

AI for theorem proving

an engineer's note on implementation details

Contents of the talk

- History
- Progress
- What excites me

Sponsored

claude.ai

<https://www.claude.ai> ⋮**Claude for Advanced Research | Ai For Research**

Research competitors efficiently with Claude's ability to process and summarize documents. Gain valuable competitive insights with...

1M+ visits in past month

Sponsored

azure.microsoft.com

<https://azure.microsoft.com> > spot > virtual_machine ⋮**Microsoft® Azure - Virtual Machine**

Create the Right Virtual Machine for Your Operating System. Start Free Today.



OpenAI

<https://openai.com> ⋮**OpenAI**

We believe our research will eventually lead to artificial general intelligence, a system that can solve human-level problems. Building safe and beneficial ...

[Introducing ChatGPT](#)[Preparing for future AI...](#)[AI to Mattel's iconic brands](#)[Sora](#)

Google AI

<https://ai.google> ⋮**Google AI - How we're making AI helpful for everyone**

Discover how Google AI is committed to enriching knowledge, solving complex challenges and helping people grow by building useful AI tools and technologies.

[AI Principles](#)[Learn essential AI skills](#)[Why AI](#)[Our AI Journey](#)

First-gen AI were theorem provers (1956)



Allen Newell



Herbert Simon



Cliff Shaw

Logic theorist

- Axioms and deduction rules
- Heuristic search
- Problem decomposition
- Symbol manipulation

Proved 38 of the first 52 theorems in chapter 2 of *Principia Mathematica*.

More recently

SCIENCE

AI achieves silver-medal standard solving International Mathematical Olympiad problems

Mathematics

DeepMind and OpenAI International Mathematical Olympiad

25 JULY 2024

AlphaProof and AlphaGeometry teams

Two AI models have achieved gold medal standard for the first time in a prestigious competition for young mathematicians – and their developers claim these AIs could soon crack tough scientific problems

By [Alex Wilkins](#)

 22 July 2025

ByteDance Seed Prover Achieves Silver Medal Score in IMO 2025

Harmonic Announces IMO Gold Medal-Level Performance & Launch of First Mathematical Superintelligence (MSI) AI App

Date
2025-07-23

Deep learning for symbolic mathematics

by Guillaume Lample and François Charton (2019)

Integration: How do we do more than mathematica by learning the mapping from $f(x)$ to $F(x)$?

- Forward generation: pump through the mathematica integration engine
- Backward generation
 - Randomly generate an expression $F(x)$
 - Differentiate it to get $f(x)$
- Integration by parts
 - Randomly generate F and G
 - if fG is easy to integrate, or the integration is already in the dataset, accept it
 - We have $\int Fg = FG - \int fG$

Article | [Open access](#) | Published: 01 December 2021

Advancing mathematics by guiding human intuition with AI

[Alex Davies](#) , [Petar Veličković](#), [Lars Buesing](#), [Sam Blackwell](#), [Daniel Zheng](#), [Nenad Tomašev](#), [Richard Tanburn](#), [Peter Battaglia](#), [Charles Blundell](#), [András Juhász](#), [Marc Lackenby](#), [Geordie Williamson](#), [Demis Hassabis](#) & [Pushmeet Kohli](#) 

Nature **600**, 70–74 (2021) | [Cite this article](#)



The signature and cusp geometry of hyperbolic knots

[Alex Davies](#), [András Juhász](#), [Marc Lackenby](#), [Nenad Tomasev](#)

Towards combinatorial invariance for Kazhdan–Lusztig polynomials

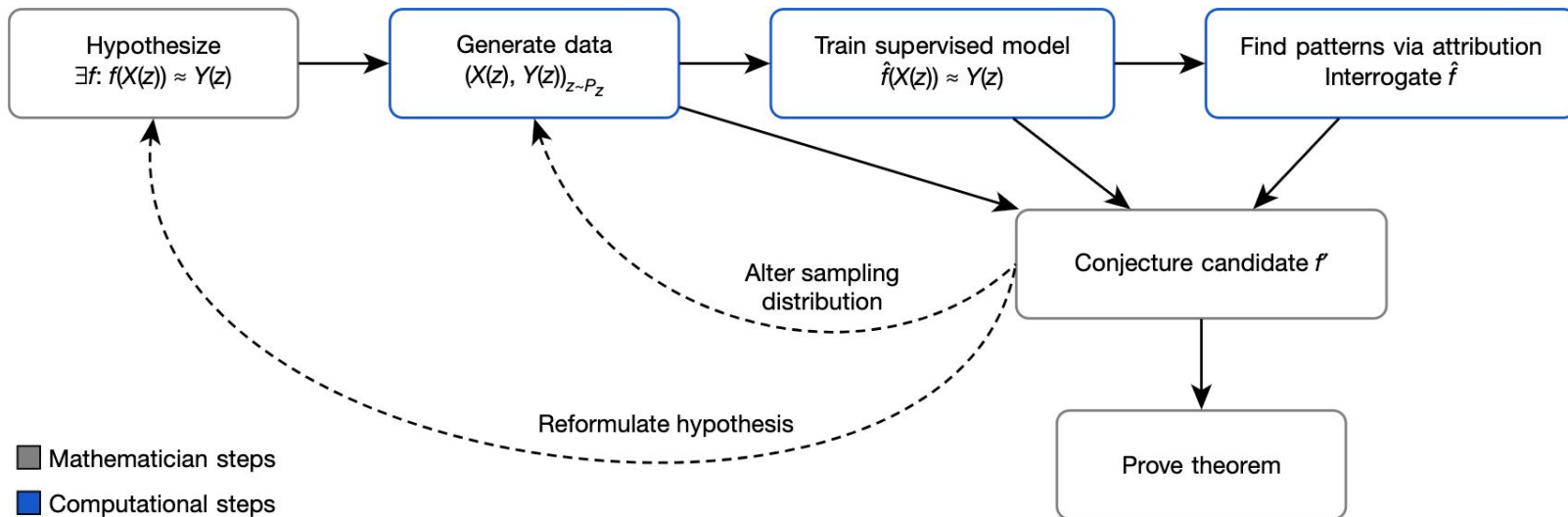
[Charles Blundell](#), [Lars Buesing](#), [Alex Davies](#), [Petar Veličković](#), [Geordie Williamson](#)

Working with mathematicians




Neural networks as feature-seeking machines.

AI scientists as AI consultants.

AI systems as an experimental mathematics tool.



Knot theory example

z: Knot	X(z): Geometric invariants				Y(z): Algebraic invariants		
	Volume	Chern–Simons	Meridional translation	...	Signature	Jones polynomial	...
	2.0299	0	i	...	0	$t^{-2} - t^{-1} + 1 - t + t^2$...
	2.8281	-0.1532	$0.7381 + 0.8831i$...	-2	$t - t^2 + 2t^3 - t^4 + t^5 - t^6$...
	3.1640	0.1560	$-0.7237 + 1.0160i$...	0	$t^{-2} - t^{-1} + 2 - 2t + t^2 - t^3 + t^4$...

The recipe

- Generate as much data as your disk allows
- Train a neural network to predict one feature from another (or a set of others)
- Maybe you find one interesting thing from millions of possible connections (use a shotgun to catch some fish)

Deep learning keeps getting deeper

What are Large Language Models (LLMs)

- Neural networks are very capable of finding patterns in data.
- A Large Language Model is:
 - Many, many floating point numbers (usually billions, can be 1-2 trillion)
 - Organised by matrix multiplications and nonlinearities
 - Trained to reduce perplexity on general text

The quick brown fox jumps over the lazy → dog

... This theorem is an apparent corollary of theorem → 2.11

LLMs can do math

Solving Quantitative Reasoning Problems with Language Models

Aitor Lewkowycz*, Anders Andreassen†, David Dohan†, Ethan Dyer†, Henryk Michalewski†,
Vinay Ramasesh†, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo,
Yuhuai Wu, Behnam Neyshabur*, Guy Gur-Ari*, and Vedant Misra*

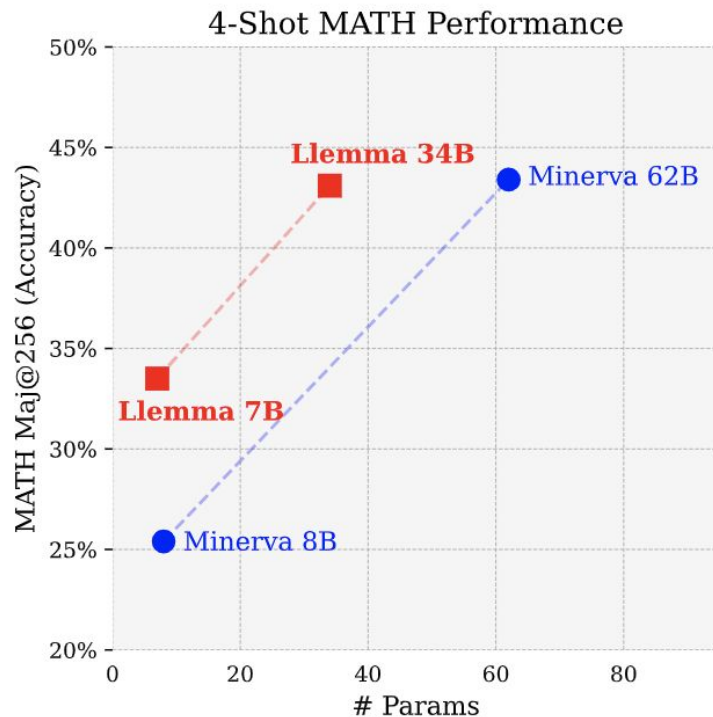
Google Research

- Above 50% on the MATH benchmark in 2022, when the prediction market median timeline was 2028
- Train on as much high-quality math data as possible (38.5B tokens)

Independent verification of this simple methodology

LLEMMA: AN OPEN LANGUAGE MODEL FOR MATHEMATICS

Zhangir Azerbayev^{1,2} Hailey Schoelkopf² Keiran Paster^{3,4}
Marco Dos Santos⁵ Stephen McAleer⁶ Albert Q. Jiang⁵ Jia Deng¹
Stella Biderman² Sean Welleck^{6,7}



Can LLMs prove theorems?

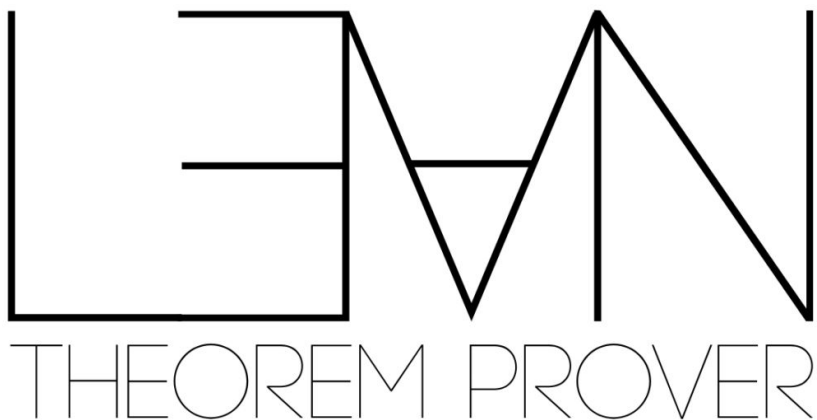
There seem to be many obstacles:

- LLMs are quite stochastic - how do you guarantee that theorems are proved?
- There simply isn't that much data of theorem proving.
- How do you even evaluate proofs to theorems?

LLMs and formal methods: a natural symbiosis

Formal methods:

- Start with axioms and deduction rules of a logic / type theory foundation.
- Build mathematics from the ground up. Approximately 2M LoC in each lang.



Generative Language Modeling for Automated Theorem Proving

Stanislas Polu
OpenAI
spolu@openai.com

Ilya Sutskever
OpenAI
ilyasu@openai.com

LISA: Language models of ISAbelle proofs

Albert Qiaochu Jiang
University of Oxford
albert594250@gmail.com

Wenda Li
University of Cambridge
wl302@cam.ac.uk

Jesse Michael Han
OpenAI
jessemichaelhan@gmail.com

Yuhuai Wu
University of Toronto
yw@cs.toronto.edu

HyperTree Proof Search for Neural Theorem Proving

Guillaume Lample^{*†} Marie-Anne Lachaux^{*†} Thibaut Lavril^{*†} Xavier Martinet^{*†}

Amaury Hayat[§] Gabriel Ebner[‡] Aurélien Rodriguez[‡] Timothée Lacroix^{*†}

Thor: Wielding Hammers to Integrate Language Models and Automated Theorem Provers

Albert Q. Jiang
University of Cambridge
qj213@cam.ac.uk

Wenda Li
University of Cambridge
wl302@cam.ac.uk

Szymon Tworkowski
University of Warsaw
szy.tworkowski@gmail.com

Konrad Czechowski
University of Warsaw
konrad.czechowski@gmail.com

Tomasz Odrzygóźdź
IDEAS NCBR
tomaszo@impan.pl

Piotr Miłoś
Polish Academy of Sciences
pmilos@mimuw.edu.pl

Yuhuai Wu
Google Research & Stanford University
yuhuai@google.com

Mateja Jamnik
University of Cambridge
mateja.jamnik@cl.cam.ac.uk

Formal Mathematics Statement Curriculum Learning

Stanislas Polu¹ Jesse Michael Han¹ Kunhao Zheng² Mantas Baksys³ Igor Babuschkin¹ Ilya Sutskever¹

The standard recipe

- Get a formal language
- Grab all of the theorems written in this language
- Let the LLM learn how to write proofs in this language
- Variants:
 - Monte-Carlo Tree Search (what AlphaGo used)
 - Integrating hammers as a usable tool
 - Expert iteration

Progress was stalling between 2023 and 2024

- Data bottleneck
 - Growing speed = human writing and reviewing speed
- Open-source base models climb but slowly
- No better way to train models than MCTS and it's hard to do

Data scarcity => Autoformalization

- Lots of informal data in arxiv papers, etc.
- If we turn them formal, we'll have lots of data.

Autoformalization with Large Language Models

Yuhuai Wu^{1,2,†} Albert Q. Jiang³ Wenda Li³
Markus N. Rabe¹ Charles Staats¹ Mateja Jamnik³ Christian Szegedy¹

DRAFT, SKETCH, AND PROVE: GUIDING FORMAL THEOREM PROVERS WITH INFORMAL PROOFS

Albert Q. Jiang^{1,2,†} Sean Welleck^{3,4,†} Jin Peng Zhou^{5,6,†}
Wenda Li² Jiacheng Liu³ Mateja Jamnik²
Timothée Lacroix¹ Yuhuai Wu^{5,7,‡} Guillaume Lample^{1,‡}

¹Meta AI ²University of Cambridge ³University of Washington ⁴Allen Institute for AI
⁵Google Research ⁶Cornell University ⁷Stanford University

Training algorithm => reasoning

- We can now reliably train models to reason in long chains of thoughts.
- Models hillclimb on tasks with verifiable results.

OpenAI o1 System Card

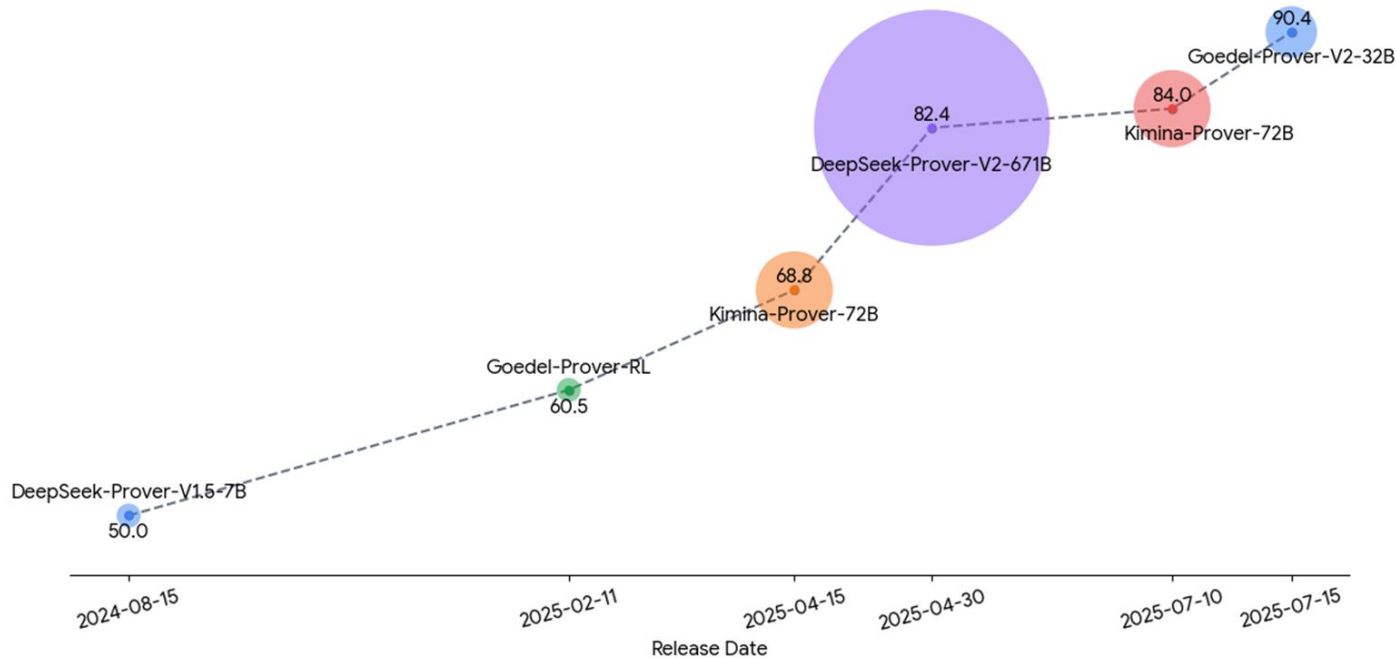
OpenAI

December 5, 2024

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

Autoformalization, reasoning, and better base models

In 2022, the state-of-the-art result on miniF2F was 41% with a 660M model.



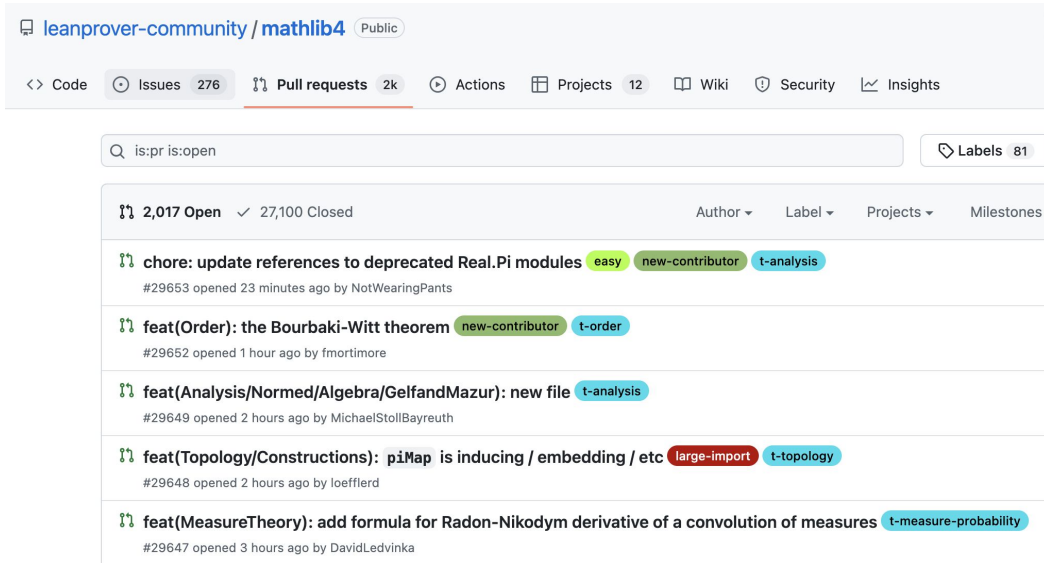
The exciting things ahead

The ingredients are ripe for new developments.

- General LLMs are very capable production tools now
- Open-source models have got seriously good

File and repo level proof engineering

How far are we from arxiv-to-proof?



leanprover-community / mathlib4 Public

<> Code Issues 276 Pull requests 2k Actions Projects 12 Wiki Security Insights

is:pr is:open Labels 81

2,017 Open 27,100 Closed Author Label Projects Milestones

- chore: update references to deprecated Real.Pi modules **easy** **new-contributor** **t-analysis**
#29653 opened 23 minutes ago by NotWearingPants
- feat(Order): the Bourbaki-Witt theorem **new-contributor** **t-order**
#29652 opened 1 hour ago by fmortimore
- feat(Analysis/Normed/Algebra/GelfandMazur): new file **t-analysis**
#29649 opened 2 hours ago by MichaelStollBayreuth
- feat(Topology/Constructions): piMap is inducing / embedding / etc **large-import** **t-topology**
#29648 opened 2 hours ago by loefflerd
- feat(MeasureTheory): add formula for Radon-Nikodym derivative of a convolution of measures **t-measure-probability**
#29647 opened 3 hours ago by DavidLedvinka

math-inc / strongpnt Public

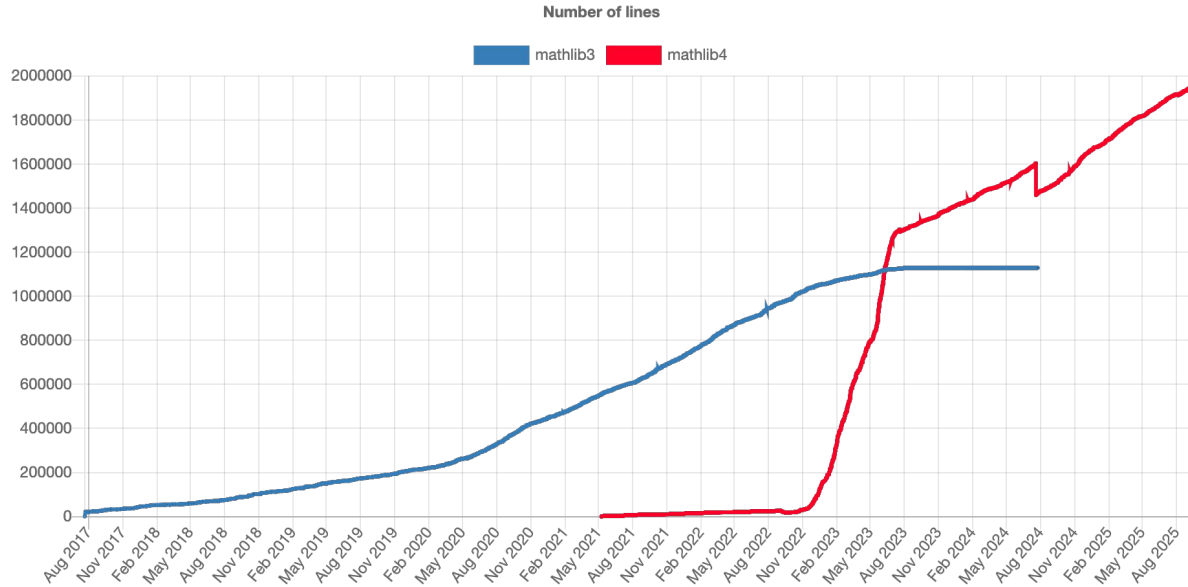
<> Code Issues 1 Pull requests

25K lines of Lean automatically generated in 3 weeks from:

- A proof in English
- A human outline

Building conjectures

What if we exhausted all the existing mathematics ever composed?



Large formal maths library

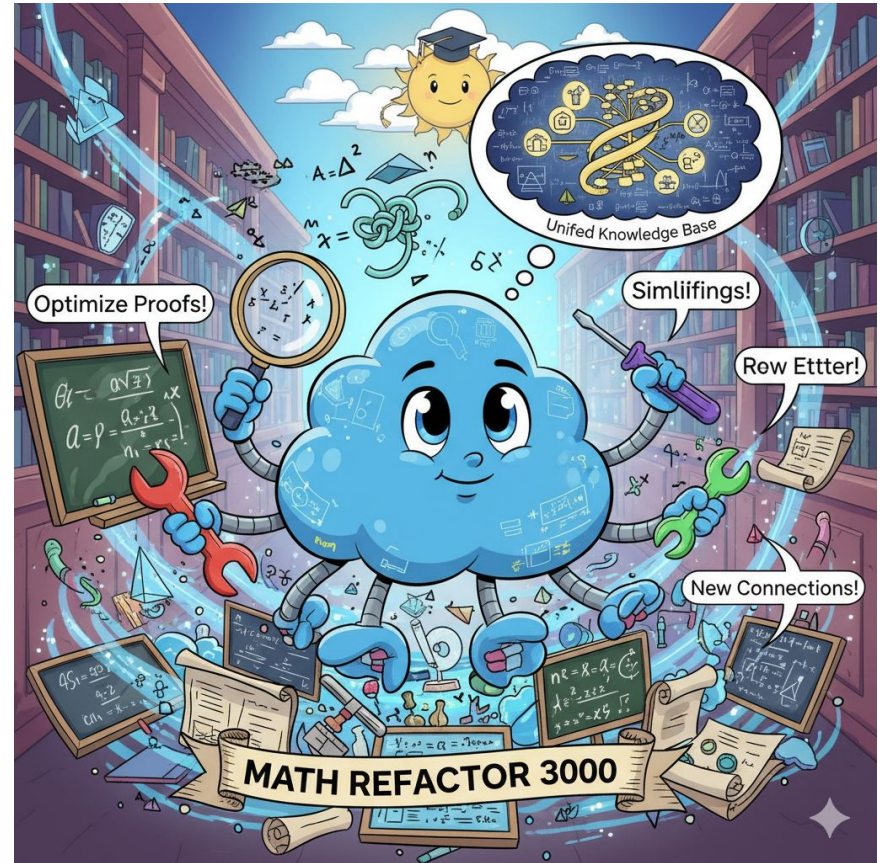


Machines aligned with human interests and preferences

Refactoring: doing math

Maybe there's something more.

Compressing mathematical library as an objective that makes machines really “do mathematics”.



Exciting times