Putnam-AXIOM: A Functional and Static Benchmark for Measuring Higher Level Mathematical Reasoning

Aryan Gulati
Department of Computer Science
Stanford University
aryangul@stanford.edu

Brando Miranda
Department of Computer Science
Stanford University
brando9@stanford.edu

Eric Chen
Department of Mathematics
Stanford University
ericc97@stanford.edu

Emily Xia
Department of Mathematics
Stanford University
emxia18@stanford.edu

Kai Fronsdal
Department of Computer Science
Stanford University
kaif@stanford.edu

Bruno de Moraes Dumont
Department of Mathematics
Stanford University
bdumont@stanford.edu

Sanmi Koyejo
Department of Computer Science
Stanford University
sanmi@stanford.edu

Abstract

As large language models (LLMs) continue to advance, many existing benchmarks designed to evaluate their reasoning capabilities are becoming less challenging. These benchmarks, though foundational, no longer offer the complexity necessary to evaluate the cutting edge of artificial reasoning. In this paper, we present the Putnam-AXIOM Original benchmark, a dataset of 236 challenging problems from the William Lowell Putnam Mathematical Competition, along with detailed step-by-step solutions. To address the potential data contamination of Putnam problems, we create functional variations for 53 problems in Putnam-AXIOM. We see that most models get a significantly lower accuracy on the variations than the original problems. Even so, our results reveal that Claude-3.5 Sonnet, the best-performing model, achieves 15.96% accuracy on the Putnam-AXIOM original but experiences more than a 50% reduction in accuracy on the variations dataset when compared to its performance on corresponding original problems. The data and evaluation code are available at `https://anonymous.4open.science/r/putnam-axiom-BD6E/`.

1 Introduction

The ability for Large Language Models (LLMs) to reason about complex problems has a plethora of applications in many fields such as economics [Zhang et al., 2024], drug discovery [Bran et al., 2023], and even simulations of human behavior and society [Park et al., 2023]. The prominence of this ability has led to significant development in the performance of LLMs on many reasoning benchmarks.

Outpacing Current Evaluations. Indeed, advanced models like GPT-4 [OpenAI, 2023] and Gemini Ultra [Team, 2023] have even surpassed human-level performance on many benchmarks like MMLU [Hendrycks et al., 2020] and MMMU [Yue et al., 2023]. Similarly, LLMs have seen astonishing

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Workshop on MATH-AI.

progress in other challenging benchmarks like GSM8K [Chen et al., 2022] and MATH [Hendrycks et al., 2021], with SOTA models attaining nearly 90% accuracy on MATH [Lei, 2024] and nearly perfect accuracy on GSM8K [Zhong et al., 2024]. Though this progress is a testament to the rapidly evolving ability and utility of LLMs, it also presents a large problem: Existing datasets are no longer sufficient to evaluate the reasoning abilities of LLMs.

Data Contamination. Compounding this issue is one of the most significant problems facing evaluation datasets today, i.e., data contamination. As LLMs are increasingly trained on more of the internet, an increasing number of the open-source problems used in evaluation benchmarks are incorporated in the training data of these models. A model can therefore display artificially high "reasoning ability" by simply memorizing the answers it has seen undermining evaluation integrity.

To address these limitations, we introduce the Putnam-AXIOM (**A**dvanced e**X**amination of **I**ntelligence in **O**perational **M**athematics) dataset, a novel and challenging compilation of high-level mathematics problems sourced from the prestigious William Lowell Putnam Mathematical Competition, an annual mathematics competition for undergraduate college students in North America which requires advanced mathematical reasoning and covers a wide range of university-level mathematical concepts. Further, we also introduce functional variations of this AXIOM dataset to combat data contamination taking inspiration from the solution employed by Srivastava et al. [2024]. These are small variations of questions on the Putnam that are equally difficult as the Putnam but unavailable anywhere on the internet. AXIOM enables fully automated evaluations by requiring models to provide final answers within "`\boxed{}`" brackets which can then be extracted and compared to the ground truth final solution using an equivalence function[1]. This approach eliminates the need for human evaluation, allows for complex open-ended answers, and avoids the limitations of multiple-choice formats, thus maintaining rigor while enabling scalability.

Initial evaluations on Putnam-AXIOM demonstrate its exceptional difficulty. Claude-3.5 Sonnet scores 15.96%, while GPT-4 achieves only 7.98%. Even math-specialized models like Qwen2-Math-7B and Qwen2-Math-7B-Instruct perform poorly, scoring 5.51% and 11.86% respectively. Performance further declines on functional variations of Putnam-AXIOM, which include significant drops for most models almost halving in many cases. These low scores highlight AXIOM's capacity to measure future improvements in LLM reasoning abilities and underscore the role of memorization in model performance.

2 Methods

2.1 Putnam-AXIOM Original Dataset

Dataset. The Putnam-AXIOM Original Dataset contains 236 problems curated from the William Lowell Putnam Mathematical Competition posed between 1985 and 2023. These problems were selected based on their ability to yield final "`\boxed{}`" solutions ensuring compatibility with our automated evaluation. The dataset encompasses various subjects within university-level mathematics categorized into 11 distinct domains - Geometry, Algebra, Trigonometry, Calculus, Linear algebra, Combinatorics, Probability, Number theory, Complex numbers, Differential equations and Analysis.

To maintain a consistent and rigorous evaluation, each problem retains its original exam ID, which indicates its difficulty level (A or B for sitting, 1-6 for increasing complexity). This categorization helps in evaluating subject-specific understanding and overall problem-solving skills at different levels of complexity. The dataset is formatted using LaTeX to accurately capture the complex equations and symbols the problems employ. Additionally, we utilize Asymptote vector graphics for encoding mathematical figures and diagrams to ensure language models can process visual elements directly. Further, we standardized the placement of boxed answers by relocating them to the end of each solution string to minimize unintended emergent behaviors leading to evaluations that are less "harsh" or prone to penalizing the model for formatting deviations rather than actual comprehension.

Model Assessment. Drawing inspiration from the MATH dataset by [Hendrycks et al., 2021], which demonstrated the effectiveness of using boxed answers for evaluating mathematical understanding in LLMs, we similarly create a dataset with final solutions being wrapped in `\boxed{}` commands. Boxed answers allow for an exact match criterion rather than relying on approximate heuristics by

[1]For instance, the equivalence function would evaluate the answers `0.5`, `1/2`, and `\frac{1}{2}` as equal

simply parsing the LLM generated string solution for the value within the box, thereby enhancing reliability and consistency of the evaluation process while being quick and cost-effective. To further ensure fair evaluation, we implemented an equivalence function that homogenizes similar answers, addressing both simple string inconsistencies and complex mathematical equivalences like $(x+1)^2$ and x^2+2x+1 or numerical expressions such as `\frac{1}{2}`, `1/2`, and `0.5` and equating them.

Modified Boxing. Given the complex nature of certain Putnam questions, some problems do not lend themselves to simple, singular boxed answers. Instead, they often include conditions, multiple possible answers, varied answer formats and elaborate proofs. These original questions would have necessitated costly and difficult human evaluations which we seek to avoid. To address this, we modified these questions by adding a trivial next step to the original questions, changing the solution accordingly. This additional step was designed so as to ensure that solvers reached the same conclusions and insights necessary to solve the problem, but then needed to perform a simpler computation to get a simplified, boxable answer. We provide an example of such a change in Figure 3. By incorporating this minor modification, we preserved the inherent difficulty and complexity of the original problems while making the answers suitable for our boxed answer evaluation criteria.

2.2 Putnam-AXIOM Variation Dataset

Models trained on snapshots of the internet have likely encountered Putnam questions, potentially inflating their performance on the Putnam-AXIOM Original dataset. Therefore, drawing inspiration from Srivastava et al. [2024], we introduce functional variations of select problems from Putnam-AXIOM Original providing an effective way of evaluating models that have been trained on the entire internet by taking advantage of weaknesses in model memorization. These variations are classified into three types.

Variable Change. The simplest variation is a variable change, where variable names are changed and the final answer is unvaried. We provide an example of a variable change in Figure 4. Variable changes slightly alter the problem from its original statement, which models could have trained on.

Constant Change. Constant changes encompass variable changes but also modify numeric properties of the question often altering constants within the solution as well as the final answer. We provide an example of a constant change in Figure 4. Constant changes significantly alter the problem from its original statement, challenging models relying on memorization, needing them to perform complex reasoning on how constant changes affect the solution and solving for the correct final answer.

Significant Change. Significant changes involve major functional or structural alterations to the question's content. Unlike variable changes in which the mathematical logic stays consistent, or constant changes where the difference is subject to purely numeric properties, significant changes require models to employ mathematical logic and reasoning beyond what was in the original question and solution. Despite changing the structure of the problem nontrivially, significant changes preserve the overall difficulty and solution logic. Figure 5 shows an example of such a significant change.

Variational Dataset Description. Not all Putnam questions are easily functionalized. Some constants are problem-specific, solutions may not generalize, and certain questions lack constants or boxable answers. In total, we have functional variations of 53 different Putnam-AXIOM questions, along with corresponding solutions and boxed answers. Of these questions, there was 1 significant change, 26 constant and variable changes, and 26 variable changes. Each variation is capable of generating an infinite number of unique, equal-difficulty questions providing a long-term solution to evaluating models that are trained on the benchmark dataset. To evaluate various SOTA models, we generated five snapshots per variation, giving us a total of 265 variations.

2.3 Model Evaluations

Using the LM Harness Evaluation framework [Gao et al., 2024], we evaluated several open and closed-source SOTA LLMs. Models were prompted to provide answers in `\boxed` format, which were then compared to Putnam ground truths using the method described in Section 2.1. Instruct models were tested with both standard and model-specific prompts. We evaluated the 236-question Putnam-AXIOM Original dataset once. For the variation dataset, we conducted five trials, each using a randomly selected variation snapshot and its corresponding 53 original questions. We then calculated mean accuracy and 95% confidence intervals.

3 Results and Analysis

Table 2 presents Putnam-AXIOM Original dataset accuracies. Most models score below 10%, with even NuminaMath, the AI Mathematics Olympiad winner [Investments, 2024], achieving only 11.787%. This underscores AXIOM's difficulty, which current SOTA models struggle to overcome despite potential data contamination. Figure 1 contrasts Putnam-AXIOM Variation dataset mean accuracies with the 56 corresponding original questions. Original accuracies typically surpass variation accuracies. For models like Claude-3.5 Sonnet , GPT-4, and NuminaMath-7B-TIR, non-overlapping confidence intervals reveal statistically significant differences, indicating artificially inflated performance on original questions due to data contamination.

Model	**Score**	**Acc.** (%)	**Model**	**Score**	**Acc.** (%)
Claude-3.5 Sonnet	38 / 236	15.96	Gemma-7B-it	8 / 236	3.38
GPT-4	22 / 236	9.322	Gemma-2B-it	2 / 236	0.85
Llama-3-8B	9 / 236	3.81	DeepSeek-Math-7B	14 / 236	5.93
Llama-3-8b Instruct	10 / 236	4.23	DeepSeek-Math-RL	19 / 236	8.05
Mistral-7B-v0.3	7 / 236	2.97	DeepSeek-Math-Instruct	12 / 236	5.08
Mistral-7B-Instruct-v0.3	8 / 236	3.38	NuminaMath-7B	11 / 236	4.66
Gemma-7B	9 / 236	3.81	Qwen2-Math-7B	13 / 236	5.51
Gemma-2B	7 / 236	2.97	Qwen2-Math-7B-Instruct	18 / 236	11.86

Table 1: Putnam-AXIOM Original results

Model	**Original Score**	**Original Acc.** (%)	**Variation Score**	**Variation Acc.** (%)
Claude-3.5 Sonnet	14 / 53	26.4	7 / 53	13.2
GPT-4	7 / 53	13.2	5 / 53	9.43
Llama-3-8B	2 /53	3.77	1 / 53	1.88
Llama-3-8b Instruct	4 / 53	7.92	1.6 / 53	3.01
Mistral-7B-v0.3	3.5 / 53	6.78	1.8 / 53	3.39
Mistral-7B-Instruct-v0.3	1.2 / 53	2.26	1.8 / 53	3.39
Gemma-7B	1.6 / 53	3.01	1.2 / 53	2.26
Gemma-2B	1.4 / 53	2.63	1.2 / 53	2.26
Gemma-7B-it	1.8 / 53	3.39	1.4 / 53	2.64
Gemma-2B-it	1.8 / 53	3.39	1 / 53	1.88
DeepSeek-Math-7B	3.2 / 53	6.03	2.2 / 53	4.15
DeepSeek-Math-RL	5.6 / 53	10.56	4.6 / 53	8.67
DeepSeek-Math-Instruct	4.2 / 53	7.92	2 / 53	3.77
NuminaMath-7B	5.6 / 53	10.55	2.4 / 53	4.53
Qwen2-Math-7B	5.2 / 53	9.81	2.8	5.28
Qwen2-Math-7B-Instruct	5.4 / 53	10.19	3.4 / 53	6.41

Table 2: Putnam-AXIOM Variation vs Corresponding Original: 5-run average

4 Conclusion

Putnam-AXIOM introduces 236 challenging Putnam problems as a benchmark for LLM reasoning. Our dataset, with complex mathematical questions and variations, reveals significant struggles even for top models. This exposes memorization's limitations and the need for genuine mathematical reasoning. Putnam-AXIOM aims to drive progress in AI reasoning as models advance.

References

Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. Llm as a mastermind: A survey of strategic reasoning with large language models, 2024. URL https://arxiv.org/abs/2404.01230.

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools, 2023. URL https://arxiv.org/abs/2304.05376.

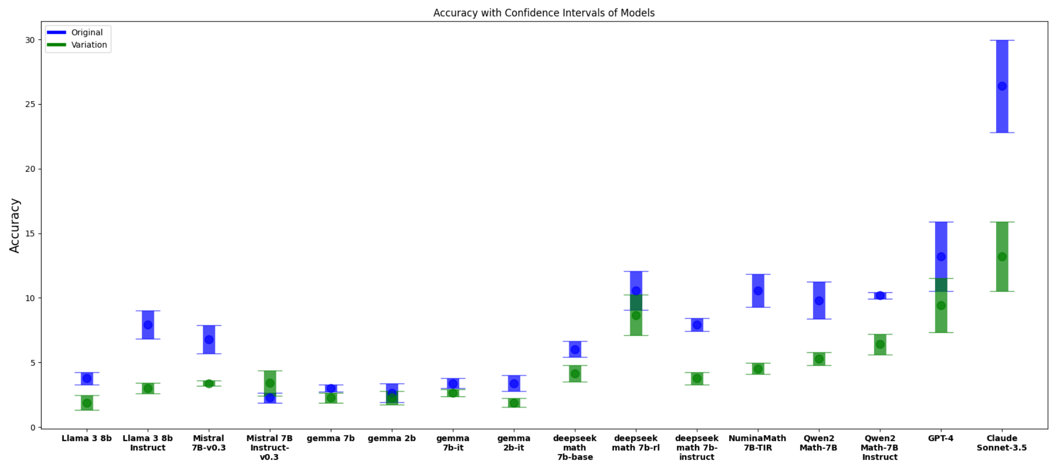

Figure 1: Mean accuracies of LLMs on random Putnam-AXIOM Variation snapshot and corresponding Original questions, with 95% confidence intervals.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior, 2023. URL `https://arxiv.org/abs/2304.03442`.

OpenAI. Gpt-4 technical report. *Preprint*, 2023.

Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv: 2312.11805*, 2023.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *International Conference on Learning Representations*, 2020.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, 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. *arXiv preprint arXiv: 2311.16502*, 2023.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. *arXiv preprint arXiv: 2210.03849*, 2022.

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 J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran, 2021. URL `https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper-round2.pdf`.

Bin Lei. Macm: Utilizing a multi-agent system for condition mining in solving complex mathematical problems, 2024. URL `https://arxiv.org/abs/2404.04735`.

Qihuang Zhong, Kang Wang, Ziyang Xu, Juhua Liu, Liang Ding, Bo Du, and Dacheng Tao. Achieving >97 URL `https://arxiv.org/abs/2404.14963`.

Saurabh Srivastava, Annarose M B, Anto P V, Shashank Menon, Ajay Sukumar, Adwaith Samod T, Alan Philipose, Stevin Prince, and Sooraj Thomas. Functional benchmarks for robust evaluation of reasoning performance, and the reasoning gap. *arXiv preprint arXiv: 2402.19450*, 2024.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 07 2024. URL https://zenodo.org/records/12608602.

XTX Investments. Ai mathematical olympiad - progress prize 1, 2024. URL https://kaggle.com/competitions/ai-mathematical-olympiad-prize.

United states copyright act. *U.S. Code Title 17*, 1976. Available at https://www.copyright.gov/title17/.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv: 2110.14168*, 2021.

Tomohiro Sawada, Daniel Paleka, Alexander Havrilla, Pranav Tadepalli, Paula Vidas, Alexander Kranias, John J. Nay, Kshitij Gupta, and Aran Komatsuzaki. Arb: Advanced reasoning benchmark for large language models, 2023. URL https://arxiv.org/abs/2307.13692.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiad-bench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv: 2402.14008*, 2024.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv: 2307.10635*, 2023.

Rylan Schaeffer. Pretraining on the test set is all you need, 2023. URL https://arxiv.org/abs/2309.08632.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark, 2023. URL https://arxiv.org/abs/2310.18018.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models, 2023. URL https://arxiv.org/abs/2304.06364.

Inbal Magar and Roy Schwartz. Data contamination: From memorization to exploitation. *Annual Meeting of the Association for Computational Linguistics*, 2022. doi: 10.48550/arXiv.2203.08242.

Leonardo Ranaldi, Elena Sofia Ruzzetti, and Fabio Massimo Zanzotto. PreCog: Exploring the relation between memorization and performance in pre-trained language models. In Ruslan Mitkov and Galia Angelova, editors, *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 961–967, Varna, Bulgaria, September 2023. INCOMA Ltd., Shoumen, Bulgaria. URL https://aclanthology.org/2023.ranlp-1.103.

A Appendix / supplemental material

A.1 Legal Compliance

We collect and modify various problems from the William Lowell Putnam Competition to create the original and variation datasets of Putnam-AXIOM. Putnam problems are created by the Mathematical Association of America (MAA), which is also the source of the AMC and AIME problems used in the MATH dataset [Hendrycks et al., 2021]. Like Hendrycks et al. [2021], we do not in any form seek to monetize or commercialize Putnam problems—only to utilize them for academic purposes.

Our use of the Putnam problems to create an evaluation dataset completely falls under the "research" section of Fair Use. Indeed, according to Section 107, of the U.S. Copyright Act [USC, 1976], our work certainly qualifies as Fair Use for the following reasons:

1. Our use of MAA problems is *only* for academic research purposes. We do not monetize or commercialize the problems.
2. Our use of Putnam problems as a reasoning evaluation benchmark for large language models is significantly different from their original use as competition problems.
3. Our use of Putnam problems is transformative. As detailed in Section 2 above, we have transformed the questions to be answered with a single numerical or algebraic "boxed answer" We have altered all of the solutions so that the final boxed answer lies at the end of the solution (so as to encourage models to explain their rationale before outputting a solution). We have also standardized the solutions: If there are many solutions given, we only use the first; if there are any references irrelevant to mathematics necessary to understand and solve the problem (such as comments like "Communicated by ..."), we have removed those.
4. Our use of Putnam problems to construct a benchmark has no effect on the demand for or supply of Putnam problems in the William Lowell Putnam Competition. The existence of our dataset does not alter the value of the original problems—as those are already freely available online—nor does it influence the market of future competitors/problem writers.

Problem: Let F_m be the mth Fibonacci number, defined by $F_1 = F_2 = 1$ and $F_m = F_{m-1} + F_{m-2}$ for all $m \geq 3$. Let $p(x)$ be the polynomial of degree 1008 such that $p(2n+1) = F_{2n+1}$ for $n = 0, 1, 2, \ldots, 1008$. Find integers j and k such that $p(2019) = F_j - F_k$ and give the answer in the form j/k.

Solution: More generally, let $p(x)$ be the polynomial of degree N such that $p(2n+1) = F_{2n+1}$ for $0 \leq n \leq N$. We will show that $p(2N+3) = F_{2N+3} - F_{N+2}$.
Define a sequence of polynomials $p_0(x), \ldots, p_N(x)$ by $p_0(x) = p(x)$ and $p_k(x) = p_{k-1}(x) - p_{k-1}(x+2)$ for $k \geq 1$. Then by induction on k, it is the case that $p_k(2n+1) = F_{2n+1+k}$ for $0 \leq n \leq N-k$, and also that p_k has degree (at most) $N-k$ for $k \geq 1$. Thus $p_N(x) = F_{N+1}$ since $p_N(1) = F_{N+1}$ and p_N is constant.
We now claim that for $0 \leq k \leq N$, $p_{N-k}(2k+3) = \sum_{j=0}^{k} F_{N+1+j}$. We prove this again by induction on k: for the induction step, we have

$$\begin{aligned} p_{N-k}(2k+3) &= p_{N-k}(2k+1) + p_{N-k+1}(2k+1) \\ &= F_{N+1+k} + \sum_{j=0}^{k-1} F_{N+1+j}. \end{aligned}$$

Thus we have

$$p(2N+3) = p_0(2N+3) = \sum_{j=0}^{N} F_{N+1+j}.$$

Now one final induction shows that $\sum_{j=1}^{m} F_j = F_{m+2} - 1$, and so $p(2N+3) = F_{2N+3} - F_{N+2}$, as claimed. In the case $N = 1008$, we thus have $p(2019) = F_{2019} - F_{1010}$. We thus prove that $(j,k) = (2019, 1010)$ is a valid solution with the final answer thus being $\boxed{2019/1010}$.

Year: 2017 **ID:** A6 **Final Answer:** 2019/1010

Figure 2: An example problem in Putnam-AXIOM. Solving this problem requires non-trivial constructions and multiple advanced reasoning chains. The format of the final answer is specified in the problem statement to make comparison simpler.

Problem: Determine which positive integers n have the following property: For all integers m that are relatively prime to n, there exists a permutation $\pi\colon \{1,2,\dots,n\} \to \{1,2,\dots,n\}$ such that $\pi(\pi(k)) \equiv mk \pmod{n}$ for all $k \in \{1,2,\dots,n\}$.

Solution: The desired property holds if and only if $\boxed{n = 1 \text{ or } n \equiv 2 \pmod{4}}$. Let $\sigma_{n,m}$ be the permutation of $\mathbb{Z}/n\mathbb{Z}$ induced by multiplication by m; the original problem asks for which n does $\sigma_{n,m}$ always have a square root.

...

By Lemma 1, $\sigma_{n,m}$ does not have a square root.

Year: 2016 **ID:** A1 **Final Answer:** ??

Problem: Determine the sum of the first k positive integers n (in terms of k) which have the following property: For all integers m that are relatively prime to n, there exists a permutation $\pi\colon \{1,2,\dots,n\} \to \{1,2,\dots,n\}$ such that $\pi(\pi(k)) \equiv mk \pmod{n}$ for all $k \in \{1,2,\dots,n\}$.

Solution: Let $\sigma_{n,m}$ be the permutation of $\mathbb{Z}/n\mathbb{Z}$ induced by multiplication by m; the original problem asks for which n does $\sigma_{n,m}$ always have a square root.

...

The desired property holds if and only if $n = 1$ or $n \equiv 2 \pmod{4}$, hence making the required sum $\boxed{2k^2 - 4k + 3}$.

Year: 2016 **ID:** A1 **Final Answer:** $2k^2 - 4k + 3$

Figure 3: A modified boxing example in Putnam-MATH. Here we see that the original problem holds true for a number of values of n conditioned on a specific property making it hard to find a boxable expression. We thus modify the solution to still require the solver to get to that conclusion and add a further computation of summing up the first k such values of n giving a boxable solution while keeping the core of the problem the same.

Problem: Define a *growing spiral* in the plane to be a sequence of points with integer coordinates $P_0 = (0,0), P_1, \dots, P_n$ such that $n \geq 2$ and:

$\dots$

How many of the points (x, y) with integer coordinates $0 \leq x \leq 2011, 0 \leq y \leq 2011$ *cannot* be the last point, P_n of any growing spiral?

Solution: We claim that the set of points with $0 \leq x \leq 2011$ and $0 \leq y \leq 2011$ that cannot be the last point of a growing spiral are as follows: $(0, y)$ for $0 \leq y \leq 2011$; $(x, 0)$ and $(x, 1)$ for $1 \leq x \leq 2011$; $(x, 2)$ for $2 \leq x \leq 2011$; and $(x, 3)$ for $3 \leq x \leq 2011$.

$\dots$

This gives a total of

$$2012 + 2011 + 2011$$
$$+2010 + 2009 = \boxed{10053}$$

excluded points.

Year: 2011 **ID:** A1 **Final Answer:** 10053

Problem: Define a *growing spiral* in the plane to be a sequence of points with integer coordinates $L_0 = (0,0), L_1, \dots, L_n$ such that $n \geq 2$ and:

$\dots$

How many of the points (w, v) with integer coordinates $0 \leq w \leq 4680, 0 \leq v \leq 4680$ *cannot* be the last point, L_n of any growing spiral?

Solution: We claim that the set of points with $0 \leq w \leq 4680$ and $0 \leq v \leq 4680$ that cannot be the last point of a growing spiral are as follows: $(0, v)$ for $0 \leq v \leq 4680$; $(w, 0)$ and $(w, 1)$ for $1 \leq w \leq 4680$; $(w, 2)$ for $2 \leq w \leq 4680$; and $(w, 3)$ for $3 \leq w \leq 4680$.

$\dots$

This gives a total of

$$4681 + 4680 + 4680$$
$$+4679 + 4678 = \boxed{23398}$$

excluded points.

Year: 2011 **ID:** A1 **Final Answer:** 23398

Figure 4: A constant change and variable change in Putnam-AXIOM. Here, we perform a variable change on the original problem/solution on the left by changing variables 'x' to 'w,' 'y' to 'v,' and 'P' to 'L.' We also perform a constant change by altering the constant '2011' to '4680'. The constant change affects the final answer, changing it from 10053 to 23398.

Problem: Determine the greatest possible value of $\sum_{i=1}^{10} \cos(3x_i)$ for real numbers $x_1, x_2, \dots, x_{10}$ satisfying $\sum_{i=1}^{10} \cos(x_i) = 0$.

Solution: Since $\cos(3x_i) = 4\cos(x_i)^3 - 3\cos(x_i)$, it is equivalent to maximize $4\sum_{i=1}^{10} y_i^3$ for $y_1, \dots, y_{10} \in [-1, 1]$ with $\sum_{i=1}^{10} y_i = 0$; note that this domain is compact, so the maximum value is guaranteed to exist.

$\dots$

The maximum value is $\boxed{480/49}$.

Year: 2018 **ID:** A3 **Final Answer:** $480/49$

Problem: Determine the least possible value of $\sum_{i=1}^{10} \sin(3c_i)$ for real numbers $c_1, c_2, \dots, c_{10}$ satisfying $\sum_{i=1}^{10} \sin(c_i) = 0$.

Solution: Since $\sin(3c_i) = 3\sin(c_i) - 4\sin(c_i)^3$, it is equivalent to minimize $4\sum_{i=1}^{10} y_i^3$ for $y_1, \dots, y_{10} \in [-1, 1]$ with $\sum_{i=1}^{10} y_i = 0$; note that this domain is compact, so the minimum value is guaranteed to exist.

$\dots$

The minimum value is $\boxed{-480/49}$.

Year: 2018 **ID:** A3 **Final Answer:** $-480/49$

Figure 5: A significant change to a question in Putnam-MATH. Here, we change the variable 'x' to 'c.' Notably, we also change $\cos$ to $\sin$, and "greatest" to "least." This constitutes a significant change to the structure of the problem.

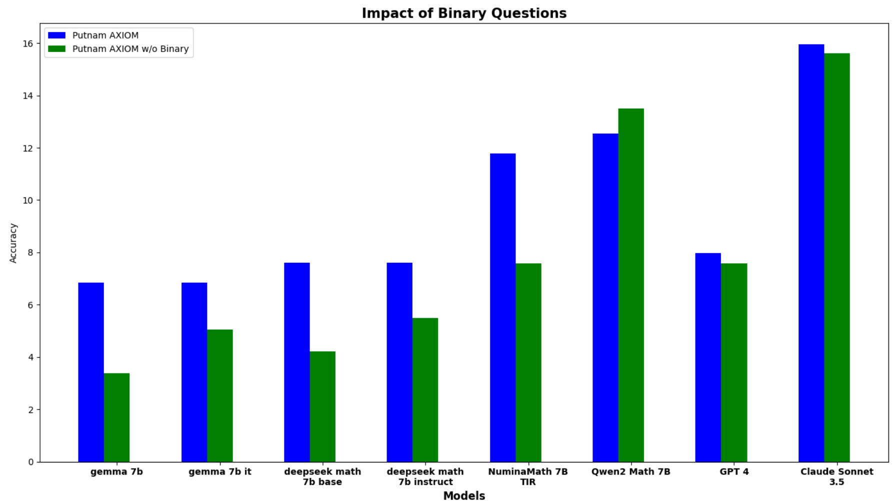

Figure 6: Putnam-AXIOM v.s. Putnam-AXIOM with only complex questions

A.2 Binary and Complex Questions

Several questions in Putnam-AXIOM are binary, meaning that the question inherently has two possible answers. These include true/false questions, questions about divergence or convergence, or questions about the winner of a two-player game. These questions make up 26 of the 262 question in Putnam-AXIOM Original; of the 60 questions of Putnam-AXIOM Variation, binary questions make up 7. We refer to all questions that are not binary as "complex" questions.

Given the guessable nature of these questions and our answer-matching evaluation method, models have a much higher chance of randomly guessing the right answer on these questions.

To discern whether the inclusion of these guessable questions significantly affects the overall difficulty of Putnam-AXIOM, we conducted an analysis of the accuracy of various models with and without the binary questions, with the overall accuracies in Figure 6.

We see that, with the exception of Qwen2 Math 7B, almost all models have a higher accuracy on Putnam-AXIOM with its binary questions than without, meaning that guessing is contributing to their success to some extent. However, we see that on the more advanced models—Qwen2 Math 7B, GPT 4, and Claude Sonnet 3.5—the gap between the accuracies on the entire dataset and the accuracies on only complex questions is much smaller. This is likely because these models are capable enough that they successfully answer a similar percentage of complex questions and binary questions; less advanced models get significantly fewer complex questions correct than binary questions, so we see a large accuracy gap.

Based on the results of this experiment, we've decided to use only the complex questions for most of our evaluations such as in Figure 1.

B Related Work

B.1 Mathematics benchmarks

Numerous benchmarks exist to assess the mathematical capabilities of models, each typically focusing on a specific task. Two notable examples are MATH [Hendrycks et al., 2021] and GSM8K [Cobbe et al., 2021]. The MATH dataset contains questions sourced from American high school mathematics competitions such as the AMC 10, AMC 12, and AIME [Hendrycks et al., 2021], while the GSM8K dataset contains 8.5K handwritten elementary school level questions Cobbe et al. [2021]. Both contain questions and answers with detailed rationale explanations.

As models have become larger and more powerful, even the most difficult existing benchmarks have become less challenging. For instance, while the MATH dataset saw 6.9% accuracy on its release, it now sees 87.92% accuracy with GPT-4 MACM [Lei, 2024]. Similarly, GPT4 has attained 97.1% accuracy on the GSM8K [Zhong et al., 2024]. This saturation necessitates the development of more challenging benchmarks.

Many contemporary data sets have been created to combat the saturation of existing benchmarks. For instance, the ARB dataset includes hundreds of challenging problems in high school and college-level math, physics, and chemistry Sawada et al. [2023]. Similarly OlympiadBench contains nearly 9,000 problems from the International Mathematics Olympiad (IMO), the Chinese GaoKao, and more He et al. [2024]. Finally, SciBench is a similar reasoning benchmark that includes hundreds of college-level scientific reasoning questions from instructional textbooks Wang et al. [2023].

Although these datasets alleviate the saturation problem, they come with many limitations. For instance, ARB Sawada et al. [2023] and OlympiadBench He et al. [2024] both contain several symbolic and proof-based questions which cannot be graded automatically and require a costly and lengthy human evaluation process. Though ARB attempts to utilize LLMs to grade their own responses with a rubric, this process is often unreliable and self-referential. Our Putnam-AXIOM dataset addresses these limitations by offering challenging Putnam problems with fully-written solutions and easily evaluable answers. It enables efficient automated assessment via frameworks like LM Harness [Gao et al., 2024], avoiding costly human evaluation or unreliable self-grading.

PutnamBench is a related benchmark that primarily focuses on formal theorem proving. Its main objective is to derive formalized proofs of mathematical statements and it provides formalizations in systems such as Lean, Isabelle, and Coq, all sourced from the prestigious Putnam competition. PutnamBench also includes 640 natural language statements and their corresponding answers where applicable. While both benchmarks draw from the same competition, *Putnam-AXIOM* focuses on the curation of natural language problems for final answer verification and introduces automatic functional variations to generate additional benchmarks addressing potential data contamination. For instance, *Putnam-AXIOM* removes questions that are easily guessable (e.g., where the final boxed answer is 0 or 1), ensuring that the benchmark better assesses the true math capabilities of models at the Putnam level.

B.2 Functional Benchmarks

Data contamination is a significant problem in creating evaluation benchmarks, as many of these problems are openly available on the Internet and are likely included in the training data for large models [Schaeffer, 2023, Sainz et al., 2023]. Thus, the MATH [Hendrycks et al., 2021], AGIEval [Zhong et al., 2023], OlympiadBench [He et al., 2024], and ARB [Sawada et al., 2023] benchmarks (which are all sourced from problems on the Internet) could potentially be contaminated. Therefore, models may achieve artificially high performance on an evaluation benchmark by memorizing the answers to the problems Magar and Schwartz [2022], Ranaldi et al. [2023].

A straightforward way of avoiding data contamination issues is to utilize problems unavailable on the Internet. However, even if problems are not currently part of model training data, it is unrealistic to expect them to remain inaccessible. At the same time, it is costly to rely on the continuous human development of new datasets.

Srivastava et al. [2024] attempts to alleviate this data contamination issue by creating *functional* variations of the MATH dataset, where new problems can be generated simply by changing numeric parameters, yielding different solutions. They observe a significant discrepancy in models' performance between standard benchmarks and these new variations. We recognize the potential of this idea and have adapted it to our more challenging dataset. We have altered the variables, constants, and phrasing of many Putnam questions while preserving their overall difficulty and requirements for logical and mathematical reasoning.