

Top AI models solve under 2% of Epoch AI's new FrontierMath research problems

Source copy kept by wheresthe.ai for the story "Top AI models solve under 2% of Epoch AI's new FrontierMath research problems" (<https://wheresthe.ai/d/epoch-frontiermath-benchmark-2024-11/>).

Source: <https://arxiv.org/abs/2411.04872>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: e82986a1494aecda31098aa9e59c614a385327878fa0acd15cbb2b4966b9ebec

-->

Computer Science > Artificial Intelligence

arXiv:2411.04872 (cs)

[Submitted on 7 Nov 2024 (v1), last revised 23 Dec 2025 (this version, v7)]

Title: FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI

Authors: Elliot Glazer , Ege Erdil , Tamay Besiroglu , Diego Chicharro , Evan Chen , Alex Gunning , Caroline Falkman Olsson , Jean-Stanislas Denain , Anson Ho , Emily de Oliveira Santos , Olli Järvinemi , Matthew Barnett , Robert Sandler , Matej Vrzala , Jaime Sevilla , Qiuyu Ren , Elizabeth Pratt , Lionel Levine , Grant Barkley , Natalie Stewart , Bogdan Grechuk , Tetiana Grechuk , Shreepranav Varma Enugandla , Mark Wildon View a PDF of the paper titled FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI, by Elliot Glazer and 23 other authors

View PDF HTML (experimental)

Abstract: We introduce FrontierMath, a benchmark of hundreds of original, exceptionally challenging mathematics problems crafted and vetted by expert mathematicians. The questions cover most major branches of modern mathematics -- from computationally intensive problems in number theory and real analysis to abstract questions in algebraic geometry and category theory. Solving a typical problem requires multiple hours of effort from a researcher in the relevant branch of mathematics, and for the upper end questions, multiple days. FrontierMath uses new, unpublished problems and automated verification to reliably evaluate models while minimizing risk of data contamination. Current state-of-the-art AI models solve under 2% of problems, revealing a vast gap between AI capabilities and the prowess of the mathematical community. As AI systems advance toward expert-level mathematical abilities, FrontierMath offers a rigorous testbed that quantifies their progress.

Subjects:

Artificial Intelligence (cs.AI)

Cite as: arXiv:2411.04872 [cs.AI]

(or arXiv:2411.04872v7 [cs.AI] for this version)

<https://doi.org/10.48550/arXiv.2411.04872>

Focus to learn more

arXiv-issued DOI via DataCite

Submission history From: Tamay Besiroglu [view email] [v1] Thu, 7 Nov 2024 17:07:35 UTC (909 KB)

[v2] Wed, 13 Nov 2024 01:06:35 UTC (911 KB)

[v3] Thu, 14 Nov 2024 16:26:03 UTC (911 KB)

[v4] Fri, 22 Nov 2024 18:00:00 UTC (911 KB)

[v5] Fri, 20 Dec 2024 03:27:06 UTC (911 KB)

[v6] Thu, 28 Aug 2025 23:10:50 UTC (991 KB)

[v7] Tue, 23 Dec 2025 02:23:47 UTC (1,021 KB)

Full-text links: Access Paper:

View a PDF of the paper titled FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI, by Elliot Glazer and 23 other authors View PDF HTML (experimental) TeX Source

view license

Current browse context:

cs.AI

< prev

| next >

new | recent | 2024-11

Change to browse by:

cs

References & Citations

NASA ADS Google Scholar

Semantic Scholar

1 blog link (what is this?)

export BibTeX citation Loading...

BibTeX formatted citation

×

loading...

Data provided by:

Bookmark

Bibliographic Tools

Bibliographic and Citation Tools

Bibliographic Explorer Toggle

Bibliographic Explorer (What is the Explorer?)

Connected Papers Toggle

Connected Papers (What is Connected Papers?)

Litmaps Toggle

Litmaps (What is Litmaps?)

scite.ai Toggle

scite Smart Citations (What are Smart Citations?)

Code, Data, Media

Code, Data and Media Associated with this Article

alphaXiv Toggle

alphaXiv (What is alphaXiv?)

Links to Code Toggle

CatalyzeX Code Finder for Papers (What is CatalyzeX?)

DagsHub Toggle

DagsHub (What is DagsHub?)

GotitPub Toggle

Gotit.pub (What is GotitPub?)

Huggingface Toggle

Hugging Face (What is Huggingface?)

ScienceCast Toggle

ScienceCast (What is ScienceCast?)

Demos

Demos

Replicate Toggle

Replicate (What is Replicate?)

Spaces Toggle

Hugging Face Spaces (What is Spaces?)

Spaces Toggle

TXYZ.AI (What is TXYZ.AI?)

Related Papers

Recommenders and Search Tools

Link to Influence Flower

Influence Flower (What are Influence Flowers?)

Core recommender toggle

CORE Recommender (What is CORE?)

Author

Venue

Institution

Topic

About arXivLabs

arXivLabs: experimental projects with community collaborators

arXivLabs is a framework that allows collaborators to develop and share new arXiv features directly on our website.

Both individuals and organizations that work with arXivLabs have embraced and accepted our values of openness, community, excellence, and user data privacy. arXiv is committed to these values and only works with partners that adhere to them.

Have an idea for a project that will add value for arXiv's community? Learn more about arXivLabs .

Which authors of this paper are endorsers? | Disable MathJax (What is MathJax?)