Artificial Intelligence and LLMs in theoretical physics research

María-José Guzmán Monsalve
Sotirios Karamitsos

Laboratory of Theoretical Physics,
Institute of Physics, University of Tartu

July 3, 2025
GeomGravX

Introduction

Interest in AI's role in theoretical physics has sparked with several workshops

- Theory + AI Workshop: Theoretical Physics for AI Apr 9-11, 2025
- AI Reasoning in Theoretical Physics, May 24-June 14, 2026
- EuCAPT Rapid Response Workshop on AI Agents in Theoretical Physics Research, Jun 10, 2026

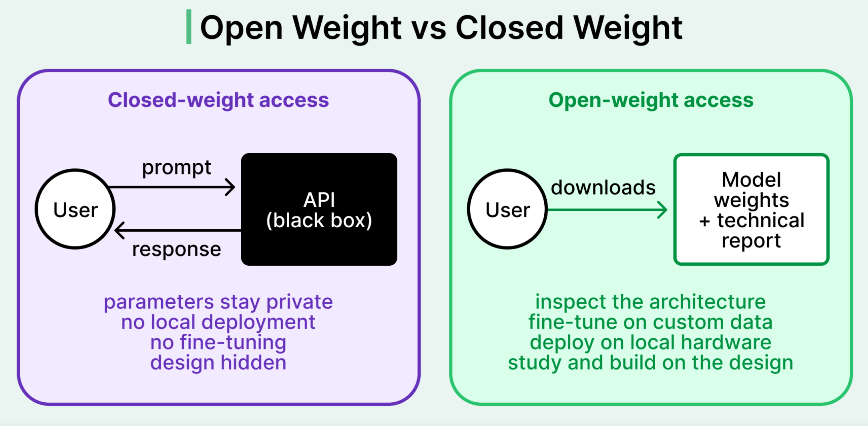
and talks/discussions

- Training Sand to Think : Artificial General Intelligence and the Future of Science | Adam Brown
- Physics Panel - Accelerating Math and Theoretical Physics with AI - IPAM at UCLA

Concepts/techniques

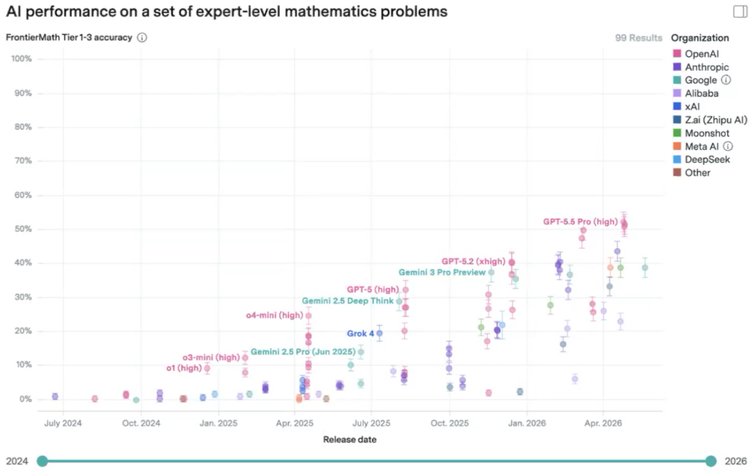
- Reasoning/test-time scaling: Instead of answering immediately, the model spends extra inference-time effort exploring intermediate steps and candidate solution paths.
- RAG (Retrieval-Augmented Generation): Access to papers and notes makes AI far more reliable than only training data.
- CoT (Chain-of-thought prompting): Encourages explicit multi-step reasoning, which helps with long derivations and structured theoretical arguments.
- Code execution/symbolic computation/verification

Open vs closed weight



https://blog.bytebytego.com/p/how-open-weight-models-changed-the

FrontierMath now



https://epoch.ai/frontiermath

Challenges in AI in theoretical physics compared to mathematics

AI reasoning in theoretical physics (TP) is different from AI reasoning in mathematics because the two fields ask for different kinds of correctness and use different styles of reasoning.

- Mathematics seeks exact proof; theoretical physics often seeks useful models.
- Physics relies more on direct computation rather than theorem-proof structures.
- Many physics results are valid only in certain regimes, such as weak fields, small couplings, low energies, or leading-order expansions.
- Physical meaning matters, not only algebra. An AI must know whether an approximation, limit, convention, or effective description is appropriate.
- Physics is harder to verify automatically. A mathematical proof can often be checked line by line, but a physics answer may depend on hidden assumptions, domains of validity, and physical judgment.

arXiv:2502.15815 [cs.LG], Price, Dominic Thomas (2023) Routines and Applications of Symbolic Algebra Software. Doctoral thesis, Durham University.

TPBench overview

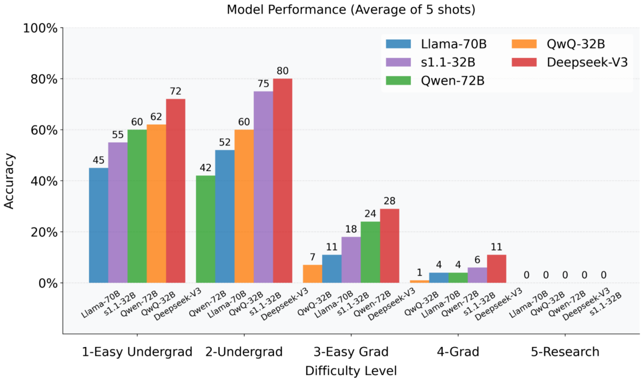
- A good **community accessible benchmark is a foundation for new work in this area.**
- **Spectrum of difficulty** to experiment with weaker models.
- Problems are **original**. E.g. **no standard textbook problems**.
- The answers are **auto-verifiable**, i.e. no need for human grading.
- It should **not be possible to guess the answer** despite a wrong reasoning chain.

Content of Version 1 of the Benchmark

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%

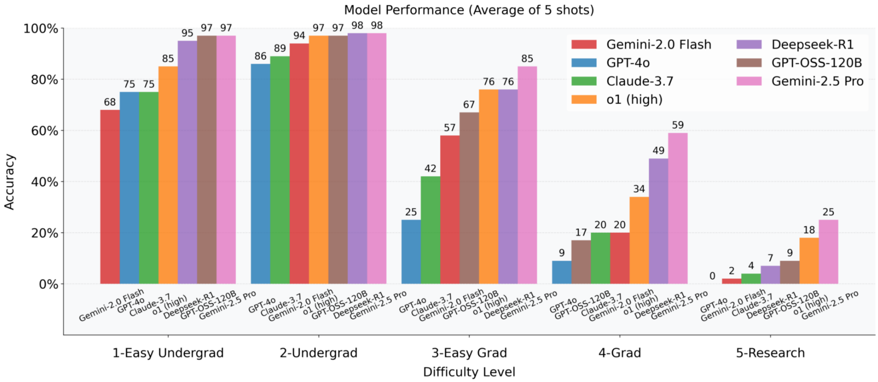
Domain	Number of Problems	Percentage
Cosmology	19	33.3%
High Energy Theory	18	31.6%
General Relativity	4	7.0%
Other	16	28.1%

Source: Moritz Muenchmeyer, EuCAPT Rapid Response Workshop on AI Agents in Theoretical Physics Research, Jun 10, 2026, and arXiv:2502.15815 [cs.LG]



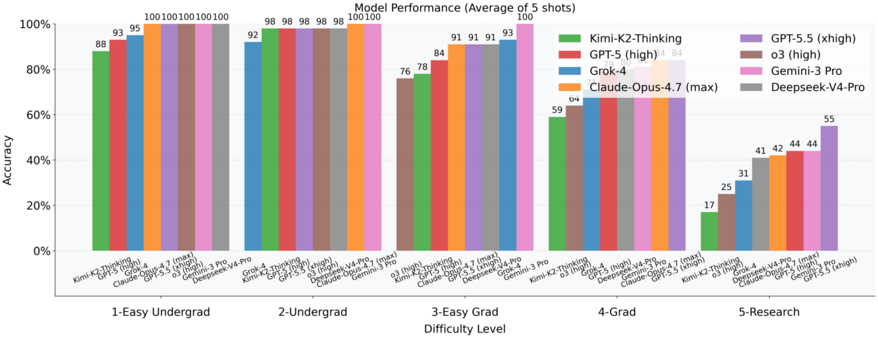
Company / organization	Model	Release date
Alibaba Cloud / Qwen	Qwen-72B	30 Nov 2023
	QwQ-32B	Mar 2025
Meta AI / Meta	Llama-70B	18 Apr 2024
DeepSeek	Deepseek-V3	26 Dec 2024
simplescaling research project	s1.1-32B	Feb 2025

LLM Performance (Aug 2025)



Company	Model	Release date
OpenAI	GPT-4o	13 May 2024
	o1 (high)	5 Dec 2024
	GPT-OSS-120B	05 Aug 2025
Google DeepMind / Google	Gemini-2.0 Flash	11 Dec 2024
	Gemini-2.5 Pro	25 Mar 2025
DeepSeek	Deepseek-R1	20 Jan 2025
Anthropic	Claude-3.7	24 Feb 2025

LLM Performance (April 2026)



Company	Model	Release date
OpenAI	o3 (high)	16 Apr 2025
	GPT-5 (high)	7 Aug 2025
	GPT-5.5 (xhigh)	23 Apr 2026
xAI	Grok-4	9 Jul 2025
Moonshot AI	Kimi-K2-Thinking	06 Nov 2025
Google DeepMind / Google	Gemini-3 Pro	18 Nov 2025
Anthropic	Claude-Opus-4.7 (max)	16 Apr 2026
DeepSeek	Deepseek-V4-Pro	24 Apr 2026

We've come a long way...

For a scalar-tensor theory without matter, what are the transformation rules for the non-minimal coupling, the kinetic prefactor, and the potential under conformal transformations?

gpt-3.5-turbo

In a scalar-tensor theory without matter, the action can be written as

$$S = \int d^4x\sqrt{-g}\left[\frac{1}{2}F(\phi)R - \frac{g^{\mu\nu}}{2}G(\phi)\partial_\mu\phi\partial_\nu\phi - V(\phi)\right],$$

where $F(\phi)$ is the non-minimal coupling, $G(\phi)$ is the kinetic prefactor, and $V(\phi)$ is the potential for the scalar field ϕ. Under a conformal transformation of the metric

$$\tilde{g}_{\mu\nu} = \Omega^2 g_{\mu\nu},$$

the transformation rules for the non-minimal coupling $F(\phi)$, the kinetic prefactor $G(\phi)$, and the potential $V(\phi)$ are as follows

$$\tilde{F}(\phi) = \frac{F(\phi)}{\Omega^2}.$$
$$\tilde{G}(\phi) = \Omega^2 G(\phi).$$
$$\tilde{V}(\phi) = V(\phi).$$

These transformation rules ensure that the action is invariant under conformal transformations, and that the field equations derived from the action are also conformally invariant.

gpt-5.5-instant

For a 4D scalar–tensor action without matter, e.g.

$$S = \int d^4x\sqrt{-g}\left[\frac{1}{2}F(\phi)R - \frac{1}{2}K(\phi)(\nabla\phi)^2 - U(\phi)\right],$$

take the conformal/scalar transformation

$$g_{\mu\nu} = e^{2\gamma(\bar{\phi})}\bar{g}_{\mu\nu}, \qquad \phi = f(\bar{\phi}).$$

Then, after dropping the boundary term from $\Box\gamma$, the functions transform as

$$\bar{F}(\bar{\phi}) = e^{2\gamma}F(f)$$
$$\bar{K}(\bar{\phi}) = e^{2\gamma}\left[(f')^2K(f) - 6F(f)(\gamma')^2 - 6\gamma' f' F_{,\phi}(f)\right]$$
$$\bar{U}(\bar{\phi}) = e^{4\gamma}U(f)$$

where primes denote derivatives with respect to $\bar{\phi}$. If there is no scalar field redefinition, $f(\bar{\phi}) = \bar{\phi}$, and $e^{\gamma} = \Omega$, then

$$\bar{F} = \Omega^2 F,$$
$$\bar{K} = \Omega^2\left[K - 6F(\partial_\phi \ln\Omega)^2 - 6F_{,\phi}\partial_\phi \ln\Omega\right],$$
$$\bar{U} = \Omega^4 U.$$

We've come a long way...

For a scalar-tensor theory without matter, what are the transformation rules for the non-minimal coupling, the kinetic prefactor, and the potential under conformal transformations?

gpt-3.5-turbo

In a scalar-tensor theory without matter, the action can be written as

$$S = \int d^4x\sqrt{-g}\left[\frac{1}{2}F(\phi)R - \frac{g^{\mu\nu}}{2}G(\phi)\partial_\mu\phi\partial_\nu\phi - V(\phi)\right],$$

where $F(\phi)$ is the non-minimal coupling, $G(\phi)$ is the kinetic prefactor, and $V(\phi)$ is the potential for the scalar field ϕ. Under a conformal transformation of the metric

$$\tilde{g}_{\mu\nu} = \Omega^2 g_{\mu\nu},$$

the transformation rules for the non-minimal coupling $F(\phi)$, the kinetic prefactor $G(\phi)$, and the potential $V(\phi)$ are as follows

$$\tilde{F}(\phi) = \cancel{\frac{F(\phi)}{\Omega^2}}.$$

$$\tilde{G}(\phi) = \cancel{\Omega^2 G(\phi)}.$$

$$\tilde{V}(\phi) = \cancel{V(\phi)}.$$

These transformation rules ensure that the action is invariant under conformal transformations, and that the field equations derived from the action are also conformally invariant.

gpt-5.5-instant

For a 4D scalar–tensor action without matter, e.g.

$$S = \int d^4x\sqrt{-g}\left[\frac{1}{2}F(\phi)R - \frac{1}{2}K(\phi)(\nabla\phi)^2 - U(\phi)\right],$$

take the conformal/scalar transformation

$$g_{\mu\nu} = e^{2\gamma(\bar{\phi})}\bar{g}_{\mu\nu}, \qquad \phi = f(\bar{\phi}).$$

Then, after dropping the boundary term from $\Box\gamma$, the functions transform as

$$\bar{F}(\bar{\phi}) = e^{2\gamma}F(f)\checkmark$$

$$\bar{K}(\bar{\phi}) = e^{2\gamma}\left[(f')^2K(f) - 6F(f)(\gamma')^2 - 6\gamma' f' F_{,\phi}(f)\right]\checkmark$$

$$\bar{U}(\bar{\phi}) = e^{4\gamma}U(f)\checkmark$$

where primes denote derivatives with respect to $\bar{\phi}$. If there is no scalar field redefinition, $f(\bar{\phi}) = \bar{\phi}$, and $e^{\gamma} = \Omega$, then

$$\bar{F} = \Omega^2 F,\ \checkmark$$

$$\bar{K} = \Omega^2\left[K - 6F(\partial_\phi \ln\Omega)^2 - 6F_{,\phi}\partial_\phi\ln\Omega\right],\ \checkmark$$

$$\bar{U} = \Omega^4 U.\checkmark$$

The GeomGravX survey on AI uses and perspectives in theoretical physics *(or GSoAIUPiTP for short)*

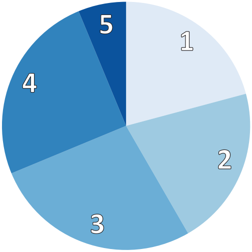
- Objectively things are improving, but how do we *feel?*
- 48 responses; $N > 100$ is (supposedly) needed for quantitative analysis...

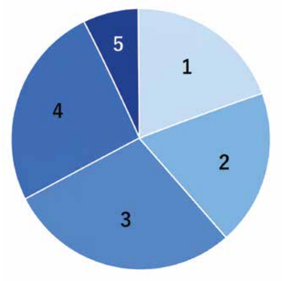
The GeomGravX survey on AI uses and perspectives in theoretical physics *(or GSoAIUPiTP for short)*

- Objectively things are improving, but how do we *feel?*
- 48 responses; $N > 100$ is (supposedly) needed for quantitative analysis...
- ...but we'll make do with qualitative for once
- We can still see some interesting trends (and they can be good starting points for discussion too)
- Quick facts:
  - 6 bachelor's/master's students
  - 15 PhD students
  - 15 postdocs
  - 7 members of faculty
- ...plus 3 "Other" (incl. visiting researchers and independent researchers)

Generative AI usage

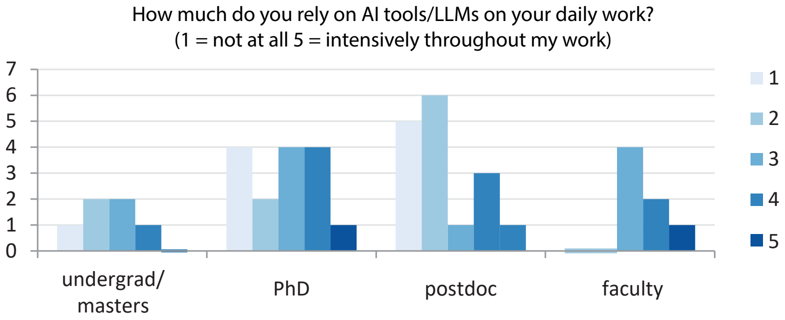
How much do you rely on AI tools/LLMs on your daily work?
(1 = not at all 5 = intensively throughout my work)





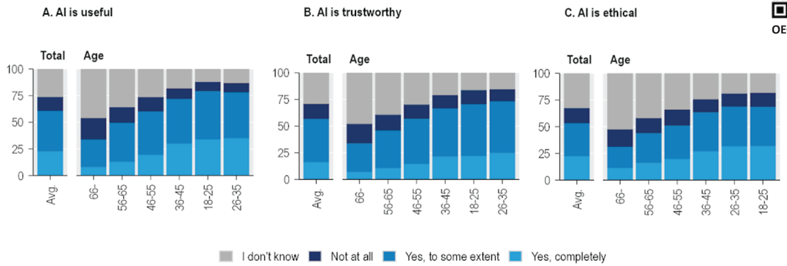
- Pretty even mix (save for the AI whales)
- Let’s split it per career stage

Generative AI usage per career stage



- Pretty good mix across the board (save for the polarized postdocs)...
- ...but note that faculty has no 1s or 2s

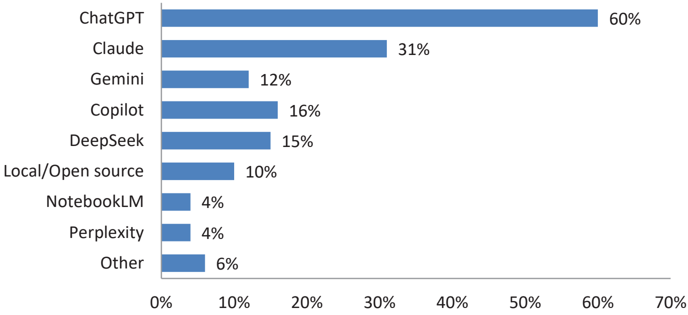
Generative AI attitudes by age (general population)



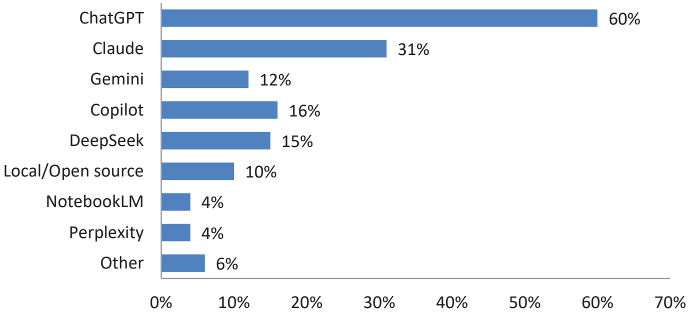
OECD report

- Definitely not enough data to run a Student's t-test, but at first blush, seems to run contrary to the pattern in the overall population (OECD report from early 2025)
- Statistical noise or is this an actual trend in academia?

Generative AI tool/services popularity

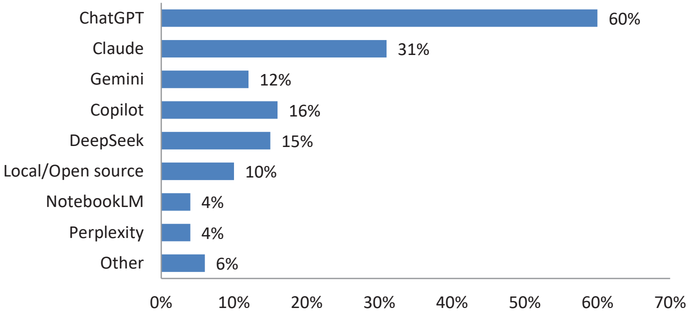


Generative AI tool/services popularity



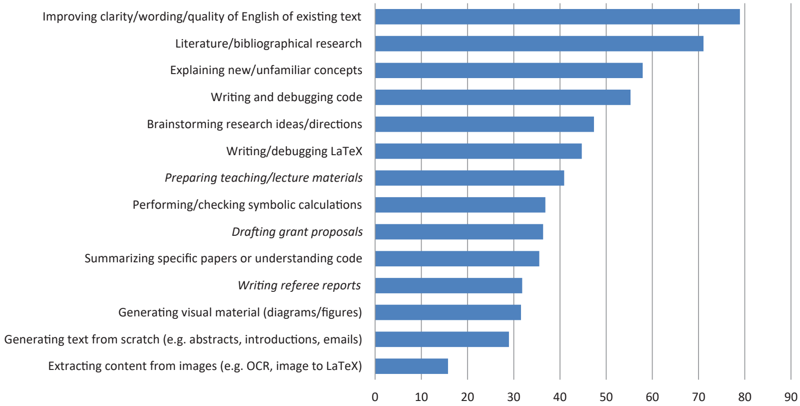
- Of the people that did not answer "Not at all" to how much they use AI, we find some clear winners
- Including Grammarly (proofreading), Scispace (literature review), other open-source models (e.g. Llama and GLM 5.2),

Generative AI tool/services popularity



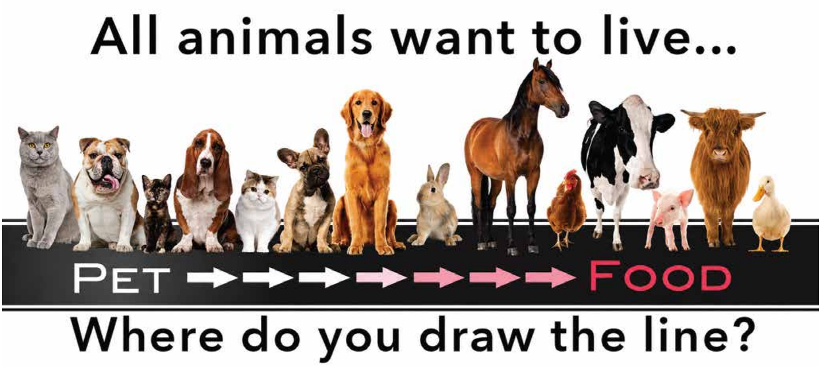
- Of the people that did not answer "Not at all" to how much they use AI, we find some clear winners
- Including Grammarly (proofreading), Scispace (literature review), other open-source models (e.g. Llama and GLM 5.2), and Denario

Generative AI tool/services uses

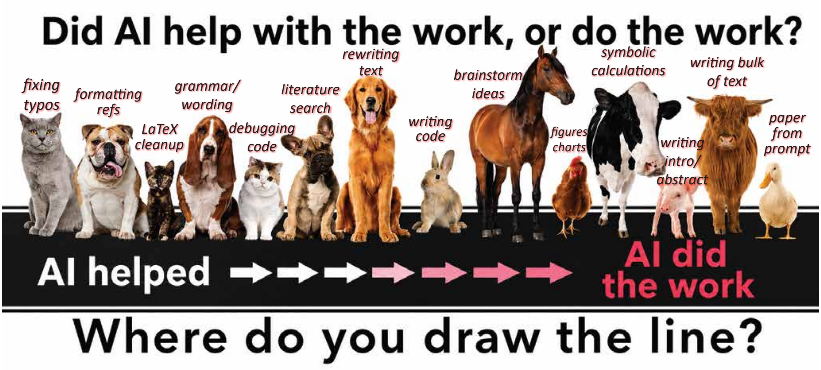


- Again we see some not-surprising trends
  - People are more likely to use AI for "busywork"; anything that gets in the way of physics...
  - ...and less likely to use it as a "replacement" for work

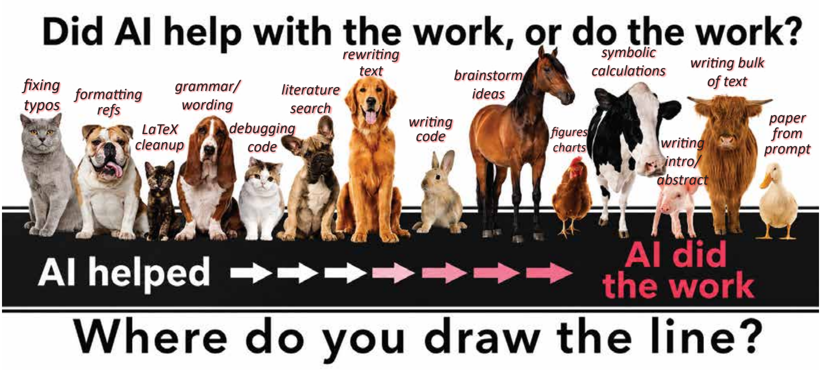
The human line



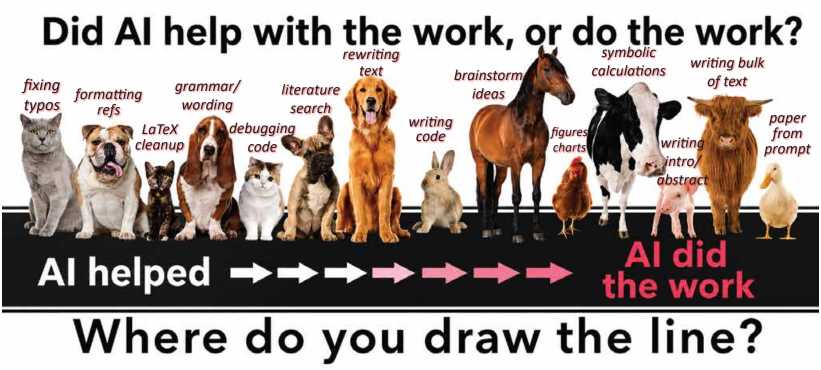
The human line



The human line



- There is a continuum suggested by the results



- There is a continuum suggested by the results
- Perhaps we'd prefer to draw the line exactly one spot to the right of where we stand...

AI use and outlook

- Doesn't look like AI use is going away ("tell us something we don't know!")

AI use and outlook

- Doesn't look like AI use is going away ("tell us something we don't know!")
- ...but it's also slowly becoming normalized
- Worth noting that even among those who self-reported as "not at all" using AI...

AI use and outlook

- Doesn't look like AI use is going away ("tell us something we don't know!")
- ...but it's also slowly becoming normalized
- Worth noting that even among those who self-reported as "not at all" using AI... about 40% still used it lightly (typo fixing and minor edits)

AI use and outlook

- Doesn't look like AI use is going away ("tell us something we don't know!")
- ...but it's also slowly becoming normalized
- Worth noting that even among those who self-reported as "not at all" using AI... about 40% still used it lightly (typo fixing and minor edits)
- So if this is the present of AI, what do you think about the future?

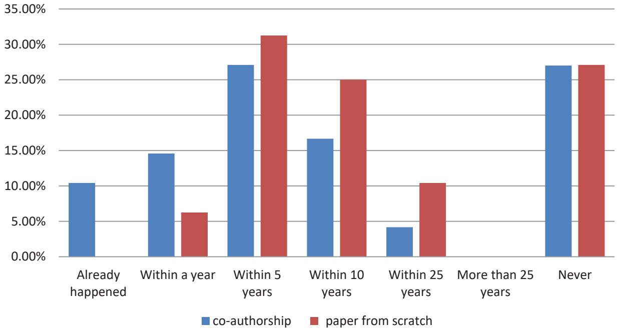
AI research predictions

- We asked you:
  - When do you expect AI to make a sufficiently novel intellectual contribution in deriving a physics result that it would merit co-authorship?
  - When do you expect an AI agent, operating without direct human scientific input after the initial prompt, to independently generate a novel physics result that is subsequently published in a reputable peer-reviewed journal?

AI research predictions

- We asked you:
  - When do you expect AI to make a sufficiently novel intellectual contribution in deriving a physics result that it would merit co-authorship?
  - When do you expect an AI agent, operating without direct human scientific input after the initial prompt, to independently generate a novel physics result that is subsequently published in a reputable peer-reviewed journal?
- (Hopefully) precisely worded, but we managed to sneak in two highly subjective qualifiers!
  - merit co-authorship
  - reputable (human?) journal

AI research predictions: now or never



- "Never" is a long time, but the split in predictions is clear: it's either happening within a generation or not at all
- And we definitely already have nascent steps towards both co-authorship and paper generation

AI research predictions: now or never

2602.12176

Single-minus gluon tree amplitudes are nonzero

Alfredo Guevara,[1] Alexandru Lupsasca,[2,3] David Skinner,[4]
Andrew Strominger,[5] and Kevin Weil[2] on behalf of OpenAI

[1]Institute for Advanced Study [2]OpenAI [3]Vanderbilt University [4]Cambridge University [5]Harvard University

Single-minus tree-level n-gluon scattering amplitudes are reconsidered. Often presumed to vanish, they are shown here to be nonvanishing for certain "half-collinear" configurations existing in Klein space or for complexified momenta. We derive a piecewise-constant closed-form expression for the decay of a single minus-helicity gluon into $n-1$ plus-helicity gluons as a function of their momenta. This formula nontrivially satisfies multiple consistency conditions including Weinberg's soft theorem.

2606.03300

A proof of an identity for the critical exponents of jamming

Giorgio Parisi[1,2,3] and Francesco Zamponi[3]

[1]*International Research Center of Complexity Sciences, Hangzhou International Innovation Institute, Beihang University, Hangzhou 311115, China*
[2]*CNR–Nanotec, Rome Unit, and INFN–Sezione di Roma 1, 00185 Rome, Italy*
[3]*Dipartimento di Fisica, Sapienza Università di Roma, 00185 Rome, Italy*

Abstract

Within the full replica-symmetry-breaking (fullRSB) solution of dense hard spheres in infinite dimension, Charbonneau, Kurchan, Parisi, Urbani, and Zamponi (CKPUZ; J. Stat. Mech. P10009, 2014) introduced three critical exponents a, b, c governing the matching region of the fullRSB profile near the jamming transition. These exponents satisfy two scaling relations. The first, $b = (1+c)/2$, was established analytically by the diffusion-drift balance in the scaling ansatz. The second, $a + b = 1$, was observed numerically to arbitrary precision but could not be proven. The exponents a, b, c of the scaling fullRSB ansatz are related to the physical exponents α, θ, κ that control the gap, force, and overlap distributions by the relations $\alpha = a/b$, $\theta = (c-a)/(b-c)$, $\kappa = c+1$. Crucially, the relation $a + b = 1$ yields the scaling relations $\alpha = 1/(2+\theta)$ and $\kappa = 2 - 2/(3+\theta)$ predicted on independent grounds by the mechanical-marginal-stability arguments of Wyart and collaborators. Here, we give an analytic proof of the identity $a + b = 1$ from the scaling fullRSB equations. The proof was obtained through interaction with Claude (Sonnet 4.6 and Opus 4.7) and verified by us.

- Papers brought to our attention as examples of "already happened": are these results enough to merit authorship?


OpenReview.net

← Go to **Agents4Science 2025 Conference** homepage

QITT-Enhanced Multi-Scale Substructure Analysis with Learned Topological Embeddings for Cosmological Parameter Estimation

Denario Astropilotai, Pablo Bermejo, Boris Bolliet, Francisco Villaescusa-Navarro, Pablo Villanueva Domingo

Published: 09 Oct 2025, Last Modified: 05 May 2026 Agents4Science Everyone

Revisions BibTeX CC BY 4.0

Keywords: Dark matter, Observational cosmology, Large-scale structure of the universe, N-body simulations, Intergalactic medium

Abstract:

Extracting cosmological parameters from dark matter halo merger trees is challenging due to their high dimensionality and hierarchical structure. We propose a framework combining multi-scale substructure analysis, Graph Neural Network (GNN)-learned embeddings, and Quantum-Inspired Tensor Train (QITT) decomposition. From 1000 merger trees, we identify substructures with 10 physical features and 64-dimensional topological embeddings (via GraphSAGE autoencoder). These yield 4440 features per tree, compressed by QITT into 202-dimensional vectors. Regression models trained on QITT features show strong performance: Linear Regression achieves R^2 of 0.923 for Ω_m and 0.621 for σ_8, while QITT-enhanced XGBoost significantly outperforms baselines without QITT ($p < 0.05$). Although global aggregate tree features reached a higher R^2 of 0.970 for Ω_m, QITT enables compact, informative representations integrating fine-grained substructure and topology. This establishes a promising pipeline for data-driven cosmology.

Supplementary Material: zip

Submission Number: 218

- And of course we have a paper fully generated with Denario, accepted for publication in the Open Conference of AI Agents for Science


OpenReview.net

← Go to **Agents4Science 2025 Conference** homepage

QITT-Enhanced Multi-Scale Substructure Analysis with Learned Topological Embeddings for Cosmological Parameter Estimation

Denario Astropilotai, Pablo Bermejo, Boris Bolliet, Francisco Villaescusa-Navarro, Pablo Villanueva Domingo

Published: 09 Oct 2025, Last Modified: 05 May 2026 Agents4Science Everyone
Revisions BibTeX CC BY 4.0

Keywords: Dark matter, Observational cosmology, Large-scale structure of the universe, N-body simulations, Intergalactic medium

Abstract:
Extracting cosmological parameters from dark matter halo merger trees is challenging due to their high dimensionality and hierarchical structure. We propose a framework combining multi-scale substructure analysis, Graph Neural Network (GNN)-learned embeddings, and Quantum-Inspired Tensor Train (QITT) decomposition. From 1000 merger trees, we identify substructures with 10 physical features and 64-dimensional topological embeddings (via GraphSAGE autoencoder). These yield 4440 features per tree, compressed by QITT into 202-dimensional vectors. Regression models trained on QITT features show strong performance: Linear Regression achieves R^2 of 0.923 for Ω_m and 0.621 for σ_8, while QITT-enhanced XGBoost significantly outperforms baselines without QITT ($p < 0.05$). Although global aggregate tree features reached a higher R^2 of 0.970 for Ω_m, QITT enables compact, informative representations integrating fine-grained substructure and topology. This establishes a promising pipeline for data-driven cosmology.

Supplementary Material: zip
Submission Number: 218

- And of course we have a paper fully generated with Denario, accepted for publication in the Open Conference of AI Agents for Science
- But do they "count"?

Some selected comments...

- We had a lot of thought-provoking comments, of which only one was LLM-generated
- Too many to go through in detail, but let's see a few outtakes...
- We have some critical comments on their efficacy:
  - *At the moment, AI agents work as advanced "rubber ducks" to spar arguments and sharpen intuition. In my opinion they have become necessary because researchers tend to be more and more isolated and they do not discuss any more ideas and thoughts, afraid of making a mistake.*
  - *Calling [LLMs] AI already reflects the layman's ignorance in this field of research, as LLM are but only a fragment of the different AI topics developed since the 50s. Moreover, contrary to other methods, like those based on logic, LLM are by their very definition incapable of sound inference. *Blindly* trusting a system with no knowledge of the semantics of the syntax it operates on is nothing else but foolish. [...] I personally suspect that the claimed increase in productivity thanks to the development of LLMs in most cases reduces to an argument of having more quantity over quality*

Some selected comments...

- And some critical comments on the ethics involved:
  - *[It doesn't make] any sense to list [AI] as an co-author, for similar reasons we don't list other tools like the world wide web, specific programs, etc. as a co-author. What is the "knowledge" generated by AI worth if humans don't understand the output?*
  - *We need to emphasize this social aspect of the human condition in our work. We need mechanisms to limit academia's dependency on an oligopoly of AI and software companies, both [...] to limit its influence on academia and to allow people and universities without the financial means [...] to access these tools.*
  - *As a free being with unlimited potential, I would find it insulting to rely on technology designed to literally steal genuine creativity. AI will one day overpass human scientists for sure, but it tells more about science than about AI.*
  - *The question is not whether AI can or cannot replace scientists, but rather: if politicians and society believe that knowledge will emerge magically from an AI, why would they pay for scientists, engineers, universities, or research centers?*

Some selected comments...

- But others are less negative (more positive?):

  - *I think those who pursue excessive mathematical and logical rigour will be replaced by AI, and it is time to cherish our humanity-oriented aspects more than ever. Because our intuition, insights based on observations and physical experience, historical context will never be replaced.*

  - *AI may offer a way to recover control over the deeper purpose of research. We should not approach science as an elite competitive sport aimed at fame, but as an expression of curiosity and love for human knowledge.*

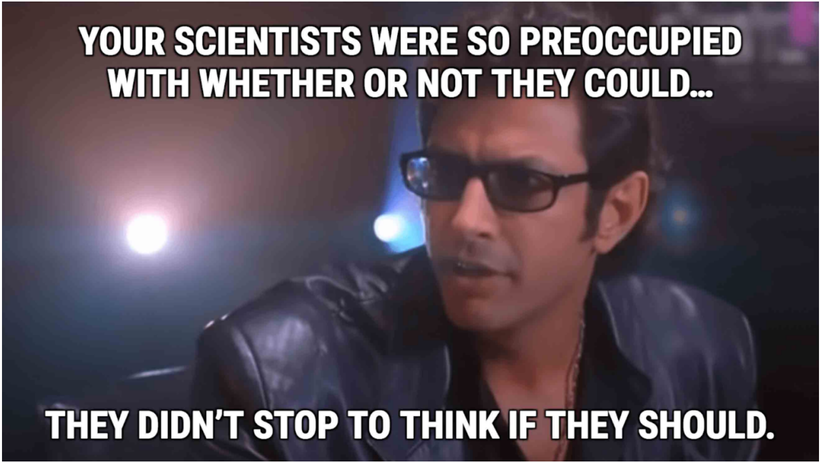
Some selected comments...

- But others are less negative (more positive?):
  - *I think those who pursue excessive mathematical and logical rigour will be replaced by AI, and it is time to cherish our humanity-oriented aspects more than ever. Because our intuition, insights based on observations and physical experience, historical context will never be replaced.*
  - *AI may offer a way to recover control over the deeper purpose of research. We should not approach science as an elite competitive sport aimed at fame, but as an expression of curiosity and love for human knowledge.*
- So what have we learned? Have we learned anything?

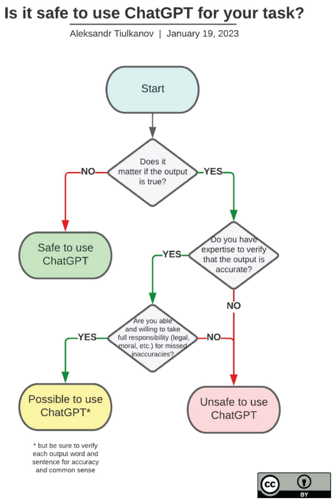
Overall opinions

- Some common threads emerge:
  - LLMs cannot replace intuition; humanity may be an inalienable part of the research process
  - LLMs are a tool, and like any tool, they can be misapplied and therefore must come with limitations (either self-imposed or through regulation)
  - Even the skeptics among us cannot deny that LLMs have their uses, and that they will probably improve, but there are real dangers involved with their unchecked development

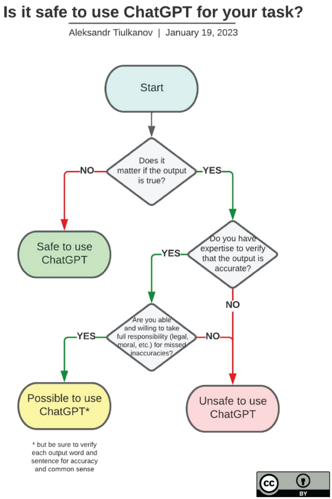
Overall opinions

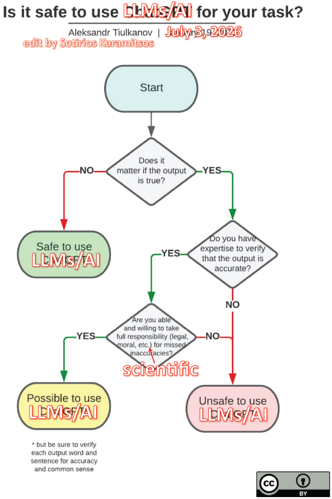


The Universal LLM Use Case Flowchart



The Universal LLM Use Case Flowchart






AI
~~Life~~, uh, finds a way.


AI
~~Life~~, uh, finds a way?