When Code Becomes a Commodity: The Mathematician's Edge

Edson Cilos Vargas Júnior

Universidade Federal de Santa Catarina

2026

Contact: edson.junior@ufsc.br

Outline

When Code Becomes “Free”

Anthropic Hackathon, Feb 2026: 13k applicants

When Code Becomes “Free”

Anthropic Hackathon, Feb 2026: 13k applicants

- **1st place:** a lawyer

When Code Becomes “Free”

Anthropic Hackathon, Feb 2026: 13k applicants

- **1st place:** a lawyer
- **3rd place:** a cardiologist who built an AI care platform in 7 days

When Code Becomes “Free”

Anthropic Hackathon, Feb 2026: 13k applicants

- **1st place:** a lawyer
- **3rd place:** a cardiologist who built an AI care platform in 7 days
- Neither wrote code

Why This Time Is Different

- Printing press, typewriters, LaTeX: automated **dissemination**
- Modern AI: automates **creation**

The question "if code is becoming a commodity, what is scarce?"
only makes sense now.

Klowden & Tao, arXiv:2603.26524 (2026), §2.

When Code Becomes “Free”

- Programmer roles (execution-focused): -27.5% (2023–2025)
- Design roles (AI Engineer): $+18.7\%$ staff premium (same period)

When Code Becomes “Free”

- Programmer roles (execution-focused): -27.5% (2023–2025)
- Design roles (AI Engineer): $+18.7\%$ staff premium (same period)

If code is becoming a commodity, what is scarce?

Outline

Glossary: AI Terms

Term	Meaning
LLM	Large Language Model (Claude, GPT-5.4, Gemini 3)
Hallucination	Confident but factually incorrect model output
Chain of thought	Step-by-step natural-language reasoning before the answer
Agentic system	Multi-step AI workflow with tool use and self-critique
Frontier model	Current public SOTA: Claude Opus 4.7, GPT-5.4, Gemini 3.1 Pro
Gated model	Capability-restricted release (e.g. Claude Mythos, April 2026)
Pass@1	Correct on the first attempt (standard eval metric)
Lean 4	Proof assistant and dependently-typed language
Mathlib	Lean's library of formally verified mathematics

Glossary: Benchmarks

Benchmark	What it tests
GSM8K	Grade-school math word problems
MATH-500	High-school to olympiad problems, free-form answers
AIME	American Invitational Mathematics Exam (high school)
Putnam	US/Canadian undergraduate competition
IMO	International Mathematical Olympiad
FrontierMath T4	Research-level problems (Epoch AI)

Outline

AI at the Math Olympiad

IMO 2024: Silver

- AlphaProof + AlphaGeometry 2
- Score: **28/42** (missed gold by 1)
- Method: formal Lean 4 proofs
- Compute: **multi-day** per problem

IMO 2025: Gold*

- Gemini Deep Think
- Score: **35/42**, 5 of 6 problems
- Method: **natural language**
- Compute: within 4.5h exam
- **Officially certified** by IMO

*OpenAI: 35/42 via **self-evaluation**, not IMO-graded. DeepSeekMath-V2: gold-level reported independently.

AI at the Math Olympiad

IMO 2024: Silver

- AlphaProof + AlphaGeometry 2
- Score: **28/42** (missed gold by 1)
- Method: formal Lean 4 proofs
- Compute: **multi-day** per problem

IMO 2025: Gold*

- Gemini Deep Think
- Score: **35/42**, 5 of 6 problems
- Method: **natural language**
- Compute: within 4.5h exam
- **Officially certified** by IMO

*OpenAI: 35/42 via **self-evaluation**, not IMO-graded. DeepSeekMath-V2: gold-level reported independently.

From formal-language specialists to general-purpose reasoning in **one year**.

AlphaGeometry 2: Geometry Specialist

Google DeepMind, 2024: IMO-AG-50 (all IMO geometry problems 2000–2024)

AlphaGeometry 2: Geometry Specialist

Google DeepMind, 2024: IMO-AG-50 (all IMO geometry problems 2000–2024)

- AG2 solved **84%** (42/50), above average gold medalist (40.9)

AlphaGeometry 2: Geometry Specialist

Google DeepMind, 2024: IMO-AG-50 (all IMO geometry problems 2000–2024)

- AG2 solved **84%** (42/50), above average gold medalist (40.9)
- General-purpose LLMs (o1, Gemini Thinking, Feb 2025): **0/50**†

AlphaGeometry 2: Geometry Specialist

Google DeepMind, 2024: IMO-AG-50 (all IMO geometry problems 2000–2024)

- AG2 solved **84%** (42/50), above average gold medalist (40.9)
- General-purpose LLMs (o1, Gemini Thinking, Feb 2025): **0/50**[†]
- Beaten by specialists since:
  Seed-Geometry **43/50** (ByteDance, Jul 2025) → InternGeometry **44/50** (Shanghai AI Lab, Dec 2025)

[†]No public re-evaluation of GPT-5/5.4, Claude 4.x, or Gemini 3 on IMO-AG-50. The zero is for Feb 2025 models only.

Competition Math: Nearly Solved

Model	MATH-500	AIME 2025
Claude Opus 4.6	n/a	99.8%
GPT-5 / GPT-5.2	99.4%	100%*
DeepSeek-R1-0528	97.3%	93.1%
Gemini 3.1 Pro	95.1%	100%*

*With code execution.

MATH-500 near ceiling. AIME 2025 saturated. Plus **IMO 2025 gold**.

Competition Math: Nearly Solved

Model	MATH-500	AIME 2025
Claude Opus 4.6	n/a	99.8%
GPT-5 / GPT-5.2	99.4%	100%*
DeepSeek-R1-0528	97.3%	93.1%
Gemini 3.1 Pro	95.1%	100%*

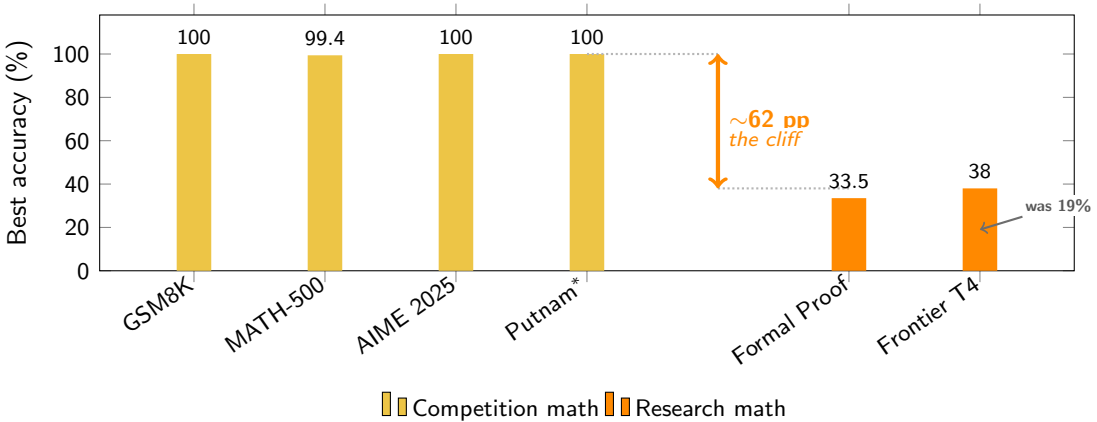
*With code execution.

MATH-500 near ceiling. AIME 2025 saturated. Plus **IMO 2025 gold**.

These benchmarks stop at **olympiad level**.

They **no longer differentiate** frontier models.

Research Math: Still Far from Solved



Putnam 2025: 12/12 formally verified (AxiomProver, Lean 4.21)

FrontierMath T4: → 38% (GPT-5.4 Pro, Mar 2026)

Where Your Expertise Matters

- Competition is **nearly** solved.

Where Your Expertise Matters

- Competition is **nearly** solved.

The cliff to research is **still there**.

That gap is where **your expertise** matters.

Where Models Fail

- **Hallucinated proof steps**
  "A single hallucinated statement can invalidate an entire derivation chain"

Where Models Fail

- **Hallucinated proof steps**
  "A single hallucinated statement can invalidate an entire derivation chain"
- **Describe-execute gap**
  Models describe algorithms correctly but fail to execute them;
  working memory collapses at 20–30 parallel branches **regardless of scale**

Where Models Fail

- **Hallucinated proof steps**
  "A single hallucinated statement can invalidate an entire derivation chain"
- **Describe-execute gap**
  Models describe algorithms correctly but fail to execute them;
  working memory collapses at 20–30 parallel branches **regardless of scale**
- **Accuracy collapse**
  Apple (2025): "complete accuracy collapse beyond certain complexities".
  Reasoning effort increases and then *declines*, as if the model gives up

The Answer-to-Proof Gap

Open Proof Corpus (arXiv, Jun 2025): 5,062 competition proofs

The Answer-to-Proof Gap

Open Proof Corpus (arXiv, Jun 2025): 5,062 competition proofs

- **6 LLMs** tested: o3, o4-mini, Gemini-2.5-Pro, Grok-3-Mini, DeepSeek-R1, Qwen3

The Answer-to-Proof Gap

Open Proof Corpus (arXiv, Jun 2025): 5,062 competition proofs

- **6 LLMs** tested: o3, o4-mini, Gemini-2.5-Pro, Grok-3-Mini, DeepSeek-R1, Qwen3
- **$\sim$30 percentage points** separate a correct answer from a correct proof

The Answer-to-Proof Gap

Open Proof Corpus (arXiv, Jun 2025): 5,062 competition proofs

- **6 LLMs** tested: o3, o4-mini, Gemini-2.5-Pro, Grok-3-Mini, DeepSeek-R1, Qwen3
- **$\sim$30 percentage points** separate a correct answer from a correct proof
- Aggregate: **43%** of proofs judged correct by 13 human graders (2,169 / 5,062)

The Gap Persists in 2026

FormalProofBench (March 2026)

Formal proofs in Lean 4 (graduate level):
best model only reaches **33.5%**.

Final-answer benchmarks: **saturated at ∼100%**.

Cliff unchanged.

Proofs Without Understanding

- AlphaProof (IMO 2024): solutions **formally verified correct**, but with redundant or inexplicable steps

Proofs Without Understanding

- AlphaProof (IMO 2024): solutions **formally verified correct**, but with redundant or inexplicable steps
- K&T call these **"odorless proofs"**: pass verification, fail the *smell test* a mathematician applies pre-verification

Proofs Without Understanding

- AlphaProof (IMO 2024): solutions **formally verified correct**, but with redundant or inexplicable steps
- K&T call these **"odorless proofs"**: pass verification, fail the *smell test* a mathematician applies pre-verification
- Thurston: good mathematics provides **a causal narrative**, not just a chain of entailment

Klowden & Tao (2026), §4.6 n.16 • Thurston (2006).

Self-Critique Blindness

The own-proof audit failure

LLMs judge proofs by *other* models more accurately than proofs by *themselves*.
Self-critique blindness $\Rightarrow$ **multi-model review pipelines** are mandatory.

- OPC Table 3: all models except Qwen3 worse on own-proof grading
- "LLMs struggle to recognize their mistakes". Critical for iterative use
- Implication: **don't trust a model to audit itself**

The Trajectory Signal

Claude Mythos Preview (April 2026)

Anthropic announces: **gated to ∼40–52 orgs** via Project Glasswing.

- USAMO 2026 (US team selection for IMO): **97.6%**‡
- **First publicly disclosed frontier model withheld from GA***

*GA (*General Availability*): release stage where the product becomes openly available to all customers, no waitlist or gated access. Withheld from GA = announced but not yet publicly released.

‡Pass@avg over 10 trials per problem, max effort + adaptive thinking, unlimited budget (Mythos system card §6.8).

Outline

The Junior Developer Crisis

Study	Data	Finding	Period
Stanford	ADP payroll	SW devs 22–25: **−20%** from peak	Aug 2025
Harvard	62M workers	Junior roles at AI firms: **−7.7%**	2025
Northeastern	Employment records	Junior/senior ratio: **−16.3%**	2025
UK market	Tech postings	Entry-level: **−46%**	2024
NY Fed	National jobless	CS grads: **∼6.1%**	Q1 2026

- 54% of engineering leaders expect fewer junior hires long-term

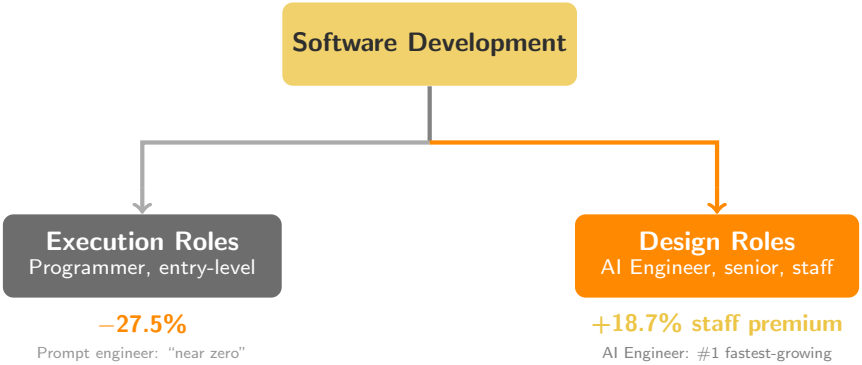
Counter-evidence: Fed Reserve (Mar 2026, 1M+ firms) found **no significant AI-to-hiring link**; mechanism contested.

Companies Reducing Engineering Teams

Company	Action	Date
Salesforce	Zero new engineers hired (FY2026)	Jan 2026
Shopify	Must prove AI can't do it first	Apr 2025
Microsoft	$\sim$6,000 layoffs; 20–30% AI code	May 2025
Amazon	14,000 corporate cuts; AI cited	Oct 2025
Klarna	Attrition 5.5k$\rightarrow$3k; **partial reversal**	2023–2025

Attribution caveat: companies have not published rigorous before/after output comparisons. CEO claims $\neq$ audited evidence.

Bifurcation, Not Extinction



BLS* still projects **15% growth** for software developers through 2034.
The aggregate hides the compositional shift.

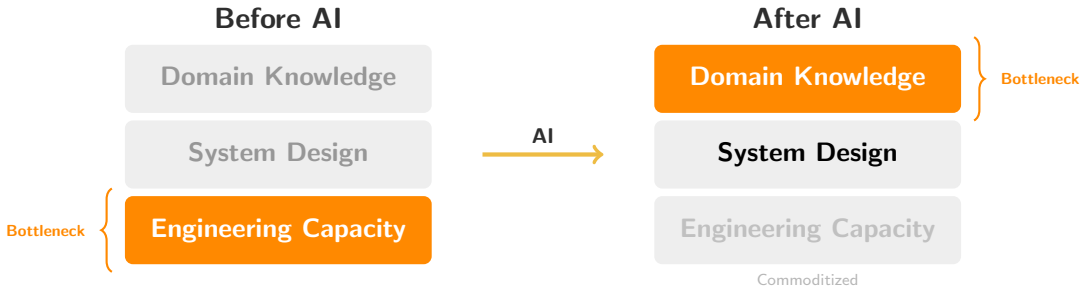
*BLS: *U.S. Bureau of Labor Statistics*, the US federal labor-statistics agency.

The Real Bottleneck

"The core problem now is knowing what the hell you actually want to build."

paraphrased from Boris Cherny, Head of Claude Code

(Lenny's Podcast, Feb 2026)



Where Mathematical Training Holds Value

Domain	**Why math matters**
AI Safety (Alignment)	Reward hacking, training dynamics
Formal Verif.	Lean proofs, AI verification
ML Theory	Interpretability, generalization
Quant Finance	Stochastic models + ML
Operations Research	**Formulate the problem** (LP/MIP)

- These are domains where **rigor and proof cannot be automated**
- AI amplifies the mathematician; it does not substitute for judgment

The Smell Test

Experienced mathematicians detect bad arguments **before** checking them line-by-line. A skill current LLMs lack by construction.

It is what you already have.

Aaronson's "Ten Signs" $\rightarrow$ your pre-verification instinct.

Aaronson (2008) • Klowden & Tao (2026), §4.2.

Outline

Tao × DeepMind (Nov 2025)

Terence Tao (UCLA, Fields Medal 2006) in collaboration with Google DeepMind

- Tao *selects and evaluates* the problems; DeepMind deploys AlphaEvolve + Gemini Deep Think

Tao × DeepMind (Nov 2025)

Terence Tao (UCLA, Fields Medal 2006) in collaboration with Google DeepMind

- Tao *selects and evaluates* the problems; DeepMind deploys AlphaEvolve + Gemini Deep Think
- **67 open problems** tested across combinatorics, analysis, and number theory

Tao × DeepMind (Nov 2025)

Terence Tao (UCLA, Fields Medal 2006) in collaboration with Google DeepMind

- Tao *selects and evaluates* the problems; DeepMind deploys AlphaEvolve + Gemini Deep Think
- **67 open problems** tested across combinatorics, analysis, and number theory
- **∼20 new results**, surpassing the best bounds previously known

Tao × DeepMind (Nov 2025)

Terence Tao (UCLA, Fields Medal 2006) in collaboration with Google DeepMind

- Tao *selects and evaluates* the problems; DeepMind deploys AlphaEvolve + Gemini Deep Think
- **67 open problems** tested across combinatorics, analysis, and number theory
- **∼20 new results**, surpassing the best bounds previously known
- A subset of the proofs passed **formal verification in Lean 4**

Source: Tao, *Mathematical exploration and discovery at scale* (personal blog, 2025-11-05).

Ernest Ryu (Seoul National University; formerly UCLA)

- Researcher in convex optimization: **operator splitting** and first-order methods

Ernest Ryu (Seoul National University; formerly UCLA)

- Researcher in convex optimization: **operator splitting** and first-order methods
- **$\sim$40-year** open problem on convergence rates

Ryu (SNU, Aug 2025): Convex Optimization

Ernest Ryu (Seoul National University; formerly UCLA)

- Researcher in convex optimization: **operator splitting** and first-order methods
- **$\sim$40-year** open problem on convergence rates
- Used GPT-5 as an assistant $\rightarrow$ **$\sim$50% improvement** over the previously known bound

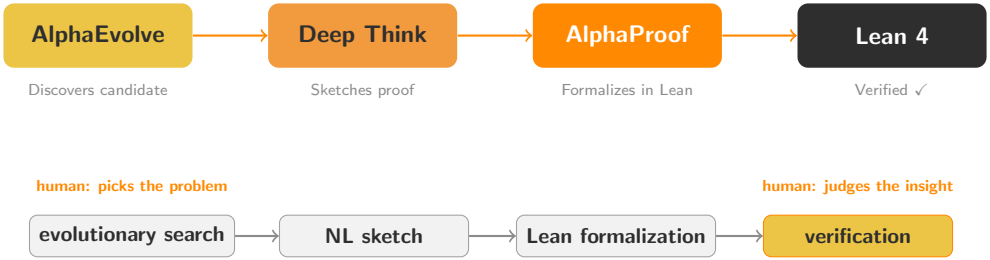
Ryu (SNU, Aug 2025): Convex Optimization

Ernest Ryu (Seoul National University; formerly UCLA)

- Researcher in convex optimization: **operator splitting** and first-order methods
- **$\sim$40-year** open problem on convergence rates
- Used GPT-5 as an assistant $\rightarrow$ **$\sim$50% improvement** over the previously known bound
- **Verified and published**: a rigorous result, not just a heuristic

Source: OpenAI, *GPT-5 mathematical discovery* (official announcement, August 2025); technical note published by E. Ryu (SNU).

The pipeline (Tao $\times$ DeepMind) that produced $\sim$*20 new results*:



Building Tools Without Coding

Andrej Karpathy's Software 3.0 (ex-OpenAI / ex-Tesla AI; Jun 2025)

- **1.0:** code humans write

Building Tools Without Coding

Andrej Karpathy's Software 3.0 (ex-OpenAI / ex-Tesla AI; Jun 2025)

- **1.0:** code humans write
- **2.0:** models humans train

Building Tools Without Coding

Andrej Karpathy's Software 3.0 (ex-OpenAI / ex-Tesla AI; Jun 2025)

- **1.0:** code humans write
- **2.0:** models humans train
- **3.0:** programs humans **describe in natural language**

Non-Developers Shipping Software

- Claude Code, Cursor: describe what you want, get working software

Non-Developers Shipping Software

- Claude Code, Cursor: describe what you want, get working software
- Non-developer at Arkance: built an AI web app solo in **48 hours** (Cursor, AU 2025)

Non-Developers Shipping Software

- Claude Code, Cursor: describe what you want, get working software
- Non-developer at Arkance: built an AI web app solo in **48 hours** (Cursor, AU 2025)
- Non-developers shipping live apps with databases, billing, APIs

Your expertise is the new programming language.

How to Use AI Today

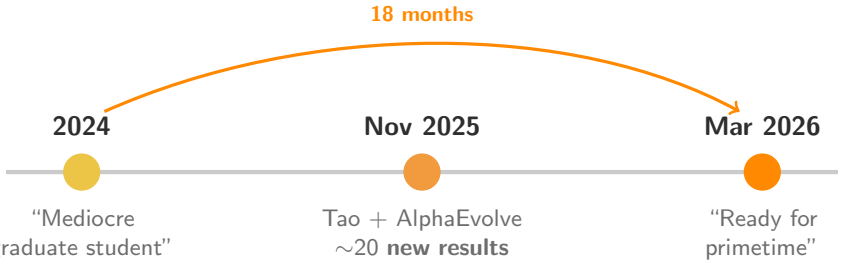
For frontier research, at the current stage of AI:

- **Red team, not blue team.**
  Let AI verify your work, not create what you cannot check.
- **Smell test first.**
  Pre-verification skepticism is your edge.

Klowden & Tao (2026), §4.2, §6.2.

Outline

From “Mediocre Grad Student” to “Ready for Primetime”



That is **18 months**.

The Mathematician's Edge

Klowden & Tao: AI is safe as a **red team** (reviewing, finding errors); but unsafe as a **blue team** (creating) without someone able to verify.

And even inside the red team, the **mathematician's smell test**, which catches bad arguments *before* line-by-line checking, is a skill LLMs lack by construction.

That is your **edge at the frontier of knowledge**.

Klowden & Tao (2026), §4.2, §6.2.

You Are the Protagonist

The best models available when you finish your PhD
don't exist yet.

You Are the Protagonist

The best models available when you finish your PhD
don't exist yet.

But they will still need someone who can tell
a valid proof from a hallucinated one.

That is you.

Thank You

Thank you. Questions?

edson.junior@ufsc.br