AI and the future of mathematics · IHP, Paris · 21 October 2026

AI and research

riding dragons

Antoine Tilloy

Warning

- not about teaching
- not about society as a whole
- **just about research**

Phase transition

LLM: an augmented database

GPT-3.5 (2022): fun, but confabulates by extrapolation

LLM: an augmented database

GPT-3.5 (2022): fun, but confabulates by extrapolation

- **Who is Sylvia Serfaty?** → [perfect answer]

LLM: an augmented database

GPT-3.5 (2022): fun, but confabulates by extrapolation

- **Who is Sylvia Serfaty?** $\longrightarrow$ [perfect answer]
- **Who is Antoine Tilloy?** $\longrightarrow$ [...] a mathematical physicist, PhD student of Sylvia Serfaty, who worked on [...]

LLM: an augmented database

GPT-3.5 (2022): fun, but confabulates by extrapolation

- **Who is Sylvia Serfaty?** ⟶ [perfect answer]
- **Who is Antoine Tilloy?** ⟶ [...] a mathematical physicist, PhD student of Sylvia Serfaty, who worked on [...]

made-up citations, elementary computations wrong, elementary logic wrong ⟶ stochastic parrot?

LLM: solves hard problems

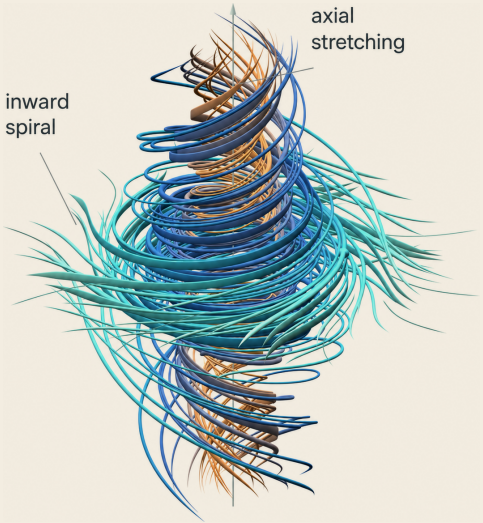
8 September 2026: an internal OpenAI model (GPT-7?) proves **finite-time blow-up** for Navier–Stokes *with a smooth force*

→ statement (C) of the Clay problem

global regularity at $f \equiv 0$ still open

166 pages, verified in Lean

$\sim 10^4$ concurrent agents



Models in a nutshell

Pre-training

Gradient descent on all the text humankind ever wrote
[$10B$ for `GPT2` $\longrightarrow$ $32T$ tokens for `DeepSeek V4 Pro`]

Pre-training

Gradient descent on all the text humankind ever wrote
[$10B$ for `GPT2` $\longrightarrow$ $32T$ tokens for `DeepSeek V4 Pro`]

Result:

- **prompt**: “The capital of France is”
- **answer**: “Paris, a city then in full expansion, which Louis XIII wished to make shine [...]”

Pre-training

Gradient descent on all the text humankind ever wrote
[$10B$ for `GPT2` $\longrightarrow$ $32T$ tokens for `DeepSeek V4 Pro`]

Result:

- **prompt**: “The capital of France is”
- **answer**: “Paris, a city then in full expansion, which Louis XIII wished to make shine [...]”

- **prompt**: “Prove that there are infinitely many primes.”

Pre-training

Gradient descent on all the text humankind ever wrote
[$10B$ for `GPT2` $\longrightarrow$ $32T$ tokens for `DeepSeek V4 Pro`]

Result:

- **prompt**: “The capital of France is”
- **answer**: “Paris, a city then in full expansion, which Louis XIII wished to make shine [...]”

- **prompt**: “Prove that there are infinitely many primes.”
- **answer**: “(3 points) 2. Prove that $\sqrt{2}$ is irrational. (4 points)”

Supervised fine-tuning

Gradient descent on carefully chosen examples of the form:

- user question
- expected answer

[$1.5M$ examples for `DeepSeek V3`]

Supervised fine-tuning

Gradient descent on carefully chosen examples of the form:

- user question
- expected answer

[$1.5M$ examples for `DeepSeek V3`]

Result:

- **prompt**: “Prove that there are infinitely many primes.”
- **answer**: “Suppose there are finitely many primes $p_1, \ldots, p_n$, and consider $N = p_1 \cdots p_n + 1$ [...]”

Reinforcement learning from verifiable rewards – RLVR

In mathematics and code, the result can be checked.

We reinforce on $\sim 1M$ problems, of increasing horizon:

- the elementary computation $3 + 6$ is correct
- the code of a simple Python function runs
- the simplification of an integral is correct
- the code compiles *and* gives the right result
- a very general mathematical question is correctly solved, with several steps and intermediate pieces of code

$\longrightarrow$ emergence of reasoning and strong generalization

Recent developments

- **Harness**: the model is given an environment (the terminal) and tools (bash, Python, internet search, APIs) + **RLVR**

  Claude Code, Codex, Antigravity, Grok Build, Pi, OpenCode, Hermes, OpenClaw

Recent developments

- **Harness**: the model is given an environment (the terminal) and tools (bash, Python, internet search, APIs) + **RLVR**
  Claude Code, Codex, Antigravity, Grok Build, Pi, OpenCode, Hermes, OpenClaw
- **Subagents**: the LLM writes the prompt of another LLM, and reads part of its output $\longrightarrow$ keeps the context reasonable

Recent developments

- **Harness**: the model is given an environment (the terminal) and tools (bash, Python, internet search, APIs) + **RLVR**
  Claude Code, Codex, Antigravity, Grok Build, Pi, OpenCode, Hermes, OpenClaw
- **Subagents**: the LLM writes the prompt of another LLM, and reads part of its output $\longrightarrow$ keeps the context reasonable
- **Skills**: recipes (in natural language) explaining how to do something

Recent developments

- **Harness**: the model is given an environment (the terminal) and tools (bash, Python, internet search, APIs) + **RLVR**
  Claude Code, Codex, Antigravity, Grok Build, Pi, OpenCode, Hermes, OpenClaw
- **Subagents**: the LLM writes the prompt of another LLM, and reads part of its output → keeps the context reasonable
- **Skills**: recipes (in natural language) explaining how to do something
- **Grounding**: internet search, literature databases, your own files

Recent developments

- **Harness**: the model is given an environment (the terminal) and tools (bash, Python, internet search, APIs) + **RLVR**
  Claude Code, Codex, Antigravity, Grok Build, Pi, OpenCode, Hermes, OpenClaw
- **Subagents**: the LLM writes the prompt of another LLM, and reads part of its output → keeps the context reasonable
- **Skills**: recipes (in natural language) explaining how to do something
- **Grounding**: internet search, literature databases, your own files
- **Chain of thought**: thinking before answering → *inference* compute buys quality too

Recent developments

- **Harness**: the model is given an environment (the terminal) and tools (bash, Python, internet search, APIs) + **RLVR**
  Claude Code, Codex, Antigravity, Grok Build, Pi, OpenCode, Hermes, OpenClaw
- **Subagents**: the LLM writes the prompt of another LLM, and reads part of its output → keeps the context reasonable
- **Skills**: recipes (in natural language) explaining how to do something
- **Grounding**: internet search, literature databases, your own files
- **Chain of thought**: thinking before answering → *inference* compute buys quality too
- **RL environments written by AI** → proto-**RSI**

Where this leaves us

Models now **reason**.

Where this leaves us

Models now **reason**.

We are no longer bottlenecked by data, only by compute.



Colossus II (xAI)



Stargate (OpenAI)

Where this leaves us

Models now **reason**.

We are no longer bottlenecked by data, only by compute.



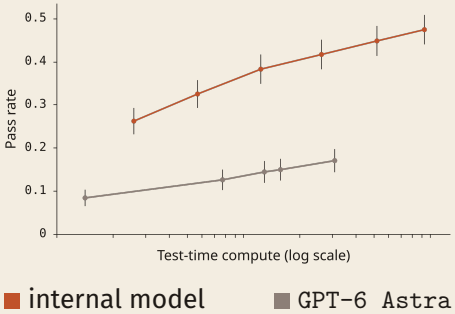
Colossus II (xAI)



Stargate (OpenAI)

It will not stop.

Not in a steady state



- decisions need to take into account improvements every 3–6 months
- the remaining weaknesses are unlikely to persist for long
- impossible to know the role of scientists in 5 years

Using modern models

Non-threatening uses

Possible since November 2025 (`Opus 4.5` + Claude Code)

- Deep literature search
- Distilling tailored explanations
- Typesetting, figures, Ti*k*Z, etc.
- Proofreading (grammar, high-level consistency)
- Writing basically any code

Scary uses: personal examples

- Nuclear reactor with flow — `Fable`, subtle perturbative calculations
  stuck for 6 months → full perturbative calculation, sims, plots, perfect
- Excitations for RCMPS — `Astra`, hard tensor network calculations
  tried for years → one-shot solution, sims, plots, perfect
- Foundations of quantum mechanics — `Astra`, remarkably subtle
  had the idea for years → killed with one surprisingly pragmatic argument

I cannot find a technical task (proof, analytical calculation, numerics) where I am better than frontier models

What should we do

The old ways are already gone

- Denial

  “It doesn’t really *think*, it’s just brute force, works on a few problems”

- Anger

  “OpenAI is desecrating mathematics!! They need to be stopped”

- Bargaining

  “Probably the technical things *I* do will remain difficult”

- Depression

  “We’ll never be able to ponder a problem for years…”

- Acceptance

  “We should ride the dragon”

Riding the dragon



Game of Thrones, HBO

- The dragon may double in size every 3 months
- The dragon may ride itself at some point
- So far, few can ride it

Letting go is necessary to do great things

Before

- Coding in assembly
- Computing integrals with manual Gauss–Legendre quadratures
- Extracting square roots with Newton’s algorithm

Letting go is necessary to do great things

Before

- Coding in assembly
- Computing integrals with manual Gauss–Legendre quadratures
- Extracting square roots with Newton's algorithm

Now

- Coding (all of it)
- Proofs (those technical/long)
- Calculations (those technical/not insightful)

What remains unsolved (for now)

qualities of a good dragon rider

- Research “taste” (what is the right question)
- Extracting theories / concepts
- Getting out of tunnel vision

General advice

Anxiety is *normal*

- Look the beast in the eyes: use the best models regularly, excessively (information matters, if only to not look like a fool)
- Destroy your own problems with it, don't let ego in the way
- Focus on the cool stuff that is now possible rather than on what's gone (to leading order, new discoveries are good!)
- Make humankind proud

Community vs society

Researchers tend to consider the community too much, instead of thinking about what they bring to society

- the community is a means to an end
- the preservation of the community cannot be the first objective
- the role of journals, conferences, PhD theses, etc. is secondary; downstream of our goals

Again, not a steady state, need to stay at the frontier to know what our goals can be

Concrete personal battle plan

Token maxing

Get near-unlimited access to frontier models

- Use the best models (`Astra`/`Fable`) in high/xhigh effort modes
  [today's `Fable` is tomorrow's `Sonnet`]
- Waste the tokens on tasks that are too difficult
- Solve problems while learning about the frontier

Find balance

Difficult, I didn’t manage

- AI-enhanced research pushes are exhausting and humiliating at first
- Sleep well / do sports / talk to your friends (or to a psychologist)

Discuss with humans, not just AIs

- Ask a strong AI together, and discuss the output live
- Do a journal club to discuss some poorly written AI breakthrough

Summary

- Models are strong, already superhuman in parts of maths
- It won’t stop or plateau soon
- Grieve fast, quickly forget what’s gone anyway
- Ride the dragon
- Find joy in the wild discoveries to come