AI Reasoning for Theoretical Physics

Yurii Kvasiuk
UW-Madison

AI4EIC, IAIFI, 29th of October 2025

TPBench https://tpbench.org/

Moritz Münchmeyer - faculty, Physics

Tianyi Li - grad student, Physics

Zhiqi Gao - grad student, CS

Yurii Kvasiuk - grad student, Physics

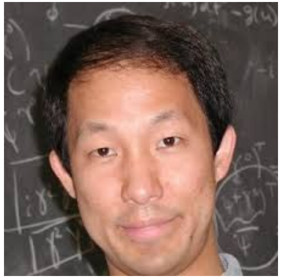
Maja Rudolph - faculty, Statistics

Fred Sala - faculty, CS

Daniel Chung - faculty, Physics

Sai C. Tadepalli - postdoc, Physics

Nate Woodward - grad student, Physics



Outline

- $\boxed{\text{Motivation}}$
- TPBench project
- Test time scaling techniques
- Ongoing and Future Directions

Large Language Models are capable of reasoning

reasoning noun

rea·son·ing ˈrēz-niŋ ˈrē-zᵊn-iŋ

Synonyms of *reasoning* >

1 : the use of reason

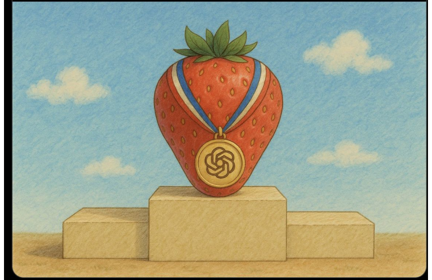
especially **:** the drawing of inferences or conclusions through the use of reason

OpenAI @OpenAI

We achieved gold medal-level performance 🥇 on the 2025 International Mathematical Olympiad with a general-purpose reasoning LLM!

Our model solved world-class math problems—at the level of top human contestants. A major milestone for AI and mathematics.

> **Alexander Wei** @alexwei_ · Jul 19
>
> 1/N I'm excited to share that our latest @OpenAI experimental reasoning LLM has achieved a longstanding grand challenge in AI: gold medal-level performance on the world's most prestigious math competition—the International Math Olympiad (IMO).



11:36 AM · Jul 19, 2025 · **660.3K** Views

RESEARCH

Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad

21 JULY 2025

Thang Luong and Edward Lockhart

Share

> “The fact that the program can come up with a non-obvious construction like this is very impressive, and well beyond what I thought was state of the art.

PROF SIR TIMOTHY GOWERS,
IMO GOLD MEDALIST AND FIELDS MEDAL WINNER

Terence Tao

@tao@mathstodon.xyz

It is tempting to view the capability of current AI technology as a singular quantity: either a given task X is within the ability of current tools, or it is not. However, there is in fact a very wide spread in capability (several orders of magnitude) depending on what resources and assistance gives the tool, and how one reports their results.

Terence Tao
@tao

Jul 19 *

But consider what happens to the difficulty level of the Olympiad if we alter the format in various ways, such as the following:

* One gives the students several days to complete each question, rather than four and half hours for three questions. (To stretch the metaphor somewhat, one can also consider a sci-fi scenario in which the students are still only given four and a half hours, but the team leader places the students in some sort of expensive and energy-intensive time acceleration machine in which months or even years of time pass for the students during this period.)
* Before the exam starts, the team leader rewrites the questions in a format that the students find easier to work with.
* The team leader gives the students unlimited access to calculators, computer algebra packages, formal proof assistants, textbooks, or the ability to search the internet.
* The team leader has the six student team work on the same problem simultaneously, communicating with each other on their partial progress and reported dead ends.
* The team leader gives the students prompts in the direction of favorable approaches, and intervenes if one of the students is spending too much time on a direction that they know to be unlikely to succeed.
* Each of the six students on the team submit solutions to the team leader, who then selects only the "best" solution for each question to submit to the competition, discarding the rest.
* If none of the students on the team obtains a satisfactory solution, the team leader does not submit any solution at all, and silently withdraws from the competition without their participation ever being noted. (2/3)

…however, this is exactly how *research* happens

- Progress happens with time (longer and longer)
- Researchers use tools (in theoretical physics - CAS, code, simulations …)
- Researchers work together and communicate
- Progress often happens from the understanding and synthesis of large amounts of information

It's interesting to evaluate current capabilities of AI in research (so far it has been extensively evaluated mostly on competition math problems)

“FrontierMATH” - AI benchmark of research math

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 Olsson[1], Jean-Stanislas Denain[1], Anson Ho[1], Emily de Oliveira Santos[5‡], Olli Järviniemi, Matthew Barnett[1], Robert Sandler[1], Jaime Sevilla[1], 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‡]

[1]Epoch AI
[2]King’s College London
[3]MIT
[4]University of Siegen
[5]ICMC, USP
[6]UC Berkeley
[7]Cornell University
[8]Harvard University
[9]University of Leicester
[10]Bristol University
†Core contributor
‡Contributing mathematician

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.

SOTA 2% in Nov 2024!

o3 scores 25% few months later

Outline

- Motivation
- $\boxed{\text{TPBench project}}$
- Test time scaling techniques
- Ongoing and Future Directions

TP Bench – Theoretical Physics Benchmark for AI

TPBench is a curated dataset and evaluation suite designed to measure the reasoning capabilities of AI models in theoretical physics. Our test problems span multiple difficulty levels—from undergraduate to frontier research—and cover topics such as cosmology, high-energy theory, general relativity, and more. By providing a unified framework for problem-solving and auto-verifiable answers, TPBench aims to drive progress in AI-based research assistance for theoretical physics.

Table 1. Distribution of problems by difficulty level.

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%

- Uncontaminated
- Novel problems
- Range of difficulty
- Open to community
- Autoverifiable
- Diverse

https://tpbench.org/

We aim to expand to include a broader range of domains

Table 2. Distribution of problems by domain. The 'Other' category includes astrophysics, electromagnetism, quantum mechanics, statistical mechanics, and classical mechanics. Many problems are in between areas. For example some Cosmology problems could also be classified as High Energy 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%

Autoverification pipeline

We formulate the problems such that the answers can be given by an executable Python function (inspired by coding competition).

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:

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

```
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)))
```

Limitations (and ongoing efforts to overcome them)

Currently, only the algebraic expressions are supported, however we also need

- **Tensors**
- **Abstract Operators**
- **Integrals, differential forms**
- **manifolds**

$$\nabla_\mu T^{\mu\nu} = 0$$

$$[A, B] = AB - BA$$

$$\int_{\mathcal{M}} d\omega = \int_{\partial M} \omega$$

Example: Difficulty 5 (a research problem that is now solved)

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:

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

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.

Currently solved by SOTA models, but that wasn't the case a year ago!

2.1 Expert Solution

Detailed Steps: Detailed Steps: The solution to this question involves some domain knowledge, parts of which were given in the problem's statement, some approximations sourced by the domain knowledge, and some mathematical calculations. The domain knowledge is very basic and should be known to anyone in the field. Approximations are intuitive and also, mostly, inspired by the domain knowledge. Following Polya, we can organize it as follows:

Understand the problem. The number density of halos $n_h(\mathbf{x})$ is defined as

$$N_h = \int_V n_h(\mathbf{x})d\mathbf{x}. \tag{1}$$

The overdensity is defined as

$$\delta_h(\mathbf{x}) = \frac{n_h(\mathbf{x}) - \langle n_h(\mathbf{x})\rangle}{\langle n_h(\mathbf{x})\rangle}. \tag{2}$$

Linear bias is defined in terms of Fourier-transformed quantities:

$$\delta_h(\mathbf{k}) = b\delta_m(\mathbf{k}). \tag{3}$$

This is an approximation that holds on sufficiently large scales (small k). $\delta_m(\mathbf{k})$ and $\delta_h(\mathbf{k})$ are Gaussian random fields with zero mean and their variance depends only on the magnitude of the wave-vector $k = |\mathbf{k}|$:

$$\delta_m \sim \mathcal{N}(0, P_{mm}(k)),\ \delta_h \sim \mathcal{N}(0, P_{hh}(k)). \tag{4}$$

The quantity $P(k)$ is called the power spectrum and is defined as

$$\langle\delta(\mathbf{k})\delta(\mathbf{k}')\rangle = (2\pi)^3\delta^D(\mathbf{k} + \mathbf{k}')P(k). \tag{5}$$

It immediately follows that

$$P_{hh}(k) = b^2P_{mm}(k). \tag{6}$$

We are given the expression in real space. In real space, the quantity $\delta_m(\mathbf{x})$ is also a Gaussian random field:

$$\delta_m(\mathbf{x}) \sim \mathcal{N}(0, \xi_m),\ \delta_h(\mathbf{x}) \sim \mathcal{N}(0, \xi_h). \tag{7}$$

Quantity ξ is called a two-point (real-space) correlation function and is defined as

$$\langle\delta(\mathbf{x})\delta(\mathbf{x}')\rangle = \xi(|\mathbf{x} - \mathbf{x}'|). \tag{8}$$

This quantity is sufficiently small when $|\mathbf{x} - \mathbf{x}'| \gg 1$. We are asked to find what is the expression for b in the equation $\delta_h(k) = b\delta_m(\mathbf{k})$, given the real-space expression for the number density $n_h(\mathbf{x})$ in terms of real-space sample of $\delta_m(\mathbf{x})$.

Devise a plan. The key point to solve this problem should be that real-space correlation function for halos ξ_h should also be equal to $b^2\xi_m$. We want to calculate that correlation function. It should be expressed in terms of $\langle n(\mathbf{x})\rangle$ and $\langle n_h(\mathbf{x})n_h(\mathbf{x}')\rangle$. We expect to be able to calculate these expectations since they are the expectations of functions of the Gaussian random variables. We are given the pixel variance σ. How does it connect to the other quantities we know? In principle, that's also the part of domain knowledge but it also can be deducted from the definitions already given. A discretized version of the correlation function is

$$\xi_{ij} = \langle\delta_{\mathbf{x_i}}\delta_{\mathbf{x_j}}\rangle. \tag{9}$$

When $i = j$, it becomes the pixel variance σ. *Aside, we could have given instead of σ, the quantity $P_{mm}(k)$, that is a common description of a cosmological dark-matter field. In that case, from the definitions of $\xi(r)$ and $P_{mm}(k)$, we could have deduced that $\sigma = \frac{1}{V}\sum_k P_{mm}(k)$.* Then we pick the ensemble of all the pixels at given fixed large distance $r = |\mathbf{x_i} - \mathbf{x_j}|$. The key is to recognize that it is fully described by a correlated bivariate Gaussian distribution.

$$(\delta_i^m, \delta_j^m) \sim \mathcal{N}(0, \Sigma) \tag{10}$$

with a covariance

$$\Sigma = \begin{pmatrix} \sigma^2 & \xi_r^m \\ \xi_r^m & \sigma^2 \end{pmatrix}. \tag{11}$$

In general, the integrals from the expectation values are cumbersome, but we should expect some simplifications from the fact that ξ is small and we can Taylor-expand the pdf.

Carry out the plan. It's more convenient to define $\hat{\delta}_i = \delta_i^m/\sigma$ and $\hat{\xi} = \xi_r^m/\sigma^2$, and ϕ_2 - a correlated bivariate Gaussian pdf - then

$$(\hat{\delta}_i, \hat{\delta}_j) \sim \frac{e^{-\frac{1}{2(1-\hat{\xi}^2)}[\hat{\delta}_i^2 + \hat{\delta}_j^2 - 2\hat{\xi}\hat{\delta}_i\hat{\delta}_j]}}{2\pi\sqrt{1 - \hat{\xi}^2}} \equiv \phi_2(\hat{\delta}_i, \hat{\delta}_j|\hat{\xi}). \tag{12}$$

We note that

$$\xi_r^n = \frac{\langle n_in_j\rangle}{\langle n\rangle^2} - 1. \tag{13}$$

The quantity $\langle n\rangle$ is the actual mean number density:

$$\bar{n}' = \langle n\rangle = \langle n_i\rangle = \int n^{loc}(\delta_i, b, \bar{n})\phi_2(\hat{\delta}_i, \hat{\delta}_j|\hat{\xi})d\hat{\delta}_id\hat{\delta}_j = \int n_i^{loc}\phi_1(\hat{\delta}_i)d\hat{\delta}_i.$$

Here, ϕ_1 - is a standard normal pdf. It is expected that it's not dependent on the correlation $\hat{\xi}$, but only on b and σ, just as the marginal of 2D correlated Gaussian distribution is 1D Gaussian that's not dependent on the cross-correlation. To the linear order in $\hat{\xi}$,

$$\phi_2(x, y|\hat{\xi}) \approx \phi_1(x)\phi_1(y)(1 + \hat{\xi}xy). \tag{14}$$

So that the two-point function neatly factorizes:

$$\begin{aligned}\langle n_in_j\rangle &= \int n^{loc}(\delta_i, b, \bar{n})n^{loc}(\delta_j, b, \bar{n})\phi_2(\hat{\delta}_i, \hat{\delta}_j|\hat{\xi})d\hat{\delta}_id\hat{\delta}_j \\ &\approx \int n_i^{loc}\phi_1(\hat{\delta}_i)d\hat{\delta}_i \int n_j^{loc}\phi_1(\hat{\delta}_j)d\hat{\delta}_j + \hat{\xi}\int n_i^{loc}\phi_1(\hat{\delta}_i)\hat{\delta}_id\hat{\delta}_i \int n_j^{loc}\phi_1(\hat{\delta}_j)\hat{\delta}_jd\hat{\delta}_j \\ &\equiv \langle n\rangle^2 + \hat{\xi}\langle n\hat{\delta}\rangle^2.\end{aligned}$$

Substituting the results for $\langle n\rangle$ and $\langle n_in_j\rangle$ in the equation for ξ_r^n, we can read off the bias:

$$b'^2 = \frac{\xi_r^n}{\sigma^2\hat{\xi}} = \frac{\langle n\hat{\delta}\rangle^2}{\sigma^2\langle n\rangle^2}. \tag{15}$$

All that is left is to calculate the expectations. One can evaluate for $b \geq 0$

$$\begin{aligned}\langle n\rangle &= \int n_i^{loc}\phi_1(\hat{\delta}_i)d\hat{\delta}_i = \int \bar{n}\max(0, 1 + b\sigma x)\phi_1(x)dx \\ &= \bar{n}\int_{-\frac{1}{b\sigma}}^{+\infty}(1 + b\sigma x)\phi_1(x)dx = \bar{n}\left[\Phi_1\left(\frac{1}{b\sigma}\right) + b\sigma\phi_1\left(\frac{1}{b\sigma}\right)\right].\end{aligned}$$

For $b < 0$ it's, however,

$$\begin{aligned}\langle n\rangle &= \bar{n}\int_{-\infty}^{+\frac{1}{|b|\sigma}}(1 - |b|\sigma x)\phi_1(x)dx \\ &= \bar{n}\left[\Phi_1\left(\frac{1}{|b|\sigma}\right) + |b|\sigma\phi_1\left(\frac{1}{|b|\sigma}\right)\right].\end{aligned}$$

So we conclude that the latter expression is valid for all b. Similarly, one can show that

$$\langle n\hat{\delta}\rangle = \bar{n}\int \max(0, 1 + b\sigma x)x\phi_1(x)dx = \bar{n}b\sigma\Phi_1\left(\frac{1}{|b|\sigma}\right) \tag{16}$$

Example: Difficulty 5 (currently unsolved by any model)

Problem Text:
Consider the conformally coupled scalar field ϕ

$$\mathcal{L} = \frac{1}{2}\left[g^{\mu\nu}\partial_\mu\phi\partial_\nu\phi - \left(m^2 - \frac{1}{6}R\right)\phi^2\right] \tag{1}$$

in curved spacetime

$$ds^2 = a^2(\eta)\left(d\eta^2 - |d\vec{x}|^2\right)$$

where the Ricci scalar is

$$R = -6\frac{a''(\eta)}{a(\eta)} \tag{2}$$

and a satisfies the differential equation

$$\frac{d}{dt}\ln a = \Theta(t_e - t)H_I + \Theta(t - t_e)\frac{H_I}{1 + \frac{3}{2}H_I(t - t_e)} \tag{3}$$

with t_e a finite positive number, the Θ function having the steplike behavior

$$\Theta(t - t_e) \equiv \begin{cases} 1 & t \geq t_e \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

and t being the comoving proper time related to η through

$$t = t_e + \int_{\eta_e}^{\eta} a(y)dy. \tag{5}$$

The boundary condition for the differential equation (in comoving proper time) is $a|_{t=t_e} = a_e$.
In the limit that $k/(a_e H_I) \to \infty$, using the steepest descent approximation starting from the dominant pole $\tilde{\eta}$ (with $\Re\tilde{\eta} > 0$) of the integrand factor $\omega_k'(\eta)/(2\omega_k(\eta))$, compute the Bogoliubov coefficient magnitude $|\beta(k)|$ approximated as

$$|\beta(k)| \approx \left|\int_{-\infty}^{\infty} d\eta \frac{\omega_k'(\eta)}{2\omega_k(\eta)} e^{-2i\int_{\eta_e}^{\eta} d\eta' \omega_k(\eta')}\right| \tag{6}$$

for particle production where the dispersion relationship given by

$$\omega_k^2(\eta) = k^2 + m^2 a^2(\eta) \tag{7}$$

with $0 < m \lesssim H_I$. Use a one pole approximation which dominates in this limit.

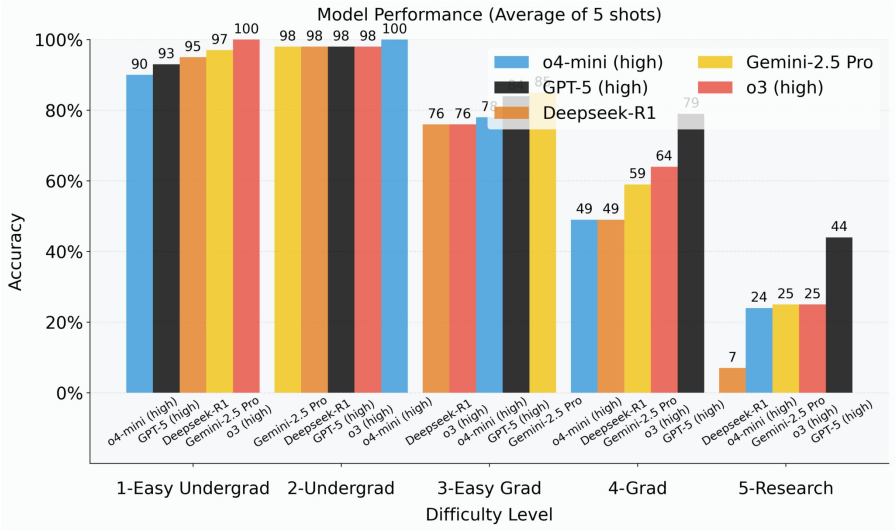
3 pages of derivation…

Final Answer:

$$\boxed{|\beta| \approx \frac{\pi}{3}\exp\left(-\frac{4}{3}\sqrt{2\pi}\frac{\Gamma(5/4)}{\Gamma(3/4)}\frac{(k/a_e)^{3/2}}{H_I\sqrt{m}}\right)}. \tag{34}$$

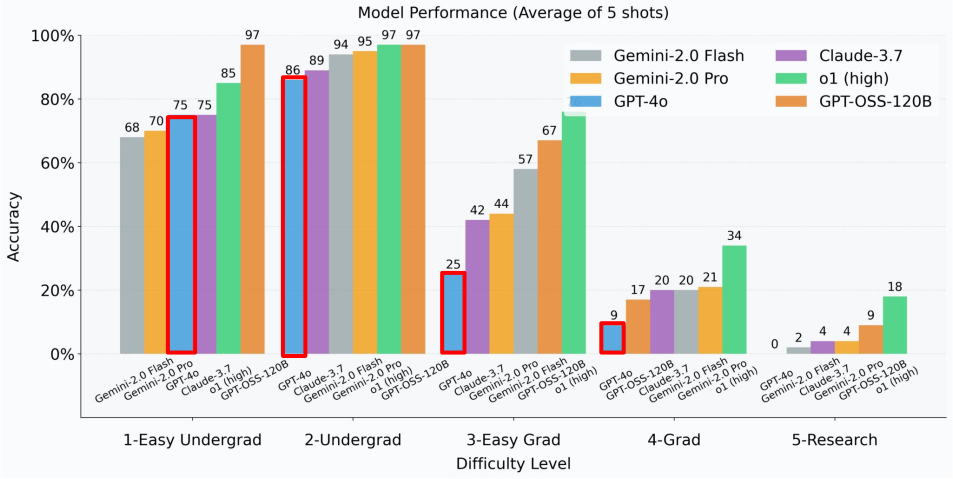
Comments about the Problem

This is an example of a difficult problem from Quantum Field Theory in curved spacetime, dealing with gravitational particle production, that appears out of reach of current models. This is part of a published research work and the solution, without steps explained, is given in a footnote of [30], but would be difficult to locate (in fact we tried, without success, with OpenAI's Deep Research).



RL finetuning drastically increases model's abilities to solve problems!

Get updates at tpbench.org



Summary of Current State

- 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.

- Both **symbolic calculation mistakes and logical reasoning mistakes are common**, but have decreased with the newest models. **See error analysis in our paper.**

- A **major problem** is that models often make solutions that **look plausible but do not follow rigid logic**.

→ Need new techniques such as **better tool usage, uncertainty quantification or different reward models**.

Contact us if you want to collaborate / contribute

- **Ten of our problems are public** on our website. Check them out!
- You sign our **“Usage Agreement” to get the entire problem set.**
  - You acknowledge and agree to **safety procedures** to keep it out of training data.
  - You **submit at least one new original problem** that we can add to the next version of TPBench.
- Anybody who submits a useable problem will be offered **co-authorship of the next TPBench update paper**.

TPBench Dataset Usage Agreement

TPBench is a private data set. It can be used for evaluation only; you may not train models on the dataset. To avoid leakage into pre-training data, the TPBench team monitors the distribution of the data set. The PI of a research group that will be given access to the private TPBench github repository must ensure that the dataset remains private and is only shared with the group members listed below. The PI further ensures that the data set will not be made public by accident (e.g. upload to a public github repository, public web storage etc) and that problems and solutions are never passed into web interfaces of LLMs that are the basis for future training data (e.g. chatGPT user interface). The PI will also inform the TPBench team in case any mistakes are found in the data set.

PI of the research group (Name, Affiliation, Email):

The dataset will be made accessible to the following group members:

Name	Role (e.g. graduate student)	Email

A larger data set of diverse theoretical physics problems would be very useful for the research community. In return for data set access, for physics research groups, the TPBench team thus requests the submission of at least one original hard problem (level 4 or 5) or three original easy problems (level 1 to 3) per research group, within three months after the dataset was shared. Suitable problems will be added to a future version of TPBench. The problems can come from any domain of theoretical physics. The author of each accepted problem will be offered co-authorship of an upcoming TPBench update paper (generally one author per problem).

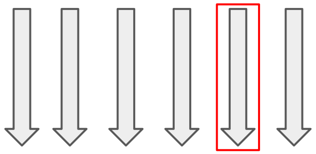
Date Signature of the PI

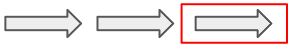
Outline

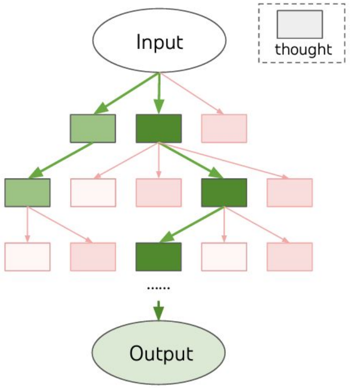
- Motivation
- TPBench project
- $\boxed{\text{Test time scaling techniques}}$
- Ongoing and Future Directions

Test-time scaling - a set of techniques to increase performance at inference time

- Parallel - generate many solutions then choose the best one according to “some rules”
- Sequential - work on the same solution during a prolonged period of time
- Search and Agentic methods (e.g. solution tree, ensemble of llms working together)



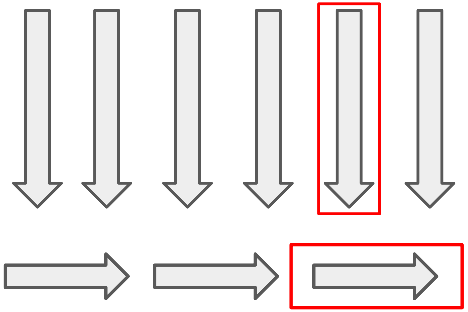




“Tree of thought” 2305.10601

Test-time scaling - a set of techniques to increase performance at inference time

- Parallel - generate many solutions then choose the best one according to “some rules”
- Sequential - work on the same solution during a prolonged period of time



Test-time Scaling Techniques in Theoretical Physics - A Comparison of Methods on the TPBench Dataset

Zhiqi Gao [*1] **Tianyi Li** [*2] **Yurii Kvasiuk** [2] **Sai Chaitanya Tadepalli** [3] **Maja Rudolph** [4] **Daniel J.H. Chung** [2]
Frederic Sala [1] **Moritz Münchmeyer** [2]

https://arxiv.org/abs/2506.20729 (accepted at NeurIPS ML4PS 2025)

Parallel scaling

Generate N responses to the same question and pick 1 by

- Choosing the most common answer (majority vote)
- Asking the LLM to verify the solution (weak verifier)
- Verifying steps with a SymPy agent (symbolic weak verifier)

The upper bound of all the approaches is **best of N** (i.e. TTS can only find the correct solution given it is present in the sample pool)

Results

Table 1. Comparison of test-time scaling approaches on TPBench With Gemini 2.5 Pro.

Method	Level 4	Level 5
Single Attempt	63.3%	29.3%
Sequential Methods		
1+Round Reasoning	65.0%	26.4%
2+Round Reasoning	65.0%	26.4%
4+Round Reasoning	68.6%	28.2%
Parallel Methods		
Simple Weak Verifier	71.4%	27.3%
Majority Vote	78.6%	36.4%
SymPy Verifier	71.4%	54.5%
Best of N	85.7%	63.6%

Note: Single attempt's accuracy is averaged over 50 attempts; multi-round reasoning over 10 attempts; parallel methods use 50 candidates.

Table 2. Comparison of test-time scaling approaches on TPBench With Gemini 2.0 Flash

Method	Level 3	Level 4	Level 5
Baseline			
Single Attempt	52.5%	13.1%	1.5%
Sequential Methods			
1+Round	61.8%	16.4%	2.7%
2+Round	60.0%	22.9%	0.9%
4+Round	60.0%	21.4%	1.8%
Parallel Methods			
Simple Weak Verifier	18.2%	7.1%	0%
Majority Vote	72.7%	35.7%	9.1%
SymPy Verifier	81.8%	57.1%	9.1%
Best of N	90.9%	85.7%	18.2%

Example

The solution then evaluates the first moment integral $I_1 = \int XP(X)dX$ over the truncated domain:

$$I_1 = \int_{-1/b}^{\infty} XP(X)dX = \frac{\sigma^2}{\sqrt{2\pi}} e^{-1/(2b^2\sigma^2)} \quad \text{(ERROR HERE)} \tag{3}$$

Combining these gives the candidate's expression for the mean halo density:

$$\bar{n}_h = \bar{n}\left[\frac{1}{2}\text{erfc}\left(-\frac{1}{b\sigma\sqrt{2}}\right) + b\frac{\sigma^2}{\sqrt{2\pi}} e^{-1/(2b^2\sigma^2)}\right] \tag{4}$$

The effective bias is formulated using the peak-background split argument, which defines b_{eff} as the response of the mean halo number density to a long-wavelength perturbation δ_L:

$$b_{\text{eff}} = \frac{1}{\bar{n}_h} \frac{d\langle n|\delta_L\rangle}{d\delta_L}\Big|_{\delta_L=0} \tag{5}$$

The derivative term is shown to be equivalent to:

$$\frac{d\langle n|\delta_L\rangle}{d\delta_L}\Big|_{\delta_L=0} = \int_{-1/b}^{\infty} \bar{n}(1+bX)\left(P(X)\frac{X}{\sigma^2}\right)dX = \frac{\bar{n}}{\sigma^2}\int_{-1/b}^{\infty}(X+bX^2)P(X)dX \tag{6}$$

To evaluate the above, the solution requires the second moment integral $I_2 = \int X^2P(X)dX$ over the truncated domain, for which it claims:

$$I_2 = \int_{-1/b}^{\infty} X^2P(X)dX = \frac{-\sigma^2/b}{\sqrt{2\pi}} e^{-1/(2b^2\sigma^2)} + \sigma^3\frac{1}{2}\text{erfc}\left(\frac{-1}{b\sigma\sqrt{2}}\right) \quad \text{(ERROR HERE)} \tag{7}$$

Combining the (incorrect) intermediate moment calculations I_1 and I_2, the candidate assembles the numerator for the b_{eff} expression:

$$D(\sigma, b) = \frac{\bar{n}}{\sigma^2}\left[I_1 + bI_2\right] = \bar{n}\left[\frac{1+b\nu}{\sqrt{2\pi}} e^{-\nu^2/(2\sigma^2)} + \frac{b\sigma}{2}\text{erfc}\left(\frac{\nu}{\sigma\sqrt{2}}\right)\right] \tag{8}$$

Verifier Limitations

Mathematical Operation	Verifiable?
Polynomial & Rational Function Algebra	Yes
Standard Differentiation & Integration	Yes
Residue Calculation at Poles	Yes
Approximations & Limit-taking	Partially
General Tensor Manipulations (GR)	No
Advanced Path Integrals (QFT)	No

The efforts to expand the domain of verifier are ongoing!

Ongoing and Future Directions

- Expanding the benchmark
- Exploring better test-time scaling methods
- Improving tool usage
- RL finetuning and process reward
- Agentic reasoning

Thanks!