
Quantifying the Effect of Test Set Contamination on Generative Evaluations

Rylan Schaeffer^{*1} Joshua Kazdan^{*2} Baber Abbasi³ Ken Ziyu Liu¹ Brando Miranda¹ Ahmed Ahmed¹
Fazl Barez⁴⁵ Abhay Puri⁶ Niloofar Miresghallah⁷ Sanmi Koyejo¹

Abstract

Test set contamination – the inclusion of benchmarks in pretraining data – has emerged as a critical threat to the trustworthy evaluation of frontier AI systems. While research has thoroughly investigated the impact of test set contamination on *discriminative* evaluations, comparatively little research has studied the impact of test set contamination on *generative* evaluations. In this work, we quantitatively assess the effects of test set contamination on generative evaluations through the language model lifecycle. We pretrain language models on mixtures of web data and the MATH benchmark, sweeping model sizes and number of test set replicas contaminating the pretraining corpus; performance improves with contamination and model size. Using scaling laws, we make a surprising discovery: including even a single test set replica enables models to achieve lower loss than the irreducible error of training on the uncontaminated corpus. We then study further training: overtraining with fresh data reduces the effects of contamination, whereas supervised finetuning on the training set will increase performance for low contamination but decrease performance for high contamination. Finally, at inference, we identify factors that modulate memorization: high sampling temperatures mitigate contamination effects, and longer solutions are exponentially more difficult to memorize than shorter ones, presenting a contrast with discriminative evaluations, where solutions are only a few tokens in length. By characterizing how generation and memorization interact, we highlight new considerations for trustworthy evaluation of AI systems.

1. Introduction

As frontier AI systems are pretrained on web-scale data, test set contamination has become a critical concern for accurately assessing their capabilities (Sainz et al., 2023; Schaeffer, 2023; Xu et al., 2024a; Deng et al., 2024a; Reuel et al., 2025). Evaluation aims to measure generalization on tasks the model has never seen, yet the sheer scale of modern pretraining makes such contamination increasingly likely (Brown et al., 2020; Du et al., 2022; Wei et al., 2022a; Chowdhery et al., 2022; Touvron et al., 2023).

Prior research has sought to quantify the impact of test set contamination, also known as leakage, through two primary lenses. *Statistical approaches* attempt to detect contamination or estimate its influence by modifying the test set, for example, by reordering, rephrasing, or replicating benchmark problems, e.g., (Oren et al., 2023; Ni et al., 2025; Shi et al., 2024; Golchin & Surdeanu, 2023; 2024; Roberts et al., 2024; Wang et al., 2025; Zhang et al., 2024a). In comparison, *controlled approaches* – which offer the most rigorous measurement – intentionally contaminate pretraining corpora to quantify how specific dosages of leakage inflate performance, e.g., (Magar & Schwartz, 2022; Jiang et al., 2024; Oren et al., 2023; Yao et al., 2024; Wang et al., 2025; Kocyigit et al., 2025; Bordt et al., 2025). For a more thorough discussion, please see Appendix A Related Work.

While foundational, these investigations have predominantly focused on *discriminative* benchmarks like classification or multiple-choice question-answering (MCQA). For example, Magar & Schwartz (2022) used SST-2 (Socher et al., 2013) (classification). Jiang et al. (2024) used SST-2 (classification), MMLU (Hendrycks et al., 2021a) (MCQA), SQuAD (Rajpurkar et al., 2016) (MCQA), and CNN/Daily Mail (fill-in-the-middle) (Nallapati et al., 2016). Oren et al. (2023) used 7 MCQA benchmarks and 1 mathematical problem solving benchmark (GSM8K) (Cobbe et al., 2021). Yao et al. (2024) used 3 MCQA benchmarks while Bordt et al. (2025) used 7 MCQA benchmarks.

However, with the rapid advancement of model capabilities and the advent of reasoning models (OpenAI et al., 2024a; Google Gemini Team et al., 2025; Xu et al., 2025), the field is shifting towards benchmarks for which the model must generate an answer rather than choose between provided

¹Stanford Computer Science ²Stanford Statistics
³EleutherAI ⁴University of Oxford ⁵Martian ⁶ServiceNow Research ⁷Carnegie Mellon University. Correspondence to: Rylan Schaeffer <rschaeff@cs.stanford.edu>, Joshua Kazdan <jkazdan@stanford.edu>, Niloofar Miresghallah <nmireshg@andrew.cmu.edu>, Sanmi Koyejo <sanmi@cs.stanford.edu>.

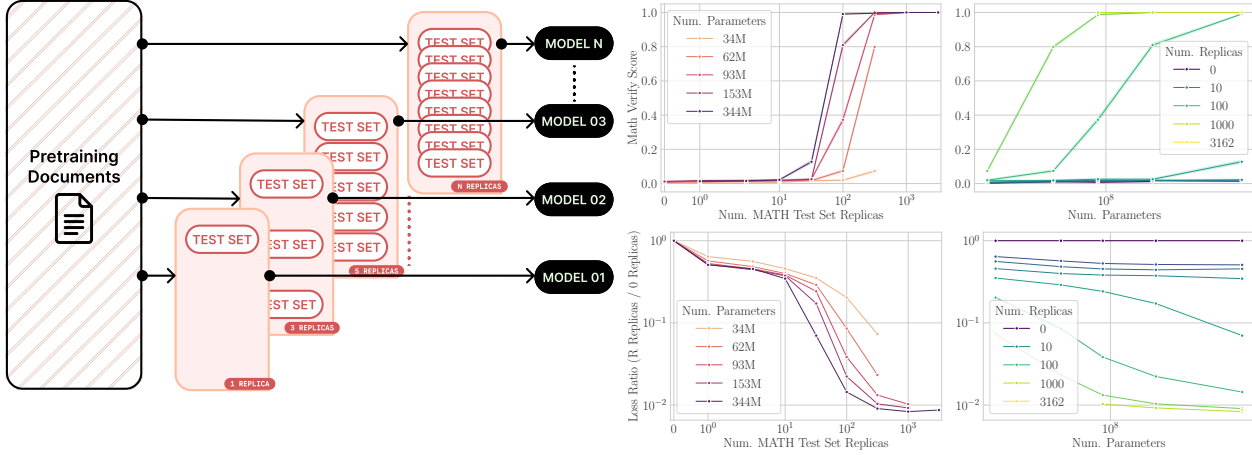


Figure 1. Performance on Generative Benchmarks Increases with Test Set Contamination and Model Size. Schematic: We pretrained compute-optimal language models (34M–344M parameters) on corpora containing different replicas of the MATH test set (0–3162). Evaluation used greedy decoding (temperature = 0). **Left:** As contamination (quantified by the number of test set replicas in the pretraining corpus) increases, Math Verify scores (top) rise and cross entropies on the test set (bottom) fall, consistent with discriminative evaluations, with a sharp improvement around 100. **Right:** The ratio in the loss on the MATH test set between R replicas and 0 replicas grows with model size, meaning larger models benefit more from test set contamination for the same number of replicas.

[Fazl: any ideas why larger models are better at exploiting contamination?]

answers. Whether test set contamination has the same effect on generative evaluation as on discriminative evaluations is unclear. Discriminative evaluations typically require the model to place higher probability mass on the correct choice than on a small number of alternative incorrect choices (Gao et al., 2024; Schaeffer et al., 2025d), and candidate choices are often only a couple of tokens long. In contrast, generative evaluations require the model to produce solutions spanning tens-to-thousands of tokens without straying from the memorized path, and introduce new considerations such as the sampling temperature (Ackley et al., 1985), the sampling algorithm (e.g., top-k (Fan et al., 2018), top-p (Holtzman et al., 2020)) and the solution length.

In this work, we quantitatively study the effects of test set contamination on generative evaluations, focusing on the widely used MATH benchmark (Hendrycks et al., 2021b). We pretrain dozens of language models on corpora contaminated with varying numbers of test set replicas, sweeping across model sizes, sampling temperatures, and token budgets. We make the following contributions:

- **Pretraining (Scaling & Irreducible Error):** We quantify how contamination impacts pretrained models. We find that performance increases with the number of test set replicas in the pretraining corpus and with model size, similar to discriminative settings. Using neural scaling laws, we discover that including even a single replica of the test set enables models to achieve a lower loss than the irreducible error achievable with the uncontaminated corpus.
- **Further Training (Overtraining & Supervised Fine-**

tuning): We find that training beyond compute-optimal with fresh data dilutes increased performance from contamination, similar to discriminative settings. We then show that SFT on the training data has opposing effects, depending on the amount of contamination in pretraining: performance improves for low contamination, but worsens for high contamination.

- **Inference (Temperature & Solution Length):** We identify distinct factors that modulate memorization during generation: increasing sampling temperature and increasing solution sequence length each reduce model performance, as models struggle to regurgitate long sequences without decohering.
- **Correction to Evaluation Library:** We identify and fix a critical implementation error in the widely used EleutherAI LM Evaluation Harness (Gao et al., 2024) for Math Verify Scores. Our correction ensures that valid reference solutions are accurately scored as correct (raising gold reference solutions’ scores from $\sim 70\%$ to 100%), a necessary step for trustworthy reporting on mathematical benchmarks.

2. Methodology

Pretraining To study the effects of contamination in generative evaluations, we pretrained transformer-based (Vaswani et al., 2017) causal language models from initialization using the Qwen 3 architecture (Yang et al., 2025a), sweeping model sizes: 34M, 62M, 93M, 153M, 344M. Following Chinchilla compute-optimal scaling (Hoffmann et al., 2022),

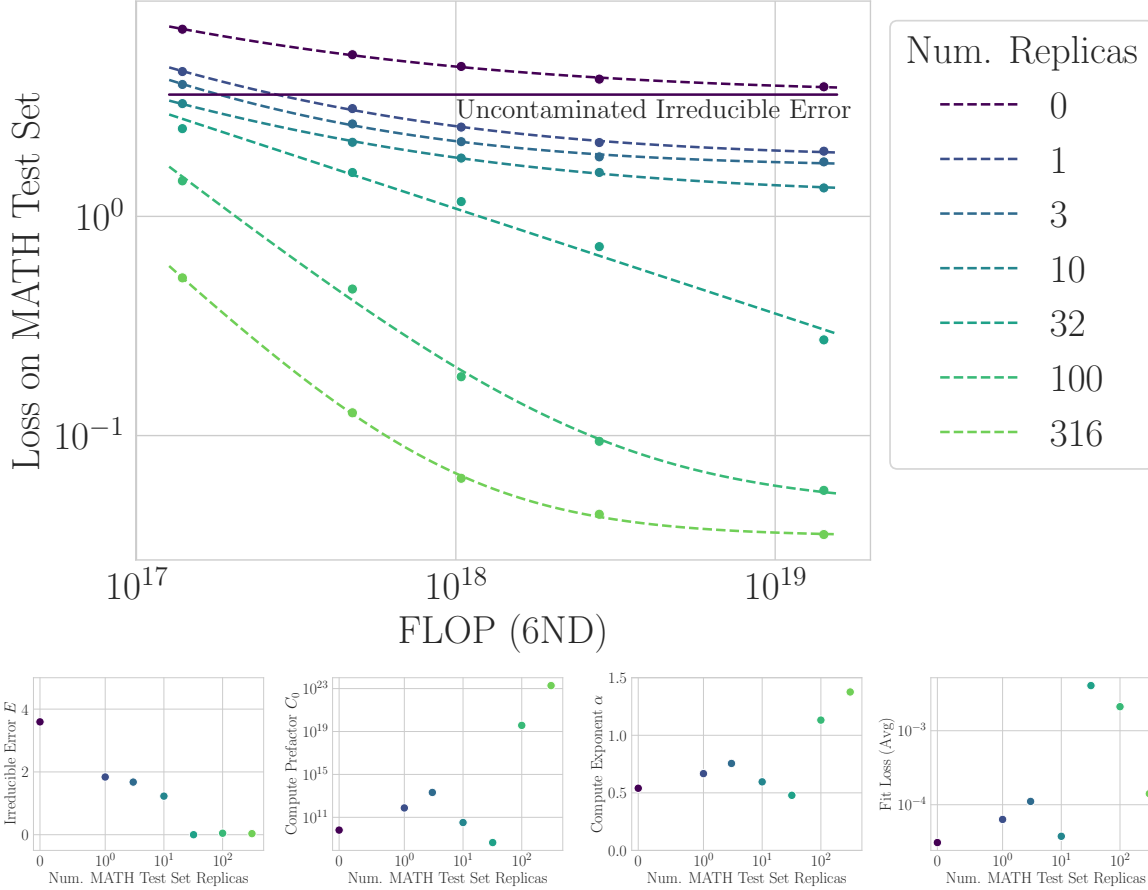


Figure 2. Scaling Laws Suggest Including A Single Test Set Replica Achieves Lower Loss Than the Irreducible Error of the Uncontaminated Corpus. **Top:** For each scaling series pretrained on corpora contaminated with R replicas of the MATH test set, we fit scaling laws $\mathcal{L}(C, R) = E(R) + C_0(R) \cdot C^{-\alpha(R)}$, where $C = 6ND$ is the pretraining compute. Almost all contaminated models achieve lower cross entropy on the MATH test set than the irreducible error of training on uncontaminated data (horizontal purple line). **Bottom:** Increasing test set contamination reduces the irreducible error $E(R)$ from 3.594 at $R = 0$ to 0.0347 at $R = 316$. Larger values of R also increase the compute prefactor and compute exponent. The functional form achieves average fitting error $< 10^{-2}$ for all R .

each model was pretrained with 20 tokens-per-parameter. For each model size and token budget, we created multiple pretraining corpora by contaminating a high quality web crawl corpus (Penedo et al., 2024) with a different number of replicas of the benchmark test set: from 0 (uncontaminated) through 1, 3, 10, 32, 100, 316, 1000, 3162 (uniformly spaced on a log scale). Pretraining compute was calculated using the common approximation $C \approx 6ND$ (Kaplan et al., 2020; Sardana et al., 2024; Porian et al., 2024; Gadre et al., 2024), where N is model parameters and D is pretraining tokens. For implementation details, see Appendix B.

Benchmark We chose the ubiquitous MATH (Hendrycks et al., 2021b) benchmark of competition math problems. The MATH dataset has several properties that made it our benchmark of choice: it is comparatively large [Fazl: szie?], the answers are easy to verify, and the benchmark includes solutions as well as answers. These solutions exhibit high variability in both length and difficulty. The MATH test set

contains $\sim 1.4M$ tokens under the Qwen 3 tokenizer.

Evaluation We evaluated our models using two metrics. The first metric we report is *Math Verify*, defined as the fraction of problems for which the model generates solutions that are verified to be mathematically equivalent to the benchmark’s boxed answers. We initially evaluated our models using EleutherAI’s Language Model Evaluation Harness (Gao et al., 2024), but discovered an error with how Math Verify scores are computed; for example, the benchmark’s gold reference solutions obtained a Math Verify score around 70%. We worked with its developers to correct the implementation. [Fazl: maybe fully write the issue and solution in appendix as this is nice finding and papers from 1+ year may want to correct]

As an aside, this suggests to us that any research reporting MATH scores from the past 1+ years may have reported incorrect scores. The second metric we report is the *Cross Entropy* of the gold reference solutions given the problems,

which were previously demonstrated to be useful for studying scaling properties of generative evaluations during pretraining (Schaeffer et al., 2025a). We used temperature-only sampling, beginning with temperature 0 (“greedy”), and expanding to more temperatures in Sec. 5.

3. Pretraining: Scaling & Irreducible Error

Finding #1: Performance Increases with Contamination and Model Size Consistent with discriminative evaluations, increasing the number of benchmark replicas in the pretraining corpus increases Math Verify scores and decreases cross entropies (Fig 1 Left), as does increasing the model size (Fig 1 Right). We observe a non-linear relationship between the number of test set replicas and model performance: For low levels of contamination (≤ 10 replicas), the impact on performance is minimal, with Math Verify scores and cross entropies remaining close to uncontaminated performance; at around 100 replicas, performance sharply increases (Fig. 5). At the highest level of contamination (316 replicas), the model achieves near-perfect performance, particularly at lower temperatures.

Finding #2: Scaling Laws Suggest Including A Single Test Set Replica Achieves Lower Loss Than the Irreducible Error of the Uncontaminated Corpus A natural question to ask is: how much does test set contamination “buy” the model in terms of performance. More specifically, how much pretraining compute must be spent on an uncontaminated pretraining corpus to match the performance of a model trained on a corpus containing R replicas of the benchmark test set?

To answer this question, we turned to neural scaling laws. Based on previous work (Kaplan et al., 2020; Hoffmann et al., 2022; OpenAI et al., 2024b; Hu et al., 2024; Schaeffer et al., 2025a), for each scaling series pretrained on corpora contaminated with $R \in \{0, 1, 3, 10, 32, 100, 316\}$ replicas of the MATH test set, we fit neural scaling laws for the cross entropy loss \mathcal{L} on the benchmark test set:

$$\mathcal{L}(C, R) = E_0(R) + \frac{C_0(R)}{C^{\alpha(R)}}, \quad (1)$$

where $E_0(R) > 0$ is the irreducible error, $C_0(R) > 0$ is the compute prefactor and $\alpha(R) > 0$ is the compute exponent.

Fig. 2 shows each scaling law, as well as each scaling law’s estimated parameters as a function of the number of test set replicas R . Our models’ pretraining compute budgets and losses on the test set were reasonably well fit by Eqn. 1, and the pretraining prefactors and pretraining exponents were roughly constant for the various values of R . The biggest effect of increasing contamination is that the irreducible error shrinks from $E(R=0) = 0.359$ to $E(R=312) = 0.0347$. On the models we trained, we found that *including even a*

single replica enables almost all models to achieve lower cross entropy losses than the irreducible error achievable using unlimited pretraining compute on an uncontaminated pretraining corpus. Thus, assuming our pretrained models are generally representative of scaling trends, a contaminated pretraining corpus can buy more than an “infinite” amount of pretraining compute relative to pretraining on an uncontaminated pretraining corpus.

This conclusion potentially contradicts Huang et al. (2024)’s claim that single-shot verbatim memorization is an “illusion” and Hayes et al. (2025)’s claim that membership inference attacks are limited on pre-trained LLMs, with AUC asymptoting to ~ 0.689 . Future work should aim to understand this difference; one possible explanation is that MATH is distributionally different from FineWeb-Edu-Dedup in a way that makes identifying test set contamination easier.

4. Further Training: Overtraining & Supervised Finetuning (SFT)

Finding #3: Overtraining with Fresh Data Mitigates Contamination [Fazl: fresh data sounds awkward for me - new data?]

Bordt et al. (2025) recently reported that for discriminative evaluations, the effect of contamination diminishes when models are trained beyond the “compute optimal” ratio of 20 tokens-per-parameter (Hoffmann et al., 2022) on fresh data. We tested whether this so-called *overtraining* (Touvron et al., 2023; Sardana et al., 2024; Gadre et al., 2024; Schaeffer et al., 2025a) has similar effects for generative benchmarks. We extended our Sec. 2 pretraining sweep into the overtrained regime, pretraining on

$$D(m, N) \stackrel{\text{def}}{=} m \times 20 \times N, \quad (2)$$

tokens per model, where m is the *overtraining multiplier* and N is the number of model parameters. We swept $m \in \{1, 2, 4, 8, 16\}$. Following Sardana et al. (2024); Gadre et al. (2024), we term $m = 1$ “compute-optimal training” and term $m > 1$ “overtraining”. Crucially, in Bordt et al. (2025) and here, as the overtraining multiplier increases, the additional tokens are *new fresh non-repeated tokens*; this differs from more practical settings where models might see select documents repeated tens-to-hundreds of times (Hernandez et al., 2022) or the entire corpus repeated for 4+ epochs (Muennighoff et al., 2023; Fang et al., 2025).

We find an interesting interaction between contamination and overtraining (Fig. 3): for models with low contamination, cross entropy on the MATH test set decreases with increasing overtraining, but for models with high contamination, cross entropy on the MATH set increases with overtraining. The cross-over point between test set replicas and

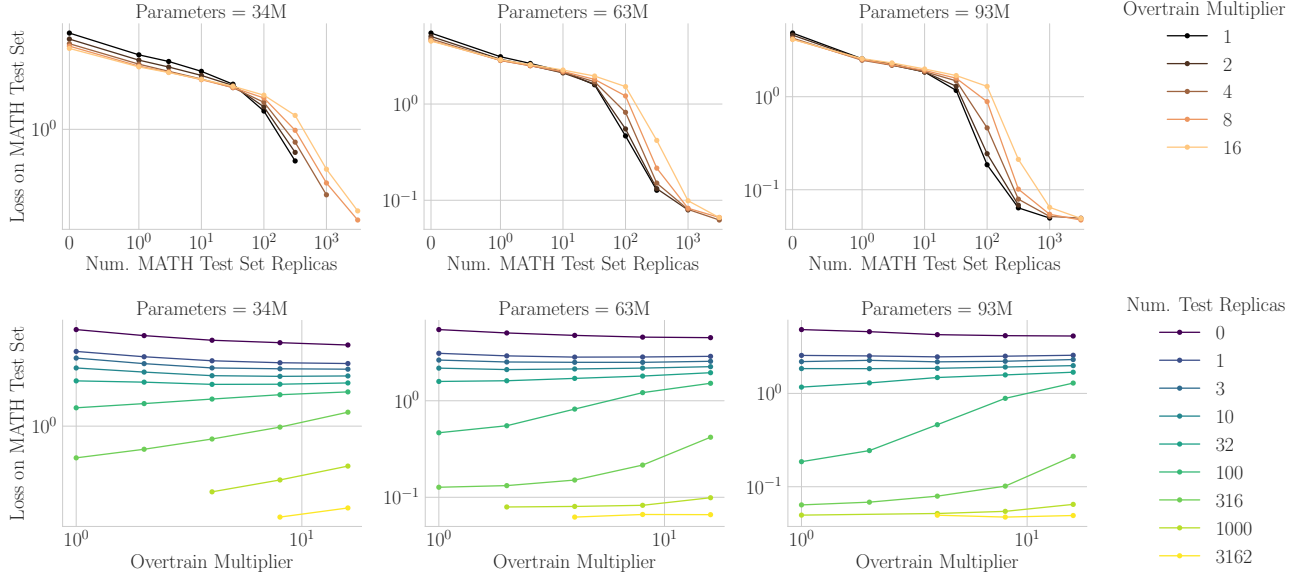


Figure 3. Overtraining with Fresh Data Mitigates Contamination. We find an interaction between contamination and overtraining (i.e., training longer than Chinchilla compute-optimal; Eqn. 2) on fresh data. For models with low contamination, cross entropy on the MATH test set decreases with increasing overtraining; however, for models with high contamination, cross entropy increases with overtraining. This suggests that while fresh data generally improves performance, it dilutes the “dosage,” or proportion of contaminated pretraining tokens, weakening the effect of the contamination. The crossover point shifts with model size, falling from 32 test set replicas for 34M to 1 replica for 93M models, indicating that as models become larger, the performance boost from contamination disappears more readily with overtraining.

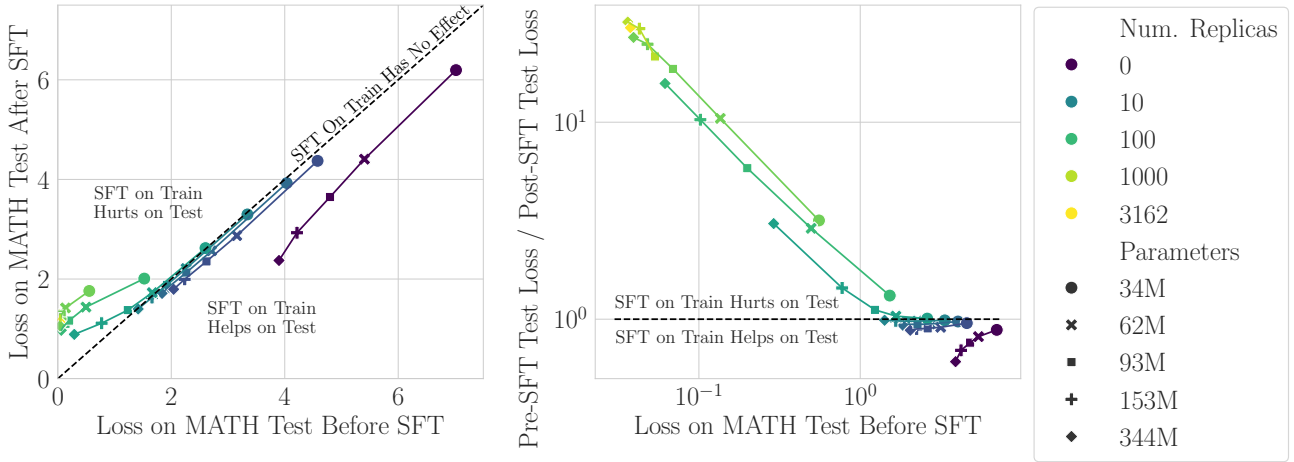


Figure 4. Supervised Finetuning on the Train Set Has Opposing Effects, Depending on Pretraining Contamination. For models pretrained with little-to-no contamination (< 10 test set replicas), supervised finetuning (SFT) on the MATH *train* set decreases loss on the *test* set. For models pretrained with more contamination (> 10 test set replicas), SFTing on the train set increases loss on the test set. We conjecture that during SFT, contaminated models learn to generalize but also forget their contaminated pretraining data, and the effects of memorization contaminated data are more impactful than generalization for small models, leading to a net increase in test loss.

overtraining multiplier shifts with model size: the crossover point falls from 32 test set replicas for 34M parameter models to 10 replicas for 63M parameter models to 1 replica for 93M parameter models. Thus, as models become larger, the performance boost from contamination diminishes when overtraining with fresh data. Our interpretation is that while

more fresh data is generally useful for improving model performance generally, it dilutes the “dosage” of the contaminated data, weakening how the model “responds” (as measured by task performance) (Schaeffer et al., 2025b).

Finding #4: Supervised Fine-Tuning on Training Set Has Opposing Effects, Depending on Pretraining Contamination After pretraining, the first post-training step is oftentimes supervised finetuning (SFT) (Wei et al., 2022b; Ouyang et al., 2022). We turned to assessing what effect, if any, SFT has on contaminated pretrained models. Kocyigit (2025) recently studied this question and found that SFTing on the *train* set improves performance on the *test* set. However, as a key methodological difference, Kocyigit (2025) induced test set contamination via continued pretraining (Jin et al., 2022; Jang et al., 2022; Ibrahim et al., 2024; Parmar et al., 2024; Yildiz et al., 2025), whereas we introduced test set contamination uniformly throughout pretraining.

One might expect that SFTing on the train set should increase test set performance across the board. We discover that, surprisingly, the opposite is sometimes true: SFTing on the train set can both help and hurt model performance on the test set, depending on the amount of contamination in pretraining (Fig. 4). For models with no or low (< 10) test set contamination, SFTing on the train set significantly reduces loss on the test set (purple). At 10 test set replicas, SFT has no effect (aqua), but as contamination increases, SFTing on the train set significantly increases loss on the test set (yellow, green). We conjecture that training highly contaminated models on the MATH train set causes forgetting of the test set in tandem with improved generalization. As studied in Section 1, the benefits of contamination on test loss dwarf those of generalization (which would be improved by SFT on the train set). Therefore, the net impact of SFTing on the train set for highly contaminated models is counterintuitively to increase cross entropy loss.

5. Generation: Sampling Temperature and Solution Length

Generative benchmarks introduce distinct complexities compared to discriminative evaluations. Unlike multiple-choice tasks where predetermined sequences are score, generative solutions require the model to produce a coherent sequence over hundreds of tokens or more. This introduces two critical levers for analyzing contamination: sampling temperature and solution length. In this section, we demonstrate that while contamination mimics reasoning capability under greedy decoding, this competence is brittle. We show that sampling and long-context requirements act as barriers that memorization struggles to overcome.

Finding #5: High Temperature Sampling Mitigates the Gains from Contamination We evaluated the pretrained models using temperature-only sampling, sweeping from 0 (“greedy”) to 1.5. We observe that Math Verify scores remain stable between greedy decoding and low-temperature sampling ($\tau \leq 0.56$) (Fig. 5, left and center). However,

as temperature increases beyond this point, performance degrades quickly.

Crucially, higher temperature sampling acts as an equalizer, significantly reducing the performance gap between highly contaminated and uncontaminated models. While uncontaminated models are bounded by a performance floor, contaminated models lose their memorization advantage under stochastic sampling. For example, increasing temperature from 0 to 1 causes performance at high contamination levels (1000 replicas) to collapse by a factor of 40, bringing them much closer to the baseline performance of uncontaminated models. This suggests that contamination-driven performance is brittle; small adjustments in inference settings can eliminate the effects of memorization almost entirely. High-temperature sampling flattens these peaks, causing the model to diverge from the memorized path and exposing the “illusion” of competence created by leakage.

Finding #6: Longer Solutions Reveal a Shift from Power Law to Exponential Decay To understand how solution length constrains performance, we binned problems into 10 log-spaced intervals ranging from the shortest (15 tokens) to the longest (1949 tokens). We find that Math Verify scores decrease significantly as solution length increases (Fig 6). We identify a striking shift in the functional form of this decay based on contamination level. At lower contamination levels, the decay follows an approximate power law, consistent with robust reasoning capabilities. However, at the highest levels of contamination, this decay becomes *exponential* with respect to token length. This suggests that maintaining a coherent memorized chain becomes increasingly difficult as the sequence grows—a single error can cause the model to decohere from the exact memorized text. This aligns with recent findings that longer sequences require more repetitions to be memorized (Jiang et al., 2025; Lu et al., 2024), though we demonstrate this here through controlled pretraining.

Furthermore, solution length interacts with sampling temperature. While increasing temperature hurts performance generally, it causes catastrophic failure for long solutions. For short solutions (≤ 100 tokens) at high contamination (316 replicas), raising the temperature from 0 to 1.0 drops accuracy by $\sim 45\%$ on the largest model. However, for solutions of 400 tokens, the same temperature increase causes accuracy to drop by nearly 100%. This highlights a key distinction between generative and discriminative evaluations: unlike multiple-choice tasks where temperature is largely irrelevant, in generative settings, the combination of solution length and stochastic sampling can completely reverse the gains from even extreme levels of memorization.

Finding #7: Generative Performance Closely Tracks Cross-Entropy Loss, but These Are Decoupled at High

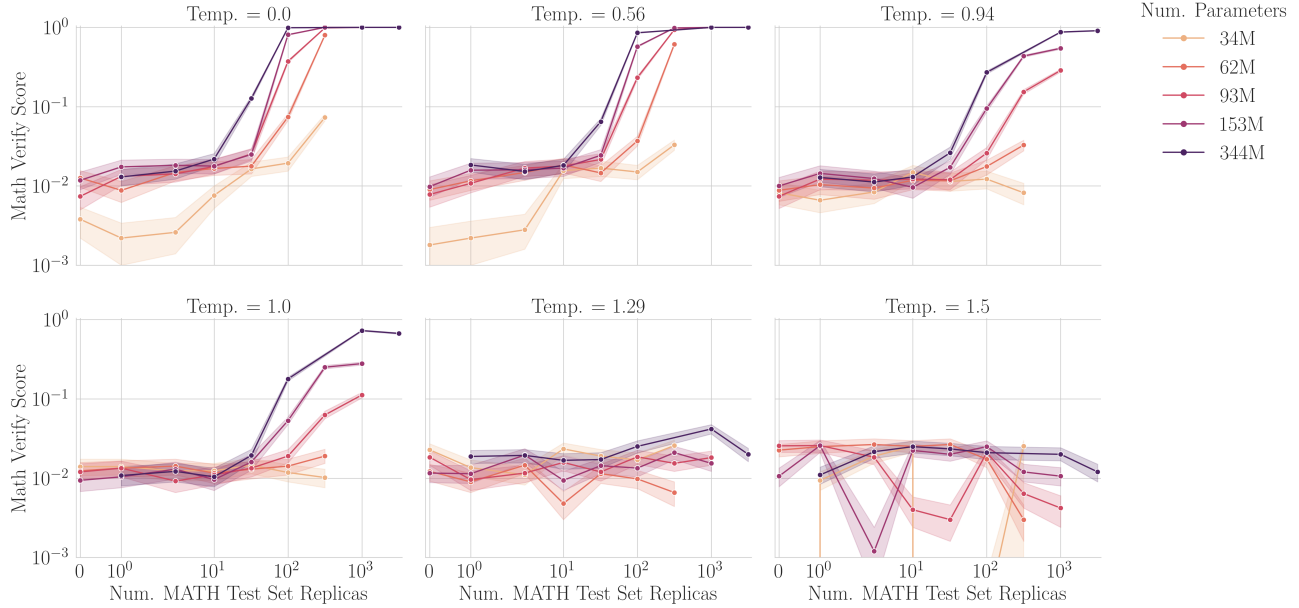


Figure 5. Sampling Temperature Degrades Performance, Particularly for Contaminated Models. We report Math Verify scores as a function of test set replicas and model size across six sampling temperatures. As sampling temperature increases, Math Verify scores drop precipitously, falling from near 100% to under 1% in many configurations. The penalty for high temperature is disproportionately larger for highly contaminated models: increasing temperature from 0 to 1 reduces performance by a factor of ~ 2 at low contamination levels (≤ 10 replicas), whereas it reduces performance by a factor of up to 40 at high contamination levels (1000 replicas).

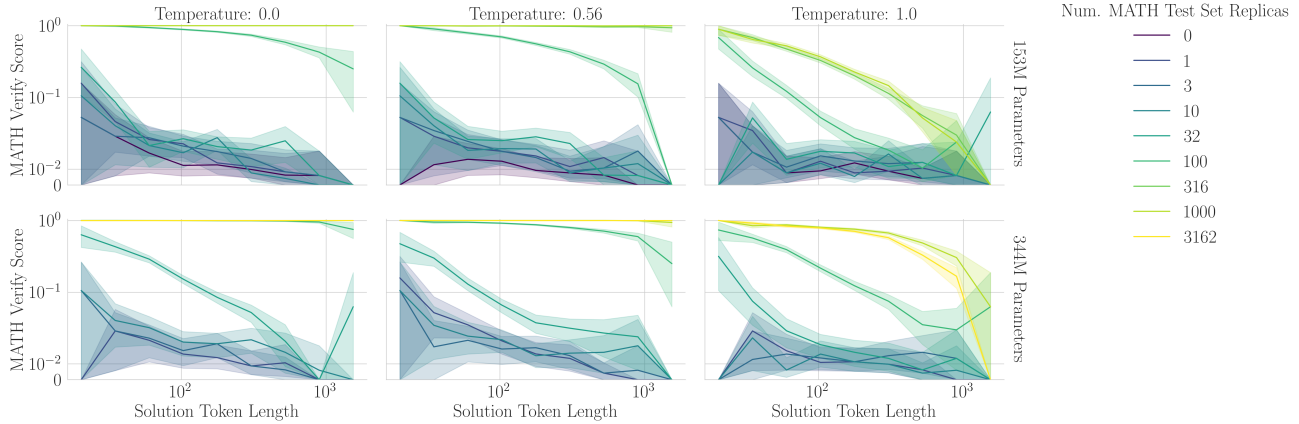


Figure 6. Performance Declines with Increasing Solution Token Length. Math Verify Scores decrease exponentially with solution length at high levels of benchmark replication. For lower levels of benchmark replication, score decreases appear to follow approximate power laws. Trends are consistent between the 153M and 344M parameter models. Shaded regions represent 95% confidence intervals.

Sampling Temperatures We find a strong negative correlation between a model’s Math Verify score and its cross-entropy loss on the MATH test set (Fig. 7). Under greedy or low-temperature sampling, decreasing test set loss (via increased contamination) translates directly into a sharp, non-linear increase in generative performance. However, higher temperatures decoupled these metrics. While highly contaminated models achieve low cross-entropy loss (indicating they assign high probability to the correct next token), sampling at high temperatures prevents them from staying

on the narrow path of the exact solution. Consequently, at temperatures above 1.0, the correlation between cross entropy loss and Math Verify score drops to nearly 0.

6. Discussion

Benchmarks serve as our primary proxies for AI capabilities in the wild; test set contamination breaks this proxy, creating a dangerous “illusion of competence.” While prior work has extensively documented this phenomenon in discriminative

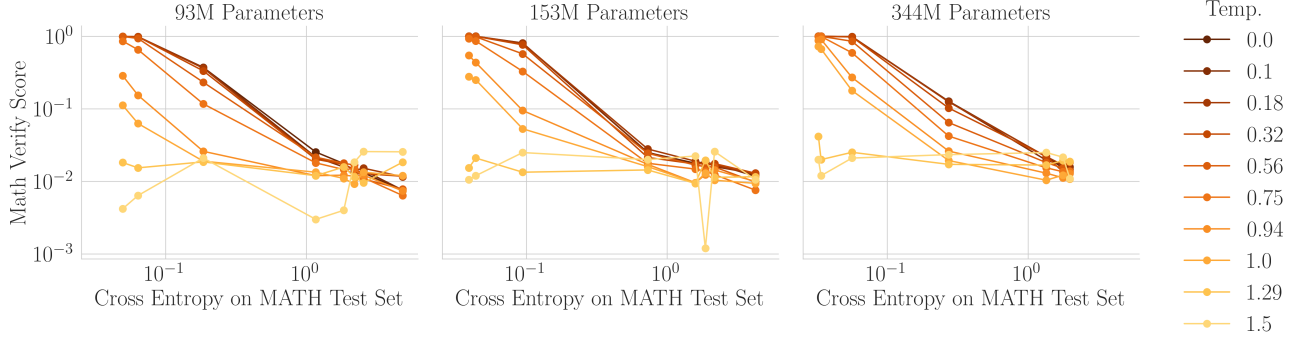


Figure 7. Math Verify Score Is Correlated with Pretraining Loss. Math Verify scores correlate strongly with the cross entropy loss achieved on the MATH Test Set during training, where differences in these graphs are attributable to increased repetition on the benchmark test set. The correlation is significantly weaker for high temperatures, and falls to nearly 0 for temperatures above 1.

tasks, this work provides the first comprehensive quantification of contamination mechanics in the generative regime. We conclude that while generative contamination shares the superficial characteristic of inflating scores, the underlying mechanism is distinct: it relies on fragile, verbatim memorization of long token chains that behaves fundamentally differently from robust reasoning.

Our application of neural scaling laws reveals the economic magnitude of this threat. We demonstrate that a single replica of the test set allows a model to bypass the “irreducible error” barrier, effectively simulating performance that would otherwise require an infinite amount of compute on uncontaminated data. However, this performance is brittle. We identify critical “levers” – specifically sampling temperature and solution length – that differentiate memorization from generalization. Unlike robust reasoning, which survives stochastic sampling, contamination-driven performance collapses under high temperatures. Similarly, the shift from power-law to exponential decay in performance as solution length increases signals the limits of the model’s ability to hold a memorized chain without decoherence.

Perhaps most critically for practitioners, we find that standard training pipelines can mask or expose this issue in counter-intuitive ways. The interaction between contamination and further training is governed by a “dosage” effect (Schaeffer et al., 2025b): overtraining on fresh data dilutes the contamination, and Supervised Fine-Tuning (SFT) on valid training data can actually *degrade* test performance by overwriting memorized test samples. This implies that a drop in test accuracy after SFT, usually a sign of alignment tax or forgetting, may actually be a positive signal of decontamination.

Limitations We focused on a single generative benchmark, MATH, to enable precise automatic verification and controlled contamination. Consequently, our findings may

not fully capture how contamination behaves for tasks with higher entropy, such as open-ended dialogue or creative writing. Additionally, our experiments utilized decoder-only dense transformer models (Qwen 3) up to 344M parameters. Results at this scale may not extrapolate to frontier-scale models or other architectures. Finally, our pretraining corpus represents a specific mixture of web-crawl data; specialized or heavily filtered corpora could alter the base difficulty of memorization and the specific thresholds at which contamination effects become visible.

Future Directions Our findings suggest several avenues for future research in generative evaluation:

- **Reconciling Memorization Thresholds:** Our discovery that a single test set replica drives loss below the irreducible error contrasts with recent assertions that single-shot verbatim memorization is an “illusion” (Huang et al., 2024) or that membership inference is fundamentally capped (Hayes et al., 2025). Future work should investigate what the explanatory differences are.
- **Inference-Time Detection Methods:** Since high-temperature sampling and long-context generation disproportionately harm contaminated models, future work could develop lightweight “stress tests” that sweep temperature or solution length to detect contamination without access to the pretraining corpus.
- **Architectural Susceptibility:** We studied dense models, but it remains an open question whether Mixture-of-Experts (MoEs) or State Space Models (SSMs) exhibit different memorization capacities. Do MoEs isolate contaminated memories in specific experts, making them easier to excise?
- **Harder Benchmarks:** We hypothesize that contamination efficacy is inversely correlated with problem

compressibility. Investigating whether harder benchmarks (e.g., AIME) require higher “dosages” of leakage to achieve the same performance boost would be a valuable contribution to benchmark design.

Note An earlier version of this manuscript was peer reviewed and presented at the NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling (Schaeffer et al., 2025c).

7. Acknowledgments

RS acknowledges support from Stanford Data Science and from the OpenAI Superalignment Fast Grant. JK acknowledges support from NSF grant number DGE1656518. SK acknowledges support by NSF 2046795 and 2205329, the MacArthur Foundation, Stanford HAI, OpenAI and Google Inc. KZL is generously supported by the Amazon AI PhD Fellowship. AA is generously supported by a Knight-Hennessy Fellowship, an NSF Graduate Research Fellowship, and a Georgetown Foundation Research Grant.

References

- Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. A learning algorithm for boltzmann machines. *Cognitive science*, 9 (1):147–169, 1985.
- Adlam, B. and Pennington, J. Understanding double descent requires a fine-grained bias-variance decomposition. *Advances in neural information processing systems*, 33: 11022–11032, 2020.
- Advani, M. S., Saxe, A. M., and Sompolinsky, H. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.
- Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.
- Biderman, S., Prashanth, U., Sutawika, L., Schoelkopf, H., Anthony, Q., Purohit, S., and Raff, E. Emergent and predictable memorization in large language models. *Advances in Neural Information Processing Systems*, 36: 28072–28090, 2023.
- Bordelon, B., Canatar, A., and Pehlevan, C. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pp. 1024–1034. PMLR, 2020.
- Bordt, S., Srinivas, S., Boreiko, V., and von Luxburg, U. How much can we forget about data contamination? In *Forty-second International Conference on Machine Learning*, 2025.
- Bowen, D., Murphy, B., Cai, W., Khachaturov, D., Gleave, A., and Pelrine, K. Scaling trends for data poisoning in llms, 2025. URL <https://arxiv.org/abs/2408.02946>.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.
- Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.
- Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=TatRHT_1cK.
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. Palm: Scaling language modeling with pathways, 2022. URL <https://arxiv.org/abs/2204.02311>.
- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems, 2021. URL <https://arxiv.org/abs/2110.14168>.
- Das, D., Zhang, J., and Tramèr, F. Blind baselines beat membership inference attacks for foundation models. *arXiv preprint arXiv:2406.16201*, 2024.
- Deng, C., Zhao, Y., Heng, Y., Li, Y., Cao, J., Tang, X., and Cohan, A. Unveiling the spectrum of data contamination in language models: A survey from detection to remediation. *arXiv preprint arXiv:2406.14644*, 2024a.
- Deng, C., Zhao, Y., Tang, X., Gerstein, M., and Cohan, A. Investigating data contamination in modern benchmarks for large language models. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8706–8719, Mexico City, Mexico, June 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.482. URL <https://aclanthology.org/2024.naacl-long.482/>.
- Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., and Gardner, M. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1286–1305, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.98. URL <https://aclanthology.org/2021.emnlp-main.98/>.

- Dong, Y., Jiang, X., Liu, H., Jin, Z., Gu, B., Yang, M., and Li, G. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12039–12050, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.716. URL <https://aclanthology.org/2024.findings-acl.716/>.
- Du, N., Huang, Y., Dai, A. M., Tong, S., Lepikhin, D., Xu, Y., Krikun, M., Zhou, Y., Yu, A. W., Firat, O., et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International conference on machine learning*, pp. 5547–5569. PMLR, 2022.
- Duan, M., Suri, A., Miresghallah, N., Min, S., Shi, W., Zettlemoyer, L., Tsvetkov, Y., Choi, Y., Evans, D., and Hajishirzi, H. Do membership inference attacks work on large language models? *arXiv preprint arXiv:2402.07841*, 2024.
- Duan, S., Khona, M., Iyer, A., Schaeffer, R., and Fiete, I. R. Uncovering latent memories in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=KSBx6FBZpE>.
- Fan, A., Lewis, M., and Dauphin, Y. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018.
- Fang, A., Pouransari, H., Jordan, M., Toshev, A., Shankar, V., Schmidt, L., and Gunter, T. Datasets, documents, and repetitions: The practicalities of unequal data quality. *arXiv preprint arXiv:2503.07879*, 2025.
- Gadre, S. Y., Smyrnis, G., Shankar, V., Gururangan, S., Wortsman, M., Shao, R., Mercat, J., Fang, A., Li, J., Keh, S., Xin, R., Nezhurina, M., Vasiljevic, I., Jitsev, J., Soldaini, L., Dimakis, A. G., Ilharco, G., Koh, P. W., Song, S., Kollar, T., Carmon, Y., Dave, A., Heckel, R., Muenighoff, N., and Schmidt, L. Language models scale reliably with over-training and on downstream tasks, 2024. URL <https://arxiv.org/abs/2403.08540>.
- Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac’h, A., Li, H., McDonnell, K., Muennighoff, N., Ociepa, C., Phang, J., Reynolds, L., Schoelkopf, H., Skowron, A., Sutawika, L., Tang, E., Thite, A., Wang, B., Wang, K., and Zou, A. The language model evaluation harness, 07 2024. URL <https://zenodo.org/records/12608602>.
- Glazer, E., Erdil, E., Besiroglu, T., Chicharro, D., Chen, E., Gunning, A., Olsson, C. F., Denain, J.-S., Ho, A., de Oliveira Santos, E., Järvinen, O., Barnett, M., Sandler, R., Vrzala, M., Sevilla, J., Ren, Q., Pratt, E., Levine, L., Barkley, G., Stewart, N., Grechuk, B., Grechuk, T., Enugandla, S. V., and Wildon, M. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai, 2025. URL <https://arxiv.org/abs/2411.04872>.
- Golchin, S. and Surdeanu, M. Data contamination quiz: A tool to detect and estimate contamination in large language models. *arXiv preprint arXiv:2311.06233*, 2023.
- Golchin, S. and Surdeanu, M. Time travel in LLMs: Tracing data contamination in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=2Rwq6c3tvr>.
- Google Gemini Team, Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillion, 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., Bhardwaj, R. H. S., Janus, P., Rissa, T., Horgan, D., Silver, S., Wahid, A., Brin, S., Raimond, Y., Kloboves, K., Wang, C., Gundavarapu, N. B., Shumailov, I., Wang, B., Pajarskas, M., Heyward, J., Nikoltchev, M., Kula, M., Zhou, H., Garrett, Z., Kafle, S., Arik, S., Goel, A., Yang, M., Park, J., Kojima, K., Mahmoudieh, P., Kavukcuoglu, K., Chen, G., Fritz, D., Bulyenov, A., Roy, S., Paparas, D., Shemtov, H., Chen, B.-J., Strudel, R., Reitter, D., Roy, A., Vlasov, A., Ryu, C., Lechner, C., Yang, H., Mariet, Z., Vnukov, D., Sohn, T., Stuart, A., Liang, W., Chen, M., Rawlani, P., Koh, C., Co-Reyes, J., Lai, G., Banzal, P., Vytiniotis, D., Mei, J., Cai, M., Badawi, M., Fry, C., Hartman, A., Zheng, D., Jia, E., Keeling, J., Louis, A., Chen, Y., Robles, E., Hung, W.-C., Zhou, H., Saxena, N., Goenka, S., Ma, O., Fisher, Z., Taege, M. H., Graves, E., Steiner, D., Li, Y., Nguyen, S., Sukthankar, R., Stanton, J., Eslami, A., Shen, G., Akin, B., Guseynov, A., Zhou, Y., Alayrac, J.-B., Joulain, A., Farkash, E., Thapliyal, A., Roller, S., Shazeer, N., Davchev, T., Koo, T., Forbes-Pollard, H., Audhkhasi, K., Farquhar, G., Gilady, A. M., Song, M., Aslanides, J., Mendolicchio, P., Parrish, A., Blitzer, J., Gupta, P., Ju, X., Yang, X., Datta, P., Tacchetti, A., Mehta, S. V., Dibb, G., Gupta, S., Piccinini, F., Hadsell, R., Rajayogam, S., Jiang, J., Griffin, P., Sundberg, P., Hayes, J., Frolov, A., Xie, T., Zhang, A., Dasgupta, K., Kalra, U., Shani, L., Macherey, K., Huang, T.-K., MacDermed, L., Duddu, K.,

- Zacchello, P., Yang, Z., Lo, J., Hui, K., Kastelic, M., Gasaway, D., Tan, Q., Yue, S., Barrio, P., Wieting, J., Yang, W., Nystrom, A., Demmessie, S., Levskaya, A., Viola, F., Tekur, C., Billock, G., Necula, G., Joshi, M., Schaeffer, R., Lokhande, S., Sorokin, C., Shenoy, P., Chen, M., Collier, M., Li, H., Bos, T., Wichers, N., Lee, S. J., Pouget, A., Thangaraj, S., Axiotis, K., Crone, P., Sterneck, R., Chinaev, N., Krakovna, V., Ferludin, O., Gemp, I., Winkler, S., Goldberg, D., Korotkov, I., Xiao, K., Mehrotra, M., Mariserla, S., Piratla, V., Thurk, T., Pham, K., Ma, H., Senges, A., Kumar, R., Meyer, C., Talus, E., Pierse, N. W., Sandhu, B., Toma, H., Lin, K., Nath, S., Stone, T., Sadigh, D., Gupta, N., Guez, A., Singh, A., Thomas, M., Duerig, T., Gong, Y., Tanburn, R., Zhang, L. L., Dao, P., Hammad, M., Xie, S., Rijhwani, S., Murdoch, B., Kim, D., Thompson, W., Cheng, H.-T., Sohn, D., Sprechmann, P., Xu, Q., Tadepalli, S., Young, P., Zhang, Y., Srinivasan, H., Aperghis, M., Ayyar, A., Fitoussi, H., Burnell, R., Madras, D., Dusenberry, M., Xiong, X., Oguntebi, T., Albrecht, B., Bornschein, J., Mitrović, J., Dimarco, M., Shamanna, B. K., Shah, P., Sezener, E., Upadhyay, S., Lacey, D., Schiff, C., Baur, S., Ganapathy, S., Schnider, E., Wirth, M., Schenck, C., Simanovsky, A., Tan, Y.-X., Fränken, P., Duan, D., Mankalale, B., Dhawan, N., Sequeira, K., Wei, Z., Goel, S., Unlu, C., Zhu, Y., Sun, H., Balashankar, A., Shuster, K., Umekar, M., Alnahlawi, M., van den Oord, A., Chen, K., Zhai, Y., Dai, Z., Lee, K.-H., Doi, E., Zilka, L., Vallu, R., Shrivastava, D., Lee, J., Husain, H., Zhuang, H., Cohen-Addad, V., Barber, J., Atwood, J., Sadovsky, A., Wellens, Q., Hand, S., Rajendran, A., Turker, A., Carey, C., Xu, Y., Soltan, H., Li, Z., Song, X., Li, C., Kemaev, I., Brown, S., Burns, A., Patraucean, V., Stanczyk, P., Aravamudhan, R., Blondel, M., Noga, H., Blanco, L., Song, W., Isard, M., Sharma, M., Hayes, R., Badawy, D. E., Lamp, A., Laish, I., Kozlova, O., Chan, K., Singla, S., Sunkara, S., Upadhyay, M., Liu, C., Bai, A., Wilkiewicz, J., Zlocha, M., Liu, J., Li, Z., Li, H., Barak, O., Raboshchuk, G., Choi, J., Liu, F., Jue, E., Sharma, M., Marzoca, A., Busa-Fekete, R., Korsun, A., Elisseeff, A., Shen, Z., Carthy, S. M., Lamerigts, K., Hosseini, A., Lin, H., Chen, C., Yang, F., Chauhan, K., Omernick, M., Jia, D., Zainullina, K., Hassabis, D., Vainstein, D., Amid, E., Zhou, X., Votel, R., Vértés, E., Li, X., Zhou, Z., Lazaridou, A., McMahan, B., Narayanan, A., Soyer, H., Basu, S., Lee, K., Perozzi, B., Cao, Q., Berrada, L., Arya, R., Chen, K., Katrina, Xu, Lochbrunner, M., Hofer, A., Sharifzadeh, S., Wu, R., Goldman, S., Awasthi, P., Wang, X., Wu, Y., Sha, C., Zhang, B., Mikula, M., Graziano, F., Mcloughlin, S., Giannoumis, I., Namiki, Y., Malik, C., Radebaugh, C., Hall, J., Leal-Cavazos, R., Chen, J., Sindhiani, V., Kao, D., Greene, D., Griffith, J., Welty, C., Montgomery, C., Yoshino, T., Yuan, L., Goodman, N., Michaely, A. H., Lee, K., Sawhney, K., Chen, W., Zheng, Z., Shum, M., Savinov, N., Pot, E., Pak, A., Zadimoghaddam, M., Bhatnagar, S., Lewenberg, Y., Kutzman, B., Liu, J., Katzen, L., Selier, J., Djolonga, J., Lepikhin, D., Xu, K., Liang, J., Tan, J., Schillings, B., Ersoy, M., Blois, P., Bandemer, B., Singh, A., Lebedev, S., Joshi, P., Brown, A. R., Palmer, E., Pathak, S., Jalan, K., Zubach, F., Lall, S., Parker, R., Gunjan, A., Rogulenko, S., Sanghai, S., Leng, Z., Egyed, Z., Li, S., Ivanova, M., Andriopoulos, K., Xie, J., Rosenfeld, E., Wright, A., Sharma, A., Geng, X., Wang, Y., Kwei, S., Pan, R., Zhang, Y., Wang, G., Liu, X., Yeung, C., Cole, E., Rosenberg, A., Yang, Z., Chen, P., Polovets, G., Nair, P., Saxena, R., Smith, J., Yiin Chang, S., Mahendru, A., Grant, S., Iyer, A., Cai, I., McGiffin, J., Shen, J., Walton, A., Girgis, A., Woodman, O., Ke, R., Kwong, M., Rouillard, L., Rao, J., Li, Z., Xu, Y., Prost, F., Zou, C., Ji, Z., Magni, A., Liechty, T., Calian, D. A., Ramachandran, D., Krivokon, I., Huang, H., Chen, T., Hauth, A., Ilić, A., Xi, W., Lim, H., Ion, V.-D., Moradi, P., Toksoz-Exley, M., Bullard, K., Allamanis, M., Yang, X., Wang, S., Hong, Z., Gergely, A., Li, C., Mittal, B., Kovalev, V., Ungureanu, V., Labanowski, J., Wassenberg, J., Lacasse, N., Cideron, G., Dević, P., Marsden, A., Nguyen, L., Fink, M., Zhong, Y., Kiyono, T., Ivanov, D., Ma, S., Bain, M., Yalasangi, K., She, J., Petrushkina, A., Lunayach, M., Bromberg, C., Hodgkinson, S., Meshram, V., Vlasic, D., Kyker, A., Xu, S., Stanway, J., Yang, Z., Zhao, K., Tung, M., Odoom, S., Fujii, Y., Gilmer, J., Kim, E., Halim, F., Le, Q., Bohnet, B., El-Sayed, S., Neyshabur, B., Reynolds, M., Reich, D., Xu, Y., Moreira, E., Sharma, A., Liu, Z., Hosseini, M. J., Raisinghani, N., Su, Y., Lao, N., Formoso, D., Gelmi, M., Gueta, A., Dey, T., Gribovskaya, E., Čevd, D., Mudgal, S., Bingham, G., Wang, J., Kumar, A., Cullum, A., Han, F., Bousmalis, K., Cedillo, D., Chu, G., Magay, V., Michel, P., Hlavnova, E., Calandriello, D., Ariafar, S., Yao, K., Sehwag, V., Vezar, A., Lago, A. D., Zhu, Z., Rubenstein, P. K., Porter, A., Baddepudi, A., Riva, O., Istina, M. D., Yeh, C.-K., Li, Z., Howard, A., Jha, N., Chen, J., de Liedekerke, R., Ahmed, Z., Rodriguez, M., Bhatia, T., Wang, B., Elqursh, A., Klinghoffer, D., Chen, P., Kohli, P., I. T., Zhang, W., Nado, Z., Chen, J., Chen, M., Zhang, G., Singh, A., Hillier, A., Lebron, F., Tao, Y., Liu, T., Dulac-Arnold, G., Zhang, J., Narayan, S., Liu, B., Firat, O., Bhowmick, A., Liu, B., Zhang, H., Zhang, Z., Rotival, G., Howard, N., Sinha, A., Grushetsky, A., Beyret, B., Gopalakrishnan, K., Zhao, J., He, K., Payrits, S., Nabulsi, Z., Zhang, Z., Chen, W., Lee, E., Fallen, N., Gollapudi, S., Zhou, A., Pavetić, F., Köppe, T., Huang, S., Pasumarthi, R., Fernando, N., Fischer, F., Ćurko, D., Gao, Y., Svensson, J., Stone, A., Qureshi, H., Sinha, A., Kulshreshtha, A., Matysiak, M., Mao, J., Saroufim, C., Faust, A., Duan, Q., Fidel, G., Katircioglu, K., Kaufman, R. L., Shah, D., Kong, W., Bapna, A., Weisz, G., Dunleavy, E., Dutta, P., Liu, T., Chaabouni, R., Parada, C., Wu, M., Belias, A., Bissacco, A., Fort, S., Xiao, L., Huot, F., Knut-

- sen, C., Blau, Y., Li, G., Prendki, J., Love, J., Chow, Y., Charoenpanit, P., Shimokawa, H., Coriou, V., Gregor, K., Izo, T., Akula, A., Pinto, M., Hahn, C., Paulus, D., Guo, J., Sharma, N., Hsieh, C.-J., Chukwuka, A., Hashimoto, K., Rauschmayr, N., Wu, L., Angermueller, C., Wang, Y., Gerlach, S., Pliskin, M., Mirylenka, D., Ma, M., Baugher, L., Gale, B., Bijwadia, S., Rakićević, N., Wood, D., Park, J., Chang, C.-C., Seal, B., Tar, C., Krasowiak, K., Song, Y., Stephanov, G., Wang, G., Maggioni, M., Lin, S. X., Wu, F., Paul, S., Jiang, Z., Agrawal, S., Piot, B., Feng, A., Kim, C., Doshi, T., Lai, J., Chuqiao, Xu, Vikram, S., Chelba, C., Krause, S., Zhuang, V., Rae, J., Denk, T., Collister, A., Weerts, L., Luo, X., Lu, Y., Garnes, H., Gupta, N., Spitz, T., Hassidim, A., Liang, L., Shafran, I., Humphreys, P., Vassigh, K., Wallis, P., Shejwalkar, V., Perez-Nieves, N., Hornung, R., Tan, M., Westberg, B., Ly, A., Zhang, R., Farris, B., Park, J., Kosik, A., Cankara, Z., Maksai, A., Xu, Y., Cassirer, A., Caelles, S., Abdolmaleki, A., Chiang, M., Fabrikant, A., Shetty, S., He, L., Giménez, M., Hashemi, H., Panthaplackel, S., Kulizhskaya, Y., Deshmukh, S., Pighin, D., Alazard, R., Jindal, D., Noury, S., S. P. K., Qin, S., Dotiwalla, X., Spencer, S., Babaeizadeh, M., Chen, B. J., Mehta, V., Lees, J., Leach, A., Koanantakool, P., Akolzin, I., Comanescu, R., Ahn, J., Svyatkovskiy, A., Mustafa, B., D'Ambrosio, D., Garlapati, S. M. R., Lamblin, P., Agarwal, A., Song, S., Sessa, P. G., Coquiot, P., Maggs, J., Masoom, H., Pitta, D., Wang, Y., Morris-Suzuki, P., Porter, B., Jia, J., Dudek, J., R. R., Paduraru, C., Ansell, A., Bolukbasi, T., Lu, T., Ganeshan, R., Wang, Z., Griffiths, H., Benenson, R., He, Y., Swirhun, J., Papamakarios, G., Chawla, A., Sengupta, K., Wang, Y., Milutinovic, V., Mordatch, I., Jia, Z., Smith, J., Ng, W., Nigam, S., Young, M., Vušak, E., Hechtman, B., Goenka, S., Zipori, A., Ayoub, K., Popat, A., Acharya, T., Yu, L., Bloxwich, D., Song, H., Roit, P., Li, H., Boag, A., Nayakanti, N., Chandra, B., Ding, T., Mehta, A., Hope, C., Zhang, J., Shtacher, I. H., Badola, K., Nakashima, R., Sozanschi, A., Comşa, I., Žužul, A., Caveness, E., Odell, J., Watson, M., de Cesare, D., Lippe, P., Lockhart, D., Verma, S., Chen, H., Sun, S., Zhuo, L., Shah, A., Gupta, P., Muzio, A., Niu, N., Zait, A., Singh, A., Gaba, M., Ye, F., Ramachandran, P., Saleh, M., Popa, R. A., Dubey, A., Liu, F., Javanmardi, S., Epstein, M., Hemsley, R., Green, R., Ranka, N., Cohen, E., Fu, C. K., Ghemawat, S., Borovik, J., Martens, J., Chen, A., Shyam, P., Pinto, A. S., Yang, M.-H., Tifrea, A., Du, D., Gong, B., Agarwal, A., Kim, S., Frank, C., Shah, S., Song, X., Deng, Z., Mikhalap, A., Chatziprimou, K., Chung, T., Creswell, T., Zhang, S., Jun, Y., Lebsack, C., Truong, W., Andačić, S., Yona, I., Fornoni, M., Rong, R., Toropov, S., Soudagar, A. S., Audibert, A., Zaiem, S., Abbas, Z., Rusu, A., Potluri, S., Weng, S., Kementsietsidis, A., Tsitsulin, A., Peng, D., Ha, N., Jain, S., Latkar, T., Ivanov, S., McLean, C., GP, A., Venkataraman, R., Liu, C., Krishnan, D., D'sa, J., Yogev, R., Collins, P., Lee, B., Ho, L., Doersch, C., Yona, G., Gao, S., Ferreira, F. T., Ozturel, A., Muckenhirn, H., Zheng, C., Balasubramaniam, G., Bansal, M., van den Driessche, G., Eiger, S., Haykal, S., Misra, V., Goyal, A., Martins, D., Leung, G., Valfridsson, J., Flynn, F., Bishop, W., Pang, C., Halpern, Y., Yu, H., Moore, L., Yuvein, Zhu, Thiagarajan, S., Drori, Y., Xiao, Z., Dery, L., Jagerman, R., Lu, J., Ge, E., Aggarwal, V., Khare, A., Tran, V., Elyada, O., Alet, F., Rubin, J., Chou, I., Tian, D., Bai, L., Chan, L., Lew, L., Misiunas, K., Bilal, T., Ray, A., Raghuram, S., Castro-Ros, A., Carpenter, V., Zheng, C., Kilgore, M., Broder, J., Xue, E., Kallakuri, P., Dua, D., Yuen, N., Chien, S., Schultz, J., Agrawal, S., Tsarfaty, R., Hu, J., Kannan, A., Marcus, D., Kothari, N., Sun, B., Horn, B., Bošnjak, M., Naeem, F., Hirsch, D., Chiang, L., Fang, B., Han, J., Wang, Q., Hora, B., He, A., Lučić, M., Changpinyo, B., Tripathi, A., Youssef, J., Kwak, C., Schlattner, P., Graves, C., Leblond, R., Zeng, W., Andreassen, A., Rasskin, G., Song, Y., Cao, E., Oh, J., Hoffman, M., Skut, W., Zhang, Y., Stritar, J., Cai, X., Khanna, S., Wang, K., Sharma, S., Reisswig, C., Jun, Y., Prasad, A., Sholokhova, T., Singh, P., Rosenthal, A. G., Ruoss, A., Beaufays, F., Kirmani, S., Chen, D., Schalkwyk, J., Herzig, J., Kim, B., Jacob, J., Vincent, D., Reyes, A. N., Balazevic, I., Hussenot, L., Schneider, J., Barnes, P., Castro, L., Babbula, S. R., Green, S., Cabi, S., Duduta, N., Driess, D., Galt, R., Velan, N., Wang, J., Jiao, H., Mauger, M., Phan, D., Patel, M., Galić, V., Chang, J., Marcus, E., Harvey, M., Salazar, J., Dabir, E., Sheth, S. S., Mandhane, A., Sedghi, H., Willcock, J., Zandieh, A., Prabhakara, S., Amini, A., Miech, A., Stone, V., Nicosia, M., Niemczyk, P., Xiao, Y., Kim, L., Kwasiborski, S., Verma, V., Oflazer, A. M., Hirschschall, C., Sung, P., Liu, L., Everett, R., Bakker, M., Ágoston Weisz, Wang, Y., Sampathkumar, V., Shaham, U., Xu, B., Altun, Y., Wang, M., Saeki, T., Chen, G., Taropa, E., Vasanth, S., Austin, S., Huang, L., Petrovic, G., Dou, Q., Golovin, D., Rozhdestvenskiy, G., Culp, A., Wu, W., Sano, M., Jain, D., Proskurnia, J., Cevey, S., Ruiz, A. C., Patil, P., Mirzazadeh, M., Ni, E., Snaider, J., Fan, L., Fréchette, A., Pierigiovanni, A., Iqbal, S., Lee, K., Fantacci, C., Xing, J., Wang, L., Irpan, A., Raposo, D., Luan, Y., Chen, Z., Ganapathy, H., Hui, K., Nie, J., Guyon, I., Ge, H., Vij, R., Zheng, H., Lee, D., Castaño, A., Baatarsukh, K., Ibagon, G., Chronopoulou, A., FitzGerald, N., Viswanadha, S., Huda, S., Moroshko, R., Stoyanov, G., Kolhar, P., Vaucher, A., Watts, I., Kuncoro, A., Michalewski, H., Kambala, S., Batsaikhan, B.-O., Andreev, A., Jurenka, I., Le, M., Chen, Q., Jishi, W. A., Chakera, S., Chen, Z., Kini, A., Yadav, V., Siddhant, A., Labzovsky, I., Lakshminarayanan, B., Bostock, C. G., Botadra, P., Anand, A., Bishop, C., Conway-Rahman, S., Agarwal, M., Donchev, Y., Singhal, A., de Chaumont Quitry, F., Ponomareva, N., Agrawal, N., Ni, B.,

- Krishna, K., Samsikova, M., Karro, J., Du, Y., von Glehn, T., Lu, C., Choquette-Choo, C. A., Qin, Z., Zhang, T., Li, S., Tyam, D., Mishra, S., Lowe, W., Ji, C., Wang, W., Faruqui, M., Slone, A., Dalibard, V., Narayanaswamy, A., Lambert, J., Manzagol, P.-A., Karliner, D., Bolt, A., Lobov, I., Kusupati, A., Ye, C., Yang, X., Zen, H., George, N., Bhutani, M., Lacombe, O., Riachi, R., Bansal, G., Soh, R., Gao, Y., Yu, Y., Yu, A., Nottage, E., Rojas-Esponda, T., Noraky, J., Gupta, M., Kotikalapudi, R., Chang, J., Deur, S., Graur, D., Mossin, A., Farnese, E., Figueira, R., Moufarek, A., Huang, A., Zochbauer, P., Ingram, B., Chen, T., Wu, Z., Puigdomènech, A., Rechis, L., Yu, D., Padmanabhan, S. G. S., Zhu, R., Ling Ko, C., Banino, A., Daruki, S., Selvan, A., Bhaswar, D., Diaz, D. H., Su, C., Scellato, S., Brennan, J., Han, W., Chung, G., Agrawal, P., Khandelwal, U., Sim, K. C., Lustman, M., Ritter, S., Guu, K., Xia, J., Jain, P., Wang, E., Hill, T., Rossini, M., Kostelac, M., Misiunas, T., Sabne, A., Kim, K., Iscen, A., Wang, C., Leal, J., Sreevatsa, A., Evci, U., Warmuth, M., Joshi, S., Suo, D., Lottes, J., Honke, G., Jou, B., Karp, S., Hu, J., Sahni, H., Taïga, A. A., Kong, W., Ghosh, S., Wang, R., Pavagadhi, J., Axelsson, N., Grigorev, N., Siegler, P., Lin, R., Wang, G., Parisotto, E., Maddineni, S., Subudhi, K., Ben-David, E., Pochernina, E., Keller, O., Avrahami, T., Yuan, Z., Mehta, P., Liu, J., Yang, S., Kan, W., Lee, K., Funkhouser, T., Cheng, D., Shi, H., Sharma, A., Kelley, J., Eyal, M., Malkov, Y., Tallec, C., Bahat, Y., Yan, S., Xintian, Wu, Lindner, D., Wu, C., Caciularu, A., Luo, X., Jenatton, R., Zaman, T., Bi, Y., Kornakov, I., Mallya, G., Ikeda, D., Karo, I., Singh, A., Evans, C., Netrapalli, P., Nallatamby, V., Tian, I., Assael, Y., Raunak, V., Carbune, V., Bica, I., Madmoni, L., Cattle, D., Grover, S., Somandepalli, K., Lall, S., Vázquez-Reina, A., Patana, R., Mu, J., Talluri, P., Tran, M., Aggarwal, R., Skerry-Ryan, R., Xu, J., Burrows, M., Pan, X., Yvinec, E., Lu, D., Zhang, Z., Nguyen, D. D., Mu, H., Barcik, G., Ran, H., Beltrone, L., Choromanski, K., Kharrat, D., Albanie, S., Purser-haskell, S., Bieber, D., Zhang, C., Wang, J., Hudson, T., Zhang, Z., Fu, H., Maurer, J., Bateni, M. H., Maschinot, A., Wang, B., Zhu, M., Pillai, A., Weyand, T., Liu, S., Akerlund, O., Bertsch, F., Premachandran, V., Jin, A., Roulet, V., de Boursac, P., Mittal, S., Ndebele, N., Karadzhov, G., Ghalebikesabi, S., Liang, R., Wu, A., Cong, Y., Ghelani, N., Singh, S., Fatemi, B., Warren, Chen, Kwong, C., Kologanov, A., Li, S., Song, R., Kuang, C., Miryoosefi, S., Webster, D., Wendt, J., Socala, A., Su, G., Mendonça, A., Gupta, A., Li, X., Tsai, T., Qiong, Hu, Kang, K., Chen, A., Girgin, S., Xian, Y., Lee, A., Ramsden, N., Baker, L., Elish, M. C., Krayvanova, V., Joshi, R., Simsa, J., Yang, Y.-Y., Ambroszczyk, P., Ghosh, D., Kar, A., Shang-guan, Y., Yamamori, Y., Akulov, Y., Brock, A., Tang, H., Vashishtha, S., Munoz, R., Steiner, A., Andra, K., Eppens, D., Feng, Q., Kobayashi, H., Goldshtein, S., Mahdy, M. E., Wang, X., Jilei, Wang, Killam, R., Kwiatkowski, T., Koppurapu, K., Zhan, S., Jia, C., Bendebury, A., Luo, S., Recasens, A., Knight, T., Chen, J., Patel, M., Li, Y., Withbroe, B., Weesner, D., Bhatia, K., Ren, J., Eisenbud, D., Songhori, E., Sun, Y., Choma, T., Kementsietsidis, T., Manning, L., Roark, B., Farhan, W., Feng, J., Tatineni, S., Cobon-Kerr, J., Li, Y., Hendricks, L. A., Noble, I., Breaux, C., Kushman, N., Peng, L., Xue, F., Tobin, T., Rogers, J., Lipschultz, J., Alberti, C., Vlaskin, A., Dehghani, M., Sharma, R., Warkentin, T., Lee, C.-Y., Uria, B., Juan, D.-C., Chandorkar, A., Sheftel, H., Liu, R., Davoodi, E., Pigem, B. D. B., Dhamdhere, K., Ross, D., Hoech, J., Mahdih, M., Liu, L., Li, Q., McCafferty, L., Liu, C., Mircea, M., Song, Y., Savant, O., Saade, A., Cherry, C., Hellendoorn, V., Goyal, S., Pucciarelli, P., Torres, D. V., Yahav, Z., Lee, H., Sjoesund, L. L., Kirov, C., Chang, B., Ghoshal, D., Li, L., Baechler, G., Pereira, S., Sainath, T., Boral, A., Grewe, D., Halumi, A., Phu, N. M., Shen, T., Ribeiro, M. T., Varma, D., Kaskasoli, A., Feinberg, V., Potti, N., Kahn, J., Wisniewski, M., Mohamed, S., Hrafnkelsson, A. M., Shahriari, B., Lespiau, J.-B., Patel, L., Yeung, L., Paine, T., Mei, L., Ramirez, A., Shivanna, R., Zhong, L., Woodward, J., Tubone, G., Khan, S., Chen, H., Nielsen, E., Ionescu, C., Prabhu, U., Gao, M., Wang, Q., Augenstein, S., Subramaniam, N., Chang, J., Iliopoulos, F., Luo, J., Khan, M., Kuo, W., Teplyashin, D., Perot, F., Kilpatrick, L., Globerson, A., Yu, H., Siddiqui, A., Sukhanov, N., Kandoor, A., Gupta, U., Andreetto, M., Ambar, M., Kim, D., Wesolowski, P., Perrin, S., Limonchik, B., Fan, W., Stephan, J., Stewart-Binks, I., Kappedal, R., He, T., Cogan, S., Datta, R., Zhou, T., Ye, J., Kieliger, L., Ramalho, A., Kastner, K., Mentzer, F., Ko, W.-J., Suggala, A., Zhou, T., Butt, S., Strejček, H., Belenki, L., Venugopalan, S., Ling, M., Eltyshv, E., Deng, Y., Kovacs, G., Raghavachari, M., Dai, H., Schuster, T., Schwarcz, S., Nguyen, R., Nguyen, A., Buttimore, G., Mallick, S. B., Gandhe, S., Benjamin, S., Jastrzebski, M., Yan, L., Basu, S., Apps, C., Edkins, I., Allingham, J., Odisho, I., Kocisky, T., Zhao, J., Xue, L., Reddy, A., Anastasiou, C., Atias, A., Redmond, S., Milan, K., Heess, N., Schmit, H., Dafoe, A., Andor, D., Gangwani, T., Dragan, A., Zhang, S., Kachra, A., Wu, G., Xue, S., Aydin, K., Liu, S., Zhou, Y., Malihi, M., Wu, A., Gopal, S., Schumann, C., Stys, P., Wang, A., Olšák, M., Liu, D., Schallhart, C., Mao, Y., Brady, D., Xu, H., Mery, T., Sitawarin, C., Velusamy, S., Cobley, T., Zhai, A., Walder, C., Katz, N., Jawahar, G., Kulkarni, C., Yang, A., Paszke, A., Wang, Y., Damoc, B., Borsos, Z., Smith, R., Li, J., Gupta, M., Kaphishnikov, A., Prakash, S., Luisier, F., Agarwal, R., Grathwohl, W., Chen, K., Han, K., Mehta, N., Over, A., Azizi, S., Meng, L., Santo, N. D., Zheng, K., Shapiro, J., Petrovski, I., Hui, J., Ghafouri, A., Snoek, J., Qin, J., Jordan, M., Sikora, C., Malmaud, J., Kuang, Y., Świetlik, A., Sang, R., Shi, C., Li, L., Rosen-

- berg, A., Zhao, S., Crawford, A., Peter, J.-T., Lei, Y., Garcia, X., Le, L., Wang, T., Amelot, J., Orr, D., Kacham, P., Alon, D., Tyen, G., Arora, A., Lyon, J., Kurakin, A., Ly, M., Guidroz, T., Yan, Z., Panigrahy, R., Xu, P., Kogohara, T., Cheng, Y., Noland, E., Lee, J., Lee, J., Yip, C., Wang, M., Nehoran, E., Bykovsky, A., Shan, Z., Bhagatwala, A., Yan, C., Tan, J., Garrido, G., Ethier, D., Hurley, N., Vesom, G., Chen, X., Qiao, S., Nayyar, A., Walker, J., Sandhu, P., Rosca, M., Swisher, D., Dekhtarev, M., Dillon, J., Muraru, G.-C., Tragut, M., Myaskovsky, A., Reid, D., Velic, M., Xiao, O., George, J., Brand, M., Li, J., Yu, W., Gu, S., Deng, X., Aubet, F.-X., Yeganeh, S. H., Alcober, F., Smith, C., Cohn, T., McKinney, K., Tschannen, M., Sampath, R., Cheon, G., Luo, L., Liu, L., Orbay, J., Peng, H., Botea, G., Zhang, X., Yoon, C., Magalhaes, C., Stradomski, P., Mackinnon, I., Hemingray, S., Venkatesan, K., May, R., Kim, J., Druinsky, A., Ye, J., Xu, Z., Huang, T., Abdallah, J. A., Dostmohamed, A., Fellingner, R., Munkhdalai, T., Maurya, A., Garst, P., Zhang, Y., Krikun, M., Bucher, S., Veerubhotla, A. S., Liu, Y., Li, S., Gupta, N., Adamek, J., Chen, H., Orlando, B., Zaks, A., van Amersfoort, J., Camp, J., Wan, H., Choe, H., Wu, Z., Olszewska, K., Yu, W., Vadali, A., Scholz, M., Freitas, D. D., Lin, J., Hua, A., Liu, X., Ding, F., Zhou, Y., Severson, B., Tsihlias, K., Yang, S., Spalink, T., Yerram, V., Pankov, H., Blevins, R., Vargas, B., Jauhari, S., Miecznikowski, M., Zhang, M., Kumar, S., Farabet, C., Lan, C. L., Flennerhag, S., Bitton, Y., Ma, A., Bražinskas, A., Collins, E., Ahuja, N., Kudugunta, S., Bortsova, A., Giang, M., Zhu, W., Chi, E., Lundberg, S., Stern, A., Puttagunta, S., Xiong, J., Wu, X., Pande, Y., Jhindal, A., Murphy, D., Clark, J., Brockschmidt, M., Deines, M., McKee, K. R., Bahir, D., Shen, J., Truong, M., McDuff, D., Gesmundo, A., Rosseel, E., Liang, B., Caluwaerts, K., Hamrick, J., Kready, J., Cassin, M., Ingale, R., Lao, L., Pollom, S., Ding, Y., He, W., Bellot, L., Iljazi, J., Boppana, R. S., Han, S., Thompson, T., Khalifa, A., Bulanova, A., Mitrevski, B., Pang, B., Cooney, E., Shi, T., Coaguila, R., Yakar, T., Ranzato, M., Momchev, N., Rawles, C., Charles, Z., Maeng, Y., Zhang, Y., Bansal, R., Zhao, X., Albert, B., Yuan, Y., Vijayanarasimhan, S., Hirsch, R., Ramasesh, V., Vodrahalli, K., Wang, X., Gupta, A., Strouse, D., Ni, J., Patel, R., Taubman, G., Huo, Z., Gharibian, D., Monteiro, M., Lam, H., Vasudevan, S., Chaudhary, A., Albuquerque, I., Gupta, K., Riedel, S., Hegde, C., Ruderman, A., György, A., Wainwright, M., Chaugule, A., Ayan, B. K., Levinboim, T., Shleifer, S., Kalley, Y., Mirokni, V., Rao, A., Radhakrishnan, P., Hartford, J., Wu, J., Zhu, Z., Bertolini, F., Xiong, H., Serrano, N., Tomlinson, H., Ott, M., Chang, Y., Graham, M., Li, J., Liang, M., Long, X., Borgeaud, S., Ahmad, Y., Grills, A., Mincu, D., Izzard, M., Liu, Y., Xie, J., O'Bryan, L., Ponda, S., Tong, S., Liu, M., Malkin, D., Salama, K., Chen, Y., Anil, R., Rao, A., Swavely, R., Bilenko, M., Anderson, N., Tan, T., Xie, J., Wu, X., Yu, L., Vinyals, O., Ryabtsev, A., Dangovski, R., Baumli, K., Keysers, D., Wright, C., Ashwood, Z., Chan, B., Shtefan, A., Guo, Y., Bapna, A., Soricut, R., Pecht, S., Ramos, S., Wang, R., Cai, J., Trinh, T., Barham, P., Friso, L., Stickgold, E., Ding, X., Shakeri, S., Ardila, D., Briakou, E., Culliton, P., Raveret, A., Cui, J., Saxton, D., Roy, S., Azizi, J., Yin, P., Loher, L., Bunner, A., Choi, M., Ahmed, F., Li, E., Li, Y., Dai, S., Elabd, M., Ganapathy, S., Agrawal, S., Hua, Y., Kunkle, P., Rajayogam, S., Ahuja, A., Conmy, A., Vasiloff, A., Beak, P., Yew, C., Mudigonda, J., Wydrowski, B., Blanton, J., Wang, Z., Dauphin, Y., Xu, Z., Polacek, M., Chen, X., Hu, H., Sho, P., Kunesch, M., Manshadi, M. H., Rutherford, E., Li, B., Hsiao, S., Barr, I., Tudor, A., Kecman, M., Nagrani, A., Pchelin, V., Sundermeyer, M., S. A. P., Karmarkar, A., Gao, Y., Chole, G., Bachem, O., Gao, I., BC, A., Dibb, M., Verzett, M., Hernandez-Campos, F., Lunts, Y., Johnson, M., Trapani, J. D., Koster, R., Brusilovsky, I., Xiong, B., Mohabey, M., Ke, H., Zou, J., Sabolić, T., Campos, V., Palowitch, J., Morris, A., Qiu, L., Ponnuramu, P., Li, F., Sharma, V., Sodhia, K., Tekelioglu, K., Chuklin, A., Yenugula, M., Gemzer, E., Strinopoulos, T., El-Husseini, S., Wang, H., Zhong, Y., Leurent, E., Natsev, P., Wang, W., Mahaarachchi, D., Zhu, T., Peng, S., Alabed, S., Lee, C.-C., Brohan, A., Szlam, A., Oh, G., Kovsharov, A., Lee, J., Wong, R., Barnes, M., Thornton, G., Gimeno, F., Levy, O., Sevenich, M., Johnson, M., Mallinson, J., Dadashi, R., Wang, Z., Ren, Q., Lahoti, P., Dhar, A., Feldman, J., Zheng, D., Ulrich, T., Panait, L., Blokzijl, M., Baetu, C., Matak, J., Harlalka, J., Shah, M., Marian, T., von Dincklage, D., Du, C., Ley-Wild, R., Brownfield, B., Schumacher, M., Stuken, Y., Noghabi, S., Gupta, S., Ren, X., Malmi, E., Weissenberger, F., Huergo, B., Bauza, M., Lampe, T., Douillard, A., Seyedhosseini, M., Frostig, R., Ghahramani, Z., Nguyen, K., Krishnakumar, K., Ye, C., Gupta, R., Nazari, A., Geirhos, R., Shaw, P., Eleryan, A., Damen, D., Palomaki, J., Xiao, T., Wu, Q., Yuan, Q., Meadowlark, P., Bilotti, M., Lin, R., Sridhar, M., Schroeder, Y., Chung, D.-W., Luo, J., Strohmaier, T., Liu, T., Zheng, A., Emond, J., Wang, W., Lampinen, A., Fukuzawa, T., Campbell-Ajala, F., Roy, M., Lee-Thorp, J., Wang, L., Naim, I., Tony, ên, N., Bensky, G., Gupta, A., Rogozińska, D., Fu, J., Pillai, T. S., Veličković, P., Drath, S., Neubeck, P., Tulsyan, V., Klimovskiy, A., Metzler, D., Stevens, S., Yeh, A., Yuan, J., Yu, T., Zhang, K., Go, A., Tsang, V., Xu, Y., Wan, A., Galatzer-Levy, I., Sobell, S., Toki, A., Salesky, E., Zhou, W., Antognini, D., Douglas, S., Wu, S., Lelkes, A., Kim, F., Cavallaro, P., Salazar, A., Liu, Y., Besley, J., Refice, T., Jia, Y., Li, Z., Sokolik, M., Kannan, A., Simon, J., Chick, J., Aharon, A., Gandhi, M., Daswani, M., Amiri, K., Birodgar, V., Ittycheriah, A., Grabowski, P., Chang, O., Sutton, C., Zhixin, Lai, Telang, U., Sargsyan, S., Jiang, T., Hoffmann, R., Brichtova, N., Hessel, M., Halcrow, J., Jerome, S.,

- Brown, G., Tomala, A., Buchatskaya, E., Yu, D., Menon, S., Moreno, P., Liao, Y., Zayats, V., Tang, L., Mah, S., Shenoy, A., Siegman, A., Hadian, M., Kwon, O., Tu, T., Khajehnouri, N., Foley, R., Haghani, P., Wu, Z., Keshava, V., Gupta, K., Bruguier, T., Yao, R., Karmon, D., Zintgraf, L., Wang, Z., Piqueras, E., Jung, J., Brennan, J., Machado, D., Giustina, M., Tessler, M., Lee, K., Zhang, Q., Moore, J., Daugaard, K., Frömmgen, A., Beattie, J., Zhang, F., Kasenberg, D., Geri, T., Qin, D., Tomar, G. S., Ouyang, T., Yu, T., Zhou, L., Mathews, R., Davis, A., Li, Y., Gupta, J., Yates, D., Deng, L., Kemp, E., Joung, G.-Y., Vassilvitskii, S., Guo, M., LV, P., Dopson, D., Lachgar, S., McConnaughey, L., Choudhury, H., Dena, D., Cohen, A., Ainslie, J., Levi, S., Gopavarapu, P., Zablotskaia, P., Vallet, H., Bahargam, S., Tang, X., Tomasev, N., Dyer, E., Balle, D., Lee, H., Bono, W., Mendez, J. G., Zubov, V., Yang, S., Rendulic, I., Zheng, Y., Hogue, A., Pundak, G., Leith, R., Bhoopchand, A., Han, M., Žanić, M., Schaul, T., Delakis, M., Iyer, T., Wang, G., Singh, H., Abdelhamed, A., Thomas, T., Brahma, S., Dib, H., Kumar, N., Zhou, W., Bai, L., Mishra, P., Sun, J., Anklin, V., Sukkerd, R., Agubuzu, L., Briukhov, A., Gulati, A., Sieb, M., Pardo, F., Nasso, S., Chen, J., Zhu, K., Sosea, T., Goldin, A., Rush, K., Hombaiah, S. A., Noever, A., Zhou, A., Haves, S., Phuong, M., Ades, J., ting Chen, Y., Yang, L., Pagadora, J., Bileschi, S., Cotruta, V., Saputro, R., Pramanik, A., Ammirati, S., Garrette, D., Villela, K., Blyth, T., Akbulut, C., Jha, N., Rrustemi, A., Wongpanich, A., Nagpal, C., Wu, Y., Rivière, M., Kishchenko, S., Srinivasan, P., Chen, A., Sinha, A., Pham, T., Jia, B., Hennigan, T., Bakalov, A., Attaluri, N., Garmon, D., Rodriguez, D., Wegner, D., Jia, W., Senter, E., Fiedel, N., Petek, D., Liu, Y., Hardin, C., Lehri, H. T., Carreira, J., Smoot, S., Prasetya, M., Akazawa, N., Stefanoiu, A., Ho, C.-H., Angelova, A., Lin, K., Kim, M., Chen, C., Sieniek, M., Li, A., Guo, T., Baltateanu, S., Tafti, P., Wunder, M., Olmert, N., Shukla, D., Shen, J., Kovelamudi, N., Venkatraman, B., Neel, S., Thoppilan, R., Connor, J., Benzing, F., Stjerngren, A., Ghiasi, G., Polozov, A., Howland, J., Weber, T., Chiu, J., Girirajan, G. P., Terzis, A., Wang, P., Li, F., Shalom, Y. B., Tewari, D., Denton, M., Aharoni, R., Kalb, N., Zhao, H., Zhang, J., Filos, A., Rahtz, M., Jain, L., Fan, C., Rodrigues, V., Wang, R., Shin, R., Austin, J., Ring, R., Sanchez-Vargas, M., Hassen, M., Kessler, I., Alon, U., Zhang, G., Chen, W., Ma, Y., Si, X., Hou, L., Mirhoseini, A., Wilson, M., Bacon, G., Roelofs, B., Shu, L., Vasudevan, G., Adler, J., Dwornik, A., Terzi, T., Lawlor, M., Askham, H., Bernico, M., Dong, X., Hidey, C., Kilgour, K., Liu, G., Bhupatiraju, S., Leonhard, L., Zuo, S., Talukdar, P., Wei, Q., Severyn, A., Listík, V., Lee, J., Tripathi, A., Park, S., Matias, Y., Liu, H., Ruiz, A., Jayaram, R., Tolins, J., Marcenac, P., Wang, Y., Seybold, B., Prior, H., Sharma, D., Weber, J., Sirotenko, M., Sung, Y., Du, D., Pavlick, E., Zinke, S., Freitag, M., Dylla, M., Arenas, M. G., Potikha, N., Goldman, O., Tao, C., Chhaparia, R., Voitovich, M., Dogra, P., Ražnatović, A., Tsai, Z., You, C., Johnson, O., Tucker, G., Gu, C., Yoo, J., Majzoubi, M., Gabeur, V., Raad, B., Rhodes, R., Kolipaka, K., Howard, H., Sampemane, G., Li, B., Asawaroengchai, C., Nguyen, D., Zhang, C., Cour, T., Yu, X., Fu, Z., Jiang, J., Huang, P.-S., Surita, G., Iturrate, I., Karov, Y., Collins, M., Baeuml, M., Fuchs, F., Shetty, S., Ramaswamy, S., Ebrahimi, S., Guo, Q., Shar, J., Barthmaron, G., Addepalli, S., Richter, B., Cheng, C.-Y., Rives, E., Zheng, F., Griesser, J., Dikkala, N., Zeldes, Y., Safarli, I., Das, D., Srivastava, H., Khan, S. M., Li, X., Pandey, A., Markeeva, L., Belov, D., Yan, Q., Rybiński, M., Chen, T., Nawhal, M., Quinn, M., Govindaraj, V., York, S., Roberts, R., Garg, R., Godbole, N., Abernethy, J., Das, A., Thiet, L. N., Tompson, J., Nham, J., Vats, N., Caine, B., Helmholz, W., Pongetti, F., Ko, Y., An, J., Hu, C. H., Ling, Y.-C., Pawar, J., Leland, R., Kinoshita, K., Khawaja, W., Selvi, M., Ie, E., Sinopalnikov, D., Proleev, L., Tripuraneni, N., Bevilacqua, M., Lee, S., Sanford, C., Suh, D., Tran, D., Dean, J., Baumgartner, S., Heitkaemper, J., Gubbi, S., Toutanova, K., Xu, Y., Thekkath, C., Rong, K., Jain, P., Xie, A., Virin, Y., Li, Y., Litchev, L., Powell, R., Bharti, T., Kraft, A., Hua, N., Ikonomidis, M., Hitron, A., Kumar, S., Matthey, L., Bridgers, S., Lax, L., Malhi, I., Skopek, O., Gupta, A., Cao, J., Rasquinha, M., Pöder, S., Stokowiec, W., Roth, N., Li, G., Sander, M., Kessinger, J., Jain, V., Loper, E., Park, W., Yarom, M., Cheng, L., Guruganesh, G., Rao, K., Li, Y., Barros, C., Sushkov, M., Ferng, C.-S., Shah, R., Aharoni, O., Kumar, R., McConnell, T., Li, P., Wang, C., Pereira, F., Swanson, C., Jamil, F., Xiong, Y., Vijayakumar, A., Shroff, P., Soparkar, K., Gu, J., Soares, L. B., Wang, E., Majmundar, K., Wei, A., Bailey, K., Kassner, N., Kawamoto, C., Žužić, G., Gomes, V., Gupta, A., Guzman, M., Dasgupta, I., Bai, X., Pan, Z., Piccinno, F., Vogel, H. N., Ponce, O., Hutter, A., Chang, P., Jiang, P.-P., Gog, I., Ionescu, V., Manyika, J., Pedregosa, F., Ragan, H., Behrman, Z., Mullins, R., Devin, C., Pyne, A., Gawde, S., Chadwick, M., Gu, Y., Tavakkol, S., Twigg, A., Goyal, N., Elue, N., Goldie, A., Venkatachary, S., Fei, H., Feng, Z., Ritter, M., Leal, I., Dasari, S., Sun, P., Rochman, A. R., O'Donoghue, B., Liu, Y., Sproch, J., Chen, K., Clay, N., Petrov, S., Sidhwani, S., Mihailescu, I., Panagopoulos, A., Piergiovanni, A., Bai, Y., Powell, G., Karkhanis, D., Yacovone, T., Mitrichev, P., Kovac, J., Uthus, D., Yazdanbakhsh, A., Amos, D., Zheng, S., Zhang, B., Miao, J., Ramabhadran, B., Radpour, S., Thakoor, S., Newlan, J., Lang, O., Jankowski, O., Bharadwaj, S., Sarr, J.-M., Ashraf, S., Mondal, S., Yan, J., Rawat, A. S., Velury, S., Kochanski, G., Eccles, T., Och, F., Sharma, A., Mahintorabi, E., Gurney, A., Muir, C., Cohen, V., Thakur, S., Bloniarz, A., Mujika, A., Pritzel, A., Caron, P., Rahman, A., Lang, F., Onoe, Y., Sirkovic, P., Hoover, J., Jian, Y., Duque, P., Narayanan, A., Soergel,

- D., Haig, A., Maggiore, L., Buch, S., Dean, J., Figotin, I., Karpov, I., Gupta, S., Zhou, D., Huang, M., Vaswani, A., Semturs, C., Shivakumar, K., Watanabe, Y., Rajendran, V. K., Lu, E., Hou, Y., Ye, W., Vashishth, S., Nti, N., Sakenas, V., Ni, D., DeCarlo, D., Bendersky, M., Bagri, S., Cano, N., Peake, E., Tokumine, S., Godbole, V., Guía, C., Lando, T., Selo, V., Ellis, S., Tarlow, D., Gillick, D., Epasto, A., Jonnalagadda, S. R., Wei, M., Xie, M., Taly, A., Paganini, M., Sundararajan, M., Toyama, D., Yu, T., Petrova, D., Pappu, A., Agrawal, R., Buthpitiya, S., Frye, J., Buschmann, T., Crocker, R., Tagliasacchi, M., Wang, M., Huang, D., Perel, S., Wieder, B., Kazawa, H., Wang, W., Cole, J., Gupta, H., Golan, B., Bang, S., Kulkarni, N., Franko, K., Liu, C., Reid, D., Dalmia, S., Whang, J., Cen, K., Sundaram, P., Ferret, J., Isik, B., Ionita, L., Sun, G., Shekhawat, A., Mohammad, M., Pham, P., Huang, R., Raman, K., Zhou, X., Mcilroy, R., Myers, A., Peng, S., Scott, J., Covington, P., Erell, S., Joshi, P., Oliveira, J. G., Noy, N., Nasir, T., Walker, J., Axelrod, V., Dozat, T., Han, P., Chu, C.-T., Weinstein, E., Shukla, A., Chandrakaladharan, S., Poklukur, P., Li, B., Jin, Y., Eruvbetine, P., Hansen, S., Dabush, A., Jacovi, A., Phatale, S., Zhu, C., Baker, S., Shomrat, M., Xiao, Y., Pouget-Abadie, J., Zhang, M., Wei, F., Song, Y., King, H., Huang, Y., Zhu, Y., Sun, R., Franco, J. V., Lin, C.-C., Arora, S., Hui, Li, Xia, V., Vilnis, L., Schain, M., Alarakya, K., Prince, L., Phillips, A., Habtegebriel, C., Xu, L., Gui, H., Ontanon, S., Aroyo, L., Gill, K., Lu, P., Katariya, Y., Madeka, D., Krishnan, S., Raghvendra, S. S., Freedman, J., Tay, Y., Menghani, G., Choy, P., Shetty, N., Abolafia, D., Kulkiansky, D., Chou, E., Lichtarge, J., Burke, K., Coleman, B., Guo, D., Jin, L., Bhattacharya, I., Langston, V., Li, Y., Kotecha, S., Yakubovich, A., Chen, X., Petrov, P., Powell, T., He, Y., Quick, C., Garg, K., Hwang, D., Lu, Y., Bhojanapalli, S., Kjems, K., Mehran, R., Archer, A., van Hasselt, H., Balakrishna, A., Kearns, J., Guo, M., Riesa, J., Sazanovich, M., Gao, X., Sauer, C., Yang, C., Sheng, X., Jimma, T., Gansbeke, W. V., Nikolaev, V., Wei, W., Millican, K., Zhao, R., Snyder, J., Bolelli, L., O'Brien, M., Xu, S., Xia, F., Yuan, W., Neelakantan, A., Barker, D., Yadav, S., Kirkwood, H., Ahmad, F., Wee, J., Grimstad, J., Wang, B., Wiethoff, M., Settle, S., Wang, M., Blundell, C., Chen, J., Duvarney, C., Hu, G., Ronneberger, O., Lee, A., Li, Y., Chakladar, A., Butryna, A., Evangelopoulos, G., Desjardins, G., Kanerva, J., Wang, H., Nowak, A., Li, N., Loo, A., Khurshudov, A., Shafey, L. E., Baddi, N., Lenc, K., Razeghi, Y., Lieber, T., Sinha, A., Ma, X., Su, Y., Huang, J., Ushio, A., Klimczak-Plucińska, H., Mohamed, K., Chen, J., Osindero, S., Ginzburg, S., Lamprou, L., Bashlovkina, V., Tran, D.-H., Khodaei, A., Anand, A., Di, Y., Eskander, R., Vuyyuru, M. R., Liu, J., Kamath, A., Goldenberg, R., Bellaiche, M., Pluto, J., Rosgen, B., Mansoor, H., Wong, W., Ganesh, S., Bailey, E., Baird, S., Deutsch, D., Baek, J., Jia, X., Lee, C., Friesen, A., Braun, N., Lee, K., Panda, A., Hernandez, S. M., Williams, D., Liu, J., Liang, E., Autef, A., Pitler, E., Jain, D., Kirk, P., Bunyan, O., Elias, J. S., Yin, T., Reid, M., Pope, A., Putikhin, N., Samanta, B., Guadarrama, S., Kim, D., Rowe, S., Valentine, M., Yan, G., Salcianu, A., Silver, D., Song, G., Singh, R., Ye, S., DeBalsi, H., Merey, M. A., Ofek, E., Webson, A., Mourad, S., Kakarla, A., Lattanzi, S., Roy, N., Sluzhaev, E., Butterfield, C., Tonioni, A., Waters, N., Kopalle, S., Chase, J., Cohan, J., Rao, G. R., Berry, R., Voznesensky, M., Hu, S., Chiafullo, K., Chikkerur, S., Scrivener, G., Zheng, I., Wiesner, J., Macherey, W., Lillicrap, T., Liu, F., Walker, B., Welling, D., Davies, E., Huang, Y., Ren, L., Shabat, N., Agostini, A., Iinuma, M., Zelle, D., Sathyanarayana, R., D'olimpio, A., Redshaw, M., Ginsberg, M., Murthy, A., Geller, M., Matejovicova, T., Chakrabarti, A., Julian, R., Chan, C., Hu, Q., Jarrett, D., Agarwal, M., Challagundla, J., Li, T., Tata, S., Ding, W., Meng, M., Dai, Z., Vezzani, G., Garg, S., Bulian, J., Jasarevic, M., Cai, H., Rajamani, H., Santoro, A., Hartmann, F., Liang, C., Perz, B., Jindal, A., Bu, F., Seo, S., Poplin, R., Goedeckemeyer, A., Ghazi, B., Khadke, N., Liu, L., Mather, K., Zhang, M., Shah, A., Chen, A., Wei, J., Shivam, K., Cao, Y., Cho, D., Scarpati, A. S., Moffitt, M., Barbu, C., Jurin, I., Chang, M.-W., Liu, H., Zheng, H., Dave, S., Kaeser-Chen, C., Yu, X., Abdagic, A., Gonzalez, L., Huang, Y., Zhong, P., Schmid, C., Petrini, B., Wertheim, A., Zhu, J., Nguyen, H., Ji, K., Zhou, Y., Zhou, T., Feng, F., Cohen, R., Rim, D., Phal, S. M., Georgiev, P., Brand, A., Ma, Y., Li, W., Gupta, S., Wang, C., Dubov, P., Tarbouriech, J., Majumder, K., Li, H., Rink, N., Suman, A., Guo, Y., Sun, Y., Nair, A., Xu, X., Elhawaty, M., Cabrera, R., Han, G., Eisenschlos, J., Bai, J., Li, Y., Bansal, Y., Sellam, T., Khan, M., Nguyen, H., Mao-Jones, J., Parotsidis, N., Marcus, J., Fan, C., Zimmermann, R., Kochinski, Y., Graesser, L., Behbahani, F., Caceres, A., Riley, M., Kane, P., Lefdal, S., Willoughby, R., Vicol, P., Wang, L., Zhang, S., Gill, A., Liang, Y., Prasad, G., Mariooryad, S., Kazemi, M., Wang, Z., Muralidharan, K., Voigtlaender, P., Zhao, J., Zhou, H., D'Souza, N., Mavalankar, A., Arnold, S., Young, N., Sarvana, O., Lee, C., Nasr, M., Zou, T., Kim, S., Haas, L., Patel, K., Bulut, N., Parkinson, D., Biles, C., Kalashnikov, D., To, C. M., Kumar, A., Austin, J., Greve, A., Zhang, L., Goel, M., Li, Y., Yaroshenko, S., Chang, M., Jindal, A., Clark, G., Taitelbaum, H., Johnson, D., Roval, O., Ko, J., Mohananey, A., Schuler, C., Dodhia, S., Li, R., Osawa, K., Cui, C., Xu, P., Shah, R., Huang, T., Gruzewska, E., Clement, N., Verma, M., Sercinoglu, O., Qian, H., Shah, V., Yamaguchi, M., Modi, A., Kosakai, T., Strohmman, T., Zeng, J., Gunel, B., Qian, J., Tarango, A., Jastrzebski, K., David, R., Shan, J., Schuh, P., Lad, K., Gierke, W., Madhavan, M., Chen, X., Kurzeja, M., Santamaria-Fernandez, R., Chen, D., Cordell, A., Chervonyi, Y., Garcia, F., Kannen, N., Perot, V., Ding, N., Cohen-Ganor,

- S., Lavrenko, V., Wu, J., Evans, G., dos Santos, C. N., Sewak, M., Brown, A., Hard, A., Puigcerver, J., Zheng, Z., Liang, Y., Gladchenko, E., Ingle, R., First, U., Ser-manet, P., Magister, C., Velimirović, M., Reddi, S., Ricco, S., Agustsson, E., Adam, H., Levine, N., Gaddy, D., Holtmann-Rice, D., Wang, X., Sathe, A., Roy, A. G., Bratanič, B., Carin, A., Mehta, H., Bonacina, S., Cao, N. D., Finkelstein, M., Rieser, V., Wu, X., Althé, F., Scandinaro, D., Li, L., Vieillard, N., Sethi, N., Tanzer, G., Xing, Z., Wang, S., Bhatia, P., Citovsky, G., Anthony, T., Lin, S., Shi, T., Jakobovits, S., Gibson, G., Apte, R., Lee, L., Chen, M., Byravan, A., Maniatis, P., Webster, K., Dai, A., Chen, P.-C., Pan, J., Fadeeva, A., Gleicher, Z., Luong, T., and Bhumihar, N. K. 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>.
- Han, Z., Mankikar, M., Michael, J., and Wang, Z. Search-time data contamination, 2025. URL <https://arxiv.org/abs/2508.13180>.
- Hayes, J., Shumailov, I., Choquette-Choo, C. A., Jagielski, M., Kaissis, G., Nasr, M., Annamalai, M. S. M. S., Miresghallah, N., Shilov, I., Meeus, M., et al. Exploring the limits of strong membership inference attacks on large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021a.
- Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset, 2021b. URL <https://arxiv.org/abs/2103.03874>.
- 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>.
- 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., Vinyals, O., Rae, J., and Sifre, L. An empirical analysis of compute-optimal large language model training. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c1e2faff6f588870935f114ebe04a3e5-Paper-Conference.pdf.
- Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.
- Hu, S., Liu, X., Han, X., Zhang, X., He, C., Zhao, W., Lin, Y., Ding, N., Ou, Z., Zeng, G., Liu, Z., and Sun, M. Predicting emergent abilities with infinite resolution evaluation, 2024. URL <https://arxiv.org/abs/2310.03262>.
- Huang, J., Yang, D., and Potts, C. Demystifying verbatim memorization in large language models. *arXiv preprint arXiv:2407.17817*, 2024.
- Ibrahim, A., Thérien, B., Gupta, K., Richter, M. L., Anthony, Q. G., Belilovsky, E., Lesort, T., and Rish, I. Simple and scalable strategies to continually pre-train large language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL <https://openreview.net/forum?id=DimPeeCxKO>.
- Jagielski, M., Nasr, M., Lee, K., Choquette-Choo, C. A., Carlini, N., and Tramer, F. Students parrot their teachers: Membership inference on model distillation. *Advances in Neural Information Processing Systems*, 36, 2024.
- Jain, N., Han, K., Gu, A., Li, W.-D., Yan, F., Zhang, T., Wang, S., Solar-Lezama, A., Sen, K., and Stoica, I. Live-codebench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=chfJJYC3iL>.
- Jang, J., Ye, S., Yang, S., Shin, J., Han, J., Kim, G., Choi, S. J., and Seo, M. Towards continual knowledge learning of language models, 2022. URL <https://arxiv.org/abs/2110.03215>.
- Jiang, M., Liu, K. Z., Zhong, M., Schaeffer, R., Ouyang, S., Han, J., and Koyejo, S. Investigating data contamination for pre-training language models, 2024. URL <https://arxiv.org/abs/2401.06059>.
- Jiang, M., Liu, K. Z., and Koyejo, S. A missing testbed for LLM pre-training membership inference attacks. In *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*, 2025. URL <https://openreview.net/forum?id=HzHUxo6KzE>.

- Jin, X., Zhang, D., Zhu, H., Xiao, W., Li, S.-W., Wei, X., Arnold, A., and Ren, X. Lifelong pretraining: Continually adapting language models to emerging corpora. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4764–4780, 2022.
- Kandpal, N., Pillutla, K., Oprea, A., Kairouz, P., Choquette-Choo, C. A., and Xu, Z. User inference attacks on large language models. *arXiv preprint arXiv:2310.09266*, 2023.
- 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. *arXiv preprint arXiv:2001.08361*, 2020.
- Kocigit, M. Y. The impact of post-training on data contamination. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, 2025. URL <https://openreview.net/forum?id=SnPmnvMQ3k>.
- Kocigit, M. Y., Briakou, E., Deutsch, D., Luo, J., Cherry, C., and Freitag, M. Overestimation in llm evaluation: A controlled large-scale study on data contamination’s impact on machine translation. In *Forty-second International Conference on Machine Learning*, 2025.
- Kong, Z., Chowdhury, A. R., and Chaudhuri, K. Can membership inferencing be refuted? *arXiv preprint arXiv:2303.03648*, 2023.
- Li, H., Chen, Y., Wang, S., He, Y., Mehrabi, N., Gupta, R., and Ren, X. Diagnosing memorization in chain-of-thought reasoning, one token at a time. *arXiv preprint arXiv:2508.02037*, 2025.
- Li, M., Wang, J., Wang, J., and Neel, S. Mope: Model perturbation-based privacy attacks on language models. *arXiv preprint arXiv:2310.14369*, 2023.
- Li, Y., Guo, Y., Guerin, F., and Lin, C. An open-source data contamination report for large language models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 528–541, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.30. URL <https://aclanthology.org/2024.findings-emnlp.30/>.
- Liu, K. Z., Choquette-Choo, C. A., Jagielski, M., Kairouz, P., Koyejo, S., Liang, P., and Papernot, N. Language models may verbatim complete text they were not explicitly trained on. *arXiv preprint arXiv:2503.17514*, 2025.
- Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.
- Lu, X., Li, X., Cheng, Q., Ding, K., Huang, X.-J., and Qiu, X. Scaling laws for fact memorization of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 11263–11282, 2024.
- Magar, I. and Schwartz, R. Data contamination: From memorization to exploitation. *arXiv preprint arXiv:2203.08242*, 2022.
- Maini, P., Yaghini, M., and Papernot, N. Dataset inference: Ownership resolution in machine learning. *arXiv preprint arXiv:2104.10706*, 2021.
- Maini, P., Jia, H., Papernot, N., and Dziedzic, A. Llm dataset inference: Did you train on my dataset? *arXiv preprint arXiv:2406.06443*, 2024.
- Mangaokar, N., Hooda, A., Li, Z., Malin, B. A., Fawaz, K., Jha, S., Prakash, A., and Chowdhury, A. R. What really is a member? discrediting membership inference via poisoning. *arXiv preprint arXiv:2506.06003*, 2025.
- Mattern, J., Mireshghallah, F., Jin, Z., Schölkopf, B., Sachan, M., and Berg-Kirkpatrick, T. Membership inference attacks against language models via neighbourhood comparison. *arXiv preprint arXiv:2305.18462*, 2023.
- Matton, A., Sherborne, T., Aumiller, D., Tommasone, E., Alizadeh, M., He, J., Ma, R., Voisin, M., Gilsonan-McMahon, E., and Gallé, M. On leakage of code generation evaluation datasets. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13215–13223, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.772. URL <https://aclanthology.org/2024.findings-emnlp.772/>.
- Meeus, M., Jain, S., Rei, M., and de Montjoye, Y.-A. Inherent challenges of post-hoc membership inference for large language models. *arXiv preprint arXiv:2406.17975*, 2024.
- Muennighoff, N., Rush, A., Barak, B., Le Scao, T., Tazi, N., Piktus, A., Pyysalo, S., Wolf, T., and Raffel, C. A. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376, 2023.
- Nallapati, R., Zhou, B., dos Santos, C., Gülçehre, Ç., and Xiang, B. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In Riezler, S. and Goldberg, Y. (eds.), *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 280–290,

- Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/K16-1028. URL <https://aclanthology.org/K16-1028/>.
- Ni, S., Kong, X., Li, C., Hu, X., Xu, R., Zhu, J., and Yang, M. Training on the benchmark is not all you need, 2025. URL <https://arxiv.org/abs/2409.01790>.
- Nie, F., Liu, K. Z., Wang, Z., Sun, R., Liu, W., Shi, W., Yao, H., Zhang, L., Ng, A. Y., Zou, J., Koyejo, S., Choi, Y., Liang, P., and Muennighoff, N. Uq: Assessing language models on unsolved questions, 2025. URL <https://arxiv.org/abs/2508.17580>.
- OpenAI, :, Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., Iftimie, A., Karpenko, A., Passos, A. T., Neitz, A., Prokofiev, A., Wei, A., Tam, A., Bennett, A., Kumar, A., Saraiva, A., Vallone, A., Duberstein, A., Kondrich, A., Mishchenko, A., Applebaum, A., Jiang, A., Nair, A., Zoph, B., Ghorbani, B., Rossen, B., Sokolowsky, B., Barak, B., McGrew, B., Minaiev, B., Hao, B., Baker, B., Houghton, B., McKinzie, B., Eastman, B., Lugaresi, C., Bassin, C., Hudson, C., Li, C. M., de Bourcy, C., Voss, C., Shen, C., Zhang, C., Koch, C., Orsinger, C., Hesse, C., Fischer, C., Chan, C., Roberts, D., Kappler, D., Levy, D., Selsam, D., Dohan, D., Farhi, D., Mely, D., Robinson, D., Tsipras, D., Li, D., Oprica, D., Freeman, E., Zhang, E., Wong, E., Proehl, E., Cheung, E., Mitchell, E., Wallace, E., Ritter, E., Mays, E., Wang, F., Such, F. P., Raso, F., Leoni, F., Tsimpourlas, F., Song, F., von Lohmann, F., Sulit, F., Salmon, G., Parascandolo, G., Chabot, G., Zhao, G., Brockman, G., Leclerc, G., Salman, H., Bao, H., Sheng, H., Andrin, H., Bagherinezhad, H., Ren, H., Lightman, H., Chung, H. W., Kivlichan, I., O’Connell, I., Osband, I., Gilaberte, I. C., Akkaya, I., Kostrikov, I., Sutskever, I., Kofman, I., Pachocki, J., Lennon, J., Wei, J., Harb, J., Twore, J., Feng, J., Yu, J., Weng, J., Tang, J., Yu, J., Candela, J. Q., Palermo, J., Parish, J., Heidecke, J., Hallman, J., Rizzo, J., Gordon, J., Uesato, J., Ward, J., Huizinga, J., Wang, J., Chen, K., Xiao, K., Singhal, K., Nguyen, K., Cobbe, K., Shi, K., Wood, K., Rimbach, K., Gu-Lemberg, K., Liu, K., Lu, K., Stone, K., Yu, K., Ahmad, L., Yang, L., Liu, L., Maksin, L., Ho, L., Fedus, L., Weng, L., Li, L., McCallum, L., Held, L., Kuhn, L., Kondraciuk, L., Kaiser, L., Metz, L., Boyd, M., Trebacz, M., Joglekar, M., Chen, M., Tintor, M., Meyer, M., Jones, M., Kaufer, M., Schwarzer, M., Shah, M., Yatbaz, M., Guan, M. Y., Xu, M., Yan, M., Glaese, M., Chen, M., Lampe, M., Malek, M., Wang, M., Fradin, M., McClay, M., Pavlov, M., Wang, M., Wang, M., Murati, M., Bavarian, M., Rohaninejad, M., McAleese, N., Chowdhury, N., Chowdhury, N., Ryder, N., Tezak, N., Brown, N., Nachum, O., Boiko, O., Murk, O., Watkins, O., Chao, P., Ashbourne, P., Izmailov, P., Zhokhov, P., Dias, R., Arora, R., Lin, R., Lopes, R. G., Gaon, R., Miyara, R., Leike, R., Hwang, R., Garg, R., Brown, R., James, R., Shu, R., Cheu, R., Greene, R., Jain, S., Altman, S., Toizer, S., Toyer, S., Miserendino, S., Agarwal, S., Hernandez, S., Baker, S., McKinney, S., Yan, S., Zhao, S., Hu, S., Santurkar, S., Chaudhuri, S. R., Zhang, S., Fu, S., Papay, S., Lin, S., Balaji, S., Sanjeev, S., Sidor, S., Broda, T., Clark, A., Wang, T., Gordon, T., Sanders, T., Patwardhan, T., Sottiaux, T., Degry, T., Dimson, T., Zheng, T., Garipov, T., Stasi, T., Bansal, T., Creech, T., Peterson, T., Eloundou, T., Qi, V., Kosaraju, V., Monaco, V., Pong, V., Fomenko, V., Zheng, W., Zhou, W., McCabe, W., Zaremba, W., Dubois, Y., Lu, Y., Chen, Y., Cha, Y., Bai, Y., He, Y., Zhang, Y., Wang, Y., Shao, Z., and Li, Z. Openai o1 system card, 2024a. URL <https://arxiv.org/abs/2412.16720>.
- OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Koscic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O’Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo,

- G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorný, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2024b. URL <https://arxiv.org/abs/2303.08774>.
- Oren, Y., Meister, N., Chatterji, N. S., Ladhak, F., and Hashimoto, T. Proving test set contamination in black-box language models. In *The Twelfth International Conference on Learning Representations*, 2023.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- Parmar, J., Satheesh, S., Patwary, M., Shoneyi, M., and Catanzaro, B. Reuse, don’t retrain: A recipe for continued pretraining of language models, 2024. URL <https://arxiv.org/abs/2407.07263>.
- Penedo, G., Kydliček, 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>.
- Porian, T., Wortsman, M., Jitsev, J., Schmidt, L., and Carmon, Y. Resolving discrepancies in compute-optimal scaling of language models. *Advances in Neural Information Processing Systems*, 37:100535–100570, 2024.
- Qian, K., Wan, S., Tang, C., Wang, Y., Zhang, X., Chen, M., and Yu, Z. Varbench: Robust language model benchmarking through dynamic variable perturbation, 2024. URL <https://arxiv.org/abs/2406.17681>.
- Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. SQuAD: 100,000+ questions for machine comprehension of text. In Su, J., Duh, K., and Carreras, X. (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL <https://aclanthology.org/D16-1264/>.
- Reuel, A., Bucknall, B., Casper, S., Fist, T., Soder, L., Aarne, O., Hammond, L., Ibrahim, L., Chan, A., Wills, P., Anderljung, M., Garfinkel, B., Heim, L., Trask, A., Mukobi, G., Schaeffer, R., Baker, M., Hooker, S., Solaiman, I., Luccioni, S., Rajkumar, N., Moës, N., Ladish, J., Bau, D., Bricman, P., Guha, N., Newman, J., Bengio, Y., South, T., Pentland, A., Koyejo, S., Kochenderfer, M., and Trager, R. Open problems in technical AI governance. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL <https://openreview.net/forum?id=lnO4qFMiS0>. Survey Certification.
- Riddell, M., Ni, A., and Cohan, A. Quantifying contamination in evaluating code generation capabilities of language models. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14116–14137, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.761. URL <https://aclanthology.org/2024.acl-long.761/>.
- Roberts, M., Thakur, H., Herlihy, C., White, C., and Dooley, S. To the cutoff... and beyond? a longitudinal perspective on LLM data contamination. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=m2NVG4Htxs>.
- Sablayrolles, A., Douze, M., Schmid, C., Ollivier, Y., and Jégou, H. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, pp. 5558–5567. PMLR, 2019.
- Sainz, O., Campos, J. A., García-Ferrero, I., Etxaniz, J., de Lacalle, O. L., and Agirre, E. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL <https://openreview.net/forum?id=KivNpBsfAS>.
- Sainz, O., García-Ferrero, I., Jacovi, A., Ander Campos, J., Elazar, Y., Agirre, E., Goldberg, Y., Chen, W.-L., Chim, J., Choshen, L., D’Amico-Wong, L., Dell, M., Fan, R.-Z., Golchin, S., Li, Y., Liu, P., Pahwa, B., Prabhu, A., Sharma,

- S., Silcock, E., Solonko, K., Stap, D., Surdeanu, M., Tseng, Y.-M., Udandara, V., Wang, Z., Xu, R., and Yang, J. Data contamination report from the 2024 CONDA shared task. In Sainz, O., García Ferrero, I., Agirre, E., Ander Campos, J., Jacovi, A., Elazar, Y., and Goldberg, Y. (eds.), *Proceedings of the 1st Workshop on Data Contamination (CONDA)*, pp. 41–56, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.conda-1.4. URL <https://aclanthology.org/2024.conda-1.4/>.
- Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M., and Backes, M. MI-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*, 2018.
- Sardana, N., Portes, J., Doubov, S., and Frankle, J. Beyond chinchilla-optimal: Accounting for inference in language model scaling laws. In *International Conference on Machine Learning*, pp. 43445–43460. PMLR, 2024.
- Schaeffer, R. Pretraining on the test set is all you need, 2023. URL <https://arxiv.org/abs/2309.08632>.
- Schaeffer, R., Robertson, Z., Boopathy, A., Khona, M., Pistunova, K., Rocks, J. W., Fiete, I. R., Gromov, A., and Koyejo, S. Double descent demystified: Identifying, interpreting & ablating the sources of a deep learning puzzle. In *The Third Blogpost Track at ICLR 2024*.
- Schaeffer, R., Levi, N., Miranda, B., and Koyejo, S. Pretraining scaling laws for generative evaluations of language models, 2025a. URL <https://arxiv.org/abs/2509.24012>.
- Schaeffer, R., Liu, K., Miranda, B., Ahmed, A. M., Mireshghallah, N., and Koyejo, S. The contamination paradox: Why test set leakage can be both potent and negligible. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, 2025b. URL <https://openreview.net/forum?id=Ajr8YtfhVO>.
- Schaeffer, R., Miranda, B., Kazdan, J., Liu, K., Ahmed, A. M., Mireshghallah, N., and Koyejo, S. Causally quantifying the effect of test set contamination on generative benchmarks. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, 2025c. URL <https://openreview.net/forum?id=RsmjshBEDP>.
- Schaeffer, R., Schoelkopf, H., Miranda, B., Mukobi, G., Madan, V., Ibrahim, A., Bradley, H., Biderman, S., and Koyejo, S. Why has predicting downstream capabilities of frontier AI models with scale remained elusive? In *Forty-second International Conference on Machine Learning*, 2025d. URL <https://openreview.net/forum?id=I1NtlLvJal>.
- Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., and Zettlemoyer, L. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=zWqr3MQUNs>.
- Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.
- Shuai, X., Wang, Y., Wu, Y., Jiang, X., and Ren, X. Scaling law for language models training considering batch size, 2024. URL <https://arxiv.org/abs/2412.01505>.
- Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In Yarowsky, D., Baldwin, T., Korhonen, A., Livescu, K., and Bethard, S. (eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL <https://aclanthology.org/D13-1170/>.
- Tirumala, K., Markosyan, A., Zettlemoyer, L., and Aghajanyan, A. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35: 38274–38290, 2022.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models, 2023. URL <https://arxiv.org/abs/2307.09288>.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

- Wang, X., Antoniadou, A., Elazar, Y., Amayuelas, A., Al-balak, A., Zhang, K., and Wang, W. Y. Generalization v.s. memorization: Tracing language models’ capabilities back to pretraining data. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=IQxBDLmVpT>.
- Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022a.
- Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners, 2022b. URL <https://arxiv.org/abs/2109.01652>.
- Xia, C. S., Deng, Y., and Zhang, L. Top leaderboard ranking= top coding proficiency, always? evoeval: Evolving coding benchmarks via llm. *arXiv preprint arXiv:2403.19114*, 2024.
- Xie, C., Huang, Y., Zhang, C., Yu, D., Chen, X., Lin, B. Y., Li, B., Ghazi, B., and Kumar, R. On memorization of large language models in logical reasoning, 2025. URL <https://arxiv.org/abs/2410.23123>.
- Xu, C., Guan, S., Greene, D., and Kechadi, M.-T. Benchmark data contamination of large language models: A survey, 2024a. URL <https://arxiv.org/abs/2406.04244>.
- Xu, F., Hao, Q., Zong, Z., Wang, J., Zhang, Y., Wang, J., Lan, X., Gong, J., Ouyang, T., Meng, F., Shao, C., Yan, Y., Yang, Q., Song, Y., Ren, S., Hu, X., Li, Y., Feng, J., Gao, C., and Li, Y. Towards large reasoning models: A survey of reinforced reasoning with large language models, 2025. URL <https://arxiv.org/abs/2501.09686>.
- Xu, R., Wang, Z., Fan, R.-Z., and Liu, P. Benchmarking benchmark leakage in large language models, 2024b. URL <https://arxiv.org/abs/2404.18824>.
- 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, 2025a. URL <https://arxiv.org/abs/2505.09388>.
- Yang, S., Chiang, W.-L., Zheng, L., Gonzalez, J. E., and Stoica, I. Rethinking benchmark and contamination for language models with rephrased samples. *arXiv preprint arXiv:2311.04850*, 2023.
- Yang, Z., Lin, H., He, Y., Xu, J., Sun, Z., Liu, S., Wang, P., Yu, Z., and Liang, Q. Rethinking the effects of data contamination in code intelligence, 2025b. URL <https://arxiv.org/abs/2506.02791>.
- Yao, F., Zhuang, Y., Sun, Z., Xu, S., Kumar, A., and Shang, J. Data contamination can cross language barriers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17864–17875, 2024.
- Yeom, S., Giacomelli, I., Fredrikson, M., and Jha, S. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pp. 268–282. IEEE, 2018.
- Yildiz, C., Ravichandran, N. K., Sharma, N., Bethge, M., and Ermiş, B. Investigating continual pretraining in large language models: Insights and implications, 2025. URL <https://arxiv.org/abs/2402.17400>.
- Zarifzadeh, S., Liu, P., and Shokri, R. Low-cost high-power membership inference attacks, 2023.
- Zhang, H., Da, J., Lee, D., Robinson, V., Wu, C., Song, W., Zhao, T., Raja, P., Zhuang, C., Slack, D., Lyu, Q., Hendryx, S., Kaplan, R., Lunati, M., and Yue, S. A careful examination of large language model performance on grade school arithmetic, 2024a. URL <https://arxiv.org/abs/2405.00332>.
- Zhang, J., Das, D., Kamath, G., and Tramèr, F. Membership inference attacks cannot prove that a model was trained on your data. *arXiv preprint arXiv:2409.19798*, 2024b.
- Zhang, Z., Liu, R., Liu, A., Liu, X., Gao, X., and Sun, H. Dynamic benchmark construction for evaluating large language models on real-world codes, 2025. URL <https://arxiv.org/abs/2508.07180>.
- Zhou, K., Zhu, Y., Chen, Z., Chen, W., Zhao, W. X., Chen, X., Lin, Y., Wen, J.-R., and Han, J. Don’t make your llm an evaluation benchmark cheater, 2023. URL <https://arxiv.org/abs/2311.01964>.

A. Related Work

Data Contamination and its Consequences Test set contamination, where benchmark data is included in pretraining corpora, is widely recognized as a threat to valid model evaluation, as it can lead to inflated performance metrics. Numerous survey and position papers have documented the various ways contamination can occur and have called for routine audits and transparent reporting for all benchmarks (Sainz et al., 2023; 2024; Deng et al., 2024a; Xu et al., 2024a; Reuel et al., 2025). Empirical studies of large web-scale datasets have confirmed significant overlap and duplication between training and test sets (Dodge et al., 2021). Research focused on ensuring benchmark integrity has identified multiple ways that language models might “cheat” on evaluations if contamination is not properly managed (Zhou et al., 2023; Dong et al., 2024). For instance, analyses of popular mathematics benchmarks have revealed signals of data leakage and potential overfitting (Zhang et al., 2024a). Ongoing community efforts and open-source audits continue to measure the extent of contamination across different models and datasets (Li et al., 2024). The risks extend beyond evaluation integrity; scaling studies indicate that poisoning risks increase with model size, as larger models learn harmful behaviors from minuscule amounts of poisoned data far more rapidly than smaller models, underscoring the necessity of robust data curation (Bowen et al., 2025). As a cautionary illustration, Schaeffer (2023) demonstrated that pretraining on the test set is a trivial path to strong benchmark performance, reinforcing the importance of rigorous decontamination and auditing.

Controlled Contamination During Pretraining A line of research directly investigates the causal effects of contamination by intentionally adding benchmark data to pretraining corpora and observing the results. Magar & Schwartz (2022) interleaved task-specific datasets into a general text corpus during pretraining, varying the duplication rate of the leaked examples. They differentiated between “memorization” (storing examples) and “exploitation” (using stored examples to boost test scores), finding that both model size and the number of repetitions increased exploitation. Jiang et al. (2024) pretrained models from scratch on corpora containing either only the inputs (“text-only”) or the full input-output pairs (“ground-truth”) of benchmark examples, sweeping the contamination frequency. They observed significant performance gains when ground-truth pairs were used and showed that simple n-gram-based detection methods could be bypassed by paraphrasing or partial data leaks. The problem also transcends language barriers; Yao et al. (2024) demonstrated a cross-lingual contamination channel where continuing to pretrain a model on non-English translations of English benchmarks led to material improvements on the original English tests, a form of contamination that

string-matching would not detect. At a larger scale, Bordt et al. (2025) varied the repetition count of leaked examples, model size (up to 1.6B parameters), and the total training token budget, finding that performance scales predictably with size and repetition. They also showed that sufficiently long training on abundant unique data could mitigate or even reverse the effects of earlier contamination. In the context of machine translation, Kocyigit et al. (2025) injected source-target pairs into the pretraining data of 1B and 8B parameter models, quantifying significant overestimation in BLEU scores, with larger models and low-resource languages showing more pronounced effects. Together, these causal intervention studies provide clear evidence that language models memorize and leverage benchmark data when it is present during pretraining.

Repeated Data and Memorization Dynamics Closely related is the study of memorization dynamics, particularly how repeated data affects model behavior. Hernandez et al. (2022) trained models where a small portion of the data was repeated many times, observing strong double descent phenomena (Advani et al., 2020; Belkin et al., 2019; Adlam & Pennington, 2020; Bordelon et al., 2020; Schaeffer et al.) and showing that repeating just 0.1% of tokens 100 times could significantly degrade generalization. Studies tracking exact-sequence memorization have shown that larger models not only memorize more content and at a faster rate but also forget less over the course of training (Tirumala et al., 2022). Carlini et al. (2023) quantified log-linear relationships between verbatim generation and model size, data duplication count, and prompt length. Other work has explored the feasibility of “forecasting” whether a model will memorize a specific string, finding that accurate prediction is possible but may require a substantial portion of the target model’s pretraining compute (Biderman et al., 2023). Beyond explicit repetition, Duan et al. (2025) discovered “latent memorization”, where memorized sequences that are not obvious at a final checkpoint can persist and be revealed later, posing privacy risks. Finally, memorization appears to be task-dependent: Wang et al. (2025) observed stronger memorization for knowledge-intensive QA, whereas machine translation and mathematical reasoning demonstrated greater novelty. Memorization also interacts with logical reasoning; using dynamically generated puzzles, Xie et al. (2025) showed that models could be fine-tuned to perfectly memorize training examples yet failed on slight variations, even as their genuine reasoning abilities also improved, revealing a complex balance between the two.

Detecting and Proving Contamination Another significant area of research focuses on detecting or proving test set contamination in existing models. Oren et al. (2023) and Ni et al. (2025) proposed statistical tests with provable control over false positives by testing if a benchmark’s

canonical ordering is statistically privileged over random shuffles. Shi et al. (2024) introduced Min- k %-Prob to determine if a sequence likely appeared in pretraining using only black-box probabilities. Two related works from Golchin & Surdeanu (2023; 2024) frame detection as a multiple-choice "quiz" and use temporal information about model training windows versus benchmark release dates, a strategy also used by Roberts et al. (2024). Broader audits have aimed to quantify leakage and decontamination across a wide range of tasks and models (Xu et al., 2024b; Deng et al., 2024b; Li et al., 2024), while Yang et al. (2023) showed that rephrasing benchmark questions can often bypass n-gram filters. In the domain of code generation, Riddell et al. (2024) quantified contamination in popular coding benchmarks and connected the degree of overlap to performance differences. Matton et al. (2024) cataloged various channels for leakage and released a dataset (LBPP) to help mitigate these issues. Complementing these audits, Yang et al. (2025b) systematically tested fine-grained contamination scenarios in code intelligence across different model types, finding that paired contamination substantially affects LLMs under a pretraining-plus-inference paradigm but has limited effect under a pretrain–finetune–inference pipeline. Other work has also provided instruments for detecting the origins of chain-of-thought sequences (Li et al., 2025).

Preventing Test Set Contamination The growing concern over contamination has spurred the development of new methods for creating benchmarks. These include dynamically updated benchmarks (Jain et al., 2025; Xia et al., 2024; Zhang et al., 2025; Qian et al., 2024) and private or restricted-access benchmarks (Zhang et al., 2024a; Glazer et al., 2025). Recently, Nie et al. (2025) released a benchmark consisting of unsolved scientific questions, which, by its nature, prevents models from being trained on the correct solutions.

Retrieval- and Agent-Time Contamination As model evaluation evolves from static prompting to using tool-augmented agents, the risk of contamination expands. Han et al. (2025) introduced search-time contamination, where an agent retrieves benchmark questions and answers from the web during its evaluation process, which can artificially inflate its performance.

Membership Inference Attacks The field of Membership Inference Attacks (MIA) aims to determine if a specific data point was used to train a model, given only access to the model itself (Shokri et al., 2017). This is highly relevant to contamination, as detection can be viewed as an MIA problem. While the MIA literature is extensive in computer vision (Yeom et al., 2018; Salem et al., 2018; Sablayrolles et al., 2019; Jagielski et al., 2024), it has more recently been applied to language models (Carlini et al., 2021; Zarifzadeh

et al., 2023; Shi et al., 2024; Mattern et al., 2023; Li et al., 2023). However, progress in sequence-level MIA for language models has been complicated by issues such as flawed evaluations (Meeus et al., 2024; Zhang et al., 2024b; Jiang et al., 2025). Duan et al. (2024) argue that membership can be inherently "blurry" for natural language. Das et al. (2024) and (Meeus et al., 2024) report that existing MIA testbeds suffer from distribution shifts. Kong et al. (2023) refute MIAs with a theoretical attack, and Liu et al. (2025) and Mangaokar et al. (2025) demonstrate fundamental limitations and exploits of n-gram based methods. Due to these challenges, recent work explores strengthening the membership signal by using multiple correlated sequences as input (Maini et al., 2021; Kandpal et al., 2023; Maini et al., 2024), which aligns more closely with detecting contamination of an entire test set rather than a single example (Golchin & Surdeanu, 2023; Oren et al., 2023).

B. Pretraining Implementation Details

We pretrained Qwen 3 (Yang et al., 2025a) architecture causal language models. We used the AdamW optimizer (Loshchilov & Hutter, 2019) with HuggingFace defaults. We used linear warmup for 250 steps followed by cosine annealing. Following Shuai et al. (2024), we set the number of tokens per optimizer step $\propto D^{0.264}$. The base learning rate was $1e-6$ and the maximum learning rate was the base learning rate times the square of the number of tokens per optimizer step. Gradients were clipped to a maximum of 1.0. For more information, please see our [public GitHub repository](#).