MATH (Hendrycks)

MATH · math benchmark · 2021

Source: https://policywindow.org/wiki/math-benchmark

Generated 2026-07-09T20:25:18 UTC

Summary

12,500 competition-math problems from AMC, AIME, etc. Evaluates step-by-step reasoning + final-answer accuracy.

At a glance

Score range

0–100 % accuracy

Contamination risk

medium

Methodology

https://arxiv.org/abs/2103.03874

Saturation

saturated

Details

Frontier reasoning models 90%+. AIME-2024 is the harder successor for unsaturated math eval. Currency (2026-06-21): MATH/MATH-500 is now even more thoroughly saturated than the article's latest cited data point (OpenAI o1, 94.8%, 2024) — current frontier models cluster at ~99% on MATH-500 (e.g. GPT-5 99.4%, o3 99.2%, LongCat-Flash-Thinking 99.2% per Artificial Analysis/llm-stats leaderboards), reinforcing (not contradicting) the article's saturation thesis; optional enrichment would add a post-2024 ceiling row, but no existing claim is stale.

How to cite this article

APA

Policy Window. (2021). MATH (Hendrycks) [Wiki article — Benchmark]. https://policywindow.org/wiki/math-benchmark

CHICAGO

Policy Window. 2021. "MATH (Hendrycks)." Wiki article (Benchmark). https://policywindow.org/wiki/math-benchmark.

HARVARD

Policy Window (2021) 'MATH (Hendrycks)', Wiki article — Benchmark, available at: https://policywindow.org/wiki/math-benchmark.

OSCOLA

Policy Window, 'MATH (Hendrycks)' (Wiki article — Benchmark, 2021) <https://policywindow.org/wiki/math-benchmark> accessed [date].

BIBTEX

```
@misc{policywindow-math-benchmark,
title  = {MATH (Hendrycks)},
author = {Policy Window},
year   = {2021},
```

```
howpublished = {MATH (2021)},
url    = {https://policywindow.org/wiki/math-benchmark},
note   = {Primary source: https://arxiv.org/abs/2103.03874}
}
```