

Form Over Content In Gradient-Based Data Attribution Methods

Sunwoo Kim Seokwon Jung Sohyung Kim Seong Joon Oh Alice Oh
 KAIST
 {sunwoo.kim, tjrdnjs0313}@kaist.ac.kr ks000225@gmail.com
 coallaoh@kaist.ac.kr alice.oh@kaist.edu

Abstract

Data attribution methods using gradient similarity are widely used to analyze and select training data for large language models, but what gradient similarity actually measures is debated. Some interpret it as identifying task-relevant skills, while other work reports that surface form is the main factor. We resolve this debate for supervised fine-tuning examples by varying task and answer format independently. Specifically, we render benchmarks in different answer formats, such that datasets can share a task without a format or a format without a task. We find that gradient alignment follows the answer format, as benchmark pairs sharing an answer format align strongly (disattenuated cosine near 0.4), while same benchmarks rendered with different answer format classes show no alignment (near 0.0). We demonstrate that this ordering holds from the earliest pretraining checkpoints through post-training, and across model scales and families. We then analyze the released selections of LESS, a gradient-based data selection method for instruction tuning, and find that each target’s selections over-represent the target’s own answer format. Hence, we demonstrate that gradient-based attribution methods track format similarity more than task semantics, meaning that such methods, as well as the semantic interpretation of the gradient, should be tested on data where answer format and task vary independently for greater robustness and reliability.

1 Introduction

Various data attribution methods have been developed to improve the analysis of training data for deep learning models such as transformer-based large language models (LLMs). Many such methods leverage the loss gradient to measure the influence of training data [1–5]. A representative method upon which many other methods are based on is TracIn, which uses the raw similarity, or dot product, between loss gradients of data points to calculate the influence of a training data point on a test data point [5].

The literature disagrees on why these methods work and what the gradient actually represents. Xia et al. [4] state that the gradient signal “goes beyond surface form cues to identify data that exemplifies the necessary reasoning skills,” while Wang et al. [6] report that answer format dominates over the semantic content of the task in the fine-tuning gradient. The confusion has persisted because format and semantics are spuriously correlated in data, especially benchmarks, so observation alone cannot separate them; for example, factual knowledge benchmarks tend to share the same multiple-choice format, so a gradient that matches format and a gradient that matches knowledge make the same predictions. The confusion is pertinent to *targeted instruction tuning*, where the gradient signal is used to pick training examples intended to exemplify the capabilities a target requires, based on the assumption that the gradient encodes semantics [4]. We resolve the confusion by varying the *task*¹

¹Note that we use the terms *task* and *benchmark* interchangeably, as the benchmarks used assess unique capabilities.

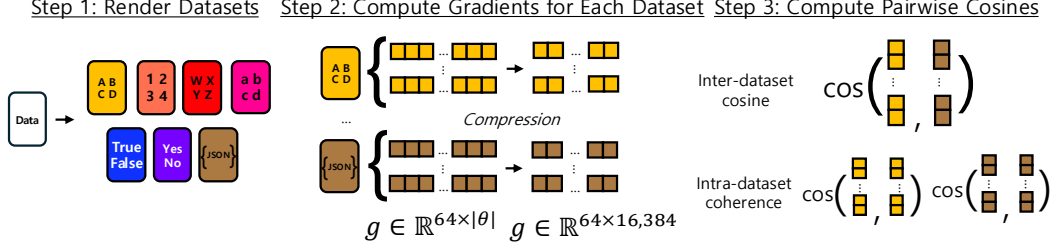


Figure 1: Overview of method used to extract pairwise cosine similarities between datasets.

and the *answer format* of the same questions independently, and measuring each factor’s effect on gradient alignment.

Our contributions are threefold:

- We show that answer format, not task content, governs gradient similarity in LLMs, meaning that influence as measured by gradient-based data attribution methods does not imply task relatedness.
- We find that such dominance of format over content is consistent across pre-training and post-training steps, as well as across model scales.
- We analyze the selection resulting from an existing gradient-based data selection method, namely LESS [4], to demonstrate that the resulting selection is skewed towards same-format samples.

2 Related Work

Gradient-based attribution and its semantic reading. The representative gradient-based data attribution method is TracIn, which estimates the influence of a training example on a test example by accumulating dot products of their loss gradients over training checkpoints [5]. Later work scales this approach to larger models and corpora [3, 1, 2, 7]. LESS applies it to data selection and interprets the signal as identifying data that exemplifies the necessary reasoning skills [4]. This semantic reading of gradient geometry is widely shared: gradients are described as task-sensitive representations of knowledge [8], and gradient clusters are treated as task experts [9], task-agnostic coresets [10], separable basic abilities [11], or attributions of reasoning ability [12]. The closest supporting evidence, which are influence analyses that find increasingly abstract relationships between influential documents and model generations [13, 1], concerns a different object: the influence of pretraining documents on model generations, not the gradients of supervised fine-tuning examples. In short, there is prior work that assumes that gradient alignment between fine-tuning examples reflects shared task content, and we test this assumption.

Evidence for surface form. Other research suggests the gradient encodes surface format more strongly than content. At the outcome level, LESS fails to beat random selection at large pool scales [14], lexical baselines remain competitive with gradient attribution at fact tracing [7], and gradient similarity fails to indicate task relatedness in multi-task classification [15]. At the signal level, the only direct gradient measurement is confined to a single dataset: Wang et al. [6] shuffle one dataset’s labels and find that early fine-tuning gradients track the format. Whether format organizes the gradients *between* datasets, the comparison that attribution and selection actually perform, remains unmeasured. The confound surfaces in deployed pipelines as well. Influence magnitudes track completion length inside LESS itself [16], gradient-matched selections of harmful data turn out to be lists, bullet points, and math questions [17], and influence retrieval for verbalized confidence latches onto surface markers of certainty rather than content [18]. We isolate the confound with a controlled design, varying task and answer format independently in the setting attribution methods operate in, where candidate and target come from different datasets.

3 Method

3.1 Preliminaries

Per-example loss gradients. Let f_θ be a language model with parameters θ , and let a supervised example $z = (x, y)$ consist of a prompt x and an answer y . The loss $\ell(z; \theta)$ is the mean cross-entropy of the answer tokens given the prompt, and the per-example gradient is $g(z; \theta) = \nabla_\theta \ell(z; \theta)$. Because the loss is masked to the answer span, g is the update that the example would contribute during instruction tuning, and it is the basic object of gradient-based attribution.

Gradient-based influence. TracIn [5] estimates the influence of a training example on a test example as a learning-rate-weighted sum of gradient dot products over training checkpoints $\theta_1, \dots, \theta_K$. LESS [4] adapts this score to instruction tuning: Notably, gradient alignments are calculated by cosine so that long answers do not dominate through gradient norm. The training pool is ranked by the aggregated similarity score with a target set across checkpoints, and the top 5% is selected for fine-tuning. We analyze the selections this method produces in Section 3.3.

3.2 Our Method: Measuring Dataset-wise Gradient Alignment

We measure gradient alignment between datasets in three steps as shown in Figure 1.

Step 1: Render datasets. We create a *dataset* by rendering the questions of a benchmark into one answer format, keeping the questions themselves fixed. We cross five benchmarks, TriviaQA [19], SQuAD [20], GSM8K [21], ARC-Challenge [22], and CRUXEval [23], with seven answer formats spanning three structure classes: binary judgment (True/False, Yes/No), four-option selection (A–D, 1–4, W–Z, a–d), and structured generation (JSON); details in Appendix C. The resulting 35 datasets let any two be compared while changing only the task or only the answer format.

Step 2: Compute gradients for each dataset. For $n = 64$ examples per dataset we compute the per-example answer-span gradient $g(z; \theta)$ of Section 3.1, taken with respect to the transformer body only; excluding the embedding and unembedding tables ensures that vocabulary overlap between answer tokens cannot drive similarity. Each gradient is compressed with a seeded CountSketch projection [24] to $k = 16,384$ dimensions, which preserves inner products in expectation. Each dataset becomes a matrix in $\mathbb{R}^{64 \times 16,384}$.

Step 3: Compute pairwise cosines. The *inter-dataset* alignment of datasets A and B is the mean cosine over all cross-dataset example pairs:

$$\hat{c}(A, B) = \frac{1}{|A||B|} \sum_{z \in A} \sum_{z' \in B} \cos(g(z; \theta), g(z'; \theta)). \quad (1)$$

The same statistic measured within a dataset gives the *intra-dataset* coherence $c_A = \hat{c}(A, A)$, evaluated over distinct question pairs. We disattenuate by coherence, as datasets differ in how noisy their example gradients are, so raw values of Equation 1 are not comparable across pairs. If each example gradient is a dataset-level signal μ plus independent noise, the measured mean is attenuated by exactly $\sqrt{c_A c_B}$, so dividing it out recovers the alignment of the signals themselves [25]:

$$\mathbb{E} \hat{c}(A, B) = \cos(\mu_A, \mu_B) \sqrt{c_A c_B} \implies \tilde{c}(A, B) = \frac{\hat{c}(A, B)}{\sqrt{c_A c_B}}. \quad (2)$$

More details on disattenuation may be found in Appendix B. As a reference for zero, the 280 dataset pairs that share neither task nor format class have mean -0.001 and standard deviation 0.008 (maximum 0.03), so we read values within ± 0.02 as null. Note, error bars in figures are ± 1 standard error across benchmark identities, so that they reflect whether the effect is consistent across tasks rather than estimator noise (Appendix E).

3.3 Experimental Setup

Models. To test whether the format ordering is a property of fully trained models only, we repeat the measurement at ten OLMo pretraining checkpoints (1B to 4T tokens) [26], at ten Pythia checkpoints

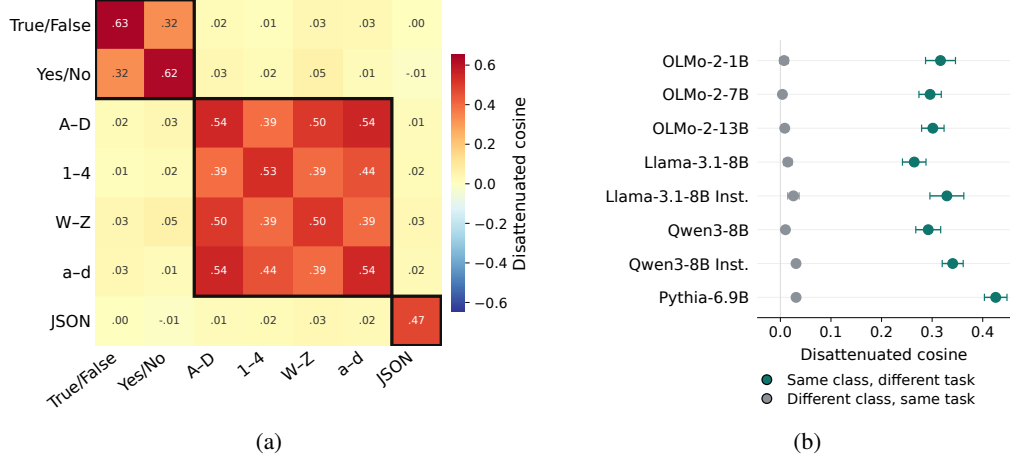


Figure 2: Gradient alignment follows answer format, not task. (a) Disattenuated cosine between all 35 datasets, averaged over four model families. Boxes mark the three answer-structure classes: alignment is high inside a class whatever the task, and at zero across classes even for the same task. Cells on the diagonal denote alignment for same answer format, different task pairs, while cells off-diagonal denote same task, different answer format pairs. (b) The same two contrasts measured per model, spanning 1B to 13B and four families.

[27], and at the base, SFT, DPO, and Instruct stages of OLMo. To test whether it is specific to a scale or model family, we repeat it at three OLMo scales (1B, 7B, 13B) and on Llama 3.1 [28] and Qwen3 [29] models, both base and instruction-tuned.

LESS selections. We analyze the selections released by Xia et al. [4] at `princeton-nlp/less_data`: the top 5% (13,533 examples) of a 270K-example pool for each of three targets (MMLU [30], BBH [31], TydiQA [32]) and three seeds. We label every selected example and the pool by answer format with a rule-based classifier (~90% agreement with manual labels on a held-out sample), and report enrichment as the share of a format among selected examples divided by its share of the pool. Refer to Appendix A for more details on the classifier and Appendix D for more details on pool composition.

4 Results

4.1 Answer Format, Not Task, Governs Gradient Alignment

Figure 2a shows cross-dataset cosine similarity while Figure 2b shows the same, aggregated for each model. The heatmap shows that same format class dataset pairs have high alignment, while dataset pairs with different format classes have near zero alignment. This implies that the gradient encodes more for answer format than task semantics. Figure 2b shows that format dominance in the gradient is consistent across model scale and family.

4.2 Format Dominance Holds Throughout Pretraining and Post-Training

We show that surface form dominance in gradient similarity is not an artifact of the amount of model pre-training or post-training in Figure 3. The figure shows that different benchmark, same answer class cosine similarity stays above the same benchmark, different answer class similarity.

4.3 LESS Selects for the Target’s Answer Format

We analyze the selections that Xia et al. [4] released for each target benchmark as shown in Figure 4. Although the pool contains no data from any of the three target benchmarks, each target’s selections are enriched in the target’s own answer format: letter answers are selected at 3.7 times their pool share for MMLU, chain-of-thought at 2.4 times for BBH, and short answers at 1.9 times for TydiQA.

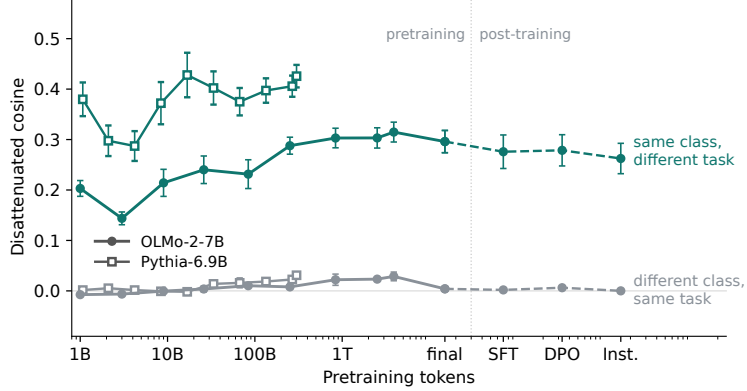


Figure 3: Answer form decides gradient similarity at every pre-training checkpoint.

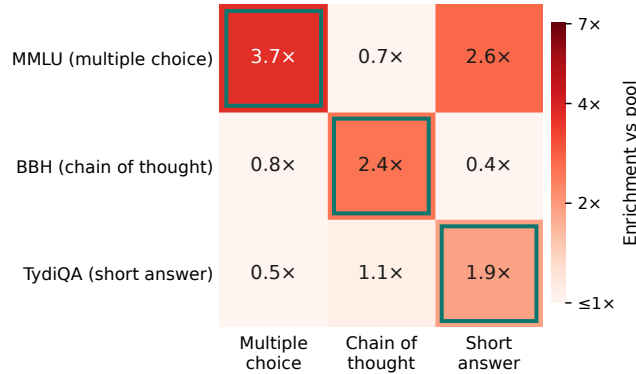


Figure 4: LESS’s released selections, by answer format, for the three target formats. Each cell is the share of a format among a target’s selections divided by its share of the pool; teal boxes mark each target’s own format.

The figure including analysis of other answer formats are in Figure 5. The pattern is stable across the three released seeds.

5 Discussion & Future Work

Our results bear on the reliability and robustness of gradient-based attribution and selection: their efficacy may partly be an artifact of format coinciding with semantics in current evaluation settings, and may not transfer to pools where the two decouple. More broadly, researchers should test gradient-based data attribution and selection methods against format-controlled baselines before relying on them. Future work includes interventional studies: For example, whether unifying a pool’s answer format makes selection more content-driven and fine-tuning more effective. Additionally, using learned parameter weightings that separate style from content may remove format dominance.

6 Limitations

We assume that task relatedness is a desirable factor in data attribution and especially selection. However, format-matched data often genuinely improves benchmark scores in the context of targeted instruction tuning; we falsify the semantic *interpretation* behind gradient-based selection to question its robustness, not its measured gains. Additionally, the scope of our experiments may be enlarged: the measurement basis is five English question-answering benchmarks and seven answer formats in three structure classes, with JSON as the only structured-generation member. The answer format classifications are also our choices rather than the result of rigorous analysis.

References

- [1] Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamilė Lukošiušė, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Samuel R. Bowman. Studying large language model generalization with influence functions, 2023. URL <https://arxiv.org/abs/2308.03296>.
- [2] Sang Keun Choe, Hwijeen Ahn, Juhan Bae, Kewen Zhao, Youngseog Chung, Adithya Pratapa, Willie Neiswanger, Emma Strubell, Teruko Mitamura, Jeff Schneider, Eduard Hovy, Roger Baker Grosse, and Eric P. Xing. What is your data worth to GPT? LLM-scale data valuation with influence functions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=zPKeJAEo27>.
- [3] Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. Trak: Attributing model behavior at scale. In *International Conference on Machine Learning (ICML)*, 2023.
- [4] Mengzhou Xia, Sathika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *International Conference on Machine Learning (ICML)*, 2024.
- [5] Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19920–19930. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/e6385d39ec9394f2f3a354d9d2b88eec-Paper.pdf.
- [6] Yihan Wang, Si Si, Daliang Li, Michal Lukasik, Felix Yu, Cho-Jui Hsieh, Inderjit S Dhillon, and Sanjiv Kumar. Two-stage LLM fine-tuning with less specialization and more generalization. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=pCEgna6Qco>.
- [7] Tyler A. Chang, Dheeraj Rajagopal, Tolga Bolukbasi, Lucas Dixon, and Ian Tenney. Scalable influence and fact tracing for large language model pretraining. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=gLa96F1Wwn>.
- [8] Yang Zhao, Li Du, Xiao Ding, Yangou Ouyang, Hepeng Wang, Kai Xiong, Jinglong Gao, Zhouhao Sun, Dongliang Xu, Qing Yang, Dongchen Li, Bing Qin, and Ting Liu. Beyond similarity: A gradient-based graph method for instruction tuning data selection. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24391–24404, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1189. URL <https://aclanthology.org/2025.acl-long.1189/>.
- [9] Yinghao Li, Vianne R. Gao, Chao Zhang, and MohamadAli Torkamani. Ensembles of low-rank expert adapters. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=l0gZS0sAlf>.
- [10] Jipeng Zhang, Yaxuan Qin, Renjie Pi, Weizhong Zhang, Rui Pan, and Tong Zhang. TAGCOS: Task-agnostic gradient clustered coreset selection for instruction tuning data. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4686–4701, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.264. URL <https://aclanthology.org/2025.findings-naacl.264/>.
- [11] Bing Wang, Ximing Li, Changchun Li, Jinjin Chi, Gang Niu, and Masashi Sugiyama. Decomposing the basic abilities of large language models: Mitigating cross-task interference in multi-task instruct-tuning. In *Forty-third International Conference on Machine Learning*, 2026. URL <https://openreview.net/forum?id=FFAHL32Wok>.

- [12] Siqi Kou, Qingyuan Tian, Hanwen Xu, Zihao Zeng, and Zhijie Deng. Which data attributes stimulate math and code reasoning? an investigation via influence functions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=b7uni0w0sZ>.
- [13] Laura Ruis, Maximilian Mozes, Juhan Bae, Siddhartha Rao Kamalakara, Dwaraknath Gnaneshwar, Acyr Locatelli, Robert Kirk, Tim Rocktäschel, Edward Grefenstette, and Max Bartolo. Procedural knowledge in pretraining drives reasoning in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=1hQKHUUsMx>.
- [14] Tingyu Xia, Bowen Yu, Kai Dang, An Yang, Yuan Wu, Yuan Tian, Yi Chang, and Junyang Lin. Rethinking data selection at scale: Random selection is almost all you need. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2698–2711, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.146. URL <https://aclanthology.org/2025.findings-emnlp.146/>.
- [15] Jingwei Ni, Zhijing Jin, Qian Wang, Mrinmaya Sachan, and Markus Leippold. When does aggregating multiple skills with multi-task learning work? a case study in financial NLP. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7465–7488, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.412. URL <https://aclanthology.org/2023.acl-long.412/>.
- [16] Qirun Dai, Dylan Zhang, Jiaqi W. Ma, and Hao Peng. Improving influence-based instruction tuning data selection for balanced learning of diverse capabilities. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 7079–7102, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.373. URL <https://aclanthology.org/2025.findings-emnlp.373/>.
- [17] Luxi He, Mengzhou Xia, and Peter Henderson. What is in your safe data? identifying benign data that breaks safety. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=Hi8jKh4HE9>.
- [18] Yuxi Xia, Loris Schoenegger, and Benjamin Roth. Influential training data retrieval for explaining verbalized confidence of llms. In Ricardo Campos, Adam Jatowt, Yanyan Lan, Mohammad Aliannejadi, Christine Bauer, Sean MacAvaney, Avishek Anand, Zhaochun Ren, Suzan Verberne, Nan Bai, and Masoud Mansoury, editors, *Advances in Information Retrieval*, pages 529–547, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-21289-4.
- [19] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL <https://aclanthology.org/P17-1147/>.
- [20] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL <https://aclanthology.org/D16-1264/>.
- [21] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

- [22] 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>.
- [23] Alex Gu, Baptiste Roziere, Hugh James Leather, Armando Solar-Lezama, Gabriel Synnaeve, and Sida Wang. CRUXEval: A benchmark for code reasoning, understanding and execution. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 16568–16621. PMLR, 21–27 Jul 2024. URL <https://proceedings.mlr.press/v235/gu24c.html>.
- [24] Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg. Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML ’09, page 1113–1120, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553516. URL <https://doi.org/10.1145/1553374.1553516>.
- [25] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. ISSN 00029556. URL <http://www.jstor.org/stable/1412159>.
- [26] Team OLMO, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious, 2025. URL <https://arxiv.org/abs/2501.00656>.
- [27] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usven Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR, 23–29 Jul 2023. URL <https://proceedings.mlr.press/v202/biderman23a.html>.
- [28] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yearry, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz

Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kam-badur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Celebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Rapparthi, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collet, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Conguet, Virginie Do, Vish Vogeti, Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyan Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkan Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippas Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ

- Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- [29] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.
- [30] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=d7KBjmI3GmQ>.
- [31] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.824. URL <https://aclanthology.org/2023.findings-acl.824/>.
- [32] Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470, 07 2020. ISSN 2307-387X. doi: 10.1162/tacl_a_00317. URL https://doi.org/10.1162/tacl_a_00317.
- [33] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and Adam Roberts. The flan collection: Designing data and methods for effective instruction tuning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22631–22648. PMLR, 23–29 Jul 2023. URL <https://proceedings.mlr.press/v202/longpre23a.html>.
- [34] Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. Openassistant conversations - democratizing large language model alignment. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 47669–47681. Curran Associates, Inc., 2023. doi: 10.52202/075280-2064.

URL https://proceedings.neurips.cc/paper_files/paper/2023/file/949f0f8f32267d297c2d4e3ee10a2e7e-Paper-Datasets_and_Benchmarks.pdf.

- [35] Databricks. Free dolly: Introducing the world’s first truly open instruction-tuned llm. Blog post, 2023. URL <https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm>.

NeurIPS Paper Checklist

1. **Claims.** Answer: Yes. The abstract and introduction state the three contributions, each matched by a results subsection.
2. **Limitations.** Answer: Yes. See the Limitations section.
3. **Theory, assumptions and proofs.** Answer: Yes. The one identity used (Eq. 2) states its assumption in the text; the derivation is in Appendix B.
4. **Experimental result reproducibility.** Answer: Yes. All models are open-weight, all datasets public, and the Method and Setup sections specify the estimator (examples per dataset, projection, parameter scope, masking); classifier rules are in Appendix A.
5. **Open access to data and code.** Answer: No. All datasets and models used are publicly available and cited; our measurement code is not yet released.
6. **Experimental setting/details.** Answer: Yes. See Method and Experimental Setup.
7. **Error bars.** Answer: Yes. Error bars are ± 1 standard error over benchmark groups, defined in Experimental Setup; the LESS analysis reports all three released seeds.
8. **Compute resources.** Answer: Yes. See Appendix F (Compute).
9. **Code of ethics.** Answer: Yes.
10. **Broader impacts.** Answer: Yes. The work analyzes existing methods; we foresee no direct negative societal impact, and a clearer picture of what drives data selection may improve curation practice.
11. **Safeguards.** Answer: NA. No models or high-risk assets are released.
12. **Licenses for existing assets.** Answer: Yes. All datasets and models are cited and used under their released licenses and terms.
13. **New assets.** Answer: NA. No new assets are released.
14. **Crowdsourcing and human subjects.** Answer: NA.
15. **IRB approvals.** Answer: NA.
16. **LLM usage.** Answer: NA. LLMs are not a component of the methodology; the format classifier is rule-based.

A Post-hoc Classifier Details

Each example is labeled by rules applied to its assistant turn (and its user turn, to detect offered options), in priority order: (1) a bare yes/no/true/false answer is *binary*; (2) an answer longer than 25 words containing reasoning markers or a trailing “the answer is (X)” is *chain of thought*, checked before the letter rule so that rationales ending in a letter are not misrouted; (3) a bare option letter, with options present in the prompt, is *choice*; (4) a bare number is *numeric*; (5) an answer that parses as JSON is *JSON*; (6) an answer of at most 12 words is *short answer*; (7) anything else is *long generation*. Numeric is merged into short answer in all reported results, as both are short constrained spans. On 30 randomly sampled examples labeled by hand, the classifier agreed on 27 (90%). Pool base rates are computed from 5,000 sampled examples per source, weighted by the true source sizes.

B Disattenuation Identity

Assume each unit-normalized example gradient of dataset A decomposes as $g_i = \sqrt{c_A} \hat{\mu}_A + \sqrt{1 - c_A} \eta_i$, where $\hat{\mu}_A$ is the unit dataset-level direction and the noise terms η_i have mean zero and are independent across examples and datasets. Then for examples $i \neq j$ within A , $\mathbb{E}\langle g_i, g_j \rangle = c_A$, so c_A is by construction the expected within-dataset cosine of Equation 1. Across datasets, all noise cross-terms vanish in expectation, leaving $\mathbb{E}\langle g_i^A, g_j^B \rangle = \sqrt{c_A c_B} \cos(\mu_A, \mu_B)$, which is Equation 2; dividing the measured mean by $\sqrt{c_A c_B}$ therefore recovers $\cos(\mu_A, \mu_B)$.

C Benchmarks and Answer Formats

Table 1 lists the five benchmarks and Table 2 the seven answer formats and their structure classes.

Table 1: The five benchmarks we cross with the answer formats of Table 2; each crossing of answer format with a benchmark serves as a *dataset*, totaling 35 datasets. For the four-option renderings, ARC supplies its own hand-written distractors and GSM8K uses near-miss arithmetic values; the rest draw distractors from other answers in the same dataset, so the options are always in-distribution.

Benchmark	Ability probed	Native answer	Example
TriviaQA	open-domain recall	short entity	PULSAR
SQuAD	reading comprehension	span of passage	1852
GSM8K	grade-school math	integer	109
ARC-Challenge	science reasoning	short phrase	a lightning strike
CRUXEval	code output prediction	Python value	True

Table 2: The seven answer formats we cross with each benchmark, grouped by the structure of the answer the model must produce.

Structure class	Answer formats	Example answer
Binary judgment	True/False, Yes/No	True
Four-option selection	A–D, 1–4, W–Z, a–d	B
Structured generation	JSON	{"answer": "Paris"}

D The LESS Selection Pool

Table 3 gives the composition of LESS’s 270,679-example selection pool: the four source datasets, their sizes, and the answer-format shares our classifier assigns within each source. The bottom row is the source-weighted pool base rate, the denominator of every enrichment in Figure 4. Note that our classifier labels only 6% of the CoT source as chain-of-thought format, since rationales without explicit reasoning markers fall under long generation. Relabeling every CoT-source example as chain-of-thought instead gives BBH $1.7\times$ on its own format (versus $2.4\times$ under the rule-based labels), leaves the diagonal of Figure 4 the maximum of every row, and changes no other conclusion.

E Aggregation and Error Bars

The trajectory and per-model figures reduce the full grid to two contrasts, and their error bars are designed to answer one question: Does the format effect depend on which benchmark was used? Each contrast is therefore aggregated in two stages, first averaging within a grouping unit defined by benchmark identity, then reporting the mean and ± 1 standard error across those units.

Same class, different task. This contrast compares datasets from two different benchmarks that share a format class, so the grouping unit is a benchmark pair. With five benchmarks there are ten pairs. Within each pair we average the disattenuated cosine over every same-class format combination (for the pair GSM8K–ARC, for example, GSM8K rendered as A–D against ARC rendered as 1–4, GSM8K as True/False against ARC as Yes/No, and so on), yielding one value per pair; the plotted point is the mean over the ten pairs and the error bar their standard error.

Table 3: Composition of the LESS selection pool [33–35]. Format shares are estimated from 5,000 sampled examples per source and weighted by true source sizes. Formats are labeled by the rule-based classifier of Appendix A; rationales without explicit reasoning markers fall under long generation, so the chain-of-thought base rate is conservative.

Source	Size	Share	Multiple choice	Yes/no	JSON	Short answer	Chain of thought	Long generation
Flan V2	100,000	36.9%	1.6%	8.1%	0.3%	54.4%	0.5%	35.0%
CoT	100,000	36.9%	0.7%	0.0%	0.0%	4.4%	6.4%	88.5%
Dolly	15,011	5.5%	0.0%	0.2%	0.0%	25.7%	1.4%	72.7%
OASST1	55,668	20.6%	0.0%	0.1%	0.0%	9.7%	2.6%	87.6%
Pool (weighted)	270,679	100.0%	0.9%	3.0%	0.1%	25.1%	3.2%	67.7%

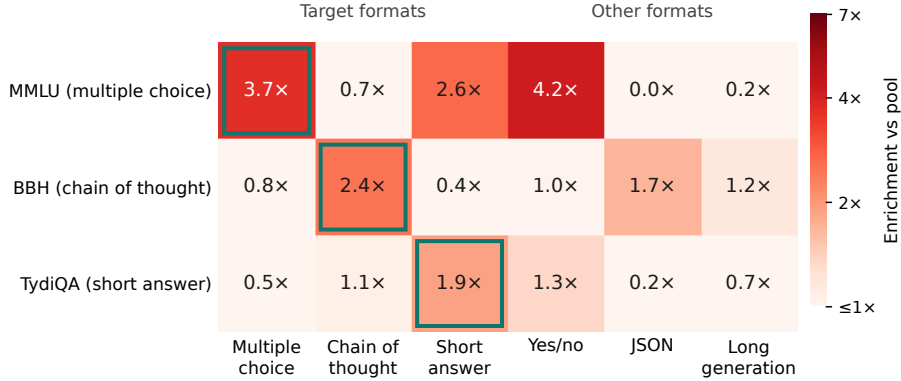


Figure 5: Enrichment of LESS’s selections over all six answer formats; the left block is Figure 4. Formats outside each target’s structure class are depleted (for MMLU, long generation at $0.2\times$ and JSON at $0.04\times$).

Different class, same task. This contrast compares a benchmark with itself rendered in different format classes, so no benchmark pair is involved and the grouping unit is the single benchmark. With five benchmarks there are five units. Within each we average over every cross-class format combination (for GSM8K, A–D against True/False, JSON against Yes/No, and so on), and the plotted point is the mean and standard error over the five.

Grouping by benchmark identity is deliberate. Each cell already averages 64×64 example pairs, so a standard error over example pairs would be very small and would measure only estimator precision. The standard error over benchmarks instead measures whether the finding generalizes across task content, which is the uncertainty relevant to our claim; the ten-versus-five asymmetry simply reflects how many benchmark identities each contrast involves.

F Compute

All gradient measurements ran on NVIDIA RTX A6000 GPUs (48 GB), using at most three concurrently. Measuring the full grid for one 7B-scale checkpoint takes about one GPU-hour; the 29 reported checkpoints total roughly 30 GPU-hours. The analysis of LESS’s released selections runs on CPU only.