Scale Dependent Data Duplication

Anonymous Authors * [1]

Abstract

Data duplication during pretraining can degrade generalization and lead to memorization, motivating aggressive deduplication pipelines. However, at web scale, it is unclear what constitutes a "duplicate": beyond surface-form matches, semantically equivalent documents (e.g. translations) may induce redundant training signals once models become sufficiently capable. Practically, this means that semantic duplicates operate increasingly like exact duplicates during training. 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, by contrast, produce gradients that reflect surface similarity (e.g., shared tokens) rather than semantic similarity. Second, we embedded all 192 million FineWeb-Edu-Dedup documents using EmbeddingGemma-300m. For moderate corpus sizes, the cosine similarity between nearest-neighbors follows an isotropic power law baseline. However, as corpus size grows to hundreds of billions of tokens, the nearest-neighbor similarities deviate sharply, indicating accelerated semantic collisions. Finally, controlled pretraining on data sampled with replacement from pools of finite unique documents shows that limited uniqueness yields mild degradation for small models, but rapidly increasing loss penalties for larger models, breaking naive scaling extrapolation. In Appendix A.3, we derive scaling corrections that allow practitioners to estimate deviation from expected scaling due to limited semantic uniqueness of the pretraining corpus. Our results identify an understudied source of scale-dependence and provide a path toward more accurate prediction at scale.

1. Introduction

Modern language models scale by increasing parameters, compute, and training tokens. For example, Llama 1 (Touvron et al., 2023) trained on $\sim$1T tokens, while the Llama4 herd (Adcock et al., 2026) trained on up to 40T tokens. At these scales, even small fractions of duplicated data can materially reduce the number of 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., 2025; 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 previously unknown source of scale dependence: as models become more capable, *semantic duplicates* induce the same gradients during training. In tandem, capable models are trained on larger corpora, in which the number of semantic collisions rapidly increases. Together, these effects create a recipe for model degradation.

Contributions:

1. We quantify the **emergence of semantic sensitivity** during training by measuring cosine similarity between per-document cross-entropy gradients across a suite of models and semantic-preserving transformations. We find that in more capable models, semantic duplicates induce similar gradients during training.
2. We study **semantic collisions** by embedding 192M documents from FineWeb-Edu-Dedup (Penedo et al., 2024) documents and analyzing nearest-neighbor (NN) statistics across dataset scales from 10^4 to 10^8 documents. We discover that power laws governing scaling for moderate corpus sizes break down for large corpora. This collapse of scaling laws occurs earlier for synthetic corpora, revealing lower semantic diversity.
3. We examine the consequences for predictability by training scaling ladders on streams sampled with replacement from finite pools of K unique documents, showing that limited uniqueness breaks naive scaling extrapolation. In Appendix A.3, we derive scaling corrections that explicitly quantify the effects of limited uniqueness and restore predictability. Furthermore, we

show how to estimate an effective K directly from mean nearest-neighbor cosine similarity.

We defer related work to Appendix B.

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. Consequently, if two documents are semantic duplicates (e.g., translations), then a sufficiently capable model will update its parameters in similar directions when trained on both documents. Practically, this means that semantic duplicates operate increasingly like exact duplicates during 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 variance due to length, each text is truncated to at most $T = 2000$ tokens using the tokenizer of the model under evaluation.

We compute the per-document full-parameter gradient

$$g(x;\theta) = \nabla_\theta \ell(x_i;\theta), \tag{1}$$

where ℓ is the mean next-token cross-entropy:

$$\ell(x;\theta) = \frac{1}{|x|}\sum_{u=1}^{|x|} \mathrm{CE}(f_\theta(x)_u, x_{u+1})\,. \tag{2}$$

To establish a null baseline, we sample unrelated English documents $(x_i, x_j), \quad i \neq j$ and compute cosine similarity

$$\mathrm{sim}(x_i, x_j) = \frac{\langle g(x_i;\theta), g(x_j;\theta)\rangle}{\|g(x_i;\theta)\|_2\,\|g(x_j;\theta)\|_2}. \tag{3}$$

We repeat this across many random pairings to estimate the baseline mean μ^- and standard deviation σ^-.

Transformations. We construct a set of transformations $\mathcal{T} = \{\tau_1, \ldots, \tau_L\}$ intended to preserve semantic content while perturbing surface form:

- Swap Characters: with probability 0.05, randomly replace each ascii character with another.
- Drop Words: Randomly delete each word with probability 0.05.
- Capitalize Humps: Capitalize every other character.
- Translate to Chinese/French/German.

For translations, we use Google's Translate API (Google).

For each document x_i and transformation τ, we compute

$$s_i^+(\tau) := \mathrm{sim}_\theta(x_i,\ \tau(x_i))\,. \tag{4}$$

Separability Metrics (Z-scores and AUC). To summarize separation between positives and negatives, we define:

$$z(\tau) := \frac{\mu^+(\tau) - \mu^-}{\sigma^-}, \tag{5}$$

$$\mu^+(\tau) := \frac{1}{N}\sum_{i=1}^{N} s_i^+(\tau), \tag{6}$$

$$\mu^- := \frac{1}{|\mathcal{S}^-|}\sum_{(i,j)\in\mathcal{S}^-} \mathrm{sim}_\theta(x_i,\ x_j)\,, \tag{7}$$

$$(\sigma^-)^2 := \mathrm{Var}_{(i,j)\in\mathcal{S}^-}\left[\mathrm{sim}_\theta(x_i,\ x_j)\right]. \tag{8}$$

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

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, positive similarities for several transformations are comparable to or below the negative baseline, indicating that gradient direction is dominated by superficial features (e.g., language identity or capitalization). As model capability increases, transformed counterparts become consistently more aligned than unrelated pairs.

To quantify separability, we compute $z(\tau)$ in Eq. (5) and AUC for the binary task described above. 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 if a model encodes meaning robustly, two semantically equivalent documents generate aligned weight updates. This allows the *same dataset* to have a smaller effective size for more capable models.

3. Semantic Collisions

When training models compute-optimally, corpus size grows in tandem with the number of parameters and model capabilities. In this section, 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 increasing exponentially.

Collision metrics. For a set of unit-normalized embeddings $\{v_i\}_{i=1}^{N}$, define nearest-neighbor (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 (i) estimates of $\mathbb{E}[M_i]$ as a function of N, and (ii) tail probabilities $\mathbb{P}(M_i \geq T)$ for fixed thresholds T.

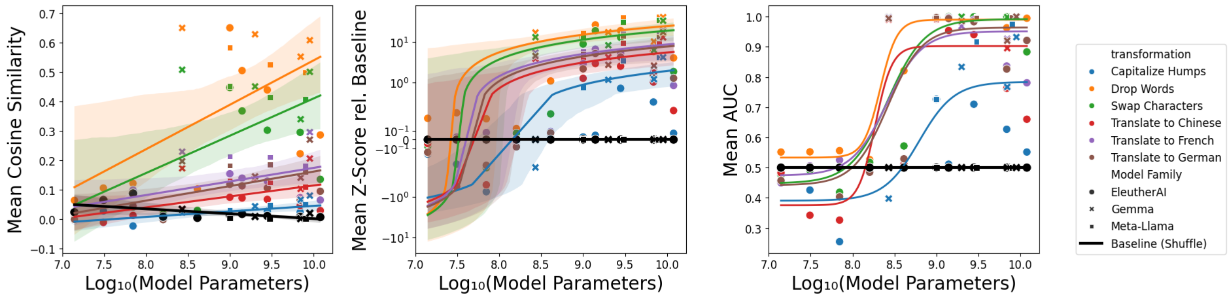

Figure 1. **Semantic-preserving transformations yield more aligned gradients for larger/stronger models.** We sample N=1000 FineWeb-Edu-Dedup documents and compute per-document gradients of normalized next-token cross-entropy (Eq. 2) for each model. We report mean cosine similarity between (i) unrelated document pairs (negative baseline) and (ii) each document and its transformed counterpart (positives), including translations and light surface perturbations. Smaller/weaker models exhibit gradient similarity dominated by surface cues (language/casing), often failing to separate positives from negatives. As capability increases, positives become consistently more aligned than the negative baseline. Error bars show per-document standard deviation. Per-model-family results are in Figure 5.

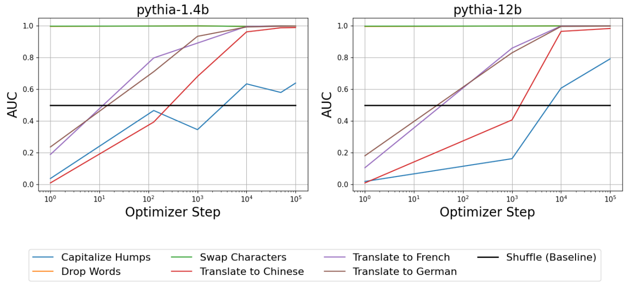

Figure 2. **Semantic sensitivity emerges over training and is accelerated by scale.** For a fixed model family, we compute AUC to detect whether a candidate gradient corresponds to a semantic-preserving transformation of the same document versus an unrelated document, with cosine similarity to the original 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.

3.1. Experimental Setup

We embed 190M texts from FineWeb-Edu-Dedup (Penedo et al., 2024) using EmbeddingGemma-300m (Vera et al., 2025). EmbeddingGemma-300m is a Matryoshka Representation Learning (Kusupati et al., 2024) model that produces embeddings of four nested sizes (768, 512, 256, and 128); sub-embeddings are obtained by slicing and re-normalizing.

We sample subsets of embeddings with cardinality ranging from 10^4 to 10^8 and estimate NN cosine similarities within each pool using FAISS (Douze et al., 2024).

3.2. Results

Figures 3 and 6 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 expected. Beyond a scale threshold, the mean cosine gap decreases *faster than any fitted power law calibrated on smaller N*. In log-linear coordinates, the decrease in NN cosine similarity is approximately linear over document corpus sizes less than 1M, before decaying much more quickly for corpora with over 10M documents. This presents a potentially 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 log-linear scaling laws would predict. Thus, models for which semantic duplicates are recognizable also experience far more of these duplicates, which could lead to loss of predictable scaling.

A Note on Synthetic Data: Recently, synthetic data has become a popular supplement for real data during pretraining and continued pretraining (Mishra et al., 2022; Chen et al., 2024a; Yang et al., 2024; Kang et al., 2025; Qin et al., 2025), though questions remain about whether it has sufficient diversity to provide a future alternative for real data. 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 10).

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), which provides a pessimistic, worst-case proxy for repeated training signals.

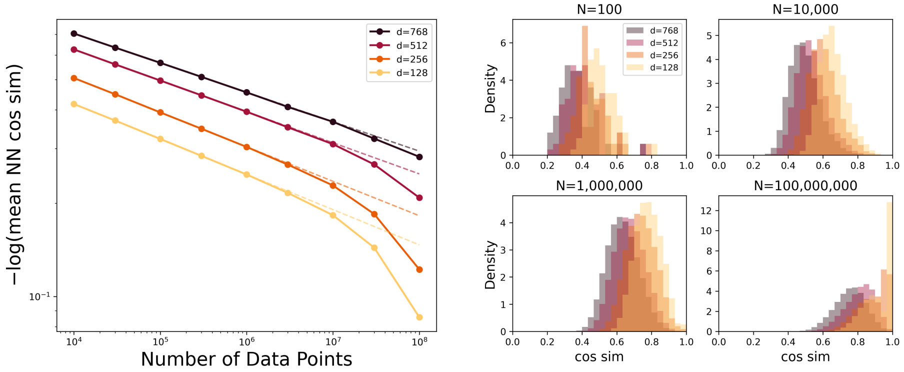

Figure 3. **NN cosine similarity scaling deviates sharply at large corpus sizes.** We embed 190M FineWeb-Edu-Dedup documents with EmbeddingGemma-300m and sample subsets of size ranging from 10^4-10^8 without replacement. For each N, we estimate the mean nearest-neighbor cosine similarity using FAISS. 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 substantially more near neighbors than expected under isotropic baselines.

4.1. Experimental Setup

We sample pools of unique data of size K ranging from 10^5 through 10^8 unique documents sampled from FineWeb-Edu-Dedup. 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.

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 7 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 trained under the same K constraint.

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:

$$\mathrm{FracInc}(K) := \frac{L(K) - L(\infty)}{L(\infty)}. \tag{9}$$

This poses a challenge for scaling prediction, since realistic settings lack an infinite-unique-data baseline. Appendix A.3 develops a scaling correction that addresses this problem and restores predictable scaling in the presence of semantic duplicates.

5. 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, in Appendix A.3, derive scaling corrections that account for semantic diversity in the dataset and restore predictable scaling.

Our experiments have profound implications for the future of language models. Until now, industry convention has been to bet trillions of dollars on the success of 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 10, 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.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Acknowledgements

RS acknowledges support from Stanford Data Science and from the OpenAI Superalignment Fast Grant. SK acknowledges support from NSF 2046795 and 2205329, IES R305C240046, ARPA-H, the MacArthur Foundation, Schmidt Sciences, HAI, OpenAI, Microsoft, and Google.

We express gratitude for helpful discussions with William Meng and first-rate copy editing from Stephanie Schneider.

References

Adcock, A., Srivastava, A., Dubey, A., Jauhri, A., Pande, A., Pandey, A., Sharma, A., Kadian, A., Kumawat, A., Kelsey, A., Stelle, A., Cheema, A., Kabiljo, A., Katz, A., Gangidi, A., Tayade, A., Victoria, A., Alastuey, A. S., Conrath, A., Mohiuddin, A., Sharif, A., Siddiqui, A., Goldstand, A., Li, A., Boyd, A., Daliri, A. K., Iqbal, A., Menon, A., Mathews, A., Mathur, A., Agarwal, A., Schelten, A., Shine, A., Muñoz, A. C., Guliaev, A., Radovic, A., Song, A., Vaughan, A., Simeonov, A., Rezende, A., Rezende, A., Baevski, A., Roubaud, A., Ma, A., Lee, A., Pereira, A., Ahmed, A., Shankar, A., Kallet, A., Budhiraja, A., Khandekar, A., Benhalloum, A., Gershman, A., Nagpal, A., Zohar, A., Sharaf, A., Desai, A., Razdaibiedina, A., Agape, A., Kurghinyan, A., Perunicic, A., Madotto, A., Darabanov, A., Alvarado, A., Brown, A., Cohen, A., Fang, A., Freeman, A., Gallagher, A., Gu, A., Jo, A. P., Ryan, A., Steffen, A., Wei, A., Rusakov, A., Golovei, A., Shang, A., Fan, A., Fan, A., Flewellen, A., Pathak, A., Goyal, A., Ramchandani, A., Pai, A., Singh, A., Garg, A., Xing, A., Cai, A., Grosul, A., Prochowska, A., Sun, A., Dong, A., Franco, A., Hu, A., Chawla, A., Hartshorn, A., Sheng, A., Thomas, A., Goyal, A., De, A., Bodiwala, A., Bodiwala, A., Yang, A., Saraf, A., Samudra, A., Mun, A., Rahnama, A., Mitra, A., Sravankumar, A., Gupta, A., Haghighi, A., Stolerman, A., Chowdhury, A., Choudhury, A., Korenev, A., Guo, A., Hinsvark, A., Mallya, A., Neelakantan, A., Talebzadeh, A., Shah, A., Shetty, A. J., Bharambe, A., Islam, A., Zhang, A., Gregerson, A., Lewis, A., Ibrahim, A., Minhas, A., Dahan, A., Dabah, A. R., Tang, B., Ulman, B., Sadeghi, B., Jedrzejewski, B., Skarabahaty, B., Zhu, B., Li, B., Bharier, B., Leonhardi, B., Muller, B., Plessala, B., Huang, B., Loyd, B., Paranjape, B., Sheth, B., Bonner, B., Holland, B., Wang, B., Liu, B., Tang, B., Liu, B., Wu, B., Li, B., Yu, B., Chen, B.-C., Araya, B., Vidolov, B., Chen, B., Peng, B., Ni, B., Davis, B., Wasti, B., Adams, B., Taylor, B., Wu, B., Swidler, B., Chiang, B., Clerkin, B., Fuller, B., Cutter, B., Novais, B., Gmyrek, B., Easton, B., Campos, C., Case, C., Fu, C. C., Burton, C., Diaz, C., Cole, C., Liu, C., Fougerat, C., Peng, C., Peng, C., Zhao, C., Wang, C., Kim, C., Shaib, C., Zhou, C., Caucheteux, C., Nguyen, C., Sitawarin, C., Nayak, C., Asher, C., Fan, C., Zhu, C., Cheng, C., Zhang, C., Zhu, C., Ruan, C., Yu, C., Hua, C., Whitehouse, C., Holloway, C., Chu, C.-H., Chuang, C.-Y., Karande, C., Nagpal, C., Bakalar, C., Bi, C., Cai, C., Marra, C., McConnell, C., Thi, C., Tindal, C., Waterson, C., Deverall, C., Fuegen, C., Keller, C., Cheng, C., Jou, C., Smith, C., Wang, C., Feichtenhofer, C., Touret, C., Luc, C., Sauper, C., Zhuge, C., Sung, C.-Y., Tang, C., Wu, C., Siegel, C., Heale, C., Wilbourn, C., White, C., Xia, C., Wong, C., Rat, C., Ferrer, C. C., Habis, C., Nikolaidis, C., Lohachov, D., Ju, D., Flanagan, D., Allonsius, D., Civin, D., Johnson, D., Bolya, D., Francisco, D., Fried, D., Hawthorne, D., Haziza, D., Ho, D., Kreymer, D., Li, D., Machlab, D., McKinnon, D., Obenshain, D., Rodriguez, D., Song, D., Tse, D., Pintz, D., Livshits, D., Rodrigo, D. J., Huynh, D., Askarov, D., Brandfonbrener, D., Esiobu, D., Kant, D., Levin, D., Renardy, D., Soofian, D., Stevens, D., Xu, D., Zhang, D., Shah, D., David, D., Douglas, D., Boyda, D., Raj, D., Hazarika, D., Mekala, D., Choudhary, D., Mahajan, D., Jin, D., Coll-Vinent, D. S., Foss, D., Garcia-Olano, D., Perino, D., Hupkes, D., Su, D., Madathil, D., Govindasamy, D., Yeduguru, D., Vengertsev, D., He, D., Li, D., Wang, D., Li, D., Le, D., Hin, D., Holland, D., Nguyen, D., Nguyen, D., Dowling, E., Litt, E., Lakomkin, E., AlBadawy, E., Ardestani, E. K., Eckstein, E., Dabir, E., Montgomery, E., Lobanova, E., Abramoviz, E., Hedeman, E., Li, E., Hilbert, E., Tan, E. X., Yun, E., Stener, E., Stoimenov, E., Garreau, E., Dinan, E., Hahn, E., Wood, E., Li, E., Ademuwagun, E., Seker, E., Alamillo, E., Gan, E., Han, E., Huang, E., Smith, E. M., Le, E.-T., Chang, E., Helenowski, E., Elnikety, E., Arcaute, E., Myers, E., Nho, E., Poliukhovych, E., Dunbar, E., Litvinenko, E., Altıntaş, E., Hochman, E., Shtrauch, E., Mastenbroek, F., Zeb, F., Ahmad, F., Farahbakhshian, F., Kou, F., Sun, F., Chen, F., Chung, F., Tian, F., Xu, F., Radenovic, F., Kokkinos, F., Barbieri, F., Caggioni, F., Esparza, F., Guzmán, F., Kanayet, F., Seide, F., Zhang, F., Lewis, F., Huang, F., Wang, F., Synnaeve, G., Jacques-Silva, G., Schwarz, G., Ghardhora, G., Elfer, G., Dickson, G., Chaurasia, G., Sewani, G., Shingi, G., Zuo, G., Jeong, G., Puthanpurackal, G., Swee, G., Bertran, G. M.-T., Keren, G., Ling, G., Stasa, G., Saha, G., Safran, G., French, G., Rajendran, G., Thattai, G., Cineas, G., Nail, G., Fletcher, G., Mialon, G., Adams, G., Sizov, G., Pang, G., Elsahar, H., Tran, H. D., Nguyen, H., Wu, H., Inan, H., Eghbalzadeh, H., Fang, H., Zou, H., Doyle, H., Korevaar, H., Wang, H., Werbel, H., Zha, H., Morsy, H., Ma, H., Zhang, H., Sun, H., Wang, H., Shah, H., Habeeb, H., Rudolph, H., Gupta, H., Poddar, H., Parikh, H., Zhang, H., Wang, H., Li, H., Sharma, H., Nguyen, H. P., Zhang, H., Qiu, H., Lv, H., Xu, H., Zhan, H., Hamooni, H., Huang, H., Xu, H., Laurençon, H., Touvron, H., Dinh, H., Goldman, H., Mehanna, H., Nguyen, H., Tsuo, H., Graves, I., Yu, I., Damlaj, I., Cohen, I., Tufanov, I., Goldenstein, I., Leontiadis, I., Zarov, I., Ahmed, I., Djiofack, I., Spulber, I., Veliche, I.-E., Ramos, I., Misra, I., Gal, I., Evtimov, I., Evtimov, I., Obraztsov, I., Wu, J., Vertino, J. R., Koo, J., Lee, J., Jung, J., Weissman, J., Beldock, J., Crnkovich, J., Grinage, J., Zeng, J. H., Kohli, J., Tian, J., Cahill, J., Geffert, J., Seidel, J., Seidel, J., Tracey, J., Cho, J. H., Wei, J., Kahn, J., Howell, J., Vu, J. L., Park, J., Yan, J., Yip, J., Li, J., Mahadeokar, J., Goluguri, J. B. R., Mehar, J., Gaya, J.-B., Shah, J., Hanson, J., Marcus, J., Walsh, J., Yang, J., van der Linde, J., Fan, J., Chan, J., Zhen, J., Lee, J., Fu, J., Reizenstein, J., Teboul, J., He, J., Zhong, J., Hou, J., Yang, J., Ding, J., Hu, J., Zhu, J., Guo, J., Wang, J., Ouyang, J., Chi, J., Huang, J., Zhao, J., Yang, J., Zhou, J., Zhao, J., Liu, J., Wang, J., You, J., Yu, J., Schwiep, J., Wu, J., Huang, J., Li, J., Koh, J. Y., Zhang, J., Chen, J., Yang, J., Shen, J., Hwang, J., Guo, J., Khatiwada, J., Bitton, J., Li, J., Quanaim, J., Beales, J., Schuijt, J., Chang, J., Quan, J., Chan, J., Shepard, J., Harris, J., Rubin, J., Janzen, J., Kaldor, J., Silva, J. L., Leitao, J., Greer, J., Moon, J., Rocca, J., Tighe, J., Fromm, J., Deng, J., Fernandes, J., Saxe, J., Zheng, J., Pino, J., Prigent, J., Chen, J., Tian, J., Qi, J., Wang, J., Jia, J., Baker, K., Londenberg, K., Wang, K., Peng, K., Peng, K., Yang, K., Alwala, K. V., Yu, K. H., Narang, K., Chadha, K., Sikka, K., Zhang, K., Schuberts, K., Mandyam, K., Sankararaman, K. A., Padthe, K., Prasad, K., Sivakumar, K., Upasani, K., Plawiak, K., Saenko, K., Žmolíková, K., Stadler, K., Matosich, K., Doulgass, K., Hassani, K., Ji, K., Li, K., Heafield, K., Yu, K., Li, K., Ma, K. C.-Y., Hannan, K., Man, K., Chen, K., El-Arini, K., Hutsulyak, K., Nash, K., Jagadeesh, K., Bartelt, K., Topaloglou-Mundy, K., Chatziioannou, K., Karanasos, K., Vougioukas, K., Tsiampouris, K., Hamill, K., Choi, K., Iyer, K., Malik, K., Chiu, K., Huang, K., Bhalla, K., Chawla, K., Li, K., Lakhotia, K., Monk, K., Garg, L., Chourey, L., Hamre, L., Gustafson, L., Deason, L., Rouesnel, L., van der Maaten, L., A, L., Chen, L., Jang, L., Silva, L., Sari, L., Hetherington, L., Zhang, L., Zhao, L., Chen, L., Li, L. C., Yang, L., Zhan, L., Corallo, L., Tan, L., Yu, L., Liu, L., Mor, L., Lin, L., Li, L., Titus, L., Jenkins, L., Madaan, L., Fang, L., Yuan, L., Nava, L., Pasqualin, L., Switzer, L., Fang, L., Sun, L., Tadic, L., Blecher, L., Landzaat, L., Zhang, L., Rao, M., Khabsa, M., Miller, M., Kariya, M., Pasupuleti, M., Luthra, M., Faruqui, M., Avlani, M., Wang, M., Singh, M., Paluri, M., Chakkaravarthy, M., Nair, M., Tiffany, M., Pawlowski, M., Wu, M., Lomeli, M., Consuegra, M., Boiteux, M., Galanis, M. A., Chen, M., Gleize, M., Fazel-Zarandi, M., Hasson, M., Oldham, M., Rita, M., Dordal, M., Setzler, M., Staats, M., Staats, M., Wilde, M., Clark, M., Grange, M., Lennie, M., Schmohl, M., Raphael, M., Naumov, M., Samoylov, M., Lecanu, M., Pavlova, M., Jawaid, M. T. B., Keneally, M., Kambadur, M., Zhang, M., Liu, M., Lin, M., Wang, M., Abraham, M., Liu, M., Au-Yeung, M., Feldergraf, M., Man, M., Matheny, M., Suo, M., Tontchev, M., Meyer, M., Ma, M., Patel, M., Kale, M. S., Vyatskov, M., Alexander, M., Andersland, M., Clark, M., Lewis, M., Li, M., Macey, M., Macey, M., Seltzer, M., Fernandez, M. J., Antonov, M., Plekhanov, M., Zhou, M., Si, M., Qiao, M., Ma, M., Zhang, M., Liang, M., Hermoso, M. J., Suzgun, M., Skarica, M., Singh, M. K., Kabbani, M., Rastegari, M., Sarantakos, M., Sim, M., Gangapuram, M., Moshe, M., Doulaty, M., Metanat, M., Chen, M., Kumar, M., Bansal, M., Ramarao, M., Li, N., Azaria, N., Malik, N., Goyal, N., Balderas, N. V., Wang, N., Kanda, N., Gimelshein, N., Neverova, N., Aclander, N., Sithiviraporn, N., Kumar, N. M., Newton, N., Bahl, N., Ghorbani, N., Patel, N., lee Golan, N., Longenbaugh, N., Egebo, N., Johri, N., Mehta, N., Naik, N., Moritz, N., Bashlykov, N., Bogoychev, N., Laptev, N. P., Chatterji, N., Jones, N., Shah, N., Dong, N., Li, N., Li, N., Zhang, N., Yadav, N., Paz, N., Cheng, N., Cheng, N., Adesanya, O., Repin, O., Maksymets, O., Salpekar, O., Harosh, O., Pednekar, O., Çelebi, O., Gafni, O., Edinger, O., Hanna, O., Mohammed, O. K., Kalinli, O., Tomasello, P., Singh, P., Quevedo, P., Jain, P., Rashidinejad, P., Tooley, P., Parekh, P., Thakkar, P., Taheri, P., Hapuarachchi, P., Kesseli, P., Alrassy, P., de Rezende Pinatti, P., Balaji, P., Sisodiya, P., Moreira, P. J. F., Rittner, P., Valenzuela, P., Sun, P., Zhang, P., Chen, P.-J., Wang, P., Zhang, P., Li, P., Vasic, P., Carras, P., Ney, P., Weng, P., Dumea, P., Hayes, P., Woods, P., Andrews, P., Ménard, P., Wu, P.-H., Liu, P., Dollar, P., Dzhelepov, P., Zvyagina, P., A, P., Agrawal, P., Rajendran, P., Prakash, P., Bhargava, P., Pramono, Shah, P., Dave, P., Jain, P., Dubal, P., Gollakota, P., Krishnan, P., Yuvraj, P., Ghosh, P., Koura, P. S., Xu, P., Qi, Q., Zhou, Q., Guan, Q., Sun, Q., Liu, Q., He, Q., Zheng, Q., Yang, Q., Guo, Q., You, Q., Carbonneaux, Q., Carbonneaux, Q., Duval, Q., Fettes, Q., Alao, R., Batish, R., Guo, R., Rodriguez, R., Bhargava, R., Asuncion, R., Murthy, R., Dutta, R., Jha, R., Kindi, R., Mitra, R., Ganapathy, R., Shah, R., Das, R., Shrivastava, R., Nishtala, R., Shankar, R., Shukhau, R., Calderer, R., Parthasarathy, R., Subramanian, R., Bensadoun, R., Bostan, R., Chaturvedi, R., Agrawal, R., Gao, R., Li, R., Kogen, R., Duran, R. J. P., Cabral, R. S., Lee, R., Pang, R. Y., Bhalodia, R., Mansour, R., Singh, R., Godugu, R., Patney, R., Boyle, R., Goldfarb, R., Caldwell, R., Kuo, R., Raileanu, R., Battey, R., Sharma, R., Sapra, R., Wang, R., Granata, R., Castro, R. D., Paim, R., Maheshwari, R., Varma, R., Girdhar, R., Patel, R., Sumbaly, R., Sheaffer, R., Silva, R., Buchillon, R. R., Hou, R., Xie, R., Mavlyutov, R., Semenov, R., Dinov, R., Bao, R., Fox, R., Kilpatrick, R., Kwan, R., Lim, R., Smith, R., Narayan, S., Qiao, S., Mehta, S., Siby, S., Jain, S., Hosseini, S., Gur-Ari, S., Chennabasappa, S., Geyik, S., Bondu, S. J., Nekkalapudi, S. M. C., Hasan, S., Okabayashi, S., Rambhatla, S., Sawhney, S., Dunster, S., Zhao, S., Keon, S., Azadi, S., Sapra, S., Dooley, S., Datta, S., Parab, S., Xie, S. M., Singh, S., Chen, S., Behn, S., Khodeir, S., Shirazyan, S., Dhillon, S., Pumma, S., Sidorov, S., Adaime, S., Khanna, S., Wani, S., Brenton, S., Bell, S., Kelly, S., Koger, S., Nunley, S., Perry, S., Caicedo, S., Dahlgren, S., Ruder, S., Yamamoto, S., Mehretu, S., Ravi, S. S., Lyu, S., Chellapan, S., Mellos, S., Edunov, S., Royt, S., Cohen, S., Peng, S., Adams, S., Nie, S., Ramaswamy, S., Narang, S., Pisupati, S., Gandham, S., Lim, S., Lindsay, S., Artrip, S., Sheynin, S., Yan, S., Feng, S., Shen, S., Zheng, S., Lin, S., Bi, S., Zha, S. C., Wan, S., Qian, S., Cai, S., Shao, S., Shahidi, S., Li, S., Bernholtz, S., Wang, S., Patil, S. G., Verma, S., P, S. S., Chen, S., Yaida, S., Debnath, S., Siravara, S., Bhosale, S., Ma, S., Zhang, S., Tang, S., Zhang, S., Zhou, S., Che, S., Srinivisan, S., Bhattacharya, S., Patki, S., Chen, S., Chen, S., Vandenhende, S., Merello, S., Wang, S., Barzily, S., Yi, S., Lin, S., Bong, S., Yin, S., Agarwal, S., Agarwal, S., Lieve, S., Sajuyigbe, S., Jiang, S., Li, S., Kim, S., Khosla, S., Maiti, S., Whitman, S., Popuri, S., Tallam, S., Vaidyanathan, S., Vaidyanathan, S., Sootla, S., Collot, S., Ding, S., Chen, S., Cai, S., Gururangan, S., Govindaprasad, S., Young, S., Dewakar, S., Gonugondla, S. K., Bhandari, S., Gumudavelli, S., Gumudavelli, S., Gupta, S., Deng, S., Cho, S., Ganapathy, S., Dhal, S., Fedynak, S., Contrera, S., Kim, S., Rebuffi, S., Chahande, T., Herman, T., Li, T., Xu, T., Fowler, T., Sheasha, T., Anand, T., Kalluri, T., Singh, T., Shavrina, T., Li, T., Rao, T., Patil, T., Li, T., Bui, T., Quach, T., Alharbash, T., Vo, T. V., Kooburat, T., Koehler, T., Georgiou, T., Scialom, T., Ye, T., Li, T., Zhang, T., Li, T., Blankevoort, T., Willi, T., Chou, T., Leung, T., Lee, T., Mihaylov, T., Heatwole, T., Xiao, T., Cao, T., Lee, T., Le, T., Rice, T., Chan, T. K. S., Tran, T., Tiplea, T., Baumgartner, T., Savagaonkar, U., Karn, U., Araiza, U. M., Farooq, U., Cohen, U., Sharif, U., Murarka, U., Phung, V., Joginpalli, V., Saravagi, V., Sharma, V., Viswamurthy, V., Goswami, V., Seth, V., Ramesh, V., Ramesh, V., Gupta, V., Montanez, V., Natarajan, V., Sarma, V., Ramanathan, V., Kerkez, V., Rao, V., Gonguet, V., Mauge, V., Do, V., Vogeti, V., Chaudhary, V., Sankaran, V., Albiero, V., Miglani, V., Pai, V., Cojanu, V., Shubin, V., Mihailescu, V. T., Petrovic, V., Ivanov, V., Vorotilov, V., Bhutada, V., Ng, W. I., Cheng, W., Sun, W., Tu, W., Wei, W., Zhou, W., Hsu, W.-N., Chu, W., Yuan, W., Wang, W., Zhao, W., Jiang, W., Fu, W., Jiang, W., Meers, W., Constable, W., Wang, W., Wong, W. R., Martinet, X., Lin, X. V., Yan, X., Yin, X., Li, X., Rui, X., Yang, X., Tang, X., Wang, X., Wang, X., Wang, X., Dai, X., Peng, X., Li, X., Meng, X., Zhang, X., Xia, X., Jin, X., xinbo Gao, Xie, X., Zhou, X., Ma, X., Ju, X., Zhao, X., Liu, X., Jia, X., Zhang, X., Cao, X., Wang, X., Wu, X., Xu, X., Ma, X., Wang, X., Cui, Y., Chen, Y., Li, Y., Shu, Y., Xia, Y., Chen, Y., Zhou, Y., Mehta, Y., Patel, Y., Tekena, Y., Gaur, Y., Babaei, Y., Zhou, Y., Hu, Y., Qi, Y., Lee, Y., Wen, Y., Liu, Y.-C., Wu, Y. B., Pan, Y., Yang, Y., Lin, Y.-H., Wang, Y., Wu, Y., Yang, Y., Huang, Y., Aharon, Y. B., Yang, Y., You, Y., Xu, Y., Zhang, Y., Yuan, Y., Liu, Y., Ma, Y., Yang, Y., Lu, Y., Komornik, Y., Lin, Y., Goyhman, Y., Mamo, Y. M., Nam, Y., Wang, Y., Lu, Y., Zhao, Y., Hsieh, Y.-H., Lo, Y.-J., Tian, Y., Zhang, Y., Xiong, Y., Yao, Y., Hao, Y., Zhang, Y., Li, Y., Cao, Y., Yu, Y., Zhao, Y., Guo, Y., Wang, Y., Huang, Y., Lu, Y., Shi, Y., Wang, Y., He, Y., Wang, Y., Qian, Y., Wang, Y., Tang, Y., Mao, Y., Li, Y., Dai, Y., Hulovatyy, Y., Hu, Y., Sun, Y., Rait, Z., Wentz, Z., Coudert, Z. D., Collins, Z., Hankir, Z., He, Z., Ahmed, Z., Ahmed, Z., RosnBrick, Z., Shu, Z., Rohalska, Z., Wen, Z., Liu, Z., Liu, Z., Qiao, Z., Xu, Z., Zhou, Z., Chen, Z., Tang, Z., Wu, Z., Ouyang, Z., Lei, Z., Hong, Z., Xiu, Z., Zhao, Z., Meng, Z., Jin, Z., Zeng, Z., Liu, Z., Meng, Z., Qiao, Z., Zheng, Z., Qi, Z., Luo, Z., Birkhead, Z. F., Sun, Z., and Achdut, Z. The llama 4 herd: Architecture, training, evaluation, and deployment notes, 2026. URL `https://arxiv.org/abs/2601.11659`.

Aljaafari, N., Carvalho, D. S., and Freitas, A. Trace for tracking the emergence of semantic representations in transformers, 2025. URL `https://arxiv.org/abs/2505.17998`.

Bhattamishra, S., Patel, A., and Goyal, N. On the ability and limitations of transformers to recognize formal languages. *arXiv preprint arXiv:2009.11264*, 2020.

Broder, A. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, pp. 21–29, 1997. doi: 10.1109/SEQUEN.1997.666900.

Cagnetta, F. and Wyart, M. Towards a theory of how the structure of language is acquired by deep neural networks. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (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`.

Cagnetta, F., Cornacchia, F., and Wyart, M. Towards a theory of how the structure of language is acquired by deep neural networks. *arXiv preprint arXiv:2406.00048*, 2024a.

Cagnetta, F., Petrini, L., Tomasini, U. M., Favero, A., and Wyart, M. How deep neural networks learn compositional data: The random hierarchy model. *Physical Review X*, 14 (3), July 2024b. ISSN 2160-3308. doi: 10.1103/physrevx.14.031001. URL `http://dx.doi.org/10.1103/PhysRevX.14.031001`.

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

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., and Raffel, C. Extracting training data from large language models, 2021. URL `https://arxiv.org/abs/2012.07805`.

Chen, H., Waheed, A., Li, X., Wang, Y., Wang, J., Raj, B., and Abdin, M. I. On the diversity of synthetic data and its impact on training large language models, 2024a. URL https://arxiv.org/abs/2410.15226.

Chen, H., Yang, X., Zhu, J., and Wang, W. Quantifying semantic emergence in language models, 2024b. URL https://arxiv.org/abs/2405.12617.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., Marris, L., Petulla, S., Gaffney, C., Aharoni, A., Lintz, N., Pais, T. C., Jacobsson, H., Szpektor, I., Jiang, N.-J., Haridasan, K., Omran, A., Saunshi, N., Bahri, D., Mishra, G., Chu, E., Boyd, T., Hekman, B., Parisi, A., Zhang, C., Kawintiranon, K., Bedrax-Weiss, T., Wang, O., Xu, Y., Purkiss, O., Mendlovic, U., Deutel, I., Nguyen, N., Langley, A., Korn, F., Rossazza, L., Ramé, A., Waghmare, S., Miller, H., Byrd, N., Sheshan, A., Hadsell, R., Bhardwaj, S., Janus, P., Rissa, T., Horgan, D., Abdagic, A., Belenki, L., Allingham, J., Singh, A., Guidroz, T., Srinivasan, S., Schmit, H., Chiafullo, K., Elisseeff, A., Jha, N., Kolhar, P., Berrada, L., Ding, F., Si, X., Mallick, S. B., Och, F., Erell, S., Ni, E., Latkar, T., Yang, S., Sirkovic, P., Feng, Z., Leland, R., Hornung, R., Wu, G., Blundell, C., Alvari, H., Huang, P.-S., Yip, C., Deur, S., Liu, L., Surita, G., Duque, P., Damen, D., Jia, J., Guez, A., Mircea, M., Sinha, A., Magni, A., Stradomski, P., Marian, T., Galić, V., Chen, W., Husain, H., Singhal, A., Grewe, D., Aubet, F.-X., Song, S., Blanco, L., Rechis, L., Ho, L., Munoz, R., Zheng, K., Hamrick, J., Mather, K., Taitelbaum, H., Rutherford, E., Lei, Y., Chen, K., Shukla, A., Moreira, E., Doi, E., Isik, B., Shabat, N., Rogozińska, D., Kolipaka, K., Chang, J., Vušak, E., Venkatachary, S., Noghabi, S., Bharti, T., Jun, Y., Zaks, A., Green, S., Challagundla, J., Wong, W., Mohammad, M., Hirsch, D., Cheng, Y., Naim, I., Proleev, L., Vincent, D., Singh, A., Krikun, M., Krishnan, D., Ghahramani, Z., Atias, A., Aggarwal, R., Kirov, C., Vytiniotis, D., Koh, C., Chronopoulou, A., Dogra, P., Ion, V.-D., Tyen, G., Lee, J., Weissenberger, F., Strohman, T., Balakrishna, A., Rae, J., Velic, M., de Liedekerke, R., Elyada, O., Yuan, W., Liu, C., Shani, L., Kishchenko, S., Alessio, B., Li, Y., Song, R., Kwei, S., Jankowski, O., Pappu, A., Namiki, Y., Ma, Y., Tripuraneni, N., Cherry, C., Ikonomidis, M., Ling, Y.-C., Ji, C., Westberg, B., Wright, A., Yu, D., Parkinson, D., Ramaswamy, S., Connor, J., Yeganeh, S. H., Grover, S., Kenwright, G., Litchev, L., Apps, C., Tomala, A., Halim, F., Castro-Ros, A., Li, Z., Boral, A., Sho, P., Yarom, M., Malmi, E., Klinghoffer, D., Lin, R., Ansell, A., S, P. K., Zhao, S., Zuo, S., Santoro, A., Cheng, H.-T., Demmessie, S., Liu, Y., Brichtova, N., Culp, A., Braun, N., Graur, D., Ng, W., Mehta, N., Phillips, A., Sundberg, P., Godbole, V., Liu, F., Katariya, Y., Rim, D., Seyedhosseini, M., Ammirati, S., Valfridsson, J., Malihi, M., Knight, T., Toor, A., Lampe, T., Ittycheriah, A., Chiang, L., Yeung, C., Fréchette, A., Rao, J., Wang, H., Srivastava, H., Zhang, R., Rhodes, R., Brand, A., Weesner, D., Figotin, I., Gimeno, F., Fellinger, R., Marcenac, P., Leal, J., Marcus, E., Cotruta, V., Cabrera, R., Luo, S., Garrette, D., Axelrod, V., Baltateanu, S., Barker, D., Chen, D., Toma, H., Ingram, B., Riesa, J., Kulkarni, C., Zhang, Y., Liu, H., Wang, C., Polacek, M., Wu, W., Hui, K., Reyes, A. N., Su, Y., Barnes, M., Malhi, I., Siddiqui, A., Feng, Q., Damaschin, M., Pighin, D., Steiner, A., Yang, S., Boppana, R. S., Ivanov, S., Kandoor, A., Shah, A., Mujika, A., Huang, D., Choquette-Choo, C. A., Patel, M., Yu, T., Creswell, T., Jerry, Liu, Barros, C., Razeghi, Y., Roy, A., Culliton, P., Xiong, B., Pan, J., Strohmann, T., Powell, T., Seal, B., DeCarlo, D., Shyam, P., Katircioglu, K., Wang, X., Hardin, C., Odisho, I., Broder, J., Chang, O., Nair, A., Shtefan, A., O'Brien, M., Agarwal, M., Potluri, S., Goyal, S., Jhindal, A., Thakur, S., Stuken, Y., Lyon, J., Toutanova, K., Feng, F., Wu, A., Horn, B., Wang, A., Cullum, A., Taubman, G., Shrivastava, D., Shi, C., Tomlinson, H., Patel, R., Tu, T., Oflazer, A. M., Pongetti, F., Yang, M., Taïga, A. A., Perot, V., Pierse, N. W., Han, F., Drori, Y., Iturrate, I., Chakrabarti, A., Yeung, L., Dopson, D., ting Chen, Y., Kulshreshtha, A., Guo, T., Pham, P., Schuster, T., Chen, J., Polozov, A., Xing, J., Zhou, H., Kacham, P., Kukliansky, D., Miech, A., Yaroshenko, S., Chi, E., Douglas, S., Fei, H., Blondel, M., Myla, P., Madmoni, L., Wu, X., Keysers, D., Kjems, K., Albuquerque, I., Yu, L., D'sa, J., Plantan, M., Ionescu, V., Elias, J. S., Gupta, A., Vuyyuru, M. R., Alcober, F., Zhou, T., Ji, K., Hartmann, F., Puttagunta, S., Song, H., Amid, E., Stefanoiu, A., Lee, A., Pucciarelli, P., Wang, E., Raul, A., Petrov, S., Tian, I., Anklin, V., Nti, N., Gomes, V., Schumacher, M., Vesom, G., Panagopoulos, A., Bousmalis, K., Andor, D., Jacob, J., Zhang, Y., Rosgen, B., Kecman, M., Tung, M., Belias, A., Goodman, N., Covington, P., Wieder, B., Saxena, N., Davoodi, E., Huang, M., Maddineni, S., Roulet, V., Campbell-Ajala, F., Sessa, P. G., Xintian, Wu, Lai, G., Collins, P., Haig, A., Sakenas, V., Xu, X., Giustina, M., Shafey, L. E., Charoenpanit, P., Garg, S., Ainslie, J., Severson, B., Arenas, M. G., Pathak, S., Rajayogam, S., Feng, J., Bakker, M., Li, S., Wichers, N., Rogers, J., Geng, X., Li, Y., Jagerman, R., Jia, C., Olmert, N., Sharon, D., Mauger, M., Mariserla, S., Ma, H., Mohabey, M., Kim, K., Andreev, A., Pollom, S., Love, J., Jain, V., Agrawal, P., Schroecker, Y., Fortin, A., Warmuth, M., Liu, J., Leach, A., Blok, I., Girirajan, G. P., Aharoni, R., Uria, B., Sozanschi, A., Goldberg, D., Ionita, L., Ribeiro, M. T., Zlocha, M., Birodkar, V., Lachgar, S., Yuan, L., Choudhury, H., Ginsberg, M., Zheng, F., Dibb, G., Graves, E., Lokhande, S., Rasskin, G., Muraru, G.-C., Quick, C., Tata, S., Sermanet, P., Chawla, A., Karo, I., Wang, Y., Zhang, S., Keller, O., Dragan, A., Su, G., Chou, I., Liu, X., Tao, Y., Prabhakara, S., Wilson, M., Liu, R., Wang, S., Evans, G., Du, D., Castaño, A., Prasad, G., Mahdy, M. E., Gerlach, S., Reid, M., Kahn, J., Zait, A., Pillai, T. S., Ulrich, T., Wang, G., Wassenberg, J., Farkash, E., Yalasangi, K., Wang, C., Bauza, M., Bucher, S., Liu, T., Yan, J., Leung, G., Sindhwani, V., Barnes, P., Singh, A., Jurin, I., Chang, J., Bhumihar, N. K., Eiger, S., Citovsky, G., Withbroe, B., Li, Z., Xue, S., Santo, N. D., Stoyanov, G., Raimond, Y., Zheng, S., Gao, Y., Listík, V., Kwasiborski, S., Saputro, R., Ozturel, A., Mallya, G., Majmundar, K., West, R., Caron, P., Wei, J., Castrejon, L., Vikram, S., Ramachandran, D., Dhawan, N., Park, J., Smoot, S., van den Driessche, G., Blau, Y., Malik, C., Liang, W., Hirsch, R., dos Santos, C. N., Weinstein, E., van den Oord, A., Lall, S., FitzGerald, N., Jiang, Z., Yang, X., Webster, D., Elqursh, A., Pope, A., Rotival, G., Raposo, D., Zhu, W., Dean, J., Alabed, S., Tran, D., Gupta, A., Gleicher, Z., Austin, J., Rosseel, E., Umekar, M., Das, D., Sun, Y., Chen, K., Misiunas, K., Zhou, X., Di, Y., Loo, A., Newlan, J., Li, B., Ramasesh, V., Xu, Y., Chen, A., Gandhe, S., Soricut, R., Gupta, N., Hu, S., El-Sayed, S., Garcia, X., Brusilovsky, I., Chen, P.-C., Bolt, A., Huang, L., Gurney, A., Zhang, Z., Pritzel, A., Wilkiewicz, J., Seybold, B., Shamanna, B. K., Fischer, F., Dean, J., Gill, K., Mcilroy, R., Bhowmick, A., Selier, J., Yang, A., Cheng, D., Magay, V., Tan, J., Varma, D., Walder, C., Kocisky, T., Nakashima, R., Natsev, P., Kwong, M., Gog, I., Zhang, C., Dieleman, S., Jimma, T., Ryabtsev, A., Brahma, S., Steiner, D., Du, D., Žužul, A., Žanić, M., Raghavachari, M., Gierke, W., Zheng, Z., Petrova, D., Dauphin, Y., Liu, Y., Kessler, I., Hand, S., Duvarney, C., Kim, S., Lee, H., Hussenot, L., Hui, J., Smith, J., Jain, D., Xia, J., Tomar, G. S., Amiri, K., Phan, D., Fuchs, F., Weyand, T., Tomasev, N., Cordell, A., Liu, X., Mallinson, J., Joshi, P., Crawford, A., Suggala, A., Chien, S., Fernando, N., Sanchez-Vargas, M., Williams, D., Crone, P., Luo, X., Karpov, I., Shan, J., Thurk, T., Strudel, R., Voigtlaender, P., Patil, P., Dozat, T., Khodaei, A., Singla, S., Ambroszczyk, P., Wu, Q., Chang, Y., Roark, B., Hegde, C., Ding, T., Filos, A., Wu, Z., Pinto, A. S., Liu, S., Khanna, S., Pandey, A., Mcloughlin, S., Li, Q., Haves, S., Zhou, A., Buchatskaya, E., Leal, I., de Boursac, P., Akazawa, N., Anderson, N., Chen, T., Somandepalli, K., Liang, C., Goenka, S., Winkler, S., Grushetsky, A., Ding, Y., Smith, J., Ye, F., Pont-Tuset, J., Li, E., Li, R., Golany, T., Wegner, D., Jiang, T., Barak, O., Shangguan, Y., Vértes, E., Wong, R., Bornschein, J., Tudor, A., Bevilacqua, M., Schaul, T., Rawat, A. S., Zhao, Y., Axiotis, K., Meng, L., McLean, C., Lai, J., Beattie, J., Kushman, N., Liu, Y., Kutzman, B., Lang, F., Ye, J., Netrapalli, P., Mishra, P., Khan, M., Goel, M., Willoughby, R., Tian, D., Zhuang, H., Chen, J., Tsai, Z., Kementsietsidis, T., Khare, A., Keeling, J., Xu, K., Waters, N., Altché, F., Popat, A., Mittal, B., Saxton, D., Badawy, D. E., Mathieu, M., Zheng, Z., Zhou, H., Ranka, N., Shin, R., Duan, Q., Salimans, T., Mihailescu, I., Shaham, U., Chang, M.-W., Assael, Y., Dikkala, N., Izzard, M., Cohen-Addad, V., Graves, C., Feinberg, V., Chung, G., Strouse, D., Karmon, D., Sharifzadeh, S., Ashwood, Z., Pham, K., Blanton, J., Vasiloff, A., Barber, J., Geller, M., Zhou, A., Zubach, F., Huang, T.-K., Zhang, L., Gupta, H., Young, M., Proskurnia, J., Votel, R., Gabeur, V., Barcik, G., Tripathi, A., Yu, H., Yan, G., Changpinyo, B., Pavetić, F., Coyle, A., Fujii, Y., Mendez, J. G., Zhou, T., Rajamani, H., Hechtman, B., Cao, E., Juan, D.-C., Tan, Y.-X., Dalibard, V., Du, Y., Clay, N., Yao, K., Jia, W., Vijaykumar, D., Zhou, Y., Bai, X., Hung, W.-C., Pecht, S., Todorov, G., Khadke, N., Gupta, P., Lahoti, P., Autef, A., Duddu, K., Lee-Thorp, J., Bykovsky, A., Misiunas, T., Flennerhag, S., Thangaraj, S., McGiffin, J., Nado, Z., Kunesch, M., Noever, A., Hertz, A., Liang, M., Stone, V., Palmer, E., Daruki, S., Pramanik, A., Põder, S., Kyker, A., Khan, M., Sluzhaev, E., Ritter, M., Ruderman, A., Zhou, W., Nagpal, C., Vodrahalli, K., Necula, G., Barham, P., Pavlick, E., Hartford, J., Shafran, I., Zhao, L., Mikuła, M., Eccles, T., Shimokawa, H., Garg, K., Vilnis, L., Chen, H., Shumailov, I., Lee, K.-H., Abdelhamed, A., Xie, M., Cohen, V., Hlavnova, E., Malkin, D., Sitawarin, C., Lottes, J., Coquinot, P., Yu, T., Kumar, S., Zhang, J., Mahendru, A., Ahmed, Z., Martens, J., Chen, T., Boag, A., Peng, D., Devin, C., Klimovskiy, A., Phuong, M., Vainstein, D., Xie, J., Ramabhadran, B., Howard, N., Yu, X., Goswami, G., Cui, J., Shleifer, S., Pinto, M., Yeh, C.-K., Yang, M.-H., Javanmardi, S., Ethier, D., Lee, C., Orbay, J., Kotecha, S., Bromberg, C., Shaw, P., Thornton, J., Rosenthal, A. G., Gu, S., Thomas, M., Gemp, I., Ayyar, A., Ushio, A., Selvan, A., Wee, J., Liu, C., Majzoubi, M., Yu, W., Abernethy, J., Liechty, T., Pan, R., Nguyen, H., Qiong, Hu, Perrin, S., Arora, A., Pitler, E., Wang, W., Shivakumar, K., Prost, F., Limonchik, B., Wang, J., Gao, Y., Cour, T., Buch, S., Gui, H., Ivanova, M., Neubeck, P., Chan, K., Kim, L., Chen, H., Goyal, N., Chung, D.-W., Liu, L., Su, Y., Petrushkina, A., Shen, J., Joulin, A., Xu, Y., Lin, S. X., Kulizhskaya, Y., Chelba, C., Vasudevan, S., Collins, E., Bashlovkina, V., Lu, T., Fritz, D., Park, J., Zhou, Y., Su, C., Tanburn, R., Sushkov, M., Rasquinha, M., Li, J., Prendki, J., Li, Y., LV, P., Sharma, S., Fitoussi, H., Huang, H., Dai, A., Dao, P., Burrows, M., Prior, H., Qin, D., Pundak, G., Sjoesund, L. L., Khurshudov, A., Zhu, Z., Webson, A., Kemp, E., Tan, T., Agrawal, S., Sargsyan, S., Cheng, L., Stephan, J., Kwiatkowski, T., Reid, D., Byravan, A., Michaely, A. H., Heess, N., Zhou, L., Goenka, S., Carpenter, V., Levskaya, A., Wang, B., Roberts, R., Leblond, R., Chikkerur, S., Ginzburg, S., Chang, M., Riachi, R., Chuqiao, Xu, Borsos, Z., Pliskin, M., Pawar, J., Lustman, M., Kirkwood, H., Anand, A., Chaudhary, A., Kalb, N., Milan, K., Augenstein, S., Goldie, A., Prince, L., Raman, K., Sun, Y., Xia, V., Cohen, A., Huo, Z., Camp, J., Ellis, S., Zilka, L., Torres, D. V., Patel, L., Arora, S., Chan, B., Adler, J., Ayoub, K., Liang, J., Jamil, F., Jiang, J., Baumgartner, S., Sun, H., Karov, Y., Akulov, Y., Zheng, H., Cai, I., Fantacci, C., Rubin, J., Acha, A. R., Wang, M., D'Souza, N., Sathyanarayana, R., Dai, S., Rowe, S., Simanovsky, A., Goldman, O., Kuang, Y., Pan, X., Rosenberg, A., Rojas-Esponda, T., Dutta, P., Zeng, A., Jurenka, I., Farquhar, G., Bansal, Y., Iqbal, S., Roelofs, B., Joung, G.-Y., Beak, P., Ryu, C., Poplin, R., Wu, Y., Alayrac, J.-B., Buthpitiya, S., Ronneberger, O., Habtegebriel, C., Li, W., Cavallaro, P., Wei, A., Bensky, G., Denk, T., Ganapathy, H., Stanway, J., Joshi, P., Bertolini, F., Lo, J., Ma, O., Charles, Z., Sampemane, G., Sahni, H., Chen, X., Askham, H., Gaddy, D., Young, P., Tan, J., Eyal, M., Bražinskas, A., Zhong, L., Wu, Z., Epstein, M., Bailey, K., Hard, A., Lee, K., Goldshtein, S., Ruiz, A., Badawi, M., Lochbrunner, M., Kearns, J., Brown, A., Pardo, F., Weber, T., Yang, H., Jiang, P.-P., Akin, B., Fu, Z., Wainwright, M., Zou, C., Gaba, M., Manzagol, P.-A., Kan, W., Song, Y., Zainullina, K., Lin, R., Ko, J., Deshmukh, S., Jindal, A., Svensson, J., Tyam, D., Zhao, H., Kaeser-Chen, C., Baird, S., Moradi, P., Hall, J., Guo, Q., Tsang, V., Liang, B., Pereira, F., Ganesh, S., Korotkov, I., Adamek, J., Thiagarajan, S., Tran, V., Chen, C., Tar, C., Jain, S., Dasgupta, I., Bilal, T., Reitter, D., Zhao, K., Vezzani, G., Gehman, Y., Mehta, P., Beltrone, L., Dotiwalla, X., Guadarrama, S., Abbas, Z., Karp, S., Georgiev, P., Ferng, C.-S., Brockschmidt, M., Peng, L., Hirnschall, C., Verma, V., Bi, Y., Xiao, Y., Dabush, A., Xu, K., Wallis, P., Parker, R., Wang, Q., Xu, Y., Safarli, I., Tewari, D., Zhang, Y., Kim, S., Gesmundo, A., Thomas, M., Levi, S., Chowdhury, A., Rao, K., Garst, P., Conway-Rahman, S., Ran, H., McKinney, K., Xiao, Z., Yu, W., Agrawal, R., Stjerngren, A., Ionescu, C., Chen, J., Sharma, V., Chiu, J., Liu, F., Franko, K., Sanford, C., Cai, X., Michel, P., Ganapathy, S., Labanowski, J., Garrett, Z., Vargas, B., Sun, S., Gale, B., Buschmann, T., Desjardins, G., Ghelani, N., Jain, P., Verma, M., Asawaroengchai, C., Eisenschlos, J., Harlalka, J., Kazawa, H., Metzler, D., Howland, J., Jian, Y., Ades, J., Shah, V., Gangwani, T., Lee, S., Ring, R., Hernandez, S. M., Reich, D., Sinha, A., Sathe, A., Kovac, J., Gill, A., Kannan, A., D'olimpio, A., Sevenich, M., Whang, J., Kim, B., Sim, K. C., Chen, J., Zhang, J., Lall, S., Matias, Y., Jia, B., Friesen, A., Nasso, S., Thapliyal, A., Perozzi, B., Yu, T., Shekhawat, A., Huda, S., Grabowski, P., Wang, E., Sreevatsa, A., Dib, H., Hassen, M., Schuh, P., Milutinovic, V., Welty, C., Quinn, M., Shah, A., Wang, B., Barth-Maron, G., Frye, J., Axelsson, N., Zhu, T., Ma, Y., Giannoumis, I., Sedghi, H., Ye, C., Luan, Y., Aydin, K., Chandra, B., Sampathkumar, V., Huang, R., Lavrenko, V., Eleryan, A., Hong, Z., Hansen, S., Carthy, S. M., Samanta, B., Ćevid, D., Wang, X., Li, F., Voznesensky, M., Hoffman, M., Terzis, A., Sehwag, V., Fidel, G., He, L., Cai, M., He, Y., Feng, A., Nikoltchev, M., Phatale, S., Chase, J., Lawton, R., Zhang, M., Ouyang, T., Tragut, M., Manshadi, M. H., Narayanan, A., Shen, J., Gao, X., Bolukbasi, T., Roy, N., Li, X., Golovin, D., Panait, L., Qin, Z., Han, G., Anthony, T., Kudugunta, S., Patraucean, V., Ray, A., Chen, X., Yang, X., Bhatia, T., Talluri, P., Morris, A., Ražnatović, A., Brownfield, B., An, J., Peng, S., Kane, P., Zheng, C., Duduta, N., Kessinger, J., Noraky, J., Liu, S., Rong, K., Veličković, P., Rush, K., Goldin, A., Wei, F., Garlapati, S. M. R., Pantofaru, C., Kwon, O., Ni, J., Noland, E., Trapani, J. D., Beaufays, F., Roy, A. G., Chow, Y., Turker, A., Cideron, G., Mei, L., Clark, J., Dou, Q., Bošnjak, M., Leith, R., Du, Y., Yazdanbakhsh, A., Nasr, M., Kwak, C., Sheth, S. S., Kaskasoli, A., Anand, A., Lakshminarayanan, B., Jerome, S., Bieber, D., Chu, C.-T., Senges, A., Shen, T., Sridhar, M., Ndebele, N., Beyret, B., Mohamed, S., Chen, M., Freitag, M., Guo, J., Liu, L., Roit, P., Chen, H., Yan, S., Stone, T., Co-Reyes, J., Cole, J., Scellato, S., Azizi, S., Hashemi, H., Jin, A., Iyer, A., Valentine, M., György, A., Ahuja, A., Diaz, D. H., Lee, C.-Y., Clement, N., Kong, W., Garmon, D., Watts, I., Bhatia, K., Gupta, K., Miecnikowski, M., Vallet, H., Taly, A., Loper, E., Joshi, S., Atwood, J., Chick, J., Collier, M., Iliopoulos, F., Trostle, R., Gunel, B., Leal-Cavazos, R., Hrafnkelsson, A. M., Guzman, M., Ju, X., Forbes, A., Emond, J., Chauhan, K., Caine, B., Xiao, L., Zeng, W., Moufarek, A., Murphy, D., Meng, M., Gupta, N., Riedel, F., Das, A., Lawal, E., Narayan, S., Sosea, T., Swirhun, J., Friso, L., Neyshabur, B., Lu, J., Girgin, S., Wunder, M., Yvinec, E., Pyne, A., Carbune, V., Rijhwani, S., Guo, Y., Doshi, T., Briukhov, A., Bain, M., Hitron, A., Wang, X., Gupta, A., Chen, K., Du, C., Zhang, W., Shah, D., Akula, A., Dylla, M., Kachra, A., Kuo, W., Zou, T., Wang, L., Xu, L., Zhu, J., Snyder, J., Menon, S., Firat, O., Mordatch, I., Yuan, Y., Ponomareva, N., Blevins, R., Moore, L., Wang, W., Chen, P., Scholz, M., Dwornik, A., Lin, J., Li, S., Antognini, D., I, T., Song, X., Miller, M., Kalra, U., Raveret, A., Akerlund, O., Wu, F., Nystrom, A., Godbole, N., Liu, T., DeBalsi, H., Zhao, J., Liu, B., Caciularu, A., Lax, L., Khandelwal, U., Langston, V., Bailey, E., Lattanzi, S., Wang, Y., Kovelamudi, N., Mondal, S., Guruganesh, G., Hua, N., Roval, O., Wesołowski, P., Ingale, R., Halcrow, J., Sohn, T., Angermueller, C., Raad, B., Stickgold, E., Lu, E., Kosik, A., Xie, J., Lillicrap, T., Huang, A., Zhang, L. L., Paulus, D., Farabet, C., Wertheim, A., Wang, B., Joshi, R., ling Ko, C., Wu, Y., Agrawal, S., Lin, L., Sheng, X., Sung, P., Breland-King, T., Butterfield, C., Gawde, S., Singh, S., Zhang, Q., Apte, R., Shetty, S., Hutter, A., Li, T., Salesky, E., Lebron, F., Kanerva, J., Paganini, M., Nguyen, A., Vallu, R., Peter, J.-T., Velury, S., Kao, D., Hoover, J., Bortsova, A., Bishop, C., Jakobovits, S., Agostini, A., Agarwal, A., Liu, C., Kwong, C., Tavakkol, S., Bica, I., Greve, A., GP, A., Marcus, J., Hou, L., Duerig, T., Moroshko, R., Lacey, D., Davis, A., Amelot, J., Wang, G., Kim, F., Strinopoulos, T., Wan, H., Lan, C. L., Krishnan, S., Tang, H., Humphreys, P., Bai, J., Shtacher, I. H., Machado, D., Pang, C., Burke, K., Liu, D., Aravamudhan, R., Song, Y., Hirst, E., Singh, A., Jou, B., Bai, L., Piccinno, F., Fu, C. K., Alazard, R., Meiri, B., Winter, D., Chen, C., Zhang, M., Heitkaemper, J., Lambert, J., Lee, J., Frömmgen, A., Rogulenko, S., Nair, P., Niemczyk, P., Bulyenov, A., Xu, B., Shemtov, H., Zadimoghaddam, M., Toropov, S., Wirth, M., Dai, H., Gollapudi, S., Zheng, D., Kurakin, A., Lee, C., Bullard, K., Serrano, N., Balazevic, I., Li, Y., Schalkwyk, J., Murphy, M., Zhang, M., Sequeira, K., Datta, R., Agrawal, N., Sutton, C., Attaluri, N., Chiang, M., Farhan, W., Thornton, G., Lin, K., Choma, T., Nguyen, H., Dasgupta, K., Robinson, D., Comşa, I., Riley, M., Pillai, A., Mustafa, B., Golan, B., Zandieh, A., Lespiau, J.-B., Porter, B., Ross, D., Rajayogam, S., Agarwal, M., Venugopalan, S., Shahriari, B., Yan, Q., Xu, H., Tobin, T., Dubov, P., Shi, H., Recasens, A., Kovsharov, A., Borgeaud, S., Dery, L., Vasanth, S., Gribovskaya, E., Qiu, L., Mahdieh, M., Skut, W., Nielsen, E., Zheng, C., Yu, A., Bostock, C. G., Gupta, S., Archer, A., Rawles, C., Davies, E., Svyatkovskiy, A., Tsai, T., Halpern, Y., Reisswig, C., Wydrowski, B., Chang, B., Puigcerver, J., Taege, M. H., Li, J., Schnider, E., Li, X., Dena, D., Xu, Y., Telang, U., Shi, T., Zen, H., Kastner, K., Ko, Y., Subramaniam, N., Kumar, A., Blois, P., Dai, Z., Wieting, J., Lu, Y., Zeldes, Y., Xie, T., Hauth, A., Ţifrea, A., Li, Y., El-Husseini, S., Abolafia, D., Zhou, H., Ding, W., Ghalebikesabi, S., Guía, C., Maksai, A., Ágoston Weisz, Arik, S., Sukhanov, N., Świetlik, A., Jia, X., Yu, L., Wang, W., Brand, M., Bloxwich, D., Kirmani, S., Chen, Z., Go, A., Sprechmann, P., Kannen, N., Carin, A., Sandhu, P., Edkins, I., Nooteboom, L., Gupta, J., Maggiore, L., Azizi, J., Pritch, Y., Yin, P., Gupta, M., Tarlow, D., Smith, D., Ivanov, D., Babaeizadeh, M., Goel, A., Kambala, S., Chu, G., Kastelic, M., Liu, M., Soltau, H., Stone, A., Agrawal, S., Kim, M., Soparkar, K., Tadepalli, S., Bunyan, O., Soh, R., Kannan, A., Kim, D., Chen, B. J., Halumi, A., Roy, S., Wang, Y., Sercinoglu, O., Gibson, G., Bhatnagar, S., Sano, M., von Dincklage, D., Ren, Q., Mitrevski, B., Olšák, M., She, J., Doersch, C., Jilei, Wang, Liu, B., Tan, Q., Yakar, T., Warkentin, T., Ramirez, A., Lebsack, C., Dillon, J., Mathews, R., Cobley, T., Wu, Z., Chen, Z., Simon, J., Nath, S., Sainath, T., Bendebury, A., Julian, R., Mankalale, B., Ćurko, D., Zacchello, P., Brown, A. R., Sodhia, K., Howard, H., Caelles, S., Gupta, A., Evans, G., Bulanova, A., Katzen, L., Goldenberg, R., Tsitsulin, A., Stanton, J., Schillings, B., Kovalev, V., Fry, C., Shah, R., Lin, K., Upadhyay, S., Li, C., Radpour, S., Maggioni, M., Xiong, J., Haas, L., Brennan, J., Kamath, A., Savinov, N., Nagrani, A., Yacovone, T., Kappedal, R., Andriopoulos, K., Lao, L., Li, Y., Rozhdestvenskiy, G., Hashimoto, K., Audibert, A., Austin, S., Rodriguez, D., Ruoss, A., Honke, G., Karkhanis, D., Xiong, X., Wei, Q., Huang, J., Leng, Z., Premachandran, V., Bileschi, S., Evangelopoulos, G., Mensink, T., Pavagadhi, J., Teplyashin, D., Chang, P., Xue, L., Tanzer, G., Goldman, S., Patel, K., Li, S., Wiesner, J., Zheng, I., Stewart-Binks, I., Han, J., Li, Z., Luo, L., Lenc, K., Lučić, M., Xue, F., Mullins, R., Guseynov, A., Chang, C.-C., Galatzer-Levy, I., Zhang, A., Bingham, G., Hu, G., Hartman, A., Ma, Y., Griffith, J., Irpan, A., Radebaugh, C., Yue, S., Fan, L., Ungureanu, V., Sorokin, C., Teufel, H., Li, P., Anil, R., Paparas, D., Wang, T., Lin, C.-C., Peng, H., Shum, M., Petrovic, G., Brady, D., Nguyen, R., Macherey, K., Li, Z., Singh, H., Yenugula, M., Iinuma, M., Chen, X., Kopparapu, K., Stern, A., Dave, S., Thekkath, C., Perot, F., Kumar, A., Li, F., Xiao, Y., Bilotti, M., Bateni, M. H., Noble, I., Lee, L., Vázquez-Reina, A., Salazar, J., Yang, X., Wang, B., Gruzewska, E., Rao, A., Raghuram, S., Xu, Z., Ben-David, E., Mei, J., Dalmia, S., Zhang, Z., Liu, Y., Bansal, G., Pankov, H., Schwarcz, S., Burns, A., Chan, C., Sanghai, S., Liang, R., Liang, E., He, A., Stuart, A., Narayanan, A., Zhu, Y., Frank, C., Fatemi, B., Sabne, A., Lang, O., Bhattacharya, I., Settle, S., Wang, M., McMahan, B., Tacchetti, A., Soares, L. B., Hadian, M., Cabi, S., Chung, T., Putikhin, N., Li, G., Chen, J., Tarango, A., Michalewski, H., Kazemi, M., Masoom, H., Sheftel, H., Shivanna, R., Vadali, A., Comanescu, R., Reid, D., Moore, J., Neelakantan, A., Sander, M., Herzig, J., Rosenberg, A., Dehghani, M., Choi, J., Fink, M., Hayes, R., Ge, E., Weng, S., Ho, C.-H., Karro, J., Krishna, K., Thiet, L. N., Skerry-Ryan, A., Eppens, D., Andreetto, M., Sarma, N., Bonacina, S., Ayan, B. K., Nawhal, M., Shan, Z., Dusenberry, M., Thakoor, S., Gubbi, S., Nguyen, D. D., Tsarfaty, R., Albanie, S., Mitrović, J., Gandhi, M., Chen, B.-J., Epasto, A., Stephanov, G., Jin, Y., Gehman, S., Amini, A., Weber, J., Behbahani, F., Xu, S., Allamanis, M., Chen, X., Ott, M., Sha, C., Jastrzebski, M., Qi, H., Greene, D., Wu, X., Toki, A., Vlasic, D., Shapiro, J., Kotikalapudi, R., Shen, Z., Saeki, T., Xie, S., Cassirer, A., Bharadwaj, S., Kiyono, T., Bhojanapalli, S., Rosenfeld, E., Ritter, S., Mao, J., Oliveira, J. G., Egyed, Z., Bandemer, B., Parisotto, E., Kinoshita, K., Pluto, J., Maniatis, P., Li, S., Guo, Y., Ghiasi, G., Tarbouriech, J., Chatterjee, S., Jin, J., Katrina, Xu, Palomaki, J., Arnold, S., Sewak, M., Piccinini, F., Sharma, M., Albrecht, B., Purser-haskell, S., Vaswani, A., Chen, C., Wisniewski, M., Cao, Q., Aslanides, J., Phu, N. M., Sieb, M., Agubuzu, L., Zheng, A., Sohn, D., Selvi, M., Andreassen, A., Subudhi, K., Eruvbetine, P., Woodman, O., Mery, T., Krause, S., Ren, X., Ma, X., Luo, J., Chen, D., Fan, W., Griffiths, H., Schuler, C., Li, A., Zhang, S., Sarr, J.-M., Luo, S., Patana, R., Watson, M., Naboulsi, D., Collins, M., Sidhwani, S., Hoogeboom, E., Silver, S., Caveness, E., Zhao, X., Rodriguez, M., Deines, M., Bai, L., Griffin, P., Tagliasacchi, M., Xue, E., Babbula, S. R., Pang, B., Ding, N., Shen, G., Peake, E., Crocker, R., Raghvendra, S. S., Swisher, D., Han, W., Singh, R., Wu, L., Pchelin, V., Munkhdalai, T., Alon, D., Bacon, G., Robles, E., Bulian, J., Johnson, M., Powell, G., Ferreira, F. T., Li, Y., Benzing, F., Velimirović, M., Soyer, H., Kong, W., Tony, Nguyên, Yang, Z., Liu, J., van Amersfoort, J., Gillick, D., Sun, B., Rauschmayr, N., Zhang, K., Zhan, S., Zhou, T., Frolov, A., Yang, C., Vnukov, D., Rouillard, L., Li, H., Mandhane, A., Fallen, N., Venkataraman, R., Hu, C. H., Brennan, J., Lee, J., Chang, J., Sundermeyer, M., Pan, Z., Ke, R., Tong, S., Fabrikant, A., Bono, W., Gu, J., Foley, R., Mao, Y., Delakis, M., Bhaswar, D., Frostig, R., Li, N., Zipori, A., Hope, C., Kozlova, O., Mishra, S., Djolonga, J., Schiff, C., Merey, M. A., Briakou, E., Morgan, P., Wan, A., Hassidim, A., Skerry-Ryan, R., Sengupta, K., Jasarevic, M., Kallakuri, P., Kunkle, P., Brennan, H., Lieber, T., Mansoor, H., Walker, J., Zhang, B., Xie, A., Žužić, G., Chukwuka, A., Druinsky, A., Cho, D., Yao, R., Naeem, F., Butt, S., Kim, E., Jia, Z., Jordan, M., Lelkes, A., Kurzeja, M., Wang, S., Zhao, J., Over, A., Chakladar, A., Prasetya, M., Jha, N., Ganapathy, S., Cong, Y., Shroff, P., Saroufim, C., Miryoosefi, S., Hammad, M., Nasir, T., Xi, W., Gao, Y., Maeng, Y., Hora, B., Cheng, C.-Y., Haghani, P., Lewenberg, Y., Lu, C., Matysiak, M., Raisinghani, N., Wang, H., Baugher, L., Sukthankar, R., Giang, M., Schultz, J., Fiedel, N., Chen, M., Lee, C.-C., Dey, T., Zheng, H., Paul, S., Smith, C., Ly, A., Wang, Y., Bansal, R., Perz, B., Ricco, S., Blank, S., Keshava, V., Sharma, D., Chow, M., Lad, K., Jalan, K., Osindero, S., Swanson, C., Scott, J., Ilić, A., Li, X., Jonnalagadda, S. R., Soudagar, A. S., Xiong, Y., Batsaikhan, B.-O., Jarrett, D., Kumar, N., Shah, M., Lawlor, M., Waters, A., Graham, M., May, R., Ramos, S., Lefdal, S., Cankara, Z., Cano, N., O'Donoghue, B., Borovik, J., Liu, F., Grimstad, J., Alnahlawi, M., Tsihlas, K., Hudson, T., Grigorev, N., Jia, Y., Huang, T., Igwe, T. P., Lebedev, S., Tang, X., Krivokon, I., Garcia, F., Tan, M., Jia, E., Stys, P., Vashishth, S., Liang, Y., Venkatraman, B., Gu, C., Kementsietsidis, A., Zhu, C., Jung, J., Bai, Y., Hosseini, M. J., Ahmed, F., Gupta, A., Yuan, X., Ashraf, S., Nigam, S., Vasudevan, G., Awasthi, P., Gilady, A. M., Mariet, Z., Eskander, R., Li, H., Hu, H., Garrido, G., Schlattner, P., Zhang, G., Saxena, R., Dević, P., Muralidharan, K., Murthy, A., Zhou, Y., Choi, M., Wongpanich, A., Wang, Z., Shah, P., Xu, Y., Huang, Y., Spencer, S., Chen, A., Cohan, J., Wang, J., Tompson, J., Wu, J., Haroun, R., Li, H., Huergo, B., Yang, F., Yin, T., Wendt, J., Bendersky, M., Chaabouni, R., Snaider, J., Ferret, J., Jindal, A., Thompson, T., Xue, A., Bishop, W., Phal, S. M., Sharma, A., Sung, Y., Radhakrishnan, P., Shomrat, M., Ingle, R., Vij, R., Gilmer, J., Istin, M. D., Sobell, S., Lu, Y., Nottage, E., Sadigh, D., Willcock, J., Zhang, T., Xu, S., Brown, S., Lee, K., Wang, G., Zhu, Y., Tay, Y., Kim, C., Gutierrez, A., Sharma, A., Xian, Y., Seo, S., Cui, C., Pochernina, E., Baetu, C., Jastrzębski, K., Ly, M., Elhawaty, M., Suh, D., Sezener, E., Wang, P., Yuen, N., Tucker, G., Cai, J., Yang, Z., Wang, C., Muzio, A., Qian, H., Yoo, J., Lockhart, D., McKee, K. R., Guo, M., Mehrotra, M., Mendonça, A., Mehta, S. V., Ben, S., Tekur, C., Mu, J., Zhu, M., Krakovna, V., Lee, H., Maschinot, A., Cevey, S., Choe, H., Bai, A., Srinivasan, H., Gasaway, D., Young, N., Siegler, P., Holtmann-Rice, D., Piratla, V., Baumli, K., Yogev, R., Hofer, A., van Hasselt, H., Grant, S., Chervonyi, Y., Silver, D., Hogue, A., Agarwal, A., Wang, K., Singh, P., Flynn, F., Lipschultz, J., David, R., Bellot, L., Yang, Y.-Y., Le, L., Graziano, F., Olszewska, K., Hui, K., Maurya, A., Parotsidis, N., Chen, W., Oguntebi, T., Kelley, J., Baddepudi, A., Mauerer, J., Shaw, G., Siegman, A., Yang, L., Shetty, S., Roy, S., Song, Y., Stokowiec, W., Burnell, R., Savant, O., Busa-Fekete, R., Miao, J., Ghosh, S., MacDermed, L., Lippe, P., Dektiarev, M., Behrman, Z., Mentzer, F., Nguyen, K., Wei, M., Verma, S., Knutsen, C., Dasari, S., Yan, Z., Mitrichev, P., Wang, X., Shejwalkar, V., Austin, J., Sunkara, S., Potti, N., Virin, Y., Wright, C., Liu, G., Riva, O., Pot, E., Kochanski, G., Le, Q., Balasubramaniam, G., Dhar, A., Liao, Y., Bloniarz, A., Shukla, D., Cole, E., Lee, J., Zhang, S., Kafle, S., Vashishtha, S., Mahmoudieh, P., Chen, G., Hoffmann, R., Srinivasan, P., Lago, A. D., Shalom, Y. B., Wang, Z., Elabd, M., Sharma, A., Oh, J., Kothawade, S., Le, M., Monteiro, M., Yang, S., Alarakyia, K., Geirhos, R., Mincu, D., Garnes, H., Kobayashi, H., Mariooryad, S., Krasowiak, K., Zhixin, Lai, Mourad, S., Wang, M., Bu, F., Aharoni, O., Chen, G., Goyal, A., Zubov, V., Bapna, A., Dabir, E., Kothari, N., Lamerigts, K., Cao, N. D., Shar, J., Yew, C., Kulkarni, N., Mahaarachchi, D., Joshi, M., Zhu, Z., Lichtarge, J., Zhou, Y., Muckenhirn, H., Selo, V., Vinyals, O., Chen, P., Brohan, A., Mehta, V., Cogan, S., Wang, R., Geri, T., Ko, W.-J., Chen, W., Viola, F., Shivam, K., Wang, L., Elish, M. C., Popa, R. A., Pereira, S., Liu, J., Koster, R., Kim, D., Zhang, G., Ebrahimi, S., Talukdar, P., Zheng, Y., Poklukar, P., Mikhalap, A., Johnson, D., Vijayakumar, A., Omernick, M., Dibb, M., Dubey, A., Hu, Q., Suman, A., Aggarwal, V., Kornakov, I., Xia, F., Lowe, W., Kolganov, A., Xiao, T., Nikolaev, V., Hemingray, S., Li, B., Iljazi, J., Rybiński, M., Sandhu, B., Lu, P., Luong, T., Jenatton, R., Govindaraj, V., Hui, Li, Dulac-Arnold, G., Park, W., Wang, H., Modi, A., Pouget-Abadie, J., Greller, K., Gupta, R., Berry, R., Ramachandran, P., Xie, J., McCafferty, L., Wang, J., Gupta, K., Lim, H., Bratanič, B., Brock, A., Akolzin, I., Sproch, J., Karliner, D., Kim, D., Goedeckemeyer, A., Shazeer, N., Schmid, C., Calandriello, D., Bhatia, P., Choromanski, K., Montgomery, C., Dua, D., Ramalho, A., King, H., Gao, Y., Nguyen, L., Lindner, D., Pitta, D., Johnson, O., Salama, K., Ardila, D., Han, M., Farnese, E., Odoom, S., Wang, Z., Ding, X., Rink, N., Smith, R., Lehri, H. T., Cohen, E., Vats, N., He, T., Gopavarapu, P., Paszke, A., Patel, M., Gansbeke, W. V., Loher, L., Castro, L., Voitovich, M., von Glehn, T., George, N., Niklaus, S., Eaton-Rosen, Z., Rakićević, N., Jue, E., Perel, S., Zhang, C., Bahat, Y., Pouget, A., Xing, Z., Huot, F., Shenoy, A., Bos, T., Coriou, V., Richter, B., Noy, N., Wang, Y., Ontanon, S., Qin, S., Makarchuk, G., Hassabis, D., Li, Z., Sharma, M., Venkatesan, K., Kemaev, I., Daniel, R., Huang, S., Shah, S., Ponce, O., Warren, Chen, Faruqui, M., Wu, J., Andačić, S., Payrits, S., McDuff, D., Hume, T., Cao, Y., Tessler, M., Wang, Q., Wang, Y., Rendulic, I., Agustsson, E., Johnson, M., Lando, T., Howard, A., Padmanabhan, S. G. S., Daswani, M., Banino, A., Kilgore, M., Heek, J., Ji, Z., Caceres, A., Li, C., Kassner, N., Vlaskin, A., Liu, Z., Grills, A., Hou, Y., Sukkerd, R., Cheon, G., Shetty, N., Markeeva, L., Stanczyk, P., Iyer, T., Gong, Y., Gao, S., Gopalakrishnan, K., Blyth, T., Reynolds, M., Bhoopchand, A., Bilenko, M., Gharibian, D., Zayats, V., Faust, A., Singh, A., Ma, M., Jiao, H., Vijayanarasimhan, S., Aroyo, L., Yadav, V., Chakera, S., Kakarla, A., Meshram, V., Gregor, K., Botea, G., Senter, E., Jia, D., Kovacs, G., Sharma, N., Baur, S., Kang, K., He, Y., Zhuo, L., Kostelac, M., Laish, I., Peng, S., O'Bryan, L., Kasenberg, D., Rao, G. R., Leurent, E., Zhang, B., Stevens, S., Salazar, A., Zhang, Y., Lobov, I., Walker, J., Porter, A., Redshaw, M., Ke, H., Rao, A., Lee, A., Lam, H., Moffitt, M., Kim, J., Qiao, S., Koo, T., Dadashi, R., Song, X., Sundararajan, M., Xu, P., Kawamoto, C., Zhong, Y., Barbu, C., Reddy, A., Verzetti, M., Li, L., Papamakarios, G., Klimczak-Plucińska, H., Cassin, M., Kavukcuoglu, K., Swavely, R., Vaucher, A., Zhao, J., Hemsley, R., Tschannen, M., Ge, H., Menghani, G., Yu, Y., Ha, N., He, W., Wu, X., Song, M., Sterneck, R., Zinke, S., Calian, D. A., Marsden, A., Ruiz, A. C., Hessel, M., Gueta, A., Lee, B., Farris, B., Gupta, M., Li, Y., Saleh, M., Misra, V., Xiao, K., Mendolicchio, P., Buttimore, G., Krayvanova, V., Nayakanti, N., Wiethoff, M., Pande, Y., Mirhoseini, A., Lao, N., Liu, J., Hua, Y., Chen, A., Malkov, Y., Kalashnikov, D., Gupta, S., Audhkhasi, K., Zhai, Y., Kopalle, S., Jain, P., Ofek, E., Meyer, C., Baatarsukh, K., Strejček, H., Qian, J., Freedman, J., Figueira, R., Sokolik, M., Bachem, O., Lin, R., Kharrat, D., Hidey, C., Xu, P., Duan, D., Li, Y., Ersoy, M., Everett, R., Cen, K., Santamaria-Fernandez, R., Taubenfeld, A., Mackinnon, I., Deng, L., Zablotskaia, P., Viswanadha, S., Goel, S., Yates, D., Deng, Y., Choy, P., Chen, M., Sinha, A., Mossin, A., Wang, Y., Szlam, A., Hao, S., Rubenstein, P. K., Toksoz-Exley, M., Aperghis, M., Zhong, Y., Ahn, J., Isard, M., Lacombe, O., Luisier, F., Anastasiou, C., Kalley, Y., Prabhu, U., Dunleavy, E., Bijwadia, S., Mao-Jones, J., Chen, K., Pasumarthi, R., Wood, E., Dostmohamed, A., Hurley, N., Simsa, J., Parrish, A., Pajarskas, M., Harvey, M., Skopek, O., Kochinski, Y., Rey, J., Rieser, V., Zhou, D., Lee, S. J., Acharya, T., Li, G., Jiang, J., Zhang, X., Gipson, B., Mahintorabi, E., Gelmi, M., Khajehnouri, N., Yeh, A., Lee, K., Matthey, L., Baker, L., Pham, T., Fu, H., Pak, A., Gupta, P., Vasconcelos, C., Sadovsky, A., Walker, B., Hsiao, S., Zochbauer, P., Marzoca, A., Velan, N., Zeng, J., Baechler, G., Driess, D., Jain, D., Huang, Y., Tao, L., Maggs, J., Levine, N., Schneider, J., Gemzer, E., Petit, S., Han, S., Fisher, Z., Zelle, D., Biles, C., Ie, E., Fadeeva, A., Liu, C., Franco, J. V., Collister, A., Zhang, H., Wang, R., Zhao, R., Kieliger, L., Shuster, K., Zhu, R., Gong, B., Chan, L., Sun, R., Basu, S., Zimmermann, R., Hayes, J., Bapna, A., Snoek, J., Yang, W., Datta, P., Abdallah, J. A., Kilgour, K., Li, L., Mah, S., Jun, Y., Rivière, M., Karmarkar, A., Spalink, T., Huang, T., Gonzalez, L., Tran, D.-H., Nowak, A., Palowitch, J., Chadwick, M., Talius, E., Mehta, H., Sellam, T., Fränken, P., Nicosia, M., He, K., Kini, A., Amos, D., Basu, S., Jobe, H., Shaw, E., Xu, Q., Evans, C., Ikeda, D., Yan, C., Jin, L., Wang, L., Yadav, S., Labzovsky, I., Sampath, R., Ma, A., Schumann, C., Siddhant, A., Shah, R., Youssef, J., Agarwal, R., Dabney, N., Tonioni, A., Ambar, M., Li, J., Guyon, I., Li, B., Soergel, D., Fang, B., Karadzhov, G., Udrescu, C., Trinh, T., Raunak, V., Noury, S., Guo, D., Gupta, S., Finkelstein, M., Petek, D., Liang, L., Billock, G., Sun, P., Wood, D., Song, Y., Yu, X., Matejovicova, T., Cohen, R., Andra, K., D'Ambrosio, D., Deng, Z., Nallatamby, V., Songhori, E., Dangovski, R., Lampinen, A., Botadra, P., Hillier, A., Cao, J., Baddi, N., Kuncoro, A., Yoshino, T., Bhagatwala, A., Ranzato, M., Schaeffer, R., Liu, T., Ye, S., Sarvana, O., Nham, J., Kuang, C., Gao, I., Baek, J., Mittal, S., Wahid, A., Gergely, A., Ni, B., Feldman, J., Muir, C., Lamblin, P., Macherey, W., Dyer, E., Kilpatrick, L., Campos, V., Bhutani, M., Fort, S., Ahmad, Y., Severyn, A., Chatziprimou, K., Ferludin, O., Dimarco, M., Kusupati, A., Heyward, J., Bahir, D., Villela, K., Millican, K., Marcus, D., Bahargam, S., Unlu, C., Roth, N., Wei, Z., Gopal, S., Ghoshal, D., Lee, E., Lin, S., Lees, J., Lee, D., Hosseini, A., Fan, C., Neel, S., Wu, M., Altun, Y., Cai, H., Piqueras, E., Woodward, J., Bissacco, A., Haykal, S., Bordbar, M., Sundaram, P., Hodkinson, S., Toyama, D., Polovets, G., Myers, A., Sinha, A., Levinboim, T., Krishnakumar, K., Chhaparia, R., Sholokhova, T., Gundavarapu, N. B., Jawahar, G., Qureshi, H., Hu, J., Momchev, N., Rahtz, M., Wu, R., S, A. P., Dhamdhere, K., Guo, M., Gupta, U., Eslami, A., Schain, M., Blokzijl, M., Welling, D., Orr, D., Bolelli, L., Perez-Nieves, N., Sirotenko, M., Prasad, A., Kar, A., Pigem, B. D. B., Terzi, T., Weisz, G., Ghosh, D., Mavalankar, A., Madeka, D., Daugaard, K., Adam, H., Shah, V., Berman, D., Tran, M., Baker, S., Andrejczuk, E., Chole, G., Raboshchuk, G., Mirzazadeh, M., Kagohara, T., Wu, S., Schallhart, C., Orlando, B., Wang, C., Rrustemi, A., Xiong, H., Liu, H., Vezer, A., Ramsden, N., yiin Chang, S., Mudgal, S., Li, Y., Vieillard, N., Hoshen, Y., Ahmad, F., Slone, A., Hua, A., Potikha, N., Rossini, M., Stritar, J., Prakash, S., Wang, Z., Dong, X., Nazari, A., Nehoran, E., Tekelioglu, K., Li, Y., Badola, K., Funkhouser, T., Li, Y., Yerram, V., Ganeshan, R., Formoso, D., Langner, K., Shi, T., Li, H., Yamamori, Y., Panda, A., Saade, A., Scarpati, A. S., Breaux, C., Carey, C., Zhou, Z., Hsieh, C.-J., Bridgers, S., Butryna, A., Gupta, N., Tulsyan, V., Woo, S., Eltyshev, E., Grathwohl, W., Parks, C., Benjamin, S., Panigrahy, R., Dodhia, S., Freitas, D. D., Sauer, C., Song, W., Alet, F., Tolins, J., Paduraru, C., Zhou, X., Albert, B., Zhang, Z., Shu, L., Bansal, M., Nguyen, S., Globerson, A., Xiao, O., Manyika, J., Hennigan, T., Rong, R., Matak, J., Bakalov, A., Sharma, A., Sinopalnikov, D., Pierson, A., Roller, S., Brown, G., Gao, M., Fukuzawa, T., Ghafouri, A., Vassigh, K., Barr, I., Wang, Z., Korsun, A., Jayaram, R., Ren, L., Zaman, T., Khan, S., Lunts, Y., Deutsch, D., Uthus, D., Katz, N., Samsikova, M., Khalifa, A., Sethi, N., Sun, J., Tang, L., Alon, U., Luo, X., Yu, D., Nayyar, A., Petrini, B., Truong, W., Hellendoorn, V., Chinaev, N., Alberti, C., Wang, W., Hu, J., Mirrokni, V., Balashankar, A., Aharon, A., Mehta, A., Iscen, A., Kready, J., Manning, L., Mohananey, A., Chen, Y., Tripathi, A., Wu, A., Petrovski, I., Hwang, D., Baeuml, M., Chandrakaladharan, S., Liu, Y., Coaguila, R., Chen, M., Ma, S., Tafti, P., Tatineni, S., Spitz, T., Ye, J., Vicol, P., Rosca, M., Puigdomènech, A., Yahav, Z., Ghemawat, S., Lin, H., Kirk, P., Nabulsi, Z., Brin, S., Bohnet, B., Caluwaerts, K., Veerubhotla, A. S., Zheng, D., Dai, Z., Petrov, P., Xu, Y., Mehran, R., Xu, Z., Zintgraf, L., Choi, J., Hombaiah, S. A., Thoppilan, R., Reddi, S., Lew, L., Li, L., Webster, K., Sawhney, K., Lamprou, L., Shakeri, S., Lunayach, M., Chen, J., Bagri, S., Salcianu, A., Chen, Y., Donchev, Y., Magister, C., Nørly, S., Rodrigues, V., Izo, T., Noga, H., Zou, J., Köppe, T., Zhou, W., Lee, K., Long, X., Eisenbud, D., Chen, A., Schenck, C., To, C. M., Zhong, P., Taropa, E., Truong, M., Levy, O., Martins, D., Zhang, Z., Semturs, C., Zhang, K., Yakubovich, A., Moreno, P., McConnaughey, L., Lu, D., Redmond, S., Weerts, L., Bitton, Y., Refice, T., Lacasse, N., Conmy, A., Tallec, C., Odell, J., Forbes-Pollard, H., Socala, A., Hoech, J., Kohli, P., Walton, A., Wang, R., Sazanovich, M., Zhu, K., Kapishnikov, A., Galt, R., Denton, M., Murdoch, B., Sikora, C., Mohamed, K., Wei, W., First, U., McConnell, T., Cobo, L. C., Qin, J., Avrahami, T., Balle, D., Watanabe, Y., Louis, A., Kraft, A., Ariafar, S., Gu, Y., Rives, E., Yoon, C., Rusu, A., Cobon-Kerr, J., Hahn, C., Luo, J., Yuvein, Zhu, Ahuja, N., Benenson, R., Kaufman, R. L., Yu, H., Hightower, L., Zhang, J., Ni, D., Hendricks, L. A., Wang, G., Yona, G., Jain, L., Barrio, P., Bhupatiraju, S., Velusamy, S., Dafoe, A., Riedel, S., Thomas, T., Yuan, Z., Bellaiche, M., Panthaplackel, S., Kloboves, K., Jauhari, S., Akbulut, C., Davchev, T., Gladchenko, E., Madras, D., Chuklin, A., Hill, T., Yuan, Q., Madhavan, M., Leonhard, L., Scandinaro, D., Chen, Q., Niu, N., Douillard, A., Damoc, B., Onoe, Y., Pedregosa, F., Bertsch, F., Leichner, C., Pagadora, J., Malmaud, J., Ponda, S., Twigg, A., Duzhyi, O., Shen, J., Wang, M., Garg, R., Chen, J., Evci, U., Lee, J., Liu, L., Kojima, K., Yamaguchi, M., Rajendran, A., Piergiovanni, A., Rajendran, V. K., Fornoni, M., Ibagon, G., Ragan, H., Khan, S. M., Blitzer, J., Bunner, A., Sun, G., Kosakai, T., Lundberg, S., Elue, N., Guu, K., Park, S., Park, J., Narayanaswamy, A., Wu, C., Mudigonda, J., Cohn, T., Mu, H., Kumar, R., Graesser, L., Zhang, Y., Killam, R., Zhuang, V., Giménez, M., Jishi, W. A., Ley-Wild, R., Zhai, A., Osawa, K., Cedillo, D., Liu, J., Upadhyay, M., Sieniek, M., Sharma, R., Paine, T., Angelova, A., Addepalli, S., Parada, C., Majumder, K., Lamp, A., Kumar, S., Deng, X., Myaskovsky, A., Sabolić, T., Dudek, J., York, S., de Chaumont Quitry, F., Nie, J., Cattle, D., Gunjan, A., Piot, B., Khawaja, W., Bang, S., Wang, S., Khodadadeh, S., R, R., Rawlani, P., Powell, R., Lee, K., Griesser, J., Oh, G., Magalhaes, C., Li, Y., Tokumine, S., Vogel, H. N., Hsu, D., BC, A., Jindal, D., Cohen, M., Yang, Z., Yuan, J., de Cesare, D., Bruguier, T., Xu, J., Roy, M., Jacovi, A., Belov, D., Arya, R., Meadowlark, P., Cohen-Ganor, S., Ye, W., Morris-Suzuki, P., Banzal, P., Song, G., Ponnuramu, P., Zhang, F., Scrivener, G., Zaiem, S., Rochman, A. R., Han, K., Ghazi, B., Lee, K., Drath, S., Suo, D., Girgis, A., Shenoy, P., Nguyen, D., Eck, D., Gupta, S., Yan, L., Carreira, J., Gulati, A., Sang, R., Mirylenka, D., Cooney, E., Chou, E., Ling, M., Fan, C., Coleman, B., Tubone, G., Kumar, R., Baldridge, J., Hernandez-Campos, F., Lazaridou, A., Besley, J., Yona, I., Bulut, N., Wellens, Q., Pierigiovanni, A., George, J., Green, R., Han, P., Tao, C., Clark, G., You, C., Abdolmaleki, A., Fu, J., Chen, T., Chaugule, A., Chandorkar, A., Rahman, A., Thompson, W., Koanantakool, P., Bernico, M., Ren, J., Vlasov, A., Vassilvitskii, S., Kula, M., Liang, Y., Kim, D., Huang, Y., Ye, C., Lepikhin, D., and Helmholz, W. 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`.

Domhan, T., Springenberg, J. T., and Hutter, F. Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pp. 3460–3468. AAAI Press, 2015. ISBN 9781577357384.

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., and Jégou, H. The faiss library. 2024.

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

Hahn, M. The computational power of transformers. *arXiv preprint arXiv:2006.09286*, 2020a.

Hahn, M. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, 2020b.

Hernandez, D., Brown, T., Conerly, T., DasSarma, N., Drain, D., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Henighan, T., Hume, T., Johnston, S., Mann, B., Olah, C., Olsson, C., Amodei, D., Joseph, N., Kaplan, J., and McCandlish, S. Scaling laws and interpretability of learning from repeated data, 2022. URL `https://arxiv.org/abs/2205.10487`.

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M. M. A., Yang, Y., and Zhou, Y. Deep learning scaling is predictable, empirically, 2017. URL `https://arxiv.org/abs/1712.00409`.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. Training compute-optimal large language models, 2022. URL `https://arxiv.org/abs/2203.15556`.

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

Ivgi, M., Carmon, Y., and Berant, J. Scaling laws under the microscope: Predicting transformer performance from small scale experiments, 2022. URL `https://arxiv.org/abs/2202.06387`.

Jerad, S., Svete, A., Hao, S., Cotterell, R., and Merrill, W. Context-free recognition with transformers. *arXiv preprint arXiv:2601.01754*, 2026.

Jin, C. and Rinard, M. Emergent representations of program semantics in language models trained on programs. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Kadra, A., Janowski, M., Wistuba, M., and Grabocka, J. Scaling laws for hyperparameter optimization, 2023. URL `https://arxiv.org/abs/2302.00441`.

Kalra, D. S. and Barkeshli, M. Why warmup the learning rate? underlying mechanisms and improvements, 2024. URL `https://arxiv.org/abs/2406.09405`.

Kang, F., Ardalani, N., Kuchnik, M., Emad, Y., Elhoushi, M., Sengupta, S., Li, S.-W., Raghavendra, R., Jia, R., and Wu, C.-J. Demystifying synthetic data in llm pretraining: A systematic study of scaling laws, benefits, and pitfalls, 2025. URL `https://arxiv.org/abs/2510.01631`.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models, 2020. URL `https://arxiv.org/abs/2001.08361`.

Khan, A., Underwood, R., Siebenschuh, C., Babuji, Y., Ajith, A., Hippe, K., Gokdemir, O., Brace, A., Chard, K., and Foster, I. Lshbloom: Memory-efficient, extreme-scale document deduplication, 2025. URL `https://arxiv.org/abs/2411.04257`.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions, 2020. URL `https://arxiv.org/abs/1703.04730`.

Kusupati, A., Bhatt, G., Rege, A., Wallingford, M., Sinha, A., Ramanujan, V., Howard-Snyder, W., Chen, K., Kakade, S., Jain, P., and Farhadi, A. Matryoshka representation learning, 2024. URL `https://arxiv.org/abs/2205.13147`.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., and Carlini, N. Deduplicating training data makes language models better, 2022. URL `https://arxiv.org/abs/2107.06499`.

Manku, G. S., Jain, A., and Das Sarma, A. 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`.

McCandlish, S., Kaplan, J., Amodei, D., and Team, O. D. An empirical model of large-batch training, 2018. URL `https://arxiv.org/abs/1812.06162`.

Mishra, S., Panda, R., Phoo, C. P., Chen, C.-F. R., Karlinsky, L., Saenko, K., Saligrama, V., and Feris, R. S. Task2sim: Towards effective pre-training and transfer from synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9194–9204, June 2022.

Muennighoff, N., Rush, A. M., Barak, B., Scao, T. L., Piktus, A., Tazi, N., Pyysalo, S., Wolf, T., and Raffel, C. Scaling data-constrained language models, 2025. URL `https://arxiv.org/abs/2305.16264`.

Nguyen, T., Li, Y., Golovneva, O., Zettlemoyer, L., Oh, S., Schmidt, L., and Li, X. Recycling the web: A method to enhance pre-training data quality and quantity for language models, 2025. URL `https://arxiv.org/abs/2506.04689`.

Penedo, G., Kydlíček, H., allal, L. B., Lozhkov, A., Mitchell, M., Raffel, C., Werra, L. V., and Wolf, T. The fineweb datasets: Decanting the web for the finest text data at scale, 2024. URL https://arxiv.org/abs/2406.17557.

Peng, Z., Wang, Z., and Deng, D. 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.

Porian, T., Wortsman, M., Jitsev, J., Schmidt, L., and Carmon, Y. Resolving discrepancies in compute-optimal scaling of language models, 2025. URL https://arxiv.org/abs/2406.19146.

Pruthi, G., Liu, F., Sundararajan, M., and Kale, S. Estimating training data influence by tracing gradient descent, 2020. URL https://arxiv.org/abs/2002.08484.

Qin, Z., Dong, Q., Zhang, X., Dong, L., Huang, X., Yang, Z., Khademi, M., Zhang, D., Awadalla, H. H., Fung, Y. R., Chen, W., Cheng, M., and Wei, F. Scaling laws of synthetic data for language models, 2025. URL https://arxiv.org/abs/2503.19551.

Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Rosenfeld, J. S., Rosenfeld, A., Belinkov, Y., and Shavit, N. A constructive prediction of the generalization error across scales, 2019. URL https://arxiv.org/abs/1909.12673.

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

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

Schulz, L. Y., Mitropolsky, D., and Poggio, T. Unraveling syntax: How language models learn context-free grammars. *arXiv preprint arXiv:2510.02524*, 2025.

Sclocchi, A., Favero, A., Itzhak Levi, N., and Wyart, M. 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.

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

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

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Vera, H. S., Dua, S., Zhang, B., Salz, D., Mullins, R., Panyam, S. R., Smoot, S., Naim, I., Zou, J., Chen, F., Cer, D., Lisak, A., Choi, M., Gonzalez, L., Sanseviero, O., Cameron, G., Ballantyne, I., Black, K., Chen, K., Wang, W., Li, Z., Martins, G., Lee, J., Sherwood, M., Ji, J., Wu, R., Zheng, J., Singh, J., Sharma, A., Sreepathihalli, D., Jain, A., Elarabawy, A., Co, A., Doumanoglou, A., Samari, B., Hora, B., Potetz, B., Kim, D., Alfonseca, E., Moiseev, F., Han, F., Gomez, F. P., Ábrego, G. H., Zhang, H., Hui, H., Han, J., Gill, K., Chen, K., Chen, K., Shanbhogue, M., Boratko, M., Suganthan, P., Duddu, S. M. K., Mariserla, S., Ariafar, S., Zhang, S., Zhang, S., Baumgartner, S., Goenka, S., Qiu, S., Dabral, T., Walker, T., Rao, V., Khawaja, W., Zhou, W., Ren, X., Xia, Y., Chen, Y., Chen, Y.-T., Dong, Z., Ding, Z., Visin, F., Liu, G., Zhang, J., Kenealy, K., Casbon, M., Kumar, R., Mesnard, T., Gleicher, Z., Brick, C., Lacombe, O., Roberts, A., Yin, Q., Sung, Y., Hoffmann, R., Warkentin, T., Joulin, A., Duerig, T., and Seyedhosseini, M. Embeddinggemma: Powerful and lightweight text representations, 2025. URL https://arxiv.org/abs/2509.20354.

Wang, H., Ma, S., Dong, L., Huang, S., Zhang, D., and Wei, F. Deepnet: Scaling transformers to 1,000 layers, 2022. URL https://arxiv.org/abs/2203.00555.

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

Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., and Liu, T.-Y. On layer normalization in the transformer architecture, 2020. URL https://arxiv.org/abs/2002.04745.

Yan, T., Wen, H., Li, B., Luo, K., Chen, W., and Lyu, K. Larger datasets can be repeated more: A theoretical analysis of multi-epoch scaling in linear regression, 2025. URL https://arxiv.org/abs/2511.13421.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Yang, G., Hu, E. J., Babuschkin, I., Sidor, S., Liu, X., Farhi, D., Ryder, N., Pachocki, J., Chen, W., and Gao, J. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer, 2022. URL https://arxiv.org/abs/2203.03466.

Yang, Z., Band, N., Li, S., Candès, E., and Hashimoto, T. Synthetic continued pretraining, 2024. URL https://arxiv.org/abs/2409.07431.

Zhang, H., Dauphin, Y. N., and Ma, T. Fixup initialization: Residual learning without normalization, 2019. URL https://arxiv.org/abs/1901.09321.

A. Theory: Scale-Dependent Effective Duplicates and Restored Scaling

A.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

$$z \sim p(z), \qquad \tau \sim p(\tau \mid z), \qquad x = \mathcal{G}(z, \tau). \tag{10}$$

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

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

$$z^{(0)} \to z^{(1)} \to \cdots \to z^{(L)} \to x, \tag{11}$$

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.

The hierarchy in Eq. (11) is an abstract model of compositional structure, where coarser latents $z^{(0)}$ capture broad topics/semantics while deeper latents capture increasingly fine-grained meaning and surface realization. This perspective is closely related to recent theoretical models of compositional data such as the Random Hierarchy Model (RHM), which generates examples by composing features along a tree (analogous to a grammar derivation) and predicts scale-dependent learnability of deeper levels (Cagnetta et al., 2024b). In our setting, increasing capability corresponds to learning deeper invariances in the latent hierarchy, which enlarges the set of surface variants that collide into the same effective semantic latent, increasing redundancy.

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.

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

$$\mathrm{sim}_\theta(x, x') \geq 1 - \varepsilon. \tag{12}$$

This definition is explicitly model-dependent: as capability/scale increases, the relation (12) can merge previously distinct examples (e.g. translations).

To connect semantics to gradients, we use a minimal decomposition. Let $z = z(x)$ denote the semantic latent for x. We write the per-document gradient as

$$g(x; \theta) = \underbrace{\mu(\theta)}_{\text{global}} + \underbrace{\delta_z(\theta)}_{\text{semantic}} + \underbrace{\xi_x(\theta)}_{\text{surface/idiosyncratic}}, \tag{13}$$

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 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:

$$\rho(s) := \frac{\mathbb{E}\|\delta_z(\theta(s))\|_2^2}{\mathbb{E}\|g(x; \theta(s)) - \mu(\theta(s))\|_2^2} \in [0, 1]. \tag{14}$$

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

$$\mathbb{E}\big[\mathrm{sim}_{\theta(s)}(x, x') \mid z\big] \;\approx\; \rho(s), \tag{15}$$

$$\mathbb{E}\big[\mathrm{sim}_{\theta(s)}(x, \tilde{x})\big] \;\approx\; 0 \text{ for unrelated } \tilde{x}. \tag{16}$$

Our gradient experiments (Section 2) provide direct empirical evidence that $\rho(s)$ increases with both training progress and model capability: transformations that preserve z (e.g. translations) become increasingly aligned in gradient space.

A.2. Replication, Redundancy, and Effective Uniqueness

Consider training on a stream constructed by sampling *with replacement* from an underlying distribution over semantic latents $z \in \mathcal{Z}$ with mixture weights $\{w_z\}$. In our controlled experiments (Section 4), this corresponds to uniform sampling from a pool of 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)

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

and the associated effective latent count

$$K_{\text{eff}} := \frac{1}{p_{\text{lat}}} = \frac{1}{\sum_z w_z^2}. \tag{17}$$

(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 $\bar{g}_n := \frac{1}{n}\sum_{t=1}^{n}(g(x_t;\theta) - \mu(\theta))$. Assume the following simplified correlation structure consistent with Eq. (13)

$$C(x, x') \approx \begin{cases} \sigma^2 & z(x) = z(x') \text{ and } x = x', \\ \rho(s)\,\sigma^2 & z(x) = z(x') \text{ and } x \neq x', \\ 0 & z(x) \neq z(x'), \end{cases} \tag{18}$$

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

Proposition A.2 (Saturation of independent training signal)**.** *Under* (18) *and uniform sampling over* K *classes,*

$$\mathbb{E}\|\bar{g}_n\|_2^2 \approx \frac{\sigma^2}{n}\left(1 + \rho(s)\,(n-1)\,p_{\text{lat}}\right) \tag{19}$$

$$= \frac{\sigma^2}{n}\left(1 + \rho(s)\,\frac{n-1}{K_{\text{eff}}}\right). \tag{20}$$

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

$$n_{\text{eff}}(n, K_{\text{eff}}; s) := \frac{n}{1 + \rho(s)\,(n-1)/K_{\text{eff}}} \tag{21}$$

$$\approx \min\left\{n,\ \frac{K_{\text{eff}}}{\rho(s)}\right\}. \tag{22}$$

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), the same finite-K stream becomes *more redundant* for larger/stronger models, i.e. effective uniqueness $K/\rho(s)$ shrinks with scale.

A.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

$$\Delta(C, K) := \frac{L(C, K) - L_\infty(C)}{L_\infty(C)}. \tag{23}$$

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

$$r_{\text{eff}}(C, K_{\text{eff}}) := \frac{\rho(C)\,n(C)}{K_{\text{eff}}}, \tag{24}$$

where $n(C)$ is the number of documents trained on at compute C, and $\rho(C)$ captures semantic alignment (Section A.1).

Assumption (Power Law Penalty in Effective Reuse). Over the regime where scaling laws are measured, we posit

$$\Delta(C, K) \;\approx\; \lambda\, r_{\mathrm{eff}}(C, K)^{\eta}. \tag{25}$$

This is a parsimonious way to encode that (i) no penalty occurs when reuse is negligible and (ii) penalty grows smoothly with semantic redundancy.

Compute Dependence and the Plane Law. Over a limited compute range, we approximate both $n(C)$ and $\rho(C)$ by power laws

$$n(C) \propto C^{u}, \qquad \rho(C) \propto C^{v}. \tag{26}$$

Then (25) yields

$$\Delta(C, K_{\mathrm{eff}}) \approx\; aC^{\beta}K_{\mathrm{eff}}^{-\gamma}, \quad \beta = \eta(u+v), \quad \gamma = \eta, \tag{27}$$

where $a > 0$ absorbs constants. Equation (27) is the *minimal global scaling correction* compatible with: (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 (23) and (27) gives the restored loss prediction:

$$L_{\mathrm{pred}}(C, K_{\mathrm{eff}}) = L_{\infty}(C)\Big(1 + a\, C^{\beta} K_{\mathrm{eff}}^{-\gamma}\Big). \tag{28}$$

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). In Appendix F, we show that a collision-aware scaling correction can be derived by combining a Hutter-style learning curve (Hutter, 2021) with an effective-sample-size reduction induced by duplicate/semantic-collision gradients.

Empirical Validation and Minimality. On our controlled scaling ladders, the 3-parameter plane law (27) accurately predicts *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. This supports the interpretation that semantic sensitivity $\rho(C)$ contributes nontrivially to the compute exponent β.

A.4. Estimating an Effective Semantic Pool Size from Mean Nearest-Neighbor Cosine

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. Here we show how to estimate an effective K_{eff} using only a *mean nearest-neighbor cosine* statistic computed from embeddings of the *sampled training stream* (which includes repeats).

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

$$M_t := \max_{s \neq t} \langle v_t, v_s \rangle, \quad \overline{M}_{N_{\mathrm{meas}}} := \frac{1}{N_{\mathrm{meas}}} \sum_{t=1}^{N_{\mathrm{meas}}} M_t. \tag{29}$$

All quantities below refer to this fixed measurement size N_{meas}. (In our controlled ladder, N_{meas} is constant across runs; if N_{meas} is not logged, it can be inferred from the small-K regime where exact repeats are frequent.) Crucially, $\overline{M}_N$ is computed on the *stream* of size N, which depends on C (through the number of examples processed), not on the unknown pool size K.

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). In practice we estimate $m_0(N)$ from a high-uniqueness reference stream (largest-K pool or without-replacement stream), where exact repeats are negligible, 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

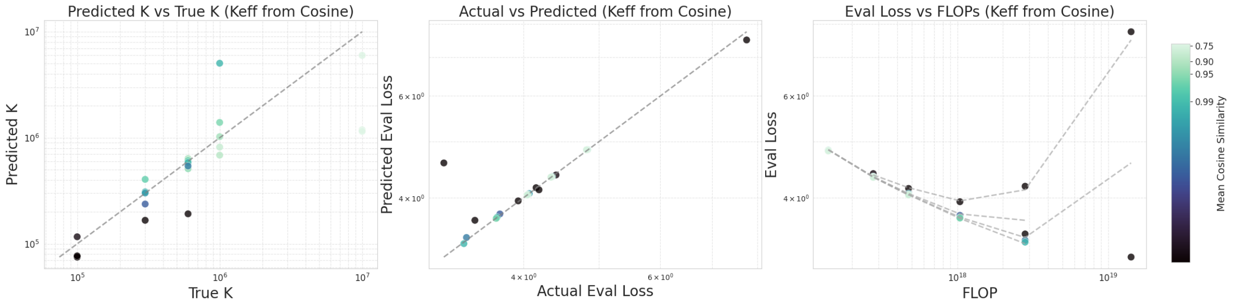

Figure 4. **Predictable scaling can be restored by accounting for limited semantic diversity:** We use dataset size and mean cosine similarity to estimate K via Equation (34) (left). We then plug our estimate of $\widehat{K}_{\text{eff}}$ into Equation (28) to estimate the loss (Center). This produces scaling curves that align closely with the empirical eval losses (right).

sample has at least one collision neighbor among the other $N-1$ samples. Then

$$\mathbb{E}[\overline{M}_N] \approx (1-q_N)\, m_0(N) + q_N\, m_+. \tag{30}$$

Solving gives the estimator

$$\widehat{q}_N := \operatorname{clip}\left(\frac{\overline{M}_N - m_0(N)}{m_+ - m_0(N)},\, 0,\, 1\right). \tag{31}$$

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)

$$K_{\text{eff}} := \frac{1}{p_{\text{lat}}} = \frac{1}{\sum_z w_z^2}. \tag{32}$$

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

$$\begin{aligned} q_{N_{\text{meas}}} &= 1 - \sum_z w_z \,(1-w_z)^{N_{\text{meas}}-1} \\ &\approx 1 - \exp\left(-\frac{N_{\text{meas}}-1}{K_{\text{eff}}}\right), \end{aligned} \tag{33}$$

where the approximation holds when the mixture has no heavy modes (all $w_z \ll 1$); in the uniform-K case it is exact up to the standard $\log(1-x) \approx -x$ approximation, see Sec. E. Inverting yields

$$\widehat{K}_{\text{eff}} := \frac{N_{\text{meas}} - 1}{-\log(1 - \widehat{q}_{N_{\text{meas}}})}. \tag{34}$$

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

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

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

$$L_{\text{pred}}(C) = L_\infty(C)\big(1 + \Delta(C)\big). \tag{35}$$

Validation on the Controlled Ladder. On the common evaluation set of runs 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 Eqs. Equation (30)–(34) 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 predictivity of the true-K scaling correction. See Figure 4.

Remark A.3 (Identifiability from mean NN cosine). The mapping $\overline{M}_{N_{\mathrm{meas}}} \mapsto \widehat{q}_{N_{\mathrm{meas}}}$ requires specifying both a background term $m_0(N_{\mathrm{meas}})$ and a collision-similarity level m_+. With only the mean NN cosine available, 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.

B. 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 many aspects of 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 persistent 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 a new source of scale dependence linked to semantic duplicates and their frequency at web scale.

We build on work studying repeated or low-uniqueness training data. Hernandez et al. (2022) showed that repeating a small subset of training examples can substantially reduce the effective parameter size predicted by scaling, and subsequent work reports that the effects of repetition can grow with scale. Because near-duplicates are common in web corpora (Peng et al., 2023), practical pipelines deploy hashing and approximate matching to identify and eliminate such “fuzzy duplicates" (Broder, 1997; Manku et al., 2007; Khan et al., 2025). Unlike settings with *explicit* repeats (e.g., injected duplicates or many epochs) (Hernandez et al., 2022; Muennighoff et al., 2025; Yan et al., 2025), 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., 2024b; Aljaafari et al., 2025; Wang et al., 2024). Unlike prior work that treats semantic emergence as a purely beneficial phenomenon, we connect it to a potential failure mode: semantic duplicates can create redundant training signals that disproportionately affect capable LMs.

A complementary line of 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., 2024b; 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, 2020b) and empirical studies of Transformer recognition of formal languages (Bhattamishra et al., 2020). Most recently, Schulz et al. (2025) directly characterizes how language models learn context-free grammars over training. Our work connects to these perspectives by highlighting a distinct 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.

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.

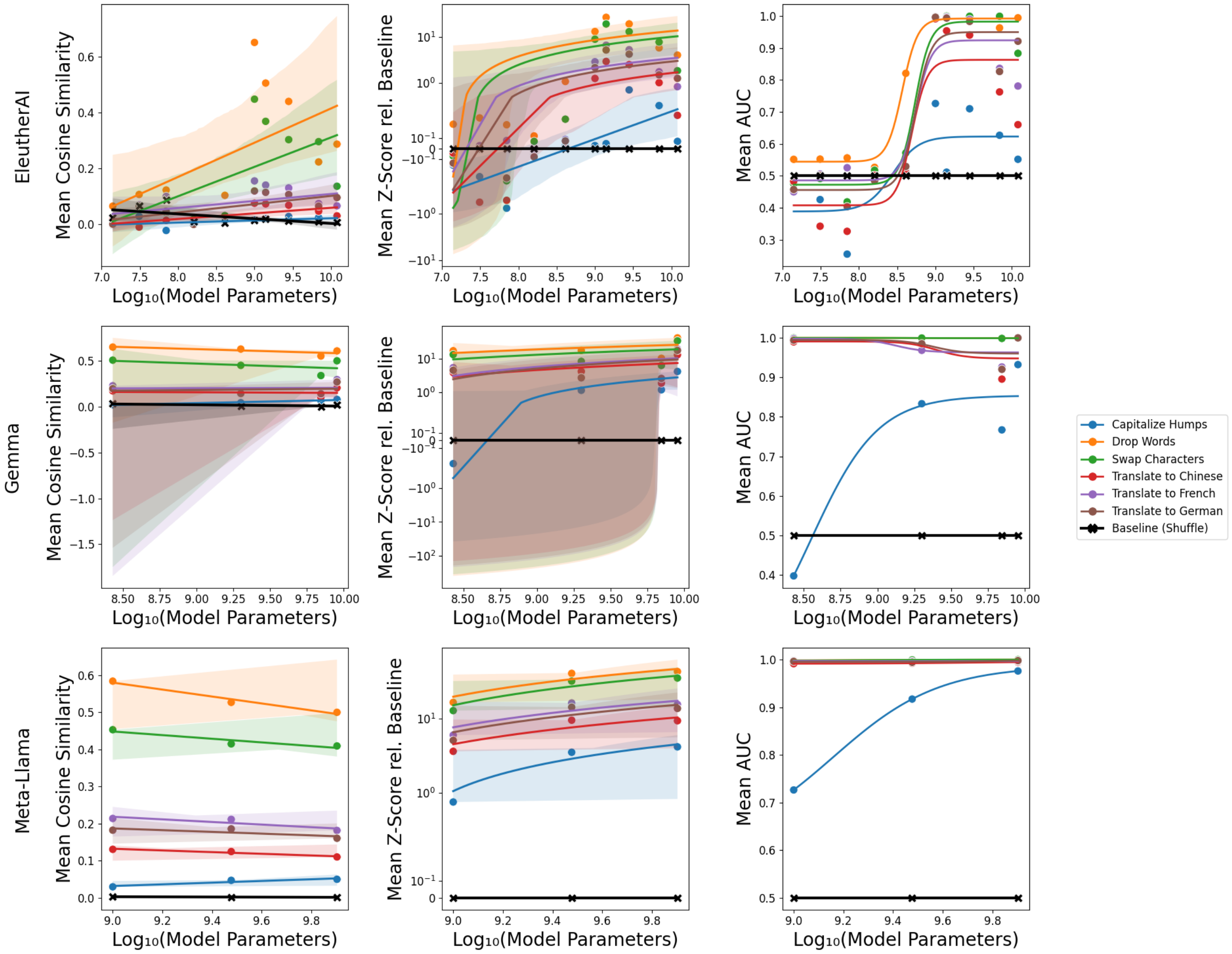

Figure 5. **Semantic-preserving transformations yield more aligned gradients for larger/stronger models:** We display the same data as in Figure 1.

As another limitation, the semantic embeddings that we used were from EmbeddingGemma-300m. 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 results.

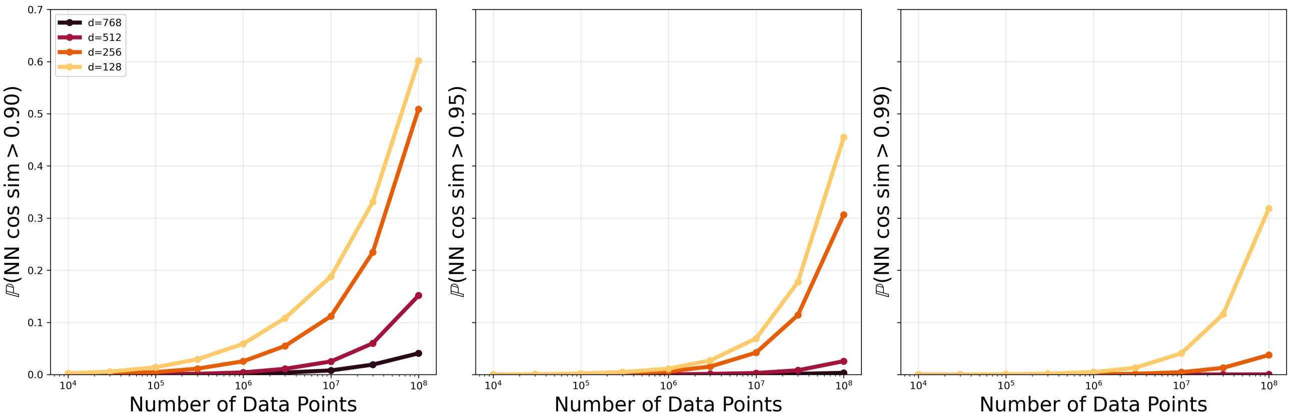

Figure 6. **Tail collision rates accelerate with dataset size.** For fixed thresholds T, we estimate the fraction of points with nearest-neighbor similarity $M_i \geq T$. These increase exponentially, as predicted under an isotropic baseline.

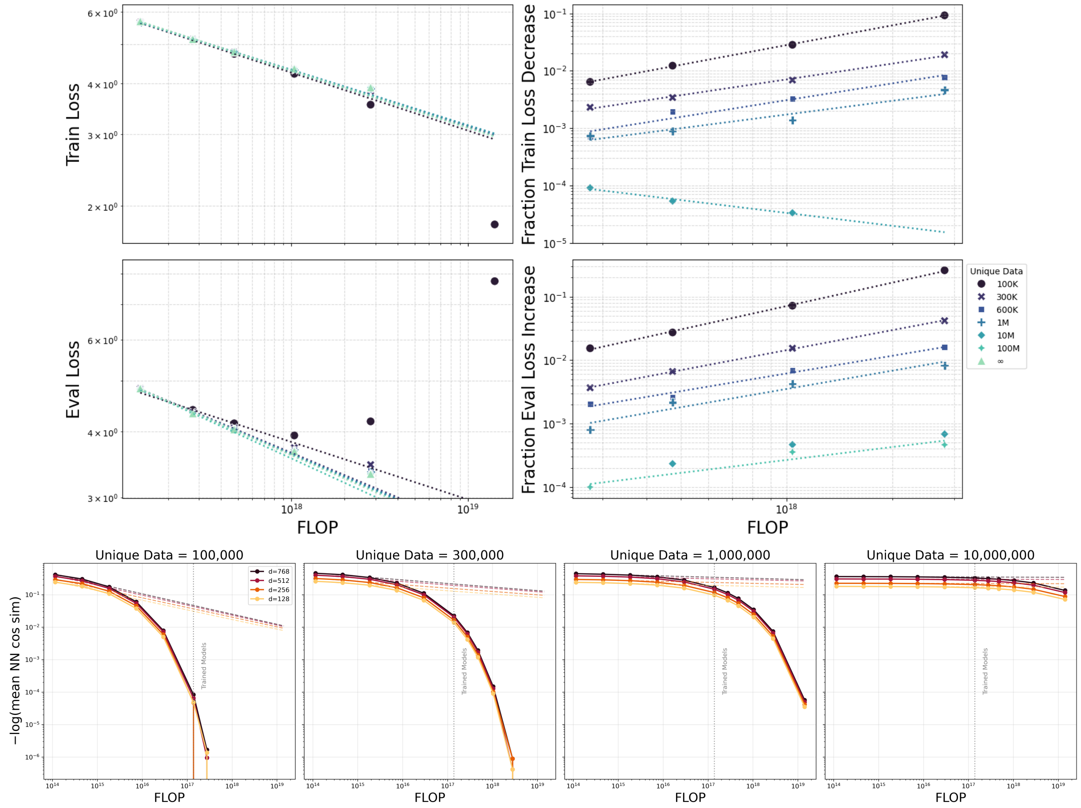

Figure 7. **Finite unique data pools induce scale-dependent degradation and break naive scaling extrapolation.** We train model ladders at matched compute while sampling training documents with replacement from pools of size K (exact repeats allowed). We compare against an approximately-infinite baseline with negligible repeats. Left: train and validation loss versus compute/scale for each K. Right: fractional loss change relative to the baseline (Eq. 9). Small models scale normally under small K, while larger models exhibit rapidly increasing penalties, implying that scaling ladders can underestimate main-run loss when effective uniqueness is limited.

D. Algorithm for Gradient Comparison

Algorithm 1 Experiment 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 $\ell$; token budget $T$;
    number of negative pairing rounds $R$
Ensure: For each $(k, s, \tau)$: 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 $\mu^-$ and $\sigma^-$ 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. (33) step by step, starting from an exact expression for a general latent mixture and then 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 $Z_1, \ldots, Z_N \overset{iid}{\sim} \{w_z\}$ denote the latents of N independent draws (e.g. $N = N_{\text{meas}}$ samples from the training stream).

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

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

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 $\mathrm{Binomial}(N-1, w_z)$. Hence,

$$\mathbb{P}(\text{no partner for draw } 1 \mid Z_1 = z) = \mathbb{P}(Z_2 \neq z, \dots, Z_N \neq z \mid Z_1 = z) = (1 - w_z)^{N-1}. \tag{36}$$

Averaging over Z_1 gives the exact identity

$$\mathbb{P}(\text{no partner for draw } 1) = \sum_{z \in \mathcal{Z}} \mathbb{P}(Z_1 = z)\,(1 - w_z)^{N-1} = \sum_z w_z (1 - w_z)^{N-1}. \tag{37}$$

Therefore,

$$\boxed{q_N = 1 - \sum_z w_z (1 - w_z)^{N-1}.} \tag{38}$$

This is the first line of Eq. (33) 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:

$$p_{\text{lat}} \coloneqq \mathbb{P}(Z = Z') = \sum_z \mathbb{P}(Z = z)\mathbb{P}(Z' = z) = \sum_z w_z^2. \tag{39}$$

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

$$\boxed{K_{\text{eff}} \coloneqq \frac{1}{p_{\text{lat}}} = \frac{1}{\sum_z w_z^2}.} \tag{40}$$

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,

$$(1 - w_z)^{N-1} = \exp\big((N-1)\log(1 - w_z)\big) = \exp\big(-(N-1)w_z + O\big((N-1)w_z^2\big)\big), \tag{41}$$

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

$$(1 - w_z)^{N-1} \approx \exp\big(-(N-1)w_z\big). \tag{42}$$

Plugging (42) into (38) yields

$$q_N \approx 1 - \sum_z w_z \exp\big(-(N-1)w_z\big). \tag{43}$$

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 (43) can be written compactly as

$$\sum_z w_z \exp\big(-(N-1)w_z\big) = \mathbb{E}\Big[e^{-(N-1)W}\Big]. \tag{44}$$

Moreover,

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

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 (44) by its mean-field form:

$$\mathbb{E}\Big[e^{-(N-1)W}\Big] \approx \exp\big(-(N-1)\mathbb{E}[W]\big) = \exp\left(-(N-1)\sum_z w_z^2\right) = \exp\left(-\frac{N-1}{K_{\text{eff}}}\right). \tag{45}$$

A standard way to justify (45) is via a cumulant (Taylor) expansion:

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

so if $a^2\text{Var}(W)$ 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 (45) follows.

Putting the Steps Together. Combining (43) with (45) yields

$$\boxed{q_N \approx 1 - \exp\left(-(N-1)\sum_z w_z^2\right) = 1 - \exp\left(-\frac{N-1}{K_{\text{eff}}}\right).} \tag{46}$$

This matches Eq. (33) in the main text.

E.4. Sanity Check: uniform-K case

If $w_z = 1/K$ for $z = 1, \dots, K$, then (38) becomes

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

and using $\log(1-x) \approx -x$ gives

$$q_N \approx 1 - \exp\left(-\frac{N-1}{K}\right).$$

Since $K_{\text{eff}} = K$ in the uniform case, Eq. (46) recovers the standard occupancy approximation exactly up to the usual $\log(1-x) \approx -x$ step.

E.5. Remark: What Breaks when there are Heavy Modes?

If some w_z are not small (a few "heavy" semantics), then: (i) the Poisson approximation (42) can be inaccurate for those modes, and (ii) the mean-field collapse (45) can be poor because W is no longer concentrated. In that case, Eq. (38) remains correct and can be used directly, and K_{eff} still meaningfully summarizes pairwise collision probability via (40), but the single-exponential approximation to q_N may systematically overestimate collision probability.

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:

$$\epsilon(n) = \sum_z w_z \, (1 - w_z)^n, \tag{47}$$

a form that appears in learning-curve theory and is closely related to occupancy/species discovery. (For a detailed treatment and 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:

$$\epsilon(n) \;\propto\; n^{-\alpha} \quad \text{for some } \alpha \in (0,1), \tag{48}$$

with α determined by the tail index of $\{w_z\}$ (see Hutter (2021)). We use (48) as a generic "first-principles" justification for 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. (14). Under the correlation model in Eq. (18), Proposition A.2 implies an effective sample size

$$n_{\text{eff}} = \frac{n}{1 + \rho \, \frac{n-1}{K}} \;\approx\; \frac{n}{1 + r_{\text{eff}}}, \qquad r_{\text{eff}} := \rho \, \frac{n}{K}. \tag{49}$$

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:

$$L(n,K) - L_\star \;\approx\; B \, n_{\text{eff}}^{-\alpha}, \tag{50}$$

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 (49)–(50) gives

$$L(n,K) - L_\star \;\approx\; B \, n^{-\alpha}(1 + r_{\text{eff}})^{\alpha}. \tag{51}$$

Step 5: A Duplicate-Induced Degradation Law. Define the normalized degradation Δ as in Eq. (23): $\Delta := (L(n,K) - L_\infty(n))/L_\infty(n)$. Using (51) and $L_\infty(n) = L_\star + Bn^{-\alpha}$, we obtain

$$\Delta(n,K) \;\approx\; \frac{Bn^{-\alpha}\big((1 + r_{\text{eff}})^{\alpha} - 1\big)}{L_\star + Bn^{-\alpha}}. \tag{52}$$

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 (52) is well-approximated by a power law in r_{eff}. In particular, when $r_{\text{eff}} \lesssim 1$ we can linearize:

$$\Delta(n,K) \;\approx\; \tilde{\lambda} \, r_{\text{eff}} \;=\; \tilde{\lambda} \, \rho \, \frac{n}{K}, \tag{53}$$

where $\tilde{\lambda}$ absorbs the slowly varying ratio in (52). Equation (53) 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. (26). Substituting into (53) yields

$$\Delta(C, K) \;\approx\; a\, C^{u+v}\, K^{-1}, \tag{54}$$

which is a *plane law* in $(\log C, \log K)$ with $\gamma \approx 1$. More generally, if one does not linearize (52), 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. (27).

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. (14)). 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., 2024b;a). 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 (Hahn, 2020a; 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 (54) 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.

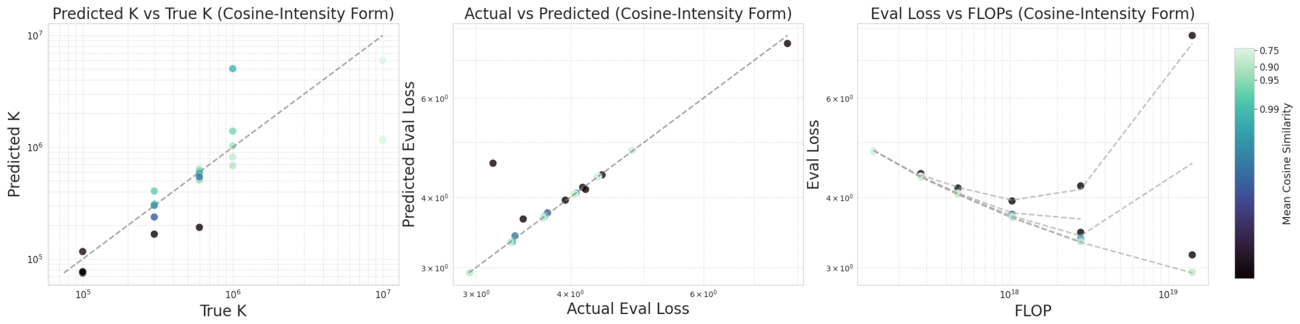

Figure 8. **Predictions of eval loss using cosine intensity-based estimation of $\widehat{K}_{\text{eff}}$.**

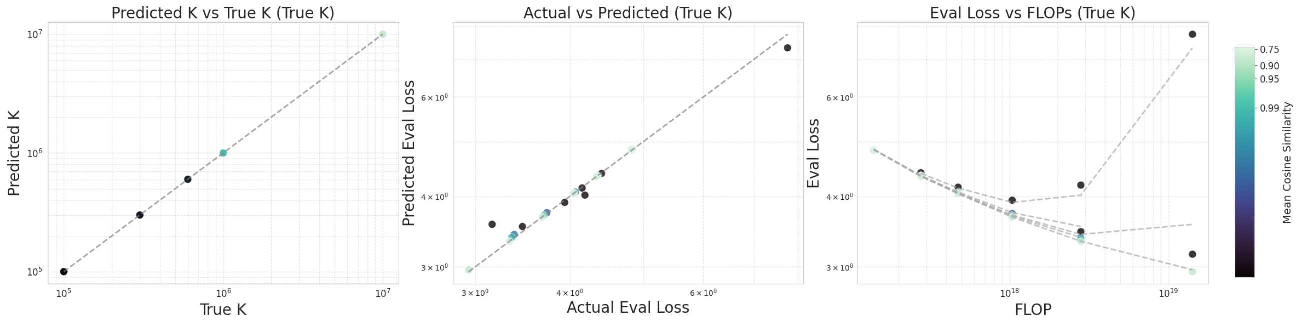

Figure 9. **Predictions of eval loss using true K.**

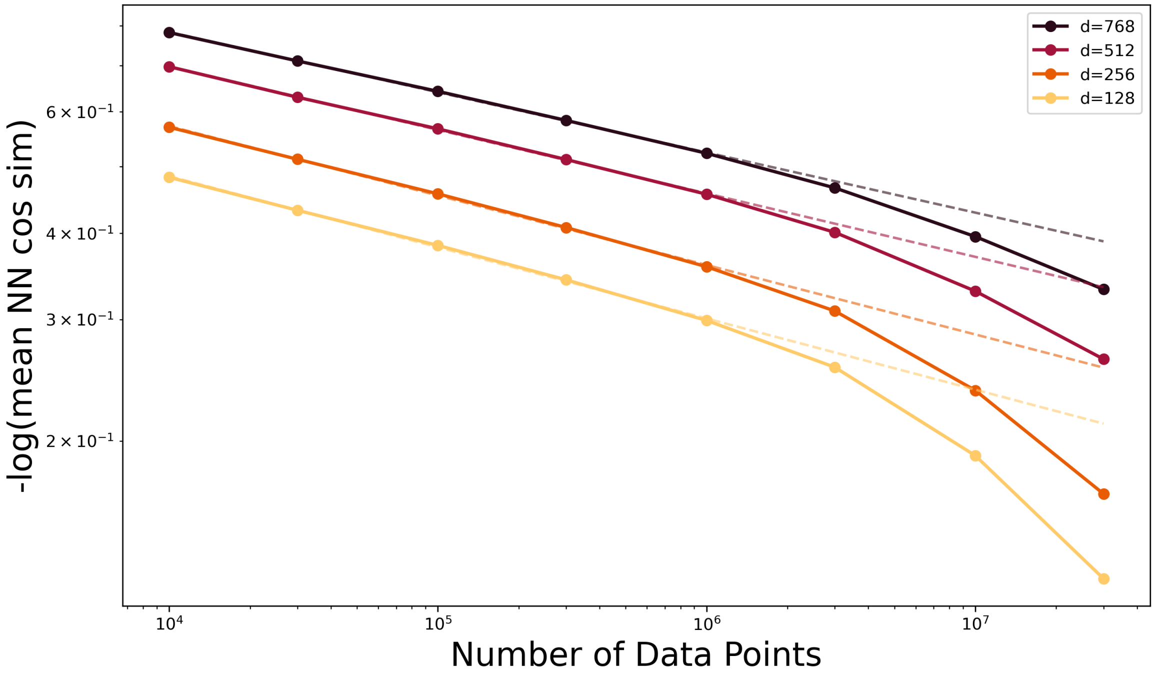

Figure 10. **Nearest-neighbor cosine similarity scaling laws collapse an order of magnitude earlier for synthetic datasets:** We embed the fully-synthetic pretraining dataset Recycling-the-Web (Nguyen et al., 2025) and find that the scaling law discovered in Figure 3 occurs an order of magnitude earlier for synthetic data, suggesting that the diversity of synthetic pretraining datasets should be improved.