The Contamination Paradox: Why Test Set Leakage Can Be Both Potent and Negligible

Rylan Schaeffer
Stanford CS

Ken Ziyu Liu
Stanford CS

Brando Miranda
Stanford CS

Ahmed Ahmed
Stanford CS

Niloofar Mireshghallah
Carnegie Mellon University

Sanmi Koyejo
Stanford CS

Abstract

Accurately evaluating the capabilities of large language models is critical for both machine learning research and society alike, but is undermined by leakage of benchmark test data into pretraining corpora. Circumstantial and causal evidence alike demonstrate that benchmark performance increases with model size and with the number of benchmark replicas in pretraining corpora. However, recent work by Bordt et al. (2025) demonstrated that test set contamination has little-to-no impact in the "overtrained" regime common to frontier AI systems, raising an apparent paradox of how test set leakage can be both potent *and* negligible. We resolve this paradox with a simple explanation: a language model memorizes a benchmark test set based on its capacity (number of parameters) and its incentive (the relative training loss reduction from memorizing test data). We introduce a novel dose-response framework to quantitatively relate how the "response" of benchmark performance depends on the "dose" of the proportion of benchmark tokens contaminating the pretraining data, mediated by model size. This allows us to extract precise scaling relationships that clarify the effect of test set contamination on model performance.

1 Introduction

Accurately evaluating large language models (LLMs) is increasingly difficult because benchmark test sets leak into web-scale pretraining corpora (Brown et al., 2020; Du et al., 2022; Wei et al., 2022; Chowdhery et al., 2022; Touvron et al., 2023). A growing body of work provides both circumstantial and causal evidence that contamination boosts performance (see Related Work in Appendix A). In particular, memorization rises predictably with model capacity and with the number of times an example is seen (Carlini et al., 2023; Tirumala et al., 2022; Biderman et al., 2023; Duan et al., 2024; Morris et al., 2025), and controlled pretraining experiments show measurable performance improvements (Magar & Schwartz, 2022; Jiang et al., 2024; Yao et al., 2024; Kocyigit et al., 2025).

However, this understanding was recently complicated by Bordt et al. (2025), who found that when models are overtrained—trained on far more tokens than Chinchilla compute-optimal (Hoffmann et al., 2022)—the effect of contamination can diminish or even vanish. Because frontier AI systems are oftentimes pretrained precisely in that regime (Touvron et al., 2023; Sardana et al., 2024; Gadre et al., 2024), the field now faces a paradox: how can contamination be both potent and negligible?

We argue these seemingly inconsistent observations can be unified by a single principle: language models memorize when they are able (i.e., sufficient model capacity by the number of parameters) and incentivized (i.e., relative loss reduction from memorizing benchmarks test sets) to do so. We make this quantitatively precise using a *dose-response relationship*, where the "dose" is the percentage of benchmark tokens in the pretraining corpus and the "response" is the pretrained model's accuracy. Fitting this relationship to the controlled scaling suit of Bordt et al. (2025) yields simple scaling laws: larger models require smaller doses to realize the same contamination-driven gains, while in the limit

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

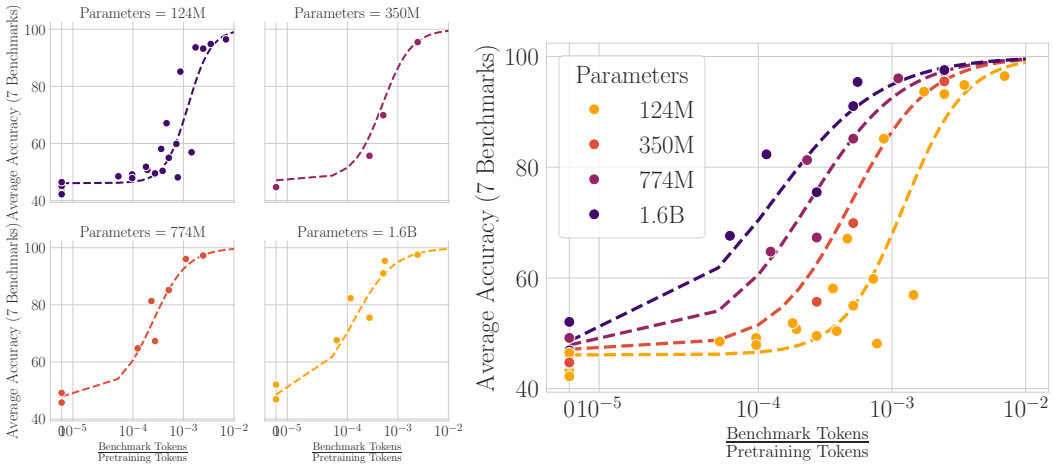

Figure 1: **A Dose-Response Model of Test Set Memorization.** We propose that language models memorize benchmark test sets based on their capacity (number of parameters) and incentive (relative loss reduction) to do so, a relationship we quantitatively capture as a dose-response relationship (Eqn. 2). We fit a functional relationship between the proportion of benchmark tokens in the pretraining data (the "dose") and the average benchmark accuracy (the "response") for different model sizes using the scaling suite from Bordt et al. (2025). These curves unify prior work: larger models have a steeper response and thus require a smaller dose of contaminated data to achieve high accuracy, but in the limit of infinite unique pretraining data (e.g., with overtrained models), the dose falls to 0% and thus test set contamination has little-to-no effect, as shown by Bordt et al. (2025). Note: The kink to the left of 10^{-4} is an artifact from symlog scaling the x-axis, not the fit response.

of unlimited unique data, the dose becomes vanishingly small and thus contamination has a negligible effect, clarifying why contamination can appear both potent and negligible.

2 Experimental Setup: Bordt et al. (2025)'s Scaling and Contamination

We study Bordt et al. (2025)'s GPT-3-like (Brown et al., 2020) language models pretrained on FineWeb-Edu (Penedo et al., 2024) in three regimes:

1. **Parameter Scaling**: Four model sizes ($N \in \{124\text{M}, 350\text{M}, 774\text{M}, 1.6\text{B}\}$) are pretrained on a fixed amount of data ($D = 7B$ tokens).
2. **Data Scaling**: One model size (124M) is pretrained on increasing data sizes ($D \in \{5\text{B}, 10\text{B}, 20\text{B}, 37\text{B}\}$, referred to respectively as 2x, 4x, 8x and 15x Chinchilla tokens).
3. **Simultaneous Parameter & Data Scaling**: Models ($N \in \{124\text{M}, 350\text{M}, 774\text{M}, 1.6\text{B}\}$) are pretrained on Chinchilla compute-optimal data ($D \in \{2.5\text{B}, 7\text{B}, 15.5\text{B}, 32\text{B}\}$ tokens, respectively).

"Chinchilla" refers to compute-optimal scaling (Hoffmann et al., 2022), taken as 20 pretraining tokens per model parameter. Each model's pretraining corpus is contaminated *uniformly at random* with test sets taken from 7 different multiple-choice question-answering benchmarks: ARC-Easy (Clark et al., 2018), BoolQ (Clark et al., 2019), HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2021), PiQA (Bisk et al., 2020), Social-I-QA (Sap et al., 2019) and WinoGrande (Sakaguchi et al., 2021). For each combination of regime, parameters, and tokens, Bordt et al. (2025) trained five models, increasing the number of test set replicas $R \in \{0$ (uncontaminated), 4, 12, 32 or 144$\}$. We report macro-average accuracy across the seven benchmarks using EleutherAI's Language Model Evaluation Harness (Gao et al., 2024). We also re-visualize the average accuracy scores from Bordt et al. (2025) here in Fig. 2 for the following reasons: (i) to provide the information conveniently to the reader, (ii) to remove the confounder introduced by Bordt et al. (2025) plotting *differences* of accuracies, and (iii) to show trends from complementary perspectives: number of test set replicas and the scaling quantity of interest (parameters, tokens, or parameters and tokens).

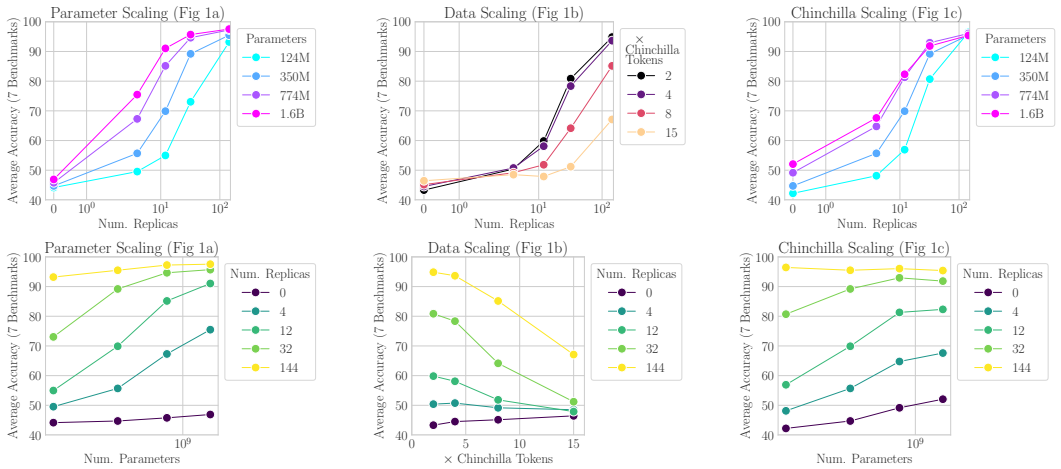

Figure 2: **Average Accuracy of Pretrained Models from Bordt et al. (2025).** We re-visualize the performance of language models pretrained by Bordt et al. (2025). **Left: Model Scaling.** Four increasing model sizes (124M, 350M, 774M, 1.6B) are pretrained on a fixed amount of data (7B tokens). **Center: Data Scaling**: One model size (124M) is pretrained on increasing data sizes (4.96B, 9.92B, 19.84B and 37.2B, referred to respectively as 2x, 4x, 8x and 15x Chinchilla tokens). **Right: Chinchilla Scaling**: Four increasing model sizes (124M, 350M, 774M, 1.6B) are pretrained on their corresponding Chinchilla compute-optimal data (2.48B, 7B, 15.48B, 32B tokens). **Key Takeaway:** When test set contamination is viewed as a function of multipliers of compute-optimal scaling, multiple complicated trends exist; once we reparameterize test set contamination as a function of the ratio of benchmark tokens to pretraining tokens, trends become cleaner (Fig. 1).

3 A Dose-Response Model of Test Set Memorization

To quantitatively understand how benchmark contamination affects model performance, we adopt the lens of dose-response relationships. In this framework, the "dose" is the proportion of contaminated test set tokens in each model's pretraining data, and the "response" is the model's resulting accuracy on the contaminated benchmarks. For each pretrained model i, we consider its number of parameters N_i, its average accuracy across the 7 benchmarks $a_i \in [0, 100]$ and the proportion of tokens in the pretraining corpus $p_i \in [0, 1]$ that originate from the benchmarks' test sets. We model the accuracy as

$$a_i \;=\; \mu(p_i, N_i;\, \theta) \;+\; \varepsilon_i, \tag{1}$$

where $\mu(\cdot)$ is the mean function and ε_i is zero-mean noise. For the mean function, we adopt the most commonly used dose-response equation called the "$E_{\max}$ equation" (Hill, 1910; Macdougall, 2006):

$$\mu(p, N;\, \theta) \stackrel{\text{def}}{=} a_{\min}(N) \;+\; \big(a_{\max}(N) - a_{\min}(N)\big) \cdot \frac{p^{\,h(N)}}{p^{\,h(N)} + p_{50}(N)^{\,h(N)}}. \tag{2}$$

This relationship has four intuitive components:

(i) $a_{\min}(N)$: The baseline accuracy of a model with no test set contamination ($p = 0$).

(ii) $a_{\max}(N)$: The maximum achievable accuracy as the dose of contamination dominates the pretraining corpus ($p \to 100$);

(iii) $p_{50}(N)$: The proportion of contamination necessary to achieve 50% of the accuracy gain.

(iv) $h(N) > 0$: A parameter that controls the curvature or steepness of the dose-response curve.

We parameterize these components as functions of model size N to ensure the relationships are interpretable and well-behaved (i.e., $0 \leq a_{\min}(N) < a_{\max}(N) \leq 100$, $p_{50}(N) > 0$, and $h(N) > 0$):

$$\begin{aligned}
a_{\min}(N) &\stackrel{\text{def}}{=} 100 \cdot \sigma(a_0 + a_1 \log N)\,, \\
a_{\max}(N) &\stackrel{\text{def}}{=} a_{\min}(N) + \big(100 - a_{\min}(N)\big) \cdot \sigma(b_0 + b_1 \log N)\,, \\
p_{50}(N) &\stackrel{\text{def}}{=} \exp(c_0 + c_1 \log N)\,, \\
h(N) &\stackrel{\text{def}}{=} \exp(h_0 + h_1 \log N)\,,
\end{aligned}$$

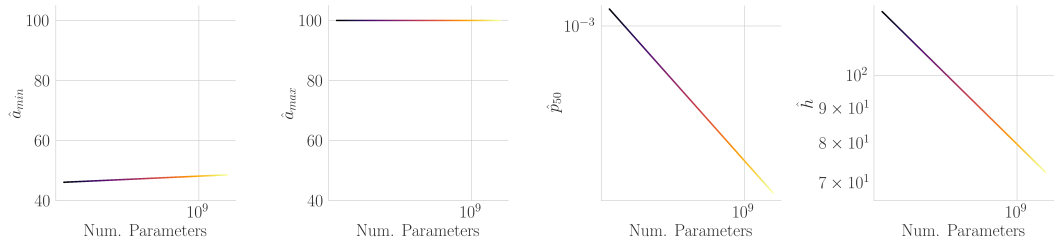

Figure 3: **Fitted Dose-Response Parameters Exhibit Predictable Scaling Laws with Model Size.** Each panel displays a parameter of our dose-response relationship (Eqn. 2) as a function of the number of model parameters N. The baseline accuracy without contamination, $a_{\min}(N)$, trends upward with model size, consistent with normal scaling without contamination. The maximum achievable accuracy, $a_{\max}(N)$, is consistently at 100%, which suggests that any model can achieve perfect accuracy with sufficient test set contamination in its training data. The proportion of contamination needed to achieve half of the possible accuracy gain scales as a power law $\log p_{50}(N) = 9.71 \cdot N^{-0.881}$, showing that larger models need a smaller proportion of benchmark data to achieve a significant accuracy boost. The steepness parameter also scales as a power law $\log h(N) = 4.63 \cdot N^{-0.214}$, indicating that the accuracy of larger models increases more sharply in response to contamination.

where $\sigma(x) = 1/(1 + e^{-x})$ is the sigmoid function. The power–law form for $p_{50}(N)$ and $h(N)$ are based on the ansatz that larger models need less benchmark proportion to reach a given fraction of the attainable improvement (Kaplan et al., 2020; Hoffmann et al., 2022). We estimate $\theta = \{a_0, a_1, b_0, b_1, c_0, c_1, h_0, h_1\}$ by robust nonlinear least squares over all points $\{(p_i, N_i, a_i)\}$ using the "soft-ℓ_1" loss to reduce sensitivity to outliers. The fit parameters are $\hat{a}_0 = -0.8877$, $\hat{a}_1 = 0.0392$, $\hat{b}_0 = -35.7432$, $\hat{b}_1 = 17.0654$, $\hat{c}_0 = 9.7100$, $\hat{c}_1 = -0.8807$, $\hat{h}_0 = 4.6277$, and $\hat{h}_1 = -0.2145$.

Fig. 1 includes all models from all three scaling regimes and all number of test set replicas ($\{0, 4, 12, 32, 144\}$). The fitted curves capture the key qualitative pattern visible in the underlying measurements: as the "dose" of benchmark contamination increases, accuracy approaches ceiling performance, and larger models achieve the same accuracy at markedly smaller doses. In contrast with Bordt et al. (2025), who find different test set memorization relationships depending on the scaling regime, Fig. 1 demonstrates that a single unifying relationship appears once one instead considers the test set dosage, i.e., the ratio of benchmark tokens to total pretraining tokens.

Fig. 3 demonstrates how dose-response parameters change with model size: The baseline accuracy with no contamination $\hat{a}_{\min}(N)$ increases gently with model size, reflecting normal scaling without contamination. The maximum achievable accuracy under heavy contamination is estimated near the upper bound ($\hat{a}_{\max}(N) \approx 100\%$ for all sizes), indicating near-perfect recall is attainable when contaminated tokens dominate pretraining. Our ansatz for sensitivity and curve steepness of power law scaling with respect to model size fit the data well: $p_{50}(N) \propto N^{-0.881}$ and $h(N) \propto N^{\hat{-0.214}}$. For practical intuition, the dose needed to realize 50% of the attainable gain is tiny and shrinks with model size: for a 1.6B parameter model, having $\sim 0.01\%$ of pretraining tokens originate from benchmarks captures 50% of the attainable performance improvements, and (assuming the relationship holds for larger models) for a 30B parameter model, a dose of $\sim 0.001\%$ benchmark tokens captures 50% of the attainable performance improvements.

4 Discussion

This work introduces a dose-response framework that resolves seemingly paradoxical findings on test set contamination within a single curve family: minor contamination can yield large gains for high-capacity models, while overtraining with unique new data reduces the dose to zero, thereby rendering contamination negligible. Moreover, our insights (1) yields actionable contamination thresholds (e.g., p_{50}) for auditors, and (2) clarifies why contamination matters most when capacity is high and unique data are scarce (Villalobos et al., 2024).

Future Directions: (1) Future work should test the generalizability of these scaling laws across different tasks and different notions of memorization (Tirumala et al., 2022; Carlini et al., 2023; Hayes et al., 2025; Duan et al., 2025). (2) Because frontier models are often pretrained on multiple epochs (Muennighoff et al., 2023), future work should study how multiple epochs complicates this

picture. (3) This framework could also be developed into a practical tool for evaluators to predict and potentially correct for performance inflation caused by contamination.

References

Ben Adlam and Jeffrey Pennington. Understanding double descent requires a fine-grained bias-variance decomposition. *Advances in neural information processing systems*, 33:11022–11032, 2020.

Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Stella Biderman, Usvsn Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. Emergent and predictable memorization in large language models. *Advances in Neural Information Processing Systems*, 36:28072–28090, 2023.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 7432–7439, 2020.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pp. 1024–1034. PMLR, 2020.

Sebastian Bordt, Suraj Srinivas, Valentyn Boreiko, and Ulrike von Luxburg. How much can we forget about data contamination? In *Forty-second International Conference on Machine Learning*, 2025.

Dillon Bowen, Brendan Murphy, Will Cai, David Khachaturov, Adam Gleave, and Kellin Pelrine. Scaling trends for data poisoning in llms, 2025. URL https://arxiv.org/abs/2408.02946.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022. URL https://arxiv.org/abs/2204.02311.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, 2019.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Debeshee Das, Jie Zhang, and Florian Tramèr. Blind baselines beat membership inference attacks for foundation models. *arXiv preprint arXiv:2406.16201*, 2024.

Chunyuan Deng, Yilun Zhao, Yuzhao Heng, Yitong Li, Jiannan Cao, Xiangru Tang, and Arman Cohan. Unveiling the spectrum of data contamination in language models: A survey from detection to remediation. *arXiv preprint arXiv:2406.14644*, 2024a.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. Investigating data contamination in modern benchmarks for large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (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/.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (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/.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (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/.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International conference on machine learning*, pp. 5547–5569. PMLR, 2022.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? *arXiv preprint arXiv:2402.07841*, 2024.

Sunny Duan, Mikail Khona, Abhiram Iyer, Rylan Schaeffer, and Ila R Fiete. Uncovering latent memories in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=KSBx6FBZpE.

Samir Yitzhak Gadre, Georgios Smyrnis, Vaishaal Shankar, Suchin Gururangan, Mitchell Wortsman, Rulin Shao, Jean Mercat, Alex Fang, Jeffrey Li, Sedrick Keh, Rui Xin, Marianna Nezhurina, Igor Vasiljevic, Jenia Jitsev, Luca Soldaini, Alexandros G. Dimakis, Gabriel Ilharco, Pang Wei Koh, Shuran Song, Thomas Kollar, Yair Carmon, Achal Dave, Reinhard Heckel, Niklas Muennighoff, and Ludwig Schmidt. Language models scale reliably with over-training and on downstream tasks, 2024. URL https://arxiv.org/abs/2403.08540.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Järviniemi,

Matthew Barnett, Robert Sandler, Matej Vrzala, Jaime Sevilla, Qiuyu Ren, Elizabeth Pratt, Lionel Levine, Grant Barkley, Natalie Stewart, Bogdan Grechuk, Tetiana Grechuk, Shreepranav Varma Enugandla, and Mark Wildon. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai, 2025. URL https://arxiv.org/abs/2411.04872.

Shahriar Golchin and Mihai Surdeanu. Data contamination quiz: A tool to detect and estimate contamination in large language models. *arXiv preprint arXiv:2311.06233*, 2023.

Shahriar Golchin and Mihai Surdeanu. 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.

Ziwen Han, Meher Mankikar, Julian Michael, and Zifan Wang. Search-time data contamination, 2025. URL https://arxiv.org/abs/2508.13180.

Jamie Hayes, Marika Swanberg, Harsh Chaudhari, Itay Yona, Ilia Shumailov, Milad Nasr, Christopher A Choquette-Choo, Katherine Lee, and A Feder Cooper. Measuring memorization in language models via probabilistic extraction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9266–9291, 2025.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

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

Archibald Vivian Hill. The possible effects of the aggregation of the molecules of hemoglobin on its dissociation curves. *j. physiol.*, 40:iv–vii, 1910.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (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.

Matthew Jagielski, Milad Nasr, Katherine Lee, Christopher A Choquette-Choo, Nicholas Carlini, and Florian Tramer. Students parrot their teachers: Membership inference on model distillation. *Advances in Neural Information Processing Systems*, 36, 2024.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: 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.

Mingjian Jiang, Ken Ziyu Liu, and Sanmi Koyejo. 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.

Minhao Jiang, Ken Ziyu Liu, Ming Zhong, Rylan Schaeffer, Siru Ouyang, Jiawei Han, and Sanmi Koyejo. Investigating data contamination for pre-training language models, 2024. URL https://arxiv.org/abs/2401.06059.

Nikhil Kandpal, Krishna Pillutla, Alina Oprea, Peter Kairouz, Christopher A Choquette-Choo, and Zheng Xu. User inference attacks on large language models. *arXiv preprint arXiv:2310.09266*, 2023.

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

Muhammed Yusuf Kocyigit, Eleftheria Briakou, Daniel Deutsch, Jiaming Luo, Colin Cherry, and Markus Freitag. 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.

Zhifeng Kong, Amrita Roy Chowdhury, and Kamalika Chaudhuri. Can membership inferencing be refuted? *arXiv preprint arXiv:2303.03648*, 2023.

Huihan Li, You Chen, Siyuan Wang, Yixin He, Ninareh Mehrabi, Rahul Gupta, and Xiang Ren. Diagnosing memorization in chain-of-thought reasoning, one token at a time. *arXiv preprint arXiv:2508.02037*, 2025.

Marvin Li, Jason Wang, Jeffrey Wang, and Seth Neel. Mope: Model perturbation-based privacy attacks on language models. *arXiv preprint arXiv:2310.14369*, 2023.

Yucheng Li, Yunhao Guo, Frank Guerin, and Chenghua Lin. An open-source data contamination report for large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (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/.

Ken Ziyu Liu, Christopher A Choquette-Choo, Matthew Jagielski, Peter Kairouz, Sanmi Koyejo, Percy Liang, and Nicolas Papernot. Language models may verbatim complete text they were not explicitly trained on. *arXiv preprint arXiv:2503.17514*, 2025.

James Macdougall. Analysis of dose–response studies—emax model. In *Dose finding in drug development*, pp. 127–145. Springer, 2006.

Inbal Magar and Roy Schwartz. Data contamination: From memorization to exploitation. *arXiv preprint arXiv:2203.08242*, 2022.

Pratyush Maini, Mohammad Yaghini, and Nicolas Papernot. Dataset inference: Ownership resolution in machine learning. *arXiv preprint arXiv:2104.10706*, 2021.

Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. Llm dataset inference: Did you train on my dataset? *arXiv preprint arXiv:2406.06443*, 2024.

Neal Mangaokar, Ashish Hooda, Zhuohang Li, Bradley A Malin, Kassem Fawaz, Somesh Jha, Atul Prakash, and Amrita Roy Chowdhury. What really is a member? discrediting membership inference via poisoning. *arXiv preprint arXiv:2506.06003*, 2025.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership inference attacks against language models via neighbourhood comparison. *arXiv preprint arXiv:2305.18462*, 2023.

Alexandre Matton, Tom Sherborne, Dennis Aumiller, Elena Tommasone, Milad Alizadeh, Jingyi He, Raymond Ma, Maxime Voisin, Ellen Gilsenan-McMahon, and Matthias Gallé. On leakage of code generation evaluation datasets. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (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/.

Matthieu Meeus, Shubham Jain, Marek Rei, and Yves-Alexandre de Montjoye. Inherent challenges of post-hoc membership inference for large language models. *arXiv preprint arXiv:2406.17975*, 2024.

John X Morris, Chawin Sitawarin, Chuan Guo, Narine Kokhlikyan, G Edward Suh, Alexander M Rush, Kamalika Chaudhuri, and Saeed Mahloujifar. How much do language models memorize? *arXiv preprint arXiv:2505.24832*, 2025.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376, 2023.

Shiwen Ni, Xiangtao Kong, Chengming Li, Xiping Hu, Ruifeng Xu, Jia Zhu, and Min Yang. Training on the benchmark is not all you need, 2025. URL https://arxiv.org/abs/2409.01790.

Fan Nie, Ken Ziyu Liu, Zihao Wang, Rui Sun, Wei Liu, Weijia Shi, Huaxiu Yao, Linjun Zhang, Andrew Y. Ng, James Zou, Sanmi Koyejo, Yejin Choi, Percy Liang, and Niklas Muennighoff. Uq: Assessing language models on unsolved questions, 2025. URL https://arxiv.org/abs/2508.17580.

Yonatan Oren, Nicole Meister, Niladri S Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. Proving test set contamination in black-box language models. In *The Twelfth International Conference on Learning Representations*, 2023.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale, 2024. URL https://arxiv.org/abs/2406.17557.

Kun Qian, Shunji Wan, Claudia Tang, Youzhi Wang, Xuanming Zhang, Maximillian Chen, and Zhou Yu. Varbench: Robust language model benchmarking through dynamic variable perturbation, 2024. URL https://arxiv.org/abs/2406.17681.

Anka Reuel, Benjamin Bucknall, Stephen Casper, Timothy Fist, Lisa Soder, Onni Aarne, Lewis Hammond, Lujain Ibrahim, Alan Chan, Peter Wills, Markus Anderljung, Ben Garfinkel, Lennart Heim, Andrew Trask, Gabriel Mukobi, Rylan Schaeffer, Mauricio Baker, Sara Hooker, Irene Solaiman, Sasha Luccioni, Nitarshan Rajkumar, Nicolas Moës, Jeffrey Ladish, David Bau, Paul Bricman, Neel Guha, Jessica Newman, Yoshua Bengio, Tobin South, Alex Pentland, Sanmi Koyejo, Mykel Kochenderfer, and Robert Trager. Open problems in technical AI governance. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=1nO4qFMiS0. Survey Certification.

Martin Riddell, Ansong Ni, and Arman Cohan. Quantifying contamination in evaluating code generation capabilities of language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 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/.

Manley Roberts, Himanshu Thakur, Christine Herlihy, Colin White, and Samuel Dooley. 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.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, pp. 5558–5567. PMLR, 2019.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 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.

Oscar Sainz, Iker García-Ferrero, Alon Jacovi, Jon Ander Campos, Yanai Elazar, Eneko Agirre, Yoav Goldberg, Wei-Lin Chen, Jenny Chim, Leshem Choshen, Luca D'Amico-Wong, Melissa Dell, Run-Ze Fan, Shahriar Golchin, Yucheng Li, Pengfei Liu, Bhavish Pahwa, Ameya Prabhu, Suryansh Sharma, Emily Silcock, Kateryna Solonko, David Stap, Mihai Surdeanu, Yu-Min Tseng, Vishaal Udandarao, Zengzhi Wang, Ruijie Xu, and Jinglin Yang. Data contamination report from the 2024 CONDA shared task. In Oscar Sainz, Iker García Ferrero, Eneko Agirre, Jon Ander Campos, Alon Jacovi, Yanai Elazar, and Yoav Goldberg (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/.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*, 2018.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4463–4473, 2019.

Nikhil Sardana, Jacob Portes, Sasha Doubov, and Jonathan Frankle. Beyond chinchilla-optimal: Accounting for inference in language model scaling laws. In *International Conference on Machine Learning*, pp. 43445–43460. PMLR, 2024.

Rylan Schaeffer. Pretraining on the test set is all you need, 2023. URL https://arxiv.org/abs/2309.08632.

Rylan Schaeffer, Zachary Robertson, Akhilan Boopathy, Mikail Khona, Kateryna Pistunova, Jason William Rocks, Ila R Fiete, Andrey Gromov, and Sanmi Koyejo. Double descent demystified: Identifying, interpreting & ablating the sources of a deep learning puzzle. In *The Third Blogpost Track at ICLR 2024*, 2024.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zWqr3MQuNs.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35:38274–38290, 2022.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Position: Will we run out of data? limits of llm scaling based on human-generated data. In *Forty-first International Conference on Machine Learning*, 2024.

Xinyi Wang, Antonis Antoniades, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, and William Yang Wang. 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.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022.

Chunqiu Steven Xia, Yinlin Deng, and Lingming Zhang. Top leaderboard ranking= top coding proficiency, always? evoeval: Evolving coding benchmarks via llm. *arXiv preprint arXiv:2403.19114*, 2024.

Chulin Xie, Yangsibo Huang, Chiyuan Zhang, Da Yu, Xinyun Chen, Bill Yuchen Lin, Bo Li, Badih Ghazi, and Ravi Kumar. On memorization of large language models in logical reasoning, 2025. URL https://arxiv.org/abs/2410.23123.

Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. Benchmark data contamination of large language models: A survey, 2024a. URL https://arxiv.org/abs/2406.04244.

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. Benchmarking benchmark leakage in large language models, 2024b. URL https://arxiv.org/abs/2404.18824.

Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E Gonzalez, and Ion Stoica. Rethinking benchmark and contamination for language models with rephrased samples. *arXiv preprint arXiv:2311.04850*, 2023.

Zhen Yang, Hongyi Lin, Yifan He, Jie Xu, Zeyu Sun, Shuo Liu, Pengpeng Wang, Zhongxing Yu, and Qingyuan Liang. Rethinking the effects of data contamination in code intelligence, 2025. URL https://arxiv.org/abs/2506.02791.

Feng Yao, Yufan Zhuang, Zihao Sun, Sunan Xu, Animesh Kumar, and Jingbo Shang. Data contamination can cross language barriers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17864–17875, 2024.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pp. 268–282. IEEE, 2018.

Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference attacks, 2023.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, 2019.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Charlotte Zhuang, Dylan Slack, Qin Lyu, Sean Hendryx, Russell Kaplan, Michele Lunati, and Summer Yue. A careful examination of large language model performance on grade school arithmetic, 2024a. URL https://arxiv.org/abs/2405.00332.

Jie Zhang, Debeshee Das, Gautam Kamath, and Florian Tramèr. Membership inference attacks cannot prove that a model was trained on your data. *arXiv preprint arXiv:2409.19798*, 2024b.

Zhe Zhang, Runlin Liu, Aishan Liu, Xingyu Liu, Xiang Gao, and Hailong Sun. Dynamic benchmark construction for evaluating large language models on real-world codes, 2025. URL https://arxiv.org/abs/2508.07180.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 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. A growing body of evidence shows that leakage of benchmark material into pretraining corpora can inflate reported performance and compromise evaluation validity. Position and survey papers argue that contamination should be routinely audited and reported for each benchmark, and they document the breadth of leakage modes and impacts (Sainz et al., 2023, 2024; Deng et al., 2024a; Xu et al., 2024a; Reuel et al., 2025). Empirical audits of large web corpora show nontrivial train–test overlap and duplication (Dodge et al., 2021), and work on systematizing benchmark integrity highlights ways LMs can "cheat" on evaluations if contamination is not addressed (Zhou et al., 2023; Dong et al., 2024). Measurements on widely used math benchmarks indicate likely leakage and overfitting signals (Zhang et al., 2024a). Community reports and open-source audits provide broader, ongoing measurements across models and datasets (Li et al., 2024). Beyond evaluation leakage, scaling studies indicate that poisoning risks increase with model size: across dozens of frontier LLMs, larger models learn harmful behaviors from tiny poisoned fractions substantially faster than smaller models, underscoring the need for robust curation and safeguards (Bowen et al., 2025). As a cautionary illustration, Schaeffer (2023) shows that pretraining on the test set trivially yields strong benchmark results, motivating rigorous decontamination and auditing.

Controlled contamination during pretraining. Several studies *causally* probe memorization by deliberately inserting evaluation items into the *pretraining* mix and varying exposure. Magar & Schwartz (2022) interleave task datasets with general text during masked-LM pretraining, systematically varying duplication; they distinguish storing examples ("memorization") from using them to improve test accuracy ("exploitation"), and show that both model size and repetition amplify exploitation on leaked items. Jiang et al. (2024) pretrain GPT-2–style models from scratch on clean corpora augmented with either *text-only* (inputs) or *ground-truth* (input–output) benchmark injections, sweeping contamination frequency; they find sizable gains under ground-truth insertion and show that paraphrases and partial leaks can evade simple n-gram decontamination. Moving beyond monolingual settings, Yao et al. (2024) demonstrate a *cross-lingual* channel: continuing pretraining on non-English translations of English benchmarks yields material improvements on the original English tests, revealing contamination undetectable by string-overlap audits. At larger scale, Bordt et al. (2025) vary (i) repetition of leaked examples, (ii) model size (up to $\sim$1.6B), and (iii) the total token budget from compute-optimal to $\gg$optimal; they recover predictable scaling with size and repeats, and also show that sufficiently long training on abundant *unique* data (with regularization) can attenuate or erase contamination measured earlier. Kocyigit et al. (2025) study machine translation, injecting held-out *source–target* pairs at controlled times and frequencies during pretraining of 1B- and 8B-parameter models; they quantify large BLEU overestimation for full-pair leakage (with weaker effects for source-only/target-only), and observe stronger inflation for larger models and lower-resource settings. Together, these pretraining-time interventions provide *causal* evidence that LMs will memorize and exploit benchmark material.

Repeated data and memorization dynamics. A complementary line of work isolates the effect of repeated training examples. Hernandez et al. (2022) train families of LMs where a small fraction of data is repeated many times, finding strong double descent (Advani et al., 2020; Belkin et al., 2019; Adlam & Pennington, 2020; Bordelon et al., 2020; Schaeffer et al., 2024) and showing that repeating even 0.1% of tokens $100\times$ can substantially degrade generalization. Tirumala et al. (2022) track exact-sequence memorization through training and across scales, showing that larger models memorize faster, memorize more, and forget less. Carlini et al. (2023) quantify log-linear relationships between verbatim emission and (i) model capacity, (ii) duplication count, and (iii) prompt length. Biderman et al. (2023) study *forecasting* whether a specific string will be memorized, showing that reliable prediction often requires using a sizable fraction of the target model's pretraining compute and providing preliminary scaling recommendations for forecast design. Beyond explicit repetition, Duan et al. (2025) uncover *latent memorization*: many memorized sequences persist and can be revealed later (e.g., by weight perturbations) even if not obviously memorized at the final checkpoint, posing privacy risks. Finally, memorization appears task-dependent: Wang et al. (2025) find stronger memorization in knowledge-intensive QA, while machine translation and mathematical reasoning show comparatively greater novelty/"true" generalization. Memorization also interacts with logical reasoning: using dynamically generated Knights & Knaves puzzles, Xie et al. (2025) show that LLMs can interpolate and memorize training puzzles to near perfection after fine-tuning yet remain brittle

to slight perturbations; importantly, fine-tuning also improves true generalization, revealing a shifting balance between reasoning and memorization. These results provide mechanistic and scaling context for the pretraining-injection studies above.

Detecting and proving contamination. Many papers focus on a complementary problem: detecting and/or proving test set contamination. Oren et al. (2023) and Ni et al. (2025) propose statistical tests with provable false-positive control by exploiting exchangeability: without contamination, canonical orderings should not be privileged relative to shuffles. Shi et al. (2024) introduce Min-$k\%$-Prob to determine if a sequence likely appeared in pretraining using only black-box probabilities. Two complementary lines from Golchin & Surdeanu (2023, 2024), respectively, frame detection as a multiple-choice "Data Contamination Quiz" (estimate contamination by asking models to pick the original among perturbations) and use temporal information about model training windows vs. benchmark releases (as well as Roberts et al. (2024)). Broader audits quantify leakage and decontamination across tasks and models (Xu et al., 2024b; Deng et al., 2024b; Li et al., 2024), while Yang et al. (2023) show that rephrasing can evade n-gram filters, underscoring the limits of surface-overlap heuristics. Riddell et al. (2024) quantify contamination across popular code suites and link overlap to performance deltas. Matton et al. (2024) catalog leakage channels (direct, synthetic-pipeline, and model-selection overfitting) and release a dataset (LBPP) to mitigate them. Complementing these audits, Yang et al. (2025) systematically test fine-grained contamination scenarios in code intelligence (input-only, output-only, unpaired, paired) across pretrained language models (RoBERTa, GPT-2) and LLMs (LLaMA, StarCoder), finding that paired contamination has limited effect under the pretrain–finetune–inference pipeline, but substantially affects LLMs under a pretraining-plus-inference paradigm, while other scenarios often have minimal impact. Work tracing the origins of chain-of-thought style sequences provides additional detection instruments (e.g., Li et al., 2025).

Preventing test set contamination Concerns with test-set contamination have led to new approaches to benchmark creation, including dynamically updating 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). Nie et al. (2025) recently released a benchmark of unsolved questions, which, while perhaps not the main motivation, has the nice benefit of preventing models from being trained on the solutions.

Retrieval-/agent-time contamination. As evaluations move from static prompting to tool-augmented agents, contamination risks expand to include search-time. Han et al. (2025) introduce search-time contamination, where an agent retrieves benchmark Q&A pages during evaluation, thereby artificially inflating scores.

Membership Inference Attacks Membership inference attacks (MIA) focus on the idea of determining whether an example has been trained on by a model based on (white-box or black-box) access to the model alone (Shokri et al., 2017). It relates to test set contamination in that detecting contamination can be cast as a membership inference problem. The MIA literature spans computer vision (e.g., Yeom et al. (2018); Salem et al. (2018); Sablayrolles et al. (2019); Jagielski et al. (2024)) and more recently to language modeling (e.g., Carlini et al. (2021); Zarifzadeh et al. (2023); Shi et al. (2024); Mattern et al. (2023); Li et al. (2023)). Despite these attempts, progress of sequence-level MIA on language models is hindered by 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 using a theoretical gradient-space attack. Liu et al. (2025) demonstrates the fundamental limitations of n-gram based membership definitions which hinder downstream tests, with Mangaokar et al. (2025) providing a concrete of exploit of existing MIA tests via poisoning. Due to these challenges, recent work also explore enhancing membership signals by leveraging multiple correlated sequences as inputs (Maini et al., 2021; Kandpal et al., 2023; Maini et al., 2024), which are closely related to detecting contamination of an entire test set rather than individual test examples (Golchin & Surdeanu, 2023; Oren et al., 2023).

Dose–response relationships. Most similar to our idea of a dose–response relationship is Hernandez et al. (2022), which argued that highly repetitive data can be severely damaging if the number of repeats incentivizes memorizing that data and if doing so consumes a meaningful fraction of

the model's capacity. Our framework connects this incentive perspective with model capacity and observed contamination effects.

Positioning. Relative to prior work that asks *whether* models are contaminated and *how* to detect or mitigate it, our contribution is a unified *dose–response* framing that quantifies *how much* performance can be attributed to contamination-driven memorization as a function of exposure (e.g., repeats, paraphrases) and training trajectory. Our measurements and fits operationalize this principle across model sizes and token budgets, connecting the controlled injections above with scaling-law regularities.