THE IMO SMALL CHALLENGE: NOT-TOO-HARD OLYMPIAD MATH DATASETS FOR LLMS*

Simon Frieder[†,1], **Mirek Olšák**[2], **Julius Berner**[3], **Thomas Lukasiewicz**[4,1]

[1]Department of Computer Science, University of Oxford
[2]Department of Pure Mathematics and Mathematical Statistics, University of Cambridge
[3]Department of Computing and Mathematical Sciences, Caltech
[4]Institute of Logic and Computation, Vienna University of Technology

ABSTRACT

We introduce the "IMO Small Challenge" (IMOSC), as opposed to the IMO Grand Challenge: A text-only, natural-language dataset consisting of competitive mathematical problems from various mathematical competitions. The IMOSC dataset exceeds the difficulty level of current datasets that are widely used for LLM evaluation, such as the MATH dataset, while not being too challenging for the current generation of LLMs. The IMOSC currently contains a carefully curated collection of the easiest possible problems from the International Mathematical Olympiad (IMO) and the Baltic Way Mathematical Contests (BWMC). Hardness is measured by applying a series of (objective) difficulty filters to the original problems. We release the full dataset under the link below to encourage transparent LLM and neurosymbolic system evaluation for mathematical proof-generating abilities:

Landing Page:
www.imo-small-challenge.io

1 INTRODUCTION AND MOTIVATION

The IMO *Grand* Challenge[1] (IMOGC)—which asks to automatically solve a full set of formalized *International Mathematical Olympiad* (IMO) problems[2] under stringent conditions–has received media attention in the last few years, although little tangible progress has been achieved. We argue that the reason for this is that solving IMO problems is very hard – for machine learning models and humans alike. However, the pace of progress in large language models' (LLMs') performance is rapid, with several models being specifically released for mathematical reasoning in the timespan of a few months (Gou et al., 2023; Luo et al., 2023; Azerbayev et al., 2023). This trend has further been accelerated by the outsized media impact of some models[3] and others.

This creates the need for a suitable evaluation dataset on an intermediate difficulty level, measured in terms of mathematical problem hardness. The MATH dataset (Hendrycks et al., 2021) and the GSM8K dataset (Cobbe et al., 2021) have both been released in 2021, about 7 months apart and make up the de facto benchmark in terms of mathematical reasoning of almost all large language models release since that time (Lightman et al., 2023; Luo et al., 2023; Azerbayev et al., 2023; Lewkowycz et al., 2022; Touvron et al., 2023). While GSM8K GPT's original motivation was to be an easier dataset than MATH, which was believed to be too hard for the language models at that time, both are now close to be considered solved. E.g., GPT-4 achieves 92% (OpenAI, 2023) by using a few-shot evaluation, while the usage of GPT-4 with tools leads to an accuracy of 83% on MATH (Zhou et al., 2023). Solving arbitrary IMO problems on the other hand, as argued by the IMOGC, which requires formal input and output, is arguably out of reach of the current generation LLMs. Recent autoformalization techniques tested a small selection of competitive mathematical problems. The miniF2F dataset (Zheng et al., 2021) achieves close to 40% Jiang et al. (2022), which is a far cry from the performance on the MATH dataset.

The right level of difficulty of a dataset is essential, in order to both stimulate research and to be an informative signal for researchers about how and where the (mathematical) failure modes of their models lie Frieder et al. (2023). If the dataset is too easy or too hard, little can be learned. Hence, we introduce the IMO Small Challenge, a dataset that is specifically tailored to proof-base mathematics and LLMs, which currently excel for text-only input and and output.

*Accepted for ICLR2024, Tiny Paper Track. S.F. conceived the project and wrote the paper. M.O. and J.B. helped with problem annotation and dataset generation, and T.L. with overall guidance.

†Corresponding author: simon.frieder@cs.ox.ac.uk.

[1]https://imo-grand-challenge.github.io/

[2]https://www.imo-official.org/problems.aspx

[3]Such as the Gemini model: https://deepmind.google/technologies/gemini

2 THE DATASET

The IMOSC is made up of competitive problems that are as easy as possible: A carefully-sourced dataset of problems that are either at the lowest level of IMO difficulty or below that (yet still count as competitive math problems), will make it possible to advance contemporary LLM systems, and serve as a stepping stone for the next generation of AI systems that solve mathematics. We measure how easy a problem is using three criteria; see Appendix A. One criterion pertains to proof lengths of the problem, where we use as proof length measure the longest of all known proofs. A secondary benefit of short proofs is that human auto-evaluation, which currently is inevitable for proof-based datasets, is less costly since it is faster to potentially reject a flawed short proof the LLM produces than a long proof. We contend that for the foreseeable time, it is unlikely that an LLM will output a solution that is not among the humanly known ones, so the existence of multiple known solutions is not (yet) problematic in this regard and likely an accurate predictor of the length of the proof that an LLM will generate. This is unlike in the general case, where we do not focus on very hard, competitive mathematical problems. For arbitrary mathematical statements that have a large number of proofs, it is not unreasonable to believe that LLMs can output new "combinations" of proofs. We note that famous theorems, such as Pythagoras' theorem, can have hundreds of proofs: https://www.cut-the-knot.org/pythagoras/.

Unlike MATH and GSM8K, IMOSC is *proof-based* to test specific problem-solving skills and mathematical creativity, which are specific to competitive mathematics. Formalization is not required for the IMOSC (which is problematic in itself, as it can occasionally lead to questions of how open-ended problems should be best formalized). Not focussing on autoformalization and a binary success criterion of whether the formal proof was correct or not easily allows raters and users of our dataset to award points for partial progress. Furthermore, because LLM's diagrammatic and visual reasoning abilities are still in their inceptions, we have excluded any problems where any graphical artifact is needed (a figure or a diagram) to either formulate the problem or understand its proof.

The table below summarizes the difference between the IMOSC and the IMO Grand Challenge (IMOGC).

	*IMO***GC**	*IMO***SC**
Any IMO difficulty level	yes	no (easy problems only)
Visual artifacts in the statement	yes	no
Visual artifacts in the proof	yes	no
Timelimit	yes (4.5 hours)	no
Querying the internet	no	yes
Formal input and output	yes	no (natural language input and output)

For this current short paper, we have focused exclusively on combinatorics as a prototype for the IMOSC; see Appendix C that explains the motivation for this choice. Besides the various measures for difficulty, no further annotations were done. IMOSC consists of 150 competitive mathematics problems, which are all annotated in terms of hardness. The criteria, as well as the dataset criterion pipeline, currently are geared toward combinatorics but will be generalized to other mathematical domains as we grow the IMOSC dataset – both in terms of domains, as well as in terms of mathematical competitions. After releasing it to the general public, further examples can be contributed.

The initial set of 100 combinatorics problems is made up as follows: 50 problems that were shortlisted for the IMO (a subset of which was used in IMO competitions) and 50 problems from 50 from the Baltic Way Mathematical Contests (BWMC).[4]. Our dataset spans the IMO shortlist from the years 2006 to 2022 (including), and 2011 to 2021 for BWMC. We focus exclusively on competitions for which statistics are available on the number of contests that solved each problem, as this gives us an objective way to assess the difficulty of each problem. It also allows users of our dataset to use it to evaluate their LLM or neurosymbolic system to assess how close their system comes to achieving (average) human performance on given problems.

The filtering process is described in detail in Appendix B, which makes the dataset generation reproducible. Some noteworthy points that arose from our filtering effort: We note that, for combinatorics problems, comparatively few became a final IMO competition problem: the 2011 C1 shortlisted problem, which became IMO 2011 problem 4, and the 2021 C3 shortlisted problem, which became IMO 2021 problem 5. Assessing solution lengths requires human inspection: We observe that for the 2006 C2 shortlisted problem, the second solution has a LaTeX character length of 900—by which it would have made it into the problems with the top six shortest solutions—but this does not represent the true solution lengths, since this second solution refers to the first, longer solution for a specific construction, thus increasing its length. We release the dataset under the CC BY-NC 4.0 license.

[4] https://www.math.olympiaadid.ut.ee/eng/html/?id=bw

URM STATEMENT

We acknowledge that one of the key authors of this work (first/last) meets the URM criteria of the ICLR 2024 Tiny Papers Track.

ACKNOWLEDGMENTS

This work was partially supported by the AXA Research Fund.

REFERENCES

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yujiu Yang, Minlie Huang, Nan Duan, Weizhu Chen, et al. ToRA: A tool-integrated reasoning agent for mathematical problem solving. *arXiv preprint arXiv:2309.17452*, 2023.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. WizardMath: Empowering mathematical reasoning for large language models via Reinforced Evol-Instruct. *arXiv preprint arXiv:2308.09583*, 2023.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q Jiang, Jia Deng, Stella Biderman, and Sean Welleck. Llemma: An open language model for mathematics. *arXiv preprint arXiv:2310.10631*, 2023.

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. *arXiv preprint arXiv:2103.03874*, 2021.

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

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857, 2022.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

OpenAI. GPT-4 technical report. *arXiv preprint 2303.0877*, 2023.

Aojun Zhou, Ke Wang, Zimu Lu, Weikang Shi, Sichun Luo, Zipeng Qin, Shaoqing Lu, Anya Jia, Linqi Song, Mingjie Zhan, et al. Solving challenging math word problems using gpt-4 code interpreter with code-based self-verification. *arXiv preprint arXiv:2308.07921*, 2023.

Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. MiniF2F: a cross-system benchmark for formal olympiad-level mathematics. *arXiv preprint arXiv:2109.00110*, 2021.

Albert Q Jiang, Sean Welleck, Jin Peng Zhou, Wenda Li, Jiacheng Liu, Mateja Jamnik, Timothée Lacroix, Yuhuai Wu, and Guillaume Lample. Draft, sketch, and prove: Guiding formal theorem provers with informal proofs. *arXiv preprint arXiv:2210.12283*, 2022.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. Mathematical capabilities of ChatGPT. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

A THE EASY-PROBLEM CRITERION

We use three different, general approaches to assess the difficulty of a competitive math problem outlined below.

- *Statistical difficulty*: For IMO and BWMC problems, statistics on the number of people that solved the problem, as well as the average score that was attained on that problem, are public[5]. Scores in the IMO are given on a scale from 0 to 7, where 0 is a completely wrong solution, and 7 is a perfect solution. We utilize this information to establish a cut-off for problems for which either a sufficiently high average score was obtained or were solved by sufficiently many people. For the BWMC the statistic are less detailed, but still give satisfactory insight into the difficulty of the problems
- *IMO difficulty*: For IMO problems, a selection of shortlisted problems (which in turn are selected from a list of problems that each participating country submits[6]) is initially made by problem creators, typically about seven problems, but the number varies between the years and the four mathematical domains of algebra, combinatorics, geometry, and number theory. Out of the shortlisted problems, the final IMO problems are selected – although minor changes can still be made at this stage, that does not affect the mathematics substantially[7]. The level of difficulty on the shortlisted problems roughly ascends in order, the first problems being the easiest problems, while the last ones are the hardest. While the previous assessment of difficulty was purely statistical, this one is subjective and reflects the IMO problem creators' assessment of what would be an easy problem. Hence, we use a cut-off on the problem numbers from the shortlists to exclude higher-numbered problems, which are harder.
  We note that there are instances where this assessment and the previous one dramatically diverged, as, for example, in the case of the "Windmill" example, shortlist problem C3 from 2011, which was the second problem in the final IMO competition. Thus, by the problem creators' assessment, this was not supposed to be a problem that was very difficult. Nonetheless, out of 563 contestants (all of which were already pre-selected for mathematical problem-solving ability at a national level), it was solved by only 22 contestants[8], indicating how statistical and subjective difficulty can diverge. For the BWMC we did not have access to shortlists, so this criterion does not apply.

  Subjective difficulty: It is plausible that the more solutions a problem has, the easier it is, as there are more possibilities to solve it. We have annotated each problem with subjective, but have not used this as a filtering criterion.
- *LLM generation difficulty*: Since language models output their tokens successively probabilistically, it is not implausible to believe that early errors can have outsized effects at later stages. Furthermore, shorter reference proofs make it easier to check correctness for humans if an LLM outputs the reference proof. Hence, we rank the problems by the length(s) of their solution(s). If a problem has multiple solutions, we thus use the *longest* one as a proxy for difficulty.

In Appendix B, we illustrate how these qualitative criteria can be turned into quantitative ones, and how to filter our initial selection of problems by applying these criteria.

B DATASET CREATION PIPELINE

Our dataset creation pipeline consists of a process of mixed human and automated elements. We start with IMO shortlists for combinatorics for the years 2006-2022 spanning 50 problems and 50 BWMC problems for the years 2011-2021, and we show in the following how the criteria from Appendix A are implemented.

For brevity, we illustrate only how our process works for IMO problems.

The IMO shortlists were used as a starting point for IMO-level problems since the lower numbered problems satisfy the *IMO difficulty* criterion mentioned in the previous section, Appendix A: For each year, we select the first three problems, C1-C3, as these have the lowest difficulty. This leaves us with 21 IMO shortlisted problems.

[5] See https://www.imo-official.org/year_statistics.aspx?year=2007 for the statistics on each problem for, e.g., the year 2007.

[6] As specified in the IMO regulations, see §6.5: https://www.imo-official.org/documents/RegulationsIMO.pdf

[7] Compare, e.g., the 2011 shortlisted problem C1, which was selected to be included in the final IMO problems, as problem 4.

[8] According to the IMO statistics for the year 2011: https://www.imo-official.org/year_statistics.aspx?year=2011. This problem was also discussed in other media channels due to its notoriety: https://www.3blue1brown.com/lessons/windmills.

Of these remaining shortlisted IMO problems, we manually extract the relevant page ranges for solutions to problems C1 to C3 and use `mathpix`[9] to convert them to LaTeX. We proceed manually to extract the proofs.

We adopt the following protocol for extracting the solutions from the previously obtained data:

- If figures, tables, or non-standard diagrams were used in the solution (we collectively refer to these as "graphical artifacts"), as is the case for various solutions in the problems from the IMO Shortlists, `mathpix` (and, we contend, most other current pdf-to-LaTeX convertors) will display them as images rather than Ti*k*Z code. We have opted to include the resulting `\includefigure` command in our solution as well as all other LaTeX-code artifacts that were produced. The reason for this is that such graphical artifacts are more information to process by a potential LLM having been trained to produce this solution or an LLM that outputs this solution. Thus, it is fair to add this code, which lengthens the proof. While for some problems, e.g. problem C2 from the IMO Shortlist 2007, the figures are essential to follow the proof, for other problems, such as problem C1 from the IMO Shortlist 2008, the figures are merely for orientation (in that problem the solution consists of certain box configurations, and the figure in the solution highlights one configuration). In the latter case, the figure is not strictly necessary to be included in the solution, but since it changes the solution length minimally, and often helps to understand the solution, we have opted to include it.
- All starting words such as "Solution.", or similar were removed, as were any comments at the end that were not relevant to the proof (e.g., comments about the proof's origin or other tangential information). We also exclude proof-ending words like "QED", should such words be used. If a problem has multiple statements to show, such as "(a)" or "(b)", and the solutions correspondingly also are split into a part "(a)" and part "(b)" (as is the case for problem C1 from the IMO Shortlist 2009), then these words are retained in the solution.
- If intermediate lemmas were formulated within a proof, these were kept unchanged, including the word *lemma*, as well as their entire proofs, including the word *proof*, as well as any proof ending words. We argue that it is fair to keep these "mathematical code words", as opposed to words such as "QED" that end a particular solution, since these denote general constructions or ideas that, for comprehension, need to be isolated.

A manual process of extracting proofs was necessary because of the diversity in which solutions are presented: For some shortlists, the solutions follow immediately after the problem statement (e.g., 2011); for others shortlists, they are at the end (e.g., 2009). Sometimes, further comments or observations are at the end of the solutions,[10] which also need to be excluded. This diversity of text structuring made automation challenging: An automatic, LLM-assisted pipeline was found not to perform well and to reliably identify only the solutions. The manual process that we followed may contain occasional errors, such as the length of the extracted solutions being off by a few characters - nonetheless, this does not alter our approach in any way.

We now apply the *LLM generation difficulty* criterion by selecting those IMO problems whose solutions are among the top half when counting the number of characters (excluding whitespaces or line breaks) their longest proof in LaTeX code has.

We now apply the *statistical difficulty* criterion from Appendix A, which we operationalize by selecting those problems that either had a score of at least 3.5 (in case of IMO) or were first or second-ranked in the number of people that solved them (IMO and BWMC). The problems that pass all these three filters receive an "IMOSC" label (but we include the full problem set in IMOSC, even those that do not have the IMOSC label).

C Mathematical Domain Choice

Our reason for focusing solely on combinatorics for this preliminary dataset is that contrary to other mathematical domains from which problems for competitive mathematics are sourced, combinatorics relies less on theoretical knowledge and more on elementary clever manipulation and new insights, with the problem helping us focus on the model's reasoning capabilities. The other three problem domains at competitions are typically algebra, geometry, and number theory. They also rely on clever insights, but sometimes these problems also have solutions that use certain theorems and methods for which prior knowledge is needed (e.g., the "bunching"[11] method, or the use of multi-variable calculus to solve certain inequalities, which often appear in the "algebra" section). Although we will include such problems in the IMOSC, we believe that combinatorics problems are the best testbed for pure mathematical reasoning, and chose to focus on this first.

[9] https://mathpix.com

[10] E.g., in case of problem C1 from the 2008 IMO Shortlist, the solution is followed by a paragraph with a comment, and by another section called "Original proposal", which discusses a variation of the given problem, C2.

[11] https://en.wikipedia.org/wiki/Muirhead%27s_Inequality