How Much Is an AI Token Worth? Scaling Laws for Wild AI-Generated Web Text

Anonymous authors
Paper under double-blind review

Abstract

Web text makes up the majority of pretraining data and is increasingly AI-generated. After applying FineWeb quality filtering, we find that 27.5% of tokens from June 2026 web data are labeled as AI-generated by Pangram, rising to 31.1% by August. Unlike synthetic data or model-collapse setups, this *wild* AI text comes from many models, is written for human readers, and arrives unlabeled in pretraining corpora. How does AI text in the wild affect language model pretraining? To answer this question, we pretrain 800 language models (19.9M to 973M parameters), varying the ratio of added AI tokens to human tokens, and fit scaling laws to held-out losses on both human and AI-generated text. For data-starved models, adding AI tokens to pretraining data initially lowers loss on human text, but the benefit saturates as more are added and quickly *reverses* into harm. For models trained on high budgets of human text, AI tokens raise loss almost immediately, while the same number of fresh human tokens keeps lowering it. Existing scaling laws such as Hoffmann et al. (2022) which treat AI tokens as no different from human tokens, fail to predict this behavior. We propose a new scaling law with separate benefit and harm terms that allows the value of an AI token (measured in human token equivalents) to change sign while also reducing to Chinchilla in the absence of AI text. When fit on smaller models, our scaling law predicts the effect of AI text on held-out human-text loss for models up to 3.6× larger with 41% lower error than the best existing law over all AI ratios. It implies that training on unfiltered 2026 web text requires 1.3× as much compute as training on its human subset at 20 tokens per parameter, a gap that grows with the human budget. We recommend filtering AI text when the target is human text, repeating human text before expanding the training dataset with AI-generated web text, and reporting validation loss on human and AI text separately, since at a 2026 crawl's AI share, mixed validation sets hide the harm in 95.5% of harmful runs. AI text remains valuable when the target is AI text. We will release WildAI, an 83B-token corpus with AI, topic, and format labels, all 800 models and code.

1 Introduction

AI-generated text is flooding the web. In the June 2026 Common Crawl, Pangram labels 27.5% of the tokens that pass FineWeb's quality filters as AI-generated, up from 10.1% two years earlier and rising to **31.1%** in August 2026 (Figure 1). At the same time, models are trained far past the compute-optimal budget of ∼20 tokens per parameter (Hoffmann et al., 2022); for example, Qwen3-32B is trained on ∼1,100 tokens per parameter (Yang et al., 2025). Models are also projected to exhaust nearly all public human-written text by 2032 (Villalobos et al., 2024). Every new crawl thus forces a choice about what to do with the AI-generated documents: keep them, filter them, or find some optimal data mixture of AI and human text.

In this paper, we measure how AI-generated web text affects pretraining and fit scaling laws that predict its impact. No previous work directly addresses this question: model-collapse studies train a model recursively on its own output (Shumailov et al., 2024; Gerstgrasser et al., 2024), and synthetic-data studies add curated rephrasings designed to help performance in specific domains (Maini et al., 2024; Kang et al., 2025). We study a third kind, which we call *wild* AI text: text that language models wrote for human readers and that occurs naturally on the web, rather than being

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.

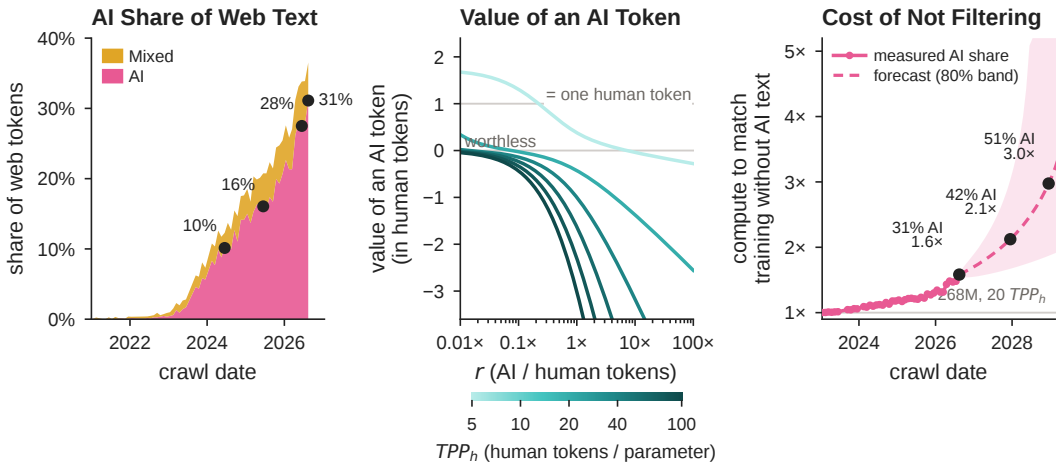


Figure 1: **AI-generated web text helps only data-starved models and costs compute otherwise.** Left: share of web tokens in documents Pangram 3.3.2 labels AI or Mixed, by month. The AI share alone is 10% in June 2024, 16% in June 2025, 28% in June 2026, and 31% in August 2026. Middle: our law's value of one added AI token at 268M, in human tokens, from 5 to 100 TPP_h: positive at 5 TPP_h, below zero from 20 TPP_h once r exceeds about 0.08. Right: compute an unfiltered crawl needs to match training on its human subset at 20 TPP_h (268M) for its AI share: 1.6× at August 2026's 31%, and 2.1× and 3.0× at the 42% and 51% forecasts for the ends of 2027 and 2028.

generated for training. It comes from many models, arrives in pretraining corpora unlabeled and mixed with human text, varies in quality, and makes up a rising share of the web.

We pretrain 800 models (19.9M to 973M parameters), adding up to 64 wild AI tokens per human token to fixed human corpora, and evaluate them on held-out human text (C4, FineWeb, and Paloma) as well as on AI text. AI text helps only data-starved models: it lowers loss on human text below about 10 human tokens per parameter, but from the Chinchilla-optimal 20 its benefit disappears, and larger amounts of AI tokens raise loss, while the same number of fresh human tokens still lowers it. Existing laws either treat AI tokens as human ones (Chinchilla), never let their value fall below zero (most repetition laws), or fit harm without separating it from benefit (most mixture laws). To address this, we propose a scaling law for wild AI-generated text that pairs a saturating benefit with a harm penalty that grows logarithmically and that reduces to Chinchilla without AI text. Fit on models up to 268M, it predicts models 3.6× larger with a paired error (the RMSE of the predicted change in log loss against each run's human-only control) of 0.83×10^{-3} on C4, against 1.41×10^{-3} for the best existing law (Shukor et al., 2025) and 4.32×10^{-3} for Chinchilla (Hoffmann et al., 2022).

Using our new law, we propose concrete recommendations for dealing with the rising share of wild AI text. Training on unfiltered web text at August 2026's AI share (31.1%) takes 1.6× the compute of training on its human subset, rising to 3.0× by 2028 at the forecast AI share, and repeating human text beats adding AI text. Current quality filters make this worse by favoring AI text. FineWeb's pipeline (Penedo et al., 2024a) keeps AI documents 2.3× as often as human ones, and the DCLM pipeline (Li et al., 2024) keeps AI documents 9.8× as often. AI text is easier to predict, so a validation set with a 2026 crawl's AI share hides the harm in 95.5% of harmful runs. For predicting AI text itself (Cosmopedia) the optimal training mix is over 90% AI. We release WILDAI, our 96M-document dataset with AI, topic, and format labels, and all 800 models.[1]

2 MEASURING AI-GENERATED WEB DATA

In this section, we describe how we create a large-scale dataset of web data, broken down into human-written and AI-generated corpora. We find that 27.5% of June 2026 tokens that pass the FineWeb filters are AI-generated, a share that is rising each month, that the AI share is uneven

[1]Data and code will be released after peer review.

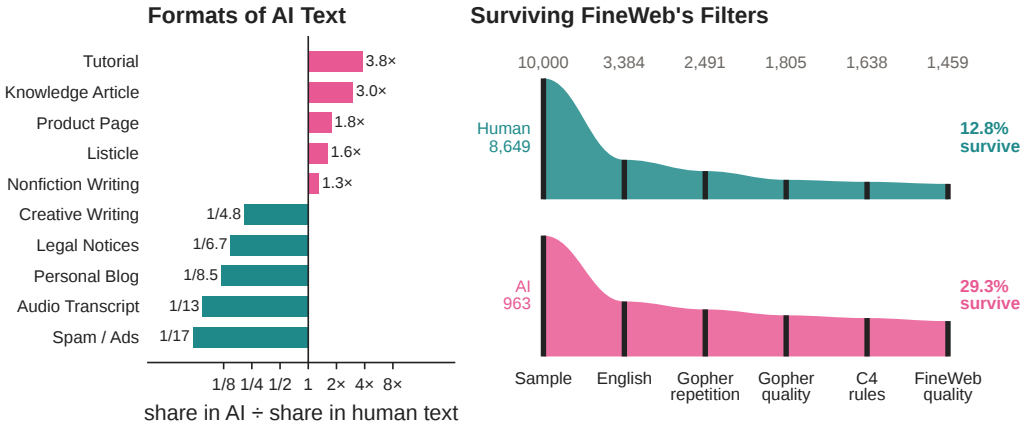


Figure 2: **AI-labeled web text leans toward a few formats and survives quality filters more often than human text.** Left: each format's share of AI-labeled tokens over its share of human-labeled tokens, January to June 2026, for the five most over- and under-represented formats. Right: share of AI- and human-labeled documents in a 2026 Common Crawl sample left after each FineWeb stage.

across topics and formats, and that existing pretraining quality filters (FineWeb and DCLM) favor it.

Collecting web data. We use the FineWeb corpus (Penedo et al., 2024a), a subset of Common Crawl (Common Crawl Foundation, 2007). We extend past the June 2025 cutoff point by replicating the FineWeb filtering process on Common Crawl data from July 2025 to June 2026.[2] We collect a dataset of 96.04M documents (83.31B tokens), which we call WILDAI.

Detecting AI-generated content. To understand the provenance of our data, we label each text as AI-generated or human-written. We first run EditLens Llama-3.2-3B (Thai et al., 2026) to label vast quantities of text (more than 280B tokens, from which we draw the 83B tokens of WILDAI). We then run the more accurate Pangram 3.3.2 on documents confidently labeled AI-generated and human-written as a second check, using those labels to create our final human-written corpus of 58.91M documents (42.19B tokens) and AI-generated corpus of 32.54M documents (35.32B tokens).[3] Pangram 3.3.2 has a reported false-positive rate of 0.05% and false-negative rate of 1.99% (Glickenhaus et al., 2026), accurate enough at this scale to separate the AI and human distributions. On pre-ChatGPT web text, Pangram 3.3.2 labels 0.062% of 60,000 documents from 2021 crawls AI.[4] We use WebOrganizer (Wettig et al., 2025) to extract topic and format labels per document.

AI-generated text is changing the makeup of the internet. To assess the increase in AI content on the web, we run Pangram on a sample of 310,000 documents, dated from January 2021 through August 2026.[5] In June of 2021, less than 0.1% of FineWeb tokens were AI-generated. By June 2024, it was 10.1%, 16.1% in June 2025, and 27.5% in June 2026 (see Figure 1, left, and Figure 8). The rise is accelerating: in August 2026 the share was 31.1%, 3.6 points above June (§A.2). Moreover, there is a large distribution mismatch between human-written and AI-generated content that continues to grow (He et al., 2026). In Figure 2, tutorials are 5.9% of human-labeled tokens and 22.3% of AI-labeled tokens, and personal blogs 9.3 and 1.1% (see Figure 7). Increasingly, the formats and topics of web data will reflect LLM outputs rather than human writing. We hypothesize that this may lead models to stray further from the human distribution on topics AI rarely writes about.

AI-generated data is more likely to survive quality filters. To understand the effects of different quality filtering processes, we audit a subset of 10,000 raw Common Crawl documents and track

[2] We use FineWeb v1.4.0; more details are in §A.

[3] Because EditLens preselects likely AI documents, AI text is 42% of WILDAI's tokens, far more than on the web. WILDAI is a pool for building training mixtures, not a natural sample of the web (§A).

[4] An upper bound on the FPR, since pre-2022 web text may itself contain some AI-generated text.

[5] 5,000 documents for each month that has an associated Common Crawl.

which portion of human- and AI-written texts make it through each step of the filtration process (Figure 2, see §A.1). We examine both FineWeb (Penedo et al., 2024a) and DCLM (Li et al., 2024) filtering processes, finding that AI-generated texts pass FineWeb's pipeline 2.3× (29.3% vs. 12.8%) and DCLM's full pipeline 9.8× (14.5% vs. 1.5%) as often as human-written documents. These filters are insufficient at filtering out AI-generated text, and in fact heavily prefer it.

3 Background and Related Work

Scaling law fundamentals. The key information needed to predict the performance of LLM training is training compute, C, model parameter count, N, and total training tokens processed, D, which one can use to estimate the loss on a held-out dataset. Training compute is commonly approximated as $C \approx 6ND$. Kaplan et al. (2020) characterized these trends, and the Chinchilla scaling law proposed by Hoffmann et al. (2022) formalized them as follows:

$$L(N, D) = E + \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}} \tag{1}$$

Scaling laws for repetition. In Chinchilla, the data term (B/D^{β}) depends only on the total token count D, without accounting for differences between unique tokens and repeated ones. To address this, Muennighoff et al. (2023) extended Chinchilla to account for repeated data by allowing for the effective value of a repeated token to be lower than a new unique token, but never below zero. Qin et al. (2026) propose that the effective value of a token relies not only on unique tokens but also on model size. Lovelace et al. (2026) alternatively propose that repetition is best modeled not by effective data but by adding an overfitting penalty.

Scaling laws for data mixtures. Other scaling laws model pretraining data from many domains or languages. Most of this body of work focuses on optimizing pretraining data mixtures (Jain et al., 2024; Ye et al., 2025; Shukor et al., 2025; Hamidieh et al., 2025; Sedova et al., 2026). Others model multilingual mixtures (He et al., 2025; Longpre et al., 2026).

Model collapse. Shumailov et al. (2024) study degradation under recursive training on model-generated data. Subsequent work examines how retaining original data or generating synthetic data from each successive model generation changes this behavior (Gerstgrasser et al., 2024; Seddik et al., 2024), and how collapse rates depend on the recursive estimation setting (Suresh et al., 2024). Schaeffer et al. (2025) show that loss increases, divergent error, and loss of distributional tails are distinct outcomes, while Dohmatob et al. (2024; 2025) show that synthetic data can affect scaling behavior even in the non-recursive setting. Kang et al. (2025) find that the benefits of synthetic rephrasings and textbooks mixed with web text depend on budget. Similarly, Zhu et al. (2025) show worsened performance on Paloma as synthetic text replaces natural text at a fixed token budget. Kazdan et al. (2025) show that the benefit of synthetic data depends on total data budget.

Wild AI text is neither synthetic nor surrogate data. Synthetic data is generated on purpose and curated to help, speeding pretraining 3–10× (Maini et al., 2024; Kang et al., 2025), although replacing natural text with it can still hurt (Zhu et al., 2025). Model collapse trains models on their own outputs, a setting that is not as harmful once human data is kept (Shumailov et al., 2024; Gerstgrasser et al., 2024; Kazdan et al., 2025). Wild AI text comes from many models, is written for readers and search engines rather than for training, and arrives mixed into human text, the setting that current pretraining pipelines actually face. Even the surrogate-data law of Jain et al. (2024) predicts our held-out C4 runs worse than Chinchilla (8.00 vs. 4.32, ours 0.83, all $\times 10^{-3}$).

4 Scaling Laws for Human-Written and AI-Generated Text

Models trained on only human text follow Chinchilla scaling laws, but when trained on a portion of AI-generated text, AI tokens initially help models with low budgets of human text but harm models with high budgets. We create five criteria for a scaling law to model this phenomenon, and propose our scaling law that fits these criteria. Our law has the lowest error predicting the effects of adding AI-generated text versus existing scaling laws (0.83 vs. the second best of 1.41 $\times 10^{-3}$ on C4).

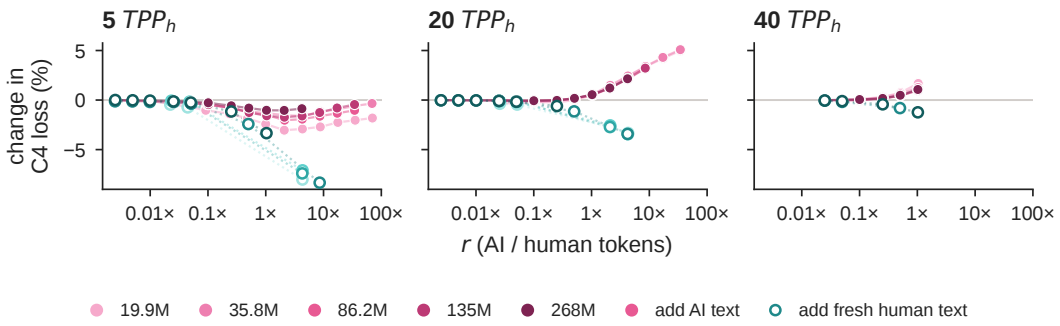


Figure 3: **Added AI text lowers loss only at small human budgets.** Change in C4 loss against each model's human-only control when AI tokens or the same number of fresh human tokens are added to a fixed human corpus, at 5, 20 and 40 TPP_h.

4.1 Empirical Observations

We train 800 models of varying sizes, human tokens per parameter (TPP_h), and ratios of AI to human data r. We find that Chinchilla scaling laws are inadequate at predicting the behavior of AI-generated text and propose a new scaling law that models both the benefits and harms of pretraining on AI-generated web text.

Experimental setup. We train 800 models varying in size from 19.9M to 973M, with 2.9 to 87.5 TPP_h, and r of 0 to 64. We use the nanochat architecture due to its fast training and high performance relative to model size (Karpathy, 2025).[6] We evaluate on human-, mixed-, and AI-generated text sets. We use C4 (Raffel et al., 2020) as our north star dataset to model human text, similar to the frontier Marin model evaluation (Marin Community, 2025). We also evaluate on Paloma (Magnusson et al., 2024), reporting the loss macro-averaged over its 16 sources, and a held-out set of pre-2022 FineWeb data (FW22). For mixed text, we evaluate on a held-out set of 2026 FineWeb data (FW26), which contains 22.3% AI-generated text. We also evaluate both the AI- and human-written splits of FineWeb 2026 (FW26-AI and FW26-H, respectively). To observe effects on purely AI text, we evaluate on Cosmopedia (Ben Allal et al., 2024, Cosmo). Human text is our main target. It is the language a pretrained model is meant to model, and loss on held-out text is a standard proxy for downstream performance (Huang et al., 2024; Gadre et al., 2025).

Models trained on AI text do not follow Chinchilla scaling laws. We find that for models where we add more human tokens, Chinchilla scaling laws perform as expected, but adapt poorly to predicting how models with AI-generated tokens will behave. On the reserved sizes, Chinchilla predicts the fresh-human additions to a paired error of 1.32×10^{-3} on C4 but the AI additions only to 4.32×10^{-3}. On models with low human token budgets, AI-generated text seems to almost always help. For Chinchilla-optimal models (20 or more TPP_h), AI text raises loss immediately, and more drastically at larger sizes, in line with the findings of Dohmatob et al. (2025). Figure 3 shows models at different human token budgets. With 5 TPP_h, AI decreases loss even at large AI budgets, but at 20 TPP_h, the AI additions increase loss with increasing AI budgets.

4.2 Criteria

Motivated by our observations, we formulate the following five criteria for a scaling law that better models the effects of adding AI-generated text:[7]

- C1 Flexible token value: the value of adding an AI token must be able to change according to N, TPP_h, and r.
- C2 Model both beneficial and harmful behavior: adding an AI token must be able to help loss in some scenarios and hurt in others.

[6] See §B for the architecture, training setup, and how N is counted, following Pearce & Song (2024).
[7] Figure 17 marks which benchmarked laws meet each of the five criteria.

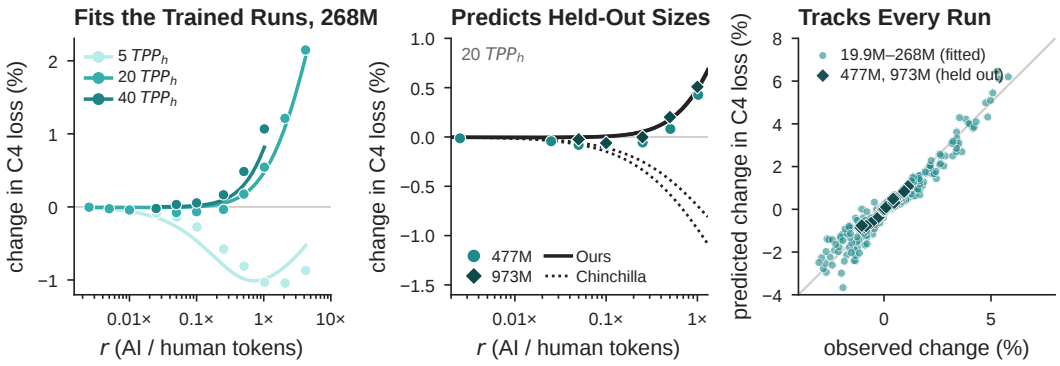


Figure 4: **Our law fits the trained models and predicts larger ones.** Left: observed change in C4 loss with our law. Middle: the held-out 477M and 973M runs at 20 TPP_h against our law and Chinchilla. Chinchilla predicts that AI text keeps lowering loss; the runs and our law turn upward. Right: predicted against observed change for every AI run.

C3 Interpretable: separate benefit and harm terms.

C4 A finite first-token value: the first AI token's value in human tokens is finite and not assumed to be 1.

C5 Recovery of Chinchilla with no AI data: when no AI data is added, the model must recover the Chinchilla scaling laws that govern unique human tokens.

4.3 Scaling Laws for AI-Generated Text

Our proposed law keeps the Chinchilla backbone of Hoffmann et al. (Equation 1), takes the saturating credit window of the compute–data (CD) law of Qin et al. (Equation 9) to represent the potential benefit of AI text (D_{eff}), and a penalty similar to Lovelace et al. (Equation 5) to represent the harm (H) of too much AI text. The credit, harm, and net value of an AI token are depicted in Figure 11. Our law has a paired error of 0.83×10^{-3} on C4's held-out sizes, lower than all other laws (Table 1). Representing human text as D_{H} and AI text as D_{A}, the loss is:

$$L(N, D_{\text{H}}, D_{\text{A}}) = E + \frac{A}{N^{\alpha}} + \frac{B}{D_{\text{eff}}^{\beta}} \big(1 + H\big) \tag{2}$$

The benefits of AI tokens. We model the saturating benefit of AI tokens as more are added. Looking to prior work on repeating tokens (Muennighoff et al., 2023; Qin et al., 2026), effective tokens increase at a decelerating rate over a window of added usefulness. Like both Muennighoff et al. and Qin et al., we use an exponential form of credit $g(r)$, but with a free first-token parameter η.[8] The benefit is as follows, where $r = D_{\text{A}}/D_{\text{H}}$, $R^{\star}$ is the saturation scale of the benefit, and $t = D_{\text{H}}/20N$:

$$D_{\text{eff}} = D_{\text{H}} \big(1 + \eta\, g(r)\big), \qquad g(r) = R^{\star}\big(1 - e^{-r/R^{\star}}\big), \qquad R^{\star} = K\, t^{\rho} \tag{3}$$

The harm of AI tokens. Qin et al. and Muennighoff et al. assume that repeated tokens can never *hurt* performance, only help less and less. Lovelace et al. (2026) model the harm of repetition with an additive penalty, finding that the harm accelerates over epochs. We observe that the harm of adding more AI tokens *decelerates* logarithmically (Figure 13), motivating our penalty:

$$H = \gamma\, t^{u}\, n^{v} \Big[\log(1 + r) - \frac{r}{1 + r}\Big] \tag{4}$$

[8] Table 8 in §D lists every benefit and harm form we tried.

Here γ sets the size of the harm, $n = N/10^8$ is the model size, and the exponent v lets the harm change with model size as u lets it change with the human budget. Subtracting the AI share $r/(1+r)$ removes the initial harm of the first AI tokens. The harm grows as $r^2/2$ at small additions of AI and as $\log r - 1$ at large ones, and the marginal harm of an AI token, proportional to $r/(1+r)^2$, peaks where AI and human tokens are equal ($r = 1$).

Relative scaling laws. Held et al. (2025) propose relative scaling laws, which measure the effect of some treatment (here, added AI text, $r > 0$) against a baseline (human text only, $r = 0$). Since our scaling law reduces to Chinchilla (Hoffmann et al., 2022) at $r = 0$, we can predict the human-only control easily, and isolate the effect of r. For every run we predict its change in log loss against its own human-only control, $\Delta = \log \widehat{\mathcal{L}}(N, D_{\mathrm{H}}, D_{\mathrm{A}}) - \log \widehat{\mathcal{L}}(N, D_{\mathrm{H}}, 0)$, and score the RMSE of predicted against observed Δ over the AI runs.[9]

Benchmarking existing scaling laws. We fit our scaling law on 726 models of 19.9M to 268M parameters, and evaluate our fit on 74 models of 477M and 973M parameters, $1.8\times$ and $3.6\times$ the largest fitted size (Table 5). We benchmark existing scaling laws developed for repetition, data mixing, and surrogate data in Table 1.[10] Our law has the lowest paired error across all four human sources (0.83 $\times 10^{-3}$ on C4).[11] Shukor et al.'s joint law (Equation 11), which treats AI and human text as two separate data mixtures, is the best of the existing laws on human text (1.41 $\times 10^{-3}$ on C4) and is the best of all laws when the target text is AI-generated.[12]

Table 1: Paired RMSE $\times 10^3$ on the held-out sizes, lower is better.

Law	k	C4	FW22	FW26-H	Paloma
Chinchilla (Hoffmann et al., 2022)	5	4.32	4.65	5.40	9.15
Muennighoff et al. (2023)	7	3.96	4.21	4.94	9.15
CD (Qin et al., 2026)	8	3.68	3.63	4.18	9.15
Lovelace et al. (2026)	9	3.55	3.74	4.13	8.38
ATLAS (Longpre et al., 2026)	6	4.36	4.72	5.51	9.15
He et al. (2025)	6	6.54	6.55	6.60	8.96
Hamidieh et al. (2025)	7	3.95	4.19	4.85	9.15
Shukor additive (Shukor et al., 2025)	9	6.77	6.70	6.26	10.05
Shukor joint (Shukor et al., 2025)	13	1.41	1.50	2.24	7.80
Sedova et al. (2026)	8	9.60	9.59	9.56	13.32
Jain et al. (2024)	7	8.00	8.03	8.03	9.34
Ours	11	**0.83**	**0.85**	**1.28**	**6.77**

5 The Effects of Training on AI Text and What to Do About It

We forecast that over half (50.7%) of FineWeb tokens will be AI-generated by the end of 2028. An unfiltered crawl at our 2026 crawl's AI share (22.3%) needs $1.3\times$ the compute, and $3.0\times$ in 2028, to match a filtered model at 20 TPP_h. We recommend filtering AI text from web data and, if data is short, repeating human tokens rather than adding AI tokens. AI text is worth training on when the target is AI text, and evaluations should report the two separately.

5.1 The Cost of Not Filtering

Filtering out AI-generated data. Mohri et al. (2026) find that with sufficient compute, it is best to not filter data at all for pretraining. To test this for AI-generated text, we train models at the 2026 natural rate of AI (22.3%), and paired models of the same size with only the human subset of pretraining data. Despite having less data, models above about 10 to 15 TPP_h (less at larger sizes) had lower loss on C4, as shown in Figure 5. Our law gets the direction of filtering right in 18 of the 19 pairs (paired error 3.28×10^{-3}), while Chinchilla predicts that removing the AI text raises loss in all 13 pairs where it lowers it. We also find that the more human-written tokens a model has, the larger the benefit in loss reduction is. As most models today are trained far past 20 tokens per parameter, we recommend removing all AI-generated web text for human next-token prediction.

Forecasting the future AI share of web text. Using our historical sample of 5,000 documents per month, we use a random walk with drift (Hyndman & Athanasopoulos, 2021) to forecast the future

[9] Absolute RMSE is in Table 10.

[10] §C.3 defines the laws; §E has full results.

[11] On human text our law's lead over the best existing law is statistically significant over all AI ratios (§E.1).

[12] Results on mixed and AI-generated target texts are in Table 9.

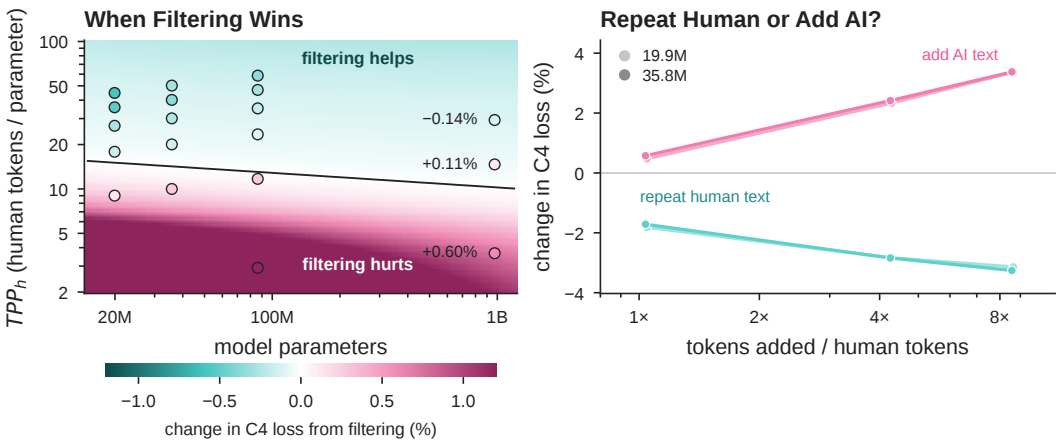


Figure 5: **Filtering AI text pays off at larger human budgets, and repeating human text beats adding AI text.** Left: change in loss from removing the AI documents of a 22.3% AI mix without replacing them, predicted by our law (color; black line: no change) and measured (circles; values for 973M). Right: change in loss when the human corpus is repeated or the same number of new AI tokens is added, for models at 20 TPP_h.

proportion of AI tokens.[13] Extrapolating the average monthly rise, it forecasts 42.3% AI tokens by the end of 2027 and 50.7% by the end of 2028. Compute-equivalent gain (CEG) (Davidson et al., 2023) compares the compute efficiency of any two models. Using a model trained with 22.3% AI (the AI share of our 2026 crawl) as our reference model, our law predicts that the reference (an unfiltered model) needs 1.3× the compute to match human-only training at 268M parameters and 20 TPP_h (Equation 17). At the forecast AI share, by the end of 2027 a model will need 2.1× the compute to match a filtered model's performance on C4, and 3.0× by the end of 2028 (Figure 1).

Is it better to repeat human data or add AI data? Muennighoff et al. (2023) and Qin et al. (2026) model repeated tokens as effective tokens, and Lovelace et al. (2026) with an overfitting penalty. To compare the effects of repeating human tokens and adding AI tokens, we train 19.9M and 35.8M models at 20 TPP_h that add 1, 4, or 8 times the human corpus in tokens, either by repeating it or by adding AI text. For C4, loss after one added epoch was 2.3% lower when repeating human tokens than when adding fresh AI tokens, a gap that expands to 6.3–6.4% after eight added epochs (nine in total).[14]

5.2 When Is AI Text Worth Training On?

It helps in data-starved settings and for AI target text. When predicting on human text, AI text helps only when human training data is scarce: the loss-minimizing AI share on C4 is 37% at 5 TPP_h but 1.0% by 20 TPP_h, and FW22 and Paloma behave similarly (Figure 6, §F.2). On AI-generated text, the optimal share stays above 90% at every budget, and our law values the first AI token at more than 20 human tokens (Figure 21).

Training on AI text changes how models write. In agentic workflows, models increasingly prompt other models and read their outputs, so a model must understand AI-written input as well as human input. Understanding such text need not mean writing it, yet training on AI text changes how a model writes: at 268M, AI-typical phrases per 1,000 words rise from 0.16 without AI text to 0.36 at $r = 1$ and 0.54 at $r = 4$, and at $r = 2$ Pangram labels twice as many stories AI (37.2% vs. 18.6%, Table 13).[15]

[13] More details about our forecasting methodology and predictions are in §A.2.
[14] See §F.3 for experimental setup details and additional results.
[15] See §G for details on AI-typical phrases and Pangram labeling.

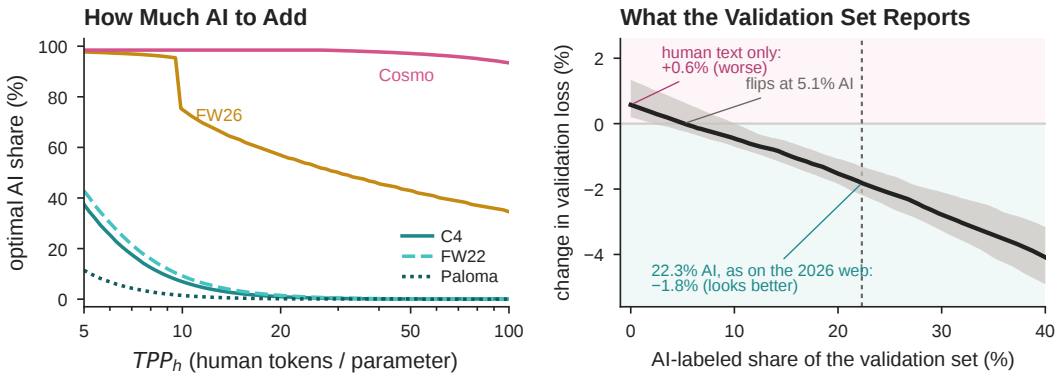


Figure 6: **Whether AI text helps depends on the evaluation set.** Left: the AI share of training tokens that minimizes our law's predicted loss with the human corpus fixed, per evaluation set: near zero for human text (C4, FW22), above 90% for AI text (Cosmo), and falling from 98% to 35% for mixed FW26; the drop near 10 TPP_h is where the optimum moves between two local minima of the predicted loss. Right: for the 244 AI additions that raise loss on human-labeled FW26 text, the median change in loss that a validation set reports as the AI-labeled share of its text grows. The sign flips at 5.1% AI; at our 2026 crawl's 22.3% the median run looks 1.8% better.

Treat AI and human text as separate domains, in training and in evaluation. Mixture laws that model the two as separate domains, such as the joint law of Shukor et al. (2025), are the most accurate existing laws in our benchmark, and the split could go further, since AI and human web text differ in format (tutorials are 22.3% of AI-labeled tokens but 5.9% of human-labeled ones). Because AI text is easier to predict, all 553 models with added AI text lower loss on AI-labeled FW26 text, while raising loss on human-labeled text in 44% of models. At our 2026 crawl's 22.3% AI share, a mixed validation set reports 95.5% of these harmful runs as improvements (Figure 6). Tuned against C4, our law recommends 1.0% AI at 20 TPP_h, and tuned against FW26, 57% but only 2.0% on the human subset. We recommend reporting loss on human and AI text separately.

6 Conclusion

Wild AI text now makes up over a quarter of filtered web crawls, and its share is rising. Across 800 pretrained models, it helps only data-starved models, increasing loss on Chinchilla-optimal models. Our law pairs a saturating benefit with a harm that grows logarithmically, reduces to Chinchilla without AI text, and predicts models 3.6× larger than those it is fit on with lower error on human text than all eleven existing laws we test. It implies that training on unfiltered 2026 web text takes 1.3× the compute of training on its human subset at our 2026 crawl's AI share (22.3%), rising to 3.0× by 2028, that human text should be repeated before AI text is added, and that validation loss should be reported on human and AI text separately. We will release WildAI, with its AI, topic, and format labels, code, and all 800 models.

Limitations and future work. We fit our law on models up to 268M parameters and test it up to 973M. Larger models memorize more of their training data (Carlini et al., 2023) and can be hurt more by model-generated data (Dohmatob et al., 2025), so they may pick up the patterns of weaker LLMs more readily. They might instead learn AI and human text as related tasks that help each other, a turn the joint law of Shukor et al. (2025) can express. We measure next-token loss, and at our sizes AI text raises CORE scores (Li et al., 2024) about as much as fresh human text does (§G), although all models are too small to have meaningful performance, and these are the tasks AI-generated text was meant to optimize. We study only pretraining, on English web text labeled by Pangram. Future work could (1) filter wild AI text, as MAI does (The Microsoft AI Team, 2026), and add targeted synthetic data, (2) find the domains in which AI text helps, (3) measure how the AI text already in pretraining corpora has affected the quality of current models, and, as agents read other models' outputs, (4) train models that understand AI-written input without writing like it, for example by tagging AI text (Caswell et al., 2019) or masking its loss (Lin et al., 2024).

AI USE STATEMENT

We used AI tools for the following tasks:

- Formulate mathematical claims: To iterate on the exact formulation of the benefit and harm terms in our scaling law, we ask coding agents to propose and test many forms of the scaling laws.
- Propose or refine hypotheses: To strengthen the paper, we ask agents to suggest additional experiments to understand our scaling laws. The repetition experiment was proposed by a coding agent.
- Design or provide feedback on research methodology or experiments: Coding agents were asked to provide feedback many times throughout the project, as well as design and run the training of all additive models.
- Implement methods: Coding agents implemented all experimental code.
- Clean and reformat dataset: To extend our FineWeb dataset through June 2026, we used a coding agent to execute the FineWeb filtration process on recent Common Crawl data.
- Support qualitative and thematic data analysis: We often asked coding agents to create tables and figures of results so we could interpret the results. Many of these figures and tables ended up in the final draft.

Additional disclosures include the use of AI for the following purposes: create or modify scientific figures or images, create or edit software code, creation of artifacts, brainstorming, sourcing/searching for information, edit a research paper to improve readability, and identify relevant literature.

We have reviewed and validated all AI-assisted work and take responsibility for the final content of the work, including text, claims, and artifacts produced with the aid of AI.

REPRODUCIBILITY STATEMENT

Data. WILDAI is FineWeb v1.4.0 extended to June 2026 by rerunning the FineWeb pipeline on Common Crawl (§2, §A); every document carries its EditLens and Pangram 3.3.2 provenance labels and WebOrganizer topic and format labels. We release all 83B tokens of WILDAI with these document-level labels.

Models. All 800 models use one training recipe (§B) at the sizes, budgets and run grid of Table 3, Table 6 and Table 5. We release every checkpoint together with its loss on the seven evaluation sets, so that all fits, tables and figures can be regenerated on a CPU in minutes without retraining.

Scaling laws. The law is Equation 2 to Equation 4 with the fitted coefficients of Table 7; the objective, starting points, bounds and the fixed development/reserved split are in §D; every comparator is written in our coordinates in §C.3 and scored by the same protocol (§E); the alternative forms we tried are in Table 8.

Code. The data pipeline, labeling, training, evaluation, law fitting and every figure and table script are released with the data and models.

REFERENCES

Loubna Ben Allal, Anton Lozhkov, Guilherme Penedo, Thomas Wolf, and Leandro von Werra. Cosmopedia, February 2024. URL https://huggingface.co/datasets/HuggingFaceTB/cosmopedia.

George E. P. Box and Gwilym M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, 1970.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=TatRHT_1cK.

Isaac Caswell, Ciprian Chelba, and David Grangier. Tagged back-translation. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor (eds.), *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pp. 53–63, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5206. URL https://aclanthology.org/W19-5206/.

Common Crawl Foundation. Common Crawl Web Archive. https://commoncrawl.org, 2007. Accessed: 2026-09-07.

Tom Davidson, Jean-Stanislas Denain, Pablo Villalobos, and Guillem Bas. AI capabilities can be significantly improved without expensive retraining, 2023. URL https://arxiv.org/abs/2312.07413.

Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. A tale of tails: Model collapse as a change of scaling laws. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=KVvku47shW.

Elvis Dohmatob, Yunzhen Feng, Arjun Subramonian, and Julia Kempe. Strong model collapse. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=et5l9qPUhm.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082. URL https://aclanthology.org/P18-1082/.

Samir Yitzhak Gadre, Georgios Smyrnis, Vaishaal Shankar, Suchin Gururangan, Mitchell Wortsman, Rulin Shao, Jean Mercat, Alex Fang, Jeffrey Li, Sedrick Keh, Rui Xin, Marianna Nezhurina, Igor Vasiljevic, Luca Soldaini, Jenia Jitsev, Alex Dimakis, Gabriel Ilharco, Pang Wei Koh, Shuran Song, Thomas Kollar, Yair Carmon, Achal Dave, Reinhard Heckel, Niklas Muennighoff, and Ludwig Schmidt. Language models scale reliably with over-training and on downstream tasks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=iZeQBqJamf.

Everette S. Gardner and Ed. McKenzie. Forecasting trends in time series. *Management Science*, 31 (10):1237–1246, 1985.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Tomasz Korbak, Henry Sleight, Rajashree Agrawal, John Hughes, Dhruv Bhandarkar Pai, Andrey Gromov, Dan 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 *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=5B2K4LRgmz.

Ben Glickenhaus, Katherine Thai, Jenna Russell, Elyas Masrour, Yue Han, Max Spero, and Bradley Emi. Pangram 4 technical report, 2026. URL https://arxiv.org/abs/2607.27183.

Kimia Hamidieh, Lester Mackey, and David Alvarez-Melis. Domain-aware scaling laws uncover data synergy. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, 2025. URL https://openreview.net/forum?id=FndNAs9s0d.

Sichang Steven He, Calvin Ardi, Ramesh Govindan, and Harsha V. Madhyastha. DeGenTWeb: A first look at LLM-dominant websites, 2026. URL https://arxiv.org/abs/2605.00087.

Yifei He, Alon Benhaim, Barun Patra, Praneetha Vaddamanu, Sanchit Ahuja, Parul Chopra, Vishrav Chaudhary, Han Zhao, and Xia Song. Scaling laws for multilingual language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4257–4273,

Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.221. URL https://aclanthology.org/2025.findings-acl.221/.

William Held, David Hall, Percy Liang, and Diyi Yang. Relative scaling laws for LLMs, 2025. URL https://arxiv.org/abs/2510.24626.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=iBBcRUlOAPR.

Yuzhen Huang, Jinghan Zhang, Zifei Shan, and Junxian He. Compression represents intelligence linearly. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=SHMj84U5SH.

Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, Melbourne, Australia, 3rd edition, 2021. URL https://otexts.com/fpp3/. Accessed: 2026-09-17.

Ayush Jain, Andrea Montanari, and Eren Sasoglu. Scaling laws for learning with real and surrogate data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=NAcHv7vtL2.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

Feiyang Kang, Newsha Ardalani, Michael Kuchnik, Youssef Emad, Mostafa Elhoushi, Shubhabrata Sengupta, Shang-Wen Li, Ramya Raghavendra, Ruoxi Jia, and Carole-Jean Wu. Demystifying synthetic data in LLM pre-training: A systematic study of scaling laws, benefits, and pitfalls. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10739–10758, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.544. URL https://aclanthology.org/2025.emnlp-main.544/.

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, 2020. URL https://arxiv.org/abs/2001.08361.

Andrej Karpathy. nanochat: The best ChatGPT that $100 can buy, 2025. URL https://github.com/karpathy/nanochat.

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 *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=buwLCdOHxO.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej

Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alexandros G. Dimakis, Yair Carmon, Achal Dav, Ludwig Schmidt, and Vaishaal Shankar. Datacomp-lm: in search of the next generation of training sets for language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, yelong shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, and Weizhu Chen. Not all tokens are what you need for pretraining. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=0NMzBwqaAJ.

Shayne Longpre, Sneha Kudugunta, Niklas Muennighoff, I-Hung Hsu, Isaac Rayburn Caswell, Alex Pentland, Sercan O Arik, Chen-Yu Lee, and Sayna Ebrahimi. ATLAS: Adaptive transfer scaling laws for multilingual pretraining, finetuning, and decoding the curse of multilinguality. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=0BkvUY61MX.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Justin Lovelace, Christian Belardi, Srivatsa R Kundurthy, Shriya Sudhakar, and Kilian Q Weinberger. Prescriptive scaling laws for data constrained training. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=9bzQi6VPS3.

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 *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=yjj8ele147.

Pratyush Maini, Skyler Seto, Richard Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. Rephrasing the web: A recipe for compute and data-efficient language modeling. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14044–14072, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.757. URL https://aclanthology.org/2024.acl-long.757/.

Marin Community. Marin 8B retrospective. Marin Documentation, 2025. URL https://marin.readthedocs.io/en/latest/reports/marin-8b-retro/. Accessed: 2026-09-16.

Christopher Mohri, John Duchi, and Tatsunori Hashimoto. A bitter lesson for data filtering, 2026. URL https://arxiv.org/abs/2605.19407.

Niklas Muennighoff, Alexander M Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=j5BuTrEj35.

OpenAI. Introducing ChatGPT. https://openai.com/index/chatgpt/, 2022. Released 30 November 2022.

Tim Pearce and Jinyeop Song. Reconciling kaplan and chinchilla scaling laws. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=NLoaLyuUUF. Reproducibility Certification.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024a. URL https://openreview.net/forum?id=n6SCkn2QaG.

Guilherme Penedo, Hynek Kydlíček, Alessandro Cappelli, Mario Sasko, and Thomas Wolf. DataTrove: Large scale data processing. https://github.com/huggingface/datatrove, 2024b.

Tian Qin, Kimia Hamidieh, and David Alvarez-Melis. Bridging compute- and data-optimal pretraining, 2026. URL https://arxiv.org/abs/2607.25271.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):5485–5551, 2020. URL http://jmlr.org.

Rylan Schaeffer, Joshua Kazdan, Alvan Caleb Arulandu, and Sanmi Koyejo. Position: Model collapse does not mean what you think, 2025. URL https://arxiv.org/abs/2503.03150.

Mohamed El Amine Seddik, Suei-Wen Chen, Soufiane Hayou, Pierre Youssef, and Merouane Abdelkader DEBBAH. How bad is training on synthetic data? a statistical analysis of language model collapse. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=t3z6UlV09o.

Anastasiia Sedova, Skyler Seto, Natalie Schluter, and Pierre Ablin. Scaling laws for mixture pretraining under data constraints, 2026. URL https://arxiv.org/abs/2605.12715.

Mustafa Shukor, Louis Béthune, Dan Busbridge, David Grangier, Enrico Fini, Alaaeldin El-Nouby, and Pierre Ablin. Scaling laws for optimal data mixtures. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=vVU1KTOsju.

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, 2024. URL https://arxiv.org/abs/2412.17646.

Katherine Thai, Bradley Emi, Elyas Masrour, and Mohit Iyyer. EditLens: Quantifying the extent of AI editing in text. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=gOkitaPCfZ.

The Microsoft AI Team. Mai-thinking-1: Building a hill-climbing machine, 2026. URL https://microsoft.ai/pdf/mai-thinking-1.pdf.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Position: Will we run out of data? Limits of LLM scaling based on human-generated data. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Lilian Weng. Scaling laws, carefully. *lilianweng.github.io*, June 2026. URL https://lilianweng.github.io/posts/2026-06-24-scaling-laws/.

Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. Organize the web: Constructing domains enhances pre-training data curation. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=boSqwdvJVC.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=jjCB27TMK3.

Xuekai Zhu, Daixuan Cheng, Hengli Li, Kaiyan Zhang, Ermo Hua, Xingtai Lv, Ning Ding, Zhouhan Lin, Zilong Zheng, and Bowen Zhou. How to synthesize text data without model collapse? In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=ihUi76a4u7.

A DATA

Collecting web data. WILDAI combines two samples of English web text. The first is drawn from the 16 released FineWeb dumps, CC-MAIN-2024-10 to CC-MAIN-2025-26 (Penedo et al., 2024a). The second extends FineWeb to the twelve Common Crawl dumps from July 2025 to June 2026 (CC-MAIN-2025-30 to CC-MAIN-2026-25): we rerun FineWeb's DataTrove pipeline (Penedo et al., 2024b) on the first 2.6% of each dump's WARC files, keeping 5.9M documents (4.6B tokens). The final dataset has 96.04M documents and 83.31B tokens.

Labeling with EditLens and Pangram. EditLens (Llama-3.2-3B) first sorts documents into four buckets from human-written to AI-generated, averaging over up to three evenly spaced 512-token windows. From its labels we take AI documents and human documents, dropping its mixed category. Pangram then labels every candidate with Pangram 3.3.2. We use the Pangram label to make the final dataset, comprising 58.91M human documents (42.19B tokens) and 32.54M AI documents (35.32B tokens), with the 4.59M Mixed documents left out. The monthly AI shares of §2 come from a separate sample of 5,000 documents drawn at random from each of the 60 crawl months between January 2021 and June 2026, taken from FineWeb through June 2025 and from our extension after.

False positives on pre-ChatGPT text. Web text crawled before ChatGPT's release in November 2022 (OpenAI, 2022) should contain almost no AI text, so the share Pangram labels AI there bounds its false-positive rate on web documents. Of the 60,000 documents in the monthly sample from the twelve 2021 crawls, Pangram 3.3.2 labels 37 AI (0.062%, 95% interval 0.045 to 0.085%; 0.04% of tokens) and 25 Mixed, in line with its reported false-positive rate of 0.05%. Some of these documents may be genuinely machine-generated, since earlier language models already wrote web text, so the false-positive rate on web text is at most this.

AI formats and topics. We find that the makeup of the web is changing over time. Product pages grow from 15 to 20% of tokens while personal blogs halve. By 2026, more than half of tutorial and knowledge-article tokens are AI-labeled, against about 10% of news and 4% of personal-blog tokens, as shown in Figure 8. In our training pool, which was assembled to supply AI text (about 34% AI documents, against 8 to 13% on the web in 2024 and 2025), knowledge articles (60% of documents), tutorials (59%) and listicles (53%) lead the formats, against about 6% for comment sections, audio transcripts and spam. Fashion and beauty, home and hobbies, games and finance (41 to 46%) lead the topics, and politics (12%) trails, as shown in Figure 7.

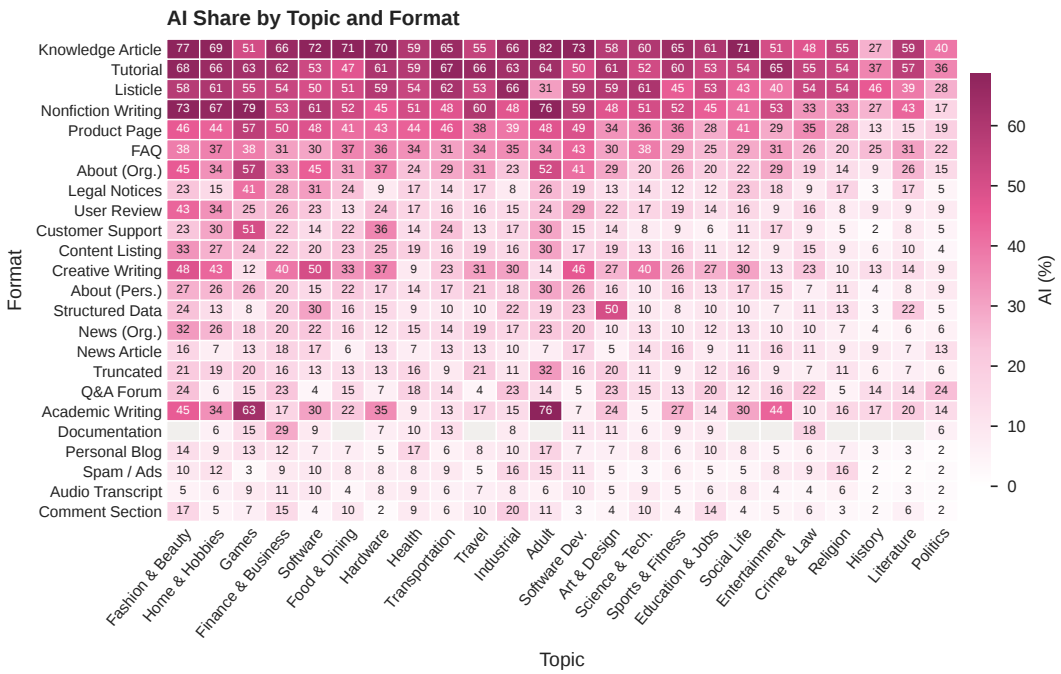


Figure 7: Knowledge articles, tutorials and listicles are the most AI-labeled formats in nearly every topic. Share of documents Pangram 3.3.2 labels AI in each WebOrganizer topic and format of our training pool documents.

A.1 Effects of Quality Filtering

To understand the effects of quality filtration systems on AI-generated text in relation to human text, we sample 10,000 documents from 2026 Common Crawl. We replicate the FineWeb and DCLM filtration processes, and trace the proportion of AI-generated text that passes through each step in relation to the proportion of human text that survives. The filtration process is depicted in Figure 9. Each pipeline runs on its own extraction of a 10,000-document sample from the same 32 WARC files of CC-MAIN-2026-30: FineWeb's on text extracted with Trafilatura, as FineWeb does, and DCLM's on text extracted with Resiliparse, as DCLM does, so the two samples hold different numbers of documents per label.

AI text survives both pipelines more often. FineWeb keeps 29.3% of AI-labeled documents (282 of 963) and 12.8% of human-labeled ones (1,104 of 8,649), 2.3× as often (95% interval 2.1 to 2.6×). AI documents pass its English filter more often (45% against 33%), and among English documents its quality heuristics keep 64% of AI documents and 39% of human ones. DCLM keeps 14.5% of AI documents (84 of 578) and 1.5% of human ones (129 of 8,697), 9.8× as often (7.5 to 12.7×). The gap opens at its learned classifier, which keeps 40% of the AI documents that reach it and 9.5% of the human ones. AI documents survive more often within the same format, so the gap is not only a matter of format mix: among English documents, DCLM keeps 53% of AI-labeled knowledge articles against 17% of human ones, and 49% of AI tutorials against 20%.

Which human documents do filters remove? Most removed human-labeled documents fail the English filters or are extraction debris (93% of truncated pages, 88% of content listings and 85% of structured data are removed after FineWeb's English filter), but DCLM's learned classifier also removes most human prose that reaches it, including 111 of 117 news articles and 106 of 117 personal blogs.

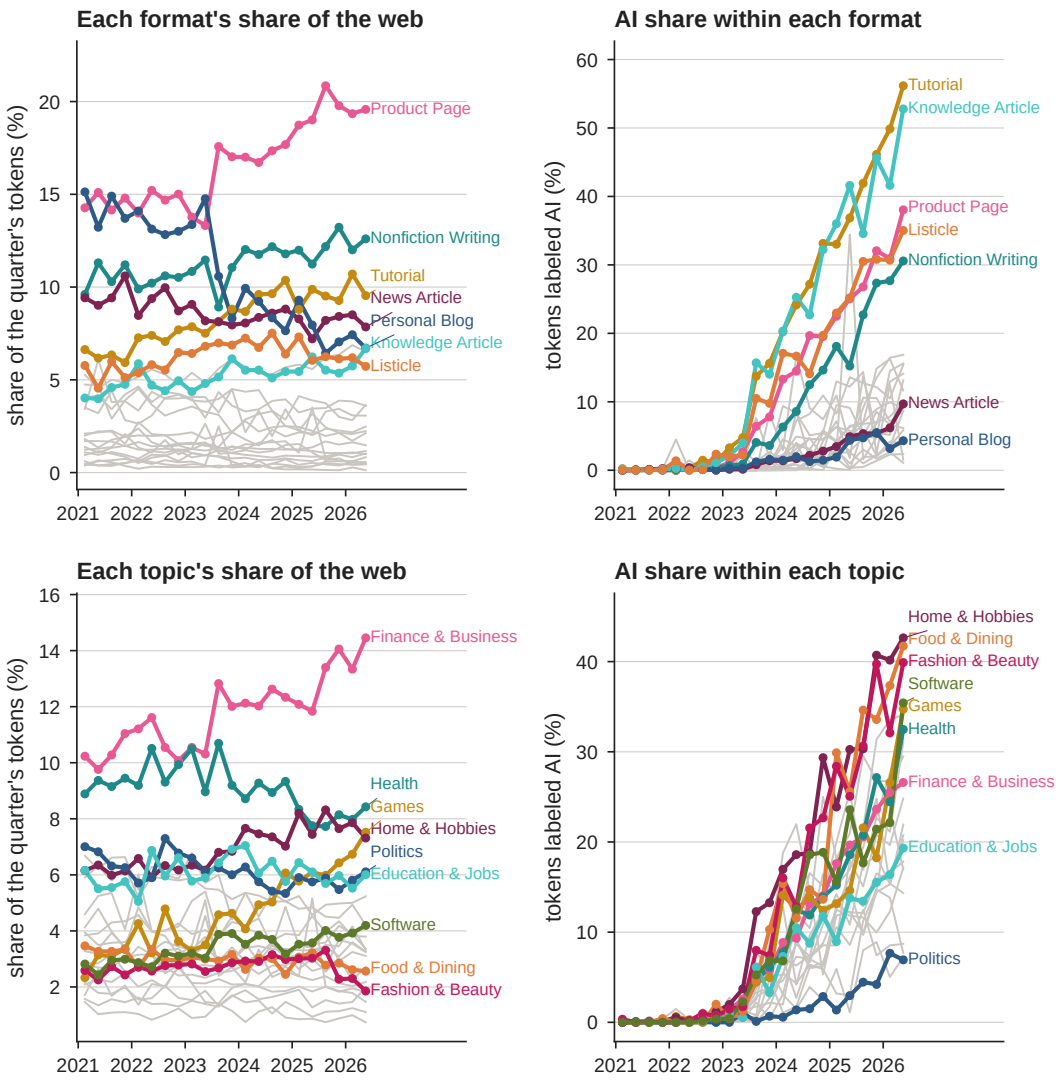


Figure 8: The web shifts toward the formats and topics AI writes most. We plot January 2021 to June 2026, Pangram 3.3.2 and WebOrganizer labels by calendar quarter. Each format's and topic's share of web tokens and the share of each format's and topic's tokens labeled AI.

A.2 Forecasting the AI Share of the Web

Every month of the series covers its whole crawl. Up to June 2025 the 5,000 documents come from FineWeb, whose dumps span every WARC file of a crawl and which we read in a seeded random shard order. From July 2025 on, they come from 32 WARC files drawn at random from each crawl and run through the FineWeb filters. To forecast our future rates of AI, we take a sample of 10,000 documents from July and August 2026, past our data cutoff dates, as an evaluation sample. We find the rate of AI tokens in the samples was 27.9% for July and 31.1% for August. We compare several forecasting methods on held-out months, including ones that respect the bounds of a share (the lowest proportion of the web AI tokens can hold is 0, and the highest 100). We try ARIMA (Box & Jenkins, 1970), random walks with drift (Hyndman & Athanasopoulos, 2021), ordinary damped Holt (Gardner & McKenzie, 1985), and logit-damped methods (a damped trend on the logit of the share; Hyndman & Athanasopoulos, 2021) and choose a random walk with drift because it gave the most accurate predictions on months since ChatGPT (OpenAI, 2022) was introduced in late November 2022. The July and August samples come from crawls after our June 2026 data cutoff (CC-MAIN-2026-30 and CC-MAIN-2026-34). They first served as a check on forecasts fit through

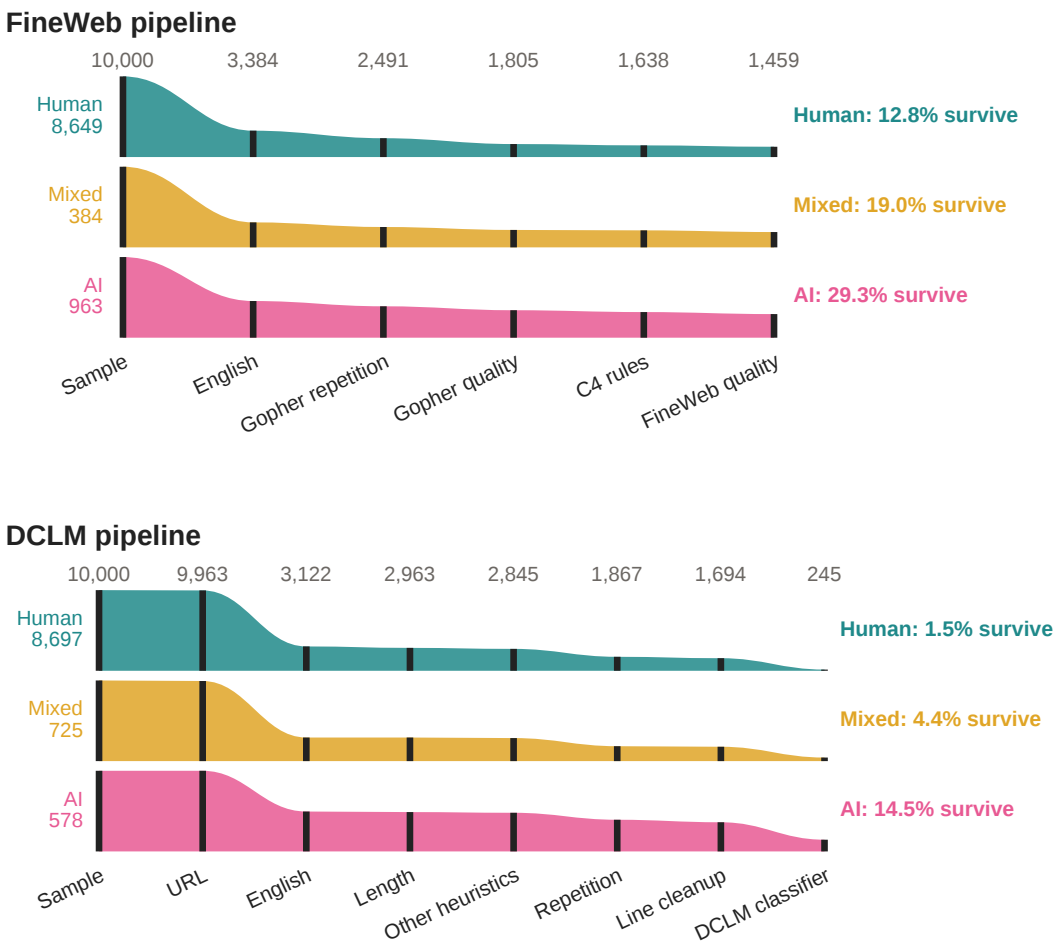


Figure 9: **DCLM's quality classifier keeps AI-labeled documents far more often than human ones.** Survival of a 10,000-document 2026 Common Crawl sample through each stage of the FineWeb (top) and DCLM (bottom) pipelines, split by Pangram 3.3.2 label.

Table 2: Forecast AI share of web tokens each December. The intervals are Student-t prediction intervals that include the uncertainty of the drift.

December	Forecast	80% interval	95% interval
2026	33.9%	30.0 to 37.8%	27.8 to 40.0%
2027	42.3%	33.5 to 51.1%	28.7 to 55.9%
2028	50.7%	38.0 to 63.4%	31.0 to 70.4%
2029	59.1%	42.7 to 75.5%	33.7 to 84.5%
2030	67.5%	47.5 to 87.4%	36.5 to 98.4%

June; the final forecasts refit the chosen random walk with drift through August 2026. We estimate 50.7% of text will be AI by the end of 2028, shown in Figure 10 and Table 2.

B Training Configurations

Architecture details. We report architecture details in Table 3. All 800 models share one nanochat training recipe (Karpathy, 2025). The input embedding and output head are untied. N counts the input embedding but not nanochat's value-embedding tables, which are also lookups (16.8M parameters at 19.9M, 709M at 973M). The tokenizer is nanochat's BPE with 32,768 tokens, and sequences hold 2,048 tokens. Muon (Jordan et al., 2024) updates the transformer matrices (learning

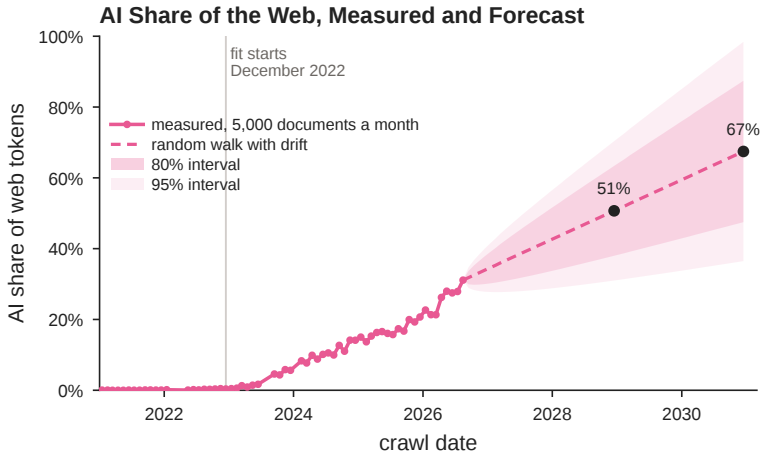


Figure 10: The AI share of web tokens, measured each crawl month and forecast to December 2030 by a random walk with drift fitted to December 2022 through August 2026, with 80 and 95% prediction intervals.

Table 3: The seven model sizes. All use the nanochat architecture. N in the scaling laws counts the input embedding; the last column subtracts it. Reserved sizes (†) enter no fit.

Model	Layers	Width	Heads	N (with input embedding)	Without
19.9M	4	256	2	19.9M	11.5M
35.8M	6	384	3	35.8M	23.2M
86.2M	9	640	5	86.2M	65.2M
135M	12	768	6	135.3M	110.1M
268M	16	1024	8	268.4M	234.9M
477M†	20	1280	10	477.1M	435.2M
973M†	26	1664	13	972.9M	918.4M

rate 0.02) and AdamW (Loshchilov & Hutter, 2019) the embeddings, output head and scalars. The learning rate warms up over 40 steps, stays constant, and decays linearly over the last 65% of steps to 5% of its peak. A step holds 262,144 tokens at 19.9M and 35.8M, 524,288 from 86.2M to 268M, and 1,048,576 at 477M and 973M. The schedule spans all of a run's tokens, so a run with added tokens trains for more steps than its control and decays later. Human and AI documents are interleaved in a fixed hashed order, and every AI or fresh-human addition trains on exactly its control's human documents, in the same order and with the same seed. 703 of the 800 runs use one seed; the other 97, all at 19.9M to 268M, are replicates with other seeds (47 cells trained with two, three or five seeds). Training uses bf16 with FP8 matrix multiplies on eight H100 GPUs. We score the final checkpoint in bits per byte on 15.7M tokens each of C4, FW22 and FW26, 12.4M of FW26-H, 3.5M of FW26-AI, 21.0M of Cosmopedia, and the validation split of Paloma's 16 sources. Table 4 lists all 800 runs, and Table 5 shows the 365 in the matched-budget grid. The 38 filtering models of Figure 5 (left) and the six repetition models of §F.3 are not among the 800 and enter no fit.

Parameter count and human budget. Following Pearce & Song (2024) (see also Weng, 2026), N counts the input embedding, as nanochat reports model sizes.[16] This choice also fits our laws better at the smallest sizes, where the two counts differ most. Figures label each human budget by its TPP_h rounded to the nearest 5, and Table 6 gives the exact values.

[16] https://github.com/karpathy/nanochat/discussions/420

Table 4: Every model in the paper. The 800 models of the scaling-law cohort by size and arm: human-only controls, AI additions and fresh-human additions, each addition trained on exactly its control's human documents. Sizes marked † are held out from every fit. TPP_h: range over the controls. *In Table 5*: runs in the matched-budget grid, controls included; the others belong to control groups at per-size budgets from earlier run generations and enter every fit and score in the same way. *Other seeds*: runs with a seed other than 1337. Below the rule, the 44 filtering and repetition models, which enter no fit.

			Additions				
Model	TPP_h	Controls	AI	Human	Total	In Table 5	Other seeds
19.9M	2.9–87.5	20	113	20	153	61	25
35.8M	3.2–87.5	16	96	16	128	60	3
86.2M	3.8–87.5	17	88	8	113	54	7
135M	4.1–87.5	18	110	45	173	73	30
268M	4.4–87.5	17	95	47	159	69	32
Fitted		88	502	136	726	317	97
477M†	4.6–73.0	8	42	12	62	36	0
973M†	4.7–37.8	3	9	0	12	12	0
Held out		11	51	12	74	48	0
All		**99**	**553**	**148**	**800**	**365**	**97**
Outside the cohort, never fitted							
Filtering (Figure 5, left): 22.3% AI mix, and AI removed					38		
Repetition (§F.3): human corpus repeated					6		

Table 5: The training grid. Each cell is one human-only control plus the number of AI-addition runs and, after the plus sign, fresh-human-addition runs on the same human corpus. Sizes marked †, below the rule, are reserved: every law is scored on them and none is fitted on them.

Model \ TPP_h	5	20	40	55	75	90	Other
19.9M	12+3	11+4	5	9+1	5	5	14: 66+12
35.8M	12+3	11+4	5	8+1	5	5	10: 50+8
86.2M	11+3	9+3	5	7	5	5	11: 46+2
135M	14+7	12+7	6+3	5+1	5+1	5+1	12: 63+25
268M	11+7	11+6	6+4	5+1	5+1	5+1	11: 52+27
477M†	7+3	7+3	5	3	3	–	3: 17+6
973M†	0	5	4	–	–	–	–

C Additional Scaling Law Background

C.1 Scaling Laws for Repetition

Lovelace et al. (2026) propose an alternative to effective data, arguing that there is an interaction between model size and repetition that the saturation term did not capture in their experiments. They instead keep the total token count $D = U_D(1 + R_D)$ in the data term and add an overfitting penalty. The penalty is fit by coefficient P, which grows with repetition and model capacity relative to the number of unique tokens, allowing for validation loss to increase if the cost of repetition outweighs the benefits. Their simplest form is:

$$L(N, U_D, R_D) = E + \frac{A}{N^{\alpha}} + \frac{B}{[U_D(1 + R_D)]^{\beta}} + \underbrace{P\, R_D\, \frac{N}{U_D}}_{\text{overfitting penalty}} \tag{5}$$

Muennighoff et al. (2023) extended Chinchilla to account for repeated data using an effective token count, D_{eff}, and an effective parameter count, N_{eff}, written as follows:

Table 6: TPP_h of the aligned budgets. 477M has no 90 budget; 973M has no 55, 75 or 90 budget.

Label in figures (TPP_h)	5	20	40	55	75	90
Actual D_{H}/N	4.7	18.9	37.8	54.7	73.0	87.5

$$L(N, U_D, R_D) = E + \frac{A}{N_{\text{eff}}^{\alpha}} + \frac{B}{D_{\text{eff}}^{\beta}} \tag{6}$$

Let U_D denote the number of unique training tokens and R_D the number of additional epochs beyond the first. Then $D = U_D(1 + R_D)$. $R_D^{\star}$ is a fitted constant controlling the saturation of repetition. Then effective data is defined as:

$$D_{\text{eff}} = U_D + \underbrace{U_D R_D^{\star} \left(1 - e^{-R_D/R_D^{\star}}\right)}_{\text{discounted credit for repeated tokens}}, \qquad R_D = \frac{D}{U_D} - 1 \tag{7}$$

In the benchmark (Table 1) the human tokens are the unique data, $U_D = D_{\mathrm{H}}$, and the added AI tokens play the part of repeated data, $R_D = D_{\mathrm{A}}/D_{\mathrm{H}}$.

Qin et al. (2026) model the dependence of data saturation on the amount of unique tokens and model size. They extend Muennighoff et al. (2023), proposing that the effective value of a token depends on N, fresh tokens per parameter $\mathrm{TPP} = U_D/N$, and R_D. For paraphrased data, $R_D = (D - U_D)/U_D$ measures added derived tokens relative to fresh tokens. Their Compute-Data (CD) law is:

$$L(N, U_D, R_D) = E + \frac{A}{N^{\alpha}} + \frac{B}{D_{\text{eff}}^{\beta}} \tag{8}$$

Unlike the constant $R_D^{\star}$ in Muennighoff et al. (2023), CD's saturation scale varies with N and TPP. The effective token count and saturation scale are detailed below:

$$D_{\text{eff}} = U_D \left[1 + R_D^{\star} \left(1 - e^{-R_D/R_D^{\star}}\right)\right], \qquad R_D^{\star} = K\, \mathrm{TPP}^{\rho} N^{\sigma} \tag{9}$$

C.2 Scaling Laws for Data Mixtures

Closest to our setting, Shukor et al. (2025) fit mixture laws for a second data source. To model the effects of AI-generated text, we adapt their law with human share $h = 1/(1 + r)$ and AI share $f = r/(1 + r)$ of the total token count $D = D_{\mathrm{H}} + D_{\mathrm{A}}$. Their additive form keeps the Chinchilla constants and adds a term in the shares:

$$L(N, D, f) = E + \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}} + \frac{1}{c_{\mathrm{H}} h^{q_{\mathrm{H}}} + c_{\mathrm{A}} f^{q_{\mathrm{A}}}}, \tag{10}$$

They also propose a joint form that lets the shares rescale the model and data amplitudes, as follows:

$$L(N, D, f) = E + \frac{(a_{\mathrm{H}} h + a_{\mathrm{A}} f)^{p_A}}{N^{\alpha}} + \frac{(b_{\mathrm{H}} h + b_{\mathrm{A}} f)^{p_B}}{D^{\beta}} + \frac{1}{c_{\mathrm{H}} h^{q_{\mathrm{H}}} + c_{\mathrm{A}} f^{q_{\mathrm{A}}}}, \tag{11}$$

We benchmark the joint form as the strongest published comparator (Table 1).

C.3 Other Benchmarked Scaling Laws

We compare published functional forms, specialized to two data sources where needed, with coefficients fitted on our experiments. We use $n = N/10^8$, $d = D_{\mathrm{H}}/10^9$, $r = D_{\mathrm{A}}/D_{\mathrm{H}}$, $f = r/(1 + r)$,

Table 7: Coefficients of Equation 2 fitted separately on each evaluation target, on the same 726 models (19.9M to 268M) as Table 1.

	Human text				Mixed	AI text	
Coefficient	C4	FW22	FW26-H	Paloma	FW26	FW26-AI	Cosmo
E	0.7115	0.6683	0.5992	0.5003	0.5541	0.1923	0.2890
A	0.2548	0.2759	0.2733	0.9020	0.2647	0.1723	0.2152
α	0.3027	0.2900	0.2912	0.1198	0.2996	0.4271	0.3714
B	0.08480	0.09414	0.09802	0.02689	0.1077	0.4186	0.2759
β	0.4722	0.4485	0.4354	0.9614	0.4104	0.1224	0.1816
η	1.729	1.629	1.680	2.886	3.393	25.22	22.51
K	0.001997	0.003274	0.004658	0.0003355	27.45	50.82	36.95
ρ	-3.293	-3.098	-2.916	-3.629	-0.3230	-0.5606	-0.5223
γ	0.4053	0.3398	0.3087	5.000	1.581	0.06199	0.1046
u	1.204	1.197	1.169	1.395	0.3862	0.6878	0.6462
v	0.4422	0.4209	0.4177	1.042	-0.1333	-0.5268	-0.5790

and $t = D_{\mathrm{H}}/(20N)$. The remaining published forms are written below in our coordinates with $M = E + A/n^{\alpha}$, $T = d(1 + r)$, $h = 1 - f$, and $q > 0$.

$$L_{\mathrm{ATLAS}} = M + B[d(1 + qr)]^{-\beta} \tag{12}$$

$$L_{\mathrm{He}} = (M + BT^{-\beta})h^{-\zeta} \tag{13}$$

$$L_{\mathrm{Hamidieh}} = M + BT^{-\beta} \exp[-a_{\mathrm{H}} h \log D_{\mathrm{H}} - a_{\mathrm{A}} f \log D_{\mathrm{A}}] \tag{14}$$

$$L_{\mathrm{Jain}} = M + B\left[\frac{1 + qr}{d(1 + r)^2}\right]^{\beta} + \gamma f^2 \tag{15}$$

$$L_{\mathrm{Sedova}} = M + Bn^{c}[d(1 + qr)]^{-\beta} + \gamma f \tag{16}$$

D Scaling-Law Fitting and Model Selection

Fitting our scaling laws. Every law, ours and each comparator, is fitted on the same 726 models at the five development sizes (19.9M to 268M). The 74 models at 477M and 973M are held out. The human-only controls contribute their own residual, and every addition run (AI or fresh human) contributes the residual of its change in log loss against its own control, so that the fit is anchored on the paired contrasts of §4. Coefficients minimize a Huber loss on these residuals with bounded least squares, from a Chinchilla fit on the human-only runs as the starting point. Table 7 lists the coefficients of our law fitted separately on each evaluation target, and Figure 11 draws its credit and harm terms on C4.

D.1 Choosing the Scaling Law

Choosing the data benefit. Table 8 fits our law on C4 for the held-out sizes with many different forms of the credit term. We choose an exponential term to make the fewest changes from Muennighoff et al. (2023) and Qin et al. (2026), just adding a free parameter η that can control the size of the benefit. This is tied for the lowest error on our held-out sizes (0.83 $\times 10^{-3}$, the same as a tanh window). Figure 12 shows how the fitted window saturates over different human token budgets and the largest amount of performance AI can help a model at different human token budgets.

Choosing the penalty. The power-law harm of Lovelace et al. (2026) would show the AI-generated harm accelerating, whereas we empirically show the increase in loss slowing down, as shown in Figure 13; fitted to our runs, a power-law penalty takes an exponent below one (0.61 on C4), so it too decelerates. We try many penalties and choose the logarithm minus the AI share, $\log(1 + r) - r/(1 + r)$; every form tried is in Table 8. Adding the same harm to the loss instead of the data term scores 0.80 on C4 against our 0.83 $\times 10^{-3}$, inside bootstrap noise (95% interval of the gap -0.07 to $+0.01$); the two forms are near-reparametrizations of each other and give the same cost of not filtering at 8B (1.58 and 1.59×). We keep the harm on the data term because it has

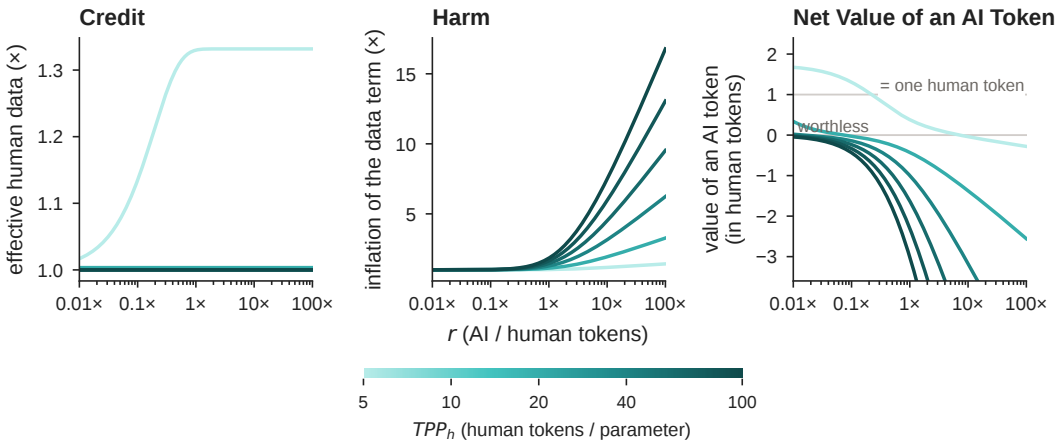


Figure 11: **The credit and harm terms of our law.** The two terms of Equation 2 and their net effect on C4 at 268M, for human budgets from 5 to 100 TPP_h.

Table 8: Our law on C4 with the harm or benefit changed one at a time: paired RMSE $\times 10^3$, lower is better, scored on the held-out sizes.

Form	k	477M and 973M
The credit AI tokens earn (harm as in our law)		
No credit: AI tokens add no data	8	3.29
Linear, ηr (no saturation)	9	1.87
Exponential window, $\eta = 1$ (CD's credit)	10	0.97
Rational window, free η	11	0.98
Tanh window, free η	11	**0.83**
Exponential window, no budget dependence ($\rho = 0$)	10	1.71
Exponential window, also size-dependent ($R^\star = Kt^\rho n^\sigma$)	12	1.41
Exponential window, free η, $R^\star = Kt^\rho$ (ours)	11	**0.83**
The harm AI tokens do (credit as in our law)		
No harm (credit only)	8	3.27
$\gamma\,[\log(1+r) - r/(1+r)]$ on the data term, no budget or size dependence	9	2.83
$\gamma t^u[\log(1+r) - r/(1+r)]$ on the data term, no size dependence	10	2.36
$\gamma n^v[\log(1+r) - r/(1+r)]$ on the data term, no budget dependence	10	2.16
$\gamma t^u n^v[\log(1+r) - r/(1+r)]$ added to the loss, not tied to the data term	11	**0.80**
$\gamma t^u n^v \log(1+r)$ on the data term (harm starts at the first AI token)	11	1.15
Box–Cox of $\log(1+r)$ with free curvature λ, on the data term	12	1.13
Power law $\gamma r^p t^u n^v$ added to the loss (Lovelace et al., 2026)	12	1.55
$\gamma t^u n^v[\log(1+r) - r/(1+r)]$ **on the data term (ours)**	11	0.83

the lower error at 973M on every target, the lower leave-one-size-out error, and the lower error on Paloma, FW26, FW26-AI and Cosmo (the last three significantly), and because it keeps the whole effect of AI text inside the data term, so that the law remains Chinchilla in an effective token count.

D.2 Law Fits Across Model Sizes and Budgets

Figure 14 compares the Chinchilla fit to ours. Figure 15 shows all TPP_h and size fits on C4. At low TPP_h added AI text first lowers loss and then raises it; from about 20 TPP_h it raises loss from the first tokens, and the harm grows roughly with $\log r$. Figure 16 compares the value of an AI token under our law with the value measured from paired AI and fresh-human runs.

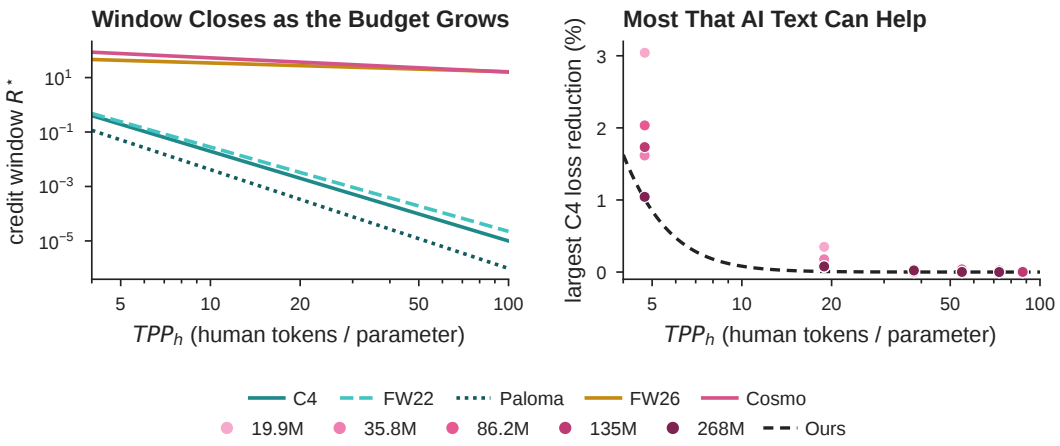


Figure 12: **The window in which AI text helps closes as the human budget grows.** Left: the fitted credit window $R^\star = K\, t^\rho$, the rate at which the benefit of AI text saturates, against the human budget for each evaluation set; it is orders of magnitude wider for the mixed and AI-text sets (FW26, Cosmo) than for human text. Right: the largest reduction in C4 loss that added AI text achieves in each fitted group and under our law: up to 3.0% at 5 TPP_h, under 0.4% at 20 TPP_h and none beyond.

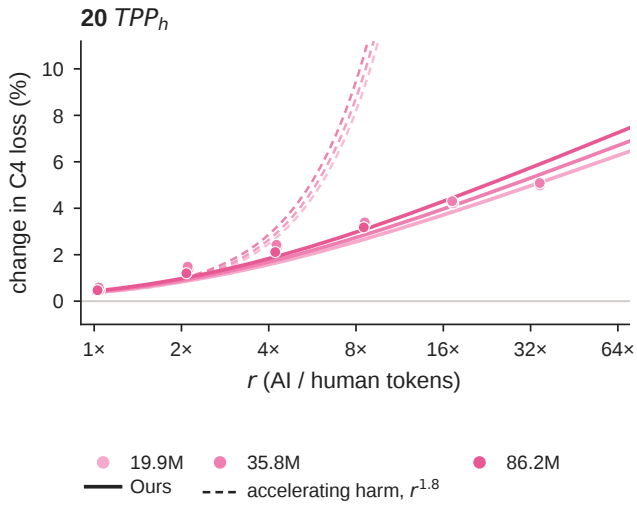


Figure 13: **The harm of AI text decelerates.** Change in C4 loss against r from 1 to 64: each doubling of AI text adds the same increase in loss. Dashed: an accelerating power-law penalty in the style of Lovelace et al. (2026) (exponent 1.8, for illustration)

.

E Benchmarking Other Scaling Laws

We benchmark 11 other scaling laws, applying laws meant for repetition, data mixing, and surrogate data to see how AI-generated text most closely acts. The requirements each law meets and their performance on held-out runs are depicted in Figure 17. Additional analyses benchmark laws on absolute loss instead of relative (Table 10), on AI-generated target text (Table 9), and on low ratios of AI-generated tokens (Table 11). On Paloma, five comparators (Chinchilla, Muennighoff et al., CD, ATLAS and Hamidieh et al.) fit a data term that is negligible at the held-out token counts (β at its upper bound of 2.5, or for CD a credit window at its lower bound, $R^\star < 0.003$). They predict no change from added AI text ($|\Delta| < 2 \times 10^{-5}$), so all five score the error of predicting zero change, 9.15 (7.75 at $r < 1$). Refitting Chinchilla with β allowed up to 10 gives the same error. On absolute

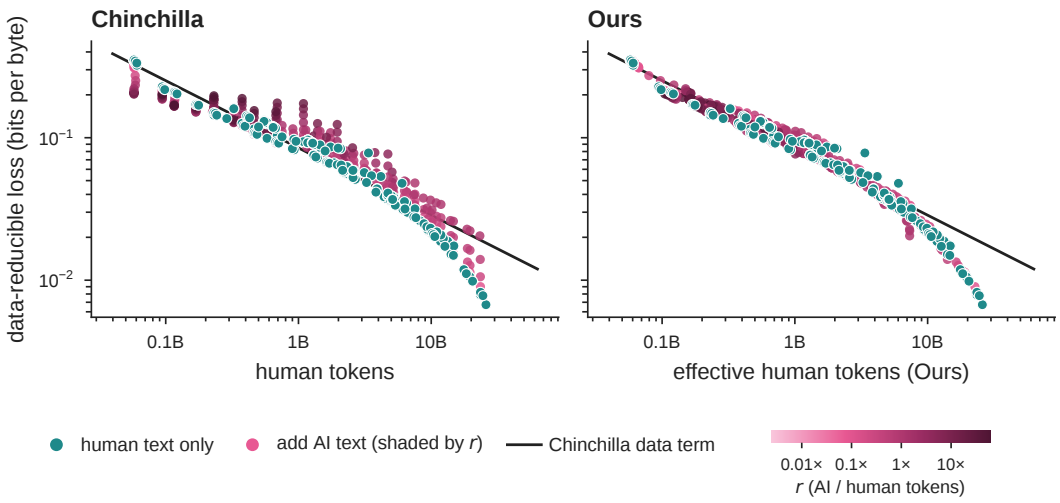


Figure 14: Effective human tokens place the AI runs on the Chinchilla data term. Counting human tokens only, AI runs scatter off the Chinchilla data term traced by the human-only runs; counted as effective tokens, they collapse onto it.

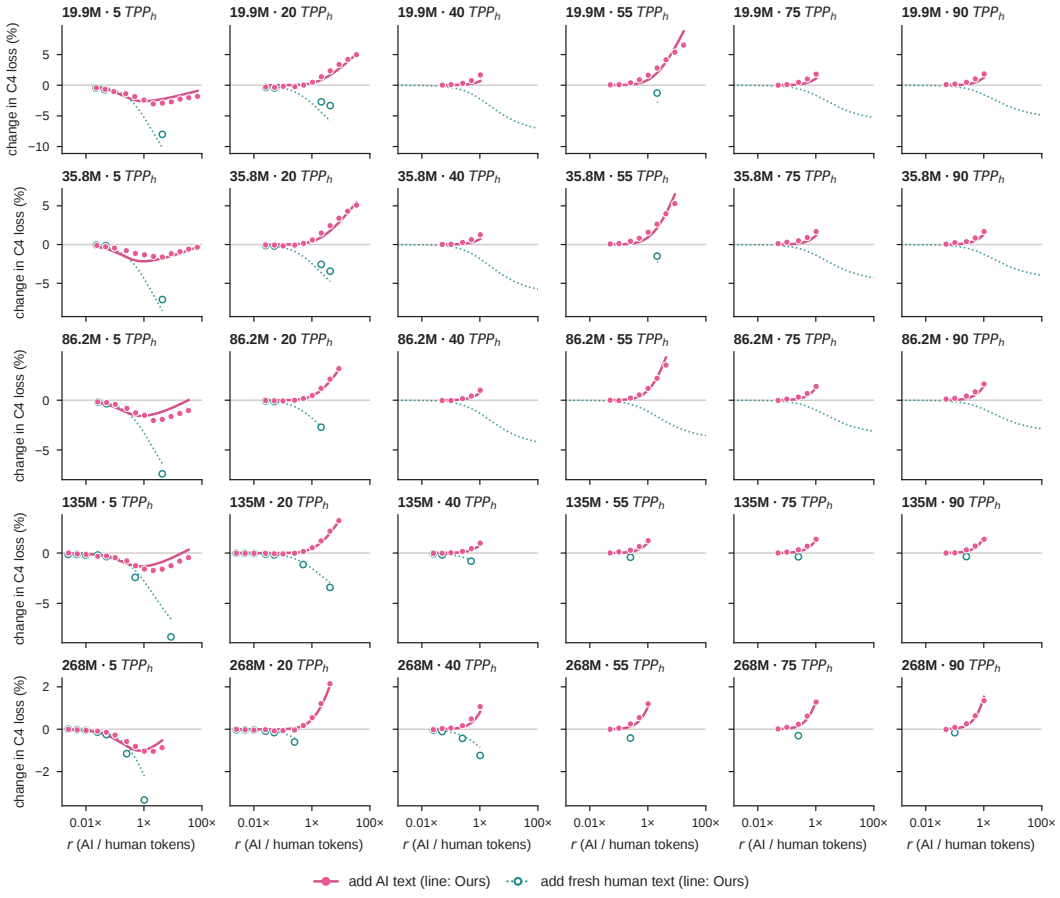


Figure 15: Our law across every fitted size and budget. Change in C4 loss against r for every fitted control group with our law's curves for added AI text and added fresh human text.

error the ranking changes: CD and ATLAS predict the loss of the human-text sets best (12.51 and

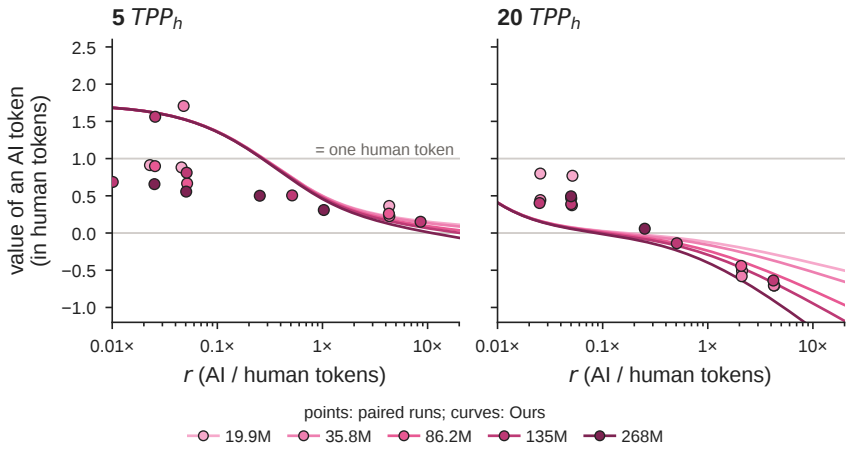


Figure 16: **Our law's value of an AI token against measured values.** Value of an added AI token in human tokens under our law and measured from paired AI and fresh-human additions to the same control, at 5 and 20 TPP_h.

Table 9: Table 1 for the mixed and AI-text targets: paired RMSE $\times 10^3$ on the held-out sizes, lower is better. On FW26-AI and Cosmo, the exponent of the human-share multiplier of He et al. (2025) sits at its bound of zero, where the law is exactly Chinchilla: the multiplier can only raise the loss as AI text is added, and added AI text lowers the loss on these targets.

		477M and 973M		
Law	k	FW26	FW26-AI	Cosmo
General				
Chinchilla (Hoffmann et al., 2022)	5	20.36	101.89	78.71
Repetition and paraphrase				
Muennighoff et al. (2023)	7	18.27	97.95	75.26
CD (Qin et al., 2026)	8	16.66	96.27	72.79
Lovelace et al. (2026)	9	17.77	97.88	73.03
Data mixtures				
ATLAS (Longpre et al., 2026)	6	17.83	26.63	24.30
He et al. (2025)	6	22.08	101.89	78.71
Hamidieh et al. (2025)	7	12.83	32.04	27.61
Shukor additive (Shukor et al., 2025)	9	6.65	9.11	8.85
Shukor joint (Shukor et al., 2025)	13	**2.13**	**6.20**	**7.97**
Sedova et al. (2026)	8	13.12	16.39	18.80
AI data				
Jain et al. (2024)	7	20.21	92.84	73.41
Ours	11	11.04	21.34	20.92

13.65 on C4, against 14.09 for ours, all $\times 10^{-3}$), while ours is best on Paloma and on the AI-text sets (Table 10). Because the paired errors are much smaller, absolute error mostly measures how well a law predicts the loss of the human-only models at the held-out sizes. The paired error removes that level and isolates the effect of AI text, which is what the laws are meant to capture, so it is our main metric.

E.1 Statistical Significance

We resample the 51 held-out AI runs by their 10 control groups (10,000 draws), hold every fit fixed, and recompute each law's paired error (Table 12). Over all AI ratios our law has lower error than the joint law of Shukor et al. (2025) in 99.99% of draws on C4 (gap 0.58 $\times 10^{-3}$, 95% interval 0.32 to 0.84) and in every draw on FW22 and FW26-H, and lower error than each other comparator in

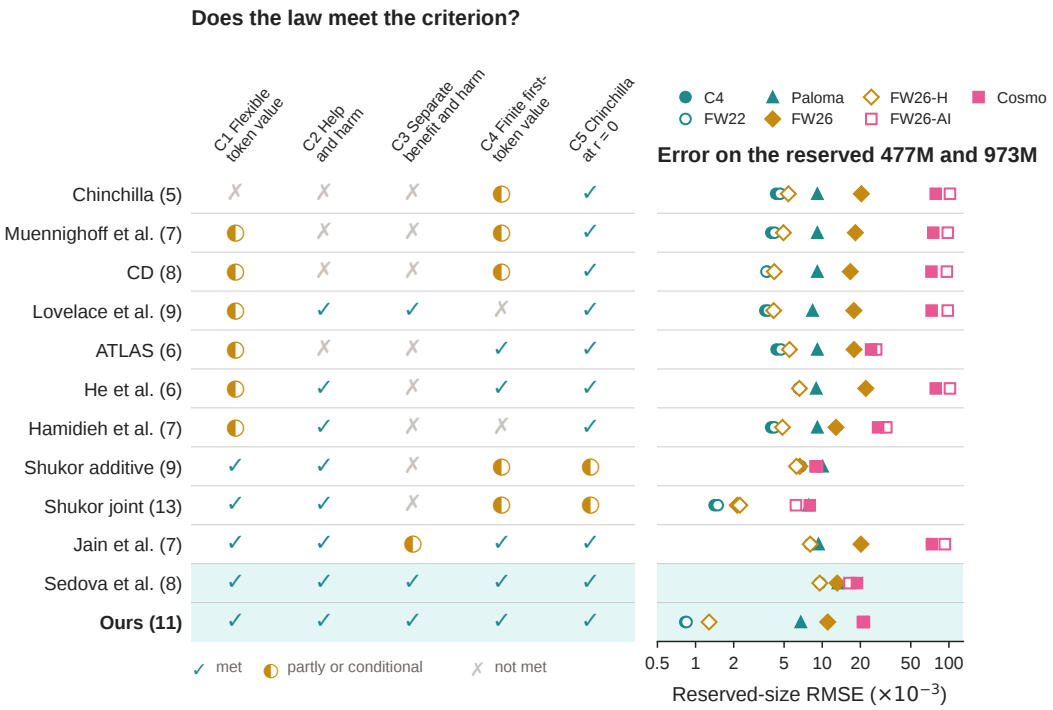


Figure 17: **Only our law meets all five criteria at low error.** Left: whether each benchmarked law meets the criteria of §4.2: always (check), partly or only for some coefficient values (half circle), or for no coefficient values (cross). For criterion C4 (a finite first-token value), a half circle marks a value that is finite but fixed, as in Chinchilla, where every AI token is worth exactly one human token, or finite only when a fitted exponent reaches one. Right: each law's paired RMSE on the held-out runs for every evaluation set, the numbers of Table 1 and Table 9. The law of Sedova et al. (2026) also meets every criterion, at 11.6 times our error on C4.

every draw. On Paloma our law beats the joint law in 98.7% of draws but is the best of all twelve laws in only 89.4%.

E.2 Comparison with the Joint Law of Shukor et al. (2025)

The joint law of Shukor et al. (2025) is the strongest published scaling law for modeling the effects of AI text. It is second to ours on human text and best on mixed and AI text (Table 1, Table 9). It has no separate benefit and harm terms, and its fits imply that a model trained on AI text reaches a lower loss floor on human text than one trained on human text (Figure 18, right). Its projections beyond the trained sizes are also unstable: fitted through 135M or through 268M, it projects opposite effects of AI text at 8B (Figure 18, middle). We trained no 8B model, so these are projections for either law.

F Details for the Recommendations

F.1 Compute-Equivalent Gain (CEG)

CEG (Davidson et al., 2023) provides a direct method to compare the compute efficiency of two models. At a fixed model size, compute is proportional to training tokens ($C \approx 6ND$), so a gain in tokens is a gain in compute. If a model trained on D tokens reaches loss L, then

$$\mathrm{CEG}_{\mathcal{R}}(D) = \frac{D_{\mathrm{ref}}(L)}{D}, \tag{17}$$

Table 10: Absolute BPB RMSE $\times 10^3$ on the held-out sizes, lower is better: the error in the predicted loss level itself rather than in the paired change against the human-only control, so the ranking can differ.

Human-text targets					
		477M and 973M			
Law	k	C4	FW22	FW26-H	Paloma
General					
Chinchilla (Hoffmann et al., 2022)	5	26.93	28.41	29.69	30.91
Repetition and paraphrase					
Muennighoff et al. (2023)	7	21.70	22.24	22.77	31.66
CD (Qin et al., 2026)	8	**12.51**	**12.86**	**13.89**	29.76
Lovelace et al. (2026)	9	18.35	18.57	19.53	30.19
Data mixtures					
ATLAS (Longpre et al., 2026)	6	13.65	14.08	15.15	30.88
He et al. (2025)	6	16.00	15.98	17.08	29.77
Hamidieh et al. (2025)	7	15.25	16.24	17.65	31.02
Shukor additive (Shukor et al., 2025)	9	15.79	15.62	16.37	30.82
Shukor joint (Shukor et al., 2025)	13	15.36	15.61	15.96	29.17
Sedova et al. (2026)	8	17.32	16.65	17.59	34.05
AI data					
Jain et al. (2024)	7	18.21	17.97	18.60	30.87
Ours	11	14.09	14.31	15.32	**28.74**

Mixed and AI-text targets				
		477M and 973M		
Law	k	FW26	FW26-AI	Cosmo
General				
Chinchilla (Hoffmann et al., 2022)	5	31.37	60.32	43.24
Repetition and paraphrase				
Muennighoff et al. (2023)	7	22.52	62.46	47.38
CD (Qin et al., 2026)	8	18.35	66.53	46.36
Lovelace et al. (2026)	9	21.47	64.53	47.68
Data mixtures				
ATLAS (Longpre et al., 2026)	6	32.40	10.10	10.87
He et al. (2025)	6	23.58	60.32	43.24
Hamidieh et al. (2025)	7	21.84	12.95	12.15
Shukor additive (Shukor et al., 2025)	9	16.69	18.53	15.79
Shukor joint (Shukor et al., 2025)	13	17.62	16.78	14.80
Sedova et al. (2026)	8	**15.24**	8.33	6.65
AI data				
Jain et al. (2024)	7	23.29	39.56	31.37
Ours	11	17.41	**6.91**	**6.55**

where $D_{\text{ref}}(L)$ is the number of tokens the reference recipe needs to reach the same loss. A gain above one means the recipe is worth more than its compute in reference tokens; below one, less. The law-based version (Figure 19) replaces both sides with the fitted law and varies the human budget.

F.2 Optimal AI Share

The optimal AI shares in Figure 6 hold the model size N and the human corpus D_{H} fixed and ask how many AI tokens to add. For each evaluation target we fit the law of Equation 2 separately and minimize its predicted loss over the ratio of AI to human tokens,

$$r_{\text{opt}}(N, D_{\text{H}}) = \operatorname*{arg\,min}_{0 \le r \le r_{\max}} \widehat{L}\big(N, D_{\text{H}}, r\, D_{\text{H}}\big), \qquad f_{\text{AI}} = \frac{r_{\text{opt}}}{1 + r_{\text{opt}}}, \tag{18}$$

Table 11: The paired errors of Table 1 restricted to the held-out runs with $r < 1$, the range a web crawl can reach our 2026 crawl is $r = 0.29$; an even split of AI and human text is $r = 1$).

		477M and 973M			
Law	k	C4	FW22	FW26-H	Paloma
General					
Chinchilla (Hoffmann et al., 2022)	5	2.93	3.18	3.82	7.75
Repetition and paraphrase					
Muennighoff et al. (2023)	7	2.47	2.63	3.22	7.75
CD (Qin et al., 2026)	8	2.22	2.12	2.60	7.75
Lovelace et al. (2026)	9	3.50	3.69	4.06	7.88
Data mixtures					
ATLAS (Longpre et al., 2026)	6	2.96	3.24	3.91	7.75
He et al. (2025)	6	4.78	4.83	5.02	7.70
Hamidieh et al. (2025)	7	2.75	2.86	3.39	7.75
Shukor additive (Shukor et al., 2025)	9	4.34	4.29	3.90	8.16
Shukor joint (Shukor et al., 2025)	13	0.98	1.01	1.80	7.06
Sedova et al. (2026)	8	7.54	7.58	7.58	10.66
AI data					
Jain et al. (2024)	7	3.51	3.43	3.26	7.27
Ours	11	**0.64**	**0.71**	**1.14**	**6.54**

Table 12: Our law's lead holds over all AI ratios on human text, but not at $r < 1$ on C4 and FW22. Paired RMSE $\times 10^3$ on the held-out runs for our law and the joint law of Shukor et al. (2025), with 95% intervals for ours and for the gap from 10,000 bootstrap draws that resample the 51 held-out AI runs (43 with $r < 1$) by their 10 control groups, every fit held fixed. The last two columns give the share of draws in which our law has lower error than the joint law, and than all eleven comparators.

					Draws where ours is (%)	
Target	AI runs	Ours	Joint law	Joint − ours	below joint	best of 12
C4	all	0.83 [0.64, 0.99]	1.41	+0.58 [+0.32, +0.84]	99.99	99.99
	$r < 1$	0.64 [0.53, 0.77]	0.98	+0.34 [-0.03, +0.65]	96.5	96.5
FW22	all	0.85 [0.74, 0.96]	1.50	+0.65 [+0.32, +0.94]	100	100
	$r < 1$	0.71 [0.56, 0.87]	1.01	+0.30 [-0.10, +0.65]	92.6	92.6
FW26-H	all	1.28 [0.96, 1.63]	2.24	+0.96 [+0.56, +1.32]	100	100
	$r < 1$	1.14 [0.75, 1.56]	1.80	+0.67 [+0.22, +1.03]	99.8	99.8
Paloma	all	6.77 [3.31, 9.57]	7.80	+1.03 [+0.14, +1.65]	98.7	89.4
	$r < 1$	6.54 [3.24, 9.27]	7.06	+0.52 [-0.07, +0.90]	96.1	71.7

where f_{AI} is the AI share of all training tokens that the figure plots. Only the data term of Equation 2 depends on r, and setting its derivative to zero balances the marginal credit of an AI token against its marginal harm.

F.3 Repeating Human Tokens vs. Adding AI Tokens

Six models (19.9M and 35.8M, 20 TPP_h) repeat the human corpus to match r = 1, 4 or 8 and are compared with the six AI additions at the same r; the repetition models are not among the 800 of the scaling-law fits. On the three human-text sets repetition beats adding AI text in all 18 comparisons, lowering loss by 1.7 to 3.8% while AI text raises it by 0.2 to 3.4%, and four extra passes recover most of what the same amount of fresh human text gives (see Figure 20). On the two AI-text sets the order reverses: AI text lowers loss by 12 to 19%, repetition by 1.6 to 4.9%. On FW26, which mixes the two, AI text wins at $r = 1$ and repetition at $r = 4$ and 8.

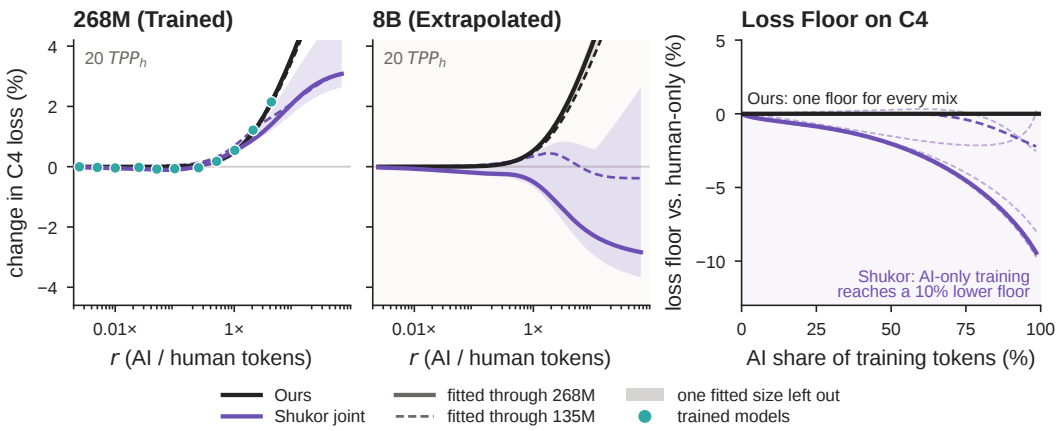


Figure 18: **The joint law of Shukor et al. (2025) extrapolates inconsistently.** Our law and the joint law on C4 at 20 TPP_h, fitted through 268M or 135M. Left: 268M, a trained size (points: trained models). Middle: extrapolated to 8B, where the joint law's predicted effect of AI text changes sign with the sizes it is fitted on, while every fit of ours predicts harm. Right: the loss floor each fit of the joint law implies for a training mix, relative to human-only training; our law has one floor for every mix.

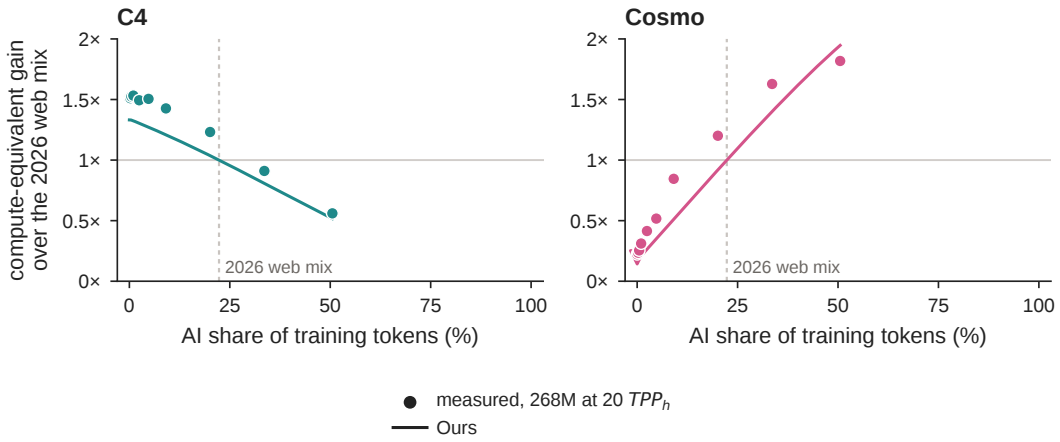


Figure 19: **Less AI text is always more compute-efficient for C4, and more AI text for Cosmopedia.** Compute-equivalent gain over our 2026 crawl mix (22.3% AI) of 268M models at 20 TPP_h as the AI share of training tokens varies, on C4 and Cosmopedia.

F.4 Loss on AI-Generated and Mixed Text

Figure 21 shows how our models learn on AI text. Figure 22 shows that models have lower loss on AI-labeled text than on human-labeled text, more so when trained with AI text, and that a mixed validation set can hide runs that look like improvements but raise loss on human text.

G Downstream Metrics

We find that CORE (Li et al., 2024) performance increases similarly when models are trained on additional AI or human tokens, as depicted in Figure 23. We hypothesize that while AI text may be good at these standard tasks, there will be greater differences in more unique tasks that may require diversity or creativity.

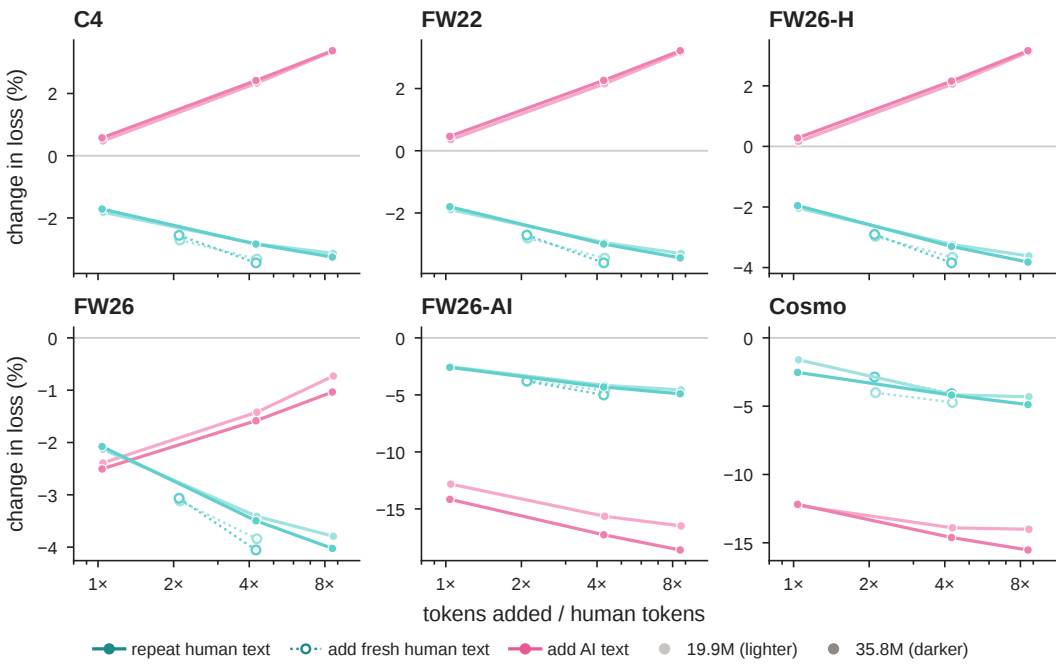


Figure 20: **Repeating human text beats adding AI text on human-text evaluation sets.** Change in loss on each evaluation set when the human corpus is repeated or the same number of new AI tokens is added, against tokens added per human token. On the human-text sets (C4, FW22, FW26-H) repetition lowers loss where AI text raises it; on the AI-text sets (FW26-AI, Cosmo) AI text lowers loss more.

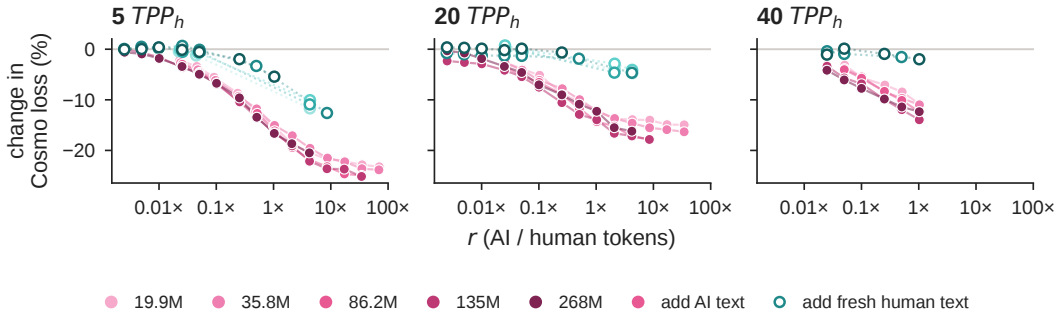


Figure 21: **On AI-generated text, added AI tokens lower loss at every budget.** Counterpart of Figure 3 for Cosmopedia: change in Cosmo loss when AI tokens or the same number of fresh human tokens are added, at 5, 20 and 40 TPP_h for every fitted size. AI text lowers the loss at every budget and size, by up to 25%; fresh human text lowers it far less.

Same benchmark scores, different fit to human text. Benchmark accuracy and fit to human text come apart. At 20 TPP_h, adding at least as many tokens as the human corpus holds raises CORE by 1.6 points with AI text and by 1.9 points with fresh human text (23 and 6 runs), yet C4 loss moves in opposite directions: up 2.3% with AI text and down 3.0% with fresh human text. Models trained with AI text answer benchmark questions about as well, but they model human text less closely and write more like AI (Table 13). CORE's tasks reward knowledge that AI web text also carries, while loss on human text also measures how well a model matches the formats and style of human writing, where AI text differs most (Figure 2, left).

AI-typical phrases increase. We generate continuations with the twelve held-out 973M additive models for 500 WritingPrompts (Fan et al., 2018) story prompts and for WebText (Radford et al.,

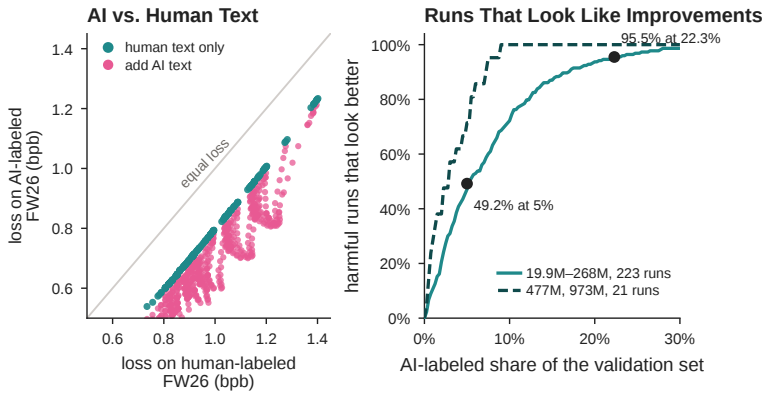


Figure 22: **Mixed validation data hides harm to human text.** Left: every model's loss on the AI-labeled against the human-labeled partition of FW26. Human-only models have 12 to 27% lower loss on AI text, and models trained with AI text fall further below the equal-loss line. Right: of the 244 AI additions that raise loss on human-labeled text, the share a validation set reports as improvements as the AI-labeled share of its text grows, for the fitted sizes and the held-out sizes: 49.2% at 5% AI and 95.5% at our 2026 crawl's 22.3%.

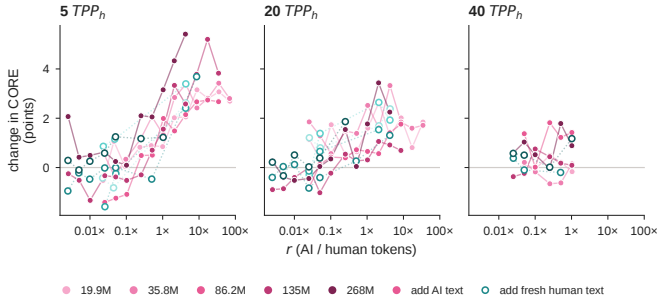


Figure 23: **Added AI tokens raise CORE about as much as fresh human tokens.** Change in CORE against each model's human-only control, as AI tokens or the same number of fresh human tokens are added. MMLU (Hendrycks et al., 2021) is not drawn because it stays at chance accuracy for every model, as expected at these sizes.

2019) article openings, and count 26 phrases typical of AI-generated text (such as "delve", "tapestry" and "it is important to note"). In story continuations the rate roughly doubles by $r = 0.5$, from 0.14–0.17 to 0.34–0.36 per 1,000 words at both budgets (Figure 24). The 268M ladders extend r to 4 (Table 13). At 20 TPP_h the phrase rate keeps rising, from 0.16 per 1,000 words without AI text to 0.36 at $r = 1$ and 0.54 at $r = 4$, and at 5 TPP_h it reaches 0.72 at $r = 4$. Pangram 3.3.2 labels 18.6% of the human-only model's stories AI, since base models often drift into generic web prose, and 28.4% at $r = 1$ and 37.2% at $r = 2$, twice the control; at $r = 4$ the share falls back to 28.9%.

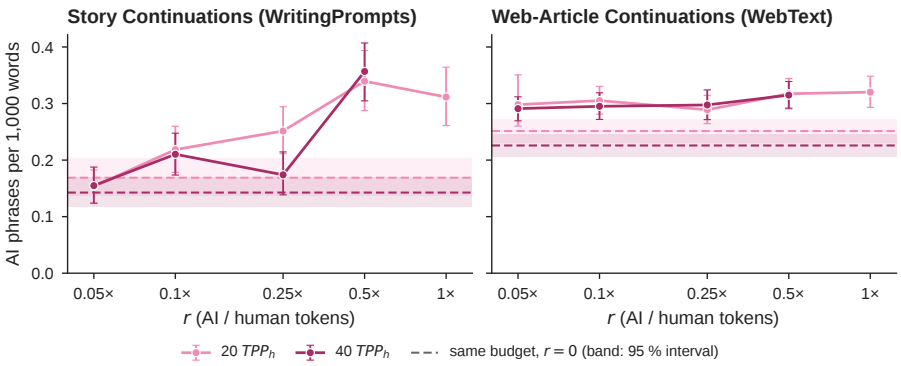


Figure 24: **Models trained with AI text use more AI-typical phrases.** Rate of 26 AI-typical phrases per 1,000 words in generations from models trained with AI text added at ratio r.

Table 13: **The 268M models trained with more AI text write more like AI.** Story continuations of 500 WritingPrompts prompts from the 268M models at 5 and 20 TPP_h, one model per r. AI-typical phrases per 1,000 words over eight continuations per prompt, and the share of one continuation per prompt that Pangram 3.3.2 labels AI, with 95% intervals from a bootstrap over prompts.

		AI phrases per 1,000 words		Stories labeled AI (%)
r	AI share (%)	5 TPP_h	20 TPP_h	20 TPP_h
0	0.0	0.16	0.16	18.6 [15.4, 22.0]
0.025	2.4	0.24	0.18	22.4 [18.6, 26.1]
0.05	4.8	0.18	0.21	21.1 [17.7, 24.7]
0.25	20.0	0.26	0.24	26.8 [23.0, 30.6]
0.5	33.3	0.23	0.24	22.7 [19.1, 26.7]
1	50.0	0.28	0.36	28.4 [24.4, 32.4]
2	66.7	0.41	0.46	37.2 [33.0, 41.6]
4	80.0	0.72	0.54	28.9 [25.1, 33.1]