Beyond Memorization: Reasoning-Driven Synthesis as a Mitigation Strategy Against Benchmark Contamination

Anonymous Author(s)
Affiliation
Address
email

Abstract

Capability evaluation of large language models (LLMs) is increasingly shadowed by rising concerns of data contamination that cast doubts on whether static benchmarks measure genuine reasoning or mere memorization. We present an empirical study using an infinitely scalable framework to synthesize research-level QA directly from arXiv papers, harnessing the natural temporal structure of research publications where performance decay after knowledge cutoffs may indicate potential contamination. We evaluated 4 frontier model represented by 2 models of different knowledge cutoff dates per family on 1,643 multi-step reasoning questions synthesized from 20,277 arXiv papers stratified over 26 months, covering at least 6 months before and after all cutoff dates. Our results consistently showed a lack of significant performance decay near knowledge cutoff dates for models of various sizes, developers, and release dates. We further performed a comparative analysis with previous longitudinal studies that reported significant post-cutoff performance decay using directly retrieved questions based on public data. we hypothesize that the multi-step reasoning required by our synthesis pipeline offered additional complexity that goes deeper than shallow memorization, which effectively serves a mitigation strategy against benchmark contamination. We fully open source our code and dataset to aid reproducibility and advocate for a paradigm shift that prioritize reasoning-driven synthesis to construct benchmarks over simply collecting newly released questions periodically.

1 Introduction

> *If you cannot measure it, you cannot improve it.*
> —**Lord Kelvin** Thomson [1889]

The reliability of large language model (LLM) evaluations faces a critical threat from data contamination, which could artificially inflate performance metrics while obscuring genuine reasoning capabilities Dong et al. [2024]. Many previous work Ding et al. [2024], Li et al. [2025], Xu et al. [2024] have demonstrated that contamination has reached concerning proportions across various evaluation benchmarks, where models achieve inflated scores through memorized patterns rather than authentic problem-solving abilities. This widespread contamination undermines the fundamental premise of capability evaluation by rewarding benchmarking memorization over genuine reasoning abilities.

The rapid saturation of established benchmarks further compounds this challenge, creating an evaluation crisis where traditional metrics no longer differentiate between reasoning advances and pattern matching. Legacy frameworks such as MATH Hendrycks et al. [2021] and GPQA Rein

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

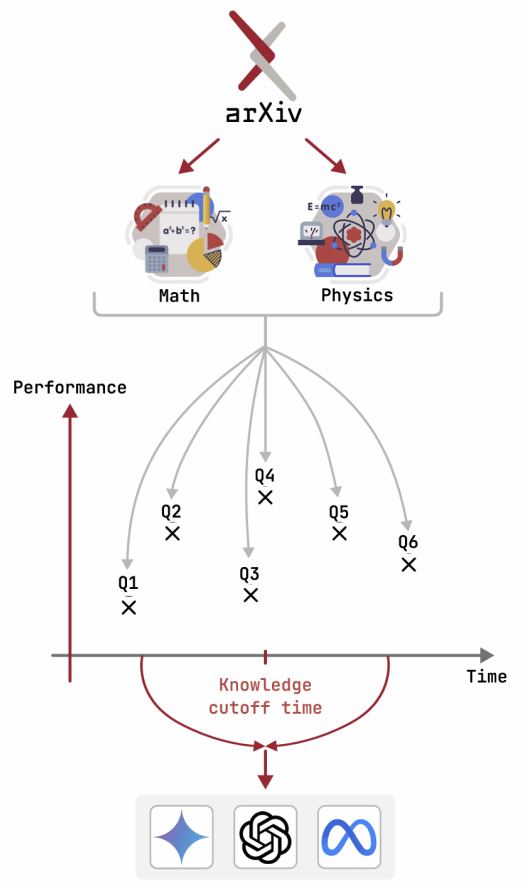


Figure 1: Overview of our framework for synthesizing research-level evaluation questions based on arXiv papers, which are temporally stratified to monthly window to assess frontier model reasoning capabilities before and after their knowledge cutoff dates. We hypothesize that potential contamination will lead to significant post-cutoff performance decay.

et al. [2023] have experienced dramatic performance saturation across model generations. This phenomenon reflects the broader inadequacy of static evaluation paradigms in measuring rapidly evolving model capabilities.

Existing mitigation strategies bear fundamental limitations in scalability and effectiveness. Manual curation approaches, exemplified by FrontierMath Glazer et al. [2024], demand extensive expert effort for problem construction and validation, creating practical barriers to frequent updates and systematic coverage. Periodic collection methods such as LiveBench White et al. [2025] promised to complete regular content updates but suffer from short shelf life and resource-intensive maintenance requirements on long time horizon. These approaches fail to address the core challenge of creating evaluation frameworks that scale automatically with advancing model capabilities while maintaining contamination resistance.

To address these challenges, we leverage the natural temporal structure of research publication to create a contamination-resistant evaluation framework that scales automatically with scientific progress. Academic publications follow strict chronological ordering, making it impossible for models to encounter evaluation content synthesized from papers published after their training cutoff dates. This temporal design enables systematic contamination detection through performance decay analysis, where post-cutoff performance decay would indicate more reliance on memorization rather than genuine reasoning capabilities.

Contributions

- We leverage a fully automated, infinitely scalable framework to synthesize 1,643 multi-step reasoning questions from 20,277 arXiv math and physics papers covering 26 months with comprehensive temporal coverage spanning at least 6 months before and after all model knowledge cutoffs.

- We comprehensively evaluated all 8 frontier models from 4 leading developers with fine-grained monthly temporal granularity across math and physics domain to improve robustness of our findings across various models and domains.

- We systematically compare longitudinal analysis based on retrieval vs. synthesis-based benchmarks to empirically show that reasoning-driven synthesis offers stronger contamination resistance compared to questions directly harvested from public sources.

Our approach addresses the fundamental trade-off between evaluation scalability and integrity that shadows current benchmarking paradigm. We further advocate for multi-step reasoning-driven transformation that creates increased cognitive distance against simple pattern matching, thereby resist contamination by design.

2 Related Work

Research-Level Reasoning Benchmarks GPQA Rein et al. [2023] present PhD-level challenges across biology, physics, and chemistry, but its static nature makes it vulnerable to memorization as training corpora expand. CURIE Anonymous [2025] advances multitask scientific evaluation across specialized domains requiring expert knowledge synthesis, yet lacks temporal stratification mechanisms for contamination detection. While HLE Jain et al. [2025] and FrontierMath Glazer et al. [2024] achieve research-level difficulty through expert curation, they suffer from scalability bottlenecks and lack systematic contamination detection that our automated synthesis pipeline addresses through temporal stratification.

Data Synthesis from Research Papers Scientific paper-based evaluation represents promising directions for scalable benchmark construction. LiveXiv Shabtay et al. [2025] pioneered temporal evaluation through chronologically ordered arXiv papers, establishing contamination detection principles, but focuses on factual extraction rather than complex reasoning chains. SPIQA Anonymous [2024] demonstrates comprehensive multimodal evaluation from scientific papers, yet emphasizes comprehension over multi-step mathematical reasoning that creates cognitive barriers against memorization. OpenScholar Ajith et al. [2024], DIGESTables Skarlinski et al. [2024], and SciQAG Wang et al. [2024] advance cross-domain synthesis with sophisticated metrics but lack systematic temporal stratification for contamination detection. PaperQA Lala et al. [2023] combines question answering with document retrieval effectively, yet introduces contamination vectors through retrieval-based approaches that our synthesis methodology avoids.

Longitudinal Analysis on Data Contamination Systematic contamination detection reveals widespread evaluation compromise, establishing temporal analysis as essential for benchmark integrity, yet existing methodologies focus on retrieval-based approaches vulnerable to memorization. Comprehensive surveys Ding et al. [2024], Li et al. [2025], Xu et al. [2024] document pervasive benchmark compromise while ContaminationCDD Fan et al. [2024] provides computational detection through output distribution analysis. Temporal boundary research Roberts et al. [2024] demonstrates effectiveness in Mathematical Olympiad contexts with performance decay on post-cutoff problems, yet relies on publicly available competition problems vulnerable to training data inclusion. Programming applications through LiveCodeBench Jain et al. [2024] and DyCodeEval Chen et al. [2025] demonstrate performance drops on post-training data but focus on code generation rather than mathematical reasoning synthesis. Our contribution uniquely combines monthly temporal resolution with reasoning-driven synthesis that generates multi-step problems requiring genuine mathematical understanding rather than memorized solution patterns.

Mathematics

arXiv:2406.19979v2 - On quantitative convergence for stochastic processes: Crossings, fluctuations and martingales

Main Field: math.PR (Probability)

Theorem: Given $\varepsilon, K > 0$ and $g : \mathbb{N} \to \mathbb{N}$ there exists some N dependent only on these parameters such that whenever $\{x_n\}$ is a monotone sequence in $[-K, K]$, there exists some $n \leq N$ such that $|x_i - x_j| < \varepsilon$ for all $n \leq i \leq j \leq n + g(n)$. Moreover, we can assign to N the following concrete value: $N := \tilde{g}^{(\lceil 2K/\varepsilon \rceil)}(0)$ for $\tilde{g}(n) := n + g(n)$, where $\tilde{g}^{(i)}$ denotes the ith iteration of $\tilde{g}$.

Q: Given parameters $\varepsilon > 0$, $K > 0$, and a function $g : \mathbb{N} \to \mathbb{N}$, what explicit bound N in terms of ε, K, and g guarantees that for any monotone sequence (x_n) in $[-K, K]$ there exists some $n \leq N$ such that $|x_i - x_j| < \varepsilon \quad$ for all $n \leq i \leq j \leq n + g(n)$?

Physics

arXiv:2407.02415v2 - The Symplectic Schur Process

Main Field: math-ph (Mathematical Physics)

Theorem: Consider the rescaling

$$i(\tau) = \left\lfloor \frac{9n}{8} + \sqrt{\frac{27n}{64}}\,\tau \right\rfloor, \qquad u(\alpha) = -n + \left\lfloor \left(\frac{n}{12}\right)^{\frac{1}{4}} \alpha \right\rfloor.$$

For any fixed $k \in \mathbb{Z}_{\geq 1}$ and $\tau_1, \ldots, \tau_k \in \mathbb{R}$ and $\alpha_1, \ldots, \alpha_k \in \mathbb{R}_+$, let $i_\ell = i(\tau_\ell)$, $u_\ell = u(\alpha_\ell)$, then

$$\lim_{n\to\infty} \det_{1\leq \ell,\ell' \leq k} \left[\left(\frac{n}{12}\right)^{\frac{1}{4}} \left(\delta_{i_\ell, i_{\ell'}} \delta_{u_\ell, u_{\ell'}} - K^{\mathrm{SSP}}(i_\ell, u_\ell; i_{\ell'}, u_{\ell'}) \right) \right] = \det_{1\leq \ell,\ell' \leq k} \left[\mathcal{K}(\tau_\ell, \alpha_\ell; \tau_{\ell'}, \alpha_{\ell'}) \right],$$

where $\mathcal{K}$ is the Pearcey-like kernel.

Q: Let $i(\tau) = \left\lfloor \frac{9n}{8} + \sqrt{\frac{27n}{64}}\,\tau \right\rfloor$ and $u(\alpha) = -n + \left\lfloor \left(\frac{n}{12}\right)^{1/4} \alpha \right\rfloor$. For a fixed $k \in \mathbb{Z}_{\geq 1}$ and real parameters $\tau_1, \ldots, \tau_k$ and $\alpha_1, \ldots, \alpha_k$, define $i_\ell = i(\tau_\ell)$ and $u_\ell = u(\alpha_\ell)$. What is the limit as $n \to \infty$ of

$$\det_{1\leq \ell,\ell' \leq k} \left[\left(\tfrac{n}{12}\right)^{1/4} \left(\delta_{i_\ell, i_{\ell'}}\, \delta_{u_\ell, u_{\ell'}} - K^{\mathrm{SSP}}(i_\ell, u_\ell; i_{\ell'}, u_{\ell'}) \right) \right]?$$

Table 1: Examples of theorem-to-question conversion with arXiv IDs, paper titles, and domain classification tags with explanations.

3 Methodology

3.1 Synthesis Pipeline

Retrieval and Theorem Identification This synthesis pipeline is adapted based on the RealMath Framework Zhang et al. [2025]. We retrieve arXiv papers directly from the arXiv API to obtain comprehensive metadata including LaTeX source code, DOI, publication date, and arXiv domain tags . Our synthesis pipeline utilizes OpenAI-GPT-4.1 and OpenAI-o4-mini OpenAI [2025] to systematically parse LaTeX source and identify constructive theorems with fixed answers, excluding theorems with multiple solutions or ambiguity to ensure automated verification while maintaining mathematical rigor.

QA Generation We employ sophisticated multi-stage generation focusing on compositional multi-step reasoning where solvers must synthesize multiple mathematical concepts, apply sequential logical transformations, and maintain coherent reasoning across extended derivation sequences. Each question undergoes automated analysis to verify solution paths requiring multiple interdependent reasoning steps including concept application, algebraic manipulation, logical inference, and result synthesis. We systematically filter problems requiring fewer than six distinct reasoning steps, ensuring substantial cognitive effort that cannot be solved through pattern recognition.

Filtering and Validation Our filtering process includes comprehensive post-processing review of each generated question-answer pair, systematically removing low-quality samples with easily guessable answers or obvious solutions from context. We extend methodology beyond mathematics to Physics, creating problems requiring interdisciplinary reasoning combining mathematical formalism with physical intuition, dimensional analysis, and connections between theoretical predictions and measurable quantities. This multi-step reasoning validation ensures our benchmark captures sophisticated abilities distinguishing advanced models from pattern-matching approaches.

3.2 Dataset Construction

Temporal Stratification We synthesized questions with monthly resolution from May 2023 to June 2025, ensuring coverage of at least 6 months before and after models' knowledge cutoffs. This 26-month window enables precise contamination detection across temporal boundaries while leveraging the natural chronological ordering of academic publications that prevents models from encountering evaluation content during training.

Cross-Domain Coverage Our dataset balances Mathematics and Physics domains, generating 20 questions per month per domain. We focus on combinatorics and number theory domains providing suitable sources for constructive theorems with fixed answers. This cross-domain approach evaluates both domain-specific knowledge and general reasoning transfer capabilities across scientific disciplines.

Quality Assurance We performed manual expert inspection for technical accuracy and LaTeX formatting as quality filters, obtaining 1,643 questions equally spanning math and physics domains. This curation process ensures evaluation integrity while maintaining the scalability advantages of automated synthesis over manually curated benchmarks.

4 Experimental

We evaluated models from four distinct families (OpenAI OpenAI [2025], Gemini Deepmind [2025], Llama Grattafiori et al. [2024], and DeepSeek Guo et al. [2025]) where each family is represented by 2 models with different knowledge cutoff dates to assess the consistency of our findings. We summarize the models and their respective cutoff dates in Table 2. We evaluated all models by calling OpenRouter API with default settings and enabled thinking modes whenever possible.

Table 2: All evaluated models and their knowledge cutoff dates HaoooWang [2025]

Model	Knowledge Cutoff
DeepSeek-R1-0528	2024.07
DeepSeek-R1	2023.10
OpenAI-o4-mini	2024.06
OpenAI-o3-mini	2023.10
Gemini-2.5-Flash	2025.01
Gemini-2.0-Flash	2024.08
Llama-4-Scout	2024.08
Llama-3.3-70B	2023.12

Figure 2 illustrates temporal performance across 26 months in both Mathematics and Physics domains. The monthly trends show no significant performance degradation around knowledge cutoff dates marked by dashed vertical lines. Rather than expected decay due to potential training data contamination, models demonstrate stable or improved performance on post-cutoff samples across both domains. This pattern suggests our multi-step reasoning synthesis pipeline creates contamination resistance by requiring genuine problem-solving rather than memorization-based pattern matching.

Figure 3 presents mean accuracy comparison across six months before and after knowledge cutoff dates for both mathematics and physics domains. Results demonstrate consistent performance patterns without significant degradation following cutoff boundaries, confirming temporal stability in synthesized evaluation content.

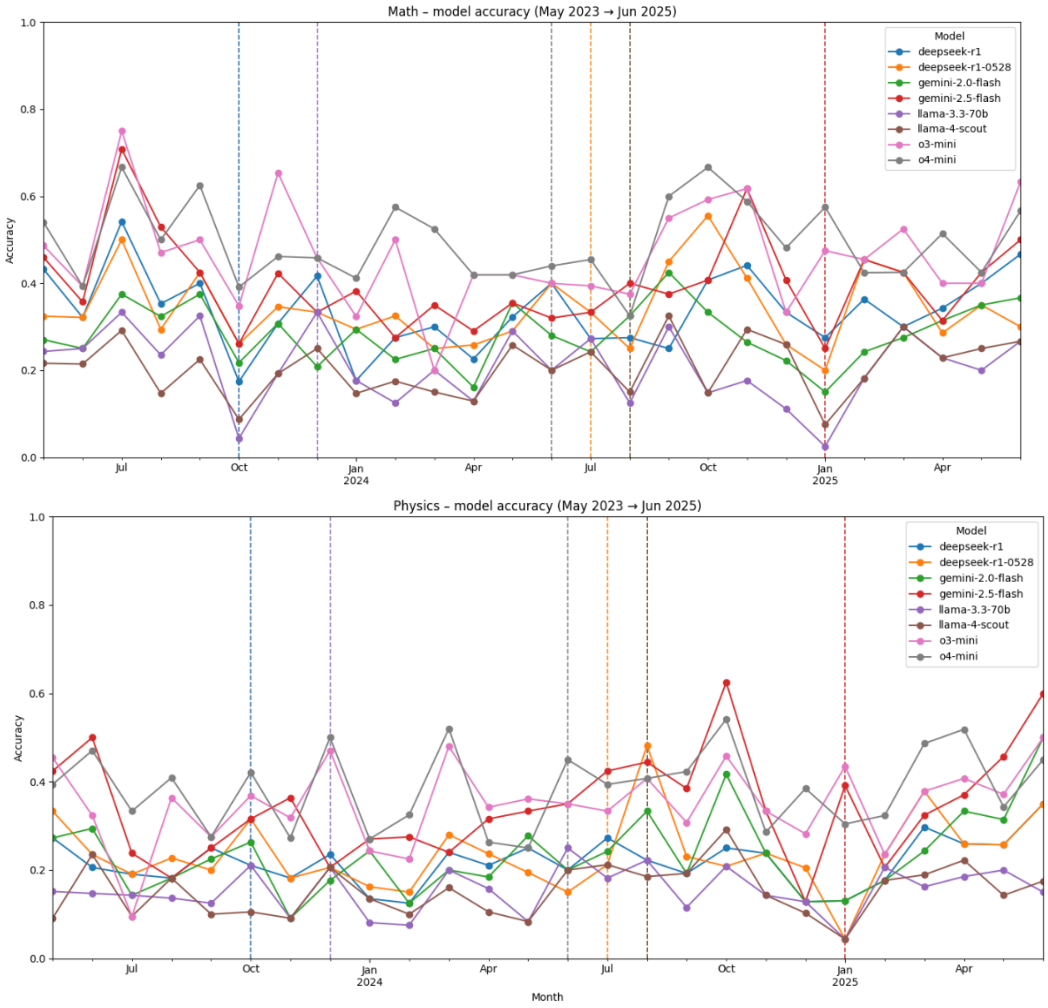


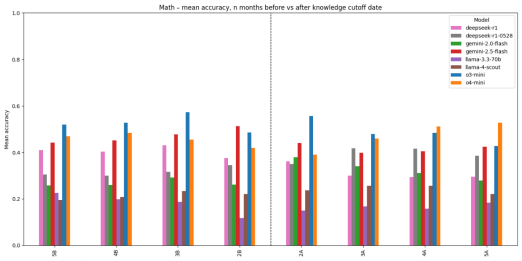
Figure 2: Model performance scores across 26 months in Math and Physics domains, with knowledge cutoff dates marked by dashed lines matching model legends. The figure demonstrates temporal performance stability without significant decay patterns around cutoff boundaries, suggesting contamination resistance through multi-step reasoning synthesis rather than performance degradation expected from contaminated benchmarks.

Figure 4 shows aggregated performance scores across temporal windows, further confirming stability of model capabilities on synthesized questions regardless of publication timing relative to training cutoffs. This consistency across all evaluated model families provides systematic evidence that reasoning-driven synthesis creates contamination resistance through cognitive complexity barriers.

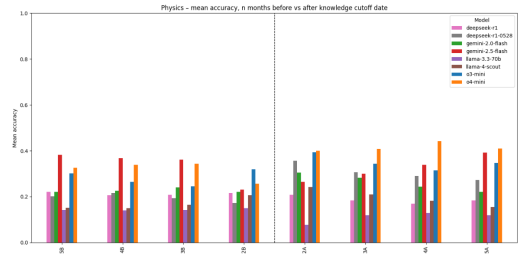
5 Discussion

We present a comparative analysis between retrieval-based and synthesis-based benchmarking approaches through their temporal contamination patterns, providing empirical evidence that reasoning-driven synthesis creates cognitive barriers against shallow memorization.

Contamination Vulnerability in Retrieval-Based Approaches Retrieval-based benchmarks consistently exhibit post-cutoff performance decay that reveals widespread contamination effects across multiple domains. Mathematical Olympiad evaluations demonstrate severe performance deterioration on problems published after training cutoffs, where models achieved artificially inflated scores on pre-cutoff content through memorized solution patterns from competition archives Roberts et al.



(a) Mean accuracy for mathematics-related QA pairs.



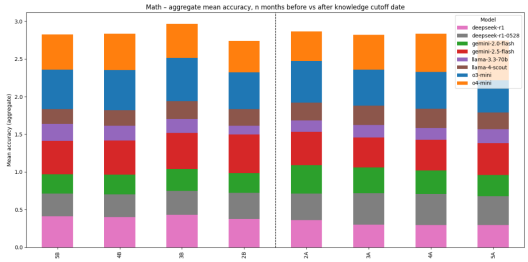
(b) Mean accuracy for physics-related QA pairs.

Figure 3: Mean accuracy for mathematics and physics QA pairs across six months before and after knowledge cutoff dates. Both domains show consistent performance without significant post-cutoff degradation, supporting our hypothesis that synthesis-based benchmarks resist contamination effects through cognitive complexity requirements that demand genuine reasoning rather than memorized pattern matching.

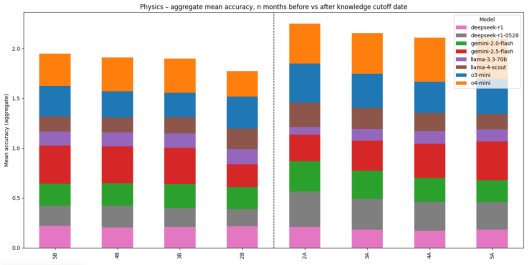
[2024]. Programming benchmarks reveal similarly concerning patterns, with coding problems showing sharp performance drops after knowledge boundaries while exhibiting suspicious correlation between problem popularity and model accuracy Huang et al. [2024]. LiveCodeBench documented systematic contamination through clear temporal performance boundaries in code generation tasks, demonstrating that models relied heavily on memorized solutions from competitive programming platforms and GitHub repositories rather than genuine algorithmic reasoning Jain et al. [2024]. These findings establish a consistent pattern where publicly sourced evaluation content becomes compromised through training data inclusion, creating evaluation frameworks that reward memorization over authentic problem-solving capabilities.

Performance Stability in Synthesis-Based Approaches In contrast, synthesis-based benchmarks performs transformation based on retrieved contents to vary the contexts and complexity of evaluation questions in question Zhu et al. [2024]. We hypothesize that the lack of post-cutoff performance decay may be attributed to the added transformation serving as an extra layer of cognitive distance that cognitively goes deeper than shallow memorization-based shortcuts. AntiLeakBench Wu et al. [2025] also highlights that data contamination before LLM cutoff dates can be effectively overcome by constructing samples with explicitly new knowledge after training cutoff instead of simply collect new data periodically. Our experimental results align well with Zhang et al. [2025] across diverse model families and extended temporal windows of 26 months, showcasing how our synthesis pipeline established contamination resistance through reasoning demands that cannot be satisfied through memorized associations alone.

Reasoning-Driven Synthesis as Mitigation Strategy We hypothesize the reason why there is a lack of significant post-cutoff performance decay in our results is because the transformation made in our synthesis pipeline goes deeper than shallow pattern-matching that enables memorization, creating effective contamination resistance through cognitive complexity requirements. Compositional reasoning tasks engage working memory integration, sequential logical processing, and abstract concept manipulation that operate through fundamentally different neural pathways than simple associative recall mechanisms established in cognitive science research Qiu et al. [2025]. The Extended Edit Distance metric provides quantitative evidence that complex reasoning problems



(a) Mathematics domain aggregated performance.



(b) Physics domain aggregated performance.

Figure 4: Aggregated performance scores for mathematics and physics domains comparing months before versus after cutoff dates, where nB indicates n months before cutoff and nA indicates n months after cutoff (e.g., 2B = 2 months before, 2A = 2 months after). Both domains maintain stable performance across temporal boundaries, demonstrating that our synthesis approach creates evaluation problems preserving discriminative power regardless of potential training data overlap.

requiring multiple transformation steps exhibit substantially lower similarity to potential training examples, creating natural barriers against memorization-based solutions. Performance stability under semantic paraphrasing correlates with genuine mathematical understanding rather than surface-level pattern matching, suggesting that our synthesis methodology creates robust evaluation frameworks resistant to contamination effects through cognitive complexity requirements Yang et al. [2023]. This theoretical foundation supports our empirical findings that reasoning-driven synthesis generates evaluation problems requiring authentic problem-solving capabilities that cannot be replicated through memorized training examples, establishing it as an effective mitigation strategy against benchmark contamination.

6 Conclusion

Our evaluation framework based on reasoning-driven synthesis scales synergistically with the ever-growing corpora scientific literature. Through a comprehensive analysis of 8 frontier models with different knowledge cutoff dates on 1643 questions across 26 months of temporally stratified questions based on 20277 arXiv papers, we empirically show that reasoning-driven synthesis produced temporally consistent performance stability across knowledge cutoff dates. We further compare our approach against various previous longitudinal analysis on retrieval-based benchmarks sourced from publicly accessible competition questions, which suffered from significant post-cutoff performance decay. The cognitive complexity inherent in multi-step reasoning synthesis naturally goes deeper than memorization shortcuts, suggesting that future benchmark construction should prioritize reasoning depth over simply collecting newly released Olympiad competition problems periodically, which suffers from increasingly shorter shelf life as the heated race for leading AI labs to develop frontier models continuously intensifies at an accelerating pace. We call on the community to further explore and improve such synthesis-based evaluation paradigms that leverage the ever-expanding corpora of research publications, harvesting the synergistic momentum from frontier scientific literature to more faithfully reflect frontier models' capability in helping with scientific research tasks reciprocally.

References

Akari Ajith, Nathan Peng, Ronald Weston, Neel Elhage, Deep Ganguli, Danny Hernandez, Tom Jones, Amanda Lovitt, Nova DasSarma, Peter Hase, et al. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *arXiv preprint arXiv:2411.14199*, 2024. URL https://arxiv.org/abs/2411.14199.

Anonymous. Spiqa: A dataset for multimodal question answering on scientific papers, 2024. URL https://neurips.cc/virtual/2024/events/datasets-benchmarks. 270K multimodal questions from scientific papers.

Anonymous. Curie: Evaluating llms on multitask scientific long context understanding and reasoning, 2025. URL https://arxiv.org/abs/2503.13517. To appear at ICLR 2025.

Simin Chen, Pranav Pusarla, and Baishakhi Ray. Dynamic benchmarking of reasoning capabilities in code large language models under data contamination, 2025. URL https://arxiv.org/abs/2503.04149.

Deepmind. Gemini flash system card, 2025. URL https://ai.google.dev/gemini-api/docs/models#gemini-2.5-flash.

Cheng Ding, Jingfeng Zhang, and Shervin Malmasi. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244*, 2024. URL https://arxiv.org/abs/2406.04244.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, and Ge Li. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. *ArXiv*, abs/2402.15938, 2024. URL https://arxiv.org/pdf/2402.15938.pdf.

Yihang Fan, Yujun Hou, Jie Zhang, Qin Liu, Xuming Hu, and Xiaolin Wang. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. *arXiv preprint arXiv:2402.15938*, 2024. URL https://arxiv.org/abs/2402.15938.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, et al. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. *arXiv preprint arXiv:2411.04872*, 2024. URL https://arxiv.org/abs/2411.04872.

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-Yeary, 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 Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne,

Onur Çelebi, 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 Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, 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 Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin 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, Changhan 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, Dingkang 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, Filippos 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, 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.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

HaoooWang. llm-knowledge-cutoff-dates: This repository contains a summary of knowledge cutoff dates for various large language models (llms), such as gpt, claude, gemini, llama, and more. https://github.com/HaoooWang/llm-knowledge-cutoff-dates, 2025. GitHub repository.

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. In Joaquin Vanschoren and Sai-Kit Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Yiming Huang, Zhenghao Lin, Xiao Liu, Yeyun Gong, Shuai Lu, Fangyu Lei, Yaobo Liang, Yelong Shen, Chen Lin, Nan Duan, and Weizhu Chen. Competition-level problems are effective LLM evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 13526–13544. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.803. URL https://doi.org/10.18653/v1/2024.findings-acl.803.

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. URL https://api.semanticscholar.org/CorpusId:268379413.

Sachin Jain, Rusheb Batra, Subbarao Kambhampati, Ellie Pavlick, William Merrill, Danqi Chen, Graham Neubig, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025. URL https://arxiv.org/abs/2501.14249.

Tanishq Lala, Odhran Pradhan, Adrian Akiki, Carolyn Wang, Noel Iskander, Amalie Trewartha, Alexander Dunn, et al. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*, 2023. URL https://arxiv.org/abs/2312.07559.

Yucheng Li, Tianshi Wang, Zhaowei Xu, Chenghua Zhang, Bo Wang, et al. A survey on data contamination for large language models. *arXiv preprint arXiv:2502.14425*, 2025. URL https://arxiv.org/abs/2502.14425.

OpenAI. Openai o3 and o4-mini system card, 2025. URL https://openai.com/index/introducing-o3-and-o4-mini/.

Shi Qiu, Shaoyang Guo, Zhuo-Yang Song, Yunbo Sun, Zeyu Cai, Jiashen Wei, Tianyu Luo, Yixuan Yin, Haoxu Zhang, Yi Hu, Chenyang Wang, Chencheng Tang, Haoling Chang, Qi Liu, Ziheng Zhou, Tianyu Zhang, Jingtian Zhang, Zhangyi Liu, Minghao Li, Yuku Zhang, Boxuan Jing, Xianqi Yin, Yutong Ren, Zizhuo Fu, Jiaming Ji, Weike Wang, Xudong Tian, Anqi Lv, Laifu Man, Jianxiang Li, Feiyu Tao, Qihua Sun, Zhou Liang, Yushu Mu, Zhongxuan Li, Jing-Jun Zhang, Shutao Zhang, Xiaotian Li, Xingqi Xia, Jiawei Lin, Zheyu Shen, Jiahang Chen, Qiuhao Xiong, Binran Wang, Fengyuan Wang, Ziyang Ni, Bohan Zhang, Fan Cui, Changkun Shao, Qing-Hong Cao, Ming xing Luo, Yaodong Yang, Muhan Zhang, and Hua Xing Zhu. Phybench:

Holistic evaluation of physical perception and reasoning in large language models, 2025. URL https://arxiv.org/abs/2504.16074.

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, 2023. URL https://arxiv.org/abs/2311.12022.

Manley Roberts, Himanshu Thakur, Christine Herlihy, Colin White, and Samuel Dooley. To the cutoff... and beyond? a longitudinal perspective on LLM data contamination. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=m2NVG4Htxs.

Nimrod Shabtay, Felipe Maia Polo, Sivan Doveh, Wei Lin, Muhammad Jehanzeb Mirza, Leshem Choshen, Mikhail Yurochkin, Yuekai Sun, Assaf Arbelle, Leonid Karlinsky, and Raja Giryes. Livexiv - A multi-modal live benchmark based on arxiv papers content. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=SulRfnEVK4.

Michael Skarlinski, Sam Chiappa, Shayne Longpre, Nelson Elhage, and Jacob Steinhardt. Digestables: Synthesizing scientific literature into tables using language models. *arXiv preprint arXiv:2410.22360*, 2024. URL https://arxiv.org/abs/2410.22360.

William Thomson. *Popular Lectures and Addresses, Vol. 1: Electrical Units of Measurement*. Macmillan and Co., London, 1889. URL https://archive.org/details/popularlecturesa01kelvuoft/page/73/mode/1up. Lecture delivered 3 May 1883.

Yuanhan Wang, Jianxi Li, Yifan Wang, Junyan Li, Dahua Lin, and Yu Qiao. Sciqag: A framework for auto-generated scientific question answering dataset with fine-grained evaluation. *arXiv preprint arXiv:2405.09939*, 2024. URL https://arxiv.org/abs/2405.09939.

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. URL https://openreview.net/forum?id=sKYHBTAxVa.

Xiaobao Wu, Liangming Pan, Yuxi Xie, Ruiwen Zhou, Shuai Zhao, Yubo Ma, Mingzhe Du, Rui Mao, Anh Tuan Luu, and William Yang Wang. Antileakbench: Preventing data contamination by automatically constructing benchmarks with updated real-world knowledge. 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), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 18403–18419. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.acl-long.901/.

Zhaowei Xu, Yapei Wang, Mengzhou Li, Jiaheng Chen, Qingxiu Dong, Hongyi Wang, Tao Gui, Qi Zhang, and Xuanjing Huang. Unveiling the spectrum of data contamination in language models: A survey from detection to remediation. *arXiv preprint arXiv:2406.14644*, 2024. URL https://arxiv.org/abs/2406.14644.

Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph Gonzalez, and Ion Stoica. Rethinking benchmark and contamination for language models with rephrased samples. *ArXiv*, abs/2311.04850, 2023. URL https://arxiv.org/pdf/2311.04850.pdf.

Jie Zhang, Cezara Petrui, Kristina Nikolić, and Florian Tramèr. Realmath: A continuous benchmark for evaluating language models on research-level mathematics, 2025. URL https://arxiv.org/abs/2505.12575.

Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. Dyval: Dynamic evaluation of large language models for reasoning tasks. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=gjfOL9z5Xr.

A Supplementary Materials and Instructions for Reproduction

The README file in the supplementary materials contains detailed instructions for how to reproduce the results, where we also provide the full synthetic dataset for evaluation as well as the codebase used for both QA synthesis and model evaluation, which ensures the reproducibility of this study. After peer review, we plan to open-source both the code and data in their entirety.

The dataset includes 1,643 QA pairs, with 856 QA pairs from mathematics papers and 787 from physics papers. These QA pairs were synthesized from 20,277 arXiv papers published between May 2023 and June 2025. Along with the datasets, we also provide the prompts that were used for question generation and evaluation.

To reproduce our results, the papers should be retrieved from the ArXiv API, and the LaTeX code should be extracted from these papers. Then, the provided prompts should be run using a large language model (LLM) to standardize the LaTeX, extract theorems, check their quality, and generate QA pairs. Another LLM is used to judge the quality of the generated QA pairs.

B Prompting Context

In this section, we detail all the prompts used for our QA generation and evaluation using LLM-as-a-Judge paradigm as adapted from RealMath Zhang et al. [2025].

System Prompt: You are an expert research scientist designing clear question-answer pairs that requires at least 6 steps of scientific reasoning from research papers. The questions should have a unique numerical or analytical answer. Some common examples: - "If and only if condition A holds, then we can get X.", then we can ask "what condition must hold for X to be true?". This is also a unique answer. - Existence and Uniqueness Theorems: e.g., "There exists a unique X that satisfies A.", then we can ask "what is the unique solution that satisfies A?". This is also a unique answer. - Exact Formula Calculations: e.g., "The answer of formula (1) is 10", then we can ask "what is the value of formula (1)?". This is also a unique answer. - Unique Maximum/Minimum Points: e.g., "The maximum value of function f is 10 at point x=1", then we can ask "what is the maximum value of function f?". This is also a unique answer. - Exact Complexity Results in Computational Complexity: e.g., "The time complexity of algorithm A is exactly $\Theta(n^2)$" (not $\Omega(n^2)$ or $O(n^2)$, because big-O and big-omega are not exact), then we can ask "what is the exact time complexity of algorithm A?". This is also a fixed answer.
If the theorem does not have a unique answer, you can skip this theorem and return empty result.
If the theorem is a good candidate, your questions should: - clearly state the context of this theorem, and clearly define all quantities to make the question statement clear and self-contained - requires at least 6 steps of scientific reasoning. - never reveal the answer in the question statement - never ask yes or no question, never ask questions that are easy to answer without any reasoning. - if the theorem says "There exists an X that satisfies A" but the numerical value of X is not unique, skip the theorem - if the conditions A under which we can get X are not unique (i.e. necessary and sufficient), skip the theorem - re-define in the question the quantities from the theorem statement (without revealing the answer) so that the question can be solved in a self-contained manner.
If the theorem is a good candidate, your answers should have: - a unique numerical or analytical answer, easy to verify without ambiguity; - if there's any approximation, the condition must be specified in the question body (e.g. to 2 decimal places)

Standardize LaTeX Prompt:You are an expert in LaTeX. Your task is to review contents from a scientific paper and ensure it can be directly rendered in standard LaTeX without requiring custom command definitions. We should only use usepackage: amsmath, amssymb, enumerate, amsfonts, mathrsfs, mathtools, logicproof. For any commonly used commands, you should not change them, e.g., mathbb, sum, prod, int, lim, frac, sin, cos, tan, ln, exp, log, etc. But if you find some words are similar to the custom command definitions but hard to parse, you can change them to the standard latex command, e.g., 'mathbb' should be changed to 'mathbb', because 'mathbb' is meaningless.
For any custom commands used in the content, please replace them with standard LaTeX notation.
Make sure to check if for each begin command, there is a corresponding end command and viceversa.
Moreover, make sure that $ is not missing and insert it when needed.
I will compile the latex content into a pdf.
IMPORTANT: You must not change the mathematical meaning of the content. Focus only on syntax corrections.
Return the standardized content in this exact JSON format:
"theorem": "the well-formatted theorem in latex format without any custom commands", "changes": "explanation of what changes were made to the theorem, don't change the theorem content"