

Δ PPL Is Not Δ CE: The Scale-Dependence Confound in Generative Continual-Learning Forgetting Metrics

✉ ✉

242015088@wxu.edu.cn

Wuxi Taihu University, School of Integrated Circuits, Department of Communication Engineering

Research Article

Keywords: continual learning, catastrophic forgetting, perplexity, cross-entropy, language models, evaluation metrics

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11481703/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Δ PPL Is Not Δ CE: The Scale-Dependence Confound in Generative Continual-Learning Forgetting Metrics

Qingjun Zhang
 Wuxi Taihu University, Wuxi 214064, China
 242015088@wxu.edu.cn

October 2026

Abstract

Forgetting in generative continual learning is routinely reported as a raw change in perplexity, $\Delta\text{PPL} = \text{PPL}_{\text{after}} - \text{PPL}_{\text{before}}$. Because perplexity is the exponential of cross-entropy, $\text{PPL} = e^{\text{CE}}$, the reported forgetting is itself scaled by the *base* perplexity at which the measurement is taken: to first order, $\Delta\text{PPL} \approx \text{PPL}_{\text{base}} \cdot \Delta\text{CE}$. A domain or stratum with a higher base perplexity therefore reports a larger ΔPPL for the *same* underlying drift, and differential-forgetting ratios across strata are scaled by exactly the ratio of their base perplexities—inflated when the harder stratum sits at a higher base, deflated when a stronger model sits at a lower base. We derive the exact three-factor decomposition and demonstrate it in both directions: on a frozen-anchor MoE language model, where a published $2.14\times$ “long-tail forgetting” ratio collapses to a $1.48\times$ effect in cross-entropy space (a 45% inflation produced entirely by the nonlinear scale of the metric), and on vanilla dense GPT-2 across three scales, where a $2.06\times$ cross-entropy differential is reported as only $1.55\times$ in perplexity space. This confound is distinct from the accuracy/backward-transfer coupling documented for task-incremental classification: the latter is an exact *linear* identity, whereas the perplexity coupling is introduced by the choice to exponentiate. We close with a perplexity instantiation of a claim-condition-control disclosure protocol and a pilot survey indicating that the generative field is heterogeneous rather than uniformly misreporting: base perplexity is usually reported and forgetting-only reporting is rare, so the residual risk is concentrated in differential comparisons across unequal bases.

1 The exponential coupling

Let $\text{CE}(t)$ denote the cross-entropy (in nats) of a fixed evaluation set at training time t , and $\text{PPL}(t) = e^{\text{CE}(t)}$ its perplexity. Forgetting is commonly reported as

$$\Delta\text{PPL} = \text{PPL}(t_{\text{after}}) - \text{PPL}(t_{\text{before}}) = e^{\text{CE}_a} - e^{\text{CE}_b}, \quad (1)$$

where $\text{CE}_b = \text{CE}(t_{\text{before}})$ and $\text{CE}_a = \text{CE}(t_{\text{after}})$. Writing $\delta = \text{CE}_a - \text{CE}_b$ and factoring e^{CE_b} ,

$$\Delta\text{PPL} = \text{PPL}_b \left(e^\delta - 1 \right) = \text{PPL}_b \cdot \underbrace{\delta \cdot \frac{e^\delta - 1}{\delta}}_{1 + \frac{\delta}{2} + \frac{\delta^2}{6} + \dots}. \quad (2)$$

To first order the reported forgetting is the product of the *base* perplexity and the actual cross-entropy drift:

$$\boxed{\Delta\text{PPL} \approx \text{PPL}_{\text{base}} \cdot \Delta\text{CE}}, \quad (3)$$

with relative error $\approx \delta/2$ (below 2% for $\delta < 0.04$ nats).

Two consequences follow immediately.

Level dependence. For a fixed underlying drift δ , the reported ΔPPL grows linearly with PPL_{base} . A domain whose base perplexity is 26 reports $1.44\times$ the forgetting of a domain at base perplexity 18 under the *identical* ΔCE .

Ratio inflation. For two strata A, B with base perplexities P_A, P_B and drifts δ_A, δ_B , the exact relation is

$$\frac{\Delta\text{PPL}_A}{\Delta\text{PPL}_B} = \underbrace{\frac{P_A}{P_B}}_{\text{scale}} \cdot \underbrace{\frac{\delta_A}{\delta_B}}_{\text{effect}} \cdot \underbrace{\frac{(e^{\delta_A} - 1)/\delta_A}{(e^{\delta_B} - 1)/\delta_B}}_{\text{2nd order } \approx 1 + \frac{\delta_A - \delta_B}{2}}. \quad (4)$$

The reported ratio confounds a pure unit/scale term P_A/P_B with the quantity of interest δ_A/δ_B . Only the second factor measures forgetting; the first is an artifact of where the measurement sits on the exponential curve.

2 Demonstrations

2.1 A frozen-anchor MoE model

We re-derive the correction on a measured long-tail probe of a Qwen2.5-0.5B model whose MLP blocks are replaced by a frozen-anchor mixture of experts (a frozen FFN deepcopy plus $E = 4$ trainable experts, $\alpha = 0.02$, hard document-shard routing), pre-trained on FineWeb and then adapted on arXiv. The FineWeb evaluation set ($N = 120$) is stratified at the median by token frequency into a *core* (high-frequency) and a *tail* (low-frequency) stratum. Forgetting per stratum is measured as ΔPPL between the pre-adaptation and post-adaptation checkpoints.

Stratum	PPL_{base}	$\text{PPL}_{\text{after}}$	ΔPPL	$\delta = \Delta\text{CE}$ (nats)
core	18.143	18.842	0.699	0.0378
tail	26.030	27.527	1.497	0.0559

Table 1: Perplexity and cross-entropy forgetting by stratum. The raw ΔPPL ratio is $1.497/0.699 = 2.14\times$.

Table 1 reports a $2.14\times$ “tail forgets more than core” ratio in perplexity space. Applying Eq. (4),

$$\frac{\Delta\text{PPL}_{\text{tail}}}{\Delta\text{PPL}_{\text{core}}} = \underbrace{\frac{26.030}{18.143}}_{1.435} \cdot \underbrace{\frac{0.0559}{0.0378}}_{1.479} \cdot \underbrace{\frac{(e^{0.0559} - 1)/0.0559}{(e^{0.0378} - 1)/0.0378}}_{1.009} = 2.142. \quad (5)$$

The scale factor 1.435 is pure exponential nonlinearity: it equals the ratio of base perplexities and is present even if the two strata had *identical* ΔCE . The genuine cross-entropy differential is $1.48\times$, not $2.14\times$; the raw metric overstated the effect by 45%.

Two caveats bound this demonstration. First, it is a single checkpoint pair and $N = 120$ eval samples, so $1.48\times$ is itself an estimate, not a precise population value; the point here is the *metric* correction, not the strength of any tail effect. Second, the exact identity $2.14 = 1.435 \times 1.479 \times 1.009$ holds by construction; the substantive claim is that the 1.435 factor—one third of the reported ratio—carries no information about forgetting whatsoever.

2.2 Vanilla dense GPT-2 across three scales

To check that the scale term is not an artifact of the frozen-anchor MoE architecture, we repeat the measurement on three vanilla dense GPT-2 checkpoints (124M, 355M, 774M). Each is continually pretrained on the tiny tinyshakespeare corpus for a fixed 150 steps and evaluated for forgetting on wikitext before and after.

Model	PPL _{base}	Δ PPL	$\delta = \Delta$ CE (nats)
GPT-2 124M	27.263	6.849	0.2241
GPT-2 355M	20.674	8.489	0.3440
GPT-2 774M	18.090	10.642	0.4627

Table 2: Dense multi-scale check. The raw Δ PPL ratio (large/small) is 1.554 against a Δ CE ratio of 2.064.

Here the base perplexity *decreases* with model scale, so the scale term of Eq. (4) runs in the opposite direction to Table 1: the stronger model’s forgetting is *deflated* in Δ PPL. For the large/small pair the exact decomposition is

$$\frac{\Delta\text{PPL}_{\text{large}}}{\Delta\text{PPL}_{\text{small}}} = \underbrace{\frac{18.090}{27.263}}_{0.664} \cdot \underbrace{\frac{0.4627}{0.2241}}_{2.064} \cdot \underbrace{\frac{(e^{0.4627} - 1)/0.4627}{(e^{0.2241} - 1)/0.2241}}_{1.135} = 1.554. \quad (6)$$

The two models forget by a factor $2.06\times$ in cross-entropy, but the raw Δ PPL comparison reports only $1.55\times$: the 0.664 scale factor—the ratio of base perplexities—conceals a third of the true differential. This is the mechanism by which a “larger models forget less” conclusion can arise in perplexity space even when cross-entropy says otherwise. A caveat bounds the claim: the fixed-step schedule gives larger models a larger effective dose, so the Δ CE *ordering* across scales is not itself asserted here; what is asserted is the metric decomposition, which holds to four significant figures on all three pairs.

3 Relation to the accuracy/backward-transfer coupling

A concurrent line of work [1] documents a structurally similar confound for task-incremental *classification*: a method can appear to forget less because it learned less before forgetting was measured, and the quantities obey an exact linear identity $\text{LearnAcc} = \text{ACC}_T - \frac{T-1}{T}\text{BWT}_T$. That coupling is intrinsic to the metric definition of backward transfer on *accuracy*, which is a linear average.

The perplexity coupling derived here is *not* that coupling. Accuracy is linear in the per-example error; perplexity is the exponential of averaged cross-entropy. The identity $\text{LearnAcc} = \text{ACC}_T - \frac{T-1}{T}\text{BWT}_T$ has no perplexity analogue, because PPL is not a linear mean. Instead, the perplexity confound is *introduced by the choice to exponentiate*: had the field reported Δ CE directly, no coupling would exist, since Δ CE is simply the difference of the two cross-entropies. The correction is therefore simpler and stronger than for accuracy: report the drift in nats (or, equivalently, the log-perplexity ratio $\ln(\text{PPL}_{\text{after}}/\text{PPL}_{\text{before}})$), and always report the base perplexity.

Two further language-model confounds are documented and should not be conflated with ours. Low-rank finetuning *learns less and forgets less* [5]: methods that acquire less on the target domain report less source-domain forgetting, and the correct report is a learning–forgetting Pareto rather

than a forgetting number alone. “Spurious forgetting” [6] refers to a drop in *task alignment* rather than in knowledge. Both are accuracy-space phenomena. A complementary third observation is that perplexity can be *insensitive* to forgetting, failing to register entity-level erasure [7]. Our coupling is orthogonal to all three: it concerns neither how much was learned nor whether knowledge was erased, but the nonlinear *scale* of the perplexity unit itself—where perplexity does move, the raw difference overstates differential forgetting.

The prior work explicitly declines to extend its result to other change metrics: “the coupling does not automatically extend to other change metrics. The coupling should be demonstrated in the target setting before the disclosure is applied” [1]. We regard Sections 1–2 as exactly that demonstration for the generative perplexity setting.

4 A perplexity disclosure, and a coded prevalence survey

Extending the claim–condition–control protocol [1] to perplexity-based forgetting yields a compact disclosure:

- **Condition.** Report the base perplexity (or base cross-entropy) for every stratum, domain, method, and baseline. Without it, a ΔPPL value cannot be compared across rows whose bases differ.
- **Metric convention.** Report $\Delta\text{CE} = \ln(\text{PPL}_{\text{after}}/\text{PPL}_{\text{before}})$ as the primary forgetting metric; keep ΔPPL as a secondary display.
- **Operating range.** State the base-perplexity range and the drift magnitudes; the first-order approximation (3) degrades as δ grows (relative error $\approx \delta/2$).
- **Controls.** A fixed ΔCE evaluated at two different base perplexities must produce ΔPPL values in exactly the base ratio; this is a zero-cost check that the scale term has been separated from the effect.

Finally, we offer a pilot coding of the reporting practice rather than leave it open. We examined eight generative continual-pretraining / LLM continual-learning papers (2023–2025) and recorded, for each, whether forgetting is reported in perplexity, whether a base perplexity is available, and whether both learning and forgetting are reported.

Three observations follow. (i) The forgetting-only pathology documented for the accuracy subfield [1] is largely absent here: every surveyed paper reports both a learning/forward-transfer and a forgetting/backward-transfer number. (ii) Base perplexity is generally reported—as a zero-shot baseline, an oracle reference, or the start of a trajectory—but only [2] normalize backward transfer by the base; the remaining perplexity-based papers report raw values or spikes. (iii) The residual risk is therefore concentrated in *differential* comparisons: a raw perplexity spike (121 $\rightarrow$ 20402 in [4]) or any cross-scale comparison of ΔPPL still carries the factor of Eq. (4) whenever base perplexities differ. The recurring claim that “larger models forget less” is instructive precisely because its validity turns on the metric convention. [2] normalize backward transfer by the zero-shot perplexity and so report a version of the claim robust to the scale term of Eq. (4); [8] measure forgetting in ceiling-limited category accuracy and so expose the claim to the complementary accuracy–ceiling confound of [1], not to Eq. (4); and [9] score forgetting as a cross-model gap in a multiple-choice “perplexity” $2^{-\log_2 P(\text{correct})} = 1/P(\text{correct})$ —a reciprocal, not exponential, nonlinearity whose scale dependence is real but distinct, and which falls outside the derivation of Eq. (4). The two nonlinearities carry the same lesson: any strictly convex change metric overstates drift at the low end of its input range.

Paper (year)	Forgetting in PPL?	Base PPL?	Learn + Forget?
Ke et al. 2023 (ICLR)	no (F1/acc.)	—	both
Yildiz et al. 2025 (TMLR)	yes	yes (ZS, <i>norm.</i>)	both
TiC-LM 2025 (ACL)	yes	oracle-subtracted	both
Emergent-CPT 2025 (ACL)	yes	yes (base row)	both
Gemma2-CPT 2025 (arXiv)	yes (PPL fluency)	yes	both
Interplay 2025 (arXiv)	yes (MCQ 1/ P)	yes	both
Replay-to-Remember 2025	yes (raw spike)	yes (121 base)	both
Low-PPL tokens 2025 (NeurIPS)	partial	yes	both
LoRA Learns Less 2024	no (acc.)	—	both (Pareto)
Spurious Forgetting 2025	no (acc., align.)	—	both
Exploring Forgetting 2024	no (entity-recall)	—	both
Mapping Forgetting 2025	no (category acc.)	—	both
Time-Incremental CPT 2026	no (acc.)	—	both
MixLoRA-DSI 2025 (EMNLP)	no (R@10)	—	both
CurLL 2025 (BabyLM)	no (accuracy)	—	both

Table 3: Expanded coding (convenience sample, $N = 15$, arXiv API + snowball, 2023–2026). “Forgetting in PPL?” records whether the forgetting claim is reported in perplexity space.

Note also that “normalizing by base perplexity” removes only the first-order scale term: $(e^\delta - 1)$ is not δ ; the exact correction is the log-ratio of Eq. (4).

Survey method. Candidates were retrieved with the arXiv API using six queries over 2023–2026 (e.g. `all:"continual learning" AND all:"language model" AND all:perplexity AND all:forgetting`), restricted to generative language-model continual learning, and coded from the paper text. This is a convenience sample, not an exhaustive denominator.

References

- [1] Anonymous. *Auditing Backward Transfer Against Learning Strength: A Reanalysis of Biologically Plausible Continual Learning*. Submitted to the AI for Meta-Science workshop, NeurIPS 2026.
- [2] Ç. Yıldız et al. *Investigating Continual Pretraining in Large Language Models: Insights and Implications*. TMLR, 2025.
- [3] J. Li et al. *TiC-LM: A Web-Scale Benchmark for Time-Continual LLM Pretraining*. ACL, 2025.
- [4] *Replay to Remember: Retaining Domain Knowledge in Streaming Language Models*. arXiv:2504.17780, 2025.
- [5] D. Biderman et al. *LoRA Learns Less and Forgets Less*. arXiv:2405.09673, 2024.
- [6] *Spurious Forgetting in Continual Learning of Language Models*. arXiv:2501.13453, 2025.
- [7] *Exploring Forgetting in Large Language Model Pre-Training*. arXiv:2410.17018, 2024.
- [8] *Mapping Post-Training Forgetting in Language Models at Scale*. arXiv:2510.17776, 2025.
- [9] *The Interplay between Domain Specialization and Model Size*. arXiv:2501.02068, 2025.