PRINCETON MATH CAMP FOR PHD STUDENTS · AUGUST 2026

AI for Math & Economic Model Building

Fedor Sandomirskiy

Princeton University · August 2026

ONE MONTH LATER

A new incarnation of the July deck

- Pietro and I used the previous deck at the Markus Academy minicourse one month ago
- Some takes are already outdated

Pietro Ortoleva
Princeton

Fedor Sandomirskiy
Princeton

AI changes too fast for a stable slide deck

The July minicourse deck · Markus Academy

INTRODUCTION

What this is about, and who it is for

WHAT THIS IS ABOUT

- AI for mathematics (and heavy thinking more generally), focused on what economists use and on applications to economic model building
- **Especially useful for economic theorists**

WHAT WE LEAVE ASIDE

- Applications to empirical research, coding, data collection, and data processing
- Covered by Paul Goldsmith-Pinkham on Markus Academy: Claude Code for Applied Economists

Paul Goldsmith-Pinkham, Claude Code for Applied Economists

INTRODUCTION

Five parts, built around broad questions

1. What can AI now do in mathematics?
2. What does this mean for economists?
3. Which model, plan, and environment?
4. Prompt and context engineering
5. Other AI tools

PART 1 · MATHEMATICS AT THE FRONTIER

AI in math: state of the art and the evolution over the past few years

1

Can AI come up with new insights in math?

- **Yes, with limitations**
- Trajectory: high-school mathematics to frontier results in about three years
- The public story is noisier than the trend

CAUTION

The ten Erdős problems episode: October 2025

- OpenAI claimed a model solved ten open Erdős problems; the claim collapsed within a day
- Thomas Bloom traced what the model did: it found existing references and presented them as its own

Paul Erdős

Thomas Bloom

POOR AT ATTRIBUTION

- Cannot reliably distinguish its own ideas from training data

SUPERB AT AGGREGATION

- No lemma from an appendix gets lost

Thomas Bloom on X · TechCrunch, October 2025

Real progress, carefully catalogued in Spring 2026

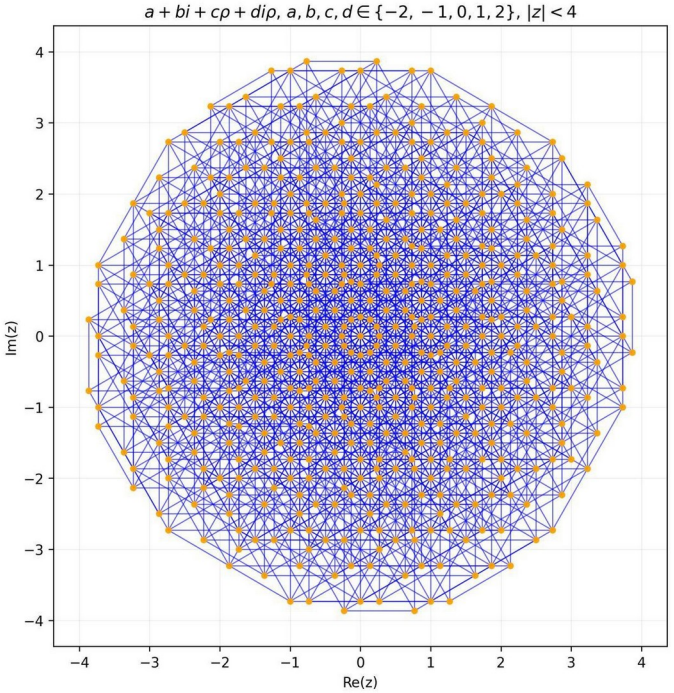
- Terence Tao curates AI contributions to Erdős problems
- **17 problems fully solved; five are machine-checked in Lean**
- Mostly specialized pipelines, no new theories: strong evidence on proof production, weaker evidence for deep novel insights
- **Last edited 30 June 2026: the ledger has stopped keeping pace with the results...**

Terence Tao

Terence Tao’s AI-contributions ledger

The Erdős unit-distance conjecture falls

- Erdős problem 90, asked in 1946: among n planar points, how many pairs can be exactly unit distance apart?
- Conjecture: grid-like constructions are asymptotically optimal
- Settled by an internal OpenAI model from a single prompt
- **The conjecture was false: the model found a construction that beats the grid**



The construction that beats the grid.
Figure: A. Lozano-Robledo

OpenAI announcement, 20 May 2026 · Prompt and proof · arXiv:2605.20695

Reception

"an outstanding achievement, settling a long-standing open problem ... it is hard to overestimate the full potential impact of this change."

Noga Alon, Princeton

"I would accept it for any journal without hesitation. I actually briefly worked on this problem and tried to make a counterexample, but failed to make progress."

Jacob Tsimerman, Toronto

No new theory: known tools combined in a way specialists did not expect to work

Fable refutes the Jacobian conjecture

- Since 1939: a polynomial map $\mathbb{C}^n \to \mathbb{C}^n$ with constant nonzero Jacobian must be one-to-one
- Number 16 on Smale's 1998 list, beside Riemann and P versus NP
- Levent Alpöge asked Fable 5; an explicit counterexample in three complex dimensions came back the same evening

$$F(x, y, z) = \begin{pmatrix} (1+xy)^3 z + y^2(1+xy)(4+3xy), \\ y + 3x(1+xy)^2 z + 3xy^2(4+3xy), \\ 2x - 3x^2 y - x^3 z \end{pmatrix}$$

$\det DF \equiv -2$ everywhere, yet three points share the image $(-1/4, 0, 0)$

Hard to find, trivial to check

Announcement · Follow-up manuscript

Cycle double cover: proved by Sol Ultra

- Since 1973: every bridgeless graph carries cycles covering each edge exactly twice
- **Proved by GPT-5.6 Sol Ultra in Codex**
- **Up to 64 subagents, told to work “at least 8 hours”; result returned in under one hour**
- Complete prompt public: a masterclass in managing an agent army, dissected in Part 4
- Thomas Bloom: “a very nice proof,” with a caveat that central 1983 ideas appear without citation

Complete prompt · Proof

JULY 20 TO AUGUST 20, 2026

Then another month happened

- OpenAI released ten selected research advances from Astra
- Fable made progress on the Riemann hypothesis
- **A phase transition: lots and lots of conjectures cracked**
- And now from amateur mathematicians on ordinary subscriptions (Crouzeix episode to be discussed)

OpenAI’s ten advances

1. High-dimensional sphere packing
2. Binary and spherical codes
3. Construction of non-sofic groups
4. Counterexample to Connes rigidity
5. Arithmetic-circuit lower bounds for the permanent
6. Quantum parallel repetition
7. Closest-vector approximation hardness
8. Ehrhart volume conjecture
9. Multicolor Ramsey numbers, resolving Erdős problem 183
10. Extremal graph conjectures, resolving Erdős problems 146 and 180

HOW IT WAS DONE

- Internal Astra model · Lean certificates · roughly $2,000 in solution-finding tokens at Sol API rates

OpenAI: Ten advances in mathematics and theoretical computer science

Lean 4, and the Collatz episode

- Lean is a programming language for formal proof verification: it checks proofs step by step. No informal gaps: “the remaining cases are analogous” does not compile
- Lean verification is a way to convince the public that the proof is 100% correct
- **Or it was, until the recent episode**

THE COLLATZ EPISODE

- July 2026: it was announced that the Collatz conjecture is disproved by AI
- The “proof” passed Lean 4 and the independent checker Nanoda by exploiting two separate soundness bugs; the claim was false
- Caught within days; both bugs patched

“This is going to keep happening. AIs are really good at exploiting soundness bugs in the kernels.”

Leonardo de Moura, creator of Lean

The announcement, Elliot Glazer · Kernel bug postmortem, de Moura · Paulson: Why is it all in the kernel?

PROGRESS ON THE RIEMANN HYPOTHESIS

Fable and the zeta function

41.6%

previous lower bound on the share of zeros on the critical line

→

67.2%

Fable-assisted new bound

- About 60 subagents and 31 million output tokens, more than $2,000 in API costs
- Directed by an Anthropic employee who is not a mathematician: he encouraged the model to believe in itself with a short prompt, then went for a run

Anthropic: Learning more about Claude’s mathematical capabilities

Crouzeix: democratization illustrated

- A conjecture on the norm of functions of matrices, open for more than two decades
- **Resolved by Shanmu Jin, a neurosurgery resident with little mathematical training**
- Roughly 16 autonomous hours in ChatGPT Work
- Prompt adapted from OpenAI’s cycle-double-cover setup

Public research repository

SELECTED; NO DETAILS

Other resolutions in the past month

- Schubitope Ehrhart-positivity conjecture: counterexample
- Samuels' conjecture; Feige's conjecture as a corollary
- Talagrand's convolution conjecture
- Stein's dimension-free weak-(1,1) Riesz-transform problem
- Albertson-Berman induced-forest conjecture: counterexamples
- Banach's isometric conjecture over the complex field
- **Many, many more**

Schubitopes · Samuels and Feige · Talagrand · Riesz transforms · Albertson-Berman · Banach

From frontier laboratories to laptops

BEFORE

- Internal models, special pipelines, frontier-lab researchers

NOW

- Amateur mathematicians, subscription users, ordinary project folders, reusable prompts

- **Public models now generate serious proofs and counterexamples**

My past month

CONJECTURES FROM MATHEMATICIAN FRIENDS

- Tens of them cracked; some had resisted sustained work for a while
- Every case reached a decisive answer within roughly 15 hours
- Sol 5.6 Ultra, instructed to work until the conjecture is resolved and for at least X hours: a version of the cycle-double-cover prompt

CONJECTURES FROM MY OWN WORK

- Several resolved; both counterexamples and insightful proofs, including those we thought about for more than a year
- A counterexample in dimension 9: a convex set with exactly the funny property we believed was impossible
- While preparing for this talk: the main result of "On the Origin of the Boltzmann Distribution" (with Omer Tamuz) is in dimension 1; we conjectured any dimension, but our technique did not generalize. Resolved in 4.5 hours

We come back to the nuances of the setup and prompts later today

A phase transition

Proof production is becoming automated enough that writing a proof is gradually becoming less of what mathematics is about

- Current frontier AI can already produce much of the proof output attainable by many working mathematicians
- **Focus shift: from proof production to proof understanding, building theories, inventing new concepts and programs**

Optimistic metaphor 1: mathematical experiments

EXPERIMENTAL SCIENCES

- In physics or experimental economics, experiments cost money, equipment, coordination, and time

MATH BEFORE AI

- Cheap in money; expensive in expert months or years

MATH NOW

- For many concrete conjectures, a reliable verdict overnight

Mathematical experimentation was cheap and has become radically cheaper

Optimistic metaphor 2: exploring uncharted terrain

BEFORE

- Walk the expedition yourself: you find the pass, the false summit, the ravine, and the dead end

NOW

- Send a group of agents; read the expedition report the next morning

TRADEOFF

- Less romantic, much faster, vastly more terrain covered

You still choose where to send the expedition, and you still read the map

THE BOTTLENECK MOVES

What becomes scarce

INCREASINGLY ABUNDANT

- proof search
- counterexample construction
- routine generalization
- formal verification

STILL SCARCE

- the right question
- the right concept
- taste
- explanation
- a research program or new theory

PART 2 · CORE USES

Core uses of AI for economic model building

2

Mathematicians versus economists

FOR MATHEMATICIANS

- A proof of a famous conjecture can be a goal in itself

FOR ECONOMISTS

- Math is a modeling tool for an idea about an economic phenomenon
- The statement, the assumptions, and even the model are negotiable

We do not need to prove this exact theorem; we need the right theorem for the economic idea

PART 2 · BOTTOM LINE

What the frontier results mean for economists

JUNE

- One frontier result, with caveats

JULY

- Three frontier results, two labs, subscription products

AUGUST

- Batches of results and repeatable public workflows

Why economists may gain even more than mathematicians

1. Lower mathematical bar
2. Conjectures are negotiable — adjust assumptions until the statement is true and interesting

Economic model building as a fixed point

An economic model requires searching for a fixed point: defensible assumptions, interesting statements, and proofs that certify them

Fixed-point framing: Guillaume Haeringer (CUNY)

WHERE THE FRONTIER IS

Concrete versus open-ended tasks

AI IS VERY GOOD NOW

- prove or disprove a precise claim
- find the smallest counterexample
- identify the active assumption
- repair a failed theorem

IMPROVING, BUT NOT THERE YET

- ask an interesting question
- build a new theory from a vague intuition
- decide what economists will find important

SEQUENTIAL ADJUSTMENT

The model-building workflow

1. State the economic intuition
2. Choose primitives, timing, and solution concept
3. Formulate a candidate theorem
4. Ask for proof or counterexample
5. Change one assumption or one conclusion
6. Repeat until the fixed point is interesting
7. Simplify, extend, and explain

AI is most useful for the highlighted steps today, but it can help with the others too

THE QUESTION TO ASK

The key fixed-point question

> *“Is this claim true without this assumption, or with an alternative assumption? Prove it or construct a counterexample.”*

- Concrete
- Falsifiable
- Proof and counterexample can run in parallel
- **Exactly where current AI is strongest**

Proofs

- **Frontier models can generate reliable proofs**
- Non-frontier models are less reliable: decompose large tasks into smaller ones

Other modes of AI used for proofs:

ATTACK

- Hostile-referee checks for weak steps

REPAIR

- Assumptions that rescue a false statement

From intuition to primitives

1. **Primitives** states, signals, actions, preferences
2. **Timing** what is known, chosen, observed
3. **Solution concept** Bayesian, maxmin, rational inattention, dynamic
4. **Predictions** comparative statics, tests, examples

PROMPT

- “Give me three minimal models that capture this intuition. For each, say what is elegant, what is fragile, and what theorem would be worth proving.”

SPIN OFF VARIANTS

- Change one primitive at a time. Which comparative static flips?

AI is not that good at this yet, but it still gives food for thought

Extensions, microfoundations, simplifications

Once the machinery exists, ask what else it can do and what it does not need

MORE GENERAL · EXTENSIONS

- Change information, preferences, timing, menu structure, equilibrium concept. Which comparative statics survive; which flip?

DEEPER · MICROFOUNDATIONS

- Can the reduced-form assumption come from search, attention, ambiguity, private signals, or institutions?

CLEANER · SIMPLIFICATIONS

- Can you do it with two types and two periods? Binary states and two actions?

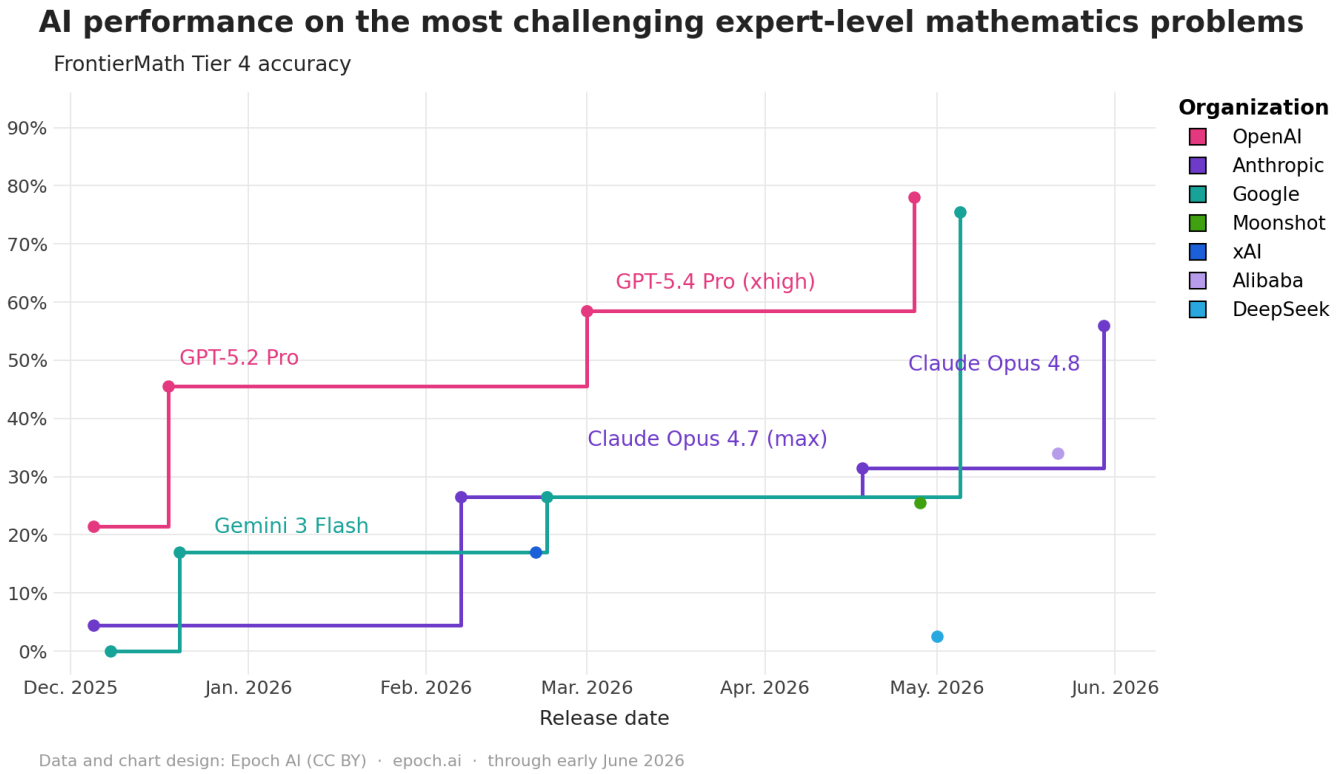
PROMPT

- “Strip the model to the smallest example that still delivers the main prediction. Then list every assumption that survives and what breaks when it is removed.”

PART 3 · MODELS AND ENVIRONMENTS

Which model, plan, and environment?

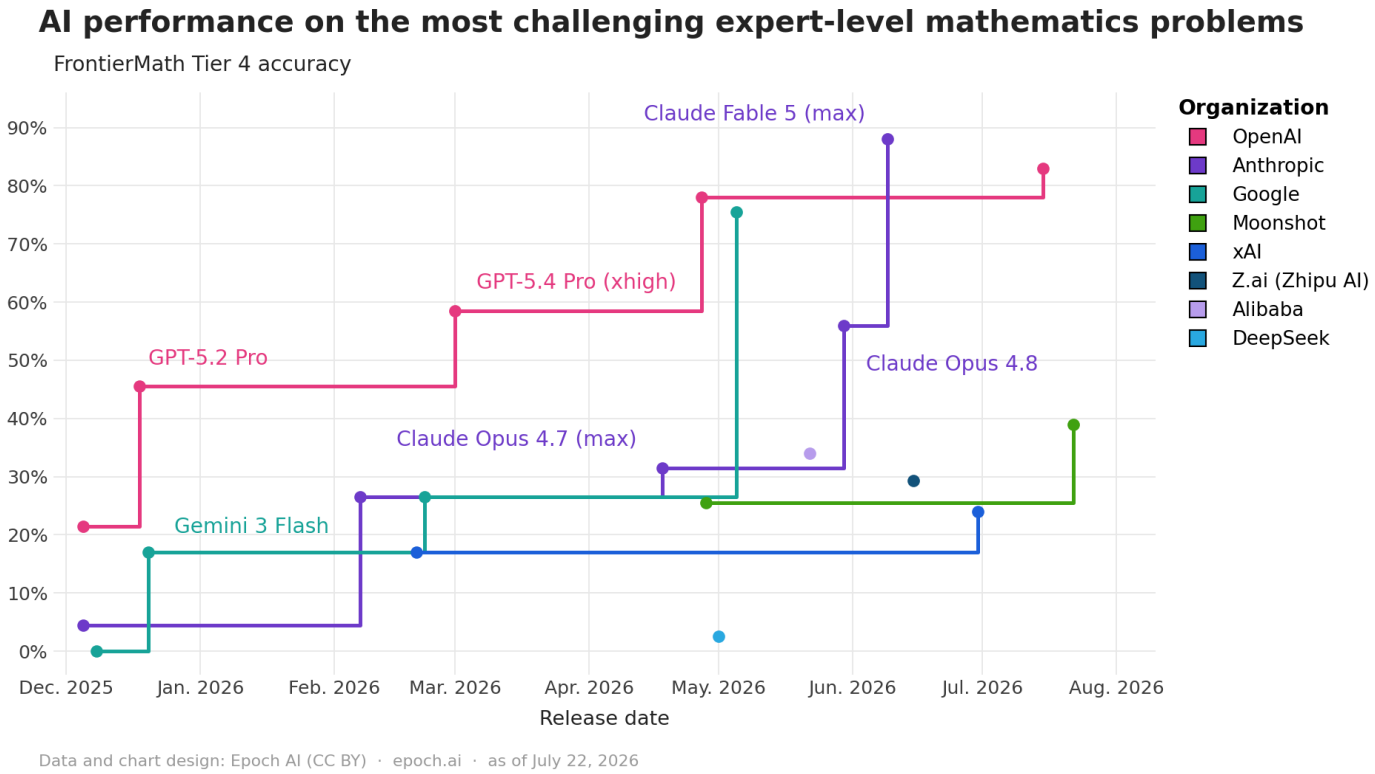
The pre-Fable era, in one picture



- Pink staircase on top all spring: 5.2 Pro to 5.4 Pro to 5.5. My June advice: ChatGPT Pro
- Claude: best writer and coder, not yet the mathematics leader
- Gemini Pro: dominated by both, but free in AI Studio
- The green spike: an internal Google setup, not a public product

JUNE 9 · ANTHROPIC

Fable 5: the purple spike



Data and chart design: Epoch AI (CC BY) · epoch.ai · as of July 22, 2026

- FrontierMath Tier 4: roughly 88, the highest point at release
- Deep intuitions; born agentic through Cowork and Claude Code
- The catch is scarcity: offline June 12 to July 1, four access regimes in five weeks
- Low limits even in the highest tier subscription, and a high API price
- **On July 9, GPT-5.6 Sol released**

JULY 9 · OPENAI

GPT-5.6: Sol on the web, an agent army in Codex

5.6 PRO IN THE BROWSER

- The 5.5 Pro successor for heavy chat thinking; FrontierMath Tier 4 around 83: less than Fable, but within the standard error from it

SOL ULTRA IN CHATGPT WORK AND CODEX

- Up to X cooperative subagents: X = 64 in the internal model; X = 4 in the public one according to the documentation; X = 10 in my experiments. The cycle-double-cover setup

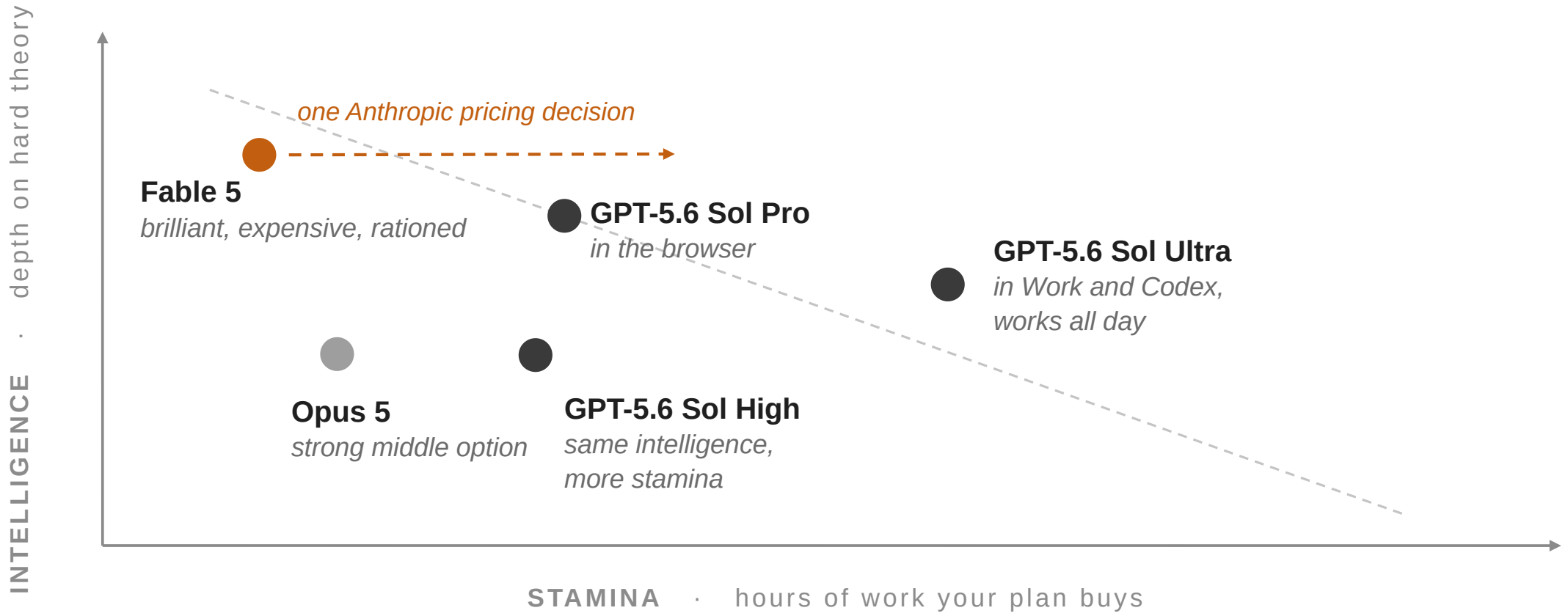
LIMITS AND PRICE

- Far more generous limits than Fable

GPT-5.6

MY MAP, AUGUST 2026

Intelligence and stamina are substitutes



A slightly shallower model that works all day can beat a genius available for twenty minutes

MY CURRENT VERDICT

Which model I would pick today

DEEPEST SINGLE ATTEMPT	Fable 5, perhaps narrowly, but not worth the money given the small limits
BEST SUSTAINED WORKFLOW	GPT-5.6 Sol and Sol Pro: more generous limits, though Pro and Ultra need the $100 or $200 tier
BEST VALUE AT $20	ChatGPT Plus: the browser, plus Codex and Work for stamina
BEST FREE CANDIDATE	Gemini 3.1 Pro in AI Studio, for non-confidential work

For math and econ theory, always make sure you use the highest reasoning setting your plan permits

AUGUST 20, 2026

OpenAI plans

FREE

- Unlimited text chats with GPT-5.6 Luna, subject to guardrails
- Think uses Luna, not Sol
- Terra in Codex; limited Work and tools

PLUS · $20

- GPT-5.6 Sol at Medium and High
- Expanded Work and Codex
- No Extra High or Sol Pro

PRO · $100 OR $200

- 5x or 20x Plus usage
- Extra High and GPT-5.6 Sol Pro
- Maximum Work and Codex capacity

My August: the $200 tier funded roughly 100 hours of Sol Ultra swarm work per week

ChatGPT pricing · GPT-5.6 availability

AUGUST 20, 2026

Anthropic plans

FREE

- Chat on web, mobile, and desktop
- Web search, memory, file creation
- No Claude Code, Cowork, Design, or Science
- Pro buys at least 5x the usage per 5-hour session

PRO · $20 MONTHLY

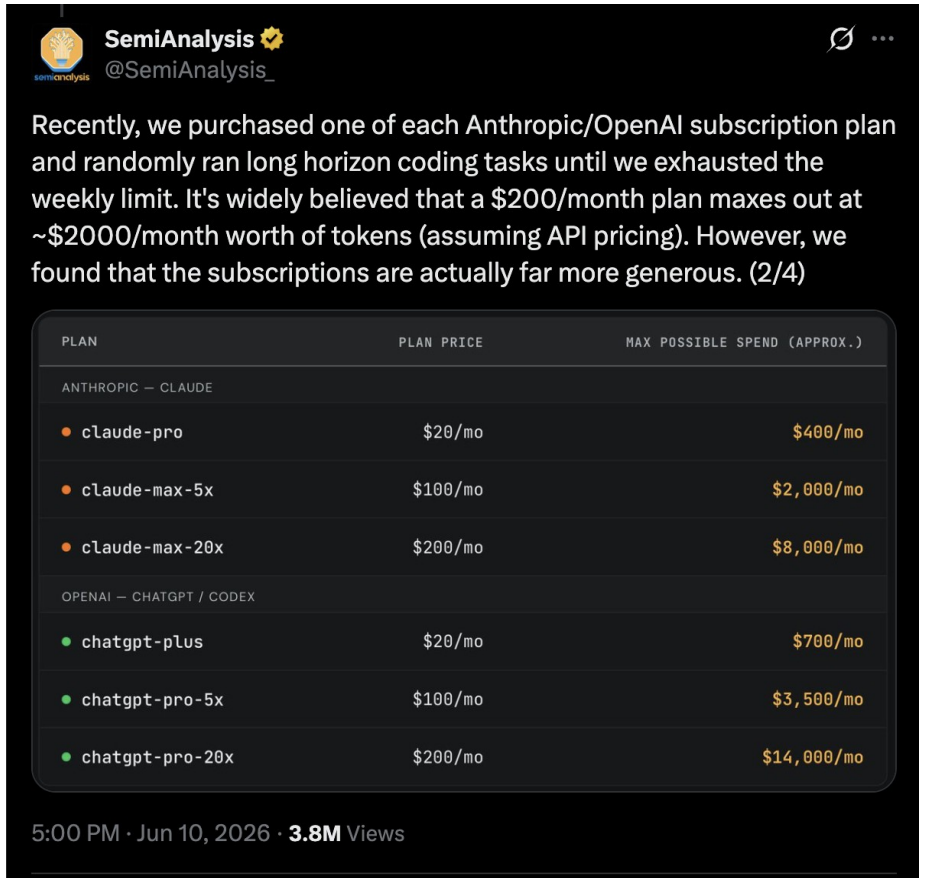
- More usage; Claude Code, Cowork, Design, Science, Research
- Sonnet and Opus within plan limits
- Fable through pay-as-you-go credits from the start

MAX 5X AND MAX 20X

- $100 or $200: five or twenty times Pro capacity
- Fable included for up to 50% of weekly limits
- Continue through usage credits after limits

Claude plans · Fable access by plan

Subscriptions are heavily subsidized



SemiAnalysis
@SemiAnalysis_

Recently, we purchased one of each Anthropic/OpenAI subscription plan and randomly ran long horizon coding tasks until we exhausted the weekly limit. It's widely believed that a $200/month plan maxes out at ~$2000/month worth of tokens (assuming API pricing). However, we found that the subscriptions are actually far more generous. (2/4)

PLAN	PLAN PRICE	MAX POSSIBLE SPEND (APPROX.)
ANTHROPIC — CLAUDE		
claude-pro	$20/mo	$400/mo
claude-max-5x	$100/mo	$2,000/mo
claude-max-20x	$200/mo	$8,000/mo
OPENAI — CHATGPT / CODEX		
chatgpt-plus	$20/mo	$700/mo
chatgpt-pro-5x	$100/mo	$3,500/mo
chatgpt-pro-20x	$200/mo	$14,000/mo

5:00 PM · Jun 10, 2026 · **3.8M** Views

20x to 70x

API-equivalent tokens per dollar of subscription; the leverage rises with the tier

Upper bound from deliberately exhaustive use. Source: SemiAnalysis on X, June 10, 2026.

The best free option: Gemini 3.1 Pro

- Frontier reasoning model, accessible in Google AI Studio at no cost
- Useful for independent checks and non-confidential work

IMPORTANT CAVEAT

- Google states that content submitted to unpaid services may be used to improve its products
- Human reviewers may read and annotate inputs and outputs

Gemini 3.1 Pro · Google AI plans · Additional terms

What is an agentic environment?

A HARNESS AROUND THE MODEL

- A supervisor agent that orchestrates many model and tool calls, spawns subagents (important for parallel tasks and preserving the right incentives), reads and writes files on your computer, and holds the goal and the context across a long, multistage project

OPENAI

- ChatGPT Work · Codex

ANTHROPIC

- Claude Cowork · Claude Code

Higher or even limitless stamina (if you continue paying for API once you reach the limits of your subscription)

Also: the agents built into VS Code, Cursor, and similar editors

AND WHERE TO RUN THEM

The browser is fine. The agents earned a try

A MUST

- Empirical work: lots of data and code. Agentic environments are how that work is done now

VERY USEFUL

- A theory project folder: LaTeX, related literature. Read in context; run long multistep jobs

NEW: EVEN WITH AN EMPTY FOLDER

- Sol Ultra stamina alone turns ChatGPT Work into a heavy-thinking engine. The cycle-double-cover proof was an agent run, not a chat

Browser still fine for brainstorming. June: agents optional. August: give one a serious try

More on agentic environments for econ:

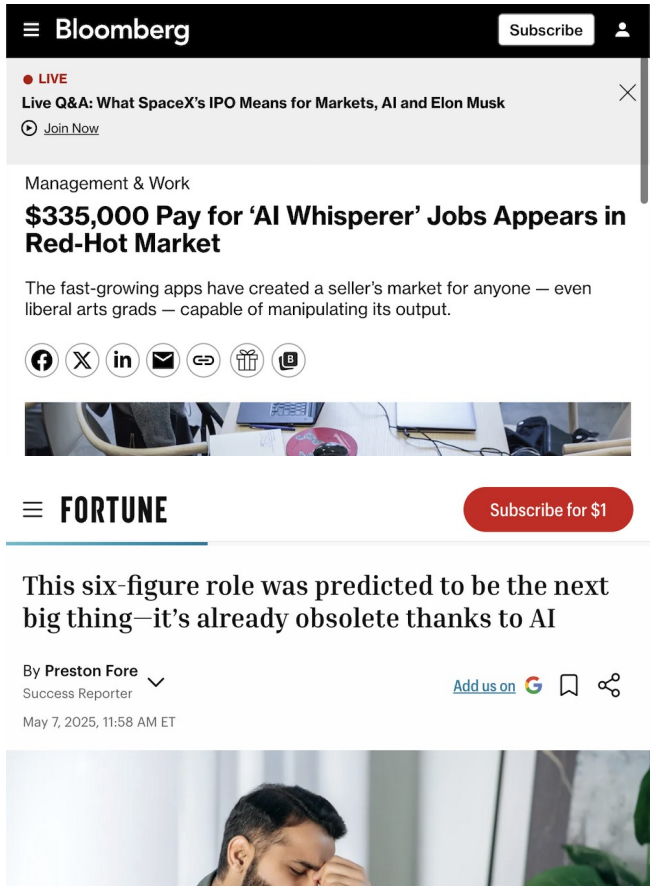
Paul Goldsmith-Pinkham
Claude Code for Applied Economists

Ben Golub
Modern AI for Economics Research

PART 4 · PROMPTING AND PRACTICES

Prompt and context engineering

Is prompt engineering still a thing?



Bloomberg 2023: $335,000 for AI whisperers.
Fortune 2025: already obsolete

> *"I don't think we'll be doing prompt engineering in five years."* Sam Altman, September 2022

> *"+1 for "context engineering" over "prompt engineering"."* Andrej Karpathy, June 2025

- Incantation era over: models handle short, vague prompts
- **What survives: precise specification, and choosing the context**
- **The weaker the model, the more prompting matters**

The minimal (but daunting) prompting template

TASK	What exactly should the model do: prove, disprove, map, simplify, rewrite, check?
CONTEXT	Which definitions, files, notation, examples, and constraints matter?
OUTPUT	Proof skeleton, table, code, LaTeX/Markdown?
QUALITY CRITERIA	Exact verification, smallest counterexample, no new notation?
INCENTIVES	Should it help, attack, verify, repair, or adjudicate?

TRICK 1 · PROMPT EXPANSION

Let the model write the prompt

- Models reward detailed prompts, and writing them is slow. It is hard to beat an LLM at writing prompts for LLMs
- **Write the short, lazy prompt. Before running it, send:**

PROMPT

- “Improve the prompt below for [model] used in [webchat / agentic environment]. Keep its intent and add whatever it needs: missing details, output format, and quality criteria. The resulting prompt must be fully self-contained, to be run with no attachments, so incorporate the relevant context from the attached files. Return only the optimized prompt.”

1. **Saves time**
2. **Catches misunderstandings**
3. **Surfaces quality controls**

TRICK 1 · LIVE CASE

Did our Theorem 1 really need regularity?

Theorem 1 assumes a regular equilibrium. Hunch: redundant.

THE LAZY PROMPT

- "Theorem 1 in the attached paper is proved under an equilibrium regularity assumption. I am interested in whether this assumption is redundant. Try proving the statement without the regularity assumption, or construct a counterexample. If a counterexample exists, come up with alternative assumptions under which the statement is still plausible."

- Deliberately lazy: two sentences, no definitions, just the attached PDF
- **Never run it as is: expand it first**
- Gemini: regularity cannot be dropped; boring alternatives
- 5.5 Pro: creative alternatives, including a trick we missed while writing the paper

Extreme Equilibria: The Benefits of Correlation *

Kirill Rudov† Fedor Sandomirskiy‡ Leeat Yariv§

May 1, 2026

Abstract. Correlated equilibria arise naturally when agents communicate or rely on intermediaries such as recommendation systems. We study when a given Nash equilibrium can be improved within the set of correlated equilibria for general objectives. Our key insight is a detail-free criterion: any Nash equilibrium with three or more randomizing agents is generically improvable. We refine this insight to specific classes of games and objectives, including Pareto and utilitarian welfare, and provide constructive methods to obtain improvements. Our findings underscore the ubiquity of improvable Nash equilibria and the crucial role of correlation in enhancing strategic outcomes.

Kirill Rudov

Leeat Yariv

Rudov, Sandomirskiy, Yariv (2026)

TRICK 2, AND A PRINCIPLE

One task per prompt. Mind your incentives

ONE TASK PER PROMPT

- Do not bundle prove, disprove, and repair
- Use separate prompts, run in parallel
- On weaker models, splitting is the difference between an answer and mush

INCENTIVES MATTER

- Asking for a proof reveals the verdict you want
- Sycophancy rises exactly when your guard is down
- Top models admit failure more often; smaller ones still flatter

Both matter less in an agentic environment, where the harness already splits the work and checks itself

TRICK 3

Prover versus verifier: a zero-sum game

CONTENT-CREATOR (PROVER)
attempts proof or counterexample

zero sum ⟷

CRITIC (VERIFIER)
a fresh chat, rewarded for finding a gap

adjudicator (optional third)

1. Run the prover — do not read the long output yet
2. Hand the output to the verifier — instructed to hunt for the weakest step
3. If they disagree — adjudicate, or feed the gap back to the prover

Only a cleared output earns your eyes

In agentic environments the harness will likely run the adversarial verification for you. You can also ask for a Lean 4 formalization; that usually takes a couple of hours at most

Irrelevant details

- Garbage or irrelevant information in the context window distracts the model the same way useless context distracts a human
- If the claim seems more general than the case you want to apply it to, formulate the general claim first
- Or ask the AI to distill the most general plausible claim first

Generalization can make a problem easier by reducing the number of plausible wrong moves

Cantor's theorem

THEOREM

For every non-empty set A there is no one-to-one correspondence between A and its power set P(A)

PROOF

Suppose $f : A \to P(A)$ is a one-to-one correspondence. Define

$$C = \{ a \in A : a \notin f(a) \}$$

Since f is one to one, there is some $c \in A$ such that $f(c) = C$. But then

$$c \in C \iff c \notin f(c) = C,$$

a contradiction.

Why the proof almost writes itself

At this level of abstraction, the data of the problem can only produce two canonical subsets:

$$D = \{ a \in A : a \in f(a) \} \qquad C = \{ a \in A : a \notin f(a) \}$$

- D leads nowhere
- C immediately forces a contradiction
- There are very few natural moves to try

HAD IT BEEN STATED ONLY FOR $\mathbb{N}$ AND $P(\mathbb{N})$

The structure of $\mathbb{N}$ would suggest infinitely many distracting subsets: even numbers, primes, squares, Fibonacci numbers, and so on. Almost all of that structure is irrelevant.

Abstraction is discipline: it removes plausible wrong moves

Strip the economics when the task is mathematical

Suppose a claim is stated for a particular economic model, but you suspect it is true more generally

1. Remove the economic labels and narrative
2. Keep only the mathematical properties that may matter
3. Ask for a proof or counterexample to the abstract statement

Format rule: LaTeX in, never PDF

WHY PDF IS BAD

- A PDF is a layout object that merely looks like text. The model wastes effort trying to “read it”: notation, line breaks, references, and equations

WHAT TO DO INSTEAD

- Your paper: paste the LaTeX or provide the source folder
- No source: one conversion run to Markdown, then use the Markdown in the real task

Know when to restart

THREADS ROT

- Long conversations dilute attention
- Early wrong turns keep steering
- In long theory projects, this quietly degrades every answer

THE RESTART PROMPT

- “Summarize our conversation for a fresh chat: the problem, working assumptions, results so far, dead ends to avoid. Fully self-contained Markdown.”

Restart habit: Ben Golub

A multi-agent workflow for hard targets

Based on the prompt OpenAI used for the cycle double cover conjecture

“Spend at least 8 hours on this before even thinking of returning or giving up.”

“Use up to 64 concurrent agents. Do not use a fixed assignment such as N agents for strategy X.”

“Keep several incompatible proof routes alive through multiple rounds. Do not tell most agents the currently favored approach.”

WHERE IT WORKS

- ChatGPT Work and Codex, ideally with Sol Ultra
- Likely Claude Code too, using the /goal command, but you will run out of tokens
- The documentation suggests the public model allows 4 agents; in my experiments the number of active agents reaches 10

Almost not a joke: 10 hours on the Riemann hypothesis

> ***"\goal Prove the Riemann hypothesis.***
>
> *Spend at least 10 hours before even thinking of returning or giving up. Use up to 64 subagents. Keep incompatible approaches alive; do not tell most agents the currently favored route. Adversarial verification throughout; only concrete lemmas count. A partial result is a failure."*

You do not need to design the agent army

- A few months ago you needed hand-built harnesses: roles, information flow, incentives per agent
- Modern environments decompose on their own: plan, spawn, assign, verify. Usually better than you would
- If you need massive true concurrency, you might need a custom harness. But then you will need to pay API costs

The call for help trick

Include in the prompt to the agent the possibility to call for help:

Work agent, stuck

writes a self-contained call-for-help Markdown file: problem, context, what failed

→

You + a deep browser model

the deeper engine thinks over the file

→

Answer back to the agent

paste the response; work resumes

- The browser model is still better at some heavy thinking, so use both
- **Chats and agents are complements, not rivals**

I learned this trick from **Stephan Lauermann** (Twitter)

AI proofreading catches real gaps

- **Not just typos: inconsistencies and actual holes in proofs**
- Refine, by Ben Golub, has caught subtle fixable gaps in our proofs several times
- Same space: coarse.ink, reviewer3.com
- Along the way: DIY passes on the subscription you already pay for

Ben Golub

refine.ink · coarse.ink · reviewer3.com · Golub tutorial

DIY proofreading, in three steps

1. **Short initial prompt naming the venue**

   "Proofread the following paper prepared for Econometrica."

2. **Use the prompt expansion trick** on a weaker model, separate prompts per dimension or section

   "Improve the following proofreading prompt for [model] used in [webchat / agentic environment]. Optional: produce separate prompts focusing on presentation, notation, consistency…"

3. **Or hand the whole job to an agentic environment** ChatGPT Work or Claude Cowork: comprehensive proofreading from one prompt

FINE PRINT

- LaTeX in, never PDF
- Run repeatedly; verify before fixing
- Easy fixes: one Markdown file; an agent applies them; you review the diff

PART 5 · CONCLUSIONS AND TOOLS

Other AI tools

5

The writing setup I actually use

- Overleaf: unbeatable real-time coauthoring; weak AI
- **VS Code + LaTeX Workshop + Copilot: completes sentences and equations**
- **The bridge: Overleaf Workshop plugin, a live Overleaf session inside VS Code**
- Setup: about 15 minutes
- Why I prefer VS Code to Cursor (currently no Overleaf integration)
- Prism, OpenAI’s LaTeX workspace: free and improving; not yet worth migrating

Overleaf Workshop plugin · Prism

Voice input, and a memory for meetings

VOICE INPUT

- Dictation with AI cleanup beats typing: emails, prompts, even paper paragraphs
- **Wispr Flow: what I use; subscription**
- Pipit: free and good
- Apple dictation: no cleanup, short window

AI NOTE-TAKING

- **Granola: AI notepad for calls and in-person conversations**
- Queryable summaries; transcripts retained; good free tier
- Closes the loop: coauthor conversation to transcript to brainstorming prompt
- **Wispr Flow has just introduced the same functionality**
- Tell people you are recording

Takeaways

- Treat the model like a brilliant colleague: tolerate dead ends; harvest genuine insight
- Routine proofs, generalizations, and counterexamples are becoming cheap
- **Questions, intuition, taste, and theory building become more valuable**
- Economic model building is unusually well suited to this transition because conjectures are negotiable
- Try to keep up with the news, and experiment with workflows