Evaluating AlphaGeometry in the Wild

Yuxi Ma **Chi Zhang** **Jiajun Song** **Wei Wang**

National Key Laboratory of General Artificial Intelligence
Beijing Institute for General Artificial Intelligence (BIGAI)
{mayuxi,zhangchi,songjiajun,wangwei}@bigai.ai

Abstract

In this report, we evaluated AlphaGeometry recently introduced from Google DeepMind on a much larger set of 125 competition-level geometry proving problems. AlphaGeometry solved **56** out of the **125** problems.

Introduction

Successfully proving geometry theorems is a remarkable intellectual milestone for humankind, indicating exceptional ability in logical reasoning and problem-solving involving graphs. Most notably, the ability is deemed one of the greatest challenges in formal reasoning for humans and tested in the most well-acknowledged intellectual competition in the world, the International Mathematical Olympiad (IMO).

As such, a mechanical geometry theorem proving system has been pursued in the artificial intelligence and mathematical community for decades from multiple perspectives [1, 3]. However, the task still remains challenging due to the following reasons: 1) data from human theorem proving is scarce, making it hard to effectively utilize machine learning approaches to solve the challenge; 2) the search space for constructing necessary auxiliary points that are fundamental to geometry theorem proving is infinitely large, adding to the difficulty of finding a proper solution [2].

On January 17th, 2024, Google DeepMind introduced AlphaGeometry, asserting its capability to tackle Olympiad-level geometry problems [2]. They evaluated the system on 30 archived IMO geometry problems from 2000 to 2020, finding that AlphaGeometry successfully solved 25 out of the 30 problems.

However, due to the limited size of the benchmarking test, it is questionable how effectively AlphaGeometry can perform on other Olympiad-level geometry problems. In this report, we tested the system on curated 125 competition-level geometry problems, covering the 30 test problems used in the AlphaGeometry benchmark and in addition, IMO (1959-2023), USAMO (1988-2023), China Team Selection Test (1986-2021), selected CMO problems, China 9th Grade competition problems, and famous geometry theorems.

AlphaGeometry solved **56** of the **125** problems.

Results

To evaluate AlphaGeometry in the wild, we used the open-sourced codebase[1] from the authors. Due to limited parallelism support in the codebase, we used a device with 128 CPU cores and 8 Nvidia 3090 GPUs for evaluation, assigning 1 GPU for each problem search. We followed the original setup in the paper and set the batch size to 32, beam size to 512, and search depth to 16. Considering timing constraint for human participants in the competition, we set the total timeout for both DD+AR and language model prediction to 1.5 hour for each problem.

In our tests, we categorized our 125 test cases into four categories: IMO, USAMO, China Team Selection Test (China TST), and Others (for selected CMO, China 9th Grade competition problems, and famous geometry theorems). During the inference process, we noted that apart from problem Solved and Timeout, the AlphaGeometry system would occasionally fail to initialize the graph after multiple tries, derive conflicting facts, or generate invalid auxiliary constructions, which are denoted as Error. Table 1 shows our experimental results.

Table 1: **Results from evaluating AlphaGeometry in the wild.** DDAR Solved denotes the number of problems solved without language model among all problems solved in the category. We show the percentage in each entry.

Test	Solved	Timeout	Error	DDAR Solved
IMO	24/42	7/42	11/42	23/24
USAMO	4/17	7/17	6/17	3/4
China TST	15/44	20/44	9/44	11/15
Others	13/22	3/22	6/22	11/13
Total	56/125	37/125	32/125	48/56

We released all the 125 testing logs in the repository[2].

Disclaimer

This paper represents the opinions of the authors and is the product of professional research. It is not meant to represent the position or opinions of the National Key Laboratory system, nor the official position of any of its staff members.

References

[1] Shang-Ching Chou. *Mechanical geometry theorem proving*. Springer, 1988.

[2] Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024.

[3] Wen-tsün Wu. *Mechanical theorem proving in geometries: Basic principles*. Springer Science & Business Media, 2012.

[1] https://github.com/google-deepmind/alphageometry
[2] https://github.com/WellyZhang/alphageometry-test