PARAMETER CONVENTIONS CONFOUND COMPUTE-OPTIMAL SCALING EXPONENTS OF DENSE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Compute-optimal scaling exponents are routinely compared across labs as if they measured the same quantity. We audit every public language-model scaling sweep we could find whose metadata determine, run by run, exactly how many parameters sit in vocabulary-sized matrices: 18 sweeps (1,673 runs) from six independent research groups. Changing only whether those matrices are counted raises the fitted data–model exponent $r = \mathrm{d}\log D^*/\mathrm{d}\log N^*$ in every sweep and every group, by $+0.54$ on average across groups in held-out fits—comparable to the entire Kaplan–Chinchilla gap—under both separable and non-separable loss forms. Public data do not consistently settle which count is right: small-to-large extrapolation leans toward transformer-only counting in most groups but not significantly, holding out the largest models splits the groups three to three, and only one lab's frontier holdout favors one count clearly. The convention changes each exponent, but we find no evidence that it explains why labs disagree: harmonizing every sweep to one convention does not detectably reduce the cross-lab spread relative to the mean exponent. The two conventions nonetheless prescribe different frontier allocations, a median $2.3\times$ difference in tokens per parameter at $100\times$ the fitted compute. We release exact reconstructions, a convention registry, and a reporting standard.

1 INTRODUCTION

Compute-optimal training chooses a model size N^* and a token count D^* for a compute budget by extrapolating a loss surface fitted to small runs. The slope $r = \mathrm{d}\log D^*/\mathrm{d}\log N^*$ decides how additional compute is split between parameters and data. Published values disagree: Kaplan et al. (2020), counting non-embedding parameters, imply $r \approx 0.37$, while Hoffmann et al. (2022), counting all parameters, support $r \approx 1$ (a roughly constant tokens-per-parameter ratio), and later public sweeps span a still wider range.

A reason these numbers need not agree is that "parameters" is not one quantity. Vocabulary-sized matrices (input embedding and output head) make up most of a small model and little of a large one, so counting them bends the horizontal axis of the scaling law in a size-dependent way. Pearce & Song (2024) and Porian et al. (2024) showed that this accounting, together with the cost of the output layer and other training details, explains much of the Kaplan–Chinchilla discrepancy, and Tao et al. (2024) and Bi et al. (2024) model vocabulary parameters or compute separately. What has not been measured is how large and how general the effect is across the modern public record, and whether the available evidence can say which count is right.

This paper. We apply a fixed inclusion rule (§3) to every public sweep we could find and keep those whose metadata determine the vocabulary-matrix parameter count exactly for every run: 18 sweeps from six research groups (Marin/Delphi, the Kempner loss-to-loss sweeps, the AI2 OLMo ladder, Farseer, the OpenLM over-training and IsoFLOP sweeps of Gadre et al. and Porian et al., and the (Mis)Fitting survey sweep). For each we verify an exact decomposition of the parameter count and refit the same loss surface with N counted with and without vocabulary-sized matrices.

Findings. (i) *The shift is universal.* Counting vocabulary matrices raises $\hat{r}$ in all 18 sweeps and all six groups (group mean $+0.54$, 95% CI $[+0.34, +0.73]$, one-sided Wilcoxon $p = 0.016$, the smallest attainable with six groups), under both the Chinchilla form and a non-separable relaxation, and also in

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.

two sources outside the rule. The mechanism is visible directly: the model-size exponent $\hat{\alpha}$ is larger under total counting in every sweep, while $\hat{\beta}$ moves by a median of 1.2%. (ii) *Public data do not consistently identify a convention.* Held-out loss at the top of the compute range, and on runs above the median of both N and D when fitting below both, leans toward transformer-only counting in most groups (five of six; four of five), but not significantly at the group level; holding out the largest models favors each convention in three groups. The Delphi frontier holdout favors transformer-only counting by $3.5\times$, a single-lab exception that the other five groups do not uniformly reproduce. (iii) *We find no evidence that the convention explains why labs disagree.* Expressing every sweep in one convention does not detectably reduce the spread of $\hat{r}$ across groups relative to the mean exponent. (iv) *It matters for allocation.* The two conventions prescribe a median $2.3\times$ difference in tokens per parameter at $100\times$ the fitted compute, on a common physical compute axis.

Contributions. (1) Exact, verified parameter-count reconstructions for 18 public sweeps under a stated inclusion rule, with a registry of every source examined. (2) A multi-lab measurement of convention sensitivity with group-level inference, held-out and extrapolation tests, a dispersion analysis, and allocation consequences. (3) A reporting standard that makes exponents and allocations comparable across studies.

2 Related Work

Compute-optimal scaling. Kaplan et al. (2020) and Hoffmann et al. (2022) set up the modern debate; Besiroglu et al. (2024) showed that Chinchilla's parametric fit is sensitive to fitting details. Pearce & Song (2024) attribute the Kaplan–Chinchilla gap largely to embedding accounting and small scale, and Porian et al. (2024) to last-layer compute, warmup, and optimizer tuning. Both study one historical pair; we measure the effect across the current public record and ask whether it can be resolved from data.

Vocabulary and compute accounting. Tao et al. (2024) fit scaling laws with separate vocabulary and non-vocabulary parameters, and Bi et al. (2024) replace N with non-embedding FLOPs per token. These choices are exactly the conventions whose consequences we quantify.

Estimation practice. Surveys and estimation guides document how fitting choices change conclusions (Li et al., 2025b; Choshen et al., 2025); Kricheli et al. (2026) show that collinear tokens-per-parameter designs are ill-conditioned; Czech et al. (2026) document biases in IsoFLOP parabola fits; and Li et al. (2025a) and Sardana et al. (2024) propose richer or inference-aware objectives. Alternative functional forms, from Farseer's coupled surface to shifted power laws with offset model or data terms (Brandfonbrener et al., 2024), change how N enters the loss but not which parameters N counts; every such form inherits the convention choice we study, and our main result holds under both the separable and the non-separable form we fit. Our registry and reporting standard therefore address an upstream choice: what N is.

3 Inclusion Rule, Sources, and Exact Reconstruction

Inclusion rule. A sweep enters the primary panel if and only if (a) its released metadata determine, for every run, the exact number of parameters in vocabulary-sized matrices; (b) N and D vary independently (at least three model sizes, with several token budgets per size or an IsoFLOP design); and (c) every (N, D, L) point is the final checkpoint of a run trained to its own budget. We searched the Hugging Face Hub, the code and data releases of scaling-law papers, and public IsoFLOP collections up to September 2026, and report every source we examined (Appendix A). No sweep that met the rule was excluded.

Primary panel. Every decomposition below holds to the parameter for every run and is asserted in the released code:

- **Marin** (the Delphi frontier ladder (Held & Marin Community, 2026) and the 2026 Comma, DCLM and Nemotron ladders, exported from the Marin project's public training logs and curated in the IsoFLOP collection accompanying Czech et al. (2026); 339 runs): untied 128,256-token embedding and head, $N_{\text{total}} = 2Vd + L(16d^2 + 2d + 256) + d$.

Table 1: Convention sensitivity by research group. All-data non-separable fits; ranges are over slices. "Held-out Δr" uses lower-80% compute fits. "Size holdout" is the geometric-mean ratio of held-out RMSE (total / core) when the largest 25% of model sizes are held out; values below 1 favor total counting.

Group	Slices	Runs	$\hat{r}_{\text{total}}$	$\hat{r}_{\text{core}}$	Δr	Held-out Δr	Size holdout
Marin	4	339	1.43–1.67	0.71–0.81	$+0.77$	$+0.85$	1.93
(Mis)Fitting	1	81	1.05	0.42	$+0.63$	$+0.28$	1.58
Kempner	6	529	0.87–1.03	0.33–0.41	$+0.57$	$+0.59$	0.34
OLMo	1	25	1.53	1.06	$+0.47$	$+0.49$	0.37
OpenLM (Gadre; Porian)	5	295	0.83–1.41	0.48–0.69	$+0.47$	$+0.54$	1.34
Farseer	1	404	1.36	0.90	$+0.45$	$+0.47$	0.98
Primary panel	18	1,673			18/18 > 0	17/17 > 0	3 of 6 favor core
Sensitivity sources (outside the inclusion rule)							
CoTok	6	503	0.64–0.73	0.46–0.54	$+0.18$	$+0.20$	0.56
OWT (ours)	1	18	1.10	0.74	$+0.36$	$+0.36$	–

- **Kempner** loss-to-loss sweeps on six corpora (Brandfonbrener et al., 2024): $N_{\text{total}} = 64{,}000\,d + L(12d^2 + 4d) + d$.
- **OLMo** model ladder (Bhagia et al., 2025): five sizes at 0.5–$10\times$ Chinchilla tokens; released sizes exclude only the input embedding of an untied, 100,352-row vocabulary.
- **Farseer** (Li et al., 2025a): released `exact_N` and `N_add_emb` differ by one $131{,}072 \times h$ matrix.
- **OpenLM**: the over-training sweeps of Gadre et al. (2024) (C4, RedPajama, RefinedWeb; released N contains exactly one $50{,}432 \times d$ matrix) and the cosine-to-budget IsoFLOP sweeps of Porian et al. (2024) (RefinedWeb, OpenWebText2; counts from each run's width and depth with the authors' own counting function). The two share authors and a codebase and are treated as one group.
- **(Mis)Fitting** survey sweep (Li et al., 2025b): released N and $N_{\text{no emb}}$ differ by $2 \times 50{,}432 \times d$; we keep the best learning rate per (N, D), as in the source.

Sensitivity sources. CoTok (Limisiewicz et al., 2026) fails (a): its alternative "global" count also removes compute-bearing local encoder and decoder blocks. Our own 18-run OWT sweep fails (c) (cumulative checkpoints over repeated data) and has only three sizes. We report both only as sensitivity checks; neither enters any primary statistic.

Units of replication. Slices from one research group share architectures and pipelines, so group-level statistics use the six groups above. As a stricter check we also collapse to four training codebases (Levanter for Marin; the OLMo codebase for Kempner and AI2; OpenLM for Gadre, Porian and (Mis)Fitting; Farseer's own).

Conventions and estimand. *Total* counts every trainable parameter; *transformer-only* ("core") removes vocabulary-sized matrices; where the head is identifiable we also report *core+head*, the FLOP-faithful count. For each sweep and convention we fit, with one fixed protocol (bounded nonlinear least squares from a fixed multi-start grid), the Chinchilla form $L = E + AN^{-\alpha} + BD^{-\beta}$ and its one-parameter relaxation $L = E + AN^{-\alpha} + BN^{\theta}D^{-\beta}$, whose compute-optimal slope is $r = (\alpha + \theta)/\beta$ (Chinchilla: $r = \alpha/\beta$). The relaxation is a diagnostic device; the main result holds under either form. The two conventions are always fitted with the same functional form, the same number of free parameters and the same runs; they differ only in the covariate N, so neither has an effective-complexity advantage. Loss targets differ across sources, so every comparison is made within a sweep.

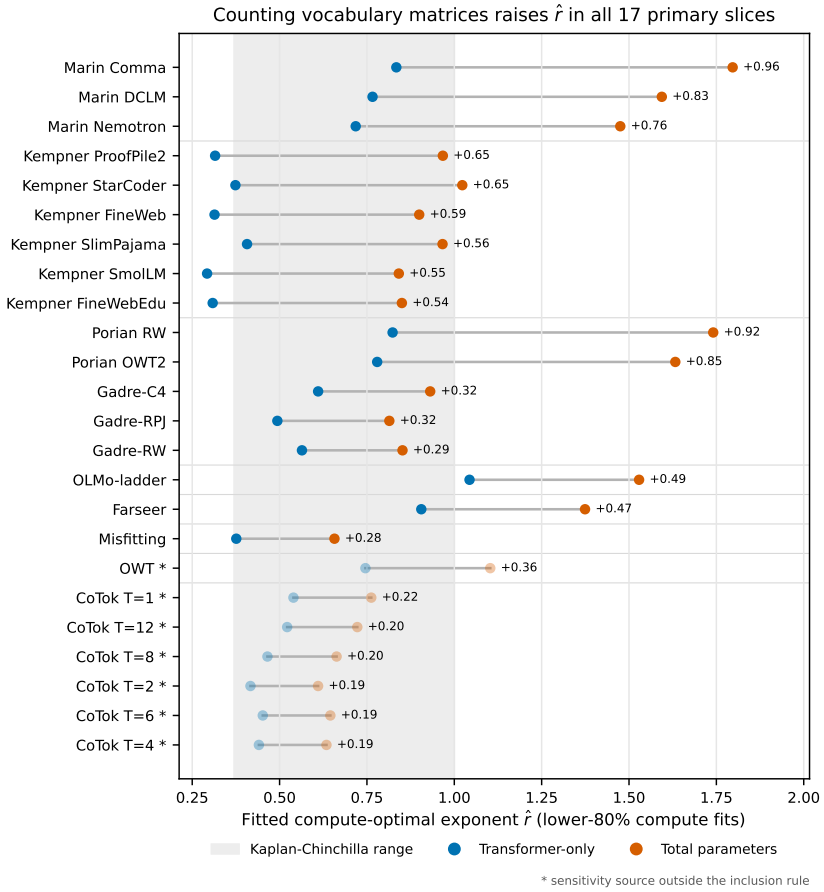

Figure 1: Fitted $\hat{r}$ under transformer-only (blue) and total (orange) counting on the lower-80% compute portion of each sweep, grouped by research group. Counting vocabulary matrices raises $\hat{r}$ in all 17 primary slices; the Delphi ladder, fitted on its own public split, adds an 18th ($+0.76$). Faded rows marked * are sensitivity sources outside the inclusion rule.

4 Results

4.1 The shift is universal

Figure 1 and Table 1 show the central result. Counting vocabulary-sized matrices raises $\hat{r}$ in every slice. With research groups as units, the mean held-out shift is $+0.54$ (95% CI $[+0.34, +0.73]$); all six groups are positive (one-sided Wilcoxon $p = 0.016$, the floor for six units), and dropping any one group leaves $p \leq 0.031$. Under the stricter codebase grouping all four units are positive (mean $+0.60$, t-interval $[+0.32, +0.87]$; with four units the Wilcoxon floor is 0.0625, which is attained). At the slice level the mean is $+0.59$ ($[+0.48, +0.70]$, $n = 17$). Both sensitivity sources shift in the same direction; including them gives eight groups, all positive ($p = 0.004$). The shifts are large: in Marin, (Mis)Fitting and Kempner they are comparable to or larger than the whole Kaplan–Chinchilla difference (≈ 0.63), and in five of six groups they move at least one sweep from above to below $r = 1$. The Delphi ladder illustrates the stakes: its released total-parameter count gives $\hat{r} = 1.57$, a slope far steeper than Chinchilla's, which becomes 0.81 under transformer-only counting. An apparently anomalous exponent from a single release can thus be largely an accounting choice.

Not an artifact of the relaxation. Refitting the standard separable Chinchilla form, $\hat{r} = \alpha/\beta$, gives a positive shift in 18 of 18 primary sweeps (range $+0.12$ to $+1.41$) and in all seven sensitivity slices.

Mechanism: a local derivation. Suppose the model-size term is $AN_{\text{core}}^{-\alpha}$ and, with the vocabulary fixed within a ladder, $N_{\text{total}} = N_{\text{core}} + V$, where $V > 0$ is the vocabulary-matrix block. Write $\gamma =$

$\mathrm{d}\log V/\mathrm{d}\log N_{\mathrm{core}}$ and $s = V/N_{\mathrm{total}}$. Because $\mathrm{d}\log N_{\mathrm{total}}/\mathrm{d}\log N_{\mathrm{core}} = (N_{\mathrm{core}} + \gamma V)/(N_{\mathrm{core}} + V) = 1 - (1-\gamma)s$, the chain rule gives the local exponent of the same term with respect to N_{total}:

$$\alpha_{\mathrm{total}} = \frac{\alpha}{1-(1-\gamma)s}.$$

Hence $\alpha_{\mathrm{total}} > \alpha$ exactly when $\gamma < 1$, that is, when the vocabulary block grows more slowly than the transformer body. With a fixed vocabulary $V \propto d$ while N_{core} grows roughly as Ld^2, so $\gamma < 1$ is expected; we verify it between every pair of consecutive model sizes in all 18 primary sweeps ($0 \le \gamma \le 0.53$). The distortion is largest where s is large, at the small-model end, and vanishes as $s \to 0$, which for a fixed vocabulary happens as models grow.

This statement is local. A single power law fitted to a ladder whose s falls from 52–97% at the smallest model to 6–19% at the largest has no closed-form exponent, so the derivation predicts the direction of the fitted shift rather than proving it; it also leaves the data exponent β unchanged, since D is not reparametrized. The prediction holds in every case: $\hat{\alpha}_{\mathrm{total}} > \hat{\alpha}_{\mathrm{core}}$ in all 18 primary Chinchilla-form fits, while $\hat{\beta}$ changes by a median of 1.2%, so $\hat{r} = \hat{\alpha}/\hat{\beta}$ rises. The size of the shift depends on each ladder's vocabulary share and width schedule, which is why it must be measured.

4.2 Public data do not consistently identify a convention

If one convention described the loss surface better, it should predict held-out runs better. We test this in three held-out designs, plus one lab's frontier holdout.

Top of the compute range. Fitting each sweep on its lower 80% of compute and predicting the top 20%, transformer-only counting predicts slightly better in five of six groups (median RMSE difference, total minus core, $+0.0035$; slice-level bootstrap CI $[+0.0001, +0.048]$), but Kempner goes the other way and the group-level test is not significant (Wilcoxon $p = 0.09$, sign test $p = 0.22$, two-sided). The same fits and folds estimate Δr precisely.

Largest models. Conventions differ most in how they extrapolate in N, so we also hold out the largest 25% of distinct model sizes in each sweep (Table 1, last column). The groups split evenly: transformer-only counting predicts the largest models better for Marin, OpenLM and (Mis)Fitting, total counting for Kempner and OLMo, and Farseer is essentially tied (ratio 0.98, marginally favoring total); across all primary slices, 9 of 17 favor core. Where the head is identifiable, the FLOP-faithful core+head count is never the most accurate of the three.

Small-to-large. Fitting only runs below the median of both N and D and predicting runs above both medians (possible for 13 primary slices from five groups), transformer-only counting predicts better in 12 of 13 slices and four of five groups (two-sided group-level Wilcoxon $p = 0.13$, the smallest attainable with five groups being 0.06). The shift itself survives this split: Δr estimated from the small-N, small-D corner alone is positive in all 13 slices, although its magnitude is poorly determined from so narrow a range.

The Delphi exception. Delphi's public frontier holdout, which extrapolates about $300\times$ beyond its fit, favors transformer-only counting clearly (RMSE 0.026 vs 0.091), consistent with the other Marin ladders. It is the strongest single piece of evidence for one convention, but it comes from one lab and recipe, and Kempner and OLMo point the other way on the largest-model test. Public data therefore do not *consistently* settle which convention is right; settling it will require a controlled intervention on vocabulary size.

4.3 No evidence that the convention explains why labs disagree

If cross-lab disagreement in $\hat{r}$ were an accounting artifact, expressing every group in one convention should shrink it. We find no evidence that it does, although the test has limited power with six groups. With groups as units, the between-group standard deviation of $\hat{r}$ is 0.25 in the conventions each source released and 0.27 under transformer-only counting (ratio 1.08, group-bootstrap 95% CI $[0.52, 3.98]$). Under the separable form the spread falls when moving from total to transformer-only counting (ratio 0.50), but transformer-only exponents are uniformly smaller (mean 0.60 vs 1.27), so the spread relative to the mean does not fall (coefficient of variation 0.65 transformer-only vs 0.62 total). The confidence intervals are wide: the data are compatible with a moderate reduction or a moderate increase, but not with the convention accounting for most of the between-lab spread.

Table 2: Allocation disagreement between conventions on one physical compute axis ($C = 6N_{\text{total}}D$), from lower-80% fits of the 17 primary slices. The transformer-only surface is evaluated at $N_{\text{core}} = N_{\text{total}} - N_{\text{vocab}}(N_{\text{total}})$ with N_{vocab} fitted per source. "Excess compute" is the extra compute needed to reach the reference path's loss when following the other convention's path, with the core or the total fit as reference.

Budget	Median TPP ratio	Interquartile range	Excess (core ref.)	Excess (total ref.)
10×	1.39×	1.16–3.40×	2.2%	2.5%
100×	2.31×	1.72–8.60×	7.1%	11.4%
1000×	4.71×	2.62–21.3×	22.6%	41.6%

Table 3: Reporting standard for compute-optimal scaling-law exponents.

Item	Required disclosure
Parameter convention	Whether N includes input embeddings, output head, tied or untied vocab matrices, encoder/decoder blocks, and other non-transformer parameters.
Reconstructability	Vocabulary size, width, depth, tying, and exact per-run configs, enough to recompute every convention.
Compute convention	How C is counted (e.g. $6ND$ with which N, head and attention FLOPs).
Exponents	$\hat{r}$ under every convention the metadata support; pool only within a matched convention and design.
Allocations	The recommended (N^*, D^*) at target budgets in total parameters and physical FLOPs, not only the exponent.

The convention thus shifts every exponent by a large, predictable amount, while the between-lab differences—recipes, data, and design geometry—remain. Resolving the convention would make exponents comparable; we see no sign that it would, by itself, make them agree.

4.4 Allocation consequences

Equal held-out loss does not mean equal decisions. We extrapolate each pair of fits to budgets 10–1000× the largest compute in its training fold, optimize both on the same physical compute axis, and compare the prescribed allocations (Table 2). At 100×, total counting prescribes a median 2.3× more tokens per parameter than transformer-only counting, in the same direction for 14 of 17 slices. Following one convention's path when the other is correct costs a median 7–11% extra compute, growing with extrapolation distance and with heavy tails (upper quartiles 48–113% at 100×, largest for the Kempner sweeps). These are counterfactual disagreements, not realized savings: they measure how much the unreported convention can change a frontier recommendation that public data do not consistently adjudicate.

5 Recommendations

Reporting. Table 3 lists the minimum disclosure needed to make exponents comparable. Of the public sources we examined, most do not meet it (Appendix A). Because an allocation stated in total parameters and physical FLOPs can be compared directly across studies, whereas an exponent is defined only relative to a parameter convention, we recommend reporting the former alongside the latter.

6 Limitations

The primary panel has six research groups, and only four training codebases; with four units, sign consistency and a t-interval rather than a rank test carry the codebase-level evidence. Slices within a group share pipelines, so all inference is at the group level. Kempner's vocabulary block is determined exactly ($64{,}000\,d$) but its split between embedding and head is not, so it is excluded from core+head analyses. The public sweeps are heterogeneous: they differ in data, tokenizers, optimizers, schedules, loss metrics and design geometry, so cross-lab confounding is real. The convention comparison is

therefore paired within each sweep (same runs, same fit, only N redefined), which removes these differences from Δr itself; they remain in any comparison of exponent levels across labs, which is one reason we do not pool exponents. All sweeps are dense decoder-only language models; mixture-of-experts, multimodal and other regimes may behave differently. The non-separable form is used only as a diagnostic: its extrapolation performance is not better than Chinchilla's (Appendix C), and none of our conclusions depend on it. Public data cannot identify which convention governs scaling; doing so conclusively will require training ladders that vary vocabulary size while holding the transformer fixed.

7 Conclusion

A compute-optimal exponent is not a property of a loss surface alone; it also depends on what is counted as a parameter. Across every public sweep we found that allows the question to be asked exactly, counting vocabulary matrices raises the exponent in every sweep and every group, by an amount comparable to the largest disagreement in the literature. Public data do not consistently say which count is right, and fixing the count would make exponents comparable without, as far as we can detect, making labs agree. Until releases report conventions and allocations in physical units, or controlled experiments identify the correct count, cross-study comparisons of scaling exponents should be read as comparisons of accounting choices as much as of scaling behavior.

AI Use Statement

In this work, we used generative AI tools for assistance with code implementation, code review, and prose editing. We designed the study, selected and audited the public data sources, ran and verified all analyses, interpreted the results, and wrote and approved the final manuscript. We did not use generative AI tools to collect data, run experiments, or generate the reported numerical results. We take responsibility for the final text, claims, code, and artifacts.

Ethics Statement

This work analyzes publicly released aggregate scaling-law data and does not involve human participants, private user data, or deployment of a learned system. Scaling-law recommendations influence how large amounts of compute are spent; we therefore report source-specific uncertainty rather than a universal prescription.

Reproducibility Statement

An anonymous reproducibility package accompanies the submission. A single command regenerates every number in the paper from public data: vendored tables where licenses permit, and files fetched from pinned commits and verified by SHA-256 otherwise. The package contains the fitting protocol, the exact parameter-count reconstructions (with assertions that fail if a source changes), and all analyses.

References

Tamay Besiroglu, Ege Erdil, Matthew Barnett, and Josh You. Chinchilla scaling: A replication attempt. *arXiv preprint arXiv:2404.10102*, 2024. URL https://arxiv.org/abs/2404.10102.

Akshita Bhagia, Jiacheng Liu, Alexander Wettig, David Heineman, Oyvind Tafjord, Ananya Harsh Jha, Luca Soldaini, Noah A. Smith, Dirk Groeneveld, Pang Wei Koh, Jesse Dodge, and Hannaneh Hajishirzi. Establishing task scaling laws via compute-efficient model ladders. In *Conference on Language Modeling (COLM)*, 2025. URL https://arxiv.org/abs/2412.04403.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. DeepSeek LLM: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024. URL https://arxiv.org/abs/2401.02954.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. URL https://arxiv.org/abs/2304.01373.

David Brandfonbrener, Nikhil Anand, Nikhil Vyas, Eran Malach, and Sham Kakade. Loss-to-loss prediction: Scaling laws for all datasets. *arXiv preprint arXiv:2411.12925*, 2024. URL https://arxiv.org/abs/2411.12925.

Leshem Choshen, Yang Zhang, and Jacob Andreas. A hitchhiker's guide to scaling law estimation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10683–10699, 2025. URL https://proceedings.mlr.press/v267/choshen25a.html.

Eric Czech, Zhiwei Xu, Yael Elmatad, Yixin Wang, and William Held. Problems with Chinchilla approach 2: Systematic biases in IsoFLOP parabola fits. *arXiv preprint arXiv:2603.22339*, 2026. URL https://arxiv.org/abs/2603.22339.

Nolan Dey, Gurpreet Gosal, Hemant Khachane, William Marshall, Ribhu Pathria, Marvin Tom, and Joel Hestness. Cerebras-GPT: Open compute-optimal language models trained on the Cerebras wafer-scale cluster. *arXiv preprint arXiv:2304.03208*, 2023. URL https://arxiv.org/abs/2304.03208.

Samir Yitzhak Gadre, Georgios Smyrnis, Vaishaal Shankar, Suchin Gururangan, Mitchell Wortsman, Rulin Shao, Jean Mercat, Alex Fang, Jeffrey Li, Sedrick Keh, et al. Language models scale reliably with over-training and on downstream tasks. *arXiv preprint arXiv:2403.08540*, 2024. URL https://arxiv.org/abs/2403.08540.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Will Held and Marin Community. Scaling laws that extrapolate 300x past the fit, 2026. URL https://openathena.ai/blog/delphi.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2203.15556.

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

Joshua Shay Kricheli, Alexander Lawrence Reid, Soumajyoti Sarkar, Venkata Gandikota, and Paulo Shakarian. Tokens-per-parameter coverage is critical for robust LLM scaling law extrapolation. *arXiv preprint arXiv:2605.08541*, 2026. URL https://arxiv.org/abs/2605.08541.

Houyi Li, Wenzhen Zheng, Qiufeng Wang, Zhenyu Ding, Haoying Wang, Zili Wang, Shijie Xuyang, Ning Ding, Shuigeng Zhou, Xiangyu Zhang, and Daxin Jiang. Predictable scale: Part II, Farseer: A refined scaling law in large language models. *arXiv preprint arXiv:2506.10972*, 2025a. URL https://arxiv.org/abs/2506.10972.

Margaret Li, Sneha Kudugunta, and Luke Zettlemoyer. (Mis)Fitting: A survey of scaling laws. In *International Conference on Learning Representations*, 2025b. URL https://arxiv.org/abs/2502.18969.

Tomasz Limisiewicz, Artidoro Pagnoni, Srini Iyer, Mike Lewis, Sachin Mehta, Alisa Liu, Margaret Li, Gargi Ghosh, and Luke Zettlemoyer. Compute optimal tokenization. *arXiv preprint arXiv:2605.01188*, 2026. URL https://arxiv.org/abs/2605.01188.

Ian Magnusson, Nguyen Tai, Ben Bogin, David Heineman, Jena D. Hwang, Luca Soldaini, Akshita Bhagia, Jiacheng Liu, Dirk Groeneveld, Oyvind Tafjord, Noah A. Smith, Pang Wei Koh, and Jesse Dodge. DataDecide: How to predict best pretraining data with small experiments. *arXiv preprint arXiv:2504.11393*, 2025. URL https://arxiv.org/abs/2504.11393.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. *arXiv preprint arXiv:2305.16264*, 2023. URL https://arxiv.org/abs/2305.16264.

Tim Pearce and Jinyeop Song. Reconciling Kaplan and Chinchilla scaling laws. *arXiv preprint arXiv:2406.12907*, 2024. URL https://arxiv.org/abs/2406.12907.

Tomer Porian, Mitchell Wortsman, Jenia Jitsev, Ludwig Schmidt, and Yair Carmon. Resolving discrepancies in compute-optimal scaling of language models. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.19146.

Nikhil Sardana, Jacob Portes, Sasha Doubov, and Jonathan Frankle. Beyond Chinchilla-optimal: Accounting for inference in language model scaling laws. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024. URL https://arxiv.org/abs/2401.00448.

Chaofan Tao, Qian Liu, Longxu Dou, Niklas Muennighoff, Zhongwei Wan, Ping Luo, Min Lin, and Ngai Wong. Scaling laws with vocabulary: Larger models deserve larger vocabularies. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2407.13623.

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. URL https://arxiv.org/abs/2205.01068.

A Source Registry

Table 4 lists every source examined and its status under the inclusion rule.

Table 4: Every source examined and its status under the inclusion rule of §3: (a) exact vocabulary-matrix count per run, (b) N and D varied independently, (c) final checkpoints of runs trained to their own budgets.

Source	Status	Reason
Marin: Delphi, Comma, DCLM, Nemotron (Held & Marin Community, 2026; Czech et al., 2026)	Primary	Per-run width and depth; exact decomposition for all 339 runs; three conventions.
Kempner, six corpora (Brandfonbrener et al., 2024)	Primary	Exact $64{,}000\,d$ vocabulary block; embedding/head split not identifiable.
OLMo ladder (Bhagia et al., 2025)	Primary	Released sizes exclude only the input embedding; three conventions.
Farseer (Li et al., 2025a)	Primary	`exact_N` and `N_add_emb` differ by one vocab matrix.
OpenLM: Gadre C4/RPJ/RW (Gadre et al., 2024)	Primary	Released N contains one $50{,}432 \times d$ matrix.
OpenLM: Porian RW/OWT2 (Porian et al., 2024)	Primary	Released width and depth; authors' counting function; three conventions.
(Mis)Fitting (Li et al., 2025b)	Primary	$N - N_{\text{no emb}} = 2 \times 50{,}432 \times d$ for every run; three conventions.
CoTok (Limisiewicz et al., 2026)	Sensitivity	Fails (a): global count also removes local encoder/decoder blocks.
OWT (ours)	Sensitivity	Fails (c): cumulative checkpoints over repeated data; three sizes.
Chinchilla (Hoffmann et al., 2022; Besiroglu et al., 2024)	Excluded	Fails (a): aggregate (N, D, L) tables without per-run architecture.
Llama 3 IsoFLOP (digitized) (Grattafiori et al., 2024)	Excluded	Fails (a): N derived from $C/6D$.
OPT (Zhang et al., 2022)	Excluded	Fails (c): one run per size; the loss table records intermediate checkpoints.
Datablations (Muennighoff et al., 2023)	Excluded	Fails (c) as released: the loss table records evaluations along training for three sizes per corpus.
Cerebras-GPT (Dey et al., 2023)	Excluded	Fails (b): one token budget per size (fixed tokens per parameter).
Pythia (Biderman et al., 2023)	Excluded	Fails (c): one schedule per size; intermediate checkpoints only.
Vocabulary scaling (Tao et al., 2024)	Excluded	Fails (b): the released runs fix non-vocabulary parameters and vary vocabulary size.
DataDecide (Magnusson et al., 2025)	Excluded	Fails (c): one run per size to 100 tokens per parameter; released results are downstream metrics without pretraining validation loss.

B Additional Convention Diagnostics

Core+head. Where the output head is identifiable, the FLOP-faithful core+head count gives exponents between the other two (Marin 1.19–1.38, Porian 1.15–1.21, OLMo 1.34, (Mis)Fitting 0.82) and, in the largest-model holdout, is never the most accurate of the three conventions.

Allocation mapping. The transformer-only surface is evaluated on the physical axis through $N_{\text{core}} = N_{\text{total}} - aN_{\text{total}}^{b}$, fitted per source to the released ladder, which keeps $0 < N_{\text{core}} < N_{\text{total}}$ at every extrapolated optimum; no optimum lies on the search boundary.

Delphi. Delphi is excluded from the top-20% comparison because its release defines a separate frontier holdout (81 IsoFLOP fit rows; seven held-out rows at three budgets from 10^{21} to 10^{23} FLOPs, four of them seed replicates).

OWT sensitivity sweep. Three decoder-only transformers with tied GPT-2 BPE embeddings: 10.2M parameters ($L = 8$, $d = 144$, 4 heads, $d_{\text{ff}} = 640$, learning rate 3×10^{-4}), 25.1M ($L = 8$, $d = 288$, 4 heads, $d_{\text{ff}} = 1152$, 2.8×10^{-4}) and 51.0M ($L = 8$, $d = 448$, 8 heads, $d_{\text{ff}} = 2048$, 2.5×10^{-4}). All use sequence length 1024, 65,536 tokens per optimizer step, AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 0.1, 300-step warmup, cosine decay to 10% of the peak rate, and one 40 GB A100-class GPU. Six checkpoints per model are cumulative passes over a fixed 500M-token OpenWebText cache, which is why the sweep fails criterion (c). Per-run logs are in the reproducibility package.

C Separability Diagnostics

The non-separable relaxation is used only to define $\hat{r}$ robustly; we report its fit quality for completeness on a separate 15-sweep panel assembled for this diagnostic. Because the separability test does not require exact parameter counts, this panel includes sources outside the inclusion rule of §3 (Chinchilla, OPT, Datablations, Cerebras-GPT and OWT) alongside the Kempner and Gadre sweeps. In-sample, a nested F-test rejects $\theta = 0$ on 13 of 15 sweeps, and on interpolating compute-stratified folds it improves held-out RMSE on 11 of 14 (median $+5.9\%$, CI $[+1.7\%, +14.0\%]$). Under extrapolation it does not: training on the lower 70% and testing on the rest, it beats Chinchilla on 8/14 sweeps by N, 8/14 by D, and 6/14 by compute, and on a joint split from below-median to above-median N and D it loses on all 11 evaluable sweeps. Separability therefore fails in-sample but the tested relaxation is not a better extrapolator, and we make no claim that it should replace the Chinchilla form.