Revive Legacy Scientific Reasoning Benchmarks by Growing Perturbation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Large language model evaluation is compromised by data contamination, where sophisticated memorization masquerades as reasoning. We propose a systematically perturbed benchmark dataset that transforms static legacy evaluations into contamination-resistant resources. Four perturbation categories enable robust assessment of authentic scientific reasoning versus pattern matching while testing contamination resistance and problem solvability recognition.

1 AI Task Definition

This dataset addresses: *How can we reliably distinguish genuine scientific reasoning from sophisticated memorization in large language models?*

The dataset enables three interconnected tasks. **Robustness Assessment** involves binary classification predicting whether performance degradation indicates memorization versus authentic reasoning limitations. **Contamination Detection** predicts data leakage likelihood by comparing performance on original versus perturbed variants, enabling assessment of genuine AI capability as scientific reasoner. **Solvability Recognition** generates and evaluates mathematically impossible problems testing genuine reasoning versus hallucination tendencies.

2 Dataset Rationale

Static benchmarks enable memorization masquerading as reasoning. While models achieve near-perfect performance on GSM8K [2], MATH [5], GPQA [12], and MMLU [4], recent studies reveal fundamental limitations. GSM-Symbolic [9] and GSM1K [17] demonstrate dramatic failures when simple numerical values change, while PertEval [8] exposes vulnerabilities to knowledge-invariant modifications that should not affect genuine understanding.

Our dataset requires 100K+ perturbed variants sourced from established benchmarks including GSM8K [2], MATH [5], GPQA [12], MMLU [4], and MMMU [16]. Each variant includes comprehensive metadata covering perturbation type, solvability labels, and formal correctness proofs verified through Lean4 [3] theorem proving.

Knowledge-Invariant Perturbations apply surface modifications like variable renaming and contextual paraphrasing while preserving underlying solution pathways. Building on GSM1K [17] and MATH-Perturb [6] methodologies, these perturbations test whether models understand fundamental logical relationships or merely memorize superficial patterns. A model demonstrating genuine reasoning should maintain consistent performance across semantically equivalent problem formulations.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Knowledge-Variant Perturbations systematically scale problem complexity through constraint additions and difficulty increases. Extending MATH-Perturb [6] hard perturbations and MMLU-Pro [13] enhancement approaches, these modifications assess whether models can adapt reasoning strategies to increased complexity or rely on memorized solution templates that fail under scaling pressure.

Solvability-Constrained Perturbations inject mathematical contradictions creating unsolvable problems while maintaining surface plausibility. These perturbations provide the most direct test of genuine scientific reasoning by distinguishing models that recognize logical impossibility from those that generate plausible-sounding but fundamentally incorrect solutions through sophisticated hallucination.

Adversarial Perturbations employ gradient-optimized semantic triggers extending GCG [20] and CatAttack [11] approaches. These perturbations reveal systematic vulnerabilities in reasoning processes while preserving problem semantic validity, uncovering failure modes that indicate reliance on brittle pattern matching rather than robust logical understanding.

3 Acceleration Potential

Automated benchmark refreshing eliminates manual curation bottlenecks, accelerating development cycles from annual to monthly updates while preventing contamination-based gaming of genuine AI scientific reasoning capabilities. Real-time perturbation generation enables systematic detection of memorization versus authentic reasoning, providing reliable assessment of models as scientific reasoners rather than sophisticated pattern matchers.

Solvability-constrained perturbations offer unprecedented diagnostic capability by testing whether models recognize mathematically impossible constraints versus hallucinating solutions. This capability directly assesses genuine scientific reasoning foundations essential for physics simulations, diagnostic reasoning systems, educational assessment, and quantitative modeling under distribution shifts.

4 Data-Creation Pathway

We leverage automated perturbation pipelines across existing benchmarks, building on contamination-resistant methodologies from LiveBench [14], AntiLeak-Bench [15], and LiveCodeBench [7]. Symbolic manipulation engines extend GSM-Infinite [19] complexity scaling approaches while fact-preserving transformations build on PertEval [8] knowledge-invariant methods. Program-of-Thought [1, 18] integration enables computational verification while adversarial generation employs semantic preservation constraints. Formal verification through Lean4 [3] and Isabelle [10] ensures correctness and solvability labeling at unprecedented scale.

5 Cost & Scalability

We train specialized perturbator models for each category through targeted approaches. Knowledge-invariant perturbators extend PertEval [8] methodologies using masked language modeling for paraphrasing tasks while knowledge-variant perturbators build on MATH-Perturb [6] and GSM-Infinite [19] using reinforcement learning for complexity-preserving modifications. Solvability-constrained perturbators employ constraint-satisfaction models for systematic reasoning testing while adversarial perturbators extend GCG [20] and CatAttack [11] optimization techniques.

Our perturbation-as-a-service platform integrates Lean4 [3] verification capabilities, democratizing contamination-resistant evaluation for assessing genuine AI scientific reasoning capabilities across the research community.

References

[1] Zhen Bi, Ningyu Zhang, Yinuo Jiang, Shumin Deng, Guozhou Zheng, and Huajun Chen. When do program-of-thought works for reasoning? In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI*

2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada, pages 17691–17699. AAAI Press, 2024.

[2] 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. *CoRR*, abs/2110.14168, 2021.

[3] Leonardo de Moura and Sebastian Ullrich. The lean 4 theorem prover and programming language. In André Platzer and Geoff Sutcliffe, editors, *Automated Deduction - CADE 28 - 28th International Conference on Automated Deduction, Virtual Event, July 12-15, 2021, Proceedings*, volume 12699 of *Lecture Notes in Computer Science*, pages 625–635. Springer, 2021.

[4] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[5] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021.

[6] Kaixuan Huang, Jiacheng Guo, Zihao Li, Xiang Ji, Jiawei Ge, Wenzhe Li, Yingqing Guo, Tianle Cai, Hui Yuan, Runzhe Wang, Yue Wu, Ming Yin, Shange Tang, Yangsibo Huang, Chi Jin, Xinyun Chen, Chiyuan Zhang, and Mengdi Wang. MATH-Perturb: Benchmarking LLMs' math reasoning abilities against hard perturbations. *arXiv preprint arXiv:2502.06453*, 2025.

[7] 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. *ArXiv*, abs/2403.07974, 2024.

[8] Jiatong Li, Renjun Hu, Kunzhe Huang, Yan Zhuang, Qi Liu, Mengxiao Zhu, Xing Shi, and Wei Lin. Perteval: Unveiling real knowledge capacity of llms with knowledge-invariant perturbations. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

[9] Seyed-Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

[10] Tobias Nipkow, Lawrence C. Paulson, and Markus Wenzel. *Isabelle/HOL - A Proof Assistant for Higher-Order Logic*, volume 2283 of *Lecture Notes in Computer Science*. Springer, 2002.

[11] Meghana Rajeev, Rajkumar Ramamurthy, Prapti Trivedi, Vikas Yadav, Oluwanifemi Bamgbose, Sathwik Tejaswi Madhusudan, James Zou, and Nazneen Rajani. Cats confuse reasoning llm: Query agnostic adversarial triggers for reasoning models, 2025.

[12] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. *CoRR*, abs/2311.12022, 2023.

[13] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

[14] Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha V. Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. Livebench: A challenging, contamination-limited LLM benchmark. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

[15] Xiaobao Wu, Liangming Pan, Yuxi Xie, Ruiwen Zhou, Shuai Zhao, Yubo Ma, Mingzhe Du, Rui Mao, A. Luu, and William Yang Wang. Antileak-bench: Preventing data contamination by automatically constructing benchmarks with updated real-world knowledge. *ArXiv*, abs/2412.13670, 2024.

[16] Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 9556–9567. IEEE, 2024.

[17] Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, William 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. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

[18] Yu Zhang, Shujun Peng, Nengwu Wu, Xinhan Lin, Yang Hu, and Jie Tang. Rm-pot: Reformulating mathematical problems and solving via program of thoughts. *CoRR*, abs/2502.12589, 2025.

[19] Yang Zhou, Hongyi Liu, Zhuoming Chen, Yuandong Tian, and Beidi Chen. Gsm-infinite: How do your llms behave over infinitely increasing context length and reasoning complexity? *CoRR*, abs/2502.05252, 2025.

[20] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.