FrontierMath

FRONTIER-MATH · math benchmark · 2024

Source: https://policywindow.org/wiki/frontiermath

Generated 2026-11-19T01:05:48 UTC

Summary

Hundreds of original research-mathematician-curated math problems requiring deep reasoning. Held-out evaluation only.

At a glance

Score range

0–100 % accuracy

Contamination risk

low

Methodology

https://epochai.org/frontiermath

Saturation

active

Details

Epoch AI eval. Top reasoning models 2-5% at launch; OpenAI o3-preview reported 25% under custom harness. Currency (2026-06-21): Article's data stops at Apr 2025 (o4-mini 17%, o3 10%); since then SOTA on Tiers 1-3 rose to >40% (GPT-5.2 / Claude Opus 4.6, per IEEE Spectrum) and Epoch's own record on Tier 4 hit 31% (GPT-5.2 Pro, 15/48, Jan 2026, up from a 19% prior max), plus Epoch shipped FrontierMath v2 on 2026-06-12 — an AI-assisted audit found critical errors in ~42% of problems (135 corrected, 12 removed ’ 338 total), a notable validity finding.

How to cite this article

APA

Policy Window. (2024). FrontierMath [Wiki article — Benchmark]. https://policywindow.org/wiki/frontiermath

CHICAGO

Policy Window. 2024. "FrontierMath." Wiki article (Benchmark). https://policywindow.org/wiki/frontiermath.

HARVARD

Policy Window (2024) 'FrontierMath', Wiki article — Benchmark, available at: https://policywindow.org/wiki/frontiermath.

OSCOLA

Policy Window, 'FrontierMath' (Wiki article — Benchmark, 2024) <https://policywindow.org/wiki/frontiermath> accessed [date].

BIBTEX

```
@misc{policywindow-frontiermath,
title  = {FrontierMath},
author = {Policy Window},
year   = {2024},
howpublished = {FRONTIER-MATH (2024)},
url    = {https://policywindow.org/wiki/frontiermath},
```

```
note   = {Primary source: https://epochai.org/frontiermath}
}
```