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]University of Oxford, [2]University of Cambridge, [3]Caltech, [4]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 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 difficult competitions, such as the International Mathematical Olympiad (IMO). Problem hardness is measured by applying a mixture of (objective and subjective) difficulty filters to the original problems. We release the full dataset under the link below to encourage transparent evaluation of LLMs and theorem provers toward their mathematical proof-generating abilities:

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 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 and make up the de facto benchmark in terms of mathematical reasoning of almost all LLMs released since that time (Lightman et al., 2023; Luo et al., 2023; Azerbayev et al., 2023; Lewkowycz et al., 2022; Touvron et al., 2023). GSM8K's original motivation was to be an easier dataset than MATH, which was believed to be too hard for the LLMs at that time. **Now GSM8K and MATH are close to be considered solved**. E.g., GPT-4 achieves 92% (OpenAI, 2023) by using few-shot evaluation, while the usage of GPT-4 with tools leads to an accuracy of 83% on MATH (Zhou et al., 2023). Yet solving arbitrary formalized IMO problems, as argued by the IMOGC, is arguably out of reach of the current generation LLMs. E.g., on the miniF2F dataset (Zheng et al., 2021) recent autoformalization techniques achieve close to only 40% accuracy (Jiang et al., 2022). (For brevity, when we subsequently refer to "LLMs", we also include other AI systems capable of solving mathematics automatically.)

The right level of difficulty of a dataset is essential in order to stimulate research and to lead to an informative signal for AI 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-based mathematics and LLMs**, which currently excel for text-only input and output.

[*]Corresponding author: simon.frieder@cs.ox.ac.uk. 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.

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

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

2 THE DATASET

The IMOSC is made up of the easiest possible IMO-level problems: A carefully sourced dataset of problems that are either at the lowest level of IMO difficulty or slightly below that will make it possible to advance contemporary LLM systems – as well as neurosymbolic solvers and theorem provers – 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 where one criterion pertains to the proof lengths of the problem, and we focus on those problems whose longest provided proof is a short one. (We elaborate in Appendix A on these three criteria and, in particular, on how proof length is defined.) A **benefit of problems with short proofs is that human evaluation of the LLMs' output is less costly**[3] since it is faster to inspect a short proof the LLM produces than a long proof.

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 allows raters and users of our dataset to award points for partial progress. Furthermore, because LLMs' diagrammatic and visual reasoning abilities are still in their inceptions, **we have excluded any problems where any graphical artifact is needed** (a figure, table, or a diagram) to either formulate the problem or understand its proof. The table below summarizes the differences between the IMOSC and the IMO Grand Challenge (IMOGC).

	*IMO***GC**	*IMO***SC**
Any competition 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 / 3 problems)	no
Querying the internet	no	yes
Formal input and output	yes	no (natural language input and output)

We focus exclusively on competitions for which at least for some problems basic statistics are available on the number of contestants[4] that solved each problem, as this gives us one objective way (out of the three mentioned above and outlined in Appendix A) to assess the difficulty of each problem. We, therefore, focus on the IMO and Baltic Way Mathematical Contests (BWMC)[5] which both fulfil this criterion. See Appendix D for a comparison of these competitions. Human statistics, where available, allow users of IMOSC to use it to evaluate their LLM to assess how close their system comes to achieving (average) human performance on given problems.

IMOSC currently consists of 100 competitive mathematics problems chosen from these competitions. Due to the two-page length limitation of this paper, we have focused exclusively on the domain of combinatorics as a prototypical illustration of the IMOSC; see Appendix C that explains the motivation for this choice. 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 the BWMC, for the years 2011 to 2021.

Each problem in IMOSC is annotated in terms of the three measures of difficulty outlined in Appendix A. **If a problem is in the top half for each difficulty measure, where available, it is labelled as "IMOSC"; for reference, the other problems are included as well.**

The filtering process is described in detail in Appendix B, in order to reproduce how we arrived at the dataset. Some parts of this process required human supervision. IMOSC will grow to encompass further mathematical domains and mathematical competitions. We will release the dataset in a versioned form to the general public, to allow datapoint submission by using GitHub pull requests.

We have released this dataset to support advancing the state-of-the-art of LLMs' abilities to solve competitive mathematical problems that require intricate reasoning. **We release the dataset under the CC BY-NC 4.0 license.**

[3] Auto-evaluation is currently an open problem for natural-language LLMs, so no alternative exists yet to human evaluation.

[4] Contestants can mean individuals, as well as teams, depending on the competition type.

[5] 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. We describe in Section B how these general approaches are instantiated for each of the IMO and BWMC. There may be problems for which there is data on all three of these measures is not available; in this case, we require that data on at least two may be available.

- **Statistical difficulty**: This type of difficulty assessment pertains to using statistics regarding the performance of the contestants that entered the competition on each problem.

  For IMO and BWMC problems, information regarding the number of people that solved the problem, as well as the average score that was attained on that problem, are public.[6] 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. For BWMC the range is 0 to 5. This information can be used to establish a cut-off for problems for which either a sufficiently high average score was obtained or which were solved by sufficiently many people. For the BWMC, the statistics are less detailed but still give satisfactory insight into the difficulty of the problems.

- **Competition difficulty**: Some competitions publish preliminary lists of problems, ranked in order of difficulty, which can be used as a source of information about problem difficulty, reflecting the organizers' difficulty assessment. By the nature of these problems, this criterion can be in conflict with the **statistical difficulty** criterion, since not all problems from such preliminary lists make it into a competition and hence information about contest performance will not be available for all such problems.

  For example, for IMO problems, a selection of shortlisted problems (which in turn are selected from a list of problems that each participating country submits[7]) 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.[8] 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, IMO 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 not 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,[9] indicating how statistical and subjective difficulty can diverge.

  For the BWMC we did not have access to shortlists, so this criterion does not apply.

- **LLM generation difficulty**: Since language models output their tokens successively, in a probabilistic manner, it is plausible to assume that early errors can have outsized effects at later stages. Furthermore, as mentioned in Section 2, shorter reference proofs make it easier to check correctness for humans if an LLM outputs a reference proof.[10] Hence, we

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

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

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

[9] 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.

[10] The assumption that short provided proofs also lead to short proofs generated by LLMs is based on the following hypotheses that we argue are currently true, due to the difficulty of IMIO-level problems: 1) The

rank the problems by the length(s) of their solution(s). We consider for this all officially provided solutions, as well as any solutions we either find in other sources or come up with ourselves. If a problem has multiple proofs, 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. The first and the last measure of difficulty are objective (although for the first one, it could be argued that the first one is also subjective to a degree, since it depends on the cohort that takes the test), while the intermediate one subjective; the Windmill problem illustrates the danger with this approach.

B Dataset Creation Pipeline

Our dataset creation pipeline consists of a process of mixed human and automated elements. We show in the following how the criteria from Appendix A are implemented. For brevity, we illustrate how our process works for IMO problems only. For BWMC problems, these are analogous, except that the **competition difficulty** criterion is not available.

We process each difficulty measure separately and, in the end, select those problems for the IMOSC that are in the lower half for *each* difficulty measure, where available. These problems receive the "IMOSC" label – but we include the full problem set in the IMOSC dataset, even those that do not have the IMOSC label, in order to allow users of our dataset to assess how these difficulty measures affect successful LLM generation.

Regarding **competition difficulty**, the IMO shortlists were used as a starting point for IMO-level problems since the lower numbered problems satisfy the **competition difficulty** criterion mentioned in the previous section: For each year, we select the first three problems, C1-C3, as there are often six shortlisted problems (although there can be more). This leaves us with 50 IMO shortlisted problems, which are our starting point.

Regarding the **LLM generation difficulty** criterion, we select those IMO problems whose solutions are among the top half when counting the number of characters (excluding whitespaces or line breaks) that their longest proof in LaTeX code has.

To assess solution length, we follow the following protocol: We manually extract the relevant page ranges for solutions and use `mathpix`[12] to convert them to LaTeX. We proceed manually to extract the proofs. We adopt the following rules:

- If figures, tables, or 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` will transform them into images rather than, e.g., Ti*k*Z code. (We contend, most other current pdf-to-LaTeX convertors will work similarly.) We included these preliminarily before deciding upon their relevance to understand a problem or its proof. If the figure is deemed to not be relevant for either the problem statement or its proof, we have opted to include the resulting `\includefigure` command in our solution, as well as all other LaTeX-code artifacts that were produced. E.g., for problem C2 from the IMO Shortlist 2007, the figures are essential to follow the proof so it was excluded; 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), so they were included.

  The reason for this is that such graphical artifacts are indicative that something that is not easy to understand from text is present, so it is fair to add this code, which lengthens the proof, to account for this information.

humanly known proofs are at most combinations of ideas contained in the provided proofs; 2) There are only a few proofs that are provided for each such problems; and 3) It is unlikely (in foreseeable time) that an LLM will output a solution that is not among the humanly known ones. This is unlike in the case of arbitrary (potentially easy) mathematical problems, where many proofs for a single problem can exist, and by combining individual proof ideas, it is possible to arrive at a large number of proofs,[11] which makes it hard to accurately determine the longest possible proof, as well as what the LLM might output.

[12] https://mathpix.com

- 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 – and thus carry more information than a "QED".

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,[13] which also need to be excluded. This diversity of text structuring made automation challenging: An automatic, GPT-4-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, but these errors are less than the errors occurring during an automated approach.

We operationalize **statistical difficulty** criterion from Appendix A, simply by using the official data available online on the number of contestants.

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" method,[14] 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 at a later stage, we believe that combinatorics problems are the best testbed for pure mathematical reasoning, and chose to focus on this first.

D Contest description

The IMO and the BWMC are both competitions on a similar level of mathematical difficulty, aimed at high-school students. The contest regulations stipulate that to participate at the BWMC one has to be a "possible candidate" for the IMO: https://balticway2023.de/regulations/.

Yet, there are significant differences. Since the BWMC is a per-country competition, where one team from each country competes against the other teams from the other countries, no information about the performance of individuals is available. Hence, the official statistics to not truly represent how well a single human is able to solve each of the given problems, but only aggregate team performance. We do not compare the difficulty of problems between competitions, so these facts are not problematic for our use case.

[13] 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.

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