Artificial Intelligence & Math

Oleksii Kuchaiev, Sr. Director of Applied Research @ NVIDIA

ICMU Conference, Odense | 26 August 2026

Introduction

I lead the applied research team responsible for LLM post-training at NVIDIA

Ph.D. from University of California, Irvine

M.S. in Applied Math from Taras Shevchenko National University of Kyiv

- **Mathematics is often the LLM proving ground — and lately, models have started giving back to mathematics**

- Everything in this talk is public research; opinions are my own

Agenda

Part 1: Math in AI

– What mathematics makes frontier AI work?

Part 2: AI in Math

– What can AI do in math today, and how should we use it?

Part 1: The mathematics inside frontier AI

All the math used should be understandable by math undergrad

Nothing at the core is exotic; the scale (~10^{26} FLOPs) and engineering is.

Part 1: what is a frontier AI model?

Large language model (LLM) is a parametrized conditional probability distribution

When you “talk” to ChatGPT, Gemini, Claude, Nemotron, etc. you are conditioning and sampling from distribution:

$$P_\theta(x_t | x_{t-1}, \ldots, x_0)$$

Hypothesis: best prediction forces structure and “understanding.”

$$P_\theta(mat | the, on, jumps, cat, the) \gg P_\theta(yellow | the, on, jumps, cat, the)$$

Everything you interact with – ChatGPT, Claude, Gemini, Nemotron, Qwen, etc. – is this mathematical object plus scaffolding

Part 1: one model, three stages of life

1 · PRE-TRAIN

fit the distribution

next-token prediction over ~10^{13}–10^{14} tokens of text, code, and mathematics

→

2 · POST-TRAIN

shape the behavior

curated demonstrations, then reinforcement learning — human preferences and verifiable rewards

→

3 · INFERENCE

search the model

sample, rank, verify, call tools — spend more compute per hard question

---

10^{11}–10^{12}
trainable parameters

~10^{26}
FLOPs for one frontier training run

10^{4}–10^{5}
GPUs, running for months

NVIDIA

Part 1: probability

The (multi) billion-dollar question

How do we find weights θ that make the most capable model?

Maximize the likelihood of human knowledge:

$$P_\theta(x_1, \ldots, x_T) = \prod_{t=1}^{T} P_\theta(x_t \mid x_{t-1}, \ldots, x_0)$$

Cross-entropy loss:

$$L(\theta) = -E_{x \sim \text{data}} \left[\sum_t \log P_\theta(x_t \mid x_{<t}) \right]$$

The surprise of the decade: pushed far enough, this humble objective yields modern AI.

Shannon, “A Mathematical Theory of Communication” (1948) and “Prediction and Entropy of Printed English” (1951) · “Language modeling is compression” (Delétang et al., 2023)

Part 1: linear algebra

Think of tokens (words) as vectors; context becomes geometry

Text is tokenized into ~10^5 discrete symbols; an embedding matrix turns symbols into geometry:

$$W = \begin{pmatrix} w_1^1 & \cdots & w_1^d \\ \vdots & \ddots & \vdots \\ w_{|V|}^1 & \cdots & w_{|V|}^d \end{pmatrix}$$

each row represents aa toke or a “piece” of a word as a vector in $\mathbb{R}^d$, d ≈ 2^{12}–2^{14}

The network is very deep and very repetitive — a few block types, stacked L ≈ 60–120 times:

$$x^{l+1} = x^l + \mathrm{Attn}\left(\mathrm{LN}\left(x^l\right)\right) + \mathrm{MLP}\left(\mathrm{LN}\left(x^l\right)\right)$$

- Directions carry meaning — v(king) − v(man) + v(woman) ≈ v(queen) — but dot products measure learned compatibility, not semantic truth.

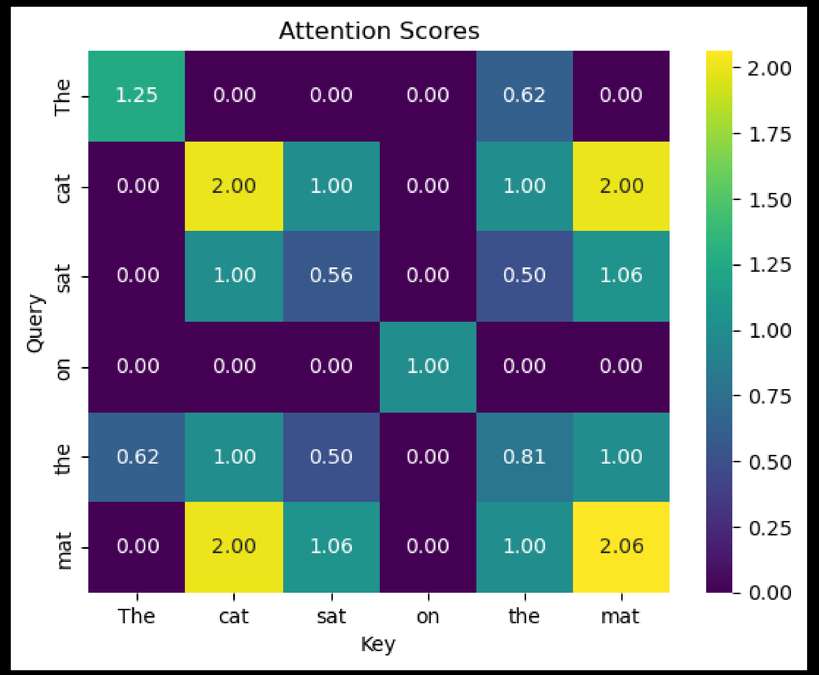
Word vectors: Mikolov et al. (2013) · Architecture: Vaswani et al. (2017)

Part 1: linear algebra — attention

The formula that replaced recurrence: a learned, data-dependent kernel

$$\text{Attn}(X) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

$Q = XW\_Q, \quad K = XW\_K, \quad V = XW\_V$ — *three learned linear maps*



- Causal masking hides the future; cost $O(T^2)$ in context length — the price of exact all-pairs communication.

Vaswani et al., "Attention Is All You Need" (2017) — still the frontier skeleton in 2026 · RoPE: Su et al. (2021)

NVIDIA

Part 1: the architecture zoo, in three formulas

Same skeleton, sparser compute

01 · DENSE TRANSFORMER

$$Y = X + Attn(X) + MLP(X)$$

every token flows through the same dense weights — simple, parallel, quadratic in context

02 · MIXTURE OF EXPERTS

$$MoE(x) = \Sigma_i \in TopK\, g_i(x) \cdot E_i(x)$$

a learned router activates a few expert MLPs per token — capacity decouples from compute

03 · STATE SPACE (MAMBA)

$$h_t = A_t h_{t-1} + B_t x_t\,, \quad y_t = C_t h_t$$

a learned linear time-varying recurrence — O(T) in length; control theory in disguise

“Frontier model” no longer means one dense architecture. **Nemotron 3 Ultra** interleaves Mamba, attention, and experts: 550B parameters total, 55B active per token, 1M-token context — pre-trained in 4-bit floating point (NVFP4).

MoE: Fedus et al. (2021) · Mamba: Gu & Dao (2023) · Nemotron 3 Ultra tech report (2026) — weights, data, and recipes open

NVIDIA

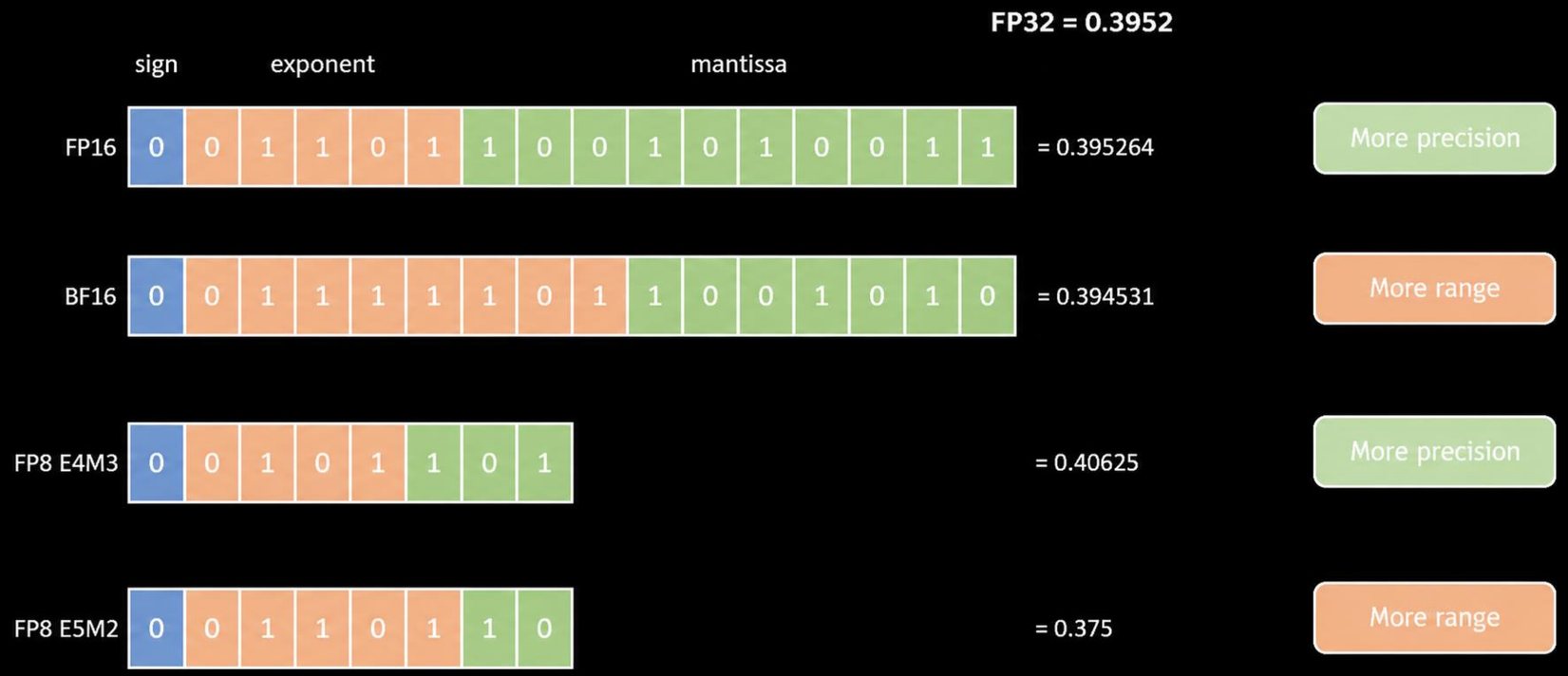
Part 1: linear algebra, at industrial scale

Numeric of computations matter in practice

~99%
of training FLOPs are matrix multiplications (GEMM)

FP8 / FP4
training precisions at the frontier — rounding-error analysis at 10^{26} FLOPs

10^4–10^5
GPUs cooperating on one distributed matrix product



Mixed precision: Micikevicius et al. (2017) · Megatron parallelism: Narayanan et al. (2021); https://developer.nvidia.com/blog/floating-point-8-an-introduction-to-efficient-lower-precision-ai-training/

Part 1: optimization

Noisy first-order methods in 10^{12} dimensions

$$g_t = \nabla_\theta L(\theta_t;\ \text{batch})$$

Backpropagation = the chain rule, organized as dynamic programming; cost ≈ 2× the forward pass

$$\theta_{t+1} = (1 - \eta\lambda)\theta_t - \eta \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon}$$

AdamW — momentum + diagonal preconditioning + decoupled weight decay: the workhorse of every frontier run

- The loss is non-convex, yet first-order methods reliably find excellent minima — why is an open problem; gradient noise itself acts as an implicit regularizer.
- Classical convergence theory does not cover the regime that works: schedules, normalization, and numerics are part of the effective algorithm, transferred across scales by empirical rules — all this lacks ”proper” math foundations.
- One run: months, no restarts — optimization as risk management (loss spikes, divergence, silent hardware faults).

Adam: Kingma & Ba (2014) · AdamW: Loshchilov & Hutter (2017)

Part 1: scaling laws

Empirical regularities — not theorems

$$L(N, D) \approx E + A \cdot N^{-\alpha} + B \cdot D^{-\beta}$$

N = parameters, D = tokens; fitted: $\alpha \approx 0.34$, $\beta \approx 0.28$; $C \approx 6ND \Rightarrow$ compute-optimal N and D grow together (~20 tokens per parameter — “Chinchilla”, 2022)

PREDICTABLE

Loss follows power laws over ~8 orders of magnitude — capability forecasting becomes curve fitting.

CONDITIONAL

Smaller + more data beats larger + undertrained; inference economics shift the optimum again.

NO GUARANTEE

A law about pre-training loss is not a theorem about truth, reasoning, or downstream ability.

This is what makes billion-dollar training runs rational: train small, fit the law, extrapolate, commit.

Deriving these exponents from first principles is an open mathematical problem.

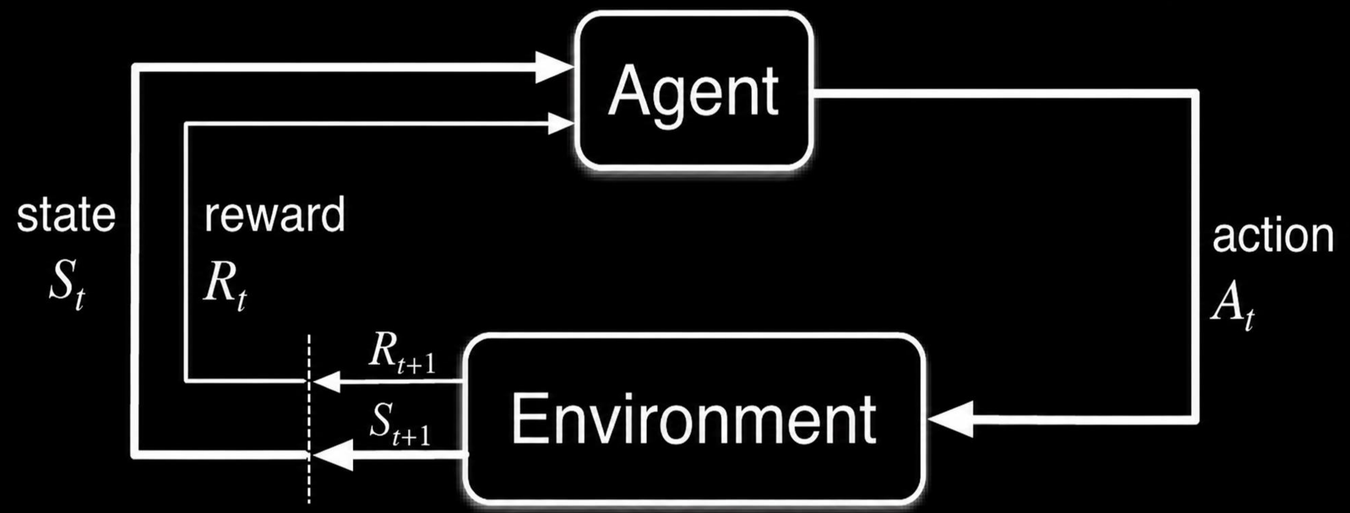
Kaplan et al. (2020) · Hoffmann et al., “Training Compute-Optimal LLMs” (2022)

Imitation learning => Exploration learning

Post-training: Reinforcement Learning

Models enter “exploration learning era”

- Pre-training: model has seen entire Internet, good initialization for θ in P_θ. **This is imitation learning**
- However, two problems:
  1. poor ”behavior”
  2. ran out of data

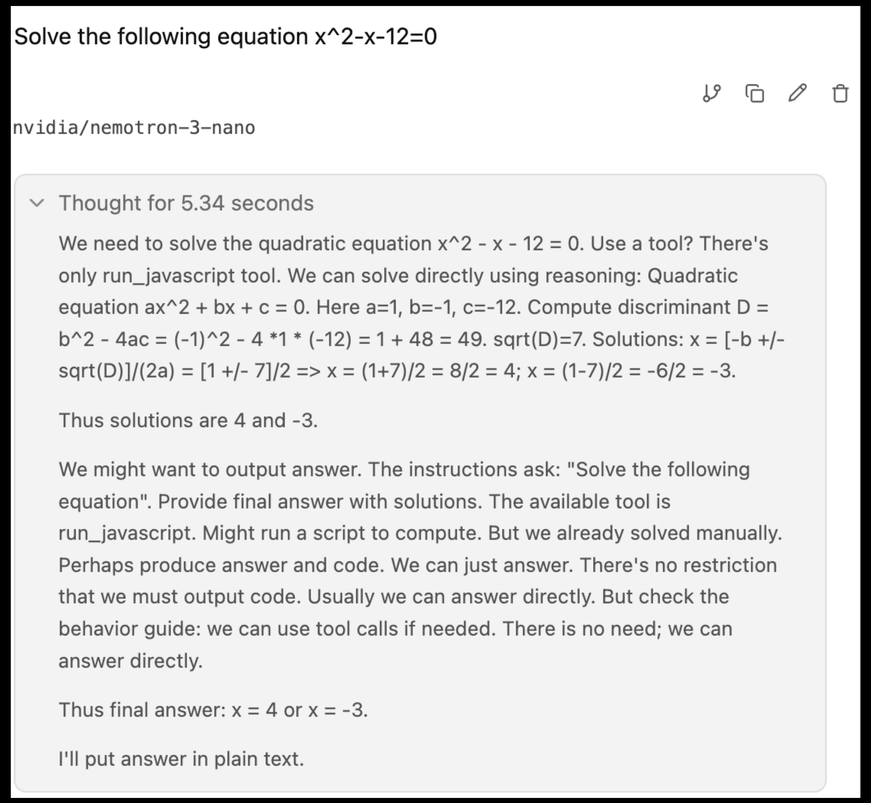


- Next (post-training), **we train model from their own experience.**

- **This is where Mathematics and Coding are particularly good “Environments”:**
  - Many things can be cheaply verified
  - Unbounded curriculum: from trivial to research grade

Part 1: inference-time compute

Chains of thoughts or ”thinking” traces



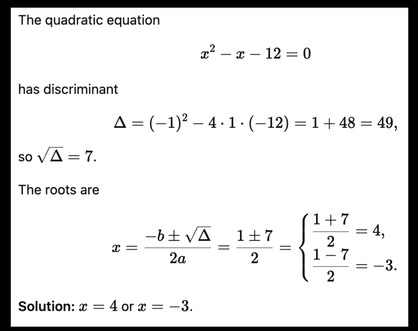
This is “thinking trace”. It is often hidden in ChatGPT/Claude. Open models (Nemotron, Qwen, etc.) expose it.

This behavior emerges during Reinforcement Learning from verifiable rewards.

Correct thinking trace maximizes likelihood of correct answer.

Amazingly, models often go on a wrong path, self-correct, step back, etc.

Inference time scaling: more compute / tokens spent on thinking => higher probability of correct answer.



This is final answer. This is a part that is often easy to verify

Part 2: AI in mathematics

Part 1 in one sentence: a differentiable probability model — scaled by data and compute, reshaped by statistical feedback, searched at inference time.

Part 2: what these systems can do in mathematics — and how to use them today

Part 2: “good at math” is not one axis

Four standards of evidence — a reading guide for every AI-math headline

Answer

EXACT MATCH

Is the final value correct? Cheap to check automatically — and saturating fast.

Informal proof

EXPERT REVIEW

Does the argument hold together? Expensive, slow, and it does not scale.

Formal proof

MACHINE CHECK

Does the kernel accept a proof term? Binary and trustworthy — for the statement as encoded.

Research contribution

COMMUNITY JUDGMENT

Novel, meaningful, understood? The slowest verifier of all — and the one that matters.

Part 2: five years, from intuition pump to coauthor

Every step: search coupled to a stronger verifier

2021	Dec 2023	Jul 2024	Jul 2025	2026
Guided intuition	FunSearch	IMO silver	IMO gold	Research frontier
ML-suggested patterns lead to human theorems in knot theory and representation theory (Nature).	First new mathematics from an LLM system: larger cap sets in dimension 8 (Nature).	AlphaProof + AlphaGeometry 2: 28/42 in formal Lean — 1 point below gold; up to 3 days per problem.	Gemini Deep Think 35/42 — first officially graded gold; natural language, contest time. Reasoning RL goes mainstream.	Perfect 42/42 at IMO 2026; decades-old conjectures disproved; autonomous agents co-author papers.

Each capability here was “decades away” five years earlier. Extrapolate with care — in both directions.

Part 2: Olympiads — the score is not the story

Read the conditions column: the scaffolding is shrinking

System	Score	Standing	Conditions
2024 · AlphaProof + AlphaGeometry 2 (Google DeepMind)	28/42	silver — 1 pt below gold	formal Lean; manual formalization; up to 3 days/problem
2025 · Gemini Deep Think (Google)	35/42	gold — first officially graded	natural language, within contest time
2025 · OpenAI experimental model	35/42	gold-level — self-reported	natural language, within contest time
2025 · best public model (MathArena)	≈13/42	below bronze (19)	independent grading of released models
2026 · Nemotron 3 Ultra (NVIDIA)	30/42	above gold cutoff (29) — graded by the IMO team	open weights; same time limit; no internet, no tools

2024: specialist formal systems and days of search → 2026: an open-weight model (Nemotron) you can download crossed the gold line under official grading.

Part 2: 2026 — Erdos problems

Decades-old conjectures are falling — and the audit is the result to remember

MAY 2026

Erdős's unit-distance conjecture (1946): disproved

An OpenAI internal model's construction — verified, then simplified and strengthened, by nine mathematicians (Alon, Gowers, Tsimerman, …).

The system used by OpenAI was remarkably similar (if not the same) as available to public. Just well trained (via RL) gigantic LLM

It considered many possibilities and tried (and succeeded) to disprove the conjecture while many people focused on proving it

Thinking trace that they published is 125 pages long!

Nine mathematicians published much more concisely written proof (~2 pages)

Part 2: 2026 — the counterexample season, audited

Decades-old conjectures are falling — and the audit is the result to remember

MAY 2026

Erdős's unit-distance conjecture (1946): disproved

An OpenAI internal model's construction — verified, then simplified and strengthened, by nine mathematicians (Alon, Gowers, Tsimerman, …).

JUL 2026

Jacobian conjecture (1939): counterexample

Found by L. Alpöge working with Claude; independently verified, Lean-formalized — called the biggest conjecture AI has played a significant role in so far.

FEB–AUG 2026

Autonomous research agents

Aletheia generated the full mathematical content of a submitted paper — and caught a flaw in a preprint that reviewers had missed. Astra: 10 long-open problems, each with a Lean certificate, ~$2,000 of compute in total.

REALITY CHECK: ONE SWEEP OF 700 “OPEN” ERDŐS PROBLEMS (GEMINI, DEC 2025)

- **700** problems marked “open,” swept autonomously
- **212** responses judged potentially correct
- **63** technically correct after expert grading
- **13** meaningfully correct — 5 seemingly novel, 8 already in the literature
- **4** new or partial resolutions (#652, #654, #1040, #1051)

50 “correct” answers were technically valid but mathematically vacuous — they solved a reading that missed Erdős's intent.

openai.com (May 2026) · Smithsonian (Jul 2026) · audit: arXiv:2601.22401 · Aletheia: arXiv:2602.10177 · xenaproject.wordpress.com

NVIDIA

Part 2: where they fail — and why

A proof-shaped answer is not a proof

FLUENT ≠ FAITHFUL ≠ VALID ≠ IMPORTANT

Wrong statement

solves a literal or altered reading of the problem — the Erdős audit's top failure mode

Wrong step

one local gap inside a convincing chain; invented lemmas — hence <5% on written proofs while acing answers

Wrong citation

sources invented, stale, or unattributed — “new to me” passed off as new

Wrong confidence

“proves” false statements ~29% of the time when prompted with them (BrokenMath); calibration error ~40–60%

Wrong self-correction

asked to revise without external feedback, accuracy can degrade — feedback must come from a verifier

Right but vacuous

technically correct, novel to no one, meaningful to nothing — correct ≠ interesting

These are corollaries of Part 1: likelihood and answer-rewards optimize plausibility, not validity. The structural fix is a verifier in the loop — yours, or Lean's.

“Proof or Bluff?”: arXiv:2503.21934 · BrokenMath: arXiv:2510.04721 · self-correction: Huang et al. (2023) · Buzzard: “one hallucination can break an entire argument… there are no guardrails”

NVIDIA

“ A proof that no human can properly explain should be viewed as incomplete, even if it has been formally verified.”

Terence Tao — “Mathematics in the Age of AI,” ICM 2026

Part 2: how to use AI

Recalibrate your AI expectations every 1 – 2 months!

ASK AI TO

- **Diverge** — many candidate strategies, constructions, counterexamples; keep what survives your check
- **Compute** — scripts, computer algebra, edge cases, simulations: the **most reliable win today**
- **Map** — a literature scout with superhuman recall; then you read the primary source
- **Formalize** — draft Lean statements and proof skeletons; fill the routine gaps
- **Attack** — “find the holes in my proof” before a referee does

KEEP HUMAN CONTROL OVER

- **The question** — definitions, assumptions, and what matters
- **The evidence** — open every source; verify every citation
- **The proof** — the statement, the gaps, the special cases; or demand Lean
- **The responsibility** — disclose material use; sign the claim yourself

> **Red line:** never upload unpublished, confidential, or referee material to a service you have not approved for it.

Henkel, “The Mathematician's Assistant” (2025): augmentation, not automation · Tao's rule of use: only where you can red-team the output

NVIDIA

Start using AI in your work!

Дякую! Так!

Oleksii Kuchaiev · okuchaiev@nvidia.com

NVIDIA


NVIDIA