Wu's Method Boosts Symbolic AI to Rival Silver Medalists and AlphaGeometry to Outperform Gold Medalists at IMO Geometry

Shiven Sinha[1*] **Ameya Prabhu**[2*] **Ponnurangam Kumaraguru**[1]
Siddharth Bhat[3+] **Matthias Bethge**[2+]
[1]IIIT Hyderabad [2] Tübingen AI Center, University of Tübingen [3]University of Cambridge

https://huggingface.co/datasets/bethgelab/simplegeometry

Abstract

Proving geometric theorems constitutes a hallmark of reasoning, combining intuitive, visual, and logical skills. This makes automated theorem proving of Olympiad-level geometry problems a milestone for human-level automated reasoning. AlphaGeometry, a neuro-symbolic model trained with 100M synthetic samples, solved 25 of 30 International Mathematical Olympiad (IMO) problems. It marked a major breakthrough compared to the reported baseline using Wu's method which solved only 10. Revisiting the IMO-AG-30 benchmark, we find that Wu's method is surprisingly strong and solves 15 problems, including some unsolved by other methods. This leads to two key findings: (i) Combining Wu's method with the classic synthetic methods of deductive databases and angle, ratio & distance chasing solves 21 out of 30 problems on a CPU-only laptop limited to 5 minutes per problem. Essentially, this classic method solves just 4 fewer problems than AlphaGeometry and establishes the first *fully symbolic* baseline that rivals the performance of IMO silver medalists. (ii) Wu's method even solves 2 of the 5 problems that AlphaGeometry failed on. Combining both, we set a new state-of-the-art for automated theorem proving on IMO-AG-30 solving 27 out of 30 problems – the first AI method which outperforms an IMO gold medalist.

1 Introduction

Automated theorem proving has been the long-term goal of developing computer programs that can match the conjecturing and proving capabilities demanded by mathematical research [10]. This field has recognized solving Olympiad-level geometry problems as a key milestone [2, 3], marking a frontier of computers to perform complex mathematical reasoning. The International Mathematical Olympiad (IMO) started in 1959 and hosts the most reputed theorem-proving competitions in the world that play an important role in identifying exceptional talents in problem solving. In fact, half of all Fields medalists participated in the IMO in their youth, and matching top human performances at the olympiad level has become a notable milestone of AI research.

Euclidean geometry is well suited to testing the reasoning skills of AI systems. It is finitely axiomatized [14] and many proof systems for Euclidean geometry have been proposed over the years which are amenable to automated theorem proving techniques [4, 5]. Furthermore, proof search can be guided by diagrammatic representations [12, 17], probabilistic verification [11, 21], and a vast array of possible deductions using human-designed heuristics for properties like angles, areas, and distances, methods affectionately called "*trig bashing*" and "*bary bashing*" [22, 23] by International Mathematical Olympiad (IMO) participants. In addition, this domain is challenging — specific proof systems need to be defined for specifying the problem, there is a shortage of data to train

[*]authors contributed equally, [+] equal advising

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

from, and problems typically contain ambiguities around degenerate cases [27, 19, 16] that are complex to resolve and have led to the humorous folklore that "*geometry problems never take care of degeneracies*".

Automated reasoning in geometry can be categorized into algebraic [26, 25, 15] and synthetic methods [12, 6, 20]. Recent focus has been on synthetic methods like Deductive Databases (DD) [6] that mimic human-like proving techniques and produce intelligible proofs by treating the problem of theorem proving as a step-by-step search problem using a set of geometry axioms. For instance, DD uses a fixed set of expert-curated geometric rules which are applied repeatedly to an initial geometric configuration. This is performed until the system reaches a fixpoint and no new facts can be deduced using the available rules. AlphaGeometry [24], a novel neuro-symbolic prover, represents a recent breakthrough advancement in this area. It adds additional rules to the prior work of DD to perform angle, ratio, and distance chasing (AR), and the resulting symbolic engine (DD+AR) is further enhanced using constructions suggested by a large language model (DD+AR+LLM-Constructions) trained on 100 million synthetic proofs. It has outclassed algebraic methods by solving 25 of 30 IMO problems, whereas the reported baseline based on Wu's method [26, 8] solved only ten [24].

Algebraic methods, such as Wu's method and the Gröbner basis method [15], transform geometric hypotheses into system of polynomials to verify conclusions. They offer powerful procedures that are proven to decide statements in broad classes of geometry [8, 15]. More precisely, Wu's method possesses the capability to address any problem for which the hypotheses and conclusion can be expressed using algebraic equations [7], while simultaneously generating non-degeneracy conditions automatically [27, 16]. This remarkable feature implies that Wu's method can handle problems not only in plane geometry but also in solid and higher-dimensional geometries, i.e. in areas where synthetic methods can be used only with great effort and additional considerations. [9].

Rather than indiscriminately tackling arbitrary problem instances, mathematicians concentrate their efforts on statements exhibiting specific properties that render them interesting, meaningful, and tractable within the broader context of mathematical inquiry [13]. In this work, we put the capabilities of Wu's method to the test on such structured problems and re-evaluate Wu's Method on the IMO-AG-30 benchmark introduced by Trinh et al. [24]. We find that it performs surprisingly strong, solving 15 problems, some of which are not solved by any of the other methods. This leads to two key findings:

- Combining Wu's method (Wu) with the classic synthetic methods of deductive databases (DD) and angle, ratio, and distance chasing (AR) solves 21 out of 30 methods by just using a CPU-only laptop with a time limit of 5 minutes per problem. Essentially, this classic method (Wu&DD+AR) solves just 4 problems less than AlphaGeometry and establishes the first *fully symbolic* baseline, strong enough to rival the performance of an IMO silver medalist.

- Wu's method even solves 2 of the 5 problems that AlphaGeometry (AG) failed to solve. Thus, by combining AlphaGeometry with Wu's method (Wu&AG) we set a new state-of-the-art for automated theorem proving on IMO-AG-30, solving 27 out of 30 problems, the first AI method which outperforms an IMO gold medalist.

2 Experiments & Results

2.1 Dataset

In January 2024, IMO-AG-30 was introduced as a new benchmark by Trinh et al. [24] to demonstrate the skill level of AlphaGeometry. IMO-AG-30 is based on geometry problems collected from the IMO competitions since 2000 and adapted to a narrower, specialized environment for classical geometry used in interactive graphical proof assistants, resulting in a test set of 30 classical geometry problems. The number of problems solved in this benchmark are related to the number of problems solved on average by IMO contestants. As indicated by the gray horizontal lines in Figure 1 (A), bronze, silver and gold medalists on average solved 19.3, 22.9 and 25.9 of these problems, and 15.2 represents the average over all contestants. The specific set of problems that have been collected for IMO-AG-30 are listed in the left column of the diagram in Figure 1 (B).

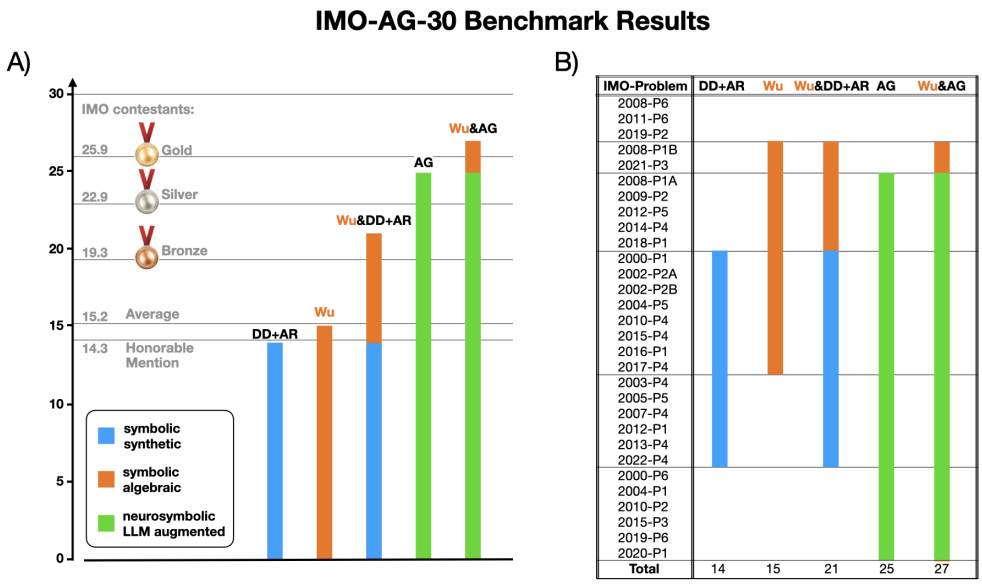


Figure 1: A) Performance across symbolic and LLM-Augmented methods on the IMO-AG-30 problem set, along with human performance. We set a strong baseline among symbolic systems at the standard of a silver medalist and outperform a gold medalist by a margin of one problem. B) Diagram showing how the different methods overlap or complement each other on the IMO-AG-30 problems.

2.2 Experimental Details

We evaluated performance using the IMO-AG-30 benchmark, with baselines and dataset all adopted from Trinh et al. [24]. We only re-implemented Wu's Method through the JGEX software [18, 28] by manual translation of the IMO-AG-30 problems into JGEX-compatible format[2]. We also successfully reproduced the DD+AR baseline, necessary for our final proposed method from the AlphaGeometry codebase. We manually verified that the hypothesis and conclusion equations generated by JGEX for several problems translated by us were indeed correct.

2.3 Results and Analysis

Our findings, are displayed in Figure 1 in combination with previous results from [24]. Figure 1 (A) compares the number of problems solved, and (B) shows which problems are solved by which method to visualize how the different methods overlap or complement each other. In Figure 1 (A), the performance levels of IMO contestants are indicated by gray horizontal lines, showing gold, silver, bronze, average, and honorable mention level. The performance levels of synthetic symbolic methods are displayed with blue bars and of LLM-augmented neurosymbolic methods are shown with green bars. Our own results obtained with Wu's method fall into the category of algebraic synthetic methods shown with orange bars. All results for synthetic symbolic methods (blue) or neurosymbolic LLM-augmented methods (green) are adopted from Trinh et al. [24].

Our combination of Wu's method with DD+AR sets a new symbolic baseline (Wu&DD+AR) that outperforms all traditional methods by a margin of 6 problems. It solves 21 of the IMO-AG-30 problems, matching the level of AlphaGeometry without fine-tuning (FT-9M only) shown in the Appendix (Figure 2). Wu's method achieves this performance with remarkably low computational requirements. On a laptop equipped with an AMD Ryzen 7 5800H processor and 16 GB of RAM, we were able to solve 14 out of 15 problems within 5 seconds. One problem (2015 P4) required 3 minutes. In our experiments, Wu's method either solves problems almost immediately or the laptop runs out of memory within 5 minutes. Remarkably, two of the fifteen problems we were able to solve with Wu's method (2021 P3, 2008 P1B) were among the five problems that were too difficult

[2]However, 4 out of 30 problems were untranslatable due to lack of appropriate constructions within the JGEX framework, hence our reported is out of 26 problems.

to solve for AlphaGeometry. Thus, by simple ensemble combination between Wu's method and AlphaGeometry, we obtain the new state-of-the-art solving 27 out of 30 problems on the IMO-AG-30 benchmark as visualized by the green/orange bar (Wu&AG) Figure 1.

3 Conclusion

Overall, our note highlights the potential of algebraic methods in automated geometric reasoning for solving International Mathematical Olympiad (IMO) geometry problems[3], raising the number of problems solved with Wu's method on IMO-AG-30 from ten to fifteen. Among those fifteen problems are several that are difficult for synthetic methods and their LLM-augmented versions that are currently most popular.

To the best of our knowledge, our symbolic baseline is the only symbolic baseline performing above the average IMO contestant and approaching the performance of an IMO silver medalist on geometry. Similarly, our combination of AlphaGeomtery with Wu's method is the first AI system to outperform a human gold-medalist at IMO geometry problems. This achievement illustrates the complementarity of algebraic and synthetic methods in this area (see Figure 1 B). The usefulness of algebraic approaches is most obvious from the two problems 2008 P1B and 2021 P3 which are currently solved by no automatic theorem prover other than Wu's method.

While algebraic methods have always been recognized for their theoretical guarantees, their usefulness has been previously questioned for being too slow and not human interpretable. Our observations indicate that on several problems Wu's Method performs more efficiently than previously recognized, and we advocate against dismissing it solely on the basis of its inability to produce human-readable proofs.

4 Limitations and Future Directions

Despite the theoretical promise, our results are a work-in-progress, currently hindered by the scarce availability of existing implementations, each with their significant inadequacies including limited constructions and suboptimal performance. We believe it might be feasible to outperform AlphaGeometry's proving capabilities through purely traditional methods and hope our note encourages improving current software for classical computational approaches in this area. Exploring improvements in the capabilities of other symbolic methods, including synthetic ones, in addition to extending the scope of geometry-specific languages and proof systems might be exciting directions to investigate.

Our exploration highlighting the complementary strengths of synthetic methods, which mimic human reasoning processes, and more abstract algebraic methods is motivated by the idea that the similarity to human reasoning and the generality of intelligence are distinct concepts, each with its own merits and applications. We believe that the strength of algebraic methods goes beyond solving Olympiad geometry problems, promising significant advancements in areas as varied as compiler verification and beyond. This potential underscores our belief in the necessity to broaden the scope of challenges addressed by automated theorem proving. The development of future benchmarks should strive for diversity and potentially open-ended testing. Embracing a wider array of problems will likely bring new insights on the usefulness, limitations, and interplay of neural and symbolic methods for general reasoning skills.

Acknowledgements

The authors would like to thank (in alphabetic order): Shashwat Goel, Shyamgopal Karthik, Yash Sharma, Matthias Tangemann, Saujas Vaduguru for helpful feedback on the draft. MB acknowledges financial support via the Open Philanthropy Foundation funded by the Good Ventures Foundation.

[3]Peter Novotný similarly proved 11 of the 17 IMO Geometry problems from 1984–2003 using the Gröbner basis method, although only after manually adding non-degeneracy conditions [1] as referenced here.

References

[1] Peter Novotný's Masters Thesis. https://skmo.sk/cvika/ukazpdf.php?pdf=diplomka.pdf.

[2] IMO Grand Challenge. https://imo-grand-challenge.github.io/, 2019. Online; accessed 29 May 2024.

[3] AIMO Prize. https://aimoprize.com/, 2023. Online; accessed 29 May 2024.

[4] Jeremy Avigad, Edward Dean, and John Mumma. A formal system for euclid's elements. *The Review of Symbolic Logic*, 2009.

[5] Michael Beeson, Pierre Boutry, Gabriel Braun, Charly Gries, and Julien Narboux. Geocoq. 2018.

[6] S.C. Chou, X.S. Gao, and J.Z. Zhang. A deductive database approach to automated geometry theorem proving and discovering. *Journal of Automated Reasoning*, 2000. doi: 10.1023/A:1006171315513.

[7] Shang-Ching Chou. Proving elementary geometry theorems using Wu's algorithm. In Woodrow Wilson Bledsoe and Donald W Loveland, editors, *Automated Theorem Proving: After 25 Years*, volume 89. American Mathematical Soc., 1984.

[8] Shang-Ching Chou. An introduction to Wu's method for mechanical theorem proving in geometry. *Journal of Automated Reasoning*, 1988.

[9] Shang-Ching Chou, Xiao-Shan Gao, and Jing-Zhong Zhang. Automated production of traditional proofs in solid geometry. *Journal of Automated Reasoning*, 14(2):257–291, 1995.

[10] Nicolaas Govert de Bruijn. *AUTOMATH, a language for mathematics*. 1983.

[11] Giuseppa Carr'a Ferro, Giovanni Gallo, and Rosario Gennaro. Probabilistic verification of elementary geometry statements. In *Automated Deduction in Geometry*, 1997. doi: 10.1007/BFb0022721.

[12] H. Gelernter. Realization of a geometry-theorem proving machine. *Computers & Thought*, 1995. doi: 10.5555/207644.207647.

[13] W. T. Gowers. How can it be feasible to find proofs? https://drive.google.com/file/d/1-FFa6nMVg18m1zPtoAQrFalwpx2YaGK4/view. Online; accessed 7 April 2024.

[14] Thomas Little Heath et al. *The thirteen books of Euclid's Elements*. 1956.

[15] Deepak Kapur. Using Gröbner bases to reason about geometry problems. *Journal of Symbolic Computation*, 1986.

[16] Deepak Kapur. A refutational approach to geometry theorem proving. *Artificial Intelligence*, 1988.

[17] Michelle Y. Kim. Visual reasoning in geometry theorem proving. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, 1989.

[18] Zoltán Kovács and Alexander Vujic. Open source prover in the attic. *arXiv preprint arXiv:2401.13702*, 2024.

[19] Zoltán Kovács, Tomas Recio, Luis F Tabera, and M Pilar Vélez. Dealing with degeneracies in automated theorem proving in geometry. *Mathematics*, 2021.

[20] Arthur J Nevins. Plane geometry theorem proving using forward chaining. *Artificial Intelligence*, 6(1): 1–23, 1975.

[21] Jürgen Richter-Gebert and Ulrich Kortenkamp. *The Interactive Geometry Software Cinderella*. 1999.

[22] Max Schindler and Evan Chen. Barycentric coordinates in olympiad geometry. *Olympiad Articles*, 2012.

[23] Justin Stevens. Coordinate and trigonometry bashing. http://services.artofproblemsolving.com/download.php?id=YXR0YWNobWVudHMvYi9kLzRmMTA5OWJhNmI1MTg2YzM2ODdkZTVhYTJjMGU0NjdmYmViNGRk&rn=Q29vcmRpbmF0ZSBhbmQgVHJpZ29ub21ldHJ5IEJhc2hpbmcucGRm. Accessed: 4 April 2024.

[24] Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 2024.

[25] Dongming Wang. Reasoning about geometric problems using an elimination method. *Automated practical reasoning: Algebraic approaches*, pages 147–185, 1995.

[26] Wu Wen-Tsün. On the decision problem and the mechanization of theorem proving in elementary geometry. *Scientia Sinica*, 1978.

[27] Wenjun Wu. On zeros of algebraic equations–an application of ritt principle. *Kexue Tongbao*, 1986.

[28] Z. Ye, S.C. Chou, and X.S. Gao. An introduction to java geometry expert – (extended abstract). 2011. doi: 10.1007/978-3-642-21046-4_10.

A Detailed Comparisons

We compare with all human and automated methods on the IMO-AG-30 benchmark [24] in Figure 2. Our evaluation includes GPT4, Full-Angle method (FA), Gröbner Basis (Gröbner), Deductive Databases (DD), Deductive Databases combined with Algebraic Rules and enhancements with GPT-4 for construction suggestions (DD+AR+GPT4). Additionally, we examined different configurations of the AlphaGeometry model: one only pretrained on 100 million samples (PT-100M) and other only finetuned on 9 million constructions (FT-9M). Note that we construct the Wu&DD+AR baseline by simply parallely running both Wu's and DD+AR methods and stopping when either method solves the problem. Similarly, we construct the Wu&AlphaGeometry baseline. We see that our Wu&DD+AR baseline matches AG (FT-9M) baseline while Wu's method alone matches the best DD+AR+GPT4 algorithm.

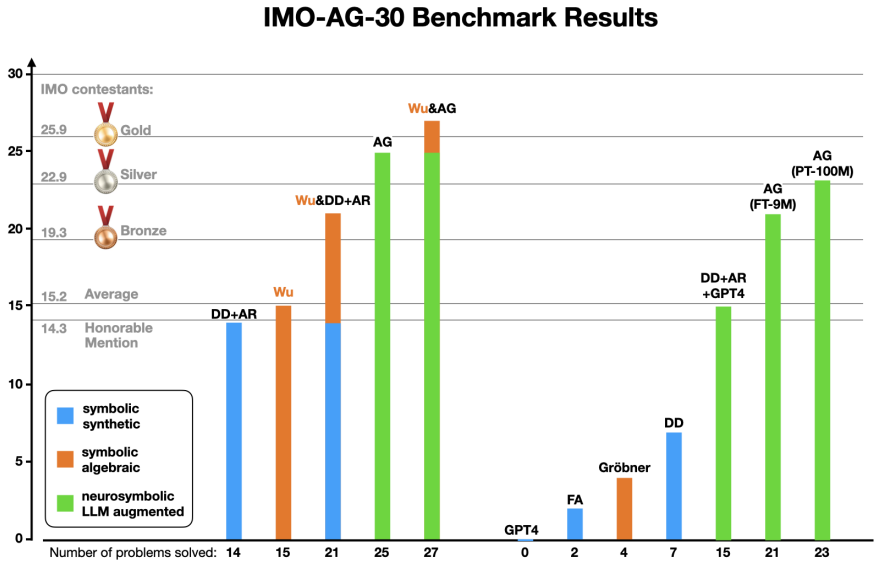
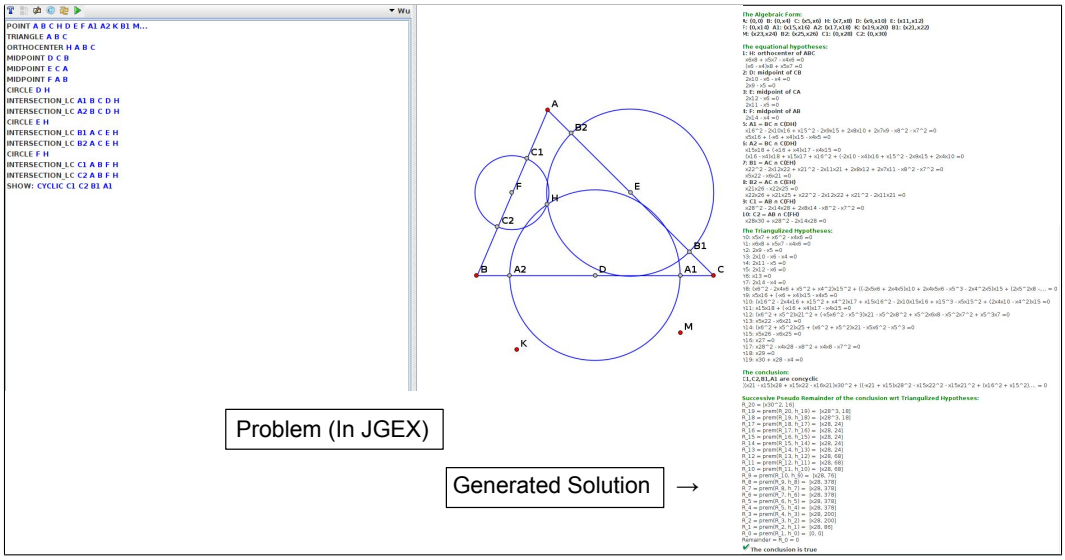


Figure 2: **Extended version of Figure 1A:** Performance across symbolic and LLM-Augmented methods on the IMO-AG-30 problem set, along with human performance. The performance of additional models adopted from Table 1 in [24] are shown on the right.

B Illustrations: 2008 P1B and 2021 P3

We provide illustrations of the solutions of Wu's method for the two problems AlphaGeometry could not solve to allow for additional scrutiny without having to reproduce the same on the JGEX solver.



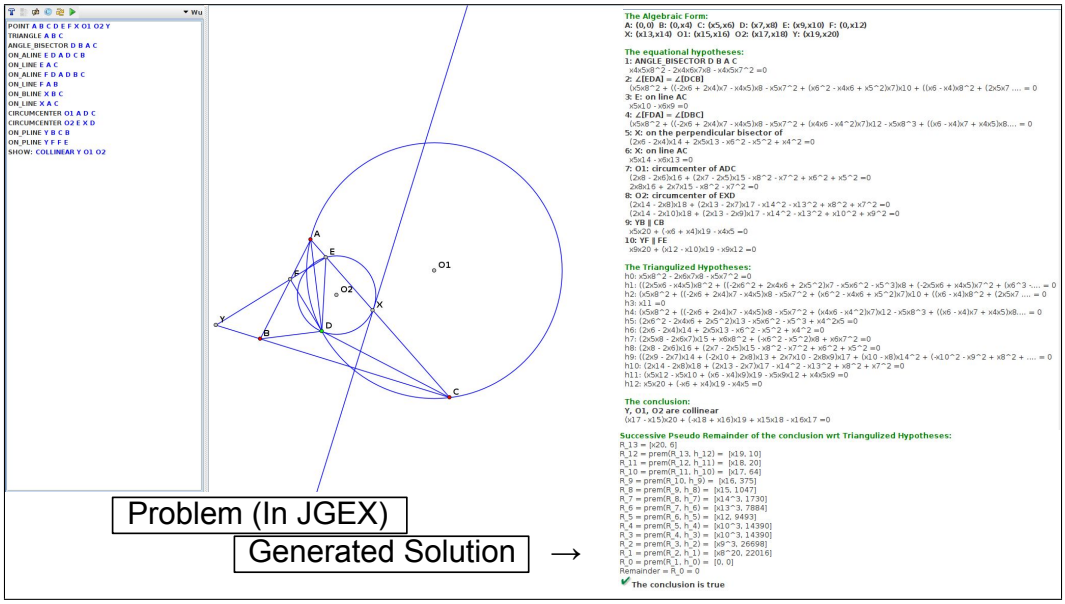


Figure 3: **Problem 2008-P1B JGEX (Above) and 2021-P3 (Below) with Input (Left) and Generated Solution (Right) for Wu's method.** This illustration can be reproduced by opening the .gex files provided alongside on the HuggingFace repository and pressing Run.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and/or introduction clearly state the claims made, including the contributions made in the paper along with important assumptions and limitations.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We acknowledge limitations concerning a non-extensive set of constructions to support the algebraic methods, underdeveloped software tools, and challenges regarding human interpretability, among other details in Sections 1 and 4.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The theoretical underpinnings of Euclidean geometry [14, 4], along with the synthetic [12, 6, 20] and algebraic methods [26, 25, 15] of automated theorem proving discussed in our work are covered extensively in past literature available in our references. We do not claim any novel theoretical contributions.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide detailed descriptions of the dataset (derived from IMO-AG-30), software (e.g., JGEX), and experimental setup, including the hardware and the methods used for the tests on the benchmark. We also release the dataset containing manual translations for each problem. This should be sufficient for reproducing the results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide access to the translated dataset of IMO problems, links to relevant software, list of methods used, and the exact steps for performing the experiments.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: In our dataset, we provide the precise translations, references to their corresponding IMO problems, and a link to the JGEX software that can parse these translations and execute the relevant methods. We also provide the time limits and computational constraints applied during testing (Section 2).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Since our evaluations are based on discrete problem-solving results on IMO problems using fixed translations and deterministic software tools, we highlight that our experiments are exactly reproducible. As a result, tests of statistical significance or error bars are not applicable.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We describe the computing resources used, specifically a CPU-only laptop with an AMD Ryzen 7 5800H processor and 16 GB RAM, as well as the time constraints per problem (Section 2).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have read the ethics guidelines and confirm that we do not use human subjects, respect the licensed use of dataset (IMO-AG-30), and do not include any personally identifiable information in the dataset of translated IMO problems introduced by us.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss the importance of advancing automated reasoning systems, which could also lead to improvements formal verification systems and educational tools. We do not foresee any negative societal impacts from this research.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The dataset of translated IMO problems released here does not pose high risks for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We credit the creators of the JGEX software used to implement Wu's method, the IMO-AG-30 dataset, as well as other cited methods and datasets, as referenced throughout the paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The dataset of translated IMO problems introduced in the paper has extensive documentation, including details on reproducing our results using relevant software.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve research with human subjects, so IRB approval is not applicable.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.