

Title: State of AI Reasoning for Theoretical Physics - Insights from the TPBench Project

Speakers: Moritz Munchmeyer

Collection/Series: Theory + AI Symposium

Date: April 08, 2025 - 9:40 AM

URL: <https://pirsa.org/25040061>

Abstract:

The newest large-language reasoning models are for the first time powerful enough to perform mathematical reasoning in theoretical physics at graduate level. In the mathematics community, data sets such as FrontierMath are being used to drive progress and evaluate models, but theoretical physics has so far received less attention. In this talk I will present our dataset TPBench (arxiv:2502.15815, tpbench.org), which was constructed to benchmark and improve AI models specifically for theoretical physics. We find extremely rapid progress of models over the last months, but also significant challenges at research level difficulty. I will also briefly outline strategies to improve these models for theoretical physics.

State of AI Reasoning for Theoretical Physics - Insights from the TPBench Project

Theory+AI Symposium, Perimeter Institute

April 8th 2025

Moritz Münchmeyer, UW Madison



LLM progress on Math

Math benchmarks such as AIME are widely used to evaluate and improve reasoning models.

MathArena: Evaluating LLMs on Uncontaminated Math Competitions

Model	Overall	AIME 2025 I	AIME 2025 II	HMMT February 2025	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
o3-mini (high)	93.33%	\$2.85			Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
o3-mini (medium)	80.00%	\$1.62			Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
o1 (medium)	80.00%	\$41.05			Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
DeepSeek-R1	75.00%	\$4.69			Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
DeepSeek-R1-Distill-Qwen-32B	65.00%	N/A			Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
DeepSeek-R1-Distill-Llama-70B	60.00%	\$1.21			Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
gemini-2.0-flash-thinking	55.00%	N/A			Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
DeepSeek-R1-Distill-Qwen-14B	48.33%	\$1.26			Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
o3-mini (low)	43.33%	\$0.65			Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
QwQ-32B-Preview	30.00%	\$0.62			Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
gemini-2.0-pro	28.33%	\$1.02			Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
gemini-2.0-flash	25.00%	\$0.08			Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
DeepSeek-V3	21.67%	\$0.21			Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
DeepSeek-R1-Distill-Qwen-1.5B	15.00%	\$0.21			Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
gpt-4o	13.33%	\$0.55			Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
claude-3.5-sonnet	3.33%	\$0.55			Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green

AIME results from <https://matharena.ai/>

Research Level: FrontierMath

FRONTIERMATH: A BENCHMARK FOR EVALUATING ADVANCED MATHEMATICAL REASONING IN AI

Elliot Glazer^{1†‡}, Ege Erdil^{1†}, Tamay Besiroglu^{1†}, Diego Chicharro^{2‡}, Evan Chen^{3‡},
Alex Gunning^{4‡}, Caroline Falkman Olsson¹, Jean-Stanislas Denain¹, Anson Ho¹, Emily de Oliveira Santos^{5‡},
Olli Järvineniemi, Matthew Barnett¹, Robert Sandler¹, Matej Vrzala¹, Jaime Sevilla¹,
Qiuyu Ren^{6‡}, Elizabeth Pratt^{6‡}, Lionel Levine^{7‡}, Grant Barkley^{8‡}, Natalie Stewart^{8‡},
Bogdan Grechuk^{9‡}, Tetiana Grechuk^{9‡}, Shreepranav Varma Enugandla^{6‡}, Mark Wildon^{10‡}

¹Epoch AI

⁵ICMC, USP

⁹University of Leicester

²King's College London

⁶UC Berkeley

¹⁰University of Bristol

[†]Core contributor

³MIT

⁷Cornell University

[‡]Contributing mathematician

⁴University of Siegen

⁸Harvard University

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.

[Submitted on 7 Nov 2024 (v1)]

FrontierMath: First research level benchmark

IMPRESSIONS OF OUR RESEARCH-LEVEL PROBLEMS
(TOP 25% OF DIFFICULTY)

“These are extremely challenging... I think they will resist AIs for several years at least.”



Terence Tao
Fields Medalist (2006)

<https://epoch.ai/frontiermath>

	Pass@1	Pass@4	Pass@8
o3-mini (high)	9.2%	16.6%	20.0%
o1-mini	5.8%	9.9%	12.8%
o1	5.5%	10%	12.8%

Research-level mathematics: OpenAI o3-mini with high reasoning performs better than its predecessor on FrontierMath. On FrontierMath, when prompted to use a Python tool, o3-mini with high reasoning effort solves over 32% of problems on the first attempt, including more than 28% of the challenging (T3) problems. These numbers are provisional, and the chart above shows performance without tools or a calculator.

<https://openai.com/index/openai-o3-mini/>

What is novel about reasoning models?

- The first reasoning model, **o1-preview** from **OpenAI**, **massively improved scores on our physics benchmark**, as you will see.
- **Modern Reasoning** models are **trained to generate reasoning** (chains of thought) using **both supervise fine-tuning** on curated chains-of-thought **and reinforcement learning**.

The fact that mathematical reasoning can now be learned with optimization (Reinforcement Learning) will be impactful for theoretical research.

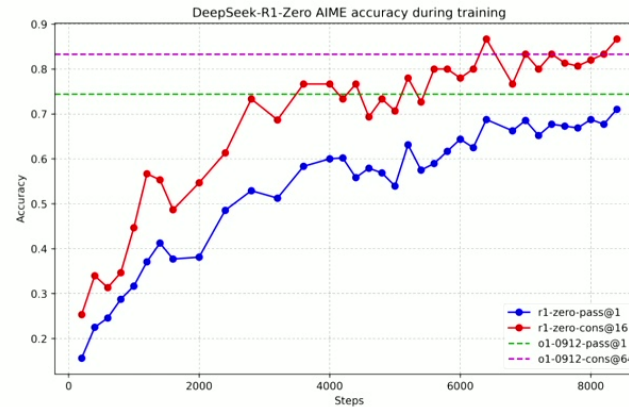


Figure 2 | AIME accuracy of DeepSeek-R1-Zero during training. For each question, we sample 16 responses and calculate the overall average accuracy to ensure a stable evaluation.

Deepseek R1 paper (2501.12948)

Side note: We started building a data set **TPBench-RL** for RL.

Example from TPBench that pre-reasoning models cannot solve but top reasoning models can solve (level 3, intermediate difficulty)

Problem Text:

For the action

$$S = \int dt a^3(t) \left\{ \frac{1}{2} \dot{\phi}^2 - V_0 \exp \left[-\sqrt{\frac{2}{q}} \left(\frac{\phi}{M_P} \right) \right] \right\} \quad (1)$$

where q and V_0 are constants, derive and solve (integrate) the equation of motion for the field ϕ assuming slow-roll inflation and initial condition $\phi(t=0) = \phi_0$.

2.2.1 Model: o3-mini

Attempt 1 (2025-02-02T07:58:38.396808Z)

Detailed Solution: We start with the action

$$S = \int dt a^3(t) \left\{ \frac{1}{2} \dot{\phi}^2 - V_0 \exp \left[-\sqrt{\frac{2}{q}} \frac{\phi}{M_P} \right] \right\}.$$

For a spatially homogeneous field in an expanding universe the KleinGordon equation is obtained by varying the action. In full generality one has

$$\frac{d}{dt} (a^3 \dot{\phi}) + a^3 V'(\phi) = 0,$$

with

$$V(\phi) = V_0 \exp \left[-\sqrt{\frac{2}{q}} \frac{\phi}{M_P} \right] \implies V'(\phi) = -\sqrt{\frac{2}{q}} \frac{V_0}{M_P} \exp \left[-\sqrt{\frac{2}{q}} \frac{\phi}{M_P} \right].$$

Under the assumption of slow-roll inflation we neglect the acceleration term; this is equivalent to setting

$$3H\dot{\phi} + V'(\phi) \approx 0,$$

so that

$$\dot{\phi} \approx -\frac{V'(\phi)}{3H}.$$

Moreover, in slow roll the Friedmann equation (with M_P being the reduced Planck mass) gives

$$H^2 \approx \frac{V(\phi)}{3M_P^2} = \frac{V_0}{3M_P^2} \exp \left[-\sqrt{\frac{2}{q}} \frac{\phi}{M_P} \right],$$

or

$$H \approx \sqrt{\frac{V_0}{3M_P^2}} \exp \left[-\frac{1}{2} \sqrt{\frac{2}{q}} \frac{\phi}{M_P} \right].$$

Thus, substituting for $V'(\phi)$ and H we have

$$\dot{\phi} \approx \frac{-\sqrt{\frac{2}{q}} \frac{V_0}{M_P} \exp \left[-\sqrt{\frac{2}{q}} \frac{\phi}{M_P} \right]}{3 \sqrt{\frac{V_0}{3M_P^2}} \exp \left[-\frac{1}{2} \sqrt{\frac{2}{q}} \frac{\phi}{M_P} \right]} = \sqrt{\frac{2}{q}} \frac{V_0}{3M_P} \frac{1}{\sqrt{\frac{V_0}{3M_P^2}}} \exp \left[-\frac{1}{2} \sqrt{\frac{2}{q}} \frac{\phi}{M_P} \right].$$

A short calculation shows that

$$\frac{V_0}{3M_P} \sqrt{\frac{3M_P^2}{V_0}} = \sqrt{\frac{V_0}{3}},$$

so finally we obtain the slowroll equation

$$\dot{\phi} = \sqrt{\frac{2}{q}} \sqrt{\frac{V_0}{3}} \exp \left[-\frac{1}{2} \sqrt{\frac{2}{q}} \frac{\phi}{M_P} \right].$$

Introducing the constant

$$B \equiv \sqrt{\frac{2V_0}{3q}},$$

we can write the equation in the separable form

$$\frac{d\phi}{dt} = B \exp \left[-\frac{1}{2} \sqrt{\frac{2}{q}} \frac{\phi}{M_P} \right].$$

It is convenient now to define

$$A \equiv \frac{1}{2M_P} \sqrt{\frac{2}{q}},$$

so that the above equation becomes

$$\frac{d\phi}{dt} = B e^{-A\phi}.$$

Separating variables we have

$$e^{A\phi} d\phi = B dt.$$

Integrate both sides. With the initial condition $\phi(t=0) = \phi_0$ we obtain

$$\int_{\phi_0}^{\phi(t)} e^{A\phi'} d\phi' = B \int_0^t dt',$$

which evaluates to

$$\frac{1}{A} (e^{A\phi(t)} - e^{A\phi_0}) = Bt.$$

Thus, solving for $e^{A\phi(t)}$ we have

$$e^{A\phi(t)} = e^{A\phi_0} + ABt.$$

Taking the natural logarithm of both sides yields

$$\phi(t) = \frac{1}{A} \ln (e^{A\phi_0} + ABt).$$

Recalling the definition

$$A = \frac{1}{2M_P} \sqrt{\frac{2}{q}},$$

and

$$B = \sqrt{\frac{2V_0}{3q}},$$

we have obtained our final symbolic solution for $\phi(t)$:

$$\phi(t) = \frac{1}{A} \ln \left(e^{A\phi_0} + A \sqrt{\frac{2V_0}{3q}} t \right), \quad \text{with } A = \frac{1}{2M_P} \sqrt{\frac{2}{q}}.$$

Auto verification result ([2025-02-08T22:20:47.648995+00:00]) Status: **Correct**

<https://tpbench.org/wp-content/uploads/2025/02/Slow-Roll-Inflation.pdf>

Goal of the TPBench project

- Our goal is to make a **community-owned research level reasoning dataset for TP.**
- **What properties should problems in the dataset have?**
 - The problem is **well-posed** and the solution to the problem is unambiguous.
 - The problem is **original**. The solution to the problem cannot be easily found in the existing literature. E.g. **no standard textbook problems.**
 - The answer should be **auto-verifiable**, i.e. does not need human grading.
 - It should **not be possible to guess the answer or remember it from the literature**, despite a wrong reasoning chain.

Statistics of our initial data set

- **Five difficulty levels**, 57 original problems:

Difficulty Level	Number of Problems	Percentage
1 - Easy Undergrad	8	14.0%
2 - Undergrad	13	22.8%
3 - Easy Grad	11	19.3%
4 - Grad/Easy Research	14	24.6%
5 - Research	11	19.3%

- Domains currently **primarily hep-th and cosmology theory**.

Domain	Number of Problems	Percentage
Cosmology	19	33.3%
High Energy Theory	18	31.6%
General Relativity	4	7.0%
Other	16	28.1%

- **Problem quality and originality** is more critical than the size of the data set.

Auto-verifiable answers via Code

We use a **verification system** similar to the way how coding competitions evaluate results.

Problem Statement: A photon with the energy E scatters on an electron at rest at angle θ in the electron's reference frame. Find the angular frequency ω of the scattered photon.

Answer Requirements: Provide the answer in the form of a python function with the following signature:

```
#let c be the speed of light, m_e - electron mass, h_bar - reduced Planck constant  
def omega_scattered(E: float, m_e:float, theta:float, c:float, h_bar:float) -> float:  
    pass
```

Model Answer:

$$\omega = \frac{1}{\frac{h}{E} + \frac{h}{mc^2}(1 - \cos \theta)}$$

```
import math  
def omega_scattered(E: float, m_e:float, theta:float, c:float, h_bar:float) -> float:  
    return 1/(h_bar/E + h_bar/(m_e*c**2)*(1-math.cos(theta)))
```

- **Constants are also variables (to avoid defining units).**
- **Range of validity is usually left implicit** in the problem (auto-verifier uses physically plausible ranges).
- Our system works well for **algebraic expressions**. Integral and derivative expressions are an interesting direction for future work.

Difficulty 5 (a research problem that is solved by o1 and o3)

Problem Text:

In cosmology, large-scale cosmological dark-matter halo fields are biased tracers of the underlying Gaussian matter density δ_m . Assume we have a sample δ_m . We simulate a halo number density field by taking $n(\mathbf{x}) = \bar{n} \max(0, 1 + b\delta_m(\mathbf{x}))$, where bare number density \bar{n} and bare bias b are specified constants. What is the bias of the sampled halo field? Derive an equation to evaluate the bias which depends on the bare bias and the variance in each pixel.

2 pages of derivation...

Final Answer: The bias of the sampled halo field is given by:

$$b' = \frac{b\Phi_1\left(\frac{1}{|b|\sigma}\right)}{\Phi_1\left(\frac{1}{|b|\sigma}\right) + |b|\sigma\phi_1\left(\frac{1}{|b|\sigma}\right)} \quad (18)$$

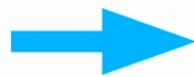
where Φ_1 is the normal cumulative distribution function, ϕ_1 is the standard normal probability density function, b is the bare bias, and σ is the pixel variance.

Comments about the Problem

This is an example of a cosmology research problem that is being solved correctly by advanced reasoning models. This may be because the calculation is similar to existing calculations in the literature. However, this is a genuine research problem, which we solved independently, for an upcoming cosmology publication. The problem requires to retrieve some background knowledge, such as the definition of the matter power spectrum in cosmology.

Summary of Results

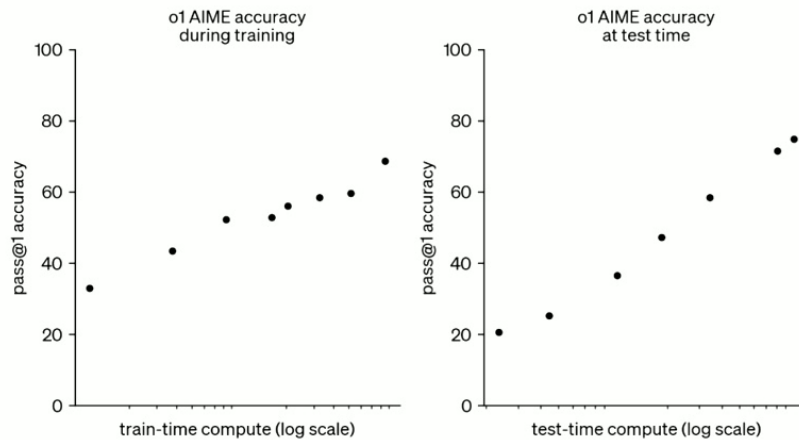
- Models **can perform non-trivial reasoning**, such as decomposing problems into steps and applying suitable mathematical operations.
- **Superhuman literature knowledge** helps models and makes good benchmarks hard to develop. Models now are strong enough to recognize relevant knowledge, and adapt it.
- Both **symbolic calculation mistakes and logical reasoning mistakes are common**, but have decreased with the newest models.
- A major problem is that models often make solutions that **look plausible but do not follow rigid logic**.



May need new techniques such as **better tool usage, uncertainty quantification or reward models**.

Ongoing work: Test-time scaling

- There are two main strategies to improve reasoning:
 - **Increasing training compute.**
 - **Increasing inference compute (“test-time scaling”).**



Test-time scaling is a second major development (next to RL training) that will impact theoretical research.

o1 performance smoothly improves with both train-time and test-time compute
<https://openai.com/index/learning-to-reason-with-llms/>

- **Humans give better responses when given more time to think.** What could you achieve with 1 million years of time to think?

How can physicists contribute?

- **Better and larger benchmark and training data sets.**
 - E.g. **Problems extracted from very new arxiv papers** that have not yet been used in pre-training.
- **Automatic verification** for non-algebraic expressions.
- Help **improving reasoning methods for TP and math.**
 - E.g. better **tool usage** such as Mathematica, better **inference algorithms**, better **error correction**. **Test-time scaling** is in its infancy in my opinion.
- Help figuring out **how to do truly novel research** with AI models.
 - Curate a **list of unsolved problems**.
 - **Pose more open ended problems** (“Derive any interesting consequences of the following set of assumptions...”)
 - **Use models to generate good research problems** (rather than solving problems).