SCALE DEPENDENT DATA DUPLICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Data duplication during pretraining can degrade generalization and lead to memorization, motivating aggressive deduplication pipelines. However, at web scale, semantically equivalent documents (e.g. translations) may induce redundant training signals once models become sufficiently capable. We present evidence that duplication is scale-dependent in two ways. First, as model capability increases, cross-entropy loss gradients for semantically equivalent documents become more aligned. Smaller models produce gradients that reflect surface similarity rather than semantic similarity. Second, we embedded all 192 million FineWeb-Edu-Dedup documents using EmbeddingGemma-300m. For moderate corpus sizes, nearest-neighbor cosine gap follows an isotropic power law baseline, but deviates sharply as corpus size grows to hundreds of billions of tokens, indicating accelerated semantic collisions. Finally, controlled pretraining on data sampled with replacement from finite pools of unique documents shows mild degradation for small models but rapidly increasing loss penalties for larger models, breaking naive scaling extrapolation. We derive scaling laws that account for limited semantic uniqueness, allowing for more accurate prediction at scale.

1 INTRODUCTION

Modern language models scale by increasing parameters, compute, and training tokens. Llama 1 (Touvron et al., 2023) trained on $\sim$1T tokens, while the Llama4 herd (Hugging Face, 2025) trained on up to 40T tokens. At these scales, even small fractions of duplicated data can reduce distinct training examples and harm downstream performance, emphasizing the importance of deduplication (Carlini et al., 2021; Hernandez et al., 2022; Lee et al., 2022; Comanici et al., 2025).

Deduplication is often framed as a dataset property: to deduplicate, simply remove exact duplicates and near-duplicates using simhashing techniques Broder (1997); Manku et al. (2007); Khan et al. (2026); Lee et al. (2022). Yet, what practically counts as a "duplicate" depends on the model as well: two documents that appear distinct may, to a sufficiently capable model, provide redundant training signal, and thus degrade training just as exact duplicates would. This work identifies a source of scale dependence: as models become more capable, *semantic duplicates* induce more aligned gradients during training. In tandem, capable models are trained on larger corpora, in which the number of semantic collisions rapidly increases. These effects can degrade models.

Contributions:

1. We quantify the **emergence of semantic sensitivity** during training using per-document cross-entropy gradient cosines across models and semantic-preserving transformations. In more capable models, semantic duplicates induce similar gradients during training.
2. We study **semantic collisions** by embedding 192M FineWeb-Edu-Dedup (Penedo et al., 2024) documents and analyzing nearest-neighbor (NN) statistics across dataset scales from 10^4 to 10^8 documents. Power laws governing moderate corpus sizes break down for large corpora. Synthetic corpora reach this collapse earlier, revealing lower semantic diversity.
3. We train scaling ladders on streams sampled with replacement from finite pools of K unique documents, showing that limited uniqueness breaks naive scaling extrapolation. We derive scaling laws that account for limited uniqueness, restoring predictability, and estimate effective K directly from mean nearest-neighbor cosine similarity.

Related work on scaling laws, deduplication, and semantics is discussed in Appendix A.

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.

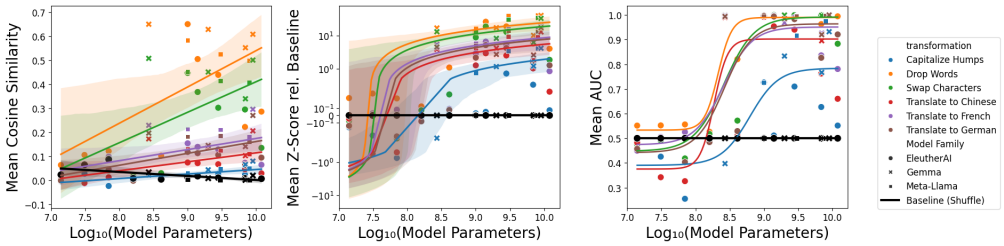


Figure 1: **Semantic-preserving transformations yield more aligned gradients for larger/stronger models.** We sample N=1000 FineWeb-Edu-Dedup documents and compute per-document gradients of normalized next-token cross-entropy (Eq. 1). We report mean cosine similarity between (i) unrelated document pairs (negative baseline) and (ii) each document and its transformed counterpart (positives), including translations. Smaller/weaker models exhibit gradient similarity dominated by surface cues (language/casing), often failing to separate positives from negatives. As capability increases, positives become more aligned than the negative baseline. Error bars show per-document standard deviation. Per-model-family results are in Figure 8.

2 Emergence of Semantics

As model capabilities increase, semantically equivalent documents induce similar training signals, as measured by the gradient of the per-document cross-entropy loss. Practically, this means that semantic duplicates operate increasingly like exact duplicates during model training.

2.1 Experimental Setup

We sample $N = 1000$ texts $\{x_i\}_{i=1}^{N}$ from FineWeb-Edu-Dedup (Penedo et al., 2024). To reduce length variance, each text is truncated to at most $T = 2000$ tokens using the model's tokenizer.

We compute the per-document full-parameter gradient of mean next-token cross-entropy:

$$g(x;\theta) = \nabla_\theta \ell(x;\theta), \qquad \ell(x;\theta) = \frac{1}{|x|}\sum_{u=1}^{|x|} \mathrm{CE}(f_\theta(x)_u, x_{u+1})\,. \tag{1}$$

For a null baseline, we sample unrelated English documents (x_i, x_j), $i \neq j$, and compute

$$\mathrm{sim}(x_i, x_j) = \frac{\langle g(x_i;\theta), g(x_j;\theta)\rangle}{\|g(x_i;\theta)\|_2 \, \|g(x_j;\theta)\|_2}. \tag{2}$$

Repeated random pairings estimate the baseline mean μ^- and standard deviation σ^-.

Transformations. We apply transformations intended to preserve semantics: randomly replacing ASCII characters or dropping words (each with probability 0.05), capitalizing every other character, and translating into Chinese, French, or German using Google's Translate API (Google).

For each document x_i and transformation τ, we compute

$$s_i^+(\tau) \coloneqq \mathrm{sim}_\theta(x_i,\ \tau(x_i))\,. \tag{3}$$

Separability Metrics (Z-scores and AUC). To summarize separation between positives and negatives, we define the Z-score using the mean positive similarity and the negative baseline:

$$z(\tau) \coloneqq \frac{\mu^+(\tau) - \mu^-}{\sigma^-}, \tag{4}$$

$$\mu^+(\tau) \coloneqq \frac{1}{N}\sum_{i=1}^{N} s_i^+(\tau), \tag{5}$$

$$\mu^- \coloneqq \frac{1}{|\mathcal{S}^-|}\sum_{(i,j)\in\mathcal{S}^-} \mathrm{sim}_\theta(x_i,\ x_j)\,, \tag{6}$$

$$(\sigma^-)^2 \coloneqq \mathrm{Var}_{(i,j)\in\mathcal{S}^-}\left[\mathrm{sim}_\theta(x_i,\ x_j)\right]. \tag{7}$$

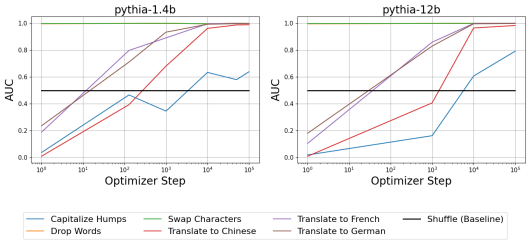


Figure 2: **Semantic sensitivity emerges over training and is accelerated by scale.** For a fixed model family, AUC distinguishes semantic-preserving transformations of the same document from unrelated documents, using cosine similarity to the original document gradient as the score. Early in training, AUC remains near 0.5 because gradients are dominated by surface-form features (language/casing). With additional optimizer steps, AUC increases, indicating that gradients increasingly reflect semantic content. Larger models reach a given AUC with fewer steps.

We also report AUC for distinguishing transformed gradients $\{g_\theta(\tau(x_i))\}$ (positives) from unrelated gradients $\{g(x_j;\theta)\}$ (negatives) using the gradient cosine similarity, $\mathrm{sim}(x_i,\cdot)$, as the score.

2.2 Results

Figure 1 reports mean gradient cosine similarities for both unrelated document pairs (negative baseline) and semantic-preserving transformations (positives). For smaller/weaker models, similarities for several transformations are comparable to or below the negative baseline, indicating that gradient direction is dominated by surface features (e.g., language or capitalization). As model capability increases, transformed counterparts become consistently more aligned than unrelated pairs.

We compute $z(\tau)$ in Eq. (4) and AUC for this binary task. Figure 2 further shows that AUC increases with training progress for a fixed family and is achieved earlier by larger models.

Interpretation. Our findings suggest that semantic and exact duplicates similarly affect capable models: if a model encodes meaning robustly, semantic duplicates generate aligned weight updates. This suggests how the *same dataset* can have a smaller effective size for more capable models.

3 Semantic Collisions

Under compute-optimal training, corpus size grows with model size and capability. We quantify the number of semantic collisions that occur in a deduplicated corpus of a given magnitude. We find that the rate of near-duplicates follows a predictable scaling law before rising more quickly.

Collision metrics. For unit-normalized embeddings $\{v_i\}_{i=1}^N$, define NN similarity

$$M_i := \max_{j\neq i}\langle v_i, v_j\rangle \quad \text{and cosine gap} \quad \Delta_i := 1 - M_i.$$

We report $\mathbb{E}[M_i]$ as a function of N and tail probabilities $\mathbb{P}(M_i \geq T)$ at fixed thresholds T.

3.1 Experimental Setup

We embed 192M texts from FineWeb-Edu-Dedup (Penedo et al., 2024) using EmbeddingGemma-300m (Vera et al., 2025). EmbeddingGemma-300m uses Matryoshka Representation Learning (Kusupati et al., 2024) to produce four nested embedding sizes (768, 512, 256, and 128); sub-embeddings are obtained by slicing and re-normalizing. Table 1 reports correlations between gradient and embedding cosine similarity for two embedders. We sample embedding subsets of size 10^4 to 10^8 and estimate within-pool NN cosine similarities using FAISS (Douze et al., 2025).

3.2 Results

Figures 3 and 4 show that nearest-neighbor collision statistics initially match a power law, but deviate sharply at larger dataset sizes. Collisions occur more quickly in smaller embedding spaces, as

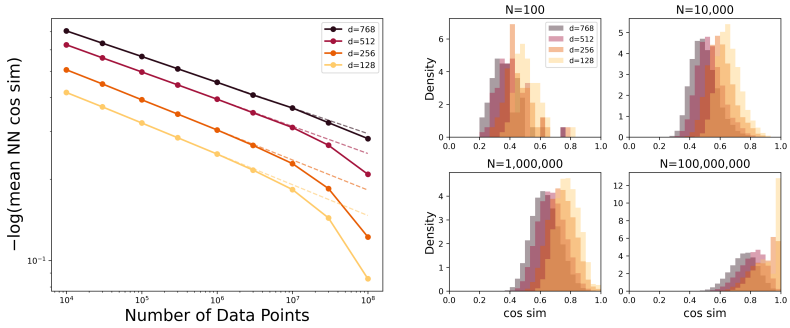


Figure 3: **NN cosine similarity scaling deviates sharply at large corpus sizes.** We embed 192M FineWeb-Edu-Dedup documents with EmbeddingGemma-300m and sample subsets of size ranging from 10^4-10^8 without replacement. For each N, we estimate the mean nearest-neighbor cosine similarity using FAISS. Dashed lines show best-fit power laws over the small-N regime where the uniform/vMF null predicts $\mathbb{E}[\Delta_i] \propto N^{-2/d}$. Beyond a scale threshold, the empirical curve steepens (smaller gaps than predicted), indicating more near neighbors than isotropic baselines predict.

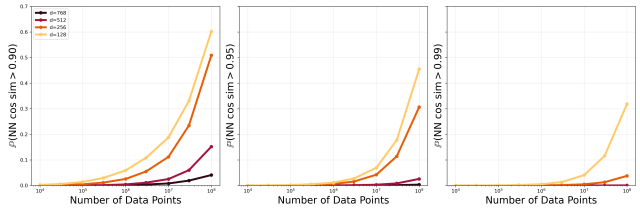


Figure 4: **Tail collision rates accelerate with dataset size.** For fixed thresholds T, we estimate the fraction of points with nearest-neighbor similarity $M_i \geq T$. These increase with N.

expected. Beyond a scale threshold, the mean cosine gap decreases *faster than any fitted power law calibrated on smaller* N. In log-log coordinates, the decrease in NN cosine gap is approximately linear over document corpus sizes less than 1M, before decaying faster for corpora with over 10M documents. This presents a compound threat to language models trained at scale: larger models that are more capable of identifying semantic duplicates are trained on more data, which contains more semantic duplicates than small-corpus scaling laws predict. Thus, models that recognize semantic duplicates also encounter more of them, potentially disrupting predictable scaling.

A Note on Synthetic Data: Synthetic data has become a popular supplement to real data during pretraining and continued pretraining (Mishra et al., 2022; Chen et al., 2024b; Yang et al., 2024; Kang et al., 2025; Qin et al., 2025), though its diversity as a future alternative to real data remains uncertain. We repeat the experiment described in Section 3.1 for the fully-synthetic, 44M-document Recycling-the-Web pretraining corpus (Nguyen et al., 2025). We find that divergence from power law scaling (Figure 3) appears an order of magnitude earlier for synthetic pretraining data (Figure 5).

4 Impact on Training

We now probe practical implications. If semantic collisions reduce effective uniqueness, scaling-ladder extrapolation can fail. Because we cannot train models at the scale where *semantic* duplicates are recognized in our controlled setting, we model semantic collisions via *exact* document repeats (sampling with replacement), a pessimistic, worst-case proxy for repeated training signals.

4.1 Experimental Setup

We sample pools of K unique documents from FineWeb-Edu-Dedup, with K ranging from 10^5 through 10^8. We construct training streams by sampling with replacement from each pool, inducing exact repeats. As a reference, we also train on streams constructed to minimize repeats ("approximately infinite unique data") by sampling without replacement from FineWeb-Edu-Dedup.

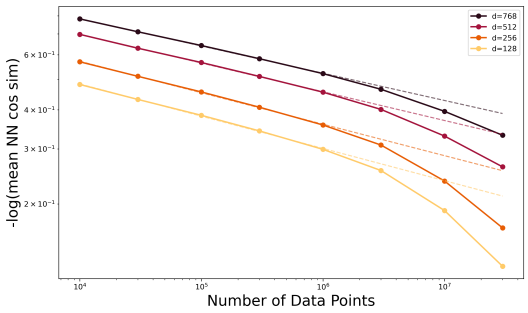


Figure 5: **Nearest-neighbor cosine similarity scaling laws collapse an order of magnitude earlier for synthetic datasets:** We embed the fully-synthetic pretraining dataset Recycling-the-Web (Nguyen et al., 2025) and find that the scaling law discovered in Figure 3 occurs an order of magnitude earlier for synthetic data, suggesting that the diversity of synthetic datasets should be improved.

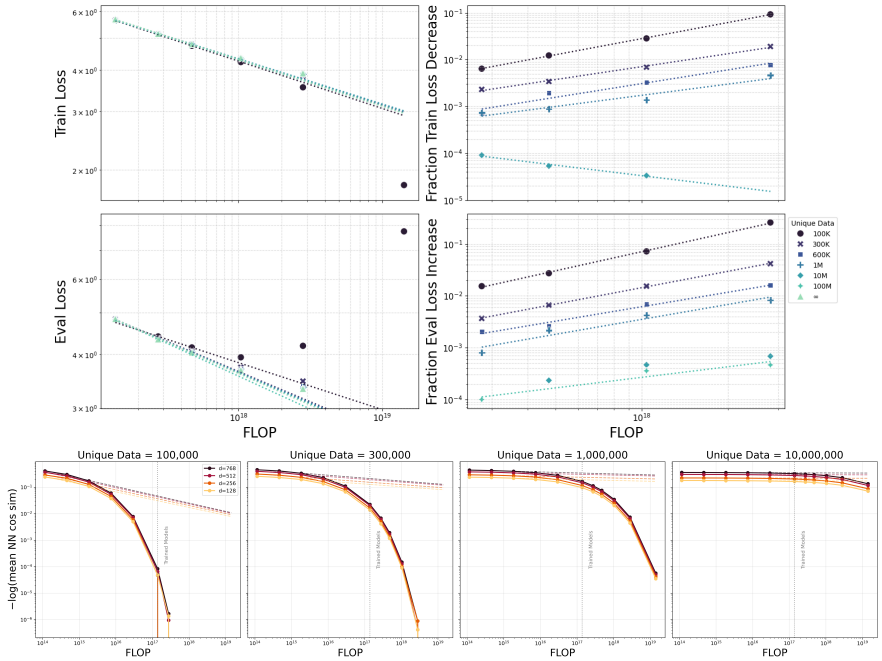


Figure 6: **Finite unique data pools induce scale-dependent degradation and break naive scaling extrapolation.** We train model ladders at matched compute, sampling with replacement from pools of K documents (exact repeats allowed). The approximately-infinite baseline has negligible repeats. Left: train and validation loss versus compute for each K. Right: fractional loss increase from the baseline (Eq. 8). Small models scale normally under small K, while larger models show increasing penalties; scaling ladders can underestimate main-run loss when effective uniqueness is limited.

We train scaling ladders of decoder-only, Chinchilla-optimal (Hoffmann et al., 2022) transformers based on the Qwen architecture ranging from 34M–344M parameters (Qwen et al., 2025; Yang et al., 2025). We match runs by compute (FLOPs) and report train and validation cross-entropy.

4.2 Results and Discussion

Figure 6 shows that limiting K produces a scale-dependent degradation pattern. For smaller models, train and validation losses are consistent with standard scaling extrapolations, even when K is small; this can mislead scaling-ladder planning. For larger FLOP budgets, finite-K streams yield increasing loss penalties, breaking naive interpolation from smaller ladders with the same K.

Although eval losses do not scale predictably with FLOP budgets under unique data constraints, *fractional loss increase* relative to the approximately-infinite baseline remains predictable:

$$\mathrm{FracInc}(K) \coloneqq \frac{L(K) - L(\infty)}{L(\infty)}. \tag{8}$$

Scaling prediction is challenging without an infinite-unique-data baseline. Section 5.1 develops theory to restore predictable scaling behavior in the presence of semantic duplicates.

5 Theory: Effective Duplicates and Restored Scaling

5.1 Semantics as Hierarchical Latents and Semantic Duplicates

We model "same meaning, different surface form" via latent semantics and transformations. Let z denote a *semantic latent* (meaning), and let τ denote a *surface transformation* (language, paraphrase, formatting, casing, etc.). A document x is generated by sampling its semantic latent and surface form

$$z \sim p(z), \qquad \tau \sim p(\tau \mid z), \qquad x = \mathcal{G}(z, \tau). \tag{9}$$

Two documents x and x' are *semantic duplicates* if they share the same z but differ in τ. This abstraction covers translations into different languages: $x = \mathcal{G}(z, \tau_{\mathrm{EN}})$ and $x' = \mathcal{G}(z, \tau_{\mathrm{ZH}})$.

To capture compositional structure, we allow the semantic latent z itself to be hierarchical:

$$z^{(0)} \to z^{(1)} \to \cdots \to z^{(L)} \to x, \tag{10}$$

where $z^{(0)}$ is coarse semantics (topic/world knowledge) and $z^{(L)}$ is closest to surface form. In this view, "duplicates" are not a binary dataset property: two documents can share an ancestor latent at some depth but not others. A model that only learns shallow latents may treat translations as distinct, while a model that learns deeper invariances collapses them to the same effective representation.

This perspective is related to the Random Hierarchy Model (RHM), which generates examples by composing features along a tree and predicts scale-dependent learnability of deeper levels (Cagnetta et al., 2024). In our setting, increasing capability corresponds to learning deeper invariances in the latent hierarchy, enlarging the set of surface variants that collide into the same semantic latent.

Using notation from Section 2, we formalize "duplication" in terms of training signal rather than surface form. Let f_θ be a language model trained by next-token prediction on documents.

Definition 1 (Effective duplicates)**.** Fix $\varepsilon \in (0, 1)$. We call x and x' *ε-effective duplicates at θ* if

$$\mathrm{sim}_\theta(x, x') \geq 1 - \varepsilon. \tag{11}$$

As capability increases, (11) can merge previously distinct examples (e.g. translations).

For a document x with semantic latent $z = z(x)$, decompose the per-document gradient as

$$g(x; \theta) = \underbrace{\mu(\theta)}_{\text{global}} + \underbrace{\delta_z(\theta)}_{\text{semantic}} + \underbrace{\xi_x(\theta)}_{\text{surface/idiosyncratic}}, \tag{12}$$

where $\mathbb{E}[\delta_z] = 0$ and $\mathbb{E}[\xi_x \mid z] = 0$. Intuitively, δ_z captures the update direction shared by all surface forms of the same meaning, while ξ_x captures surface-specific gradient variations.

A convenient summary of semantic sensitivity at scale s (parameters/compute/training time) is the fraction of gradient energy explained by the semantic component in this decomposition:

$$\rho(s) \coloneqq \frac{\mathbb{E}\|\delta_z(\theta(s))\|_2^2}{\mathbb{E}\|g(x; \theta(s)) - \mu(\theta(s))\|_2^2} \in [0, 1]. \tag{13}$$

Assuming ξ_x is approximately isotropic and independent across surface forms of the same latent z, $\rho(s)$ controls expected gradient cosine similarity: $\mathbb{E}[\mathrm{sim}_{\theta(s)}(x, x') \mid z] \approx \rho(s)$ for semantic duplicates, and approximately zero for unrelated documents. Concretely, for semantic duplicates $x = \mathcal{G}(z, \tau)$ and $x' = \mathcal{G}(z, \tau')$ with the same z, the numerator satisfies $\mathbb{E}\langle g(x) - \mu,\, g(x') - \mu\rangle \approx \mathbb{E}\|\delta_z\|_2^2$, while the denominator is $\mathbb{E}\|g(x) - \mu\|_2^2 \approx \mathbb{E}\|\delta_z\|_2^2 + \mathbb{E}\|\xi_x\|_2^2$. See Appendix G for the derivation. Our gradient experiments (Section 2) show that transformations preserving z (e.g. translations) become increasingly aligned with training progress and model capability.

5.2 Replication, Redundancy, and Effective Uniqueness

Consider a stream sampled *with replacement* over semantic latents $z \in \mathcal{Z}$ with mixture weights $\{w_z\}$. In our experiments (Section 4), this corresponds to uniform sampling from K unique documents (so $w_z = 1/K$), but the latent view also covers non-uniform frequencies.

A key quantity is the (Simpson) latent collision probability Simpson (1949)

$$p_{\text{lat}} \coloneqq \mathbb{P}(z = z') = \sum_z w_z^2,$$

and the associated effective latent count

$$K_{\text{eff}} \coloneqq \frac{1}{p_{\text{lat}}} = \frac{1}{\sum_z w_z^2}. \tag{14}$$

(When $w_z \equiv 1/K$, we have $K_{\text{eff}} = K$.)

Let $x_1, \ldots, x_n$ be n iid draws from this mixture and define the averaged centered gradient $\bar{g}_n \coloneqq \frac{1}{n}\sum_{t=1}^{n}(g(x_t;\theta) - \mu(\theta))$. Assume a simplified correlation structure consistent with Eq. (12)

$$C(x, x') \approx \begin{cases} \sigma^2 & z(x) = z(x') \text{ and } x = x', \\ \rho(s)\,\sigma^2 & z(x) = z(x') \text{ and } x \neq x', \\ 0 & z(x) \neq z(x'), \end{cases} \tag{15}$$

where $C(x, x') \equiv E\langle g(x;\theta) - \mu,\, g(x';\theta) - \mu\rangle$.

Proposition 1 (Saturation of training signal)**.** *For K uniformly sampled classes under (15),*

$$\begin{aligned} \mathbb{E}\|\bar{g}_n\|_2^2 &\approx \frac{\sigma^2}{n}\left(1 + \rho(s)(n-1)p_{\text{lat}}\right) \\ &= \frac{\sigma^2}{n}\left(1 + \rho(s)\frac{n-1}{K_{\text{eff}}}\right). \end{aligned} \tag{16}$$

Equivalently, the averaged gradient behaves like an iid average with effective *sample size*

$$n_{\text{eff}}(n, K_{\text{eff}}; s) \coloneqq \frac{n}{1 + \rho(s)(n-1)/K_{\text{eff}}} \approx \min\left\{n, \frac{K_{\text{eff}}}{\rho(s)}\right\}. \tag{17}$$

Interpretation. When $n \ll K/\rho(s)$, redundancy is negligible and signal scales like $1/n$. When $n \gg K/\rho(s)$, semantic redundancy dominates and the number of effectively independent update directions saturates at $K/\rho(s)$. Because $\rho(s)$ increases with capability (Section 2), a finite-K stream becomes *more redundant* for larger models; effective uniqueness $K/\rho(s)$ shrinks with scale.

5.3 From Effective Reuse to a Restored Scaling Law

Let C denote training compute. Let $L(C, K_{\text{eff}})$ be eval loss when sampling with replacement from an effective semantic pool size K_{eff}, and let $L_\infty(C)$ be the baseline with effectively infinite uniqueness (negligible repeats). Define the normalized degradation relative to this high-uniqueness baseline

$$\Delta(C, K) \coloneqq \frac{L(C, K) - L_\infty(C)}{L_\infty(C)}. \tag{18}$$

The redundancy picture suggests that the relevant control variable is an *effective reuse ratio*

$$r_{\text{eff}}(C, K_{\text{eff}}) \coloneqq \frac{\rho(C)\, n(C)}{K_{\text{eff}}}, \tag{19}$$

where $n(C)$ counts training documents and $\rho(C)$ captures semantic alignment (Section 5.1).

Assumption (Power Law Penalty in Effective Reuse). Within the fitted regime, we posit

$$\Delta(C, K) \;\approx\; \lambda\, r_{\text{eff}}(C, K)^{\eta}. \tag{20}$$

This encodes no penalty for negligible reuse and a smooth increase with semantic redundancy.

Compute Dependence. Locally, we approximate $n(C)$ and $\rho(C)$ by power laws

$$n(C) \propto C^u, \qquad \rho(C) \propto C^v. \tag{21}$$

Then (20) yields

$$\Delta(C, K_{\text{eff}}) \approx \; aC^{\beta} K_{\text{eff}}^{-\gamma}, \quad \beta = \eta(u+v), \quad \gamma = \eta, \tag{22}$$

for $a > 0$. Equation (22) combines (i) reuse increasing with compute ($u > 0$), and (ii) semantic sensitivity increasing with compute ($v \geq 0$). A special *ratio-only* law $\Delta \propto (\sqrt{C}/K)^{\eta}$ corresponds to $u = 1/2$ and $v = 0$, which can be too restrictive when $\rho(C)$ grows with scale.

Restored Predictivity. Combining (18) and (22) gives the restored loss prediction:

$$L_{\text{pred}}(C, K_{\text{eff}}) = L_{\infty}(C)\Big(1 + a\, C^{\beta} K_{\text{eff}}^{-\gamma}\Big). \tag{23}$$

In our experiments (Section 4), $L_{\infty}(C)$ is measured directly from the "approximately infinite unique data" runs at the same compute, so restoring predictivity requires fitting only (a, β, γ). Appendix F derives a collision-aware scaling correction by combining a Hutter-style learning curve Hutter (2021) with an effective-sample-size reduction induced by duplicate/semantic-collision gradients.

Empirical Fit. On our controlled scaling ladders, the 3-parameter plane law (22) fits *all* eval losses across (C, K), including the breakdown regime, with small average relative error, whereas the 2-parameter ratio-only constraint can substantially underpredict the catastrophic $K = 10^5$ main run. Fitting the compute and uniqueness exponents separately improves agreement in this regime.

5.4 Estimating an Effective Semantic Pool Size

In real pretraining, the "number of unique semantic items" K is not directly observable. However, our restored scaling law only requires an *effective uniqueness*—the rate at which training samples collide under the model's semantic resolution. We estimate an effective K_{eff} using a *mean nearest-neighbor cosine* statistic from embeddings of the *training stream*, including repeats.

Embed and unit-normalize a fixed subsample of N_{meas} training-stream documents, including repeats. Their nearest-neighbor cosines and their mean are computed within this fixed subsample:

$$M_t \coloneqq \max_{s \neq t} \langle v_t, v_s \rangle, \quad \overline{M}_{N_{\text{meas}}} \coloneqq \frac{1}{N_{\text{meas}}} \sum_{t=1}^{N_{\text{meas}}} M_t. \tag{24}$$

Here $N = N_{\text{meas}}$. A mixture separates background neighbors with mean $m_0(N)$ from same-latent neighbors with similarity m_+. Estimate $m_0(N)$ from a high-uniqueness reference stream using the same embedding pipeline. Let q_N denote the probability of at least one same-latent neighbor among $N - 1$ other samples. The two-component model expresses the mean similarity as

$$\mathbb{E}[\overline{M}_N] \approx (1 - q_N)\, m_0(N) + q_N\, m_+. \tag{25}$$

Solving gives the estimated probability

$$\widehat{q}_N \coloneqq \text{clip}\left(\frac{\overline{M}_N - m_0(N)}{m_+ - m_0(N)},\, 0,\, 1\right). \tag{26}$$

For exact repeats, $m_+ = 1$; semantic collisions require calibration from known duplicate pairs. For mixture weights $\{w_z\}$, the probability of a same-latent partner among the other $N - 1$ samples is

$$\begin{aligned} q_{N_{\text{meas}}} &= 1 - \sum_z w_z\, (1 - w_z)^{N_{\text{meas}} - 1} \\ &\approx 1 - \exp\left(-\frac{N_{\text{meas}} - 1}{K_{\text{eff}}}\right), \end{aligned} \tag{27}$$

under the small-weight and concentration conditions in Appendix E. Inverting yields

$$\widehat{K}_{\text{eff}} \coloneqq \frac{N_{\text{meas}} - 1}{-\log(1 - \widehat{q}_{N_{\text{meas}}})}. \tag{28}$$

Substituting $\widehat{K}_{\text{eff}}$ into Equation (23) gives a scaling correction based on observable stream geometry. The measurement procedure and detailed derivation are deferred to Appendix H.

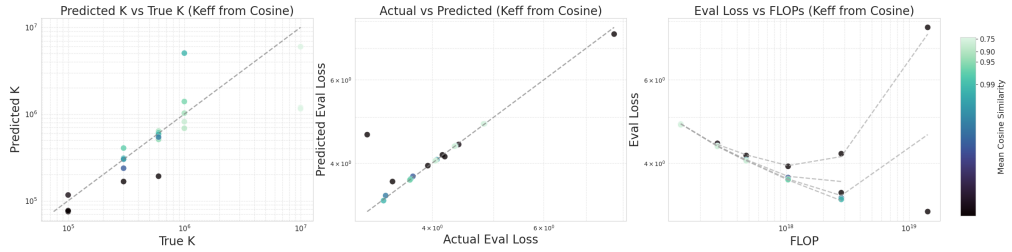


Figure 7: **Predictable scaling can be restored by accounting for limited semantic diversity:** Dataset size and mean NN cosine give K via Equation (28) (left). We substitute $\widehat{K}_{\text{eff}}$ into Equation (23) to estimate loss (center). The scaling curves closely match empirical eval losses (right).

Fit to the Controlled Ladder. On the runs used to fit the law in our controlled K-pool experiment, the plane law using the true pool size K achieves mean absolute relative error $\approx 0.77\%$ (median $\approx 0.28\%$). Replacing K with $\widehat{K}_{\text{eff}}$ estimated from mean NN cosine via Equations (26) and (28) achieves $\approx 0.90\%$ (median $\approx 0.24\%$). Thus, even with access only to a mean cosine statistic, $\widehat{K}_{\text{eff}}$ recovers most of the fit accuracy of the true-K scaling correction. See Figure 7.

Uniform Pools and Heavy Modes. If $w_z = 1/K$ for $z = 1, \ldots, K$, then (27) becomes

$$q_N = 1 - \sum_{z=1}^{K} \frac{1}{K}\left(1 - \frac{1}{K}\right)^{N-1} = 1 - \left(1 - \frac{1}{K}\right)^{N-1},$$

using $\log(1-x) \approx -x$ gives $q_N \approx 1 - \exp(-(N-1)/K)$, consistent with $K_{\text{eff}} = K$.

If some w_z are not small (a few "heavy" semantics), then: (i) the Poisson approximation (35) can be inaccurate for those modes, and (ii) the mean-field collapse (38) can be poor because the size-biased latent weight W is no longer concentrated. In that case, Eq. (31) remains correct and can be used directly, and K_{eff} still meaningfully summarizes pairwise collision probability via (33), but the single-exponential approximation to q_N may systematically overestimate collision probability.

Remark (Identifiability from mean NN cosine). The mapping $\overline{M}_{N_{\text{meas}}} \mapsto \widehat{q}_{N_{\text{meas}}}$ requires specifying both a background term $m_0(N_{\text{meas}})$ and a collision-similarity level m_+. From mean NN cosine alone, m_+ cannot be identified without external calibration; in the controlled with-replacement experiment, exact repeats imply $m_+ \approx 1$, while for semantic (non-exact) collisions one can calibrate m_+ using known semantic-duplicate pairs (e.g. translations) embedded by the same model.

6 Discussion and Future Directions

We discover an insidious source of scale-dependence that can impact the training of large language models, but not smaller language models: as model capabilities increase, training signals from semantically equivalent documents align. Thus, semantically equivalent documents in the corpora may act similarly to exact duplicates, harming model quality. Moreover, as training data scale, the number of semantic collisions increases far more quickly than one would expect based on trends gleaned from small corpora. We model this effect on small language models and propose scaling laws that account for semantic diversity in the dataset, restoring predictable scaling.

Our experiments have important implications for the future of language models. Industry convention has been to bet trillions of dollars on the bitter lesson: scale, scale, scale, and super-intelligence will follow (Sutton, 2019). The only obstruction on this path has been the limited number of training data in web-scale corpora. Frontier labs have tried to sidestep this obstacle by synthesizing massive corpora comprised of LLM-generated text. Our findings tell a cautionary tale about this approach: even if one can scale the raw number of tokens to asymptotically high regimes, semantic diversity may be just as important as data volume. As we show in Figure 5, synthetic data scales poorly with respect to semantic diversity. Our experiments emphasize the importance of seeding semantic diversity in synthetic data. There is only one other path: if the sum total of extant semantically distinct human thoughts is insufficient to train modern LMs, then labs must invest in more data-efficient training and architectures. We discuss limitations and future work in Appendix C.

AI USE STATEMENT

We used Codex and Claude Code to assist with experimental code, and GPT Pro for proof assistance. We also used Codex for editorial revisions and LaTeX formatting of the manuscript.

REFERENCES

Nura Aljaafari, Danilo S. Carvalho, and André Freitas. Trace for tracking the emergence of semantic representations in transformers, 2025. URL https://arxiv.org/abs/2505.17998.

Satwik Bhattamishra, Kabir Ahuja, and Navin Goyal. On the ability and limitations of transformers to recognize formal languages, 2020a. URL https://arxiv.org/abs/2009.11264.

Satwik Bhattamishra, Arkil Patel, and Navin Goyal. On the computational power of transformers and its implications in sequence modeling, 2020b. URL https://arxiv.org/abs/2006.09286.

Andrei Z. Broder. On the resemblance and containment of documents. In Bruno Carpentieri, Alfredo De Santis, Ugo Vaccaro, and James A. Storer (eds.), *Compression and Complexity of SEQUENCES 1997, Positano, Amalfitan Coast, Salerno, Italy, June 11-13, 1997, Proceedings*, pp. 21–29. IEEE, 1997. doi: 10.1109/SEQUEN.1997.666900. URL https://doi.org/10.1109/SEQUEN.1997.666900.

Francesco Cagnetta and Matthieu Wyart. Towards a theory of how the structure of language is acquired by deep neural networks. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 83119–83163. Curran Associates, Inc., 2024. doi: 10.52202/079017-2645. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/9740da1c07c7b451af14e11523f95271-Paper-Conference.pdf.

Francesco Cagnetta, Leonardo Petrini, Umberto M. Tomasini, Alessandro Favero, and Matthieu Wyart. How deep neural networks learn compositional data: The random hierarchy model. *Phys. Rev. X*, 14:031001, Jul 2024. doi: 10.1103/PhysRevX.14.031001. URL https://link.aps.org/doi/10.1103/PhysRevX.14.031001.

Francesco Cagnetta, Hyunmo Kang, and Matthieu Wyart. Learning curves theory for hierarchically compositional data with power-law distributed features, 2025. URL https://arxiv.org/abs/2505.07067.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models, 2021. URL https://arxiv.org/abs/2012.07805.

Hang Chen, Xinyu Yang, Jiaying Zhu, and Wenya Wang. Quantifying semantic emergence in language models, 2024a. URL https://arxiv.org/abs/2405.12617.

Hao Chen, Abdul Waheed, Xiang Li, Yidong Wang, Jindong Wang, Bhiksha Raj, and Marah I. Abdin. On the diversity of synthetic data and its impact on training large language models, 2024b. URL https://arxiv.org/abs/2410.15226.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, Kornraphop Kawintiranon, Tania Bedrax-Weiss, Oliver Wang, Ya Xu, Ollie Purkiss, Uri Mendlovic, Ilaï Deutel, Nam Nguyen, Adam Langley, Flip Korn, Lucia Rossazza, Alexandre Ramé, Sagar Waghmare, Helen Miller, Nathan Byrd, Ashrith Sheshan, Raia Hadsell, Sangnie Bhardwaj, Pawel Janus, Tero Rissa, Dan Horgan, Alvin Abdagic, Lior Belenki, James Allingham, Anima Singh, Theo Guidroz, Srivatsan Srinivasan, Herman Schmit, Kristen Chiafullo, Andre Elisseeff, Nilpa Jha, Prateek Kolhar, Leonard Berrada, Frank Ding, Xiance Si, Shrestha Basu Mallick, Franz Och, Sofia Erell, Eric Ni, Tejasi Latkar, Sherry Yang, Petar Sirkovic, Ziqiang Feng, Robert Leland, Rachel Hornung, Gang Wu, Charles Blundell, Hamidreza Alvari, Po-Sen Huang, Cathy Yip, Sanja Deur, Li Liu, Gabriela Surita, Pablo Duque, Dima Damen, Johnson Jia, Arthur Guez, Markus Mircea, Animesh Sinha, Alberto Magni, Paweł Stradomski, Tal Marian, Vlado Galić, Wenhu Chen, Hisham Husain, Achintya Singhal, Dominik Grewe, François-Xavier Aubet, Shuang Song, Lorenzo Blanco,

Leland Rechis, Lewis Ho, Rich Munoz, Kelvin Zheng, Jessica Hamrick, Kevin Mather, Hagai Taitelbaum, Eliza Rutherford, Yun Lei, Kuangyuan Chen, Anand Shukla, Erica Moreira, Eric Doi, Berivan Isik, Nir Shabat, Dominika Rogozińska, Kashyap Kolipaka, Jason Chang, Eugen Vušak, Srinivasan Venkatachary, Shadi Noghabi, Tarun Bharti, Younghoon Jun, Aleksandr Zaks, Simon Green, Jeshwanth Challagundla, William Wong, Muqthar Mohammad, Dean Hirsch, Yong Cheng, Iftekhar Naim, Lev Proleev, Damien Vincent, Aayush Singh, Maxim Krikun, Dilip Krishnan, Zoubin Ghahramani, Aviel Atias, Rajeev Aggarwal, Christo Kirov, Dimitrios Vytiniotis, Christy Koh, Alexandra Chronopoulou, Pawan Dogra, Vlad-Doru Ion, Gladys Tyen, Jason Lee, Felix Weissenberger, Trevor Strohman, Ashwin Balakrishna, Jack Rae, Marko Velic, Raoul de Liedekerke, Oded Elyada, Wentao Yuan, Canoee Liu, Lior Shani, Sergey Kishchenko, Bea Alessio, Yandong Li, Richard Song, Sam Kwei, Orion Jankowski, Aneesh Pappu, Youhei Namiki, Yenai Ma, Nilesh Tripuraneni, Colin Cherry, Marissa Ikonomidis, Yu-Cheng Ling, Colin Ji, Beka Westberg, Auriel Wright, Da Yu, David Parkinson, Swaroop Ramaswamy, Jerome Connor, Soheil Hassas Yeganeh, Snchit Grover, George Kenwright, Lubo Litchev, Chris Apps, Alex Tomala, Felix Halim, Alex Castro-Ros, Zefei Li, Anudhyan Boral, Pauline Sho, Michal Yarom, Eric Malmi, David Klinghoffer, Rebecca Lin, Alan Ansell, Pradeep Kumar S, Shubin Zhao, Siqi Zuo, Adam Santoro, Heng-Tze Cheng, Solomon Demmessie, Yuchi Liu, Nicole Brichtova, Allie Culp, Nathaniel Braun, Dan Graur, Will Ng, Nikhil Mehta, Aaron Phillips, Patrik Sundberg, Varun Godbole, Fangyu Liu, Yash Katariya, David Rim, Mojtaba Seyedhosseini, Sean Ammirati, Jonas Valfridsson, Mahan Malihi, Timothy Knight, Andeep Toor, Thomas Lampe, Abe Ittycheriah, Lewis Chiang, Chak Yeung, Alexandre Fréchette, Jinmeng Rao, Huisheng Wang, Himanshu Srivastava, Richard Zhang, Rocky Rhodes, Ariel Brand, Dean Weesner, Ilya Figotin, Felix Gimeno, Rachana Fellinger, Pierre Marcenac, José Leal, Eyal Marcus, Victor Cotruta, Rodrigo Cabrera, Sheryl Luo, Dan Garrette, Vera Axelrod, Sorin Baltateanu, David Barker, Dongkai Chen, Horia Toma, Ben Ingram, Jason Riesa, Chinmay Kulkarni, Yujing Zhang, Hongbin Liu, Chao Wang, Martin Polacek, Will Wu, Kai Hui, Adrian N Reyes, Yi Su, Megan Barnes, Ishaan Malhi, Anfal Siddiqui, Qixuan Feng, Mihai Damaschin, Daniele Pighin, Andreas Steiner, Samuel Yang, Ramya Sree Boppana, Simeon Ivanov, Arun Kandoor, Aditya Shah, Asier Mujika, Da Huang, Christopher A. Choquette-Choo, Mohak Patel, Tianhe Yu, Toni Creswell, Jerry, Liu, Catarina Barros, Yasaman Razeghi, Aurko Roy, Phil Culliton, Binbin Xiong, Jiaqi Pan, Thomas Strohmann, Tolly Powell, Babi Seal, Doug DeCarlo, Pranav Shyam, Kaan Katircioglu, Xuezhi Wang, Cassidy Hardin, Immanuel Odisho, Josef Broder, Oscar Chang, Arun Nair, Artem Shtefan, Maura O'Brien, Manu Agarwal, Sahitya Potluri, Siddharth Goyal, Amit Jhindal, Saksham Thakur, Yury Stuken, James Lyon, Kristina Toutanova, Fangxiaoyu Feng, Austin Wu, Ben Horn, Alek Wang, Alex Cullum, Gabe Taubman, Disha Shrivastava, Chongyang Shi, Hamish Tomlinson, Roma Patel, Tao Tu, Ada Maksutaj Oflazer, Francesco Pongetti, Mingyao Yang, Adrien Ali Taïga, Vincent Perot, Nuo Wang Pierse, Feng Han, Yoel Drori, Iñaki Iturrate, Ayan Chakrabarti, Legg Yeung, Dave Dopson, Yi ting Chen, Apoorv Kulshreshtha, Tongfei Guo, Philip Pham, Tal Schuster, Junquan Chen, Alex Polozov, Jinwei Xing, Huanjie Zhou, Praneeth Kacham, Doron Kukliansky, Antoine Miech, Sergey Yaroshenko, Ed Chi, Sholto Douglas, Hongliang Fei, Mathieu Blondel, Preethi Myla, Lior Madmoni, Xing Wu, Daniel Keysers, Kristian Kjems, Isabela Albuquerque, Lijun Yu, Joel D'sa, Michelle Plantan, Vlad Ionescu, Jaume Sanchez Elias, Abhirut Gupta, Manish Reddy Vuyyuru, Fred Alcober, Tong Zhou, Kaiyang Ji, Florian Hartmann, Subha Puttagunta, Hugo Song, Ehsan Amid, Anca Stefanoiu, Andrew Lee, Paul Pucciarelli, Emma Wang, Amit Raul, Slav Petrov, Isaac Tian, Valentin Anklin, Nana Nti, Victor Gomes, Max Schumacher, Grace Vesom, Alex Panagopoulos, Konstantinos Bousmalis, Daniel Andor, Josh Jacob, Yuan Zhang, Bill Rosgen, Matija Kecman, Matthew Tung, Alexandra Belias, Noah Goodman, Paul Covington, Brian Wieder, Nikita Saxena, Elnaz Davoodi, Muhuan Huang, Sharath Maddineni, Vincent Roulet, Folawiyo Campbell-Ajala, Pier Giuseppe Sessa, Xintian, Wu, Guangda Lai, Paul Collins, Alex Haig, Vytenis Sakenas, Xiaowei Xu, Marissa Giustina, Laurent El Shafey, Pichi Charoenpanit, Shefali Garg, Joshua Ainslie, Boone Severson, Montse Gonzalez Arenas, Shreya Pathak, Sujee Rajayogam, Jie Feng, Michiel Bakker, Sheng Li, Nevan Wichers, Jamie Rogers, Xinyang Geng, Yeqing Li, Rolf Jagerman, Chao Jia, Nadav Olmert, David Sharon, Matthew Mauger, Sandeep Mariserla, Hongxu Ma, Megha Mohabey, Kyuyeun Kim, Alek Andreev, Scott Pollom, Juliette Love, Vihan Jain, Priyanka Agrawal, Yannick Schroecker, Alisa Fortin, Manfred Warmuth, Ji Liu, Andrew Leach, Irina Blok, Ganesh Poomal Girirajan, Roee Aharoni, Benigno Uria, Andrei Sozanschi, Dan Goldberg, Lucian Ionita, Marco Tulio Ribeiro, Martin Zlocha, Vighnesh Birodkar, Sami Lachgar, Liangzhe Yuan, Himadri Choudhury, Matt Ginsberg, Fei Zheng, Gregory Dibb, Emily Graves, Swachhand Lokhande, Gabriel Rasskin, George-Cristian Muraru, Corbin Quick, Sandeep

Tata, Pierre Sermanet, Aditya Chawla, Itay Karo, Yan Wang, Susan Zhang, Orgad Keller, Anca Dragan, Guolong Su, Ian Chou, Xi Liu, Yiqing Tao, Shruthi Prabhakara, Marc Wilson, Ruibo Liu, Shibo Wang, Georgie Evans, David Du, Alfonso Castaño, Gautam Prasad, Mona El Mahdy, Sebastian Gerlach, Machel Reid, Jarrod Kahn, Amir Zait, Thanumalayan Sankaranarayana Pillai, Thatcher Ulrich, Guanyu Wang, Jan Wassenberg, Efrat Farkash, Kiran Yalasangi, Congchao Wang, Maria Bauza, Simon Bucher, Ting Liu, Jun Yan, Gary Leung, Vikas Sindhwani, Parker Barnes, Avi Singh, Ivan Jurin, Jichuan Chang, Niket Kumar Bhumihar, Sivan Eiger, Gui Citovsky, Ben Withbroe, Zhang Li, Siyang Xue, Niccolò Dal Santo, Georgi Stoyanov, Yves Raimond, Steven Zheng, Yilin Gao, Vít Listík, Sławek Kwasiborski, Rachel Saputro, Adnan Ozturel, Ganesh Mallya, Kushal Majmundar, Ross West, Paul Caron, Jinliang Wei, Lluis Castrejon, Sharad Vikram, Deepak Ramachandran, Nikhil Dhawan, Jiho Park, Sara Smoot, George van den Driessche, Yochai Blau, Chase Malik, Wei Liang, Roy Hirsch, Cicero Nogueira dos Santos, Eugene Weinstein, Aäron van den Oord, Sid Lall, Nicholas FitzGerald, Zixuan Jiang, Xuan Yang, Dale Webster, Ali Elqursh, Aedan Pope, Georges Rotival, David Raposo, Wanzheng Zhu, Jeff Dean, Sami Alabed, Dustin Tran, Arushi Gupta, Zach Gleicher, Jessica Austin, Edouard Rosseel, Megh Umekar, Dipanjan Das, Yinghao Sun, Kai Chen, Karolis Misiunas, Xiang Zhou, Yixian Di, Alyssa Loo, Josh Newlan, Bo Li, Vinay Ramasesh, Ying Xu, Alex Chen, Sudeep Gandhe, Radu Soricut, Nikita Gupta, Shuguang Hu, Seliem El-Sayed, Xavier Garcia, Idan Brusilovsky, Pu-Chin Chen, Andrew Bolt, Lu Huang, Alex Gurney, Zhiying Zhang, Alexander Pritzel, Jarek Wilkiewicz, Bryan Seybold, Bhargav Kanagal Shamanna, Felix Fischer, Josef Dean, Karan Gill, Ross Mcilroy, Abhishek Bhowmick, Jeremy Selier, Antoine Yang, Derek Cheng, Vladimir Magay, Jie Tan, Dhriti Varma, Christian Walder, Tomas Kocisky, Ryo Nakashima, Paul Natsev, Mike Kwong, Ionel Gog, Chiyuan Zhang, Sander Dieleman, Thomas Jimma, Andrey Ryabtsev, Siddhartha Brahma, David Steiner, Dayou Du, Ante Žužul, Mislav Žanić, Mukund Raghavachari, Willi Gierke, Zeyu Zheng, Dessie Petrova, Yann Dauphin, Yuchuan Liu, Ido Kessler, Steven Hand, Chris Duvarney, Seokhwan Kim, Hyo Lee, Léonard Hussenot, Jeffrey Hui, Josh Smith, Deepali Jain, Jiawei Xia, Gaurav Singh Tomar, Keyvan Amiri, Du Phan, Fabian Fuchs, Tobias Weyand, Nenad Tomasev, Alexandra Cordell, Xin Liu, Jonathan Mallinson, Pankaj Joshi, Andy Crawford, Arun Suggala, Steve Chien, Nick Fernando, Mariella Sanchez-Vargas, Duncan Williams, Phil Crone, Xiyang Luo, Igor Karpov, Jyn Shan, Terry Thurk, Robin Strudel, Paul Voigtlaender, Piyush Patil, Tim Dozat, Ali Khodaei, Sahil Singla, Piotr Ambroszczyk, Qiyin Wu, Yifan Chang, Brian Roark, Chaitra Hegde, Tianli Ding, Angelos Filos, Zhongru Wu, André Susano Pinto, Shuang Liu, Saarthak Khanna, Aditya Pandey, Siobhan Mcloughlin, Qiujia Li, Sam Haves, Allan Zhou, Elena Buchatskaya, Isabel Leal, Peter de Boursac, Nami Akazawa, Nina Anderson, Terry Chen, Krishna Somandepalli, Chen Liang, Sheela Goenka, Stephanie Winkler, Alexander Grushetsky, Yifan Ding, Jamie Smith, Fan Ye, Jordi Pont-Tuset, Eric Li, Ruichao Li, Tomer Golany, Dawid Wegner, Tao Jiang, Omer Barak, Yuan Shangguan, Eszter Vértes, Renee Wong, Jörg Bornschein, Alex Tudor, Michele Bevilacqua, Tom Schaul, Ankit Singh Rawat, Yang Zhao, Kyriakos Axiotis, Lei Meng, Cory McLean, Jonathan Lai, Jennifer Beattie, Nate Kushman, Yaxin Liu, Blair Kutzman, Fiona Lang, Jingchen Ye, Praneeth Netrapalli, Pushkar Mishra, Myriam Khan, Megha Goel, Rob Willoughby, David Tian, Honglei Zhuang, JD Chen, Zak Tsai, Tasos Kementsietsidis, Arjun Khare, James Keeling, Keyang Xu, Nathan Waters, Florent Altché, Ashok Popat, Bhavishya Mittal, David Saxton, Dalia El Badawy, Michael Mathieu, Zheng Zheng, Hao Zhou, Nishant Ranka, Richard Shin, Qingnan Duan, Tim Salimans, Ioana Mihailescu, Uri Shaham, Ming-Wei Chang, Yannis Assael, Nishanth Dikkala, Martin Izzard, Vincent Cohen-Addad, Cat Graves, Vlad Feinberg, Grace Chung, DJ Strouse, Danny Karmon, Sahand Sharifzadeh, Zoe Ashwood, Khiem Pham, Jon Blanton, Alex Vasiloff, Jarred Barber, Mark Geller, Aurick Zhou, Fedir Zubach, Tzu-Kuo Huang, Lei Zhang, Himanshu Gupta, Matt Young, Julia Proskurnia, Ronny Votel, Valentin Gabeur, Gabriel Barcik, Aditya Tripathi, Hongkun Yu, Geng Yan, Beer Changpinyo, Filip Pavetić, Amy Coyle, Yasuhisa Fujii, Jorge Gonzalez Mendez, Tianhao Zhou, Harish Rajamani, Blake Hechtman, Eddie Cao, Da-Cheng Juan, Yi-Xuan Tan, Valentin Dalibard, Yilun Du, Natalie Clay, Kaisheng Yao, Wenhao Jia, Dimple Vijaykumar, Yuxiang Zhou, Xinyi Bai, Wei-Chih Hung, Steven Pecht, Georgi Todorov, Nikhil Khadke, Pramod Gupta, Preethi Lahoti, Arnaud Autef, Karthik Duddu, James Lee-Thorp, Alexander Bykovsky, Tautvydas Misiunas, Sebastian Flennerhag, Santhosh Thangaraj, Jed McGiffin, Zack Nado, Markus Kunesch, Andreas Noever, Amir Hertz, Marco Liang, Victor Stone, Evan Palmer, Samira Daruki, Arijit Pramanik, Siim Põder, Austin Kyker, Mina Khan, Evgeny Sluzhaev, Marvin Ritter, Avraham Ruderman, Wenlei Zhou, Chirag Nagpal, Kiran Vodrahalli, George Necula, Paul Barham, Ellie Pavlick, Jay Hartford, Izhak Shafran, Long Zhao, Maciej Mikuła, Tom Eccles, Hidetoshi Shimokawa,

Kanav Garg, Luke Vilnis, Hanwen Chen, Ilia Shumailov, Kuang-Huei Lee, Abdelrahman Abdelhamed, Meiyan Xie, Vered Cohen, Ester Hlavnova, Dan Malkin, Chawin Sitawarin, James Lottes, Pauline Coquinot, Tianli Yu, Sandeep Kumar, Jingwei Zhang, Aroma Mahendru, Zafarali Ahmed, James Martens, Tao Chen, Aviel Boag, Daiyi Peng, Coline Devin, Arseniy Klimovskiy, Mary Phuong, Danny Vainstein, Jin Xie, Bhuvana Ramabhadran, Nathan Howard, Xinxin Yu, Gitartha Goswami, Jingyu Cui, Sam Shleifer, Mario Pinto, Chih-Kuan Yeh, Ming-Hsuan Yang, Sara Javanmardi, Dan Ethier, Chace Lee, Jordi Orbay, Suyog Kotecha, Carla Bromberg, Pete Shaw, James Thornton, Adi Gerzi Rosenthal, Shane Gu, Matt Thomas, Ian Gemp, Aditya Ayyar, Asahi Ushio, Aarush Selvan, Joel Wee, Chenxi Liu, Maryam Majzoubi, Weiren Yu, Jake Abernethy, Tyler Liechty, Renke Pan, Hoang Nguyen, Qiong, Hu, Sarah Perrin, Abhinav Arora, Emily Pitler, Weiyi Wang, Kaushik Shivakumar, Flavien Prost, Ben Limonchik, Jing Wang, Yi Gao, Timothee Cour, Shyamal Buch, Huan Gui, Maria Ivanova, Philipp Neubeck, Kelvin Chan, Lucy Kim, Huizhong Chen, Naman Goyal, Da-Woon Chung, Lu Liu, Yao Su, Anastasia Petrushkina, Jiajun Shen, Armand Joulin, Yuanzhong Xu, Stein Xudong Lin, Yana Kulizhskaya, Ciprian Chelba, Shobha Vasudevan, Eli Collins, Vasilisa Bashlovkina, Tony Lu, Doug Fritz, Jongbin Park, Yanqi Zhou, Chen Su, Richard Tanburn, Mikhail Sushkov, Mitchelle Rasquinha, Jinning Li, Jennifer Prendki, Yiming Li, Pallavi LV, Shriya Sharma, Hen Fitoussi, Hui Huang, Andrew Dai, Phuong Dao, Mike Burrows, Henry Prior, Danfeng Qin, Golan Pundak, Lars Lowe Sjoesund, Art Khurshudov, Zhenkai Zhu, Albert Webson, Elizabeth Kemp, Tat Tan, Saurabh Agrawal, Susie Sargsyan, Liqun Cheng, Jim Stephan, Tom Kwiatkowski, David Reid, Arunkumar Byravan, Assaf Hurwitz Michaely, Nicolas Heess, Luowei Zhou, Sonam Goenka, Viral Carpenter, Anselm Levskaya, Bo Wang, Reed Roberts, Rémi Leblond, Sharat Chikkerur, Stav Ginzburg, Max Chang, Robert Riachi, Chuqiao, Xu, Zalán Borsos, Michael Pliskin, Julia Pawar, Morgane Lustman, Hannah Kirkwood, Ankit Anand, Aditi Chaudhary, Norbert Kalb, Kieran Milan, Sean Augenstein, Anna Goldie, Laurel Prince, Karthik Raman, Yanhua Sun, Vivian Xia, Aaron Cohen, Zhouyuan Huo, Josh Camp, Seher Ellis, Lukas Zilka, David Vilar Torres, Lisa Patel, Sho Arora, Betty Chan, Jonas Adler, Kareem Ayoub, Jacky Liang, Fayaz Jamil, Jiepu Jiang, Simon Baumgartner, Haitian Sun, Yael Karov, Yaroslav Akulov, Hui Zheng, Irene Cai, Claudio Fantacci, James Rubin, Alex Rav Acha, Mengchao Wang, Nina D'Souza, Rohit Sathyanarayana, Shengyang Dai, Simon Rowe, Andrey Simanovsky, Omer Goldman, Yuheng Kuang, Xiaoyue Pan, Andrew Rosenberg, Tania Rojas-Esponda, Praneet Dutta, Amy Zeng, Irina Jurenka, Greg Farquhar, Yamini Bansal, Shariq Iqbal, Becca Roelofs, Ga-Young Joung, Parker Beak, Changwan Ryu, Ryan Poplin, Yan Wu, Jean-Baptiste Alayrac, Senaka Buthpitiya, Olaf Ronneberger, Caleb Habtegebriel, Wei Li, Paul Cavallaro, Aurora Wei, Guy Bensky, Timo Denk, Harish Ganapathy, Jeff Stanway, Pratik Joshi, Francesco Bertolini, Jessica Lo, Olivia Ma, Zachary Charles, Geta Sampemane, Himanshu Sahni, Xu Chen, Harry Askham, David Gaddy, Peter Young, Jiewen Tan, Matan Eyal, Arthur Bražinskas, Li Zhong, Zhichun Wu, Mark Epstein, Kai Bailey, Andrew Hard, Kamyu Lee, Sasha Goldshtein, Alex Ruiz, Mohammed Badawi, Matthias Lochbrunner, JK Kearns, Ashley Brown, Fabio Pardo, Theophane Weber, Haichuan Yang, Pan-Pan Jiang, Berkin Akin, Zhao Fu, Marcus Wainwright, Chi Zou, Meenu Gaba, Pierre-Antoine Manzagol, Wendy Kan, Yang Song, Karina Zainullina, Rui Lin, Jeongwoo Ko, Salil Deshmukh, Apoorv Jindal, James Svensson, Divya Tyam, Heri Zhao, Christine Kaeser-Chen, Scott Baird, Pooya Moradi, Jamie Hall, Qiuchen Guo, Vincent Tsang, Bowen Liang, Fernando Pereira, Suhas Ganesh, Ivan Korotkov, Jakub Adamek, Sridhar Thiagarajan, Vinh Tran, Charles Chen, Chris Tar, Sanil Jain, Ishita Dasgupta, Taylan Bilal, David Reitter, Kai Zhao, Giulia Vezzani, Yasmin Gehman, Pulkit Mehta, Lauren Beltrone, Xerxes Dotiwalla, Sergio Guadarrama, Zaheer Abbas, Stefani Karp, Petko Georgiev, Chun-Sung Ferng, Marc Brockschmidt, Liqian Peng, Christoph Hirnschall, Vikas Verma, Yingying Bi, Ying Xiao, Avigail Dabush, Kelvin Xu, Phil Wallis, Randall Parker, Qifei Wang, Yang Xu, Ilkin Safarli, Dinesh Tewari, Yin Zhang, Seungyeon Kim, Andrea Gesmundo, Mackenzie Thomas, Sergey Levi, Ahmed Chowdhury, Kanishka Rao, Peter Garst, Sam Conway-Rahman, Helen Ran, Kay McKinney, Zhisheng Xiao, Wenhao Yu, Rohan Agrawal, Axel Stjerngren, Catalin Ionescu, Jingjing Chen, Vivek Sharma, Justin Chiu, Fei Liu, Ken Franko, Clayton Sanford, Xingyu Cai, Paul Michel, Sanjay Ganapathy, Jane Labanowski, Zachary Garrett, Ben Vargas, Sean Sun, Bryan Gale, Thomas Buschmann, Guillaume Desjardins, Nimesh Ghelani, Palak Jain, Mudit Verma, Chulayuth Asawaroengchai, Julian Eisenschlos, Jitendra Harlalka, Hideto Kazawa, Don Metzler, Joshua Howland, Ying Jian, Jake Ades, Viral Shah, Tynan Gangwani, Seungji Lee, Roman Ring, Steven M. Hernandez, Dean Reich, Amer Sinha, Ashutosh Sathe, Joe Kovac, Ashleah Gill, Ajay Kannan, Andrea D'olimpio, Martin Sevenich, Jay Whang, Been Kim, Khe Chai Sim, Jilin Chen, Jiageng Zhang, Shuba Lall, Yossi Matias, Bill Jia, Abe

Friesen, Sara Nasso, Ashish Thapliyal, Bryan Perozzi, Ting Yu, Anna Shekhawat, Safeen Huda, Peter Grabowski, Eric Wang, Ashwin Sreevatsa, Hilal Dib, Mehadi Hassen, Parker Schuh, Vedrana Milutinovic, Chris Welty, Michael Quinn, Ali Shah, Bangju Wang, Gabe Barth-Maron, Justin Frye, Natalie Axelsson, Tao Zhu, Yukun Ma, Irene Giannoumis, Hanie Sedghi, Chang Ye, Yi Luan, Kevin Aydin, Bilva Chandra, Vivek Sampathkumar, Ronny Huang, Victor Lavrenko, Ahmed Eleryan, Zhi Hong, Steven Hansen, Sara Mc Carthy, Bidisha Samanta, Domagoj Ćevid, Xin Wang, Fangtao Li, Michael Voznesensky, Matt Hoffman, Andreas Terzis, Vikash Sehwag, Gil Fidel, Luheng He, Mu Cai, Yanzhang He, Alex Feng, Martin Nikoltchev, Samrat Phatale, Jason Chase, Rory Lawton, Ming Zhang, Tom Ouyang, Manuel Tragut, Mehdi Hafezi Manshadi, Arjun Narayanan, Jiaming Shen, Xu Gao, Tolga Bolukbasi, Nick Roy, Xin Li, Daniel Golovin, Liviu Panait, Zhen Qin, Guangxing Han, Thomas Anthony, Sneha Kudugunta, Viorica Patraucean, Aniket Ray, Xinyun Chen, Xiaochen Yang, Tanuj Bhatia, Pranav Talluri, Alex Morris, Andrija Ražnatović, Bethanie Brownfield, James An, Sheng Peng, Patrick Kane, Ce Zheng, Nico Duduta, Joshua Kessinger, James Noraky, Siqi Liu, Keran Rong, Petar Veličković, Keith Rush, Alex Goldin, Fanny Wei, Shiva Mohan Reddy Garlapati, Caroline Pantofaru, Okwan Kwon, Jianmo Ni, Eric Noland, Julia Di Trapani, Françoise Beaufays, Abhijit Guha Roy, Yinlam Chow, Aybuke Turker, Geoffrey Cideron, Lantao Mei, Jon Clark, Qingyun Dou, Matko Bošnjak, Ralph Leith, Yuqing Du, Amir Yazdanbakhsh, Milad Nasr, Chester Kwak, Suraj Satishkumar Sheth, Alex Kaskasoli, Ankesh Anand, Balaji Lakshminarayanan, Sammy Jerome, David Bieber, Chun-Te Chu, Alexandre Senges, Tianxiao Shen, Mukund Sridhar, Ndaba Ndebele, Benjamin Beyret, Shakir Mohamed, Mia Chen, Markus Freitag, Jiaxian Guo, Luyang Liu, Paul Roit, Heng Chen, Shen Yan, Tom Stone, JD Co-Reyes, Jeremy Cole, Salvatore Scellato, Shekoofeh Azizi, Hadi Hashemi, Alicia Jin, Anand Iyer, Marcella Valentine, András György, Arun Ahuja, Daniel Hernandez Diaz, Chen-Yu Lee, Nathan Clement, Weize Kong, Drew Garmon, Ishaan Watts, Kush Bhatia, Khyatti Gupta, Matt Miecnikowski, Hugo Vallet, Ankur Taly, Edward Loper, Saket Joshi, James Atwood, Jo Chick, Mark Collier, Fotis Iliopoulos, Ryan Trostle, Beliz Gunel, Ramiro Leal-Cavazos, Arnar Mar Hrafnkelsson, Michael Guzman, Xiaoen Ju, Andy Forbes, Jesse Emond, Kushal Chauhan, Ben Caine, Li Xiao, Wenjun Zeng, Alexandre Moufarek, Daniel Murphy, Maya Meng, Nitish Gupta, Felix Riedel, Anil Das, Elijah Lawal, Shashi Narayan, Tiberiu Sosea, James Swirhun, Linda Friso, Behnam Neyshabur, Jing Lu, Sertan Girgin, Michael Wunder, Edouard Yvinec, Aroonalok Pyne, Victor Carbune, Shruti Rijhwani, Yang Guo, Tulsee Doshi, Anton Briukhov, Max Bain, Ayal Hitron, Xuanhui Wang, Ashish Gupta, Ke Chen, Cosmo Du, Weiyang Zhang, Dhruv Shah, Arjun Akula, Max Dylla, Ashyana Kachra, Weicheng Kuo, Tingting Zou, Lily Wang, Luyao Xu, Jifan Zhu, Justin Snyder, Sachit Menon, Orhan Firat, Igor Mordatch, Yuan Yuan, Natalia Ponomareva, Rory Blevins, Lawrence Moore, Weijun Wang, Phil Chen, Martin Scholz, Artur Dwornik, Jason Lin, Sicheng Li, Diego Antognini, Te I, Xiaodan Song, Matt Miller, Uday Kalra, Adam Raveret, Oscar Akerlund, Felix Wu, Andrew Nystrom, Namrata Godbole, Tianqi Liu, Hannah DeBalsi, Jewel Zhao, Buhuang Liu, Avi Caciularu, Lauren Lax, Urvashi Khandelwal, Victoria Langston, Eric Bailey, Silvio Lattanzi, Yufei Wang, Neel Kovelamudi, Sneha Mondal, Guru Guruganesh, Nan Hua, Ofir Roval, Paweł Wesołowski, Rishikesh Ingale, Jonathan Halcrow, Tim Sohn, Christof Angermueller, Bahram Raad, Eli Stickgold, Eva Lu, Alec Kosik, Jing Xie, Timothy Lillicrap, Austin Huang, Lydia Lihui Zhang, Dominik Paulus, Clement Farabet, Alex Wertheim, Bing Wang, Rishabh Joshi, Chu ling Ko, Yonghui Wu, Shubham Agrawal, Lily Lin, XiangHai Sheng, Peter Sung, Tyler Breland-King, Christina Butterfield, Swapnil Gawde, Sumeet Singh, Qiao Zhang, Raj Apte, Shilpa Shetty, Adrian Hutter, Tao Li, Elizabeth Salesky, Federico Lebron, Jonni Kanerva, Michela Paganini, Arthur Nguyen, Rohith Vallu, Jan-Thorsten Peter, Sarmishta Velury, David Kao, Jay Hoover, Anna Bortsova, Colton Bishop, Shoshana Jakobovits, Alessandro Agostini, Alekh Agarwal, Chang Liu, Charles Kwong, Sasan Tavakkol, Ioana Bica, Alex Greve, Anirudh GP, Jake Marcus, Le Hou, Tom Duerig, Rivka Moroshko, Dave Lacey, Andy Davis, Julien Amelot, Guohui Wang, Frank Kim, Theofilos Strinopoulos, Hui Wan, Charline Le Lan, Shankar Krishnan, Haotian Tang, Peter Humphreys, Junwen Bai, Idan Heimlich Shtacher, Diego Machado, Chenxi Pang, Ken Burke, Dangyi Liu, Renga Aravamudhan, Yue Song, Ed Hirst, Abhimanyu Singh, Brendan Jou, Liang Bai, Francesco Piccinno, Chuyuan Kelly Fu, Robin Alazard, Barak Meiri, Daniel Winter, Charlie Chen, Mingda Zhang, Jens Heitkaemper, John Lambert, Jinhyuk Lee, Alexander Frömmgen, Sergey Rogulenko, Pranav Nair, Paul Niemczyk, Anton Bulyenov, Bibo Xu, Hadar Shemtov, Morteza Zadimoghaddam, Serge Toropov, Mateo Wirth, Hanjun Dai, Sreenivas Gollapudi, Daniel Zheng, Alex Kurakin, Chansoo Lee, Kalesha Bullard, Nicolas Serrano, Ivana Balazevic, Yang Li, Johan Schalkwyk, Mark Murphy, Mingyang Zhang, Kevin Sequeira, Romina Datta, Nishant Agrawal, Charles Sut-

ton, Nithya Attaluri, Mencher Chiang, Wael Farhan, Gregory Thornton, Kate Lin, Travis Choma, Hung Nguyen, Kingshuk Dasgupta, Dirk Robinson, Iulia Comşa, Michael Riley, Arjun Pillai, Basil Mustafa, Ben Golan, Amir Zandieh, Jean-Baptiste Lespiau, Billy Porter, David Ross, Sujeevan Rajayogam, Mohit Agarwal, Subhashini Venugopalan, Bobak Shahriari, Qiqi Yan, Hao Xu, Taylor Tobin, Pavel Dubov, Hongzhi Shi, Adrià Recasens, Anton Kovsharov, Sebastian Borgeaud, Lucio Dery, Shanthal Vasanth, Elena Gribovskaya, Linhai Qiu, Mahdis Mahdieh, Wojtek Skut, Elizabeth Nielsen, CJ Zheng, Adams Yu, Carrie Grimes Bostock, Shaleen Gupta, Aaron Archer, Chris Rawles, Elinor Davies, Alexey Svyatkovskiy, Tomy Tsai, Yoni Halpern, Christian Reisswig, Bartek Wydrowski, Bo Chang, Joan Puigcerver, Mor Hazan Taege, Jian Li, Eva Schnider, Xinjian Li, Dragos Dena, Yunhan Xu, Umesh Telang, Tianze Shi, Heiga Zen, Kyle Kastner, Yeongil Ko, Neesha Subramaniam, Aviral Kumar, Pete Blois, Zhuyun Dai, John Wieting, Yifeng Lu, Yoel Zeldes, Tian Xie, Anja Hauth, Alexandru Ţifrea, Yuqi Li, Sam El-Husseini, Dan Abolafia, Howard Zhou, Wen Ding, Sahra Ghalebikesabi, Carlos Guía, Andrii Maksai, Ágoston Weisz, Sercan Arik, Nick Sukhanov, Aga Świetlik, Xuhui Jia, Luo Yu, Weiyue Wang, Mark Brand, Dawn Bloxwich, Sean Kirmani, Zhe Chen, Alec Go, Pablo Sprechmann, Nithish Kannen, Alen Carin, Paramjit Sandhu, Isabel Edkins, Leslie Nooteboom, Jai Gupta, Loren Maggiore, Javad Azizi, Yael Pritch, Pengcheng Yin, Mansi Gupta, Danny Tarlow, Duncan Smith, Desi Ivanov, Mohammad Babaeizadeh, Ankita Goel, Satish Kambala, Grace Chu, Matej Kastelic, Michelle Liu, Hagen Soltau, Austin Stone, Shivani Agrawal, Min Kim, Kedar Soparkar, Srinivas Tadepalli, Oskar Bunyan, Rachel Soh, Arvind Kannan, DY Kim, Blake JianHang Chen, Afief Halumi, Sudeshna Roy, Yulong Wang, Olcan Sercinoglu, Gena Gibson, Sijal Bhatnagar, Motoki Sano, Daniel von Dincklage, Qingchun Ren, Blagoj Mitrevski, Mirek Olšák, Jennifer She, Carl Doersch, Jilei, Wang, Bingyuan Liu, Qijun Tan, Tamar Yakar, Tris Warkentin, Alex Ramirez, Carl Lebsack, Josh Dillon, Rajiv Mathews, Tom Cobley, Zelin Wu, Zhuoyuan Chen, Jon Simon, Swaroop Nath, Tara Sainath, Alexei Bendebury, Ryan Julian, Bharath Mankalale, Daria Ćurko, Paulo Zacchello, Adam R. Brown, Kiranbir Sodhia, Heidi Howard, Sergi Caelles, Abhinav Gupta, Gareth Evans, Anna Bulanova, Lesley Katzen, Roman Goldenberg, Anton Tsitsulin, Joe Stanton, Benoit Schillings, Vitaly Kovalev, Corey Fry, Rushin Shah, Kuo Lin, Shyam Upadhyay, Cheng Li, Soroush Radpour, Marcello Maggioni, Jing Xiong, Lukas Haas, Jenny Brennan, Aishwarya Kamath, Nikolay Savinov, Arsha Nagrani, Trevor Yacovone, Ryan Kappedal, Kostas Andriopoulos, Li Lao, YaGuang Li, Grigory Rozhdestvenskiy, Kazuma Hashimoto, Andrew Audibert, Sophia Austin, Daniel Rodriguez, Anian Ruoss, Garrett Honke, Deep Karkhanis, Xi Xiong, Qing Wei, James Huang, Zhaoqi Leng, Vittal Premachandran, Stan Bileschi, Georgios Evangelopoulos, Thomas Mensink, Jay Pavagadhi, Denis Teplyashin, Paul Chang, Linting Xue, Garrett Tanzer, Sally Goldman, Kaushal Patel, Shixin Li, Jeremy Wiesner, Ivy Zheng, Ian Stewart-Binks, Jie Han, Zhi Li, Liangchen Luo, Karel Lenc, Mario Lučić, Fuzhao Xue, Ryan Mullins, Alexey Guseynov, Chung-Ching Chang, Isaac Galatzer-Levy, Adam Zhang, Garrett Bingham, Grace Hu, Ale Hartman, Yue Ma, Jordan Griffith, Alex Irpan, Carey Radebaugh, Summer Yue, Lijie Fan, Victor Ungureanu, Christina Sorokin, Hannah Teufel, Peiran Li, Rohan Anil, Dimitris Paparas, Todd Wang, Chu-Cheng Lin, Hui Peng, Megan Shum, Goran Petrovic, Demetra Brady, Richard Nguyen, Klaus Macherey, Zhihao Li, Harman Singh, Madhavi Yenugula, Mariko Iinuma, Xinyi Chen, Kavya Kopparapu, Alexey Stern, Shachi Dave, Chandu Thekkath, Florence Perot, Anurag Kumar, Fangda Li, Yang Xiao, Matthew Bilotti, Mohammad Hossein Bateni, Isaac Noble, Lisa Lee, Amelio Vázquez-Reina, Julian Salazar, Xiaomeng Yang, Boyu Wang, Ela Gruzewska, Anand Rao, Sindhu Raghuram, Zheng Xu, Eyal Ben-David, Jieru Mei, Sid Dalmia, Zhaoyi Zhang, Yuchen Liu, Gagan Bansal, Helena Pankov, Steven Schwarcz, Andrea Burns, Christine Chan, Sumit Sanghai, Ricky Liang, Ethan Liang, Antoine He, Amy Stuart, Arun Narayanan, Yukun Zhu, Christian Frank, Bahar Fatemi, Amit Sabne, Oran Lang, Indro Bhattacharya, Shane Settle, Maria Wang, Brendan McMahan, Andrea Tacchetti, Livio Baldini Soares, Majid Hadian, Serkan Cabi, Timothy Chung, Nikita Putikhin, Gang Li, Jeremy Chen, Austin Tarango, Henryk Michalewski, Mehran Kazemi, Hussain Masoom, Hila Sheftel, Rakesh Shivanna, Archita Vadali, Ramona Comanescu, Doug Reid, Joss Moore, Arvind Neelakantan, Michaël Sander, Jonathan Herzig, Aviv Rosenberg, Mostafa Dehghani, JD Choi, Michael Fink, Reid Hayes, Eric Ge, Shitao Weng, Chia-Hua Ho, John Karro, Kalpesh Krishna, Lam Nguyen Thiet, Amy Skerry-Ryan, Daniel Eppens, Marco Andreetto, Navin Sarma, Silvano Bonacina, Burcu Karagol Ayan, Megha Nawhal, Zhihao Shan, Mike Dusenberry, Shantanu Thakoor, Sagar Gubbi, Duc Dung Nguyen, Reut Tsarfaty, Samuel Albanie, Jovana Mitrović, Meet Gandhi, Bo-Juen Chen, Alessandro Epasto, Georgi Stephanov, Ye Jin, Samuel Gehman, Aida Amini, Jack Weber, Feryal Behba-

hani, Shawn Xu, Miltos Allamanis, Xi Chen, Myle Ott, Claire Sha, Michal Jastrzebski, Hang Qi, David Greene, Xinyi Wu, Abodunrinwa Toki, Daniel Vlasic, Jane Shapiro, Ragha Kotikalapudi, Zhe Shen, Takaaki Saeki, Sirui Xie, Albin Cassirer, Shikhar Bharadwaj, Tatsuya Kiyono, Srinadh Bhojanapalli, Elan Rosenfeld, Sam Ritter, Jieming Mao, João Gabriel Oliveira, Zoltan Egyed, Bernd Bandemer, Emilio Parisotto, Keisuke Kinoshita, Juliette Pluto, Petros Maniatis, Steve Li, Yaohui Guo, Golnaz Ghiasi, Jean Tarbouriech, Srimon Chatterjee, Julie Jin, Katrina, Xu, Jennimaria Palomaki, Séb Arnold, Madhavi Sewak, Federico Piccinini, Mohit Sharma, Ben Albrecht, Sean Purser-haskell, Ashwin Vaswani, Chongyan Chen, Matheus Wisniewski, Qin Cao, John Aslanides, Nguyet Minh Phu, Maximilian Sieb, Lauren Agubuzu, Anne Zheng, Daniel Sohn, Marco Selvi, Anders Andreassen, Krishan Subudhi, Prem Eruvbetine, Oliver Woodman, Tomas Mery, Sebastian Krause, Xiaoqi Ren, Xiao Ma, Jincheng Luo, Dawn Chen, Wei Fan, Henry Griffiths, Christian Schuler, Alice Li, Shujian Zhang, Jean-Michel Sarr, Shixin Luo, Riccardo Patana, Matthew Watson, Dani Naboulsi, Michael Collins, Sailesh Sidhwani, Emiel Hoogeboom, Sharon Silver, Emily Caveness, Xiaokai Zhao, Mikel Rodriguez, Maxine Deines, Libin Bai, Patrick Griffin, Marco Tagliasacchi, Emily Xue, Spandana Raj Babbula, Bo Pang, Nan Ding, Gloria Shen, Elijah Peake, Remi Crocker, Shubha Srinivas Raghvendra, Danny Swisher, Woohyun Han, Richa Singh, Ling Wu, Vladimir Pchelin, Tsendsuren Munkhdalai, Dana Alon, Geoff Bacon, Efren Robles, Jannis Bulian, Melvin Johnson, George Powell, Felipe Tiengo Ferreira, Yaoyiran Li, Frederik Benzing, Mihajlo Velimirović, Hubert Soyer, William Kong, Tony, Nguyên, Zhen Yang, Jeremiah Liu, Joost van Amersfoort, Daniel Gillick, Baochen Sun, Nathalie Rauschmayr, Katie Zhang, Serena Zhan, Tao Zhou, Alexey Frolov, Chengrun Yang, Denis Vnukov, Louis Rouillard, Hongji Li, Amol Mandhane, Nova Fallen, Rajesh Venkataraman, Clara Huiyi Hu, Jennifer Brennan, Jenny Lee, Jerry Chang, Martin Sundermeyer, Zhufeng Pan, Rosemary Ke, Simon Tong, Alex Fabrikant, William Bono, Jindong Gu, Ryan Foley, Yiran Mao, Manolis Delakis, Dhruva Bhaswar, Roy Frostig, Nick Li, Avital Zipori, Cath Hope, Olga Kozlova, Swaroop Mishra, Josip Djolonga, Craig Schiff, Majd Al Merey, Eleftheria Briakou, Peter Morgan, Andy Wan, Avinatan Hassidim, RJ Skerry-Ryan, Kuntal Sengupta, Mary Jasarevic, Praveen Kallakuri, Paige Kunkle, Hannah Brennan, Tom Lieber, Hassan Mansoor, Julian Walker, Bing Zhang, Annie Xie, Goran Žužić, Adaeze Chukwuka, Alex Druinsky, Donghyun Cho, Rui Yao, Ferjad Naeem, Shiraz Butt, Eunyoung Kim, Zhipeng Jia, Mandy Jordan, Adam Lelkes, Mark Kurzeja, Sophie Wang, James Zhao, Andrew Over, Abhishek Chakladar, Marcel Prasetya, Neha Jha, Sriram Ganapathy, Yale Cong, Prakash Shroff, Carl Saroufim, Sobhan Miryoosefi, Mohamed Hammad, Tajwar Nasir, Weijuan Xi, Yang Gao, Young Maeng, Ben Hora, Chin-Yi Cheng, Parisa Haghani, Yoad Lewenberg, Caden Lu, Martin Matysiak, Naina Raisinghani, Huiyu Wang, Lexi Baugher, Rahul Sukthankar, Minh Giang, John Schultz, Noah Fiedel, Minmin Chen, Cheng-Chun Lee, Tapomay Dey, Hao Zheng, Shachi Paul, Celine Smith, Andy Ly, Yicheng Wang, Rishabh Bansal, Bartek Perz, Susanna Ricco, Stasha Blank, Vaishakh Keshava, Deepak Sharma, Marvin Chow, Kunal Lad, Komal Jalan, Simon Osindero, Craig Swanson, Jacob Scott, Anastasija Ilić, Xiaowei Li, Siddhartha Reddy Jonnalagadda, Afzal Shama Soudagar, Yan Xiong, Bat-Orgil Batsaikhan, Daniel Jarrett, Naveen Kumar, Maulik Shah, Matt Lawlor, Austin Waters, Mark Graham, Rhys May, Sabela Ramos, Sandra Lefdal, Zeynep Cankara, Nacho Cano, Brendan O'Donoghue, Jed Borovik, Frederick Liu, Jordan Grimstad, Mahmoud Alnahlawi, Katerina Tsihlas, Tom Hudson, Nikolai Grigorev, Yiling Jia, Terry Huang, Tobenna Peter Igwe, Sergei Lebedev, Xiaodan Tang, Igor Krivokon, Frankie Garcia, Melissa Tan, Eric Jia, Peter Stys, Shikhar Vashishth, Yu Liang, Balaji Venkatraman, Chenjie Gu, Anastasios Kementsietsidis, Chen Zhu, Junehyuk Jung, Yunfei Bai, Mohammad Javad Hosseini, Faruk Ahmed, Aditya Gupta, Xin Yuan, Shereen Ashraf, Shitij Nigam, Gautam Vasudevan, Pranjal Awasthi, Adi Mayrav Gilady, Zelda Mariet, Ramy Eskander, Haiguang Li, Hexiang Hu, Guillermo Garrido, Philippe Schlattner, George Zhang, Rohun Saxena, Petar Dević, Kritika Muralidharan, Ashwin Murthy, Yiqian Zhou, Min Choi, Arissa Wongpanich, Zhengdong Wang, Premal Shah, Yuntao Xu, Yiling Huang, Stephen Spencer, Alice Chen, James Cohan, Junjie Wang, Jonathan Tompson, Junru Wu, Ruba Haroun, Haiqiong Li, Blanca Huergo, Fan Yang, Tongxin Yin, James Wendt, Michael Bendersky, Rahma Chaabouni, Javier Snaider, Johan Ferret, Abhishek Jindal, Tara Thompson, Andrew Xue, Will Bishop, Shubham Milind Phal, Archit Sharma, Yunhsuan Sung, Prabakar Radhakrishnan, Mo Shomrat, Reeve Ingle, Roopali Vij, Justin Gilmer, Mihai Dorin Istin, Sam Sobell, Yang Lu, Emily Nottage, Dorsa Sadigh, Jeremiah Willcock, Tingnan Zhang, Steve Xu, Sasha Brown, Katherine Lee, Gary Wang, Yun Zhu, Yi Tay, Cheolmin Kim, Audrey Gutierrez, Abhanshu Sharma, Yongqin Xian, Sungyong Seo, Claire Cui, Elena Pochernina, Cip Baetu, Krzysztof Jastrzebski, Mimi Ly, Mohamed Elhawaty, Dan Suh, Eren Sezener, Pidong Wang, Nancy Yuen, George Tucker, Jiahao Cai, Zuguang

Yang, Cindy Wang, Alex Muzio, Hai Qian, Jae Yoo, Derek Lockhart, Kevin R. McKee, Mandy Guo, Malika Mehrotra, Artur Mendonça, Sanket Vaibhav Mehta, Sherry Ben, Chetan Tekur, Jiaqi Mu, Muye Zhu, Victoria Krakovna, Hongrae Lee, AJ Maschinot, Sébastien Cevey, Hyun-Jeong Choe, Aijun Bai, Hansa Srinivasan, Derek Gasaway, Nick Young, Patrick Siegler, Dan Holtmann-Rice, Vihari Piratla, Kate Baumli, Roey Yogev, Alex Hofer, Hado van Hasselt, Svetlana Grant, Yuri Chervonyi, David Silver, Andrew Hogue, Ayushi Agarwal, Kathie Wang, Preeti Singh, Four Flynn, Josh Lipschultz, Robert David, Lizzetth Bellot, Yao-Yuan Yang, Long Le, Filippo Graziano, Kate Olszewska, Kevin Hui, Akanksha Maurya, Nikos Parotsidis, Weijie Chen, Tayo Oguntebi, Joe Kelley, Anirudh Baddepudi, Johannes Mauerer, Gregory Shaw, Alex Siegman, Lin Yang, Shravya Shetty, Subhrajit Roy, Yunting Song, Wojciech Stokowiec, Ryan Burnell, Omkar Savant, Robert Busa-Fekete, Jin Miao, Samrat Ghosh, Liam MacDermed, Phillip Lippe, Mikhail Dektiarev, Zach Behrman, Fabian Mentzer, Kelvin Nguyen, Meng Wei, Siddharth Verma, Chris Knutsen, Sudeep Dasari, Zhipeng Yan, Petr Mitrichev, Xingyu Wang, Virat Shejwalkar, Jacob Austin, Srinivas Sunkara, Navneet Potti, Yan Virin, Christian Wright, Gaël Liu, Oriana Riva, Etienne Pot, Greg Kochanski, Quoc Le, Gargi Balasubramaniam, Arka Dhar, Yuguo Liao, Adam Bloniarz, Divyansh Shukla, Elizabeth Cole, Jong Lee, Sheng Zhang, Sushant Kafle, Siddharth Vashishtha, Parsa Mahmoudieh, Grace Chen, Raphael Hoffmann, Pranesh Srinivasan, Agustin Dal Lago, Yoav Ben Shalom, Zi Wang, Michael Elabd, Anuj Sharma, Junhyuk Oh, Suraj Kothawade, Maigo Le, Marianne Monteiro, Shentao Yang, Kaiz Alarakyia, Robert Geirhos, Diana Mincu, Håvard Garnes, Hayato Kobayashi, Soroosh Mariooryad, Kacper Krasowiak, Zhixin, Lai, Shibl Mourad, Mingqiu Wang, Fan Bu, Ophir Aharoni, Guanjie Chen, Abhimanyu Goyal, Vadim Zubov, Ankur Bapna, Elahe Dabir, Nisarg Kothari, Kay Lamerigts, Nicola De Cao, Jeremy Shar, Christopher Yew, Nitish Kulkarni, Dre Mahaarachchi, Mandar Joshi, Zhenhai Zhu, Jared Lichtarge, Yichao Zhou, Hannah Muckenhirn, Vittorio Selo, Oriol Vinyals, Peter Chen, Anthony Brohan, Vaibhav Mehta, Sarah Cogan, Ruth Wang, Ty Geri, Wei-Jen Ko, Wei Chen, Fabio Viola, Keshav Shivam, Lisa Wang, Madeleine Clare Elish, Raluca Ada Popa, Sébastien Pereira, Jianqiao Liu, Raphael Koster, Donnie Kim, Gufeng Zhang, Sayna Ebrahimi, Partha Talukdar, Yanyan Zheng, Petra Poklukar, Ales Mikhalap, Dale Johnson, Anitha Vijayakumar, Mark Omernick, Matt Dibb, Ayush Dubey, Qiong Hu, Apurv Suman, Vaibhav Aggarwal, Ilya Kornakov, Fei Xia, Wing Lowe, Alexey Kolganov, Ted Xiao, Vitaly Nikolaev, Steven Hemingray, Bonnie Li, Joana Iljazi, Mikołaj Rybiński, Ballie Sandhu, Peggy Lu, Thang Luong, Rodolphe Jenatton, Vineetha Govindaraj, Hui, Li, Gabriel Dulac-Arnold, Wonpyo Park, Henry Wang, Abhinit Modi, Jean Pouget-Abadie, Kristina Greller, Rahul Gupta, Robert Berry, Prajit Ramachandran, Jinyu Xie, Liam McCafferty, Jianling Wang, Kilol Gupta, Hyeontaek Lim, Blaž Bratanič, Andy Brock, Ilia Akolzin, Jim Sproch, Dan Karliner, Duhyeon Kim, Adrian Goedeckemeyer, Noam Shazeer, Cordelia Schmid, Daniele Calandriello, Parul Bhatia, Krzysztof Choromanski, Ceslee Montgomery, Dheeru Dua, Ana Ramalho, Helen King, Yue Gao, Lynn Nguyen, David Lindner, Divya Pitta, Oleaser Johnson, Khalid Salama, Diego Ardila, Michael Han, Erin Farnese, Seth Odoom, Ziyue Wang, Xiangzhuo Ding, Norman Rink, Ray Smith, Harshal Tushar Lehri, Eden Cohen, Neera Vats, Tong He, Parthasarathy Gopavarapu, Adam Paszke, Miteyan Patel, Wouter Van Gansbeke, Lucia Loher, Luis Castro, Maria Voitovich, Tamara von Glehn, Nelson George, Simon Niklaus, Zach Eaton-Rosen, Nemanja Rakićević, Erik Jue, Sagi Perel, Carrie Zhang, Yuval Bahat, Angéline Pouget, Zhi Xing, Fantine Huot, Ashish Shenoy, Taylor Bos, Vincent Coriou, Bryan Richter, Natasha Noy, Yaqing Wang, Santiago Ontanon, Siyang Qin, Gleb Makarchuk, Demis Hassabis, Zhuowan Li, Mandar Sharma, Kumaran Venkatesan, Iurii Kemaev, Roxanne Daniel, Shiyu Huang, Saloni Shah, Octavio Ponce, Warren, Chen, Manaal Faruqui, Jialin Wu, Slavica Andačić, Szabolcs Payrits, Daniel McDuff, Tom Hume, Yuan Cao, MH Tessler, Qingze Wang, Yinan Wang, Ivor Rendulic, Eirikur Agustsson, Matthew Johnson, Tanya Lando, Andrew Howard, Sri Gayatri Sundara Padmanabhan, Mayank Daswani, Andrea Banino, Michael Kilgore, Jonathan Heek, Ziwei Ji, Alvaro Caceres, Conglong Li, Nora Kassner, Alexey Vlaskin, Zeyu Liu, Alex Grills, Yanhan Hou, Roykrong Sukkerd, Gowoon Cheon, Nishita Shetty, Larisa Markeeva, Piotr Stanczyk, Tejas Iyer, Yuan Gong, Shawn Gao, Keerthana Gopalakrishnan, Tim Blyth, Malcolm Reynolds, Avishkar Bhoopchand, Misha Bilenko, Dero Gharibian, Vicky Zayats, Aleksandra Faust, Abhinav Singh, Min Ma, Hongyang Jiao, Sudheendra Vijayanarasimhan, Lora Aroyo, Vikas Yadav, Sarah Chakera, Ashwin Kakarla, Vilobh Meshram, Karol Gregor, Gabriela Botea, Evan Senter, Dawei Jia, Geza Kovacs, Neha Sharma, Sebastien Baur, Kai Kang, Yifan He, Lin Zhuo, Marija Kostelac, Itay Laish, Songyou Peng, Louis O'Bryan, Daniel Kasenberg, Girish Ramchandra Rao, Edouard Leurent, Biao Zhang, Sage Stevens, Ana Salazar, Ye Zhang, Ivan Lobov, Jake Walker, Allen Porter, Morgan Redshaw, Han Ke, Abhishek Rao, Alex Lee, Hoi

Lam, Michael Moffitt, Jaeyoun Kim, Siyuan Qiao, Terry Koo, Robert Dadashi, Xinying Song, Mukund Sundararajan, Peng Xu, Chizu Kawamoto, Yan Zhong, Clara Barbu, Apoorv Reddy, Mauro Verzetti, Leon Li, George Papamakarios, Hanna Klimczak-Plucińska, Mary Cassin, Koray Kavukcuoglu, Rigel Swavely, Alain Vaucher, Jeffrey Zhao, Ross Hemsley, Michael Tschannen, Heming Ge, Gaurav Menghani, Yang Yu, Natalie Ha, Wei He, Xiao Wu, Maggie Song, Rachel Sterneck, Stefan Zinke, Dan A. Calian, Annie Marsden, Alejandro Cruzado Ruiz, Matteo Hessel, Almog Gueta, Benjamin Lee, Brian Farris, Manish Gupta, Yunjie Li, Mohammad Saleh, Vedant Misra, Kefan Xiao, Piermaria Mendolicchio, Gavin Buttimore, Varvara Krayvanova, Nigamaa Nayakanti, Matthew Wiethoff, Yash Pande, Azalia Mirhoseini, Ni Lao, Jasmine Liu, Yiqing Hua, Angie Chen, Yury Malkov, Dmitry Kalashnikov, Shubham Gupta, Kartik Audhkhasi, Yuexiang Zhai, Sudhindra Kopalle, Prateek Jain, Eran Ofek, Clemens Meyer, Khuslen Baatarsukh, Hana Strejček, Jun Qian, James Freedman, Ricardo Figueira, Michal Sokolik, Olivier Bachem, Raymond Lin, Dia Kharrat, Chris Hidey, Pingmei Xu, Dennis Duan, Yin Li, Muge Ersoy, Richard Everett, Kevin Cen, Rebeca Santamaria-Fernandez, Amir Taubenfeld, Ian Mackinnon, Linda Deng, Polina Zablotskaia, Shashank Viswanadha, Shivanker Goel, Damion Yates, Yunxiao Deng, Peter Choy, Mingqing Chen, Abhishek Sinha, Alex Mossin, Yiming Wang, Arthur Szlam, Susan Hao, Paul Kishan Rubenstein, Metin Toksoz-Exley, Miranda Aperghis, Yin Zhong, Junwhan Ahn, Michael Isard, Olivier Lacombe, Florian Luisier, Chrysovalantis Anastasiou, Yogesh Kalley, Utsav Prabhu, Emma Dunleavy, Shaan Bijwadia, Justin Mao-Jones, Kelly Chen, Rama Pasumarthi, Emily Wood, Adil Dostmohamed, Nate Hurley, Jiri Simsa, Alicia Parrish, Mantas Pajarskas, Matt Harvey, Ondrej Skopek, Yony Kochinski, Javier Rey, Verena Rieser, Denny Zhou, Sun Jae Lee, Trilok Acharya, Guowang Li, Joe Jiang, Xiaofan Zhang, Bryant Gipson, Ethan Mahintorabi, Marco Gelmi, Nima Khajehnouri, Angel Yeh, Kayi Lee, Loic Matthey, Leslie Baker, Trang Pham, Han Fu, Alex Pak, Prakhar Gupta, Cristina Vasconcelos, Adam Sadovsky, Brian Walker, Sissie Hsiao, Patrik Zochbauer, Andreea Marzoca, Noam Velan, Junhao Zeng, Gilles Baechler, Danny Driess, Divya Jain, Yanping Huang, Lizzie Tao, John Maggs, Nir Levine, Jon Schneider, Erika Gemzer, Samuel Petit, Shan Han, Zach Fisher, Dustin Zelle, Courtney Biles, Eugene Ie, Asya Fadeeva, Casper Liu, Juliana Vicente Franco, Adrian Collister, Hao Zhang, Renshen Wang, Ruizhe Zhao, Leandro Kieliger, Kurt Shuster, Rui Zhu, Boqing Gong, Lawrence Chan, Ruoxi Sun, Sujoy Basu, Roland Zimmermann, Jamie Hayes, Abhishek Bapna, Jasper Snoek, Weel Yang, Puranjay Datta, Jad Al Abdallah, Kevin Kilgour, Lu Li, SQ Mah, Yennie Jun, Morgane Rivière, Abhijit Karmarkar, Tammo Spalink, Tao Huang, Lucas Gonzalez, Duc-Hieu Tran, Averi Nowak, John Palowitch, Martin Chadwick, Ellie Talius, Harsh Mehta, Thibault Sellam, Philipp Fränken, Massimo Nicosia, Kyle He, Aditya Kini, David Amos, Sugato Basu, Harrison Jobe, Eleni Shaw, Qiantong Xu, Colin Evans, Daisuke Ikeda, Chaochao Yan, Larry Jin, Lun Wang, Sachin Yadav, Ilia Labzovsky, Ramesh Sampath, Ada Ma, Candice Schumann, Aditya Siddhant, Rohin Shah, John Youssef, Rishabh Agarwal, Natalie Dabney, Alessio Tonioni, Moran Ambar, Jing Li, Isabelle Guyon, Benny Li, David Soergel, Boya Fang, Georgi Karadzhov, Cristian Udrescu, Trieu Trinh, Vikas Raunak, Seb Noury, Dee Guo, Sonal Gupta, Mara Finkelstein, Denis Petek, Lihao Liang, Greg Billock, Pei Sun, David Wood, Yiwen Song, Xiaobin Yu, Tatiana Matejovicova, Regev Cohen, Kalyan Andra, David D'Ambrosio, Zhiwei Deng, Vincent Nallatamby, Ebrahim Songhori, Rumen Dangovski, Andrew Lampinen, Pankil Botadra, Adam Hillier, Jiawei Cao, Nagabhushan Baddi, Adhi Kuncoro, Toshihiro Yoshino, Ankit Bhagatwala, Marcáurelio Ranzato, Rylan Schaeffer, Tianlin Liu, Shuai Ye, Obaid Sarvana, John Nham, Chenkai Kuang, Isabel Gao, Jinoo Baek, Shubham Mittal, Ayzaan Wahid, Anita Gergely, Bin Ni, Josh Feldman, Carrie Muir, Pascal Lamblin, Wolfgang Macherey, Ethan Dyer, Logan Kilpatrick, Víctor Campos, Mukul Bhutani, Stanislav Fort, Yanif Ahmad, Aliaksei Severyn, Kleopatra Chatziprimou, Oleksandr Ferludin, Mason Dimarco, Aditya Kusupati, Joe Heyward, Dan Bahir, Kevin Villela, Katie Millican, Dror Marcus, Sanaz Bahargam, Caglar Unlu, Nicholas Roth, Zichuan Wei, Siddharth Gopal, Deepanway Ghoshal, Edward Lee, Sharon Lin, Jennie Lees, Dayeong Lee, Anahita Hosseini, Connie Fan, Seth Neel, Marcus Wu, Yasemin Altun, Honglong Cai, Enrique Piqueras, Josh Woodward, Alessandro Bissacco, Salem Haykal, Mahyar Bordbar, Prasha Sundaram, Sarah Hodkinson, Daniel Toyama, George Polovets, Austin Myers, Anu Sinha, Tomer Levinboim, Kashyap Krishnakumar, Rachita Chhaparia, Tatiana Sholokhova, Nitesh Bharadwaj Gundavarapu, Ganesh Jawahar, Haroon Qureshi, Jieru Hu, Nikola Momchev, Matthew Rahtz, Renjie Wu, Aishwarya P S, Kedar Dhamdhere, Meiqi Guo, Umang Gupta, Ali Eslami, Mariano Schain, Michiel Blokzijl, David Welling, Dave Orr, Levent Bolelli, Nicolas Perez-Nieves, Mikhail Sirotenko, Aman Prasad, Arjun Kar, Borja De Balle Pigem, Tayfun Terzi, Gellért Weisz, Dipankar Ghosh, Aditi Mavalankar, Dhruv Madeka, Kaspar Daugaard, Hartwig Adam, Viraj Shah, Dana Berman, Mag-

gie Tran, Steven Baker, Ewa Andrejczuk, Grishma Chole, Ganna Raboshchuk, Mahdi Mirzazadeh, Thais Kagohara, Shimu Wu, Christian Schallhart, Bernett Orlando, Chen Wang, Alban Rrustemi, Hao Xiong, Hao Liu, Arpi Vezer, Nolan Ramsden, Shuo yiin Chang, Sidharth Mudgal, Yan Li, Nino Vieillard, Yedid Hoshen, Farooq Ahmad, Ambrose Slone, Amy Hua, Natan Potikha, Mirko Rossini, Jon Stritar, Sushant Prakash, Zifeng Wang, Xuanyi Dong, Alireza Nazari, Efrat Nehoran, Kaan Tekelioglu, Yinxiao Li, Kartikeya Badola, Tom Funkhouser, Yuanzhen Li, Varun Yerram, Ramya Ganeshan, Daniel Formoso, Karol Langner, Tian Shi, Huijian Li, Yumeya Yamamori, Amayika Panda, Alaa Saade, Angelo Scorza Scarpati, Chris Breaux, CJ Carey, Zongwei Zhou, Cho-Jui Hsieh, Sophie Bridgers, Alena Butryna, Nishesh Gupta, Vaibhav Tulsyan, Sanghyun Woo, Evgenii Eltyshev, Will Grathwohl, Chanel Parks, Seth Benjamin, Rina Panigrahy, Shenil Dodhia, Daniel De Freitas, Chris Sauer, Will Song, Ferran Alet, Jackson Tolins, Cosmin Paduraru, Xingyi Zhou, Brian Albert, Zizhao Zhang, Lei Shu, Mudit Bansal, Sarah Nguyen, Amir Globerson, Owen Xiao, James Manyika, Tom Hennigan, Rong Rong, Josip Matak, Anton Bakalov, Ankur Sharma, Danila Sinopalnikov, Andrew Pierson, Stephen Roller, Geoff Brown, Mingcen Gao, Toshiyuki Fukuzawa, Amin Ghafouri, Kenny Vassigh, Iain Barr, Zhicheng Wang, Anna Korsun, Rajesh Jayaram, Lijie Ren, Tim Zaman, Samira Khan, Yana Lunts, Dan Deutsch, Dave Uthus, Nitzan Katz, Masha Samsikova, Amr Khalifa, Nikhil Sethi, Jiao Sun, Luming Tang, Uri Alon, Xianghong Luo, Dian Yu, Abhishek Nayyar, Bryce Petrini, Will Truong, Vincent Hellendoorn, Nikolai Chinaev, Chris Alberti, Wei Wang, Jingcao Hu, Vahab Mirrokni, Ananth Balashankar, Avia Aharon, Aahil Mehta, Ahmet Iscen, Joseph Kready, Lucas Manning, Anhad Mohananey, Yuankai Chen, Anshuman Tripathi, Allen Wu, Igor Petrovski, Dawsen Hwang, Martin Baeuml, Shreyas Chandrakaladharan, Yuan Liu, Rey Coaguila, Maxwell Chen, Sally Ma, Pouya Tafti, Susheel Tatineni, Terry Spitz, Jiayu Ye, Paul Vicol, Mihaela Rosca, Adrià Puigdomènech, Zohar Yahav, Sanjay Ghemawat, Hanzhao Lin, Phoebe Kirk, Zaid Nabulsi, Sergey Brin, Bernd Bohnet, Ken Caluwaerts, Aditya Srikanth Veerubhotla, Dan Zheng, Zihang Dai, Petre Petrov, Yichong Xu, Ramin Mehran, Zhuo Xu, Luisa Zintgraf, Jiho Choi, Spurthi Amba Hombaiah, Romal Thoppilan, Sashank Reddi, Lukasz Lew, Li Li, Kellie Webster, KP Sawhney, Lampros Lamprou, Siamak Shakeri, Mayank Lunayach, Jianmin Chen, Sumit Bagri, Alex Salcianu, Ying Chen, Yani Donchev, Charlotte Magister, Signe Nørly, Vitor Rodrigues, Tomas Izo, Hila Noga, Joe Zou, Thomas Köppe, Wenxuan Zhou, Kenton Lee, Xiangzhu Long, Danielle Eisenbud, Anthony Chen, Connor Schenck, Chi Ming To, Peilin Zhong, Emanuel Taropa, Minh Truong, Omer Levy, Danilo Martins, Zhiyuan Zhang, Christopher Semturs, Kelvin Zhang, Alex Yakubovich, Pol Moreno, Lara McConnaughey, Di Lu, Sam Redmond, Lotte Weerts, Yonatan Bitton, Tiziana Refice, Nicolas Lacasse, Arthur Conmy, Corentin Tallec, Julian Odell, Hannah Forbes-Pollard, Arkadiusz Socala, Jonathan Hoech, Pushmeet Kohli, Alanna Walton, Rui Wang, Mikita Sazanovich, Kexin Zhu, Andrei Kapishnikov, Rich Galt, Matthew Denton, Ben Murdoch, Caitlin Sikora, Kareem Mohamed, Wei Wei, Uri First, Tim McConnell, Luis C. Cobo, James Qin, Thi Avrahami, Daniel Balle, Yu Watanabe, Annie Louis, Adam Kraft, Setareh Ariafar, Yiming Gu, Eugénie Rives, Charles Yoon, Andrei Rusu, James Cobon-Kerr, Chris Hahn, Jiaming Luo, Yuvein, Zhu, Niharika Ahuja, Rodrigo Benenson, Raphaël Lopez Kaufman, Honglin Yu, Lloyd Hightower, Junlin Zhang, Darren Ni, Lisa Anne Hendricks, Gabby Wang, Gal Yona, Lalit Jain, Pablo Barrio, Surya Bhupatiraju, Siva Velusamy, Allan Dafoe, Sebastian Riedel, Tara Thomas, Zhe Yuan, Mathias Bellaiche, Sheena Panthaplackel, Klemen Kloboves, Sarthak Jauhari, Canfer Akbulut, Todor Davchev, Evgeny Gladchenko, David Madras, Aleksandr Chuklin, Tyrone Hill, Quan Yuan, Mukundan Madhavan, Luke Leonhard, Dylan Scandinaro, Qihang Chen, Ning Niu, Arthur Douillard, Bogdan Damoc, Yasumasa Onoe, Fabian Pedregosa, Fred Bertsch, Chas Leichner, Joseph Pagadora, Jonathan Malmaud, Sameera Ponda, Andy Twigg, Oleksii Duzhyi, Jingwei Shen, Miaosen Wang, Roopal Garg, Jing Chen, Utku Evci, Jonathan Lee, Leon Liu, Koji Kojima, Masa Yamaguchi, Arunkumar Rajendran, AJ Piergiovanni, Vinodh Kumar Rajendran, Marco Fornoni, Gabriel Ibagon, Harry Ragan, Sadh MNM Khan, John Blitzer, Andrew Bunner, Guan Sun, Takahiro Kosakai, Scott Lundberg, Ndidi Elue, Kelvin Guu, SK Park, Jane Park, Arunachalam Narayanaswamy, Chengda Wu, Jayaram Mudigonda, Trevor Cohn, Hairong Mu, Ravi Kumar, Laura Graesser, Yichi Zhang, Richard Killam, Vincent Zhuang, Mai Giménez, Wael Al Jishi, Ruy Ley-Wild, Alex Zhai, Kazuki Osawa, Diego Cedillo, Jialu Liu, Mayank Upadhyay, Marcin Sieniek, Roshan Sharma, Tom Paine, Anelia Angelova, Sravanti Addepalli, Carolina Parada, Kingshuk Majumder, Avery Lamp, Sanjiv Kumar, Xiang Deng, Artiom Myaskovsky, Tea Sabolić, Jeffrey Dudek, Sarah York, Félix de Chaumont Quitry, Jiazhong Nie, Dee Cattle, Alok Gunjan, Bilal Piot, Waleed Khawaja, Seojin Bang, Simon Wang, Siavash Khodadadeh, Raghavender R, Praynaa Rawlani, Richard Powell, Kevin Lee, Johannes Griesser, GS Oh, Ce-

sar Magalhaes, Yujia Li, Simon Tokumine, Hadas Natalie Vogel, Dennis Hsu, Arturo BC, Disha Jindal, Matan Cohen, Zi Yang, Junwei Yuan, Dario de Cesare, Tony Bruguier, Jun Xu, Monica Roy, Alon Jacovi, Dan Belov, Rahul Arya, Phoenix Meadowlark, Shlomi Cohen-Ganor, Wenting Ye, Patrick Morris-Suzuki, Praseem Banzal, Gan Song, Pranavaraj Ponnuramu, Fred Zhang, George Scrivener, Salah Zaiem, Alif Raditya Rochman, Kehang Han, Badih Ghazi, Kate Lee, Shahar Drath, Daniel Suo, Antonious Girgis, Pradeep Shenoy, Duy Nguyen, Douglas Eck, Somit Gupta, Le Yan, Joao Carreira, Anmol Gulati, Ruoxin Sang, Daniil Mirylenka, Emma Cooney, Edward Chou, Mingyang Ling, Cindy Fan, Ben Coleman, Guilherme Tubone, Ravin Kumar, Jason Baldridge, Felix Hernandez-Campos, Angeliki Lazaridou, James Besley, Itay Yona, Neslihan Bulut, Quentin Wellens, AJ Pierigiovanni, Jasmine George, Richard Green, Pu Han, Connie Tao, Geoff Clark, Chong You, Abbas Abdolmaleki, Justin Fu, Tongzhou Chen, Ashwin Chaugule, Angad Chandorkar, Altaf Rahman, Will Thompson, Penporn Koanantakool, Mike Bernico, Jie Ren, Andrey Vlasov, Sergei Vassilvitskii, Maciej Kula, Yizhong Liang, Dahun Kim, Yangsibo Huang, Chengxi Ye, Dmitry Lepikhin, and Wesley Helmholz. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Tobias Domhan, Jost Tobias Springenberg, and Frank Hutter. Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves. In Qiang Yang and Michael J. Wooldridge (eds.), *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pp. 3460–3468. AAAI Press, 2015. URL http://ijcai.org/Abstract/15/487.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library, 2025. URL https://arxiv.org/abs/2401.08281.

Google. Cloud translation api reference. https://cloud.google.com/translate/docs/reference/rest. Accessed: 2026-01-24.

Michael Hahn. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, December 2020. ISSN 2307-387X. doi: 10.1162/tacl_a_00306. URL http://dx.doi.org/10.1162/tacl_a_00306.

Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, Scott Johnston, Ben Mann, Chris Olah, Catherine Olsson, Dario Amodei, Nicholas Joseph, Jared Kaplan, and Sam McCandlish. Scaling laws and interpretability of learning from repeated data, 2022. URL https://arxiv.org/abs/2205.10487.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically, 2017. URL https://arxiv.org/abs/1712.00409.

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, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. URL https://arxiv.org/abs/2203.15556.

Hugging Face. Llama4. https://huggingface.co/docs/transformers/en/model_doc/llama4, 2025. Hugging Face Transformers Documentation.

Marcus Hutter. Learning curve theory, 2021. URL https://arxiv.org/abs/2102.04074.

Maor Ivgi, Yair Carmon, and Jonathan Berant. Scaling laws under the microscope: Predicting transformer performance from small scale experiments, 2022. URL https://arxiv.org/abs/2202.06387.

Selim Jerad, Anej Svete, Sophie Hao, Ryan Cotterell, and William Merrill. Context-free recognition with transformers, 2026. URL https://arxiv.org/abs/2601.01754.

Charles Jin and Martin Rinard. Emergent representations of program semantics in language models trained on programs, 2024. URL https://arxiv.org/abs/2305.11169.

Arlind Kadra, Maciej Janowski, Martin Wistuba, and Josif Grabocka. Scaling laws for hyperparameter optimization, 2023. URL https://arxiv.org/abs/2302.00441.

Dayal Singh Kalra and Maissam Barkeshli. Why warmup the learning rate? underlying mechanisms and improvements, 2024. URL https://arxiv.org/abs/2406.09405.

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, 2025. URL https://arxiv.org/abs/2510.01631.

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.

Arham Khan, Robert Underwood, Carlo Siebenschuh, Yadu Babuji, Aswathy Ajith, Kyle Hippe, Ozan Gokdemir, Alexander Brace, Kyle Chard, and Ian Foster. Lshbloom: Memory-efficient, extreme-scale document deduplication, 2026. URL https://arxiv.org/abs/2411.04257.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions, 2020. URL https://arxiv.org/abs/1703.04730.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning, 2024. URL https://arxiv.org/abs/2205.13147.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better, 2022. URL https://arxiv.org/abs/2107.06499.

Gurmeet Singh Manku, Arvind Jain, and Anish Das Sarma. Detecting near-duplicates for web crawling. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pp. 141–150, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595936547. doi: 10.1145/1242572.1242592. URL https://doi.org/10.1145/1242572.1242592.

Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. An empirical model of large-batch training, 2018. URL https://arxiv.org/abs/1812.06162.

Samarth Mishra, Rameswar Panda, Cheng Perng Phoo, Chun-Fu Chen, Leonid Karlinsky, Kate Saenko, Venkatesh Saligrama, and Rogerio S. Feris. Task2sim : Towards effective pre-training and transfer from synthetic data, 2022. URL https://arxiv.org/abs/2112.00054.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models, 2025. URL https://arxiv.org/abs/2305.16264.

Thao Nguyen, Yang Li, Olga Golovneva, Luke Zettlemoyer, Sewoong Oh, Ludwig Schmidt, and Xian Li. Recycling the web: A method to enhance pre-training data quality and quantity for language models, 2025. URL https://arxiv.org/abs/2506.04689.

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, 2024. URL https://arxiv.org/abs/2406.17557.

Zhencan Peng, Zhizhi Wang, and Dong Deng. Near-duplicate sequence search at scale for large language model memorization evaluation. *Proc. ACM Manag. Data*, 1(2), June 2023. doi: 10.1145/3589324. URL https://doi.org/10.1145/3589324.

Tomer Porian, Mitchell Wortsman, Jenia Jitsev, Ludwig Schmidt, and Yair Carmon. Resolving discrepancies in compute-optimal scaling of language models, 2025. URL https://arxiv.org/abs/2406.19146.

Garima Pruthi, Frederick Liu, Mukund Sundararajan, and Satyen Kale. Estimating training data influence by tracing gradient descent, 2020. URL https://arxiv.org/abs/2002.08484.

Zeyu Qin, Qingxiu Dong, Xingxing Zhang, Li Dong, Xiaolong Huang, Ziyi Yang, Mahmoud Khademi, Dongdong Zhang, Hany Hassan Awadalla, Yi R. Fung, Weizhu Chen, Minhao Cheng, and Furu Wei. Scaling laws of synthetic data for language models, 2025. URL https://arxiv.org/abs/2503.19551.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Jonathan S. Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A constructive prediction of the generalization error across scales, 2019. URL https://arxiv.org/abs/1909.12673.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage?, 2023. URL https://arxiv.org/abs/2304.15004.

Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation, 2017. URL https://arxiv.org/abs/1611.01232.

Laura Ying Schulz, Daniel Mitropolsky, and Tomaso Poggio. Unraveling syntax: Language modeling and the substructure of grammars, 2026. URL https://arxiv.org/abs/2510.02524.

Antonio Sclocchi, Alessandro Favero, Noam Itzhak Levi, and Matthieu Wyart. Probing the latent hierarchical structure of data via diffusion models*. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(8):084005, aug 2025. doi: 10.1088/1742-5468/aded6c. URL https://doi.org/10.1088/1742-5468/aded6c.

E. H. Simpson. Measurement of diversity. *Nature*, 163(4148):688–688, 1949. doi: 10.1038/163688a0. URL https://doi.org/10.1038/163688a0.

Richard Sutton. The bitter lesson. http://www.incompleteideas.net/IncIdeas/BitterLesson.html, March 2019.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, Weiyi Wang, Zhe Li, Gus Martins, Jinhyuk Lee, Mark Sherwood, Juyeong Ji, Renjie Wu, Jingxiao Zheng, Jyotinder Singh, Abheesht Sharma, Divyashree Sreepathihalli, Aashi Jain, Adham Elarabawy, AJ Co, Andreas Doumanoglou, Babak Samari, Ben Hora, Brian Potetz, Dahun Kim, Enrique Alfonseca, Fedor Moiseev, Feng Han, Frank Palma Gomez, Gustavo Hernández Ábrego, Hesen Zhang, Hui Hui, Jay Han, Karan Gill, Ke Chen, Koert Chen, Madhuri Shanbhogue, Michael Boratko, Paul Suganthan, Sai Meher Karthik Duddu, Sandeep Mariserla, Setareh Ariafar, Shanfeng Zhang, Shijie Zhang, Simon Baumgartner, Sonam Goenka, Steve Qiu, Tanmaya Dabral, Trevor Walker, Vikram Rao, Waleed Khawaja, Wenlei Zhou, Xiaoqi Ren, Ye Xia, Yichang Chen, Yi-Ting Chen, Zhe Dong, Zhongli Ding, Francesco Visin, Gaël Liu, Jiageng Zhang, Kathleen Kenealy, Michelle Casbon, Ravin Kumar, Thomas Mesnard, Zach Gleicher, Cormac Brick, Olivier

Lacombe, Adam Roberts, Qin Yin, Yunhsuan Sung, Raphael Hoffmann, Tris Warkentin, Armand Joulin, Tom Duerig, and Mojtaba Seyedhosseini. Embeddinggemma: Powerful and lightweight text representations, 2025. URL https://arxiv.org/abs/2509.20354.

Hetong Wang, Pasquale Minervini, and Edoardo M. Ponti. Probing the emergence of cross-lingual alignment during llm training, 2024. URL https://arxiv.org/abs/2406.13229.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. Deepnet: Scaling transformers to 1,000 layers, 2022. URL https://arxiv.org/abs/2203.00555.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture, 2020. URL https://arxiv.org/abs/2002.04745.

Tingkai Yan, Haodong Wen, Binghui Li, Kairong Luo, Wenguang Chen, and Kaifeng Lyu. Larger datasets can be repeated more: A theoretical analysis of multi-epoch scaling in linear regression, 2026. URL https://arxiv.org/abs/2511.13421.

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.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer, 2022. URL https://arxiv.org/abs/2203.03466.

Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candès, and Tatsunori Hashimoto. Synthetic continued pretraining, 2024. URL https://arxiv.org/abs/2409.07431.

Hongyi Zhang, Yann N. Dauphin, and Tengyu Ma. Fixup initialization: Residual learning without normalization, 2019. URL https://arxiv.org/abs/1901.09321.

A Related Work

Predictable scaling has been central to deep learning since early work on learning curves and performance prediction (Domhan et al., 2015; Hestness et al., 2017). Despite the analytic intractability of modern neural networks, empirical scaling laws often predict loss as a function of model size, data size, and compute with high accuracy (Schoenholz et al., 2017; Rosenfeld et al., 2019; Kaplan et al., 2020; Hoffmann et al., 2022). Scaling predictability governs training recipes, including parameterization, learning rates, depth-to-width ratios, initialization, warmup, and batch size (Kadra et al., 2023; Yang et al., 2022; Xiong et al., 2020; Wang et al., 2022; Zhang et al., 2019; Kalra & Barkeshli, 2024; McCandlish et al., 2018). A challenge is identifying *scale-dependent* factors that undermine predictable extrapolation (Ivgi et al., 2022; Schaeffer et al., 2023; Porian et al., 2025). Our work highlights scale dependence from semantic duplicates and their frequency at web scale.

Hernandez et al. (2022) showed that repeating a small subset can reduce the effective parameter size predicted by scaling, and later work reports stronger repetition effects at scale. Because near-duplicates are common in web corpora (Peng et al., 2023), practical pipelines deploy hashing and approximate matching to remove "fuzzy duplicates" (Broder, 1997; Manku et al., 2007; Khan et al., 2026). Unlike settings with *explicit* repeats (e.g., injected duplicates or many epochs) (Hernandez et al., 2022; Muennighoff et al., 2025; Yan et al., 2026), we emphasize an *implicit* and *scale-dependent* notion of repetition: as models become semantically sensitive, semantically equivalent documents may function as duplicates, and the prevalence of semantic collisions grows with corpus scale.

Our gradient-based measurements relate to work that treats gradients as training signals and influence proxies (Pruthi et al., 2020; Koh & Liang, 2020). Our observation that semantic structure becomes more salient over training aligns with prior work on the emergence and probing of semantic representations during pretraining (Jin & Rinard, 2024; Chen et al., 2024a; Aljaafari et al., 2025; Wang et al., 2024). We connect semantic emergence to a potential failure mode: semantic duplicates can create redundant training signals that disproportionately affect capable LMs.

Related work studies how neural networks learn hierarchical and compositional structure. Recent theory introduces stylized latent-data models such as the Random Hierarchy Model (RHM), in which examples are generated by composing features along a tree (analogous to a grammar derivation), yielding sharp predictions about which levels of the hierarchy are learnable at a given scale (Cagnetta et al., 2024; 2025; Cagnetta & Wyart, 2024; Sclocchi et al., 2025). In language modeling, formal-language and grammar-based probes have been used to analyze whether attention-based architectures can represent and generalize hierarchical dependencies, including theoretical limitations of self-attention (Hahn, 2020) and empirical studies of Transformer recognition of formal languages (Bhattamishra et al., 2020a). Most recently, Schulz et al. (2026) directly characterizes how language models learn context-free grammars over training. Our work highlights a consequence of learning deeper invariances: as models become semantically/compositionally sensitive, semantically equivalent documents increasingly behave as effective duplicates, amplifying the impact of semantic collisions at corpus scale.

B Robustness to the Choice of Embedding Model

We compared embedding similarity with gradient similarity using Qwen3-Embedding-0.6B alongside EmbeddingGemma-300m, and found similar correlations across language models. We construct 3,000 document pairs: 1,000 original-document/matching-rewrite pairs, 1,000 pairs of rewrites of different documents, and 1,000 pairs of different original documents. For each of 25 language models, we compute the cosine similarity between per-document loss gradients, then measure its Pearson correlation with embedding cosine similarity for each embedder.

Agreement between the two embedders suggests the correspondence between embedding similarity and gradient alignment is not specific to EmbeddingGemma-300m. For Pythia-1B, the correlations are 0.8343 with Qwen3-Embedding-0.6B and 0.8230 with EmbeddingGemma-300m, while both yield correlations near zero for Pythia-31M, 70M, and 160M. Both also reproduce exceptions to a simple model-size trend, including strong correlations for Gemma-3-270M and weaker correlations for GPT-OSS-20B. These results support the robustness of the measured association across two em-

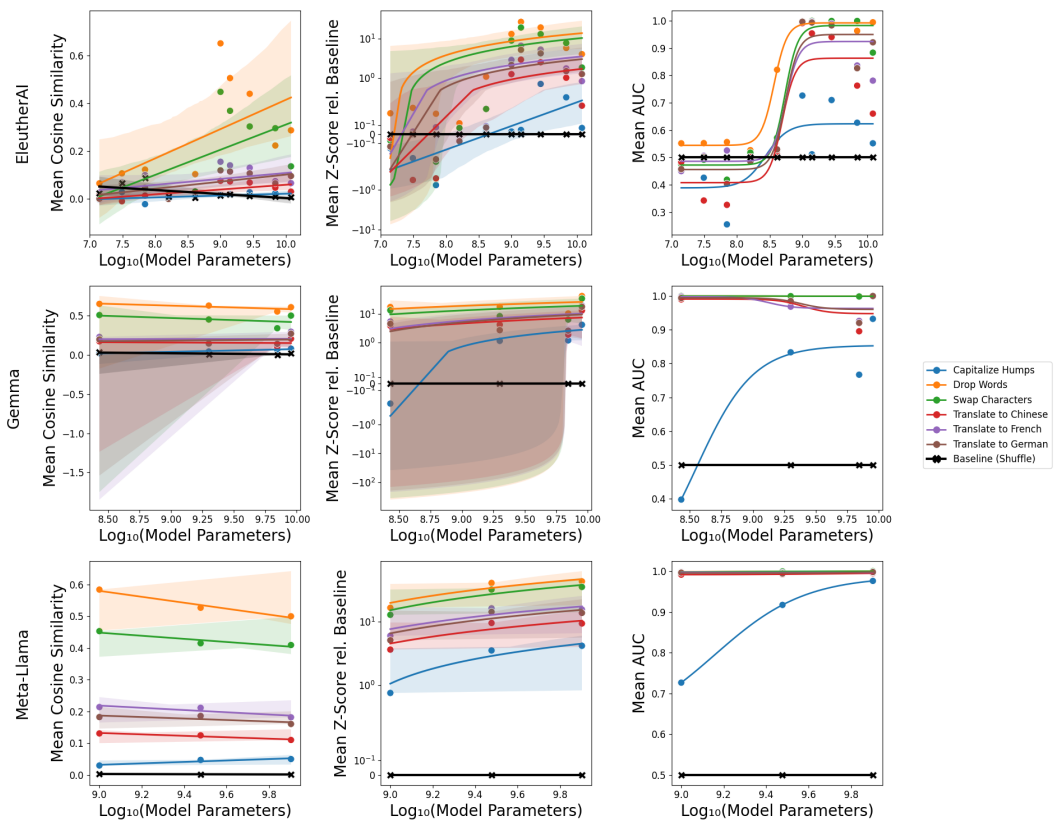


Figure 8: **Semantic-preserving transformations yield more aligned gradients for larger/stronger models:** We display the same data as in Figure 1, by model family.

Table 1: **Robustness to embedding model choice.** Pearson correlations of gradient and embedding cosine similarity for 3,000 document pairs across 25 language models. Columns give correlations for Qwen3-Embedding-0.6B (Qwen3) and EmbeddingGemma-300m (Gemma).

Model	Qwen3	Gemma	Model	Qwen3	Gemma
Pythia 14M	0.5687	0.5696	Gemma 2B	0.7824	0.7731
Pythia 31M	−0.0061	−0.0020	Pythia 2.8B	0.8466	0.8362
Pythia 70M	−0.0343	−0.0388	Llama 3.2 3B	0.7960	0.7833
Pythia 160M	−0.0074	−0.0039	Qwen2.5 3B	0.7862	0.7738
Gemma 3 270M	0.8290	0.8189	Gemma 3 4B	0.8098	0.7971
Pythia 410M	0.4597	0.4569	Pythia 6.9B	0.8346	0.8299
Qwen2.5 0.5B	0.8011	0.7889	Gemma 7B	0.7586	0.7461
Pythia 1B	0.8343	0.8230	Qwen2.5 7B	0.8099	0.7960
Gemma 3 1B	0.8217	0.8096	Pythia 12B	0.7743	0.7716
Llama 3.2 1B	0.8181	0.8062	Gemma 3 12B	0.8114	0.7989
Pythia 1.4B	0.8407	0.8311	Qwen2.5 14B	0.8001	0.7864
Qwen2.5 1.5B	0.7987	0.7866	GPT-OSS 20B	0.2856	0.2842
Gemma 2 2B	0.8271	0.8148			

bedding models, although the pooled correlations do not establish that semantic duplication harms training or validate naturally occurring nearest-neighbor pairs from the full corpus.

C Limitations and Future Work

This work has several limitations. First, because we are unable to train multi-billion-parameter models on our compute budget, we simulated the impact of semantic duplicates using exact duplicates. We drew appropriate comparisons by basing estimates on mean cosine similarity of semantic embeddings, but behavior may differ for large models trained on semantic rather than exact duplicates. Future work could validate our findings on large models by training on semantically deduplicated datasets. The formation of such datasets requires future research attention and resources.

Our corpus-scale analysis uses EmbeddingGemma-300m embeddings. Although this is a state-of-the-art embedding model used by many frontier labs today for data exploration, it still does not produce perfectly isotropic or representative embeddings. Training better embedding models that utilize the latent space more efficiently could improve confidence in the corpus-scale results.

D Algorithm for Gradient Comparison

The algorithm computes gradient similarities and separability metrics defined in Section 2.

Algorithm 1 Gradient similarity for semantic duplicates with negative baseline distribution

Require: Base texts $\{x_i\}_{i=1}^{N}$; transformation set $\mathcal{T}$; models $\{f^{(k)}\}_{k=1}^{K}$ with checkpoints $\{\theta_{k,s}\}$; loss ℓ; token budget T; number of negative pairing rounds R
Ensure: For each (k, s, τ): positives $\{s_i^+\}$, baseline negatives $\mathcal{S}^-$, AUC, Z-score summary

Truncate each x_i to at most T tokens (model tokenizer)
for $k \leftarrow 1$ **to** K **do**
  for all checkpoints s of model k **do**
    Load parameters $\theta \leftarrow \theta_{k,s}$
    ▷ Compute gradients for all base texts once
    for $i \leftarrow 1$ **to** N **do**
      $g_i \leftarrow \nabla_\theta \, \ell(x_i; \theta)$
    end for
    ▷ Build a negative baseline distribution from many random pairings
    $\mathcal{S}^- \leftarrow [\,]$
    for $r \leftarrow 1$ **to** R **do**
      Sample pairing map $j_r(\cdot)$ such that $j_r(i) \neq i$ for all i
      for $i \leftarrow 1$ **to** N **do**
        $\mathcal{S}^-.\mathrm{append}\big(\cos(g_i, g_{j_r(i)})\big)$
      end for
    end for
    Compute μ^- and σ^- from $\mathcal{S}^-$
    ▷ Evaluate each transformation on all texts
    for all $\tau \in \mathcal{T}$ **do**
      for $i \leftarrow 1$ **to** N **do**
        $x_i^{(\tau)} \leftarrow \tau(x_i)$
        $g_i^{(\tau)} \leftarrow \nabla_\theta \, \ell(x_i^{(\tau)}; \theta)$
        $s_i^+ \leftarrow \cos(g_i, g_i^{(\tau)})$
        $z_i \leftarrow (s_i^+ - \mu^-)/\sigma^-$
      end for
      Compute AUC using positives $\{s_i^+\}_{i=1}^{N}$ vs negatives $\mathcal{S}^-$
      Summarize Z-scores (e.g., mean/median over i)
    end for
  end for
end for

E Deriving The Partner Probability and the Effective Latent Count Approximation

This appendix derives Eq. (27) starting from an exact expression for a general latent mixture and giving a controlled approximation that yields the compact form $1 - \exp(-(N_{\text{meas}} - 1)/K_{\text{eff}})$.

Setup (Latent-Mixture Model). Let Z be a discrete semantic latent taking values in an index set $\mathcal{Z}$ with mixture weights $\{w_z\}_{z\in\mathcal{Z}}$, i.e. $\mathbb{P}(Z = z) = w_z$ and $\sum_z w_z = 1$. Let $Z_1, \dots, Z_N \overset{iid}{\sim} \{w_z\}$ denote the latents of N independent draws (e.g. $N = N_{\text{meas}}$ samples from the training stream).

For draw i, define the event that it has at least one same-latent partner among the other $N - 1$ draws:

$$A_i \coloneqq \{\exists j \neq i : \ Z_j = Z_i\}.$$

By exchangeability, $\mathbb{P}(A_i)$ does not depend on i, so we analyze A_1.

E.1 Exact Expression for q_N

Define

$$q_N \coloneqq \mathbb{P}(A_1) = \mathbb{P}\big(\exists j \neq 1 : \ Z_j = Z_1\big).$$

Condition on $Z_1 = z$. Then each of the remaining $N - 1$ draws matches z with probability w_z, independently, so the number of matches among draws $2, \dots, N$ is $\text{Binomial}(N - 1, w_z)$. Hence,

$$\mathbb{P}(\text{no partner for draw } 1 \mid Z_1 = z) = \mathbb{P}(Z_2 \neq z, \dots, Z_N \neq z \mid Z_1 = z) = (1 - w_z)^{N-1}. \quad (29)$$

Averaging over Z_1 gives the exact identity

$$\mathbb{P}(\text{no partner for draw } 1) = \sum_{z\in\mathcal{Z}} \mathbb{P}(Z_1 = z)\,(1 - w_z)^{N-1} = \sum_z w_z(1 - w_z)^{N-1}. \quad (30)$$

Therefore,

$$\boxed{q_N = 1 - \sum_z w_z(1 - w_z)^{N-1}.} \quad (31)$$

This is the first line of Eq. (27) and is *exact* for any discrete mixture.

E.2 From Mixture Weights to K_{eff}

A key quantity is the probability that *two independent draws* share the same latent:

$$p_{\text{lat}} \coloneqq \mathbb{P}(Z = Z') = \sum_z \mathbb{P}(Z = z)\mathbb{P}(Z' = z) = \sum_z w_z^2. \quad (32)$$

This is the Simpson collision probability. It induces the *Simpson effective number of latents*

$$\boxed{K_{\text{eff}} \coloneqq \frac{1}{p_{\text{lat}}} = \frac{1}{\sum_z w_z^2}.} \quad (33)$$

In the uniform-K case ($w_z = 1/K$ for $z = 1, \dots, K$), we have $p_{\text{lat}} = 1/K$ and thus $K_{\text{eff}} = K$.

E.3 Approximation: Rare-Collision / No-Heavy-Modes Regime

We now explain the approximation

$$q_N \approx 1 - \exp\left(-(N - 1)\sum_z w_z^2\right) = 1 - \exp\left(-\frac{N - 1}{K_{\text{eff}}}\right).$$

Step 1: Poissonizing the Binomial for Small w_z. For small w_z, the binomial $\mathrm{Binomial}(N-1, w_z)$ is well-approximated by $\mathrm{Poisson}(\lambda_z)$ with rate $\lambda_z = (N-1)w_z$. In particular,

$$(1-w_z)^{N-1} = \exp\big((N-1)\log(1-w_z)\big) = \exp\big(-(N-1)w_z + O\big((N-1)w_z^2\big)\big), \tag{34}$$

so when $\max_z w_z \ll 1$ and $(N-1)\max_z w_z^2$ is not too large, we may use

$$(1-w_z)^{N-1} \approx \exp\big(-(N-1)w_z\big). \tag{35}$$

Plugging (35) into (31) yields

$$q_N \approx 1 - \sum_z w_z \exp\big(-(N-1)w_z\big). \tag{36}$$

Step 2: Collapsing the Mixture to a Single Effective rate. Let W be the random variable $W := w_{Z_1}$ when $Z_1 \sim \{w_z\}$, i.e. $\mathbb{P}(W = w_z) = w_z$. Then (36) can be written compactly as

$$\sum_z w_z \exp\big(-(N-1)w_z\big) = \mathbb{E}\Big[e^{-(N-1)W}\Big]. \tag{37}$$

Moreover,

$$\mathbb{E}[W] = \sum_z w_z \cdot w_z = \sum_z w_z^2 = p_{\mathrm{lat}} = \frac{1}{K_{\mathrm{eff}}}.$$

If the mixture has *no heavy modes* (informally: $w_z \ll 1$ and the distribution of W is not extremely spread out), we can approximate the expectation in (37) by its mean-field form:

$$\mathbb{E}\Big[e^{-(N-1)W}\Big] \approx \exp\big(-(N-1)\mathbb{E}[W]\big) = \exp\left(-(N-1)\sum_z w_z^2\right) = \exp\left(-\frac{N-1}{K_{\mathrm{eff}}}\right). \tag{38}$$

A standard way to justify (38) is via a cumulant (Taylor) expansion:

$$\log \mathbb{E}[e^{-aW}] = -a\,\mathbb{E}[W] + \frac{a^2}{2}\mathrm{Var}(W) + O\big(a^3\mathbb{E}[|W - \mathbb{E}W|^3]\big), \qquad a := N-1,$$

so if $a^2\mathrm{Var}(W)$ is small compared to $a\mathbb{E}[W]$ (i.e. W is concentrated around its mean at the scale relevant for a), then $\log \mathbb{E}[e^{-aW}] \approx -a\mathbb{E}[W]$ and (38) follows.

Putting the Steps Together. Combining (36) with (38) yields

$$\boxed{q_N \approx 1 - \exp\left(-(N-1)\sum_z w_z^2\right) = 1 - \exp\left(-\frac{N-1}{K_{\mathrm{eff}}}\right).} \tag{39}$$

This matches Eq. (27).

The uniform-pool specialization and the heavy-mode limitations are stated in Section 5.4.

F A First-Principles Model of Duplicate-Limited Scaling via Hutter-Style Learning Curves

This appendix derives a collision-aware scaling correction by combining: (i) a Hutter-style learning-curve model in which performance improves as a power law in the number of *independent* training signals, and (ii) a reduction of independent signal due to duplicates/semantic collisions. The goal is not a fully realistic theory of language modeling, but a minimal mechanism that explains why a plane law of the form $\Delta(C, K) \approx a\, C^{\beta} K^{-\gamma}$ arises naturally.

Step 1: A Hutter-Style "New Information" Learning Curve. A classic abstraction (learning curve theory) models learning progress as driven by discovering previously unseen "features" or "types" in a heavy-tailed environment. Concretely, let z denote a latent "type" (semantic class, rule, or pattern) with weights $\{w_z\}$. Consider the idealized memorization learner that, upon seeing *one* example of type z, can thereafter predict z perfectly, while unseen types incur a fixed excess loss. In this model, the expected excess risk after n iid draws is proportional to the probability mass of unseen types:

$$\epsilon(n) = \sum_z w_z\,(1-w_z)^n, \tag{40}$$

a form that appears in learning-curve theory and is closely related to occupancy/species discovery. (For a detailed treatment of conditions under which heavy tails yield power laws, see Hutter (2021).)

Step 2: Power Laws from Heavy Tails. If the type weights follow a Zipf/regularly varying tail, $\epsilon(n)$ follows a power law:

$$\epsilon(n) \;\propto\; n^{-\alpha} \quad \text{for some } \alpha \in (0,1), \tag{41}$$

with α determined by the tail index of $\{w_z\}$ (see Hutter (2021)). We use (41) to motivate a power law dependence of excess loss on the amount of *independent* training signal.

Step 3: Duplicates Reduce the Effective Number of Independent Signals. In our setting, training examples are not independent sources of new information: duplicates (exact or semantic) induce correlated gradients and therefore reduce the number of effectively independent update directions. Let n denote the number of training documents processed. Let K denote the number of effective semantic classes available (or K_{eff} in the main text). Let $\rho \in [0,1]$ summarize semantic sensitivity (gradient alignment within a class) as in Eq. (13). Under the correlation model in Eq. (15), Proposition 1 implies an effective sample size

$$n_{\text{eff}} = \frac{n}{1+\rho\,\frac{n-1}{K}} \;\approx\; \frac{n}{1+r_{\text{eff}}}, \qquad r_{\text{eff}} := \rho\,\frac{n}{K}. \tag{42}$$

Intuitively, r_{eff} is an *effective reuse ratio*: when $r_{\text{eff}} \ll 1$ the stream is mostly novel, and when $r_{\text{eff}} \gg 1$ the stream is dominated by redundant semantics.

Step 4: Substitute n_{eff} into the Learning Curve. Assume the excess loss (or excess cross-entropy) is a power law in the *independent* signal count:

$$L(n,K) - L_\star \;\approx\; B\,n_{\text{eff}}^{-\alpha}, \tag{43}$$

where $L_\star$ is an irreducible floor and $B > 0$. For the high-uniqueness baseline (negligible collisions), $n_{\text{eff}} \approx n$ and $L_\infty(n) - L_\star \approx Bn^{-\alpha}$. For finite K, combining (42)–(43) gives

$$L(n,K) - L_\star \;\approx\; B\,n^{-\alpha}(1+r_{\text{eff}})^{\alpha}. \tag{44}$$

Step 5: A Duplicate-Induced Degradation Law. Define the normalized degradation Δ as in Eq. (18): $\Delta := (L(n,K) - L_\infty(n))/L_\infty(n)$. Using (44) and $L_\infty(n) = L_\star + Bn^{-\alpha}$, we obtain

$$\Delta(n,K) \;\approx\; \frac{Bn^{-\alpha}\big((1+r_{\text{eff}})^{\alpha}-1\big)}{L_\star + Bn^{-\alpha}}. \tag{45}$$

In the regime where $Bn^{-\alpha}$ is not negligible relative to $L_\star$ (typical for the losses in our controlled ladders), the prefactor is slowly varying and (45) is well-approximated by a power law in r_{eff}. In particular, when $r_{\text{eff}} \lesssim 1$ we can linearize:

$$\Delta(n,K) \;\approx\; \tilde{\lambda}\,r_{\text{eff}} \;=\; \tilde{\lambda}\,\rho\,\frac{n}{K}, \tag{46}$$

where $\tilde{\lambda}$ absorbs the slowly varying ratio in (45). Equation (46) recovers the main-text intuition that degradation is (approximately) proportional to an effective reuse ratio.

Step 6: Translating to Compute and the Plane Law. Let C denote compute. Over restricted ranges, it is empirically accurate to approximate

$$n(C) \propto C^{u}, \qquad \rho(C) \propto C^{v},$$

as in Eq. (21). Substituting into (46) yields

$$\Delta(C,K) \;\approx\; a\,C^{u+v}\,K^{-1}, \tag{47}$$

which is a *plane law* in $(\log C, \log K)$ with $\gamma \approx 1$. More generally, if one does not linearize (45), the same substitution yields a plane $\Delta(C,K) \propto C^{\beta}K^{-\gamma}$ with $\beta = \eta(u+v)$ and $\gamma = \eta$ for some effective exponent η, matching Eq. (22).

Discussion: Why $\rho(C)$ Should Grow with Scale (and why a Power Law is a Reasonable Local Model). The parameter $\rho(C)$ captures the fraction of gradient energy explained by invariances to surface form (Eq. (13)). A growing body of theory and empirical work on *hierarchical/compositional* data suggests that neural networks learn coarse, high-level structure before finer structure, and that deeper invariances require more data/compute. For example, the random hierarchy model (RHM) formalizes language-like hierarchical generation and yields staged learning dynamics where progressively deeper variables become learnable as sample size increases (Cagnetta et al., 2024; Cagnetta & Wyart, 2024). Separately, work on formal-language recognition by transformers highlights a connection between model depth/recurrence and the ability to represent hierarchical (context-free) structure, which is a canonical form of compositional invariance (Bhattamishra et al., 2020b; Jerad et al., 2026). Taken together, these results motivate modeling $\rho(C)$ as monotone increasing with scale; over the narrow compute ranges used in scaling ladders, a power law approximation $\rho(C) \propto C^v$ is a parsimonious local model.

A Reduced-Parameter Variant. Equation (47) suggests a two-degree-of-freedom correction: fix $\gamma = 1$ and fit only (a, β) (or even fit v with u known from the compute-to-sample mapping). In our controlled ladders, the fitted γ is close to 1, consistent with this linear-reuse regime.

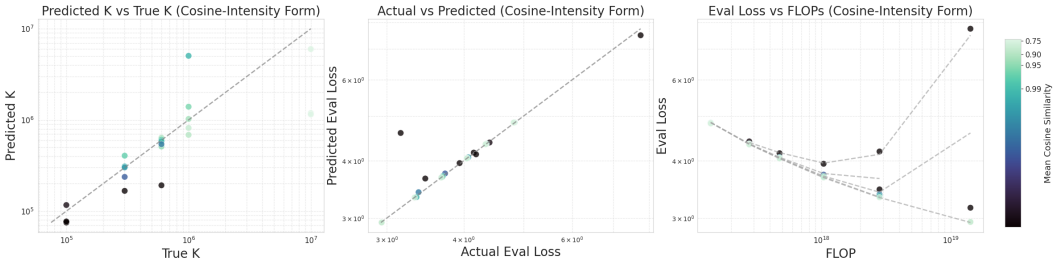


Figure 9: **Predictions of eval loss using cosine intensity-based estimation of $\widehat{K}_{\text{eff}}$.**

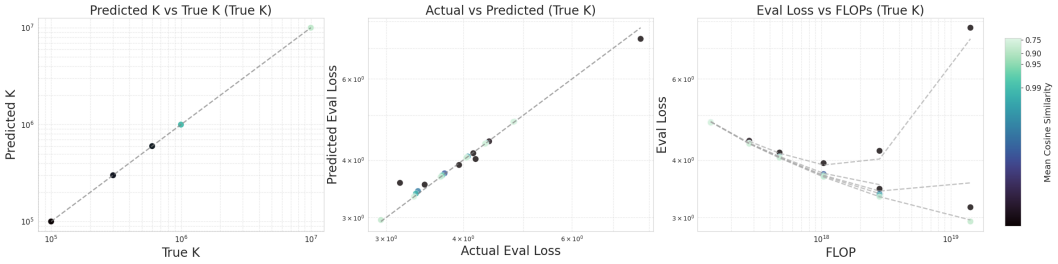


Figure 10: **Predictions of eval loss using true K.**

G Semantic Sensitivity and Gradient Alignment

Under mild assumptions that the surface/idiosyncratic term ξ_x is approximately isotropic and independent across different surface forms of the same latent z, $\rho(s)$ controls expected gradient cosine similarity. Concretely, for semantic duplicates $x = \mathcal{G}(z, \tau)$ and $x' = \mathcal{G}(z, \tau')$ with the same z, the numerator satisfies $\mathbb{E}\langle g(x) - \mu,\, g(x') - \mu\rangle \approx \mathbb{E}\|\delta_z\|_2^2$, while the denominator is $\mathbb{E}\|g(x) - \mu\|_2^2 \approx \mathbb{E}\|\delta_z\|_2^2 + \mathbb{E}\|\xi_x\|_2^2$, yielding the approximation

$$\mathbb{E}\big[\mathrm{sim}_{\theta(s)}(x, x') \mid z\big] \;\approx\; \rho(s), \tag{48}$$

$$\mathbb{E}\big[\mathrm{sim}_{\theta(s)}(x, \tilde{x})\big] \;\approx\; 0 \text{ for unrelated } \tilde{x}. \tag{49}$$

H Estimating Effective Uniqueness from Mean NN Cosine

Setup: Cosine is Measured on a Fixed Embedding Subsample of the Stream. For each training run we take a subsample of N_{meas} training documents from the run's data stream (including repeats), embed each document with a fixed embedding model, and unit-normalize to obtain vectors $v_t \in \mathbb{S}^{d-1}$. We then compute the nearest-neighbor cosine for each embedded sample as defined in the main text. All quantities below refer to this fixed measurement size N_{meas}. In our controlled ladder, N_{meas} is constant across runs. Crucially, $\overline{M}_N$ is computed on a fixed-size subsample of the *training stream*, including repeats, not on the unknown pool of unique documents.

Step 1: Background NN Similarity without Collisions. Let $m_0(N)$ denote the expected mean NN cosine when the nearest neighbor is *not* a semantic collision (i.e., no same-latent partner appears among the $N - 1$ other samples). We estimate $m_0(N)$ from a high-uniqueness reference stream (largest-K pool or without-replacement stream), with negligible exact repeats, using the same embedding pipeline.

Step 2: A Two-Component Model for $\overline{M}_N$. We model each M_t as either: (i) a background neighbor with mean $m_0(N)$, or (ii) a collision neighbor (same latent) with typical similarity $m_+ \in (m_0(N), 1]$. Let q_N be the probability that a given sample has at least one collision neighbor among the other $N - 1$ samples. Equation (25) gives the estimator in Equation (26). In our controlled experiment where collisions correspond to *exact repeats*, we take $m_+ = 1$ (up to numerical precision). For semantic (non-exact) collisions, $m_+ < 1$ can be calibrated using known semantic-duplicate pairs (e.g. translations).

Step 3: Invert $\widehat{q}_N$ into an Effective Latent Count K_{eff}. Let z be the semantic latent with mixture weights $\{w_z\}$ and collision probability $p_{\text{lat}} = \mathbb{P}(z = z') = \sum_z w_z^2$. Define the effective number of latents (Simpson effective size)

$$K_{\text{eff}} \coloneqq \frac{1}{p_{\text{lat}}} = \frac{1}{\sum_z w_z^2}. \tag{50}$$

For a latent mixture with weights $\{w_z\}$, the probability that a given draw has at least one same-latent partner among the other $N_{\text{meas}} - 1$ draws is given by Equation (27), where the approximation uses the small-weight and concentration conditions of Appendix E; in the uniform-K case it uses the approximation $\log(1 - x) \approx -x$. Inverting gives the estimator in Equation (28).

Step 4: A K-free Restored Scaling Law. Our restored degradation model is

$$\Delta(C, K) \coloneqq \frac{L(C, K) - L_\infty(C)}{L_\infty(C)} \approx a\, C^\beta K^{-\gamma}.$$

Replacing K by $\widehat{K}_{\text{eff}}$ gives a correction depending only on observable stream geometry $\Delta(C) \approx aC^\beta\, \widehat{K}_{\text{eff}}^{-\gamma}$, with predicted loss

$$L_{\text{pred}}(C) = L_\infty(C)\big(1 + \Delta(C)\big). \tag{51}$$