

MATH (Hendrycks)

MATH · math benchmark · 2021

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

Generated 2026-05-30T22:07:31 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

Methodology

<https://arxiv.org/abs/2103.03874>

Contamination risk

medium

Saturation

saturated

Details

Frontier reasoning models 90%+. AIME-2024 is the harder successor for unsaturated math eval.

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}  
}
```