Carnegie Mellon University

Language Technologies Institute

Guest Lecture: AI for and as Science

Sireesh Gururaja

October 29, 2026

About me

- Current PhD Student at CMU
- I work on understanding the incentives that shape NLP, and building systems that preserve individual expertise
- I spent six years in industry before coming back to grad school, working on “AI” for such fascinating problems as reinsurance and corporate tax

Outline

- Table Setting

- AI for Math

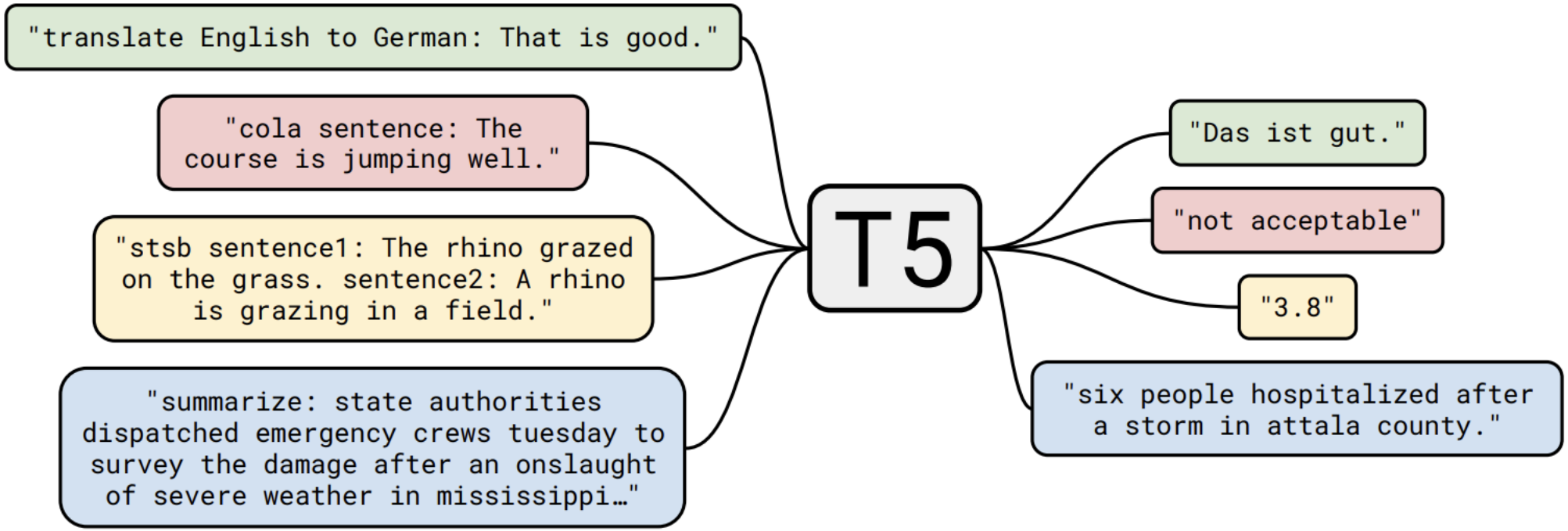
- AI for Materials Science

- AI for AI, and AI as Science

What do we mean by “AI”?

- Historically, AI is a *broad* term.
- Here, we’re using AI specifically to refer to deep learning models, not just LLMs!
- An LLM is still a good model to think about the AI we’re using here, though

What does “AI” do?



Raffel et al. (2020) *Exploring the limits of transfer learning*

AI for Math

August 1, 2026 Publication

Ten advances in mathematics and theoretical computer science

Read the paper ↗

Read the reasoning walkthroug

Learning more about Claude's mathematical capabilities



Google DeepMind

Advancing Mathematics Research with AI-Driven Formal Proof Search

Why do we do math?

Research mathematics deals with understanding basic structures of shapes, numbers, and natural phenomena. Over the course of generations, it has built a large corpus of sophisticated ideas, methods, abstractions, and other tools to comprehend the mathematical landscape. In turn, modern technologies and sciences are based on mathematical tools.

mathandai.org

Why might we want to do math with AI?

- Proofs for open math problems can potentially have major downstream impacts across fields like physics and engineering

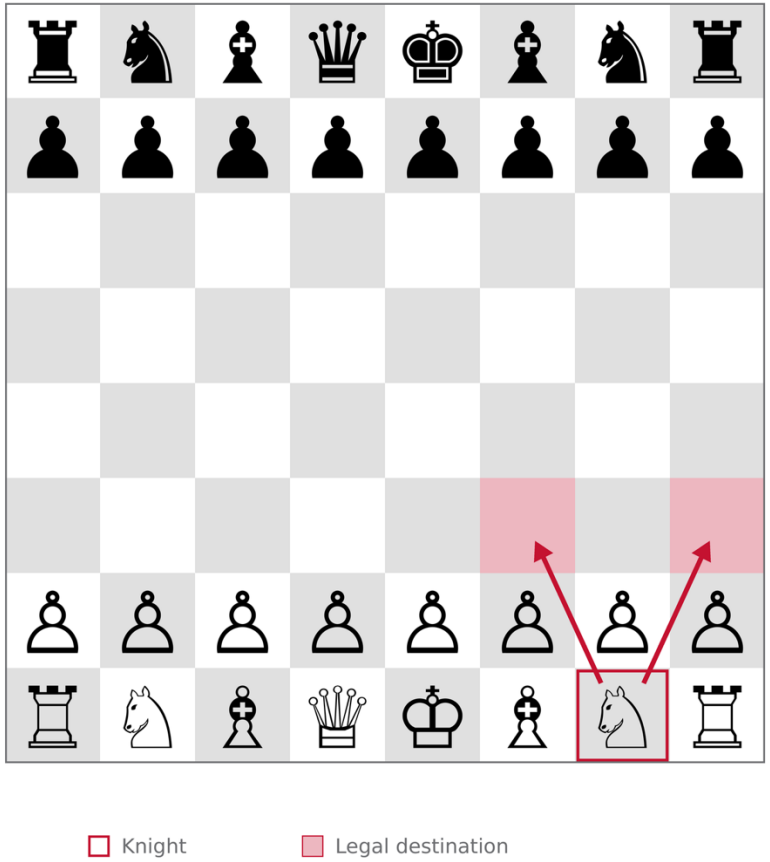
- Advances in AI for math also pave the way for *formal verification* of software: proving that software behaves in the ways that you want

- AI allows us to scale the process of proving theorems*

What does it mean to do math with AI?

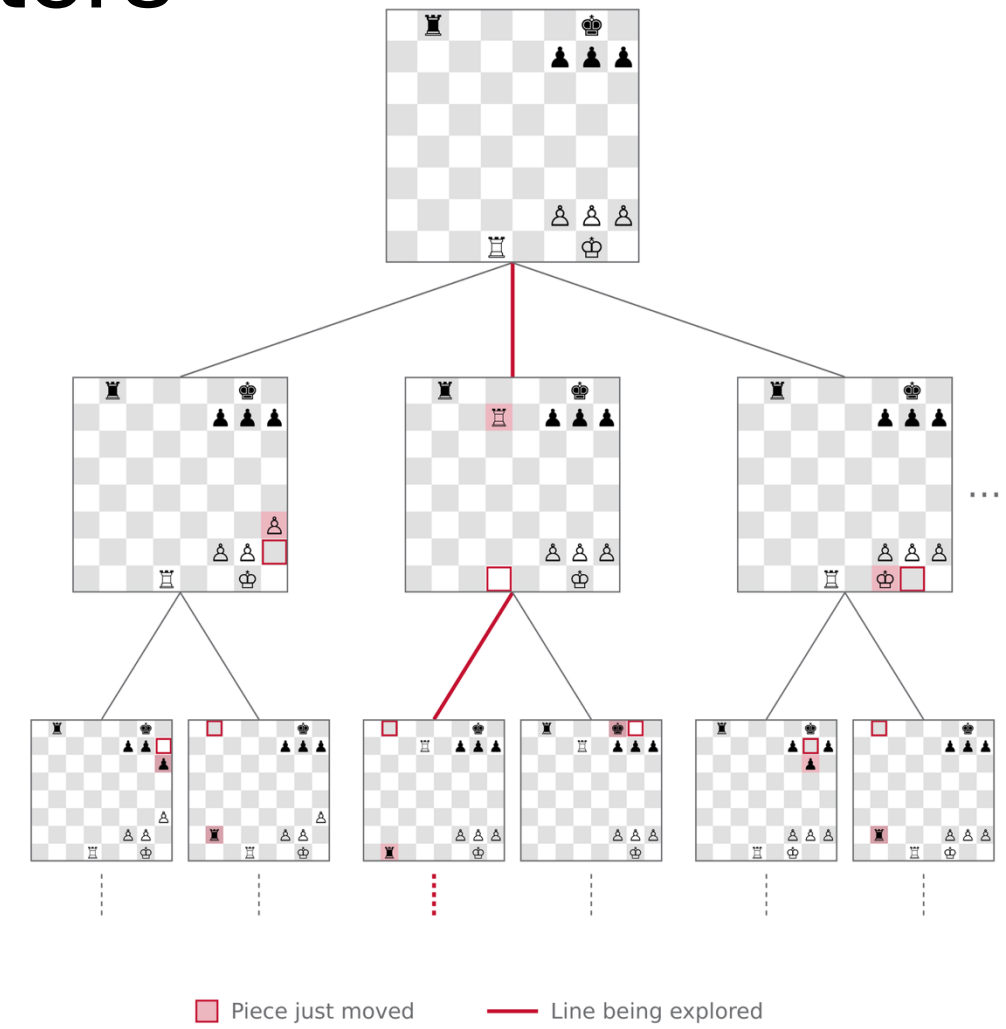
- Math can be seen as a *formal system:* there is a set of rules you can follow to expand on a set of axioms

- As an analogy: consider chess



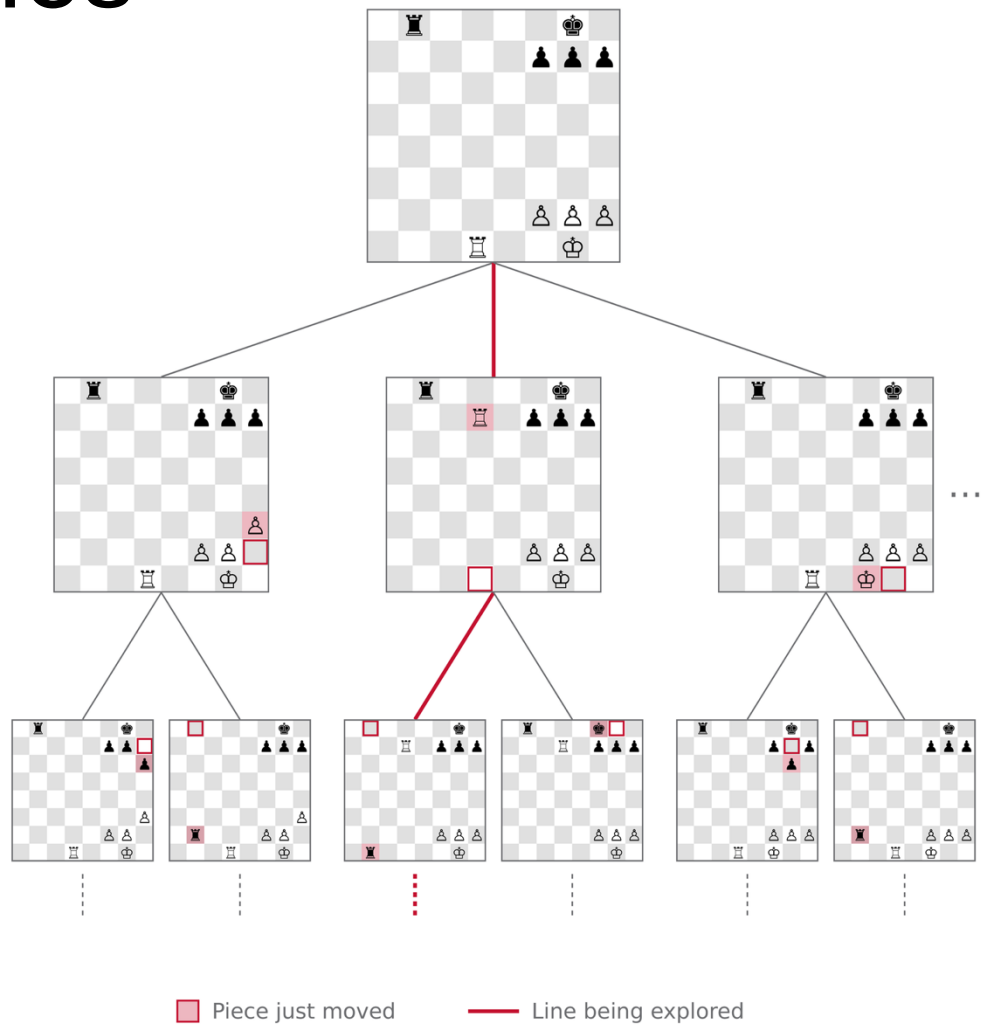
Playing chess with computers

- A chess board has a finite number of ways it can evolve
- Chess engines will search through the possible valid next moves to pick one
- Figuring out which moves are *good*…



The challenge: combinatorics

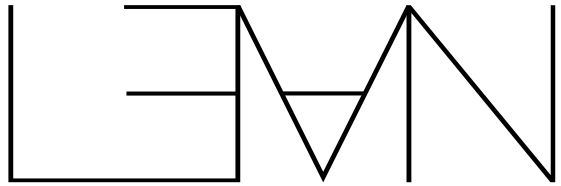
- Despite a finite number of rules and evolutions, the space to search is *huge*
- An early estimate by Claude Shannon argued for 10^{120} possible games



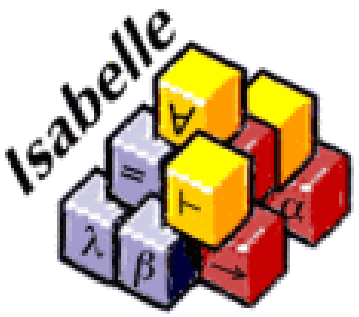
Computational math

- Math can also be formalized such that a computer can check whether the state of your proof is “valid”

- This can also be used to work forwards or backwards from particular statements!

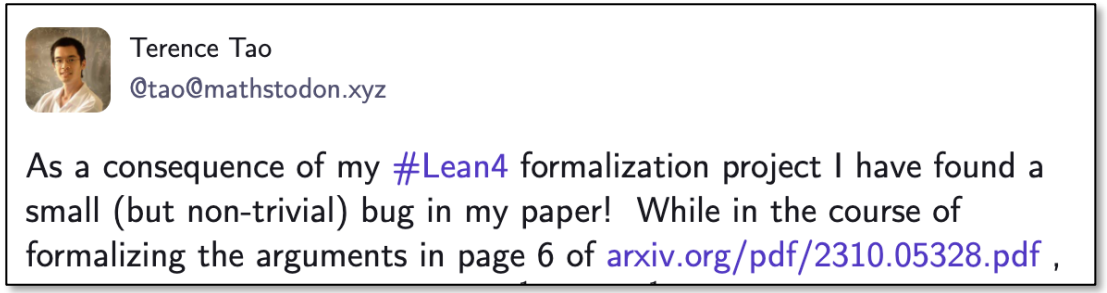






Computational math, historically

- No AI (by current definitions) involved
- Mostly mathematicians confirming that their proofs were right



mastodon

```
n : ℕ
s : ℕ → ℝ
h1 : n > 2
h2 : attainable n s
h1' : 2 < ↑n
⊢ 0 < ↑n - 3

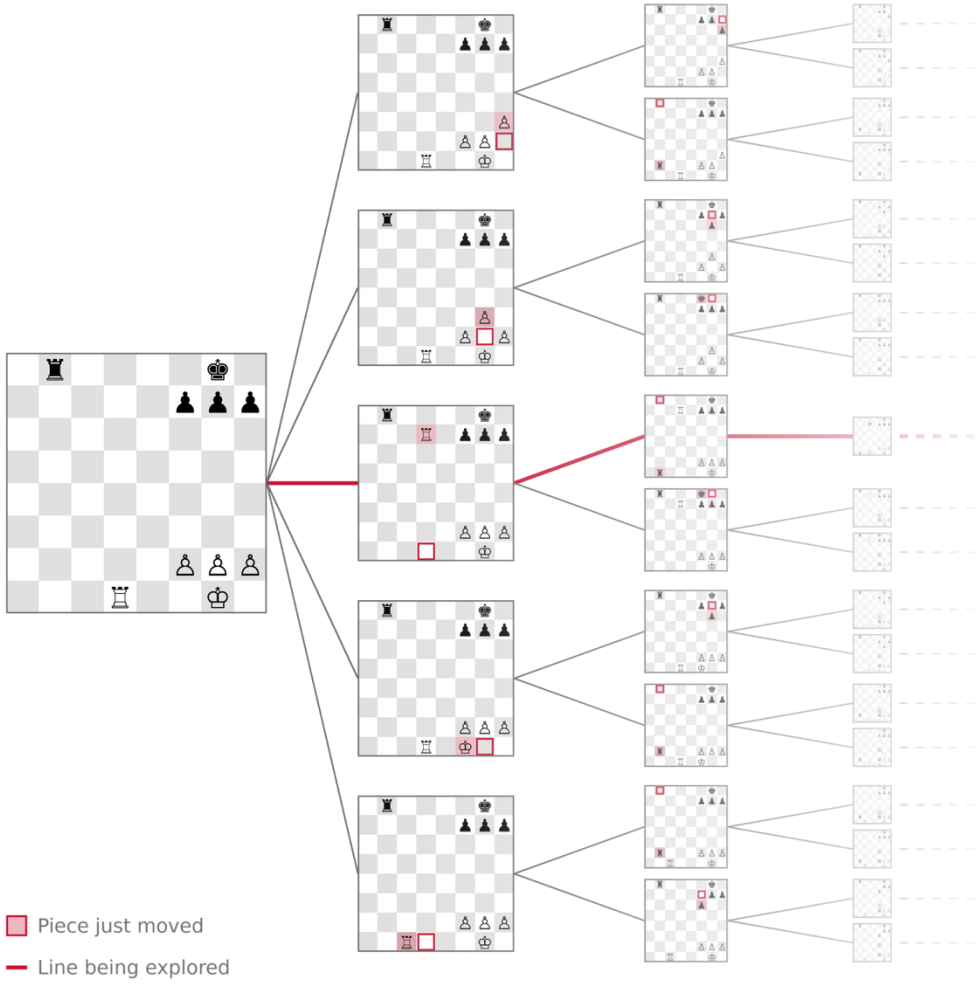
Messages (1)
prev_bound.lean:222:6

linarith failed to find a contradiction
case h
n : ℕ
s : ℕ → ℝ
h1 : n > 2
h2 : attainable n s
h1' : 2 < ↑n
at : 0 ≥ ↑n - 3
⊢ False
```

The combinatorics are even worse

- Chess and go both have *finite* action spaces, and it's still not tractable to search the entire tree of possibilities

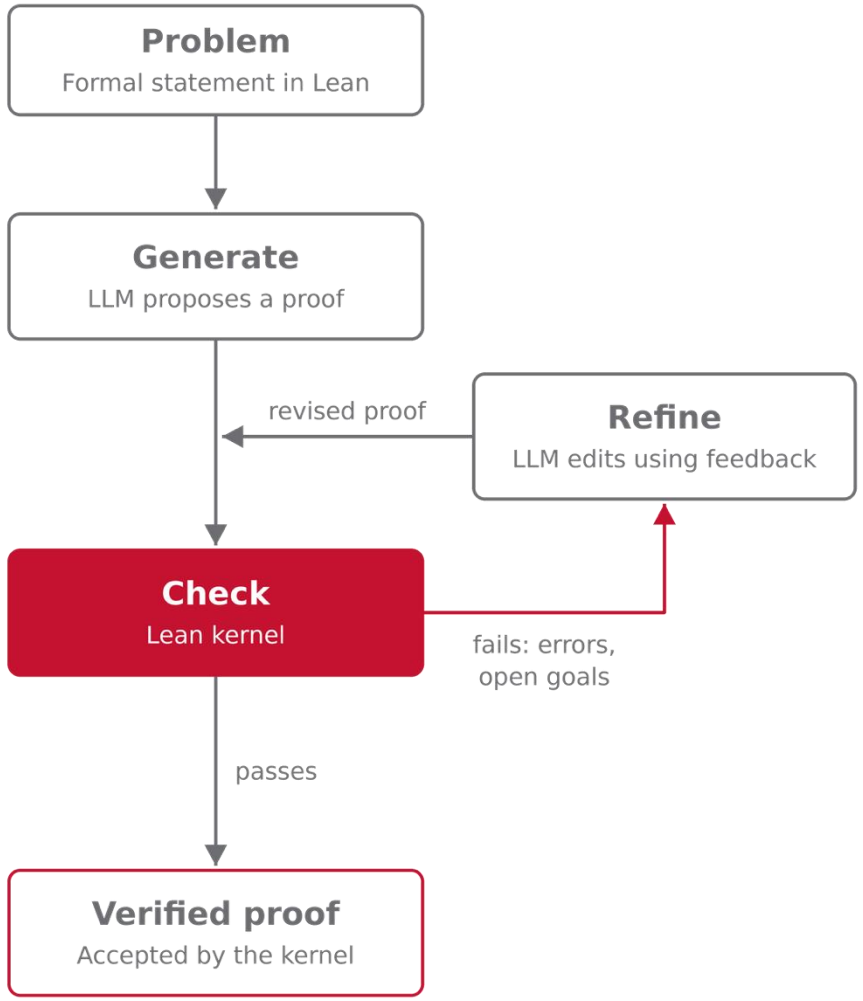
- Even if we *only* use theorems already in a popular library, we still have hundreds of thousands of actions to check



Where LLMs come in

- LLMs generate *candidate* proofs that the checker can then validate

- As LLMs have gotten better at this, the number of steps they can take in between invocations of the checker have increased



This is still *enormously* expensive

- Even where candidates are *valid,* it might not get you closer to a proof that is interesting or useful

- For the Millennium Prize Navier-Stokes solutions, estimates at retail prices ran to ~$6.5m

> The agents arrived at their resolution on Saturday, September 5, about 88 hours after the first agents were launched. Lean formalization and verification took an additional 17 hours via GPT-6 Astra.
>
> Across all attempted problems, the agents sent 4.9 million messages and used about 300 billion output tokens. In the process of resolving the Navier–Stokes problem, the agents sent 2.7 million messages and used approximately 130 billion output tokens.

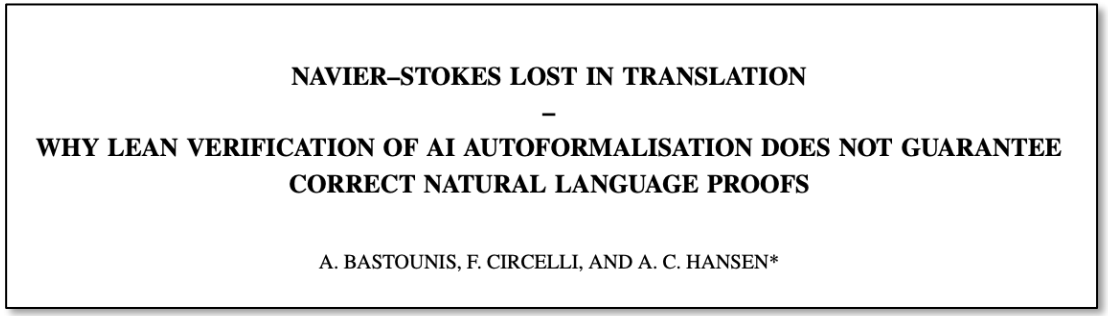
OpenAI

Drawbacks of AI theorem proving: understanding

- Formal versions of math theorems aren't very readable

- They still require interpretation and re-presentation to be understood like traditional™ math proofs

- There's some doubt introduced by gaps between the Lean and the "natural language" proof



NAVIER–STOKES LOST IN TRANSLATION

–

WHY LEAN VERIFICATION OF AI AUTOFORMALISATION DOES NOT GUARANTEE CORRECT NATURAL LANGUAGE PROOFS

A. BASTOUNIS, F. CIRCELLI, AND A. C. HANSEN*

Drawbacks of AI theorem proving: correctness

- Theorem checking relies on the checker to be correct!
- Bugs have been found, and theorem checkers are works in progress
- Checkers (the non-AI part of this loop) are seen as added rigor on human math

> Lean 4 was released in September 2023. Prior to release, two soundness bugs were found and corrected. In May 2025, another soundness bug was reported, caused by overflow. All hell broke loose in the spring and summer of 2026, which is now being called the "Summer of Soundness Bugs". Several soundness bugs in Lean were uncovered in July and August. The summer madness affected various proof assistants, but my focus is Lean. One Lean bug led to an illicit disproof of the Collatz conjecture. I learned of the bug this summer when it produced a short illicit proof of the Kepler conjecture in Lean. All these bugs were quickly repaired, and mathlib has been verified by the repaired kernel. An analysis of the soundness bugs is found in de Moura's postmortem.

Thomas Hales, on Terry Tao's Blog

Theorem proving…isn’t the point?

- A person proving a theorem develops an intuition, and potentially new mechanisms that can then potentially be applied to other problems

- Math has also historically used “grand challenges” as one way to motivate the field. Will AI proving still generate new problems to solve?

Drawbacks of AI theorem proving: violating community norms

- The use of AI trained on unclear data makes credit attribution *much* more complicated

- What does contemporaneous work mean if a rival can prove something you're working on by throwing compute at it?

> We (the researchers and the agents) did not see any of their work through any means until they released it publicly — in particular, no specific user data was accessed in order to solve this problem. While unlikely, we cannot rule out that de-identified data derived from their usage of our products helped improve our models. However, our proofs differ significantly and even the precise results proved are different in the Euler case (forced vs unforced).

Wayback machine snapshot, 9/9/26

> We (the researchers and the agents) did not see any of their work through any means until they released it publicly — in particular, no specific user data was accessed in order to solve this problem.
>
> Following an investigation, we have confirmed that Buckmaster's Codex prompts over the two months preceding this announcement and paper on September 8, 2026, could not have influenced the system in any way, including through training. The OpenAI internal model used for this result was developed through large-scale reinforcement learning on top of a previously pretrained model. Our proofs also differ significantly. In the Euler case, Alpöge and Buckmaster proved a result with external forcing, while OpenAI's system proved a result without external forcing.

Current page

AI builds on the work of people in automating the proving of theorems, both through the literature and proof checkers like Lean.

But what about the implicit benefits that we assumed about proving theorems before, like the new tools that emerge?

What about the use of problems not just to advance the state of understanding, but for other goals, like pedagogy?

AI for Materials Science

November 29, 2023 Science

Millions of new materials discovered with deep learning

Amil Merchant and Ekin Dogus Cubuk

Trump Administration Announces More Than $5 Billion for the Genesis Mission, a National Mission on AI for Science

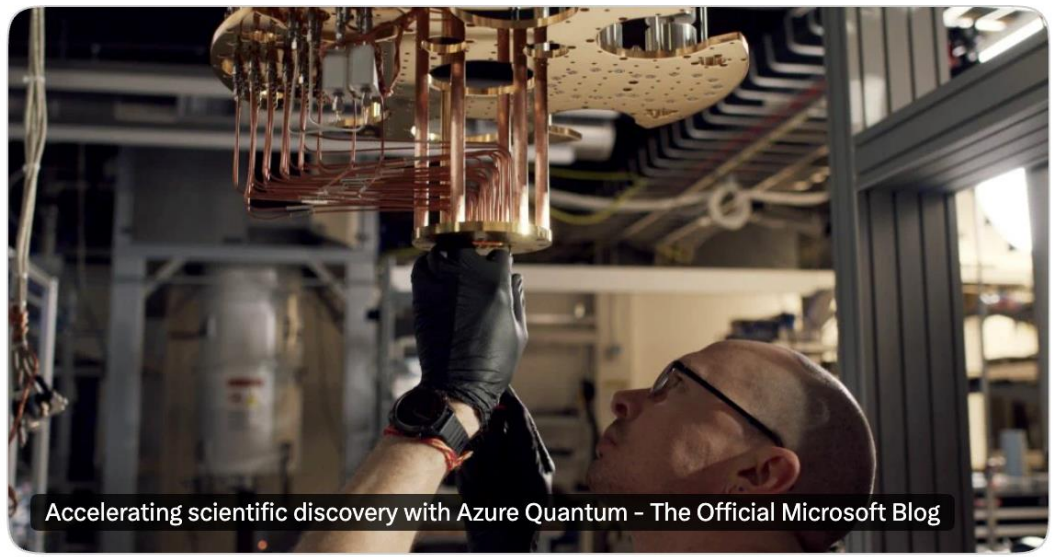
The White House | July 22, 2026

Designing Materials with Predictable Functionality: DOE will pair its world-leading experimental and computational capabilities, materials datasets, and physics-informed AI models in closed-loop experimentation that couples prediction, synthesis, and characterization to accelerate the design of new materials.



Satya Nadella
@satyanadella

Today, we are bringing together AI and quantum with Azure Quantum Elements, ushering in a new era of scientific discovery. Our goal is to compress the next 250 years of chemistry and materials science progress into the next 25.



AI in materials science: the goals

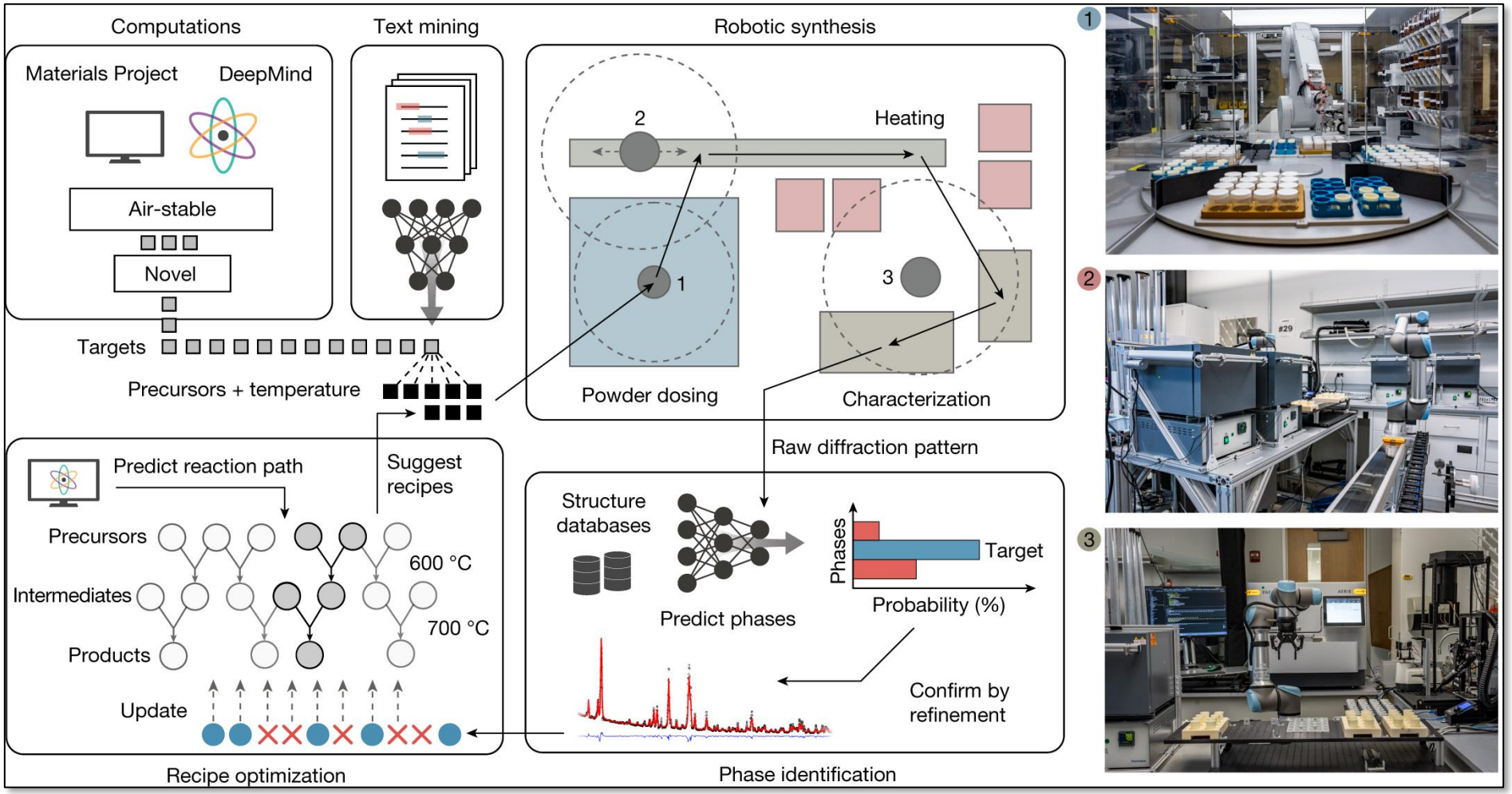
How can we produce materials that have the properties that we want?

AI in materials science: the goals

How can we produce materials that have the properties that we want?

- What materials have people synthesized in the past?
- What methods did they use to synthesize those materials?
- What properties did those materials have?

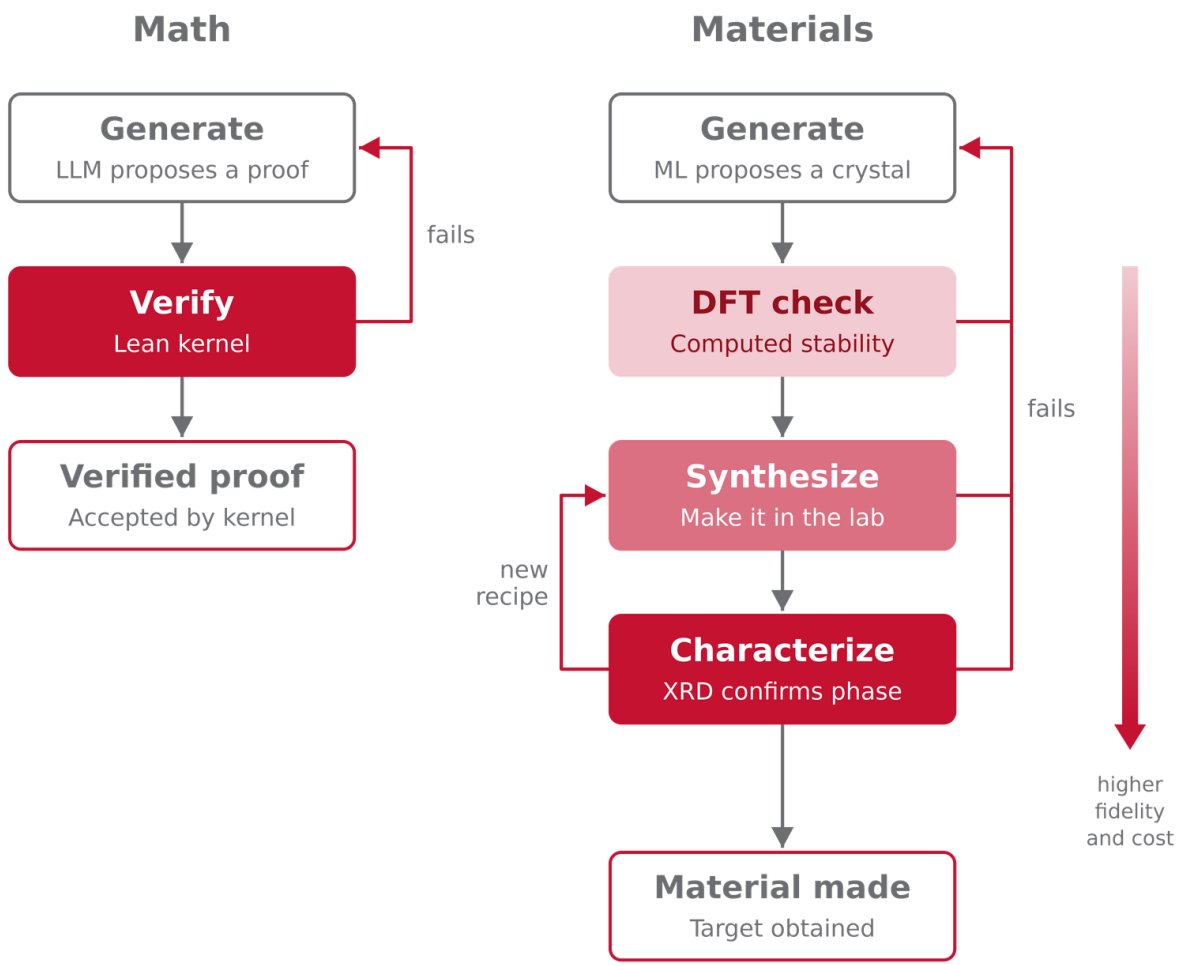
Autonomous Materials Synthesis



[Szymanski et al., An autonomous laboratory for the accelerated synthesis of inorganic materials]

This looks kinda like AI in math!

- The loop remains a generate-verify loop
- Materials are not a formal system, though, so there are multiple stages of “checking” that have different fidelity and cost



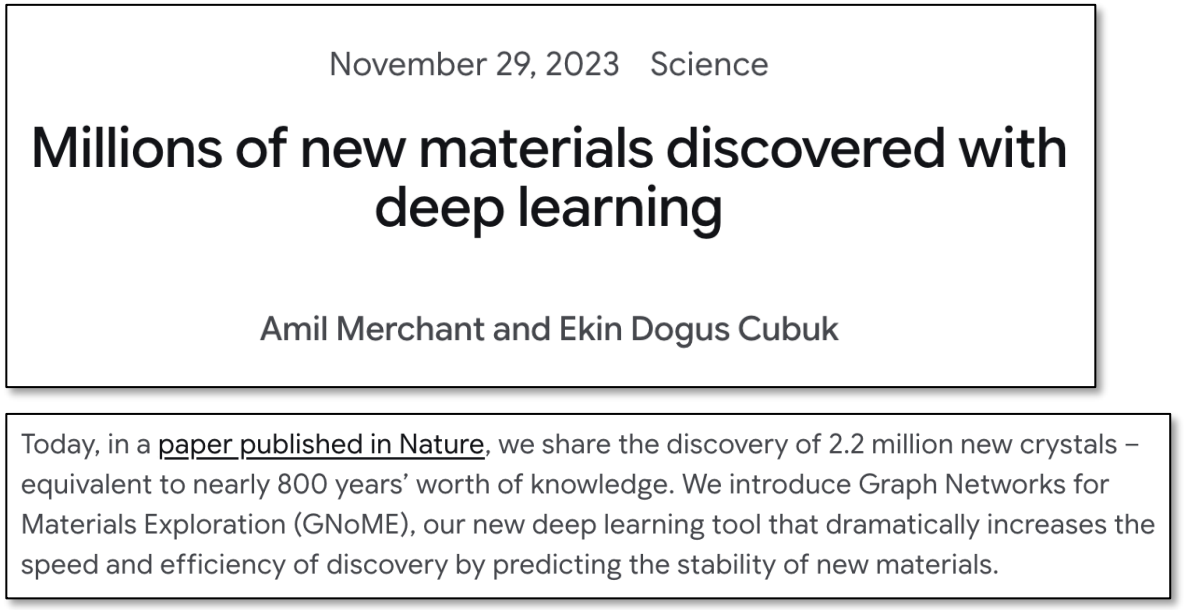
Materials are not a formal system

- At each stage, the materials domain adds indeterminacy
- Candidate materials can look good, but not be synthesizable, or well modeled by DFT
- Even if they are synthesizable, they may not have the properties you want!

Case study: Google's GNoME

- Google released a model in 2023, which they claimed "discovered" 380,000 novel, stable crystals

- This approach used a form of deep learning known as Graph Neural Networks



November 29, 2023 Science

Millions of new materials discovered with deep learning

Amil Merchant and Ekin Dogus Cubuk

Today, in a paper published in Nature, we share the discovery of 2.2 million new crystals – equivalent to nearly 800 years' worth of knowledge. We introduce Graph Networks for Materials Exploration (GNoME), our new deep learning tool that dramatically increases the speed and efficiency of discovery by predicting the stability of new materials.

Google Deepmind Blog

GNoME: Criticism

- Cheetham and Seshadri (2024) is a thoughtful critique of Google's approach

- While they still see promise in the method, they identify numerous issues with the work as presented

> *"We also note that, while many of the new compositions are trivial adaptations of known materials, the computational approach delivers credible overall compositions, which gives us confidence that the underlying approach is sound."*

GNoME: Criticism

- Cheetham and Seshadri (2024) is a thoughtful critique of Google's approach

- While they still see promise in the method, they identify numerous issues with the work as presented

> *"We have yet to find any strikingly novel compounds in the GNoME and Stable Structure listings, although we anticipate that there must be some among the 384,870 compositions."*

AI's primary value is *scale*

- Like in math, AI's primary value claim for materials is being able to predict *many, plausible* candidates

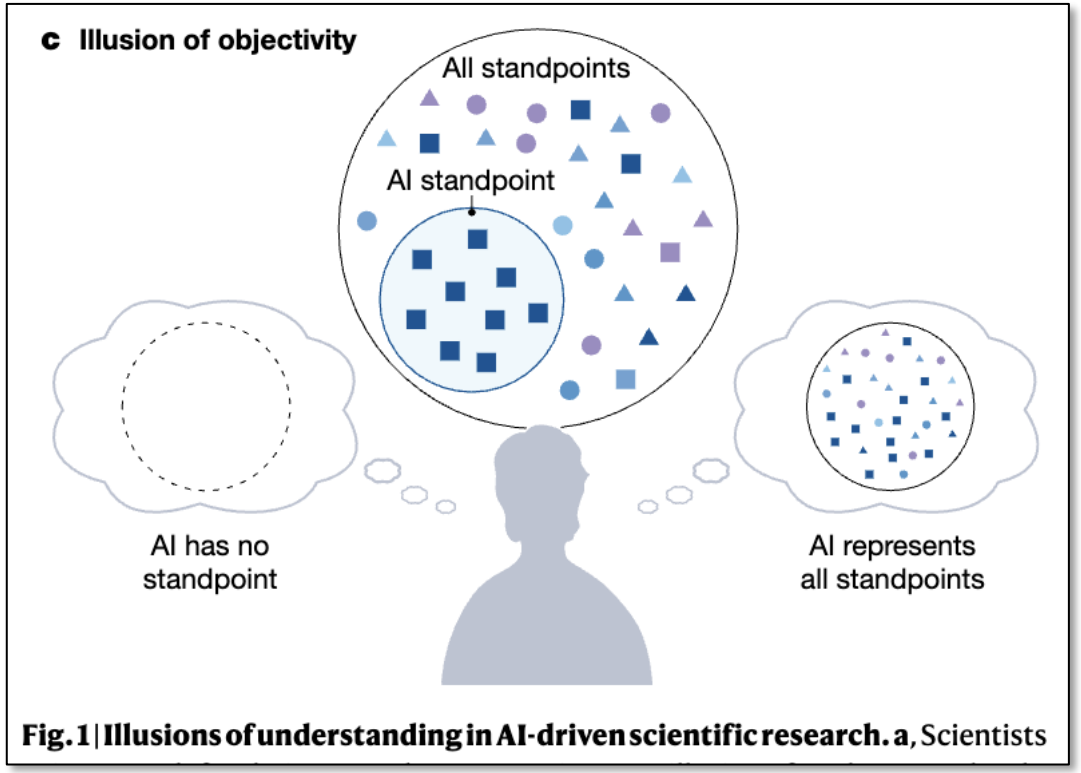
- Whether or not the community is able to assimilate them…

"What is now needed is greater effort to…filter out the many candidates that are not truly novel. It is impractical to do this manually with a list of almost 400,000 new compositions."

Data imposes its own distortions

- Hypotheses based on the literature are constrained in opaque ways by how the literature is used to train the AI model

- Even if the dataset used to train the AI were perfect, there's no guarantee that the trained model would be



Messeri and Crockett, *Artificial intelligence and illusions of understanding in scientific research* (2024)

What is a scientific paper?

- Latour and Woolgar (1986) argue that papers elide the material context that produced them
- AI's tendency to reduce a paper to text reifies this
- Crucially: actual scientists don't take papers at face value!

> "Once the data sheet has been taken to the office for discussion, one can forget the several weeks of work by technicians and the hundreds of dollars which have gone into its production. After the paper which incorporates these figures has been written…it is easy to forget that the construction of the paper depended on material factors."

What is a scientific paper?

- Papers are a partial representation of the research process
- Even for what they do capture, there is a wide spectrum of uncertainty that accompanies what's reported
- Papers belong to a social framework: they are understood as partial representations of research process, and they depend on interpretation to be read. Different scientists read papers differently!

Heterogeneity in scientific thought

- Scientists are *deeply* idiosyncratic in how they read and understand papers
- This is a good thing! We *want* people to develop different ways of thinking about problems
- If AI models are all exposed to the same data, do we retain a good distribution?

AI might expand our ability to understand what's in the literature: what we've done before, and what might be reasonable to do in the future.

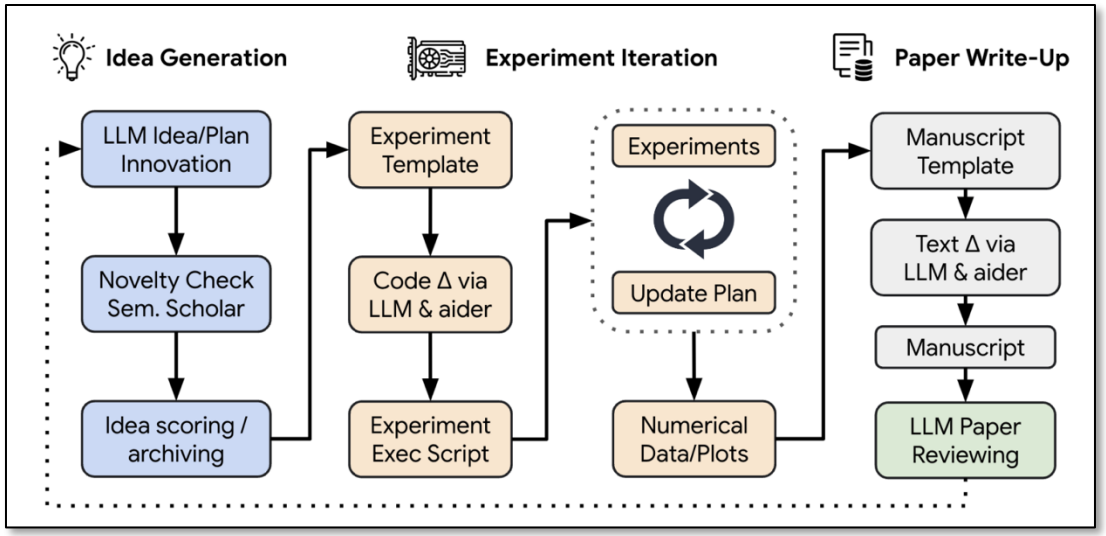
But does it promise a vision of complete literature understanding that it can't really deliver? Is that something that we want?

If we do get synthesizable materials out of this process, how do we deal with maybe not understanding how we got there, or how they work?

AI for AI, and AI as Science

AI for AI looks like…generation and verification

- Similar to what we've seen, AI has also been used to do AI research
- Verification here relies on our ability to quantify the performance of AI models

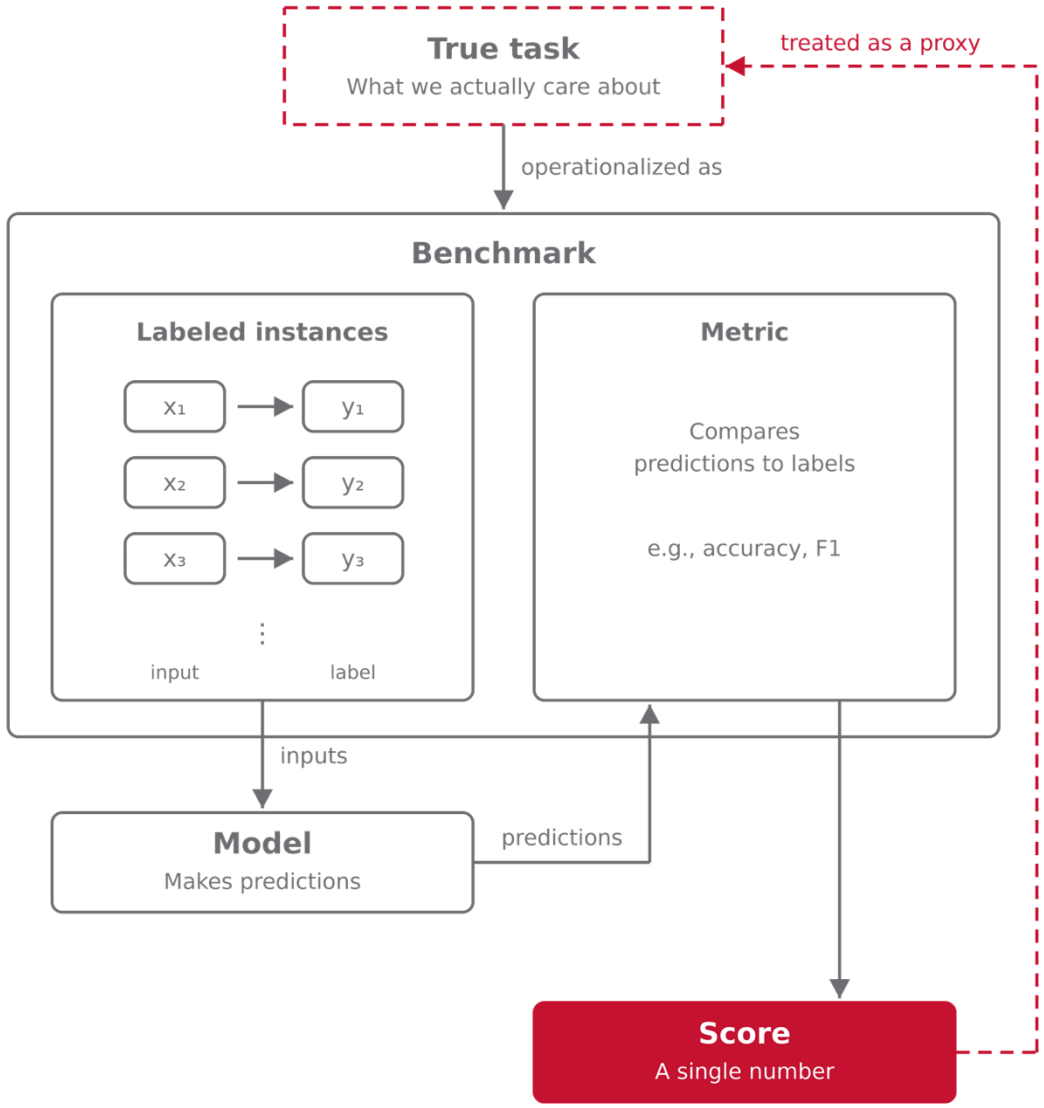


Lu et al. (2024) *The AI Scientist: Towards Fully Automated Scientific Discovery*

Verification for AI: Benchmarks

- Benchmarks are a concrete way to proxy a measure of *task* performance

- *Tasks* are operationalized into prediction tests on *collected data*, measured numerically with a *metric*



Benchmarking has a whole *host* of issues

- Are these datasets carefully constructed to be real reflections of tasks we care about?

Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks

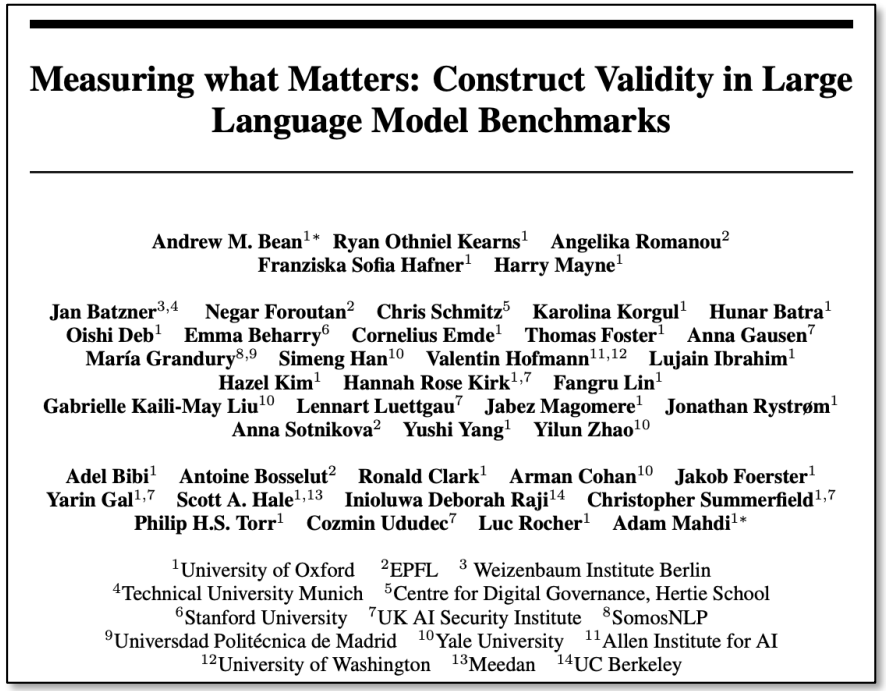
Curtis G. Northcutt* ChipBrain, MIT, Cleanlab
Anish Athalye MIT, Cleanlab
Jonas Mueller AWS

Are We Done with MMLU?

Aryo Pradipta Gema[1] **Joshua Ong Jun Leang**[1] **Giwon Hong**[1] **Alessio Devoto**[2] **Alberto Carlo Maria Mancino**[2,3] **Rohit Saxena**[1] **Xuanli He**[4] **Yu Zhao**[1] **Xiaotang Du**[1] **Mohammad Reza Ghasemi Madani**[5] **Claire Barale**[1] **Robert McHardy**[6] **Joshua Harris**[7] **Jean Kaddour**[4] **Emile van Krieken**[1] **Pasquale Minervini**[1,8]
[1]University of Edinburgh [2]Sapienza University of Rome [3]Polytechnic University of Bari
[4]University College London [5]University of Trento [6]AssemblyAI
[7]UK Health Security Agency [8]Miniml.AI

Benchmarking has a whole *host* of issues

- Does the data actually represent what we want to measure?



Measuring what Matters: Construct Validity in Large Language Model Benchmarks

Andrew M. Bean[1*] Ryan Othniel Kearns[1] Angelika Romanou[2]
Franziska Sofia Hafner[1] Harry Mayne[1]

Jan Batzner[3,4] Negar Foroutan[2] Chris Schmitz[5] Karolina Korgul[1] Hunar Batra[1]
Oishi Deb[1] Emma Beharry[6] Cornelius Emde[1] Thomas Foster[1] Anna Gausen[7]
María Grandury[8,9] Simeng Han[10] Valentin Hofmann[11,12] Lujain Ibrahim[1]
Hazel Kim[1] Hannah Rose Kirk[1,7] Fangru Lin[1]
Gabrielle Kaili-May Liu[10] Lennart Luettgau[7] Jabez Magomere[1] Jonathan Rystrøm[1]
Anna Sotnikova[2] Yushi Yang[1] Yilun Zhao[10]

Adel Bibi[1] Antoine Bosselut[2] Ronald Clark[1] Arman Cohan[10] Jakob Foerster[1]
Yarin Gal[1,7] Scott A. Hale[1,13] Inioluwa Deborah Raji[14] Christopher Summerfield[1,7]
Philip H.S. Torr[1] Cozmin Ududec[7] Luc Rocher[1] Adam Mahdi[1*]

[1]University of Oxford [2]EPFL [3] Weizenbaum Institute Berlin
[4]Technical University Munich [5]Centre for Digital Governance, Hertie School
[6]Stanford University [7]UK AI Security Institute [8]SomosNLP
[9]Universdad Politécnica de Madrid [10]Yale University [11]Allen Institute for AI
[12]University of Washington [13]Meedan [14]UC Berkeley

Position: Evaluating Generative AI Systems Is a Social Science Measurement Challenge

Hanna Wallach [1] Meera Desai [2] A. Feder Cooper [1] Angelina Wang [3] Chad Atalla [1] Solon Barocas [1]
Su Lin Blodgett [1] Alexandra Chouldechova [1] Emily Corvi [1] P. Alex Dow [1] Jean Garcia-Gathright [1]
Alexandra Olteanu [1] Nicholas Pangakis [1] Stefanie Reed [1] Emily Sheng [1] Dan Vann [1]
Jennifer Wortman Vaughan [1] Matthew Vogel [1] Hannah Washington [1] Abigail Z. Jacobs [2]

Benchmarking has a whole *host* of issues

- Is it reasonable to claim that models have a capability through a finite number of instances?

AI and the Everything in the Whole Wide World Benchmark

Inioluwa Deborah Raji
Mozilla Foundation, UC Berkeley
rajiinio@berkeley.edu

Emily M. Bender
Department of Linguistics
University of Washington

Amandalynne Paullada
Department of Linguistics
University of Washington

Emily Denton
Google Research

Alex Hanna
Google Research

Benchmarks as Microscopes: A Call for Model Metrology

Michael Saxon **Ari Holtzman** **Peter West**
William Yang Wang **Naomi Saphra**
University of California, Santa Barbara University of Chicago
University of Washington Kempner Institute at Harvard University

AI-for-science in the shadow of AI-as-science

- AI approaches in science reflect the way AI itself is performed as a science: *quantitatively*, at *scale*

- But communities construct knowledge in different ways, and that's not always compatible with the values that AI brings. It's under active negotiation!



What is lost when we focus only on numbers?

- Scientific work makes assumptions about how numbers are generated: what does it mean to cite someone? What qualifications does this score need to make sense?

- Accelerating only the parts of science that are quantifiable will have some weird consequences

Slop Papers

- When the citation/paper becomes a metric to optimize, we see a surge in “papers” being created

A Retrospective on the ICLR 2026 Review Process

ICLR 2026 received 19,525 valid, format-compliant submissions. 779 submissions were desk rejected for procedural/content violations and 5,042 were withdrawn, so that ultimately 13,763 submissions received an accept or reject decision based on 76,139 reviews provided by 18,054 reviewers. These

ICLR Blog

Table 1: Main EEG-to-image retrieval results on THINGS-EEG and EEG-ImageNet. All values are simulated placeholder results and should be replaced by actual experimental outputs. The best results are in bold and the second best are underlined.

Method	THINGS-EEG		EEG-ImageNet	
	Top-1 (%)	Top-5 (%)	Top-1 (%)	Top-5 (%)
EEGNet-Retrieval	18.6	37.9	24.2	46.7
TSception-Retrieval	20.9	41.5	26.0	48.9
EEG-Transformer	24.7	46.8	30.8	55.1
ATM-style Matcher	27.5	50.9	33.7	58.9
SeeEEG-style Matcher	29.1	53.4	35.4	61.2
NeuroClaim	**32.8**	**58.6**	**39.6**	**66.8**

Nihar Shah’s Blog

AI and concrete signals

- AI provides a way to mechanize and scale concrete signals

- These concrete signals are *incomplete* descriptions of the activity of a field



Article | Published: 14 January 2026

Artificial intelligence tools expand scientists' impact but contract science's focus

Qianyue Hao, Fengli Xu ✉, Yong Li ✉ & James Evans ✉

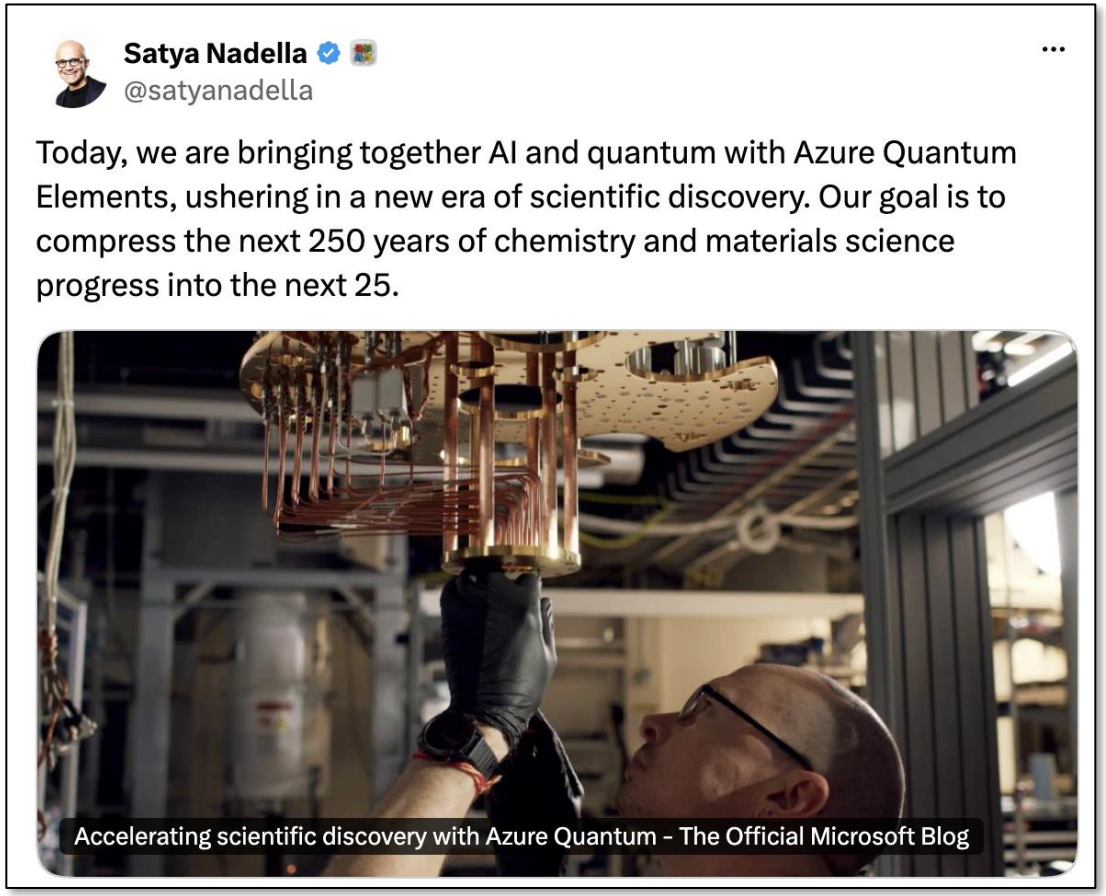
AI for science is opportunistic

- Most prominent AI results in science in the past several months are in domains where the solutions are *quickly* and *computationally* verifiable.

- How will AI impact fields where that is less the case, like materials or biochemistry? Will fields have enough time to catch up?

Hype and AI in Science

- The mechanics of hype both drive and are driven by AI “progress” in science.
- The need for hype pushes AI for science as a prestigious, proximal goal
- Results from AI in turn drive hype



Understanding

- The three fields we've looked at today expect decreasing levels of understanding: math > materials science > AI

- To what extent is understanding a necessary component of science?

Questions to think about

- What should *knowledge* mean in different fields of inquiry? How do we decide what we value in a field, and prevent that from being overwhelmed?

- What parts of our scientific processes should remain? What parts should be replaced?

- How do you encourage doing science in the spirit of the endeavor, and not turn the job of scientists into policing each other?

Further reading

- Messeri and Crockett (2024), *Artificial intelligence and illusions of understanding in scientific research*
- Evans and Duede (2025), *After Science*
- Ted Chiang (2000), *Catching Crumbs from the Table*
- Terence Tao (2007), *What is good mathematics?*