree has an overfit ratio between $1.00$ and $1.06$, so a well-fit 3B model is predicted to be a disagreeing condition and would be the first above 1B.

| prediction | registered | measured | verdict |
|---|---|---|---|
| P1 aggregate gain | near zero or negative | $-0.121$ ($0.431$ vs $0.553$) | pass |
| P2 rare band, slope | $> 0$, slope $< 0$ | $+0.118$, slope $-0.042$ | pass |
| P3 a ninth disagreeing condition | aggregate vs tail | opposite signs | pass |
| P4 cloze rare-band accuracy | $>$ $+0.2$ pp, CI excludes 0 | $-0.34$ pp, $[-0.49, -0.20]$ | **failed**, sign reversed |
| P5 band weights, 128k vocabulary | reported, no prediction | $b8/b1 = 17.4$, group $2.53$ | as elsewhere |

Manipulation check: generation-0 validation perplexity $16.91$ against the fit check's $16.93$, so the recipe reproduced. The weights under SmolLM3's 128k vocabulary match the universality sweep's independent measurement for the same tokenizer ($17.1$ and $2.50$).

**Interpretation of the sign reversal.** P4 is the paper's only sign-reversed registered prediction and Section 5.4 reports it as one. It is also the outcome the calibration control predicts: that control finds a unigram tilt reproducing up to $93\%$ of the rare-band NLL advantage in well-fit models, and this is the best-fit model in the paper. A probe immune to calibration by construction should therefore find little here, and it finds nothing. We record that this reading was written down after seeing the result, unlike P1 to P5, and Section 5.4 is the out-of-sample test of it. The measurement claim is unaffected: P1 to P3 are the measurement claim, and they hold.

**Amendment 4: the overfitted branch.** The first registration's conditional prediction applied at an overfit ratio above $1.2$, which no 3B recipe we could afford had reached. Amendment 4 fixed four predictions and the recipe-selection rule, the highest overfit ratio still beating the $41.25$ zero-shot guard, chosen by `experiments/fit_check_3b.py` rather than by inspection, before any arm ran. Amendment 5 then recorded the operating conditions, again before any outcome existed, including that the machine turned out to be shared contrary to Amendment 4's assumption. The sweep selected 3 epochs at $\mathrm{lr} = 5 \times 10^{-6}$, the highest overfit ratio beating the guard, giving generation-0 perplexity $31.09$ and a ratio of $1.84$. P1, aggregate gain at least $+0.15$, held at $+0.601$. P2, rare band positive with negative slope, held ($+0.666$, slope $-0.068$). P3, the discriminating one, cloze above $+0.5$ pp with a CI excluding zero, held ($+1.28$ pp, CI $[+1.03, +1.52]$). P4, the best tilt reproducing under $60\%$, held (best tilt reproduces $24\%$, residual $+0.509$ nats). 4 of 4 held.

# W PRE-REGISTRATION OF THE BAND-CROSSOVER TEST

Registered while the arms were at generation 2 of 9; the discriminating comparison is generation $0$ against generation $9$ and did not exist at the time of writing.

**Hypotheses.** $H_{\text{reg}}$: the advantage $A(b)$ of $\tau = 1.15$ over $\tau = 1.0$ is flat in band index, since label smoothing lifts likelihood roughly uniformly. $H_{\text{attr}}$: $A(b)$ is positive in rare bands and decreases with band index, approaching zero or turning negative in the commonest, because attrition removes rare items and inflation replaces them. The controlled model motivated $H_{\text{attr}}$: at $\epsilon = 0.01$, $\lambda = 0.6$, $\beta = 2$, per-band cross-entropy excess at $\tau = 1.15$ has slope $-0.146$ against band index, with rare bands improving and common bands degrading: over matched quartiles of the 35 populated bands the means are $+1.79$ and $-1.03$.

**Outcome.** $H_{\text{attr}}$ confirmed on the primary test: slope $-0.161$, rare mean $+0.617$ against common $+0.043$, a ratio of $14.25\times$, and a slope that deepens monotonically from generation 1 to 8 ($-0.050$ to $-0.166$) before flattening at generation 9 ($-0.161$). $H_{\text{reg}}$ is falsified.

**Departures from the registered prediction.** The registration allowed for the advantage "turning negative" in the commonest bands. It does, but only transiently: the commonest band is $-0.026$, $-0.018$, $-0.013$ at generations 1 to 3 and then crosses back, ending at $+0.028$. At generation 9 tail inflation is weakly better even for common tokens, so the correct claim is concentration, not sign reversal. The controlled model does show a terminal sign reversal and the language model does not; the discrepancy is reported at generation 9, the registered endpoint. A secondary prediction, that the logit correction would leave ref-NLL near its control, is falsified outright: it improves $\tau = 1.6$ from 12.35 to 10.41.

## X Pre-registration of the Causality Test

Recorded before the `pd_t0.7` and `pd_t1.6` arms were launched, and reproduced here in substance.

**Manipulation.** Two-pass logit correction applied at every generation including generation 0: generate a probe corpus with no correction, measure $\hat{\delta}$, then generate the training corpus with $-\hat{\delta} \log f$ added to the logits. Controls are `acc_t0.7` and `acc_t1.6`, matched in every other respect.

**Manipulation check, specified in advance.** $|\hat{\delta}_0|$ in the treated arm must be at most half the control's, i.e. below 0.055 at $\tau = 0.7$ and below 0.172 at $\tau = 1.6$. Outcome: $\tau = 1.6$ passes at 0.149; $\tau = 0.7$ reaches 0.0587, a 46% reduction against the 50% required, and marginally fails.

**Hypotheses, specified in advance.** H1, $\hat{\delta}$ is causal: driving it toward zero moves the exponent to the value observed at the naturally unbiased setting, 0.433, predicting 0.35 to 0.55 for `pd_t1.6` against a control of 0.794. H2, $\hat{\delta}$ is a marker: the exponent stays above 0.70. We recorded in advance that we expected H2 at $\tau = 1.6$, on the grounds that the bigram metric saturates there and text which is 100% novel at the bigram level is noise whose damage a unigram tilt cannot repair.

**Analysis, specified in advance.** Exponent by the same estimator as every other arm, single seed matching the controls, no re-running and re-reporting. Interpretable effect sizes were bounded below by the observed seed spread of 0.02 to 0.03.

**Outcome.** Neither hypothesis as stated. 0.671, which is 34% of the predicted movement: the drop is $4\times$ the seed spread, and the shortfall from H1 is $8\times$ it. The $\tau = 0.7$ arm was found post hoc to have no discriminating power, since the V-law is flat over the interval it manipulated and H1 predicted a change of 0.004 there. The arm was not diagnostic as designed, and is reported unchanged rather than re-run.

## Y Controlled-Model Details for the Interior Optimum

**Smoothing and count discounting: dose-response.** Both effects named in Appendix J are monotone in the discount $\lambda$. At smoothing $\epsilon = 0.003$, raising $\lambda$ through $0, 0.3, 0.6, 1.0$ moves $\hat{\delta}$ at $\tau = 1$ through $-0.035, -0.009, +0.027, +0.226$ and the exponent at $\tau = 1$ through $0.007, 0.050, 0.091, 0.110$. A faithful decoder on a discounting learner is a biased generator,

and that residual suppression is what mild inflation cancels. Across all 24 conditions ($\epsilon \in \{0.003, 0.01, 0.03\}$, $\lambda \in \{0, 0.3, 0.6, 1.0\}$, $\beta \in \{2.0, 2.5\}$) the optimum sits at negative measured bias in 23, median $\hat{\delta}^\star = -0.087$. Individual conditions range from $-0.298$ to $+0.051$ and the exponent is flat near its minimum, so the optimum is identified only to the first decimal.

**The likelihood channel and the head-dominance hypothesis.** In the controlled model the coverage channel (missing mass) shows the interior optimum and the likelihood channel (cross-entropy of the source under the model) does not: excess cross-entropy after 500 generations is minimized at $\tau = 1.0$ for every combination of smoothing $\epsilon \in \{0.003, 0.01, 0.03\}$, discount $\lambda \in \{0, 0.3, 0.6, 1.0\}$ and tail index we ran. The language model behaves differently, with held-out perplexity minimized at $\tau = 1.15$.

The first candidate explanation was head dominance. A global temperature flattens the head along with the tail, and cross-entropy is mass-weighted, so at $\beta = 2.0$, where the top item alone carries $61\%$ of the mass and the top ten carry $94\%$, head distortion should swamp any tail benefit. Natural language is not like that: unigram frequencies are Zipfian with exponent near 1.

We tested this directly by rerunning at $\beta \in \{1.1, 1.3\}$, where the top item carries $12.4\%$ and $25.8\%$ of the mass and the unigram entropy is $7.55$ and $4.75$ nats, the former comparable to WikiText-2 (`experiments/results_lowbeta`). The prediction is not borne out. The coverage optimum survives and remains at negative measured bias in both, though its precise location is not identified, since the exponent is flat near the minimum and the argmin moves between adjacent $\tau$ across reruns. Cross-entropy, however, is still minimized at $\tau = 1.0$ in both ($\Delta\mathrm{CE} = 0.374$ and $0.172$, rising monotonically on both sides). We therefore reject this explanation.

No replacement explanation is established here. The most likely missing ingredient is sequential structure: the controlled model is a single global multinomial, while a language model's perplexity is a statement about conditionals, and a per-token temperature does not act on the corpus distribution the way a global temperature does. A sequential treatment is left to future work, and the discrepancy is reported alongside the coverage-channel agreement.

## Z Complete Results

Table 12 lists every distinct language-model run in the paper. It is generated directly from the result files by `experiments/make_table.py` so that it cannot drift from the data. "interv" records any intervention beyond the decoder: `keepreal` retains a fraction of real data in every training set, `verify` keeps only the most probable fraction of generated blocks, `blockfix` reweights generated blocks toward the real frequency profile, `logitfix` applies a fixed-offset correction at the logits, `noise` replaces a fraction of generated tokens with context-free draws from the real unigram distribution, and `probefix` is the two-pass probe-measured correction of Appendix M. "corpus" is WikiText-2 unless stated. Arms differing only in seed are listed separately.

**Duplicate runs.** 94 result files satisfy the five-generation minimum but only 84 are distinct runs. The pipeline is deterministic, so an arm re-run later with extra instrumentation reproduces its original to every digit and is the same run recorded twice; `make_table.py` detects those automatically by trajectory signature and the released `docs/RESULTS_TABLE.md` marks each with the arm it duplicates. Their existence is a check on the determinism claim of Section 5 rather than additional evidence, and the cross-stack re-run of Appendix C is excluded by name because its trajectory differs in the fourth decimal and the signature detector cannot see it.

**Fit quality.** $R^2$ is reported for every arm and we flag rather than drop poor fits. Five arms have $R^2 < 0.75$ and their exponents should not be read as rates: `p410ep5_t1.15` (0.288), `ep10_t1.15` (0.373), `ep10_t1.3` (0.603), `rep_t0.7` (0.704) and `scale_pythia-160m_t1.0` (0.229). The first three matter for Section 5.1 and are discussed there: a trajectory that flattens produces a small excess with little dynamic range, which the power-law estimator fits poorly by construction, so for those arms we report the perplexity trajectory as the primary evidence and the exponent as a summary.

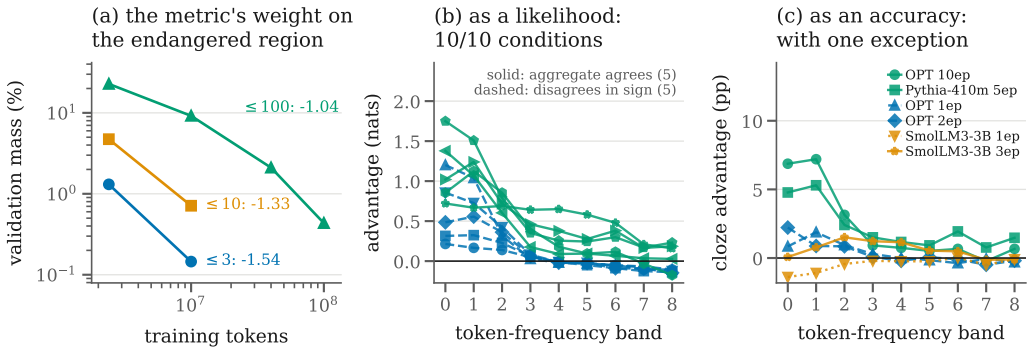


Figure 4: **(a)** Proposition 1 measured on WikiText-103, evaluation held fixed; each line is one count window, labelled with its fitted slope in $T$. Points resting on fewer than 300 validation tokens are dropped, which is why the two narrow windows stop at $10^7$, and every slope is steeper than the $c < 1$ the proposition guarantees. **(b)** Advantage of $\tau = 1.15$ over $\tau = 1.0$ by band as a likelihood, all ten conditions: decreasing in $10/10$, dashed where the aggregate reports the opposite sign to the tail. **(c)** The same as a count-matched accuracy; in both dashed conditions perplexity prefers $\tau = 1.0$ while the accuracy prefers $\tau = 1.15$. The one exception is the dotted best-fit condition, where the calibration control predicts no effect, and the same model overfitted is the solid line of the same colour that recovers it.

**The per-band result across all ten conditions.** Figure 4 gives the per-band advantage for every condition of Table 1, of which Figure 1b is one row, with the measured validation-mass decay of Proposition 1 and the count-matched accuracy.

Table 12: All 84 distinct language-model runs, generated directly from the result files by `experiments/make_table.py`. "$a$" is the fitted collapse exponent, "div" the percentage change in distinct bigrams, $q/f$ the final mass ratio on the rarest frequency band, and $\hat{\delta}_0$ the measured generator bias at generation 0. Arms with $R^2 < 0.75$ are discussed in the text and their exponents should not be read as rates.

| arm | model | ep | workflow | dec | interv | corp | $\text{ppl}_0$ | $\text{ppl}_N$ | $a$ | $R^2$ | div% | $q/f$ | $\hat{\delta}_0$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| qv_t0.7 | Qwen2.5-0.5B | 1 | acc-sub | tau0.7 | - | wt2 | 23.8 | 27.1 | 0.604 | 0.980 | -53.9 | 0.272 | +0.1028 |
| qwen_acc_t1.0 | Qwen2.5-0.5B | 1 | acc-sub | tau1 | - | wt2 | 23.8 | 25.0 | 0.545 | 0.970 | +2.3 | 1.192 | -0.0063 |
| qv_t1.3 | Qwen2.5-0.5B | 1 | acc-sub | tau1.3 | - | wt2 | 23.8 | 26.3 | 0.583 | 0.989 | +40.9 | 10.235 | -0.1778 |
| s3b_of_t1.0 | SmolLM3-3B | 3 | acc-sub | tau1 | - | wt2 | 31.1 | 64.2 | 0.544 | 0.994 | -0.5 | 1.849 | -0.0011 |
| s3b_t1.0 | SmolLM3-3B | 1 | acc-sub | tau1 | - | wt2 | 16.9 | 17.7 | 0.431 | 0.976 | +5.1 | 1.417 | -0.0110 |
| s3b_tilt_t1.0 | SmolLM3-3B | 1 | acc-sub | tau1 | - | wt2 | 16.9 | 17.7 | 0.391 | 0.973 | +5.6 | 1.420 | -0.0116 |
| s3b_of_t1.15 | SmolLM3-3B | 3 | acc-sub | tau1.15 | - | wt2 | 31.1 | 41.3 | -0.057 | 0.076 | +117.8 | 11.211 | -0.0335 |
| s3b_t1.15 | SmolLM3-3B | 1 | acc-sub | tau1.15 | - | wt2 | 16.9 | 18.3 | 0.553 | 0.991 | +88.6 | 12.509 | -0.0767 |
| beam5_acc | opt-125m | 5 | acc-sub | beam5 | - | wt2 | 43.3 | 62.8 | 0.515 | 0.980 | -17.0 | 0.353 | +0.1468 |
| mx_beam5 | opt-125m | 5 | acc-sub | beam5 | - | wt2 | 43.1 | 62.0 | 0.527 | 0.986 | -16.8 | 0.341 | +0.1459 |
| acc_t0.7 | opt-125m | 5 | acc-sub | tau0.7 | - | wt2 | 43.3 | 61.2 | 0.469 | 0.995 | -51.0 | 0.230 | +0.1087 |
| fix_t0.7 | opt-125m | 5 | acc-sub | tau0.7 | blockfix1 | wt2 | 43.3 | 63.4 | 0.439 | 0.988 | -48.8 | 0.260 | +0.1087 |
| fixhalf_t0.7 | opt-125m | 5 | acc-sub | tau0.7 | blockfix0.5 | wt2 | 43.3 | 62.9 | 0.434 | 0.994 | -50.9 | 0.230 | +0.1087 |
| ld_t0.7 | opt-125m | 5 | acc-sub | tau0.7 | logitfix1 | wt2 | 43.3 | 61.7 | 0.476 | 0.992 | -42.8 | 0.318 | +0.1087 |
| mit_keep10_t0.7 | opt-125m | 5 | acc-sub | tau0.7 | keepreal0.1 | wt2 | 43.3 | 57.0 | 0.307 | 0.986 | -46.2 | 0.263 | +0.1087 |
| mit_verify50_t0.7 | opt-125m | 5 | acc-sub | tau0.7 | verify0.5 | wt2 | 43.3 | 57.7 | 0.600 | 0.998 | -59.4 | 0.255 | +0.1087 |
| mx_t0.7 | opt-125m | 5 | acc-sub | tau0.7 | - | wt2 | 43.1 | 59.3 | 0.472 | 0.996 | -50.9 | 0.219 | +0.1037 |
| pd_t0.7 | opt-125m | 5 | acc-sub | tau0.7 | probefix0.1 | wt2 | 43.3 | 61.9 | 0.486 | 0.994 | -46.8 | 0.434 | +0.0587 |
| big_t0.85 | opt-125m | 2 | acc-sub | tau0.85 | - | wt2 | 35.6 | 43.7 | 0.574 | 0.986 | -53.4 | 0.159 | +0.0714 |
| ep1_t0.85 | opt-125m | 1 | acc-sub | tau0.85 | - | wt2 | 37.4 | 43.7 | 0.657 | 0.988 | -53.0 | 0.348 | +0.0557 |
| v_t0.85 | opt-125m | 5 | acc-sub | tau0.85 | - | wt2 | 43.3 | 60.6 | 0.473 | 0.989 | -42.7 | 0.292 | +0.0594 |
| acc_t1.0 | opt-125m | 5 | acc-sub | tau1 | - | wt2 | 43.3 | 54.8 | 0.433 | 0.990 | +0.4 | 1.298 | -0.0083 |
| bep10_t1.0 | opt-125m | 10 | acc-sub | tau1 | - | wt2 | 60.7 | 95.3 | 0.409 | 0.978 | +2.4 | 1.420 | -0.0079 |
| bep1_t1.0 | opt-125m | 1 | acc-sub | tau1 | - | wt2 | 37.4 | 40.8 | 0.495 | 0.998 | -0.2 | 1.547 | -0.0224 |
| bep2_t1.0 | opt-125m | 2 | acc-sub | tau1 | - | wt2 | 37.7 | 41.4 | 0.411 | 0.976 | -2.3 | 1.273 | -0.0129 |
| big_t1.0 | opt-125m | 2 | acc-sub | tau1 | - | wt2 | 35.6 | 39.3 | 0.419 | 0.973 | -0.8 | 1.065 | -0.0080 |
| cz_bep1_t1.0_s1 | opt-125m | 1 | acc-sub | tau1 | - | wt2 | 37.3 | 40.7 | 0.500 | 0.993 | +0.8 | 1.647 | -0.0257 |
| cz_bep2_t1.0_s1 | opt-125m | 2 | acc-sub | tau1 | - | wt2 | 37.7 | 41.3 | 0.409 | 0.983 | -1.3 | 1.346 | -0.0154 |
| mx_keep10 | opt-125m | 5 | acc-sub | tau1 | keepreal0.1 | wt2 | 43.1 | 52.2 | 0.288 | 0.985 | +2.5 | 1.410 | -0.0123 |
| mx_keep10_s1 | opt-125m | 5 | acc-sub | tau1 | keepreal0.1 | wt2 | 43.1 | 52.5 | 0.286 | 0.962 | +1.9 | 1.425 | -0.0108 |
| mx_t1.0 | opt-125m | 5 | acc-sub | tau1 | - | wt2 | 43.1 | 53.7 | 0.423 | 0.983 | +1.3 | 1.418 | -0.0123 |
| mx_t1.0_s1 | opt-125m | 5 | acc-sub | tau1 | - | wt2 | 43.1 | 53.9 | 0.415 | 0.986 | +2.4 | 1.472 | -0.0108 |
| mx_verify50 | opt-125m | 5 | acc-sub | tau1 | verify0.5 | wt2 | 43.1 | 53.6 | 0.542 | 0.990 | -14.3 | 0.976 | -0.0123 |

*Table 12, continued*

| arm | model | ep | workflow | dec | interv | corp | $\text{ppl}_0$ | $\text{ppl}_N$ | $a$ | $R^2$ | div% | $q/f$ | $\hat{\delta}_0$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mx_verify50_s1 | opt-125m | 5 | acc-sub | tau1 | verify0.5 | wt2 | 43.1 | 53.7 | 0.526 | 0.995 | -14.5 | 0.947 | -0.0108 |
| noise0.15_ep10 | opt-125m | 10 | acc-sub | tau1 | noise0.15 | wt2 | 60.7 | 72.6 | 0.470 | 0.995 | +17.4 | 1.351 | -0.0066 |
| noise0.30_ep10 | opt-125m | 10 | acc-sub | tau1 | noise0.3 | wt2 | 60.7 | 75.0 | 0.651 | 0.989 | +13.6 | 1.263 | -0.0055 |
| w103_t1.0 | opt-125m | 1 | acc-sub | tau1 | - | wt103 | 34.7 | 38.4 | 0.530 | 0.994 | +2.1 | 1.275 | -0.0137 |
| w103_t1.0_s1 | opt-125m | 1 | acc-sub | tau1 | - | wt103 | 34.6 | 38.3 | 0.472 | 0.990 | +2.3 | 1.328 | -0.0106 |
| big_t1.15 | opt-125m | 2 | acc-sub | tau1.15 | - | wt2 | 35.6 | 39.6 | 0.516 | 0.990 | +95.6 | 15.547 | -0.0939 |
| cz_ep1_t1.15_s1 | opt-125m | 1 | acc-sub | tau1.15 | - | wt2 | 37.3 | 42.0 | 0.605 | 0.992 | +64.0 | 16.388 | -0.1111 |
| cz_ep2_t1.15_s1 | opt-125m | 2 | acc-sub | tau1.15 | - | wt2 | 37.7 | 41.5 | 0.525 | 0.992 | +67.9 | 15.468 | -0.0997 |
| ep10_t1.15 | opt-125m | 10 | acc-sub | tau1.15 | - | wt2 | 60.7 | 70.3 | 0.037 | 0.373 | +80.7 | 16.893 | -0.0809 |
| ep1_t1.15 | opt-125m | 1 | acc-sub | tau1.15 | - | wt2 | 37.4 | 41.8 | 0.554 | 0.992 | +65.0 | 16.300 | -0.1088 |
| ep2_t1.15 | opt-125m | 2 | acc-sub | tau1.15 | - | wt2 | 37.7 | 41.5 | 0.479 | 0.984 | +69.5 | 15.757 | -0.0968 |
| mx_t1.15 | opt-125m | 5 | acc-sub | tau1.15 | - | wt2 | 43.1 | 48.2 | 0.257 | 0.912 | +72.3 | 16.127 | -0.0912 |
| mx_t1.15_s1 | opt-125m | 5 | acc-sub | tau1.15 | - | wt2 | 43.1 | 48.0 | 0.271 | 0.963 | +71.4 | 15.748 | -0.0921 |
| out115_s1 | opt-125m | 5 | acc-sub | tau1.15 | - | wt2 | 43.4 | 48.7 | 0.284 | 0.960 | +72.3 | 14.810 | -0.0895 |
| out115_s2 | opt-125m | 5 | acc-sub | tau1.15 | - | wt2 | 43.3 | 48.1 | 0.257 | 0.962 | +73.2 | 15.504 | -0.0886 |
| v_t1.15 | opt-125m | 5 | acc-sub | tau1.15 | - | wt2 | 43.3 | 48.4 | 0.229 | 0.959 | +73.0 | 15.906 | -0.0880 |
| w103_t1.15 | opt-125m | 1 | acc-sub | tau1.15 | - | wt103 | 34.7 | 39.0 | 0.570 | 0.995 | +100.3 | 33.999 | -0.1106 |
| acc_t1.3 | opt-125m | 5 | acc-sub | tau1.3 | - | wt2 | 43.3 | 49.0 | 0.505 | 0.973 | +45.8 | 22.823 | -0.1709 |
| big_t1.3 | opt-125m | 2 | acc-sub | tau1.3 | - | wt2 | 35.6 | 40.7 | 0.599 | 0.994 | +55.7 | 22.780 | -0.1833 |
| ep10_t1.3 | opt-125m | 10 | acc-sub | tau1.3 | - | wt2 | 60.7 | 67.3 | 0.335 | 0.603 | +53.5 | 23.845 | -0.1597 |
| ep1_t1.3 | opt-125m | 1 | acc-sub | tau1.3 | - | wt2 | 37.4 | 42.8 | 0.591 | 0.997 | +34.6 | 22.407 | -0.1969 |
| ep10_t1.45 | opt-125m | 10 | acc-sub | tau1.45 | - | wt2 | 60.7 | 68.9 | 1.228 | 0.936 | +27.3 | 25.741 | -0.2424 |
| v_t1.45 | opt-125m | 5 | acc-sub | tau1.45 | - | wt2 | 43.3 | 50.7 | 0.734 | 0.994 | +22.0 | 25.065 | -0.2572 |
| acc_t1.6 | opt-125m | 5 | acc-sub | tau1.6 | - | wt2 | 43.3 | 52.6 | 0.794 | 0.993 | +9.2 | 25.897 | -0.3442 |
| fix_t1.6 | opt-125m | 5 | acc-sub | tau1.6 | blockfix1 | wt2 | 43.3 | 51.9 | 0.657 | 0.997 | +9.2 | 25.715 | -0.3442 |
| ld_t1.6 | opt-125m | 5 | acc-sub | tau1.6 | logitfix1 | wt2 | 43.3 | 51.7 | 0.746 | 0.993 | +6.2 | 6.914 | -0.3442 |
| pd_t1.6 | opt-125m | 5 | acc-sub | tau1.6 | probefix0.1 | wt2 | 43.3 | 51.9 | 0.671 | 0.998 | -19.2 | 1.740 | -0.1490 |
| v_t1.9 | opt-125m | 5 | acc-sub | tau1.9 | - | wt2 | 43.3 | 54.6 | 0.799 | 0.980 | +1.6 | 26.438 | -0.4860 |
| v_topp0.7 | opt-125m | 5 | acc-sub | tau1/topp0.7 | - | wt2 | 43.3 | 61.3 | 0.466 | 0.993 | -51.8 | 0.237 | +0.0991 |
| acc_topp0.9 | opt-125m | 5 | acc-sub | tau1/topp0.9 | - | wt2 | 43.3 | 60.6 | 0.501 | 0.980 | -34.9 | 0.302 | +0.0505 |
| mx_topp0.9 | opt-125m | 5 | acc-sub | tau1/topp0.9 | - | wt2 | 43.1 | 58.8 | 0.486 | 0.988 | -33.5 | 0.327 | +0.0478 |
| v_topp0.95 | opt-125m | 5 | acc-sub | tau1/topp0.95 | - | wt2 | 43.3 | 58.9 | 0.489 | 0.991 | -19.1 | 0.455 | +0.0315 |
| beam5 | opt-125m | 5 | replace | beam5 | - | wt2 | 43.3 | 83.4 | 0.150 | 0.867 | -32.3 | 0.255 | +0.1468 |
| rep_t0.7 | opt-125m | 5 | replace | tau0.7 | - | wt2 | 43.3 | 121.6 | 0.509 | 0.704 | -59.6 | 0.253 | +0.1087 |
| sample_t1 | opt-125m | 5 | replace | tau1 | - | wt2 | 43.3 | 123.4 | 1.083 | 0.999 | -14.3 | 1.757 | -0.0083 |
| sample_t1_s1 | opt-125m | 5 | replace | tau1 | - | wt2 | 43.4 | 122.6 | 1.068 | 0.998 | -12.4 | 1.717 | -0.0109 |
| rep_t1.3 | opt-125m | 5 | replace | tau1.3 | - | wt2 | 43.3 | 1521.5 | 2.366 | 0.856 | +46.7 | 26.769 | -0.1709 |
| rep_topp0.9 | opt-125m | 5 | replace | tau1/topp0.9 | - | wt2 | 43.3 | 142.1 | 0.980 | 0.839 | -66.6 | 0.284 | +0.0505 |
| scale_pythia-160m_t1.0 | pythia-160m | 2 | acc-sub | tau1 | - | wt2 | 78.8 | 96.4 | 0.203 | 0.229 | +3.6 | 0.432 | +0.0163 |
| p1b_t0.85 | pythia-1b | 1 | acc-sub | tau0.85 | - | wt2 | 22.0 | 24.6 | 0.770 | 0.996 | -35.3 | 0.574 | +0.0359 |
| p1b_of_t1.0 | pythia-1b | 2 | acc-sub | tau1 | - | wt2 | 38.5 | 57.4 | 0.433 | 0.966 | -11.6 | 0.769 | +0.0170 |
| p1b_t1.0_s1 | pythia-1b | 1 | acc-sub | tau1 | - | wt2 | 22.0 | 23.0 | 0.534 | 0.980 | +2.6 | 1.463 | -0.0153 |
| scale_pythia-1b_t1.0 | pythia-1b | 1 | acc-sub | tau1 | - | wt2 | 22.0 | 22.9 | 0.530 | 0.980 | +2.5 | 1.455 | -0.0153 |
| p1b_of_t1.15 | pythia-1b | 2 | acc-sub | tau1.15 | - | wt2 | 38.5 | 50.9 | 0.255 | 0.673 | +76.3 | 7.741 | -0.0345 |
| p1b_t1.15 | pythia-1b | 1 | acc-sub | tau1.15 | - | wt2 | 22.0 | 24.2 | 0.648 | 0.982 | +83.8 | 17.886 | -0.0810 |
| p1b_t1.15_s1 | pythia-1b | 1 | acc-sub | tau1.15 | - | wt2 | 22.0 | 24.1 | 0.647 | 0.987 | +84.4 | 17.744 | -0.0804 |
| p1b_t1.3 | pythia-1b | 1 | acc-sub | tau1.3 | - | wt2 | 22.0 | 24.5 | 0.592 | 0.994 | +48.6 | 22.614 | -0.1635 |
| bench_p410ep5_t1.0 | pythia-410m | 5 | acc-sub | tau1 | - | wt2 | 36.3 | 59.3 | 0.810 | 0.917 | +13.1 | 2.117 | -0.0078 |
| scale_pythia-410m_t1.0 | pythia-410m | 1 | acc-sub | tau1 | - | wt2 | 26.8 | 28.2 | 0.641 | 0.970 | +1.8 | 1.528 | -0.0174 |
| bench_p410ep5_t1.15 | pythia-410m | 5 | acc-sub | tau1.15 | - | wt2 | 36.3 | 42.9 | 0.243 | 0.288 | +87.6 | 15.125 | -0.0653 |
| scale_pythia-70m_t1.0 | pythia-70m | 2 | acc-sub | tau1 | - | wt2 | 377.8 | 579.8 | 0.411 | 0.825 | +3.2 | 0.464 | +0.0259 |