Proposal: Revive Legacy Benchmarks By Self-Evolving Perturbations

Terry Jingchen Zhang

Abstract

Capability evaluation of large language models faces growing challenges from potential benchmark contamination, which may obscure the distinction between genuine reasoning and memorization. Benchmark perturbation represents one promising approach to address these concerns by systematically modifying evaluation problems through various techniques such as altering numerical values, modifying contextual elements, or introducing controlled variations. However, the effectiveness of these approaches remains understudied across different domains and benchmark types.

We propose to investigate benchmark perturbation methods within a structured comparative framework, examining how different perturbation techniques perform across a selection of commonly used benchmarks. While previous work has typically focused on single benchmarks or specific perturbation types, our approach aims to provide a more comprehensive understanding of these methods' strengths and limitations. We acknowledge significant challenges in this endeavor, including the complexity of maintaining semantic equivalence across perturbations, the risk of introducing new biases, and the substantial validation requirements across different domains. Our goal is to contribute empirical insights that may inform the development of more robust evaluation practices, while recognizing the inherent limitations and ongoing debates in this area.

1 Introduction

The evaluation of large language models has become increasingly complex as models demonstrate remarkable performance across established benchmarks spanning mathematical reasoning tasks like GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021b), scientific knowledge assessments such as MMLU (Hendrycks et al., 2021a) and GPQA (Rein et al., 2023), and multimodal understanding through benchmarks like MMMU (Yue et al., 2024). However, this apparent success has been accompanied by growing concerns about benchmark data contamination, where language models may inadvertently incorporate evaluation benchmark information from their training data, leading to potentially inflated or unreliable performance measures.

Recent investigations have revealed concerning patterns in model behavior. Apple's GSM-Symbolic benchmark demonstrates that LLMs exhibit noticeable variance when responding to different instantiations of the same question, with performance declining when only numerical values are altered. Similarly, studies using GSM-Plus (Li et al., 2024) and other perturbation approaches suggest that models may be more fragile than initially apparent, though the extent and implications of these findings remain subjects of active research and debate.

The current evaluation landscape presents several interconnected challenges that we hope to better understand through systematic investigation. First, most established benchmarks consist of static problem sets that may become less reliable over time as models are exposed to similar data during training. Second, while dynamic benchmarks like LiveBench (White et al., 2025) offer one solution, they require substantial ongoing resources and may have limited scope. Third, existing robustness studies often examine specific perturbation methods applied to individual benchmarks, making it difficult to draw broader conclusions about model vulnerabilities or perturbation effectiveness across domains.

Our research aims to contribute to this evolving field by conducting a systematic comparison of perturbation techniques across multiple benchmarks, while carefully acknowledging the significant methodological challenges and limitations inherent in such work. We recognize that this investigation may raise as many questions as it answers,

and we approach this research with appropriate caution regarding the generalizability of our findings.

1.1 Research Questions

This investigation is guided by three research questions that reflect both our aspirations and the significant uncertainties in this domain:

RQ1: Comparative Analysis Across Domains. How do different perturbation techniques compare when applied consistently across various benchmark types? While previous studies like GSM-Symbolic (Mirzadeh et al., 2025) and MATH-Perturb (Huang et al., 2025) have demonstrated the potential of specific approaches, we seek to understand whether these findings generalize and how different methods compare within a controlled evaluation framework. We acknowledge that such comparisons face substantial methodological challenges and that our findings may be limited by benchmark selection and evaluation constraints.

RQ2: Vulnerability Pattern Investigation. Can we identify consistent patterns in how models respond to different types of perturbations, and what might these patterns reveal about underlying model behaviors? We plan to examine whether certain perturbation categories consistently expose similar weaknesses across different models and domains. However, we recognize that interpreting such patterns requires caution, as apparent vulnerabilities may reflect evaluation artifacts rather than fundamental model limitations.

RQ3: Practical Implementation Considerations. What are the practical challenges and limitations of implementing benchmark perturbation at scale? Rather than promising revolutionary advances, we aim to document the real-world complexities of maintaining semantic equivalence, ensuring quality control, and validating perturbations across different domains. Our goal is to contribute honest insights about both the potential and the limitations of these approaches.

2 Related Work

Benchmark Perturbation and Robustness Studies The field of benchmark perturbation has emerged from growing concerns about evaluation reliability. Early work like NL-Augmenter (Dhole et al., 2022) established systematic approaches to text transformation across various NLP tasks, providing over 100 transformation operations that serve as foundational tools for benchmark modification. This comprehensive framework demonstrated both the potential and complexity of systematic perturbation approaches.

Recent mathematical reasoning studies have revealed concerning fragility patterns. GSM1K (Zhang et al., 2024) provided initial evidence of potential overfitting by altering numerical values and contextual elements, while GSM-Infinite (Zhou et al., 2025) explored template-based generation approaches. The GSM-Symbolic framework represents a significant advancement in this area, generating diverse mathematical questions through symbolic templates and demonstrating that model performance declines significantly when numerical values or question complexity increases. However, questions remain about how these findings extend to other domains and whether the observed patterns reflect fundamental limitations or evaluation artifacts.

The MATH-Perturb framework (Huang et al., 2025) introduced important distinctions between simple perturbations (rephrasing, variable swapping) and more complex modifications (increased polynomial degrees, non-linear constraints). These categorizations provide useful conceptual frameworks, though their effectiveness across different mathematical domains and their applicability to non-mathematical reasoning tasks require further investigation.

Beyond mathematical reasoning, studies like MMLU-Pro (Wang et al., 2024) have attempted to enhance evaluation difficulty through answer option expansion and prompt reformatting, while MMMU-Pro (Yue et al., 2025) has explored multimodal complexity increases. These efforts highlight both the potential for benchmark enhancement and the substantial challenges in maintaining equivalence across different modalities and knowledge domains.

Contamination Detection and Mitigation Recent contamination studies have revealed varying effects across benchmarks, with some showing significant accuracy increases on contaminated datasets while others show minimal impact. This variability suggests that contamination effects may be more complex and domain-specific than initially understood. Detection methods continue to evolve, though each approach faces distinct limitations and may miss certain types of contamination.

Dynamic benchmarking approaches like LiveBench (White et al., 2025) and LiveCodeBench (Jain et al., 2024) offer promising

alternatives through continuous content updating, but these methods face scalability challenges and may introduce their own biases through selection processes. The trade-offs between manual curation quality and automated generation scalability remain an active area of investigation.

Connection to Our Proposed Approach Our work builds upon these existing foundations while acknowledging their limitations and the uncertainties in this field. Rather than claiming to solve fundamental problems, we aim to contribute systematic comparisons that may inform future evaluation practices. We plan to leverage existing tools and frameworks where possible while documenting the practical challenges and limitations we encounter.

3 Methodology

3.1 Benchmark Selection and Scope

We plan to focus our investigation on a carefully selected set of benchmarks, chosen based on practical considerations rather than comprehensive coverage. Our selection criteria emphasize widespread adoption in recent model evaluations, documented concerns about contamination vulnerability, availability of reliable ground truth, and feasibility for systematic perturbation within our resource constraints.

Initial Focus Areas: We intend to begin with mathematical reasoning benchmarks including GSM8K and selected problems from MATH, as these offer clearer correctness criteria and more straightforward perturbation validation. Pending successful development of our methodology, we hope to extend to scientific knowledge tasks through MMLU subsets and basic multimodal problems from MMMU, though we recognize that each domain expansion introduces significant additional complexity.

We acknowledge that this selective approach limits the generalizability of our findings and may introduce sampling biases. Our choice to prioritize depth over breadth reflects both resource constraints and our belief that careful analysis of fewer benchmarks may provide more reliable insights than superficial coverage of many.

3.2 Perturbation Approach and Categorization

Rather than developing entirely new perturbation methods, we plan to systematically apply and compare existing approaches across our selected benchmarks. This approach allows us to focus on comparative analysis while avoiding the substantial challenges of novel method development.

Surface-Level Modifications: We will implement straightforward modifications using established tools. These include variable renaming, unit conversions, and syntactic rephrasing using existing libraries like NL-Augmenter. For mathematical problems, we will explore numerical value changes and basic algebraic transformations using symbolic mathematics tools. While these modifications may seem simple, maintaining semantic equivalence while ensuring natural language quality presents non-trivial challenges that we expect to document carefully.

Content and Context Variations: Building on approaches from GSM-Symbolic and related work, we plan to modify narrative contexts and problem settings while preserving underlying logical structures. This involves entity substitution, scenario changes, and format variations. However, we recognize that validating semantic equivalence for these modifications, particularly across different knowledge domains, presents substantial difficulties that may limit our scope.

Complexity and Structure Changes: Where feasible, we will explore modifications that alter problem complexity or presentation structure while maintaining solvability. This represents our most ambitious and uncertain objective, as ensuring that such changes preserve the intended evaluation target while avoiding the introduction of confounding factors requires careful validation that may exceed our capabilities.

3.3 Quality Control and Validation Challenges

We anticipate that quality control will represent one of the most significant challenges in our investigation, and we plan to document these difficulties as thoroughly as our findings.

Automated Validation Efforts: We will attempt to implement automated checks for mathematical consistency using constraint solvers and symbolic manipulation tools, though we expect these automated approaches to have limited effectiveness for more complex or subjective perturbations. For factual content, we plan to leverage existing knowledge bases where possible, while acknowledging their incompleteness and potential inaccuracies.

Human Validation Requirements: Recognizing the limitations of automated approaches, we plan to incorporate human review processes, though we anticipate that scaling such review across different domains and perturbation types will present substantial challenges. We expect that resource constraints will require us to sample rather than comprehensively validate all perturbations, introducing potential bias in our quality assessment.

Error Propagation and Bias Introduction: We are particularly concerned about the risk of amplifying existing benchmark errors or introducing new biases through our perturbation process. We plan to document these risks and implement monitoring approaches, though we acknowledge that our ability to detect and correct all such issues may be limited.

3.4 Evaluation Metrics and Statistical Considerations

Our evaluation approach emphasizes careful measurement while acknowledging the limitations and potential artifacts in our metrics.

Performance Consistency Measures: We will examine model performance across original and perturbed problems, measuring both absolute performance changes and response consistency. However, we recognize that interpreting such measures requires caution, as performance differences may reflect evaluation artifacts, perturbation quality issues, or genuine reasoning limitations.

Statistical Analysis Approach: We plan to apply appropriate statistical tests to assess the significance of observed differences, including multiple comparison corrections where relevant. However, we acknowledge that statistical significance does not necessarily imply practical importance, and we will interpret our findings within the broader context of evaluation uncertainty.

Limitations and Confounding Factors: We expect our analysis to be influenced by numerous confounding factors, including perturbation quality variations, benchmark selection effects, and evaluation methodology limitations. We plan to document these limitations transparently rather than attempting to eliminate them entirely.

3.5 Implementation Timeline and Risk Assessment

We propose a 42-month timeline that prioritizes careful development over ambitious scope, with significant time allocated for addressing unexpected challenges and limitations.

Phase 1 - Foundation and Pilot (Months 1-18): We will begin with basic mathematical reasoning tasks, focusing on developing reliable perturbation and validation processes rather than achieving broad coverage. This extended timeline reflects our expectation that seemingly straightforward tasks may reveal substantial complexity. We plan to use this phase to assess the feasibility of our broader objectives and adjust our scope accordingly.

Phase 2 - Selective Expansion (Months 19-30): Based on our Phase 1 findings and remaining resources, we will selectively expand to additional domains or perturbation types. This expansion will be contingent on successful resolution of challenges encountered in Phase 1 and may involve significant scope modifications.

Phase 3 - Analysis and Documentation (Months 31-42): We will focus on comprehensive analysis of our findings while documenting the limitations and challenges we have encountered. This phase emphasizes honest assessment of both our results and the broader challenges in benchmark perturbation research.

Risk Mitigation Strategies: We plan for regular scope reassessment, with willingness to reduce ambitions if challenges exceed our capabilities. We will prioritize documentation of negative results and limitations as valuable contributions to the field. Our modular approach will allow us to produce useful outputs even if certain objectives prove infeasible.

4 Expected Contributions and Limitations

Rather than promising transformative advances, we aim to contribute useful empirical insights while documenting the substantial challenges in this research area.

Anticipated Positive Contributions: We hope to provide systematic comparisons of perturbation techniques across multiple benchmarks, offering insights that may inform future evaluation practices. Our documentation of practical implementation challenges and quality control requirements may help other researchers planning similar investigations. Additionally, our validation of existing

perturbation methods across different domains may reveal both their strengths and limitations.

Acknowledged Limitations and Risks: We recognize that our findings may have limited generalizability due to benchmark selection constraints and resource limitations. Our perturbation quality may vary significantly across domains, potentially undermining comparative analyses. The computational and human resources required for comprehensive validation may exceed our capabilities, forcing difficult trade-offs between scope and rigor. Furthermore, our results may be influenced by rapidly evolving model capabilities and evaluation practices, potentially limiting their long-term relevance.

Potential Negative Outcomes: We acknowledge the possibility that our investigation may reveal fundamental limitations in current perturbation approaches, or that the challenges of maintaining semantic equivalence across domains may prove more substantial than anticipated. Such findings, while disappointing, would still represent valuable contributions to the field by documenting these limitations and informing future research directions.

5 Conclusion

We propose a systematic investigation of benchmark perturbation techniques that prioritizes careful empirical analysis over ambitious claims. By comparing existing perturbation approaches across selected benchmarks while documenting the substantial challenges and limitations we encounter, we aim to contribute honest insights that may inform the ongoing development of robust evaluation practices. We approach this research with appropriate humility regarding both the complexity of the challenges and the limitations of our proposed solutions. Our goal is to advance understanding in this important area while maintaining realistic expectations about what such research can achieve given current methodological constraints and the evolving nature of language model capabilities.

References

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. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

Kaustubh D. Dhole, Varun Gangal, Sebastian Gehrmann, Aadesh Gupta, Zhenhao Li, Saad Mahamood, Abinaya Mahendiran, Simon Mille, Ashish Shrivastava, Samson Tan, Tongshuang Wu, Jascha Sohl-Dickstein, Jinho D. Choi, Eduard Hovy, Ondrej Dusek, Sebastian Ruder, and 1 others. 2022. Nl-augmenter: A framework for task-sensitive natural language augmentation. *Preprint*, arXiv:2112.02721.

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

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *Preprint*, arXiv:2103.03874.

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. 2025. MATH-Perturb: Benchmarking LLMs' math reasoning abilities against hard perturbations. *arXiv preprint arXiv:2502.06453*.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. Livecodebench: Holistic and contamination free evaluation of large language models for code. *ArXiv*, abs/2403.07974.

Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. 2024. Gsm-plus: A comprehensive benchmark for evaluating the robustness of llms as mathematical problem solvers. *Preprint*, arXiv:2402.19255.

Seyed-Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2025. 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.

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

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. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural*

Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024.

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. 2025. Livebench: A challenging, contamination-limited LLM benchmark. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

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, and 3 others. 2024. 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.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. 2025. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 15134–15186. Association for Computational Linguistics.

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. 2024. A careful examination of large language model performance on grade school arithmetic. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

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