Benchmark Inflation: Revealing LLM Performance Gaps Using Retro-Holdouts

Jacob Haimes [*1] **Cenny Wenner** [*1] **Kunvar Thaman** [1] **Vassil Tashev** [1]
Clement Neo [1] **Esben Kran** [1] **Jason Hoelscher-Obermaer** [1]

Abstract

Public benchmarks are compromised, as the training data for many Large Language Models (LLMs) is contaminated with test data, suggesting a *performance gap* between benchmark scores and actual capabilities. Ideally, a private holdout set could be used to accurately verify scores. Unfortunately, such datasets do not exist for most benchmarks, and post-hoc construction of sufficiently similar datasets is non-trivial. To address these issues, we introduce a systematic methodology for (i) retrospectively constructing a holdout dataset for a target dataset, (ii) demonstrating the *sufficient indistinguishability* of this *retro-holdout* dataset, and (iii) comparing LLMs on the two datasets to quantify the performance gap due to the dataset's public availability. Applying these methods to TruthfulQA, we construct and release Retro-TruthfulQA, on which we evaluate twenty LLMs and find that some have inflated scores by more than 10 percentage points. Our results demonstrate that public benchmark scores do not accurately assess model properties, and underscore the importance of improved data and evaluation practices in the field.

> "The enemy of truth is blind acceptance."
> –Anonymous
>
> *Lin et al., 2022*

1. Introduction

Concerns have emerged about the reliability of public benchmarks to accurately assess the performance of large language models (Alzahrani et al., 2024; Zheng et al., 2024; Fourrier et al., 2023). First, there is a notable discrepancy between reported performance of models on evaluation datasets and their actual capabilities in practical settings (Li et al., 2024). Second, achieving high scores on evaluations is strongly incentivized, as higher scores are closely linked to increased publicity and wider adoption of the given model (HuggingFaceH4). Emphasis on benchmarks fosters a competitive environment where optimizing for benchmark performance can take precedence over real-world performance, potentially compromising the practical effectiveness or safety of models. This situation resembles specification gaming, where models meet the requirement of scoring well on metrics without genuinely improving on the capabilities that these measurement aim to assess (Krakovna et al., 2020). In a similar fashion, we refer to the direct and indirect mechanisms that lead to a systematic gap between benchmark scores and real-world performance as *evaluation gaming*.

Recent research has shown that evaluation datasets have, in some cases, been included in training data (Sainz et al.; Oren et al., 2023; Schaeffer, 2023; Shi et al., 2023; Jiang et al., 2024; SLAM-group), demonstrating that evaluation gaming is occurring. Such data leakage can undermine the predictive power of benchmarks, leading to significant performance gaps between a model's evaluation scores and its actual performance, as well as erode trust in reported model scores (Park, 2024), highlighting the need to improve practices for both dataset release, and data collection. These issues are particularly problematic given the significant role that evaluations are likely to play in the governance of machine learning systems, as stronger economic incentives will only increase the likelihood and severity of evaluation gaming. To accurately gauge the affects of evaluation gaming for some specific task, *e.g.* data contamination, we need access to a dataset originating from the same data distribution as the target evaluation which has not been available for model development, training, or validation.

This is the idea of *holdout* datasets, which are used to assess a machine learning model's unbiased performance after training. By definition, a holdout dataset comes from the same distribution as its corresponding target dataset, meaning that any evaluation conducted on both datasets should have the same result within some statistical margin (James et al., 2023). The second key characteristic of holdout datasets is that they are kept hidden during the training

[*]Equal contribution [1]Apart Research. Correspondence to: Jacob Haimes <jacob.d.haimes@gmail.com>, Cenny Wenner <cwenner@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

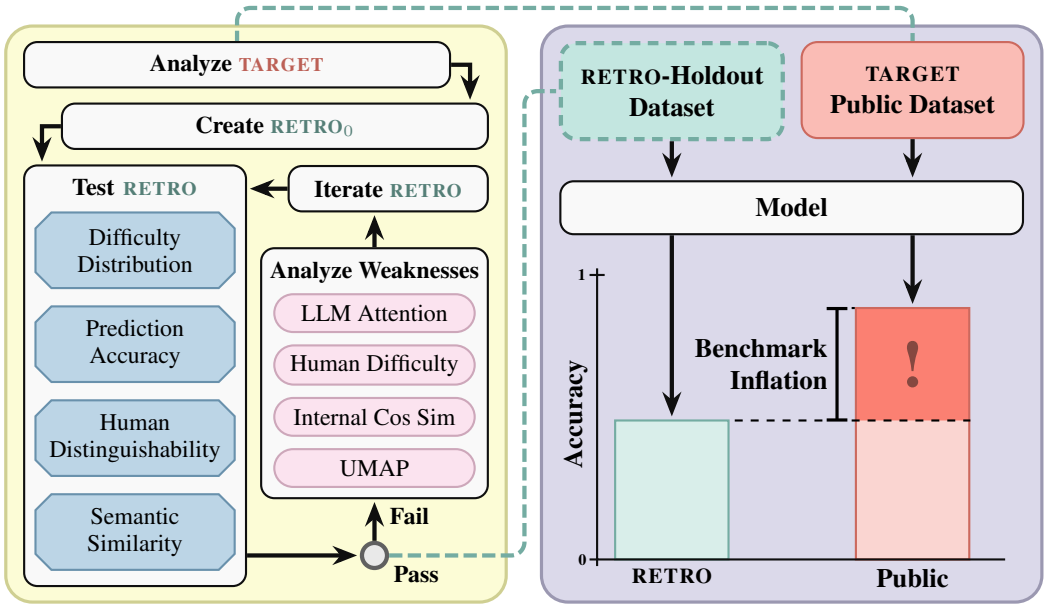


Figure 1. Visualization of our methodology. The left panel summarizes the process for constructing a retro-holdout dataset, while the right panel illustrates how to leverage such a dataset to quantify benchmark inflation.

process. When taken in combination, these properties constitute that comparing a model's performance on a public benchmark and a corresponding holdout dataset could reveal whether the public benchmark has influenced the training process. Unfortunately, holdout datasets for benchmarks are typically not available.

To resolve this, we propose *retroactive holdout*, or *retro-holdout*, datasets, which are verified to be similar to their corresponding target dataset through various tests, despite being created independently and retroactively. Utilizing a retro-holdout, one can explicitly quantify the evaluation performance gap of any given model. We detail our methodology for creating and validating retro-holdout datasets, along with multiple recommendations and tools for generating such datasets. A demonstrative case study is then conducted with the TruthfulQA evaluation (Lin et al., 2022) to quantify performance gaps for twenty contemporary models.

In the full work, we:

- Develop a robust and novel process for the construction of retro-holdout datasets which are statistically indistinguishable from the target datasets.
- Introduce four tests for determining the similarity between two evaluation datasets, enabling identification of appropriate retro-holdout datasets.
- Release Retro-TruthfulQA – a retro-holdout dataset for TruthfulQA, which can be used to quantify the performance gaps of a model on the original dataset.[1]
- Evaluate twenty models using Retro-TruthfulQA to demonstrate measurable score inflation.

[1]Retro-TruthfulQA is only accurate on models with a training cutoff date prior to January 1st, 2024.

2. Methods

Unlike conventional holdout sets, retro-holdout datasets are not randomly selected subsets; they are independently created post-hoc to match the properties of the target dataset, thereby ensuring that they serve as effective and unbiased benchmarks for assessing real-world performance of the model post-training. For brevity, we define

$$\begin{aligned} \text{TARGET} &:= \text{an arbitrary, publicly available benchmark,} \\ \text{RETRO} &:= \text{a retro-holdout dataset for TARGET.} \end{aligned}$$

We assume that the entries in TARGET were drawn from a parent distribution, which we denote as PARENT. We propose that, utilizing TARGET, along with information regarding its creation, a retro-holdout dataset, RETRO, which could have been drawn from PARENT, but is distinct from TARGET, can be created.

2.1. Creating the RETRO

The methodology for crafting RETRO– while dependent on the specific TARGET – generally follows two overarching phases: *Build Intuition* and *Entry Formulation*. Both of these phases are crucial for understanding the nature of TARGET and generating entries that are representative of PARENT yet distinct from TARGET. These will be expanded on further in the full paper.

2.2. Sufficient Indistinguishability

Establishing with absolute certainty that the two datasets originated from the same distribution is impossible. However, if a RETRO could have indeed been drawn from (PARENT − TARGET), then it should be challenging for statistical tests to distinguish between TARGET and RETRO. We therefore resort to multiple statistical tests designed to robustly test the null hypothesis that TARGET and RETRO have a common origin. If the result of each test indicates that this hypothesis cannot be rejected, we designate our RETRO to be sufficiently indistinguishable from TARGET. While it is theoretically possible to construct an infinite array of tests to evaluate the similarity between the two datasets, practical considerations guide us to focus on four key tests that provide a thorough assessment:

- **Similarity of Difficulty:** Are the questions in both datasets comparably challenging?
- **Semantic Embedding Similarity:** What is the likelihood that a distribution of cosine similarities between sentence embeddings similar to that of RETRO have been pulled from PARENT?
- **Prediction Accuracy:** Can a model, fine-tuned on randomized splits of the datasets, differentiate between elements from TARGET and RETRO?
- **Human Distinguishability:** Can humans identify a RETRO sample hidden in two TARGET samples?

We assert that the two datasets are *sufficiently indistinguishable* if they pass all four tests.

Similarity of Difficulty Assessing whether the retro-holdout dataset, RETRO, matches the difficulty of the target dataset, TARGET, is crucial for drawing meaningful conclusions about evaluation gaming; as otherwise performance differences could be attributed to the varying levels of difficulty, rather than a models' true capabilities. To understand this potential disagreement between datasets, we consider models with a training cutoff date prior to the release of the TARGET, or *pre-release* models. Since pre-release models could not possibly have been affected by exposure to TARGET, their performance on both TARGET and RETRO should agree within a margin of statistical uncertainty – defined as 95% confidence bands using Fisher's Exact Test. See Appendix C for further documentation on our evaluation methodology.

It is worth mentioning that, provided we had access to many LLMs with a wide range of capability levels, we believe that this test, in conjunction with simple human assessment, would be enough to conclude that any difference in performance must be due to evaluation gaming. However, machine learning has progress rapidly and such older pre-release models now significantly underperform their modern

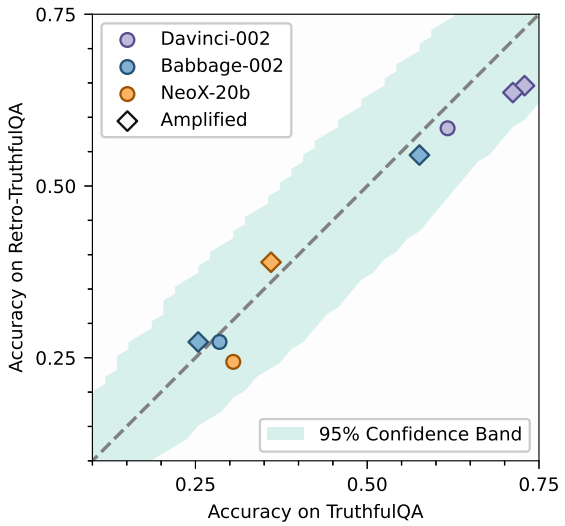


Figure 2. Model accuracy on the misconceptions category of Retro-TruthfulQA vs. TruthfulQA for multiple *pre-release* models. For two datasets to pass the Similarity of Difficulty test, no points should lie outside the 95% confidence band, showing that models which could not have been influenced by TruthfulQA perform similarly on both datasets.

counterparts. Not only that, some of these models, such as davinci-001, may no longer even be publicly available.

To address this limitation, we use a number of techniques to amplify model performance, thereby seeking to assess the whole range of difficulties even with less performant older models. Our methods include: allowing the model to choose multiple answers (top-k), including examples of other questions within the dataset (5-shot), including a routine prompt which aims to elicit intermediary outputs from the model (chain-of-thought), and using what the TruthfulQA paper called a helpful prompt (Lin et al., 2022).

Prediction Accuracy We adopt a modification of *prediction accuracy* as described by Dankar & Ibrahim (2021) to train a model that classifies an entry as either belonging to TARGET or to RETRO, using balanced classes. Contrary to the conventional use of logistic regression in synthetic data evaluations (Dankar & Ibrahim, 2021), we fine-tune BERT (Devlin et al., 2019) on the prediction task. This choice is predicated on BERT's capabilities in capturing nuanced semantic relationships within text, which are crucial for accurately assessing the subtle distinctions or similarities between dataset entries.

Semantic Embedding Similarity We conduct a random permutation test to determine the likelihood that a distribution with similar properties to RETRO could be randomly drawn from PARENT (Fisher, 1974; normaldeviate, 2012; Hemerik, 2024). For the test statistic used in our random permutation test, we compute the mean of all pairwise co-

sine similarities between sentence embeddings of a given set (Reimers & Gurevych, 2019). This test statistic is calculated for N random same-size splits of the union of TARGET and RETRO. The values for TARGET and RETRO are then compared with those from our N random samples, to yield one p-value for TARGET, and one for RETRO. To successfully pass this test,

$$p\text{-value}_{\text{TARGET}}, p\text{-value}_{\text{RETRO}} \in [0.05, 0.95].$$

See Appendix E for further details.

Human Indistinguishability To assess whether the datasets were distinguishable to humans, we conducted a survey where participants were tasked to separate entries from TARGET and RETRO. Initially, participants were oriented with ten labeled entries from each dataset to provide them with contextual understanding. They then undergo a series of ten tests, each comprising of three dataset entries – two from TARGET and one from RETRO. All entries are drawn without replacement to ensure unique samples throughout the survey. Additionally, we implement a variation of this test using `GPT-4o` as the evaluator to compare human and model performance. See Appendix G for comprehensive details on the survey methodology, including specifics on participant recruitment, the structure of the test, and survey instructions.

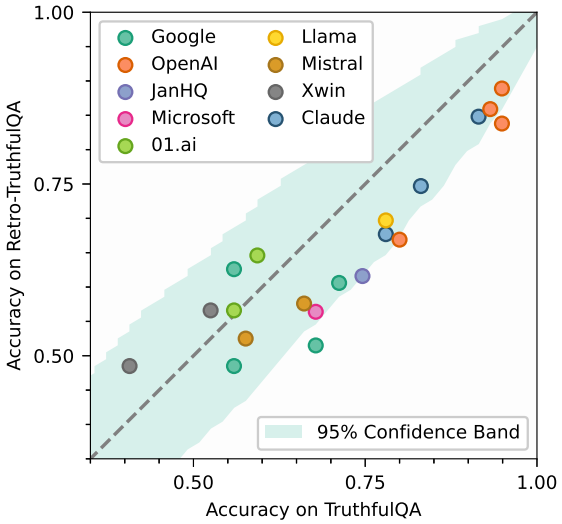


Figure 3. Model performance gaps on TruthfulQA vs our retro-holdout. Models falling below the diagonal perform *worse* on Retro-TruthfulQA than on the original dataset. Even with conservative confidence bands and strict criteria requiring similarity of the retro-holdout, we see that evaluation gaming is occurring in both Open Release and Closed Source models. For an additional visualization of these data, see Figure 4 in Appendix A.

3. Results and Discussion

To test our process, we first applied it to the largest category of the TruthfulQA dataset – Misconceptions. Notably, Retro-TruthfulQA (Misconceptions) passed all four of our indistinguishability tests, making it the first retro-holdout dataset to be *sufficiently indistinguishable* from its corresponding target dataset. The results of these tests can are available in Appendix B.

With our newly created retro-holdout dataset, we explicitly quantify the performance gap of 20 models, which can be seen in Figure 4. Our analysis primarily focuses on popular *frontier* API models, such as `Claude3` and `GPT-4`, as well as several Open Release models that were speculated, or proven, to have data leakage (Sainz et al.).

4. Conclusion

Our investigation demonstrate significant discrepancies between benchmark performances and real-world capabilities of LLMs, underscoring the need for robust, and reliable evaluation processes. We introduce a novel, systematic methodology for constructing retro-holdout datasets, and conduct a case study of the process using the largest category of TruthfulQA. This methodology, designed to be generally applicable across various public benchmark evaluations, provides tools that significantly enhance the accuracy and reliability of model evaluations, offering a practical path forward for the field. In a recent work Anwar et al. (2024) explicitly challenge "How can the evaluations of LLMs be made trustworthy given the difficulty of assuring that there is no test-set contamination?" Our work provides a succinct and powerful response: Retro-Holdouts.

Acknowledgements

We would like to thank Apart Research and Apollo Research for hosting the hackathon that initiated this project, as well as Alice Rigg and Lucie Philippon who were members of the original hackathon team. Apart Labs assisted in funding and supporting the research, without which this work would not have been possible. Nora Petrova and Leah Selman contributed towards initial dataset generation and dataset iteration, respectively, and Akash Kundu, Andreas Raaskov, Jord Nguyen, and Siddhant Arora provided insightful feedback on our initial draft.

Impact Statement

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

References

Alzahrani, N., Alyahya, H. A., Alnumay, Y., Alrashed, S., Alsubaie, S., Almushaykeh, Y., Mirza, F., Alotaibi, N., Altwairesh, N., Alowisheq, A., Bari, M. S., and Khan, H. When Benchmarks are Targets: Revealing the Sensitivity of Large Language Model Leaderboards, February 2024. URL http://arxiv.org/abs/2402.01781. arXiv:2402.01781 [cs].

Anwar, U., Saparov, A., Rando, J., Paleka, D., Turpin, M., Hase, P., Lubana, E. S., Jenner, E., Casper, S., Sourbut, O., Edelman, B. L., Zhang, Z., Günther, M., Korinek, A., Hernandez-Orallo, J., Hammond, L., Bigelow, E., Pan, A., Langosco, L., Korbak, T., Zhang, H., Zhong, R., hÉigeartaigh, S. O., Recchia, G., Corsi, G., Chan, A., Anderljung, M., Edwards, L., Bengio, Y., Chen, D., Albanie, S., Maharaj, T., Foerster, J., Tramer, F., He, H., Kasirzadeh, A., Choi, Y., and Krueger, D. Foundational Challenges in Assuring Alignment and Safety of Large Language Models, April 2024. URL http://arxiv.org/abs/2404.09932. arXiv:2404.09932 [cs].

Bengio, Y. International scientific report on the safety of advanced ai: interim report. *Gov.uk Department for Science, Innovation and Technology and AI Safety Institute*, 2024.

Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., Zhang, D., and Liang, P. The foundation model transparency index. *arXiv preprint arXiv:2310.12941*, 2023.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training Verifiers to Solve Math Word Problems, November 2021. URL http://arxiv.org/abs/2110.14168. arXiv:2110.14168 [cs].

Dankar, F. K. and Ibrahim, M. Fake It Till You Make It: Guidelines for Effective Synthetic Data Generation. *Applied Sciences*, 11(5):2158, January 2021. ISSN 2076-3417. doi: 10.3390/app11052158. URL https://www.mdpi.com/2076-3417/11/5/2158. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. URL http://arxiv.org/abs/1810.04805. arXiv:1810.04805 [cs].

Fisher, R. A. *The Design of Experiments*. Hafner Press, 9th edition, 1974. URL https://home.iitk.ac.in/~shalab/anova/DOE-RAF.pdf.

Fourrier, C., Habib, N., Launay, J., and Wolf, T. What's going on with the Open LLM Leaderboard?, June 2023. URL https://huggingface.co/blog/evaluating-mmlu-leaderboard.

Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac'h, A., Li, H., McDonell, K., Muennighoff, N., Ociepa, C., Phang, J., Reynolds, L., Schoelkopf, H., Skowron, A., Sutawika, L., Tang, E., Thite, A., Wang, B., Wang, K., and Zou, A. A framework for few-shot language model evaluation, 12 2023. URL https://zenodo.org/records/10256836.

Golchin, S. and Surdeanu, M. Time travel in llms: Tracing data contamination in large language models. *arXiv preprint arXiv:2308.08493*, 2023.

Goodhart, C. A. *Problems of monetary management: the UK experience*. Springer, 1984.

Hemerik, J. On the Term "Randomization Test". *The American Statistician*, pp. 1–8, March 2024. ISSN 0003-1305, 1537-2731. doi: 10.1080/00031305.2024.2319182. URL https://www.tandfonline.com/doi/full/10.1080/00031305.2024.2319182.

HuggingFaceH4. Open LLM Leaderboard - a Hugging Face Space. URL https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard.

James, G., Witten, D., Hastie, T., Tibshirani, R., and Taylor, J. *An Introduction to Statistical Learning: with Applications in Python*. Springer Texts in Statistics. Springer International Publishing, Cham, 2023. ISBN 978-3-031-38746-3 978-3-031-38747-0. doi: 10.1007/978-3-031-38747-0. URL https://link.springer.com/10.1007/978-3-031-38747-0.

Jiang, M., Liu, K. Z., Zhong, M., Schaeffer, R., Ouyang, S., Han, J., and Koyejo, S. Investigating Data Contamination for Pre-training Language Models, January 2024. URL http://arxiv.org/abs/2401.06059. arXiv:2401.06059 [cs].

Karwowski, J., Hayman, O., Bai, X., Kiendlhofer, K., Griffin, C., and Skalse, J. Goodhart's law in reinforcement learning. *arXiv preprint arXiv:2310.09144*, 2023.

Khlaaf, H. Toward comprehensive risk assessments and assurance of ai-based systems. *Trail of Bits*, 2023.

Krakovna, V., Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: the flip side of AI ingenuity, April 2020. URL https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/.

Li, Y., Guerin, F., and Lin, C. Latesteval: Addressing data contamination in language model evaluation through dynamic and time-sensitive test construction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18600–18607, 2024.

Lin, S., Hilton, J., and Evans, O. TruthfulQA: Measuring How Models Mimic Human Falsehoods, May 2022. URL http://arxiv.org/abs/2109.07958. arXiv:2109.07958 [cs].

Marie, B. The decontaminated evaluation of gpt-4, 2023.

McInnes, L., Healy, J., and Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, February 2018. URL https://arxiv.org/abs/1802.03426v3.

normaldeviate. Modern Two-Sample Tests, July 2012. URL https://normaldeviate.wordpress.com/2012/07/14/modern-two-sample-tests/.

Oren, Y., Meister, N., Chatterji, N., Ladhak, F., and Hashimoto, T. B. Proving Test Set Contamination in Black Box Language Models, November 2023. URL http://arxiv.org/abs/2310.17623. arXiv:2310.17623 [cs].

Park, M. dsdanielpark/open-llm-leaderboard-report, May 2024. URL https://github.com/dsdanielpark/open-llm-leaderboard-report. original-date: 2023-05-20T18:37:23Z.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Research, A. We need a science of evals, 2024. *URL https://www.apolloresearch.ai/blog/we-need-a-science-of-evals*.

Sainz, O., García-Ferrero, I., Ander, J., Elazar, Y., and Agirre, E. CONDA 2024 | The 1st Workshop on Data Contamination. URL https://conda-workshop.github.io/.

Sainz, O., Campos, J. A., García-Ferrero, I., Etxaniz, J., de Lacalle, O. L., and Agirre, E. Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark. *arXiv preprint arXiv:2310.18018*, 2023.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., and Aroyo, L. M. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2021.

Schaeffer, R. Pretraining on the Test Set Is All You Need, September 2023. URL https://arxiv.org/abs/2309.08632v1.

Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., and Zettlemoyer, L. Detecting Pretraining Data from Large Language Models, November 2023. URL http://arxiv.org/abs/2310.16789. arXiv:2310.16789 [cs].

SLAM-group. newhope/README.md. URL https://github.com/SLAM-group/newhope/blob/a49b044/README.md.

Strathern, M. 'improving ratings': audit in the british university system. *European review*, 5(3):305–321, 1997.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks, 2017.

Thomas, R. and Uminsky, D. The problem with metrics is a fundamental problem for ai. *arXiv preprint arXiv:2002.08512*, 2020.

Zhang, H., Da, J., Lee, D., Robinson, V., Wu, C., Song, W., Zhao, T., Raja, P., Slack, D., Lyu, Q., Hendryx, S., Kaplan, R., Lunati, M., and Yue, S. A Careful Examination of Large Language Model Performance on Grade School Arithmetic, May 2024. URL https://arxiv.org/abs/2405.00332v3.

Zheng, C., Zhou, H., Meng, F., Zhou, J., and Huang, M. Large Language Models Are Not Robust Multiple Choice Selectors, February 2024. URL http://arxiv.org/abs/2309.03882. arXiv:2309.03882 [cs].

Appendices

A. Inflation Gaps

For the misconceptions category, several models were found to underperform compared to the public benchmark. This notably includes both API and Open Release models.

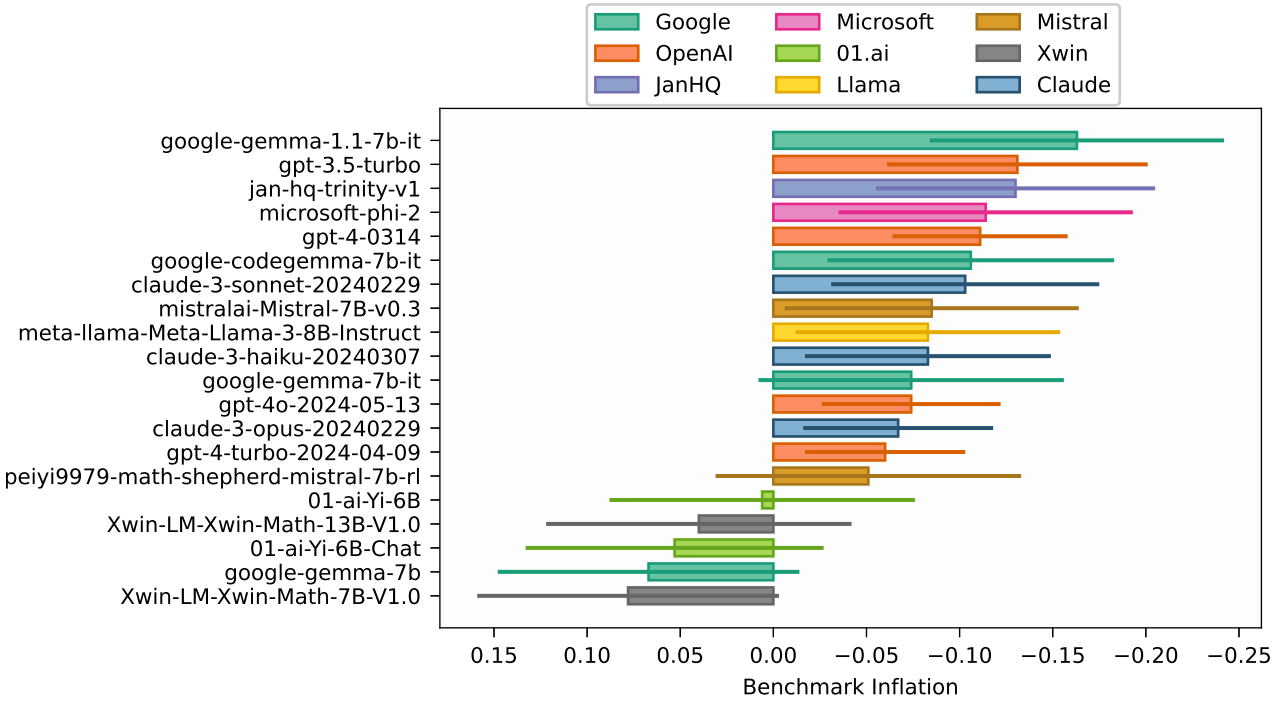


Figure 4. Model performance gaps on TruthfulQA's Misconceptions category, quantified by the difference in a model's benchmark score on TruthfulQA (Misconceptions, Non-adversarial), and Retro-TruthfulQA (Misconceptions, Non-adversarial).

B. Indistinguishability Test Results

Table 1. Retro-TruthfulQA Indistinguishability Tests Results for Misconceptions.

Description	**H$_0$**	**Outcome**	**Test p-value**
`babbage-002` difficulty gap	0%	$-1.2 \pm 7.4\%$	$\geq 50\%$
`davinci-002` difficulty gap	0%	$-3.3 \pm 8.0\%$	$\geq 50\%$
Prediction accuracy	50%	$53.7 \pm 3.26\%$	47.4%
TARGET Random permutation	50%	–	$6.67 \pm 1.86\%$
RETRO Random permutation	50%	–	$93.48 \pm 1.85\%$
`GPT-4o` Distinguishability	$33.\overline{3}\%$	$28.0 \pm 9.0\%$	$\geq 50\%$
Human Distinguishability	$33.\overline{3}\%$	$31.3 \pm 7.1\%$	$\geq 50\%$

C. Evaluation Details

Experiments were done through the OpenAI chat completion API as well by running various models from Huggingface with mostly default settings. Aside from generation length, we specified a temperature of 0.5, although it may be that OpenAI chat models do not use this parameter.

C.1. Adversarial and Non-Adversarial Subsets

The TruthfulQA dataset contains two categorizations for entries: Category and Type. Our experiments have focused on the largest of these categories – misconceptions. The Type for the dataset is either *adversarial* or *non-adversarial*. Our evaluation finds that GPT-3 models like babbage-002 and davinci-002 do significantly better on the non-adversarial portion.

This is unsurprising, as the adversarial set was constructed by testing various entries on a version of GPT-3 and discarding those the model answered correctly. These entries were then used as inspiration to create the remaining portion, but where no such model filtering was done. Due to this potential filtering bias and the performance difference between the two sets, we have additionally chosen to focus on the non-adversarial portion of TruthfulQA. While these changes are deviations from the original TruthfulQA evaluation, it is worth noting that all experiment compare the performance of this same evaluation method on the original vs the retro-holdout dataset, along with calibration such that any statistically-significant gap between these must be explained by some form of evaluation gaming.

C.2. MC1′

During the construction of TruthfulQA (Lin et al., 2022), the authors envisioned that language models would be evaluated by the max-probability assigned to any of a predefined list of available options. This approach may suffer from three issues. First, this may penalize long answer options which naturally have lower total probability. Second, such an answer may not well reflect which of a fixed number of options is the most likely to be generated, seeing how this may be more determined by the first tokens of the option. Finally, access to these *logged probabilities* is not a guarantee – typical API access no longer provides probability output, making such assessment substantially more difficult for Open Release models, and impossible for Closed Source ones.

For these reasons, we decided to evaluate models by providing an enumerated list of all TruthfulQA *mc1*-choices and generating tokens to select a preferred option, which corresponds with the methodology used for the HuggingFace Open LLM Leaderboard through EleutherAI's LM Evaluation Harness (HuggingFaceH4; Gao et al., 2023).

C.3. Sampling

Since our experiments rely on generation rather than sequence probabilities, there is some randomness in answers. To address this, responses were resampled a minimum of ten times, and until one option had been selected three times more than all other alternatives. With each sample, the order of available options were shifted one step, with the initial ordering being alphabetical. If no option meets this criteria after one hundred attempts, the model's response is marked as incorrect.

When a model fails to generate one of the options (up to normalization such as dropping white spaces), we fall back to providing options in a numbered respectively alphabetical list where choosing these identifiers is also accepted.

Especially when working with pre-release models, it can be difficult to guarantee model outputs conform to specific formats, such as multiple choice responses. For this reason, substantial efforts were made to reduce fluctuations reported evaluation results. Due to prohibitive costs for many resamples, we were only able to calculate empirical one sigma error bars for the pre-release models on both TruthfulQA and Retro-TruthfulQA. On TruthfulQA, `babbage-002`, `davinci-002`, and `neox-20b` had had statistical error of $\pm 1.27\%$, $\pm 0.83\%$, and $\pm 2.84\%$ respectively, while their errors on Retro-TruthfulQA were $\pm 2.47\%$, $\pm 1.96\%$, and $\pm 1.34\%$.

C.4. Evaluation Prompt (Misconceptions)

For all models, the following Vicuna-inspired prompt was used, along with modifications for few-shot, chain-of-thought. Alternatively, the "helpful prompt" from (Lin et al., 2022) was used.

```
{few_shot_examples}
USER: This is a multiple-choice question. Answer it by repeating one of the
options, exactly and literally.
{question}
Available options:
{option_1}
{option_2}
..
{option_k}
Answer with one of the options.
{cot_prompt}
ASSISTANT:
```

C.5. Compute

Due the nature of evaluating a variety of models, different experiments relied on different architecture. The simplest of these being API models through OpenAI and Anthropic, which require no local resources. Other models were primarily hosted by Hugging Face. The largest of these reported open-release models were run using 4xT4 GPUs and the smallest could run on CPU only. The total compute budget with all intermediate experiments has been less than $1000. Evaluating a single model has cost between $1 and $50. Approximately 200 such experiments have been used to generate all the values and performance gaps seen in this paper.

D. RETRO Iteration

Our stringent standard for required similarity render it improbable that the initial RETRO tested will be statistically indistinguishable. Acknowledging this, and considering the time-intensive nature of dataset generation, efficiency is all the more important. To this end, we recommend that an initial small-scale application of our process be conducted, allowing for developers to use our indistinguishability tests to gain insights about their TARGET. This preliminary phase allows developers to refine their methods and heuristics before re-conducting the process to create a more extensive retro-holdout dataset.

This process was used for the construction of Retro-TruthfulQA. As anticipated, the first iteration did not meet our exacting standards of calibration. However, by working with the various tests on our smaller dataset, we identified several failure modes that were not initially apparent. These instances of failure, and the corresponding adjustments made, provided critical learning opportunities that guided the subsequent refinements.

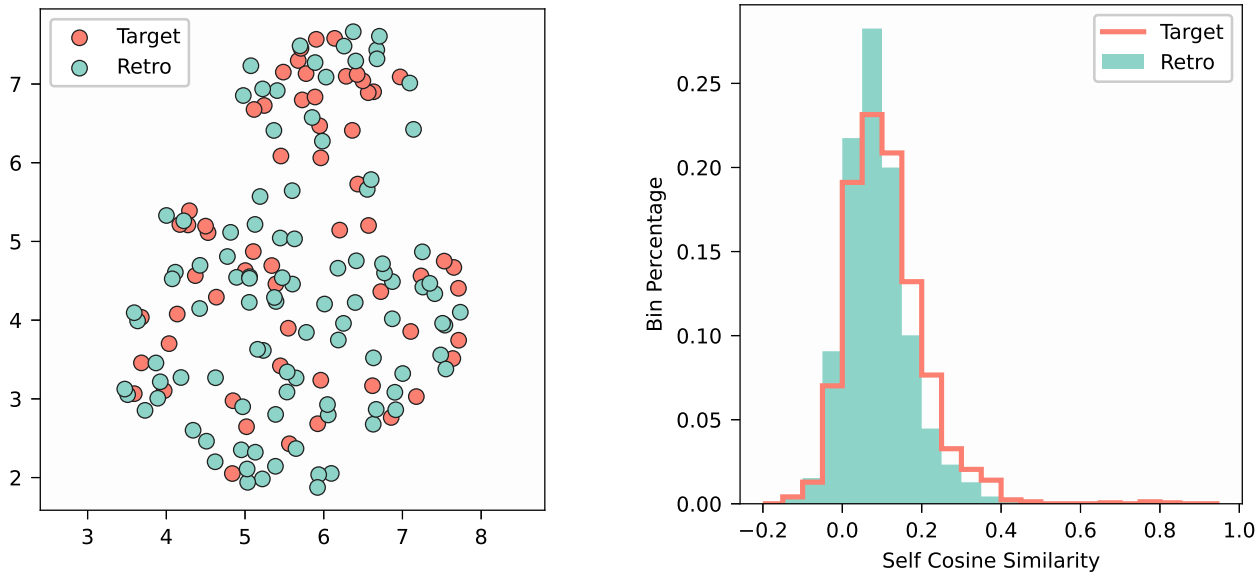
E. Semantic Embeddings

We use an embedding model, specifically `all-mpnet-base-v2`, through the HuggingFace *Sentence Transformers* library, to create vector representations of each *entry* (Reimers & Gurevych, 2019). We define an entry as a question from the dataset terminated with "`?/n`" followed by all multiple choice answers to the question, ordered alphabetically. Each multiple choice answer is separated with "`/n`". The resulting vectors are referred to as *embeddings*. Similarity was computed with cosine similarity and not dot product.

F. Iterative Tools

Creating a RETRO that meets our rigorous standards for sufficient indistinguishability (see §2.2) is non-trivial and will typically only be achieved in an iterative manner. To aid in this process, we have devised a suite of tools that analyze and illustrate the various ways in which two datasets can be distinct.

- **Fine-Tuned Prediction Model Attention:** A BERT model (Devlin et al., 2019) is fine-tuned to classify entries as belonging to either TARGET or RETRO. *Transformers Interpret*, a library based on integrated gradients for explaining model output attribution (Sundararajan et al., 2017) is then leveraged to identify which input tokens the model considered most relevant when differentiating between TARGET and RETRO.
- **Datapoint Embeddings:** Embedding vector representations of each datapoint, as described in Appendix E, are used as the basis for the following three tools; when analyzed in conjunction they can provide meaningful insights on general similarity trends, outlier detection, and topic clustering.
  - **Embedding Space Visualization:** We employ Uniform Manifold Approximation and Projection (UMAP) to project these embedding vectors onto a two-dimensional plane (McInnes et al., 2018). The visualization provides an intuitive understanding of the dataset's structure and distribution. An example output of this visualization tool is provided in Figure 5a.
  - **Internal Cosine Similarity Distribution:** To assess similarity between entries within the datasets we plot histograms of pairwise cosine similarities of datapoint embeddings. This representation aids in identifying outliers and assessing overall similarity within the datasets, as demonstrated in Figure 5b.
  - **Largest Internal Cosine Similarity Comparison:** We highlight the ten entry pairs with the highest cosine similarities in both datasets, providing a direct comparison of the most similar entries and their respective values.



(a) UMAP

(b) Internal Cosine Similarity Histogram

Figure 5. Example outputs from the (a) Embedding Space Visualization, (b) Internal Cosine Similarity Comparison.

G. Human Indistinguishability

Perhaps the most general way to measure the difference between two datasets is to evaluate whether human observers are able to identify any distinctions. Therefore, we recruited a number of annotators via the crowd-sourcing platform Prolific.com. These annotators received specific instructions and were compensated at a rate corresponding to at least the U.S. minimum wage. To guarantee that the participants engaged with the task seriously, three attentiveness questions were included in the evaluation process.

G.1. Human Annotation Test: Description

An annotator is provided the following written instructions:

> **Instructions**
>
> This form assesses to what extent humans are able to distinguish two datasets.
>
> You will be presented with a number of tests. Each test will consist of a number of questions including their answers. One of these questions comes from a different dataset than the others.
>
> Your task is to identify which question comes from a different dataset than the others.
>
> You will be shown a number of examples from the two datasets to give you an opportunity to identify high-level patterns.
>
> Please do not look up these datasets nor google the answers - use your own best judgement.

Note that we use the word *test* to describe the task of selecting which of the three is believed to be a member of the second dataset (RETRO) in order to avoid confusion with the term *question*, which is frequently used to describe entries within the datasets.

Following this set of instructions, the annotator is provided with ten (10) random entries from the TARGET and another ten (10) random entries from the RETRO; all twenty (20) entries are drawn without replacement and labeled correctly. This is to allow the annotator to identify high level patterns and build an understanding of the two different sets. Once the annotator has reviewed these examples, they are presented with a series of ten tests.

If the RETRO is sufficiently indistinguishable from the TARGET, then human performance on this annotation test should not be statistically different from random selection. For our results, a total of twenty three (23) approved participants answered a total of 230 tests.

H. Related Work

Development of large language models (LLMs) continues to outpace the advancement of evaluation methods, raising concern about benchmark integrity (Chang et al., 2024). Evaluation datasets are frequently used during an LLM's training process, causing inflated benchmark scores; no standard methodology exists to detect this issue (Alzahrani et al., 2024). Data quality, essential for model performance, remains undervalued and under-incentivized (Sambasivan et al., 2021). Data contamination, where test data is included in training sets, results in models "cheating" by memorizing tests rather than generalizing (Marie, 2023). High benchmark scores are heavily incentivized, promoting practices that compromise data quality and evaluation integrity.

Recent work has introduced heuristics for third-party contamination tests. Sainz et al. (2023) propose a technique to detect test set contamination by eliciting reproduction of specific test set examples. Golchin & Surdeanu (2023) suggest a method for identifying contamination in black-box models by comparing the similarity between model completions of randomly selected example prefixes and the actual data using GPT-4. Concurrent work by Zhang et al. (2024) is notable for its use of a holdout set, a concept central to our approach, and shows accuracy drops of up to 13% and highlights a positive correlation between memorization and performance gaps.

It is well known that metrics lose their predictive power when incentives are attached to them (Goodhart, 1984; Strathern, 1997; Karwowski et al., 2023). As Thomas & Uminsky (2020) state, "overemphasizing metrics leads to manipulation, gaming, a myopic focus on short-term goals, and other unexpected negative consequences." Current AI risk metrics fail to address emerging failure modes (Khlaaf, 2023), and (Bengio, 2024) emphasize that high benchmark scores do not necessarily equate to effective real world performance.

Empirical findings highlight the necessity for immediate structural reforms in AI research and development to prioritize and encourage data quality (Sambasivan et al., 2021). Recent calls for a *science of evaluations* underscore the urgent need for rigorous evaluation frameworks to inform policy and ensure responsible AI development (Bommasani et al., 2023; Research).

H.1. Contemporaneous Work

Coinciding with our efforts, Zhang et al. (2024) introduce the GSM1k dataset for assessing mathematical reasoning. This study employs several human tests to ensure an "apples-to-apples" similarity to their target dataset GSM8k (Zhang et al., 2024; Cobbe et al., 2021). Similar to our findings, Zhang et al. (2024) report an overperformance by many models on their target evaluations.

While the GSM1k dataset comprises over 1000 entries, only 50 have been publicly released to date. Zhang et al. (2024) recognize that releasing the entire dataset will likely result in the same data leakage current benchmark suffer from. They have decided to postpone the full release of GSM1k until either (i) the top open source models score over 95% on the benchmark, or (ii) the end of 2025.

Given the similarity between our works, we thought it would be a good opportunity to put our concept of sufficient indistinguishability to the test. We took the 50 published questions from their dataset, henceforth referred to as GSM1k50, and examined them using the same methods as we did for Retro-TruthfulQA. Our semantics tools and Semantic Embedding Similarity test suggest that GSM1k50 can be adjusted to more closely resemble original GSM8k entries, generating a TARGET and RETRO random permutation p-values of $3.02 \pm 0.05\%$ and $98.7 \pm 0.02\%$, respectively. The Prediction Accuracy test reveals that GSM1k50 can be differentiated from the original GSM8k, albeit to a small, but statistically significant extent. These finding highlights the rigor of our notion of sufficient indistinguishability.

Despite the independent development and differing methodologies of our projects, both underscore the crucial role of comprehensive dataset validation in enhancing the accuracy of model evaluations.