

On Compressing Context Beyond the Window

Rishabh Agarwal

COLM 2026 Workshop

Representing the work done by the entire team at Periodic Labs

AI learned to model our digital world.

Next, it can help us build the physical world.

AI learned to model our digital world.

Next, it can help us build the physical world.

Digital world

- Learned from the internet: the finished record
- Language, math, code

AI learned to model our digital world.

Next, it can help us build the physical world.

Digital world

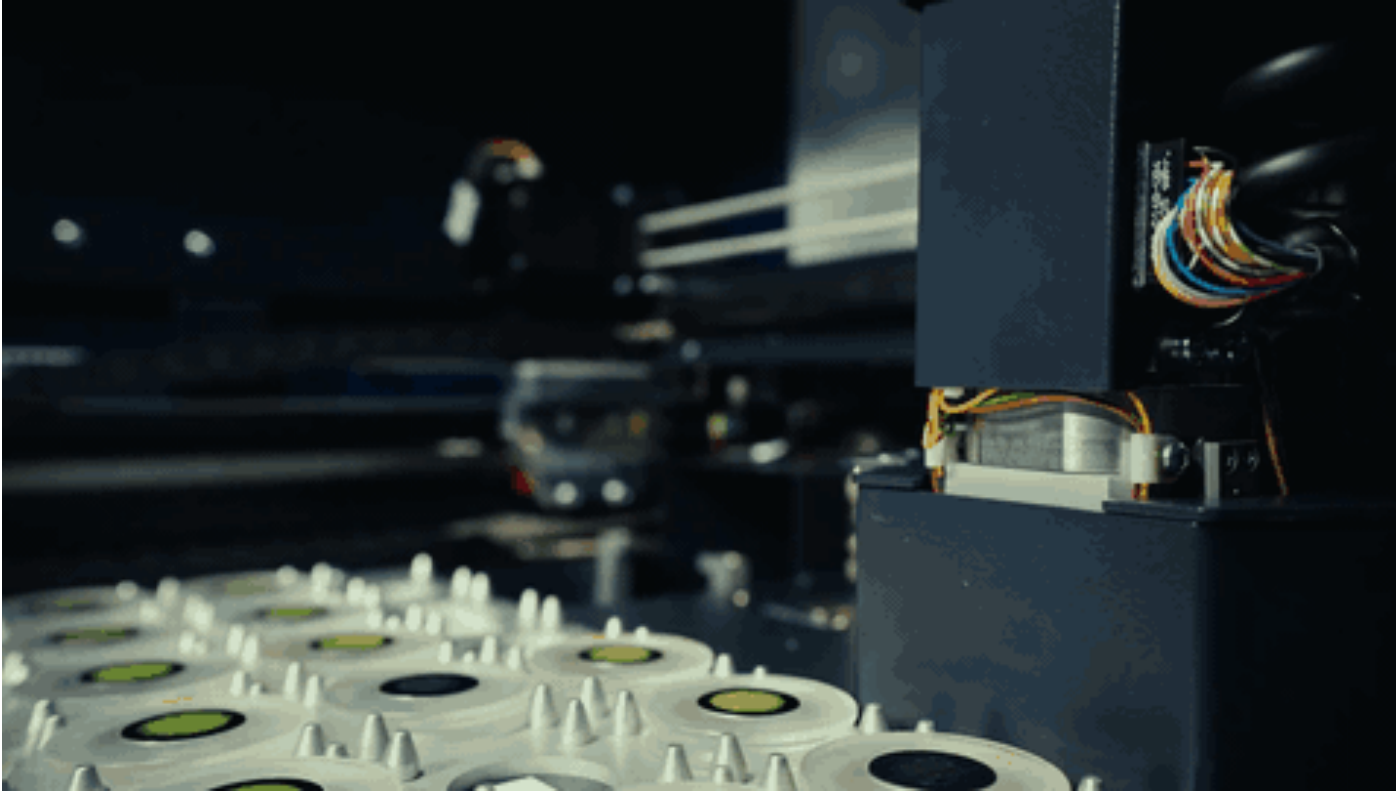
- Learned from the internet
- Verifiable tasks: Language, math, code

Physical world

- Learned from the scientific process, in our labs
- Experience beyond any window, learned continually

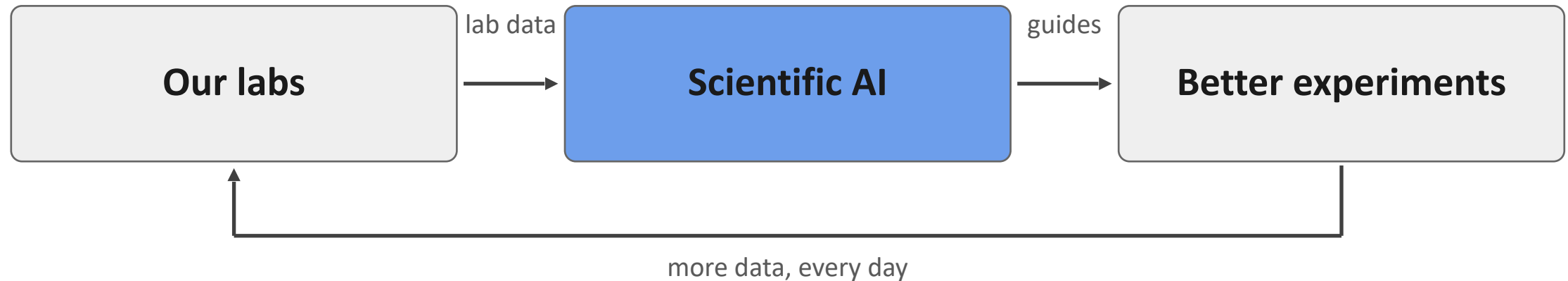
Goal: if physics allows it, we can build it

Periodic Labs: AI in the Physical World



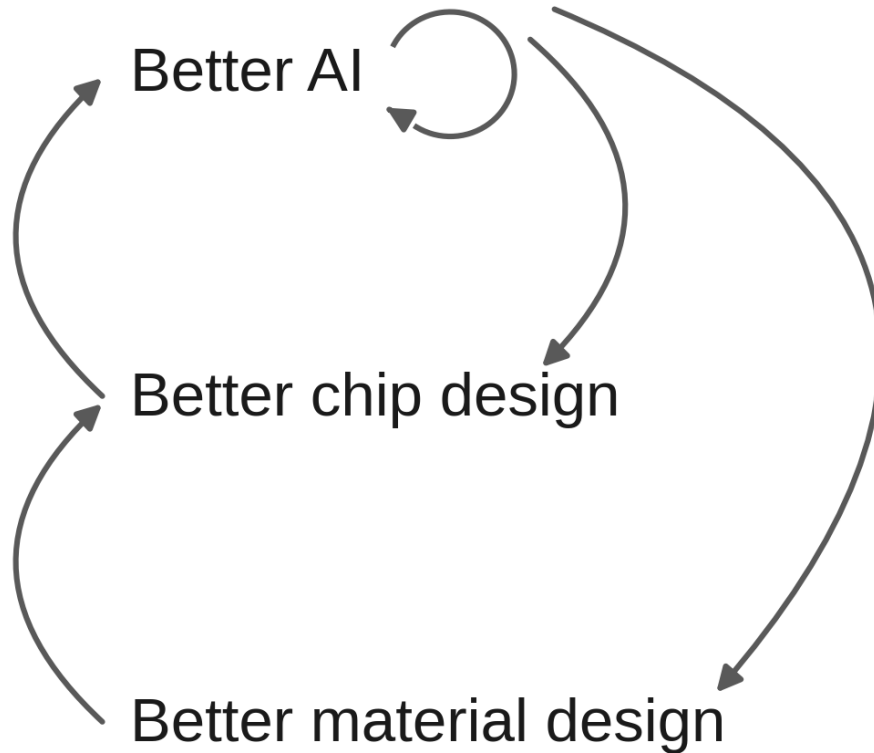
- High-throughput labs in Menlo Park, running 24/7, plus large-scale physics simulations
- Initial Goal: **discover new materials**, like superconductors and magnets
- AI embeds in our equipment, flags issues in data, and learns the scientific process

Building ~~Lab~~ Labs that Learn



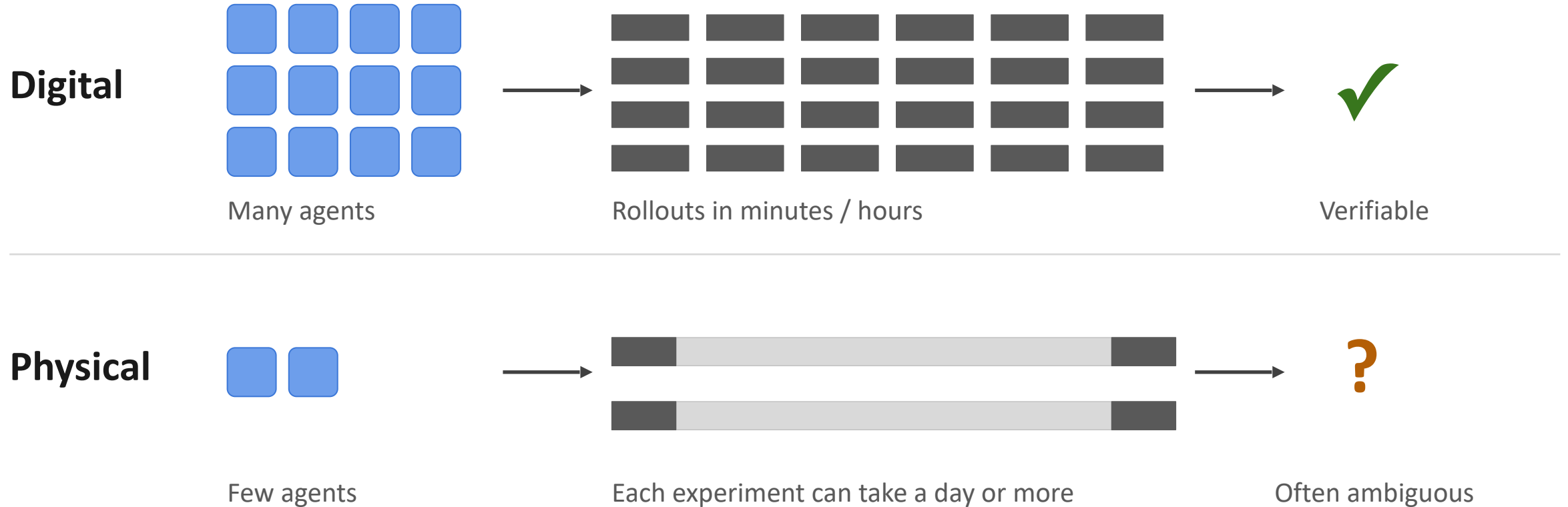
- Experiments from our labs train increasingly capable scientific AI, which in turn guides better experiments
- We benefit as frontier models improve; **when they fall short or cost too much, we train our own**

Why Materials? It's Flywheels All the Way Down



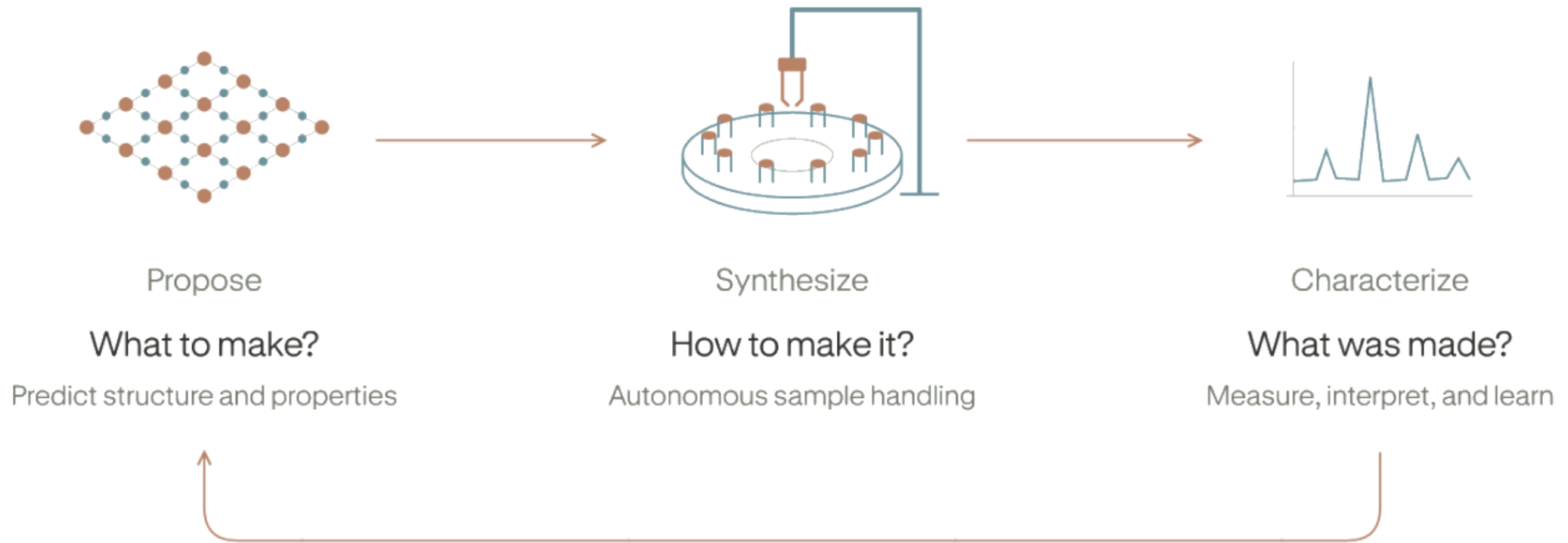
- Computation is physical: every add and multiply is engineered atoms
- Each new generation of chips is limited by materials
- Better materials → better chips → better AI

RL in the Physical World



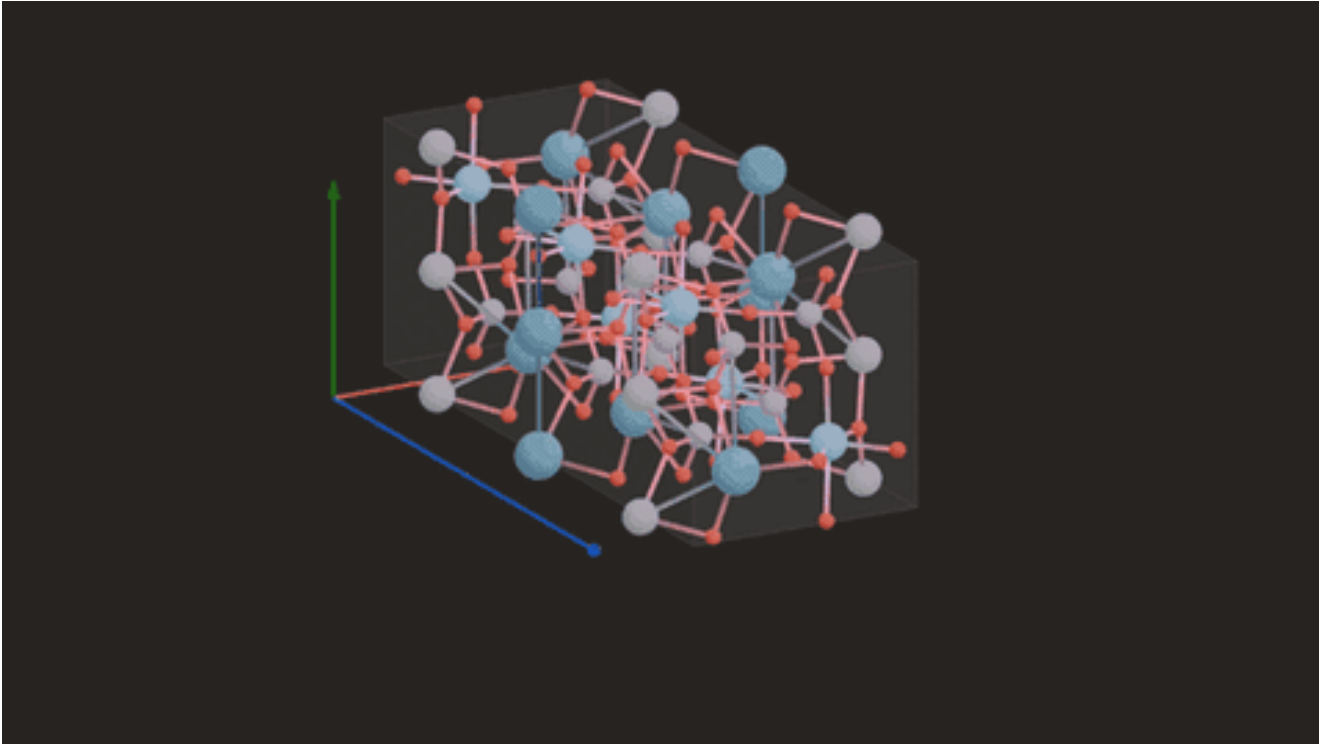
Between experiments, we **learn from the data we already have**

Scientific Discovery Loop @ Periodic

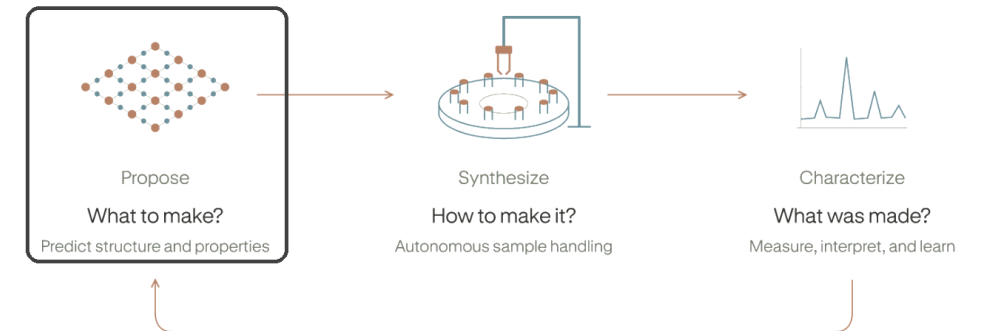


- Every turn of the loop adds **new experience**
- What was made feeds back into what to make

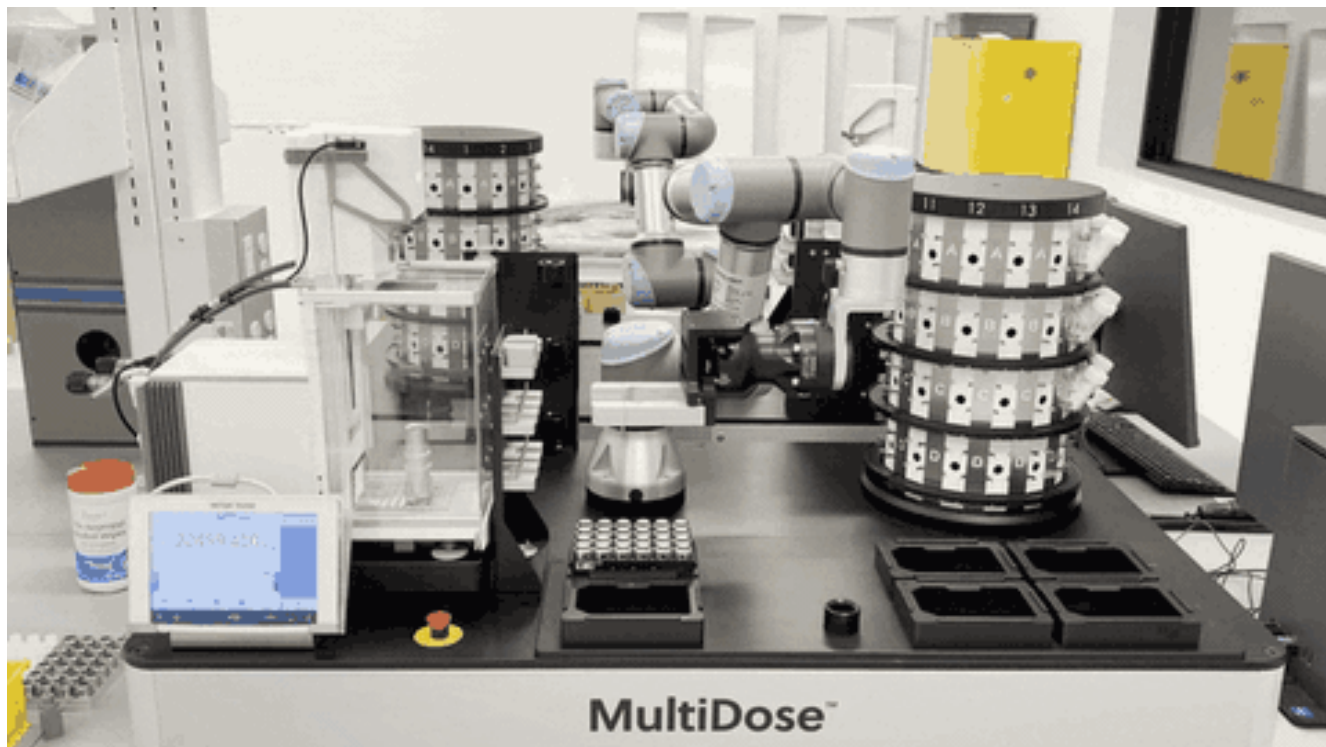
What to Make?



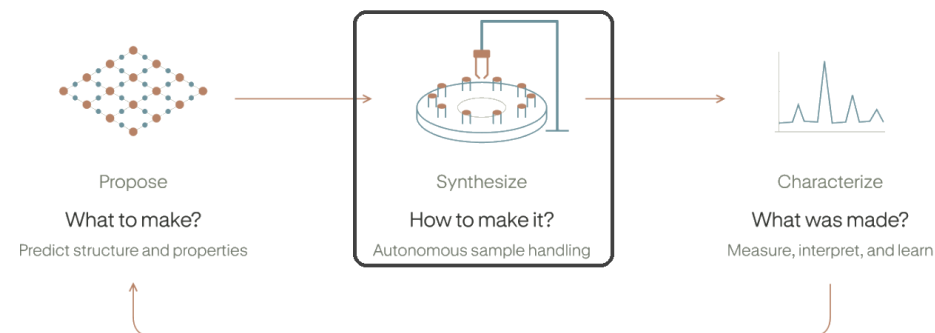
- Materials are arrangements of atoms; the arrangement sets the properties
- Predict which arrangements are stable, and their properties
- Target: a new superconductor or magnet



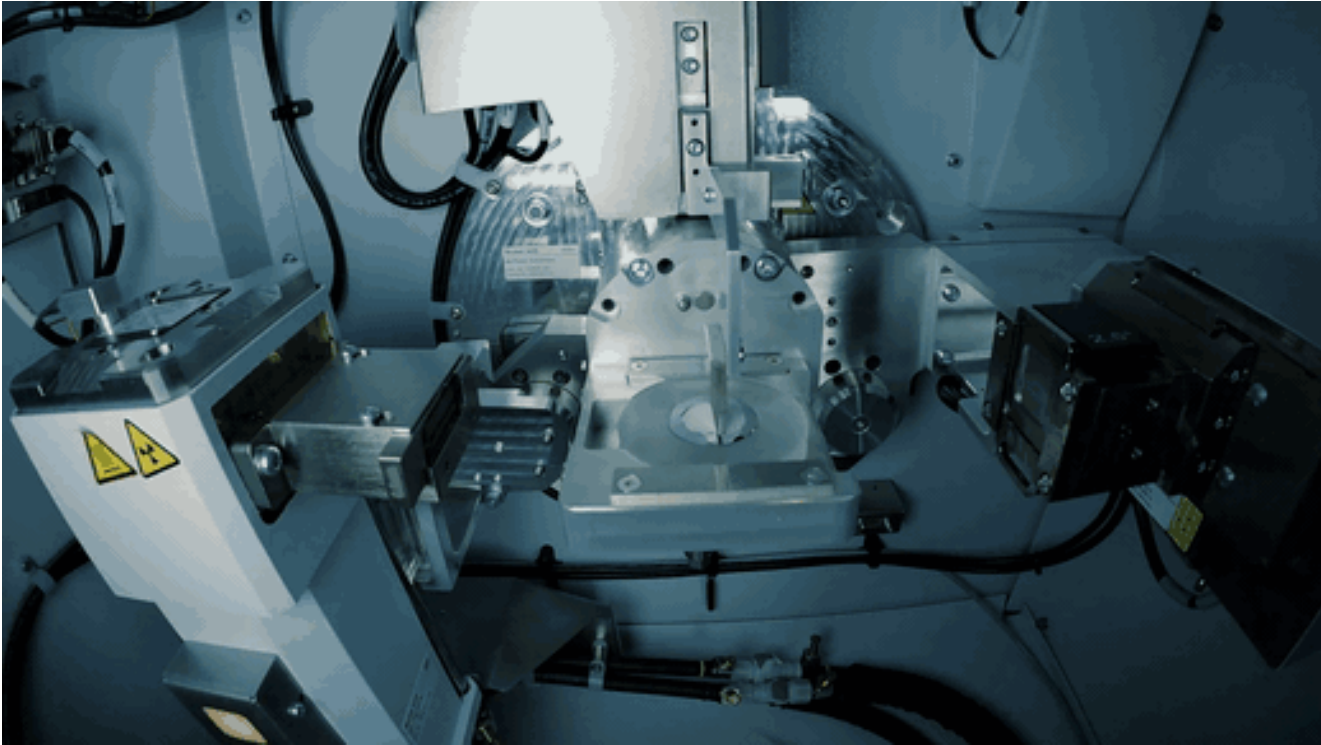
How to Make It?



- Knowing it can exist isn't enough: we need a recipe
- Precursors, ratios, atmosphere, mixing, heating
- Robots run it 24/7; every attempt is logged, including failures

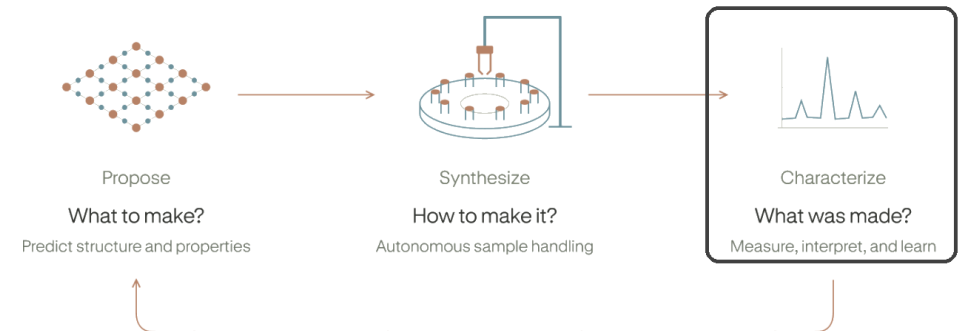


What Was Made?



An X-ray diffractometer in our labs

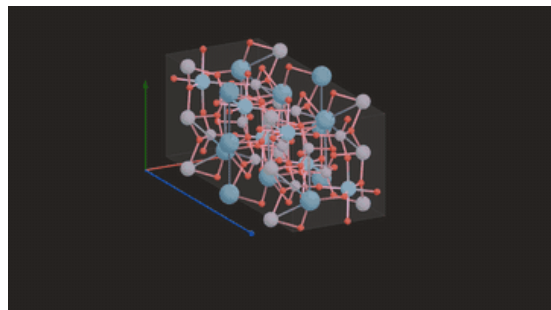
- Things don't come out labeled
- X-ray diffraction gives a fingerprint of the atomic structure
- Reading that fingerprint became our bottleneck



Every Experiment Adds Useful Context

One experiment

What to make?

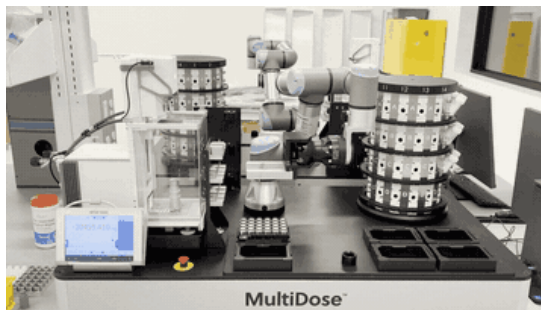


Literature, databases

Stability predictions

Past attempts

How to make it?

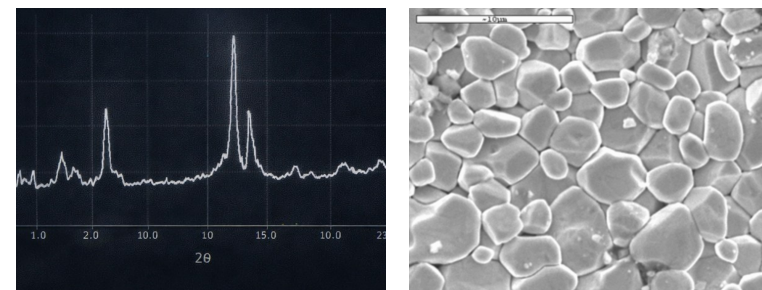


Precursors, recipes

Robot, furnace logs

Failed syntheses

What was made?

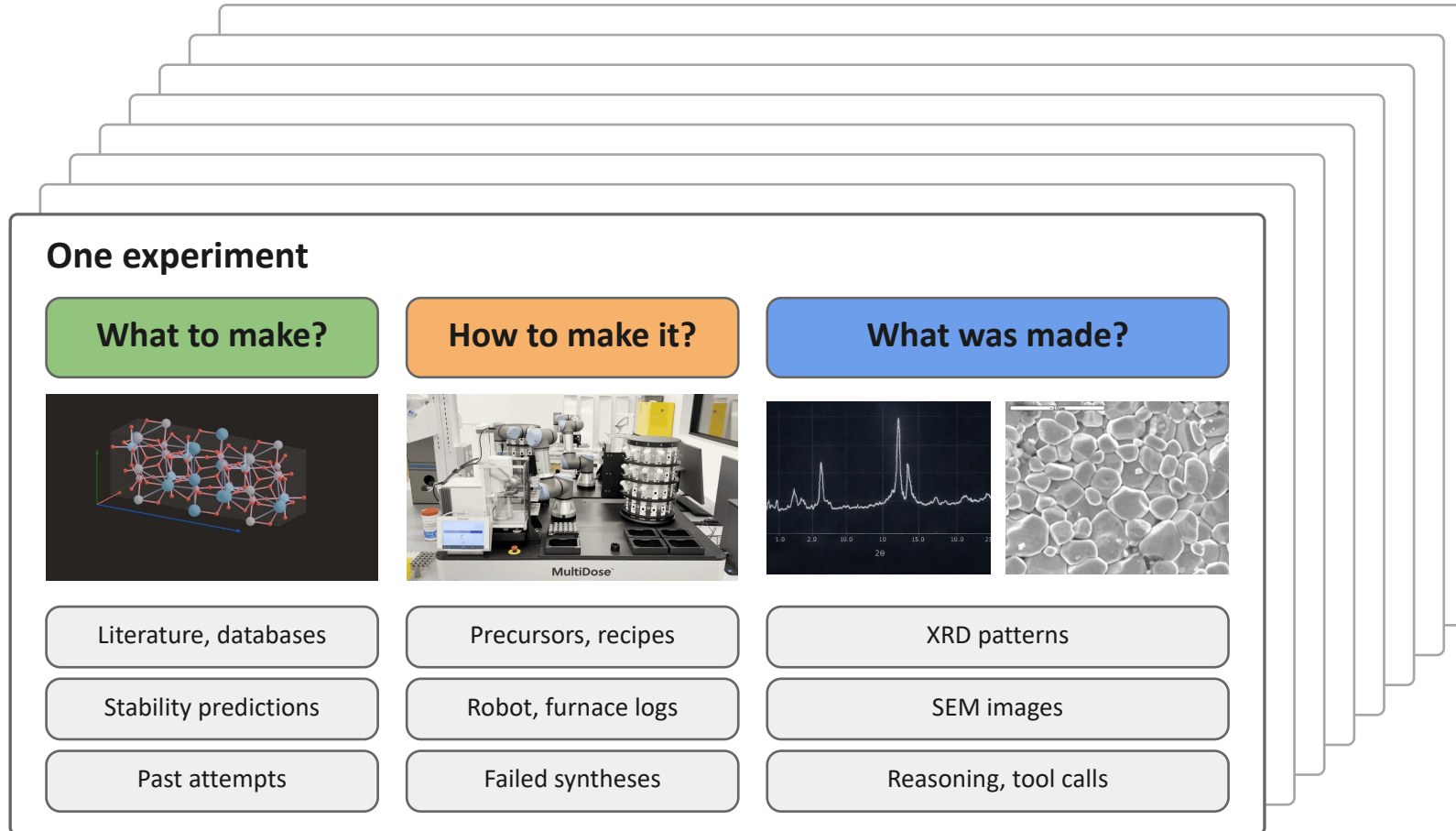


XRD patterns

SEM images

Reasoning, tool calls

Every Experiment Adds Useful Context



- Every experiment adds a card
- Results, failures, expert corrections
- Grows every day

Both outgrow the context window

AI That Knows How to Do Science

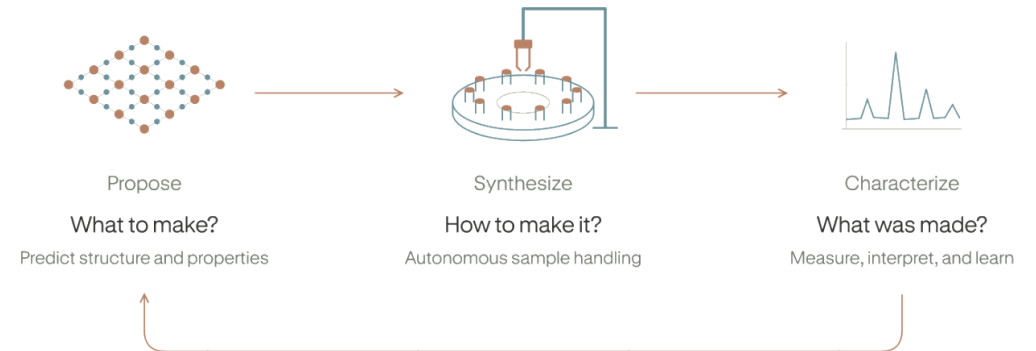
Language, math, code: **lots of data + end-to-end optimization**. Same for science.

Today's models



Trained on the **finished record** of science
→ AI that knows science

Periodic's AI



Trained on the **process** of science
→ AI that knows **how to do** science

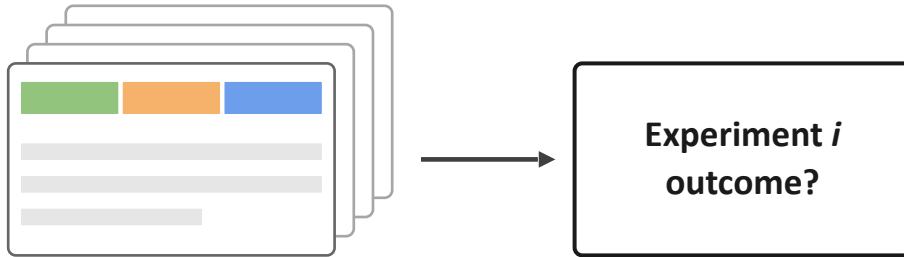
The process of science is **context that never reaches the finished record**

Context is Key: Prior Experiments Halve the Error

Predict experiment i outcome before running it

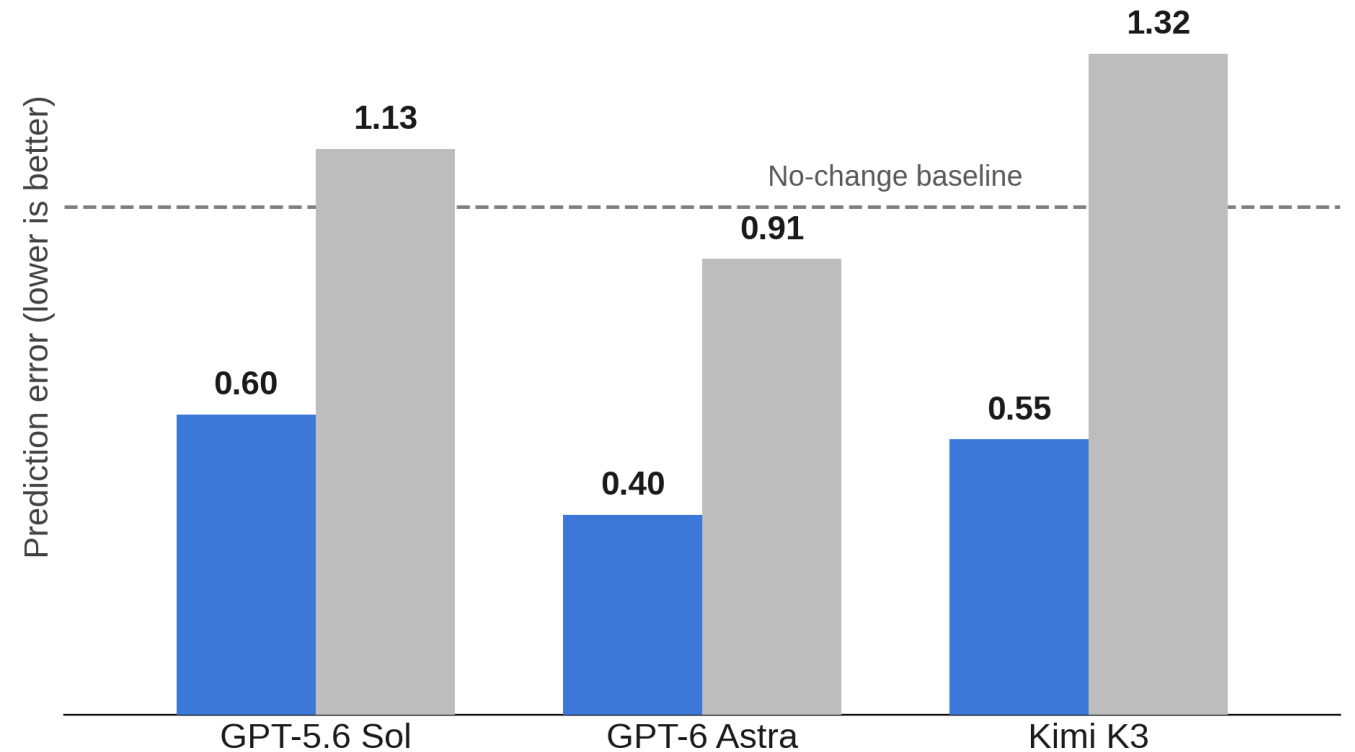
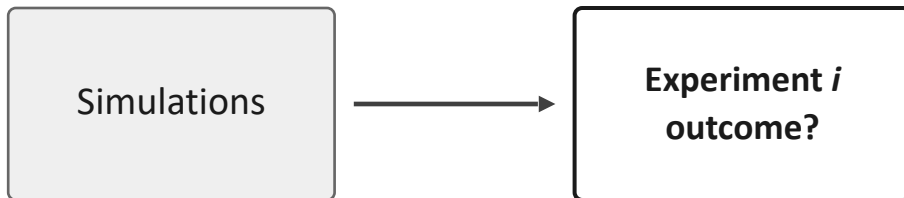
■ Prior experiments ■ No experiments

Prior experiments

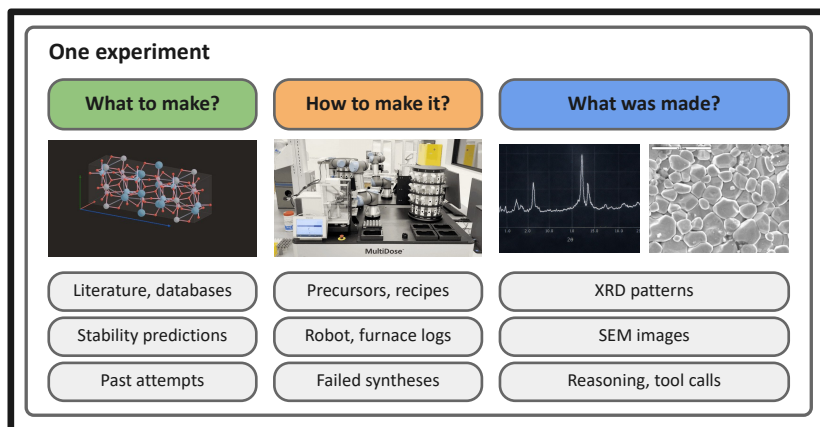


our other experiments + simulations

No experiments

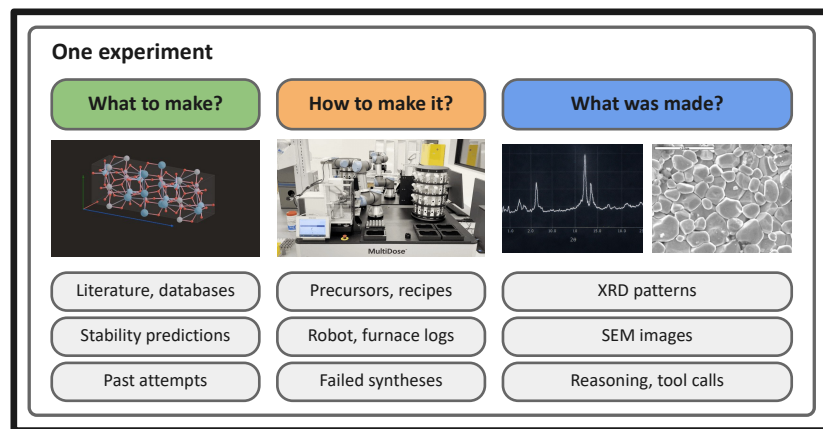


Thought Experiment: A Year of Lab Context



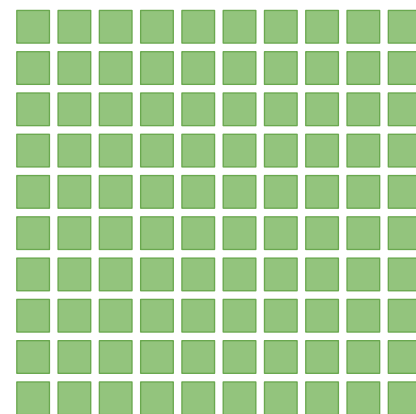
1 experiment $\approx$ 1M tokens
= **one context window**

Thought Experiment: A Year of Lab Context



1 experiment $\approx$ 1M tokens
= **one context window**

×



100 experiments / day
= **100 windows / day**

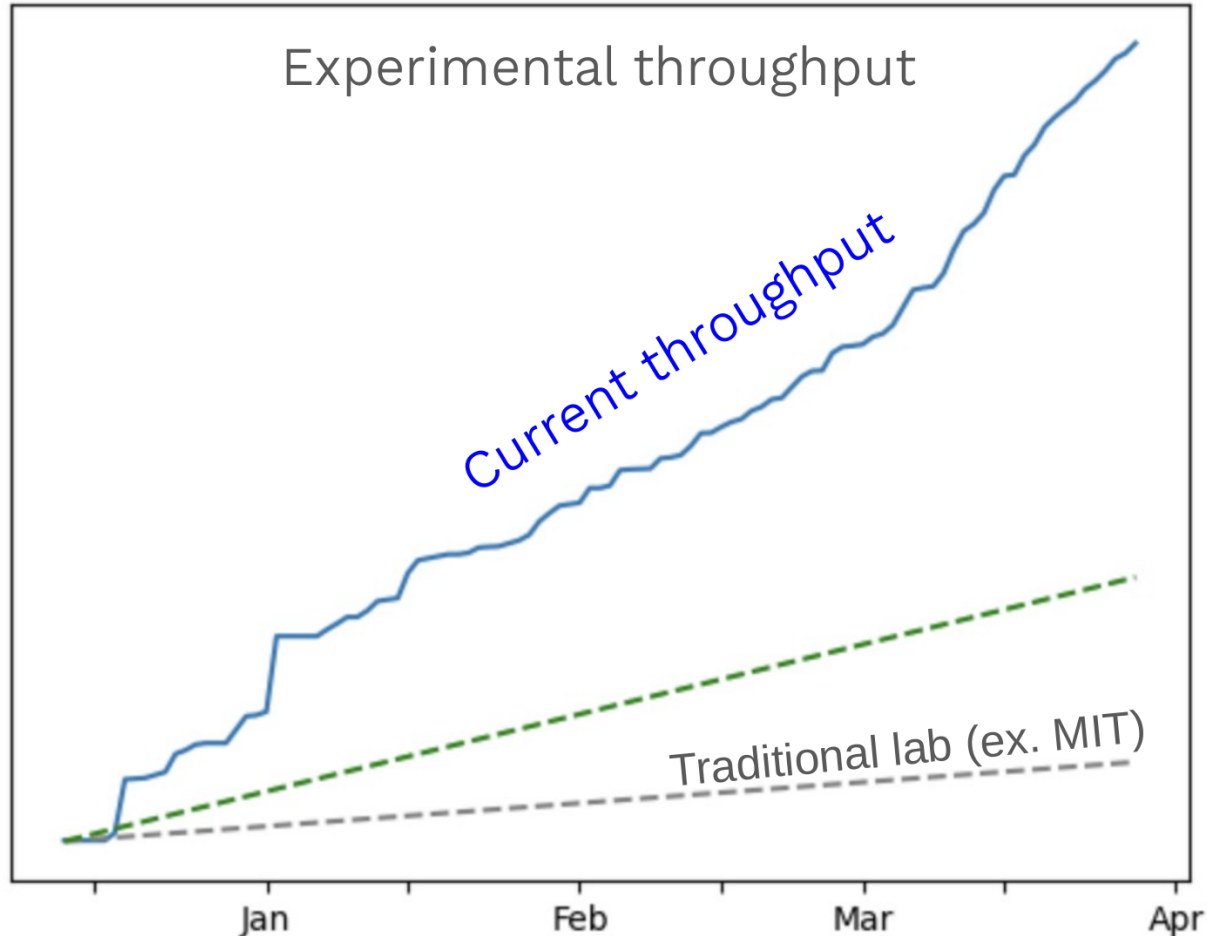
×

365 days =

36,500
windows / year

Context Is Already Beyond the Window

Our throughput today



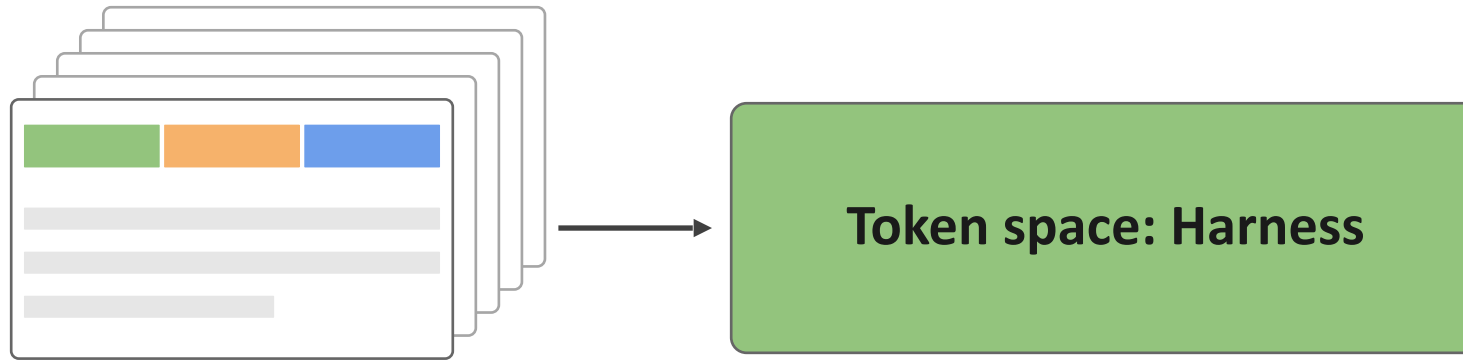
- Millions of tokens per experiment
- **Context scales with throughput**

How to handle context beyond the window?

How to handle context beyond the window?
Compression

Token Space Compression

Our labs' context



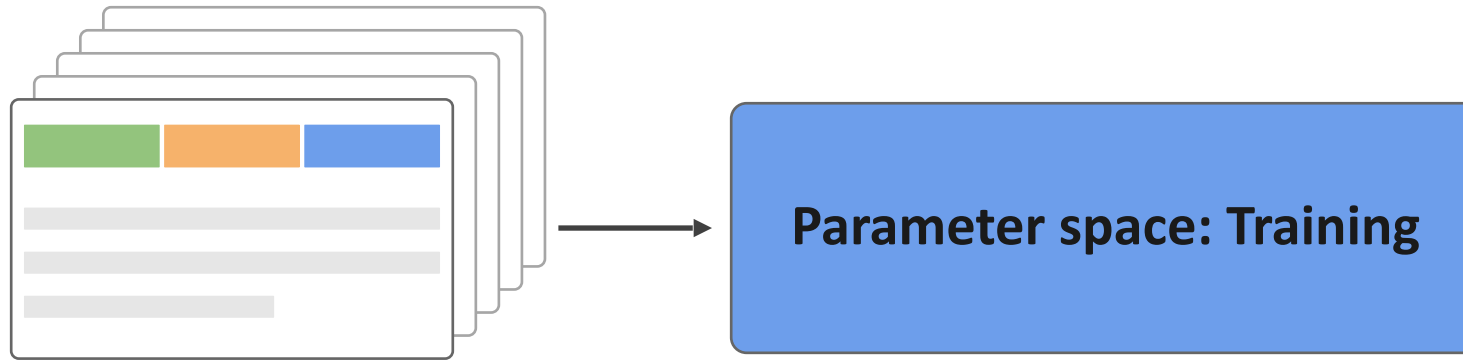
grows every day

- Summaries, notes, memory
- Learned prompts
- Retrieval

- Write it down shorter; pull it into the window when needed
- Instant and readable; works with any model
- **But you pay to re-read it on every query**

Parameter Space Compression

Our labs' context

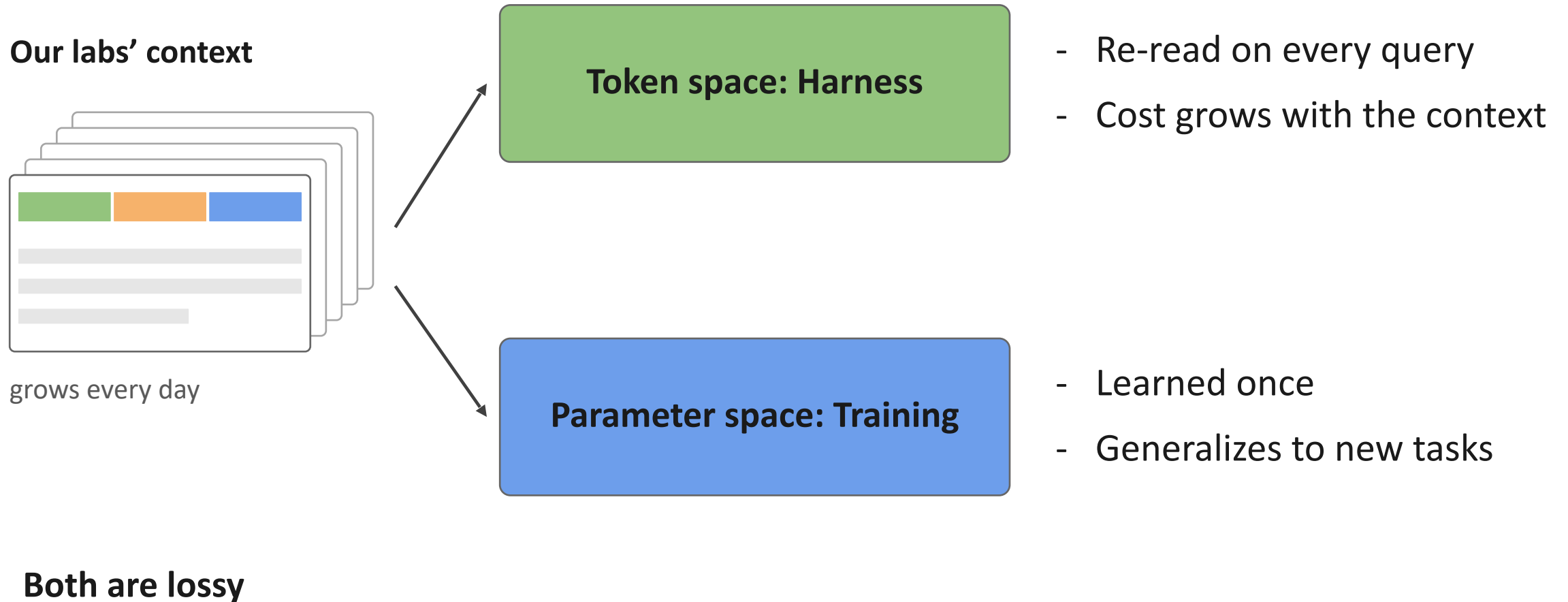


grows every day

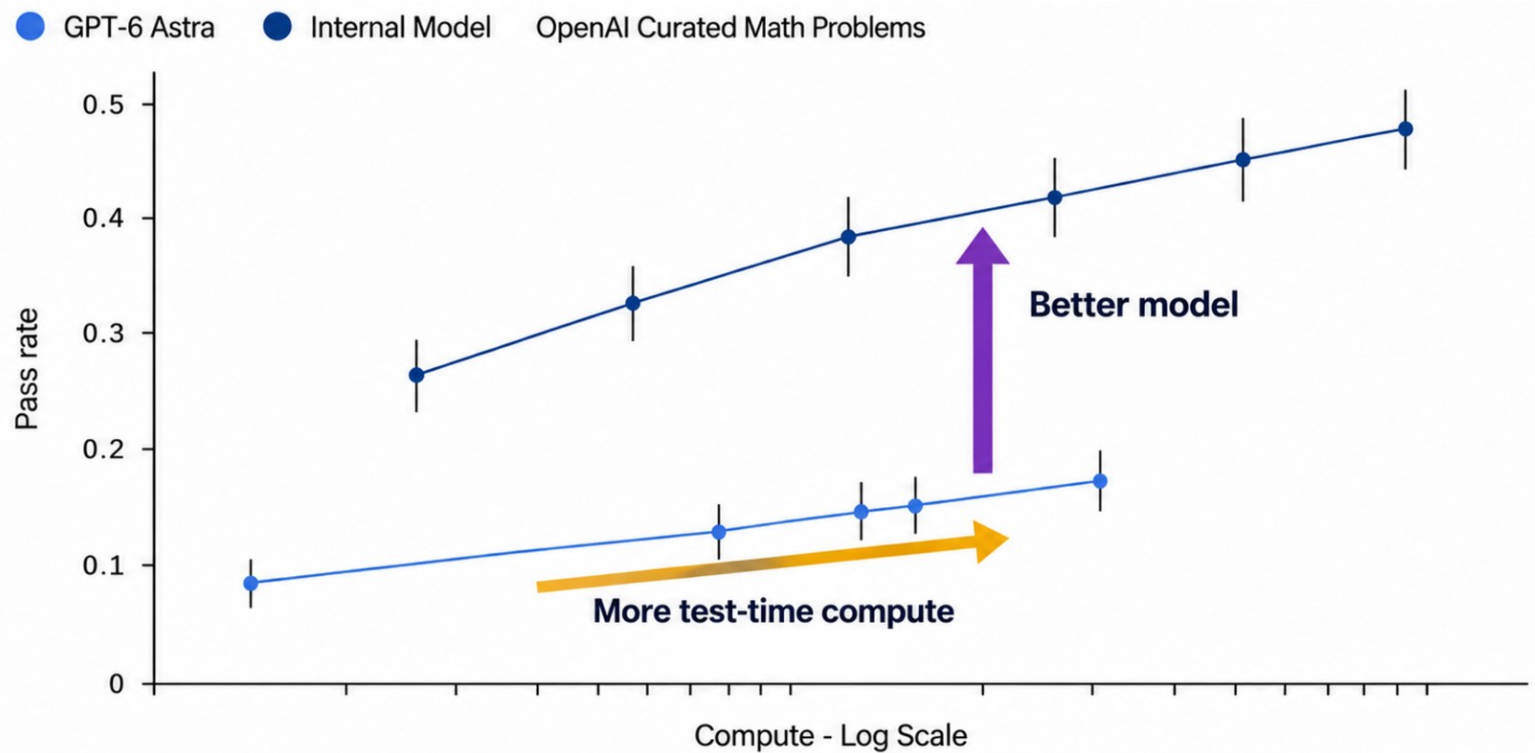
- Midtraining, RL
- Context distillation
- Adapters

- Learn it into the weights, by gradient descent
- Paid once; generalizes to new tasks
- **But slow to write, and it can forget**

Which One Keeps Up as Context Piles Up?



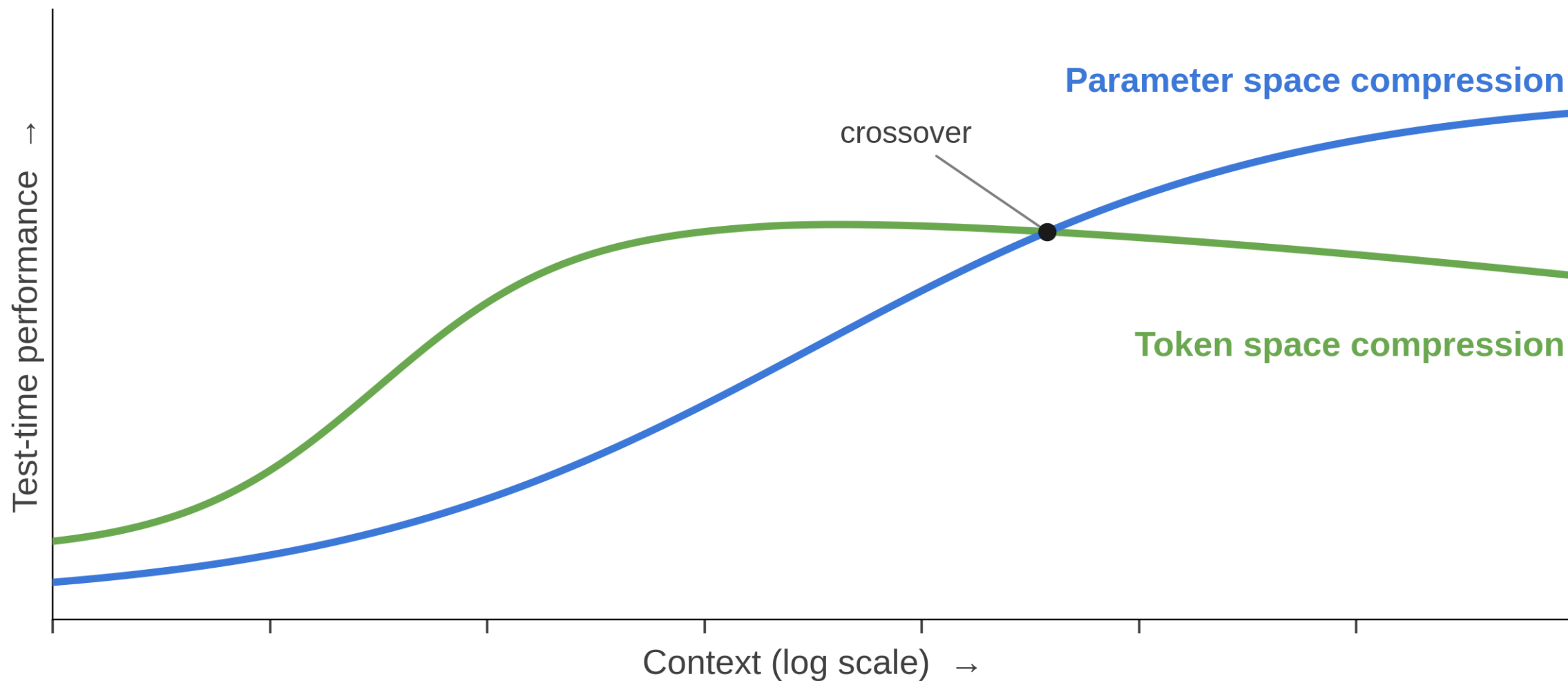
Why Not Stop at GPT-4?



- If harness / test-time compute were enough, we could have stopped at GPT-4
- More test-time compute: **slow gains**
- A better-trained model: **a leap**, and it still scales with test-time compute

Test-time compute is **complementary** to learning

The Crossover: Token vs. Parameter Space Compression



Illustrative

The Bitter Lesson

*“The two methods that seem to scale arbitrarily in this way are **search and learning**.” – Sutton*

Search = test-time compute

- Retrieval, sub-agents, longer reasoning, more samples
- Paid on every query; nothing carries over

Learning = training

- Compresses what's shared into the weights
- Paid once; generalizes to new tasks

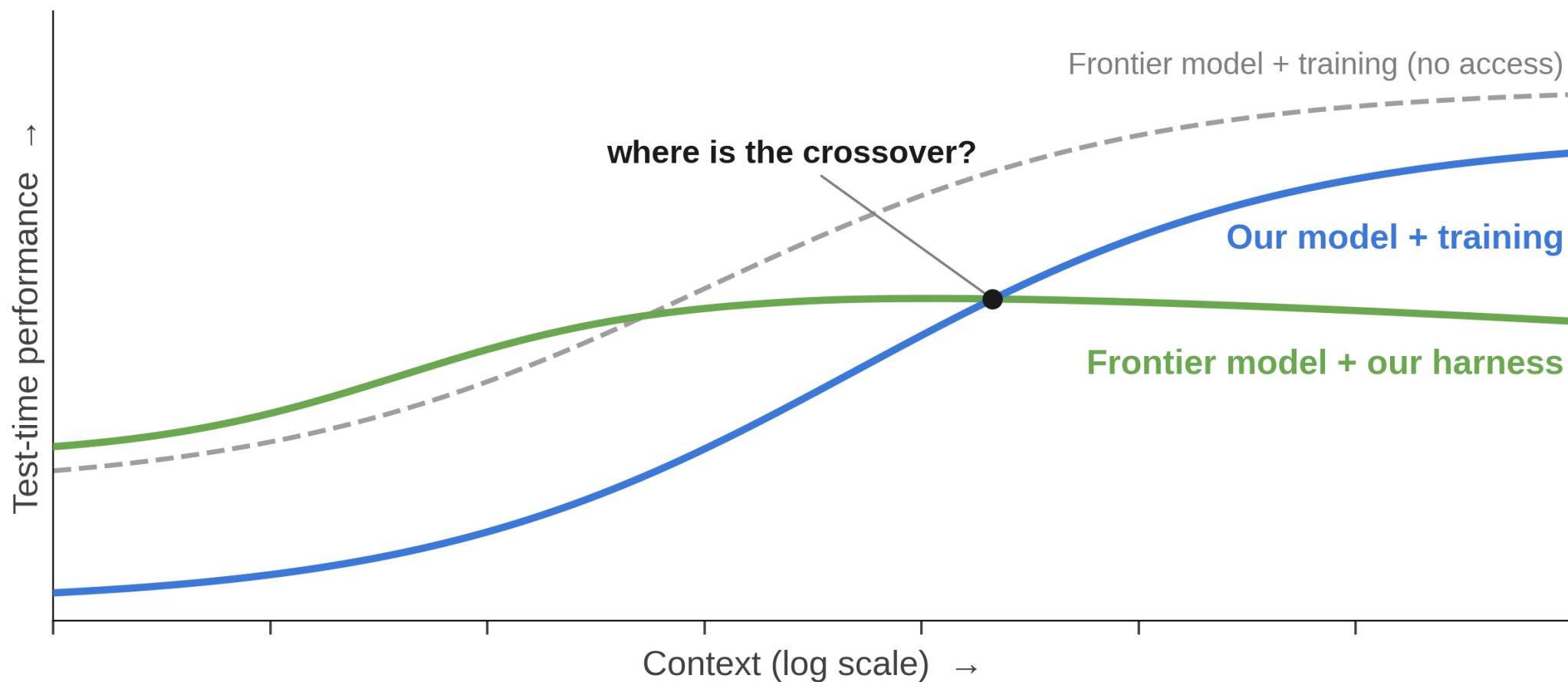
What context to use / retrieve is itself an intelligence problem

The window holds the **task**.

The weights hold the **experience**.

The Billion Dollar Question

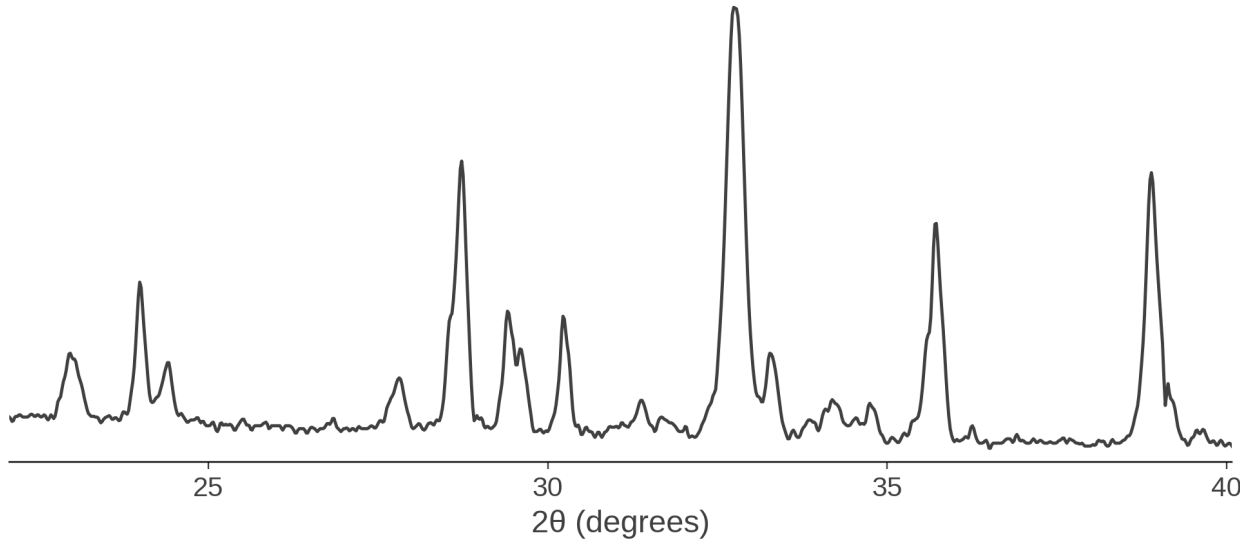
Frontier model + our harness, or **train our own model**?



Illustrative

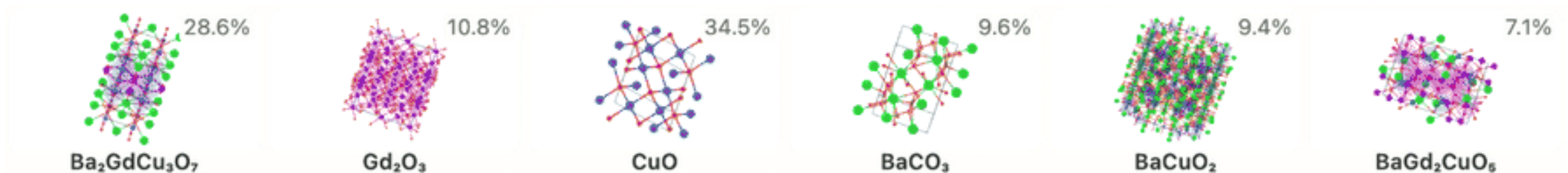
What Did We Actually Make?

A sample comes out unlabeled



- The product can be a mixture of the target, unreacted precursors, and unexpected byproducts
- We need to know what formed, and in what proportions
- At our scale, scientists can no longer keep up by hand

What formed, and how much (Neon's answer)



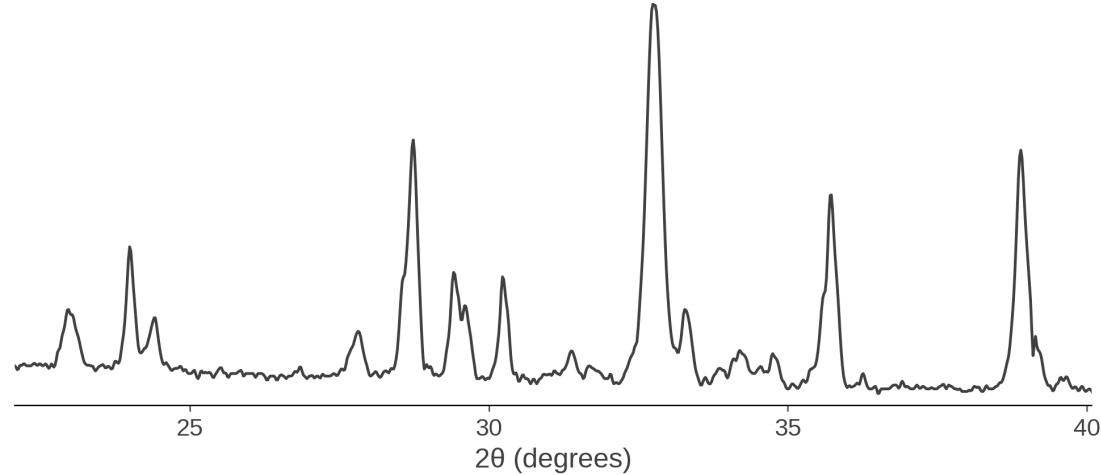
The Right Answer Depends on Context

Precursors

Temperature

Atmosphere

Handling history



Related samples

Literature

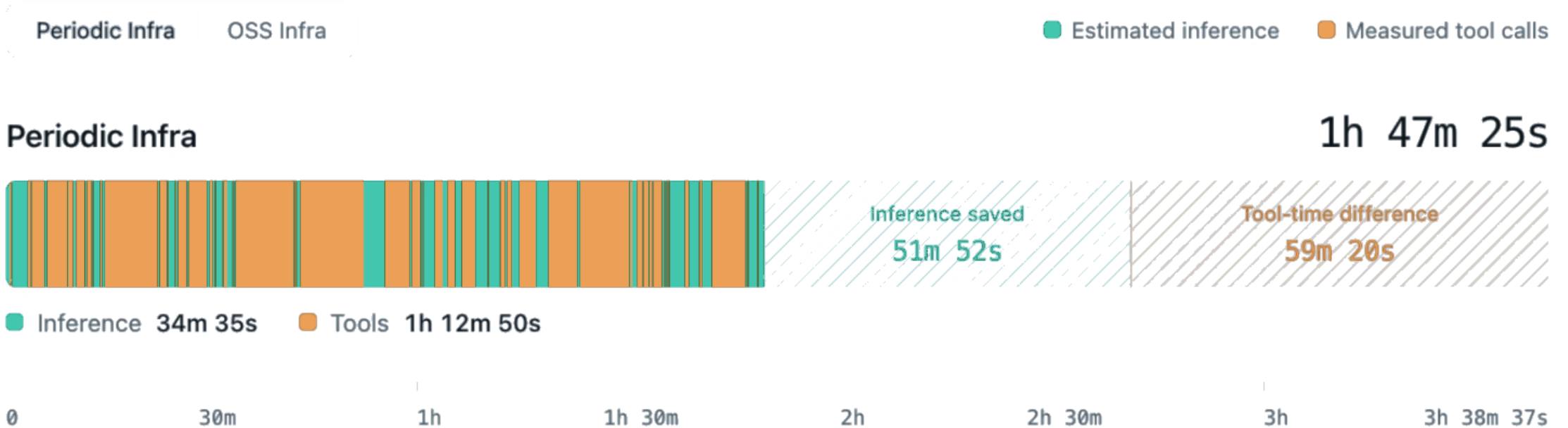
Thermodynamics

Simulations

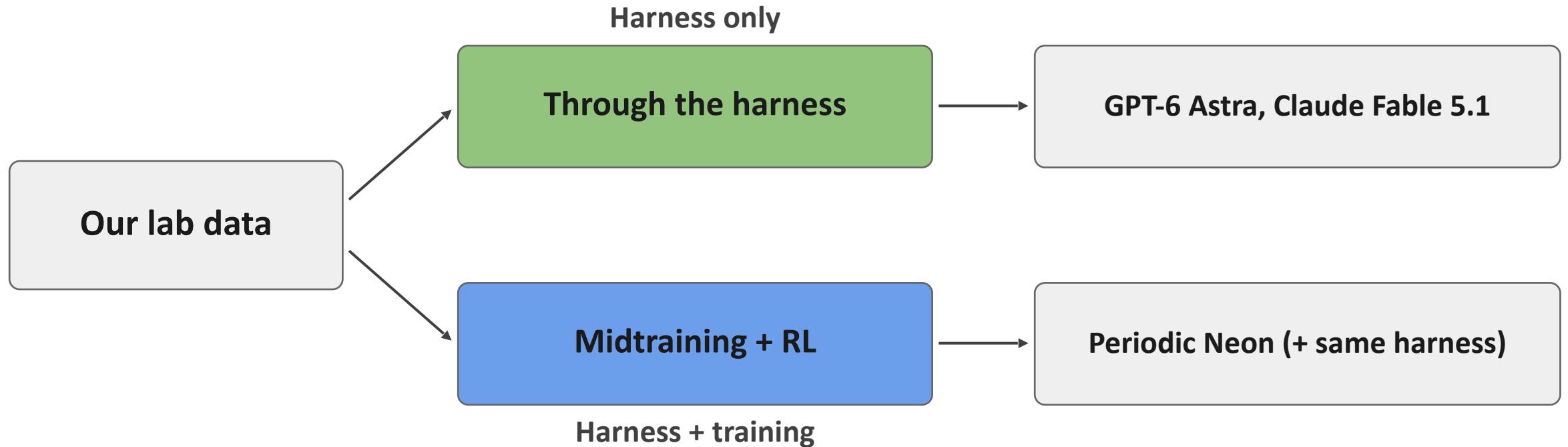
- An oxide may make sense if the sample was heated in air; a hydroxide only if moisture exposure was possible
- A fit can look good mathematically but fail scientifically
- Several hours of expert time per pattern

One XRD Analysis Can Run for Hours

Rollout timing



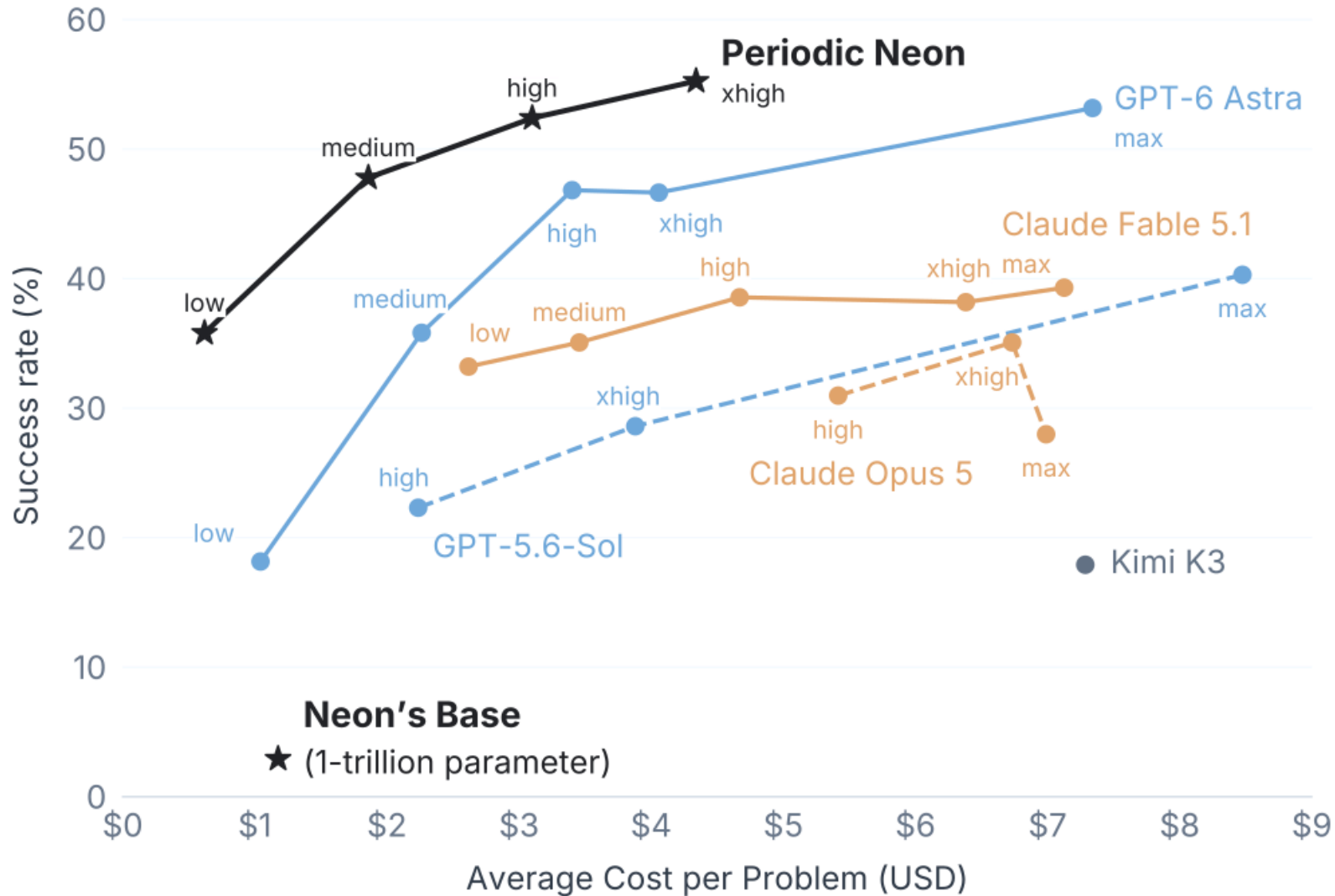
Same Lab Data, Same Harness, Different Weights



The only difference: **did our lab data also go into the weights?**

What context to retrieve is itself an intelligence problem

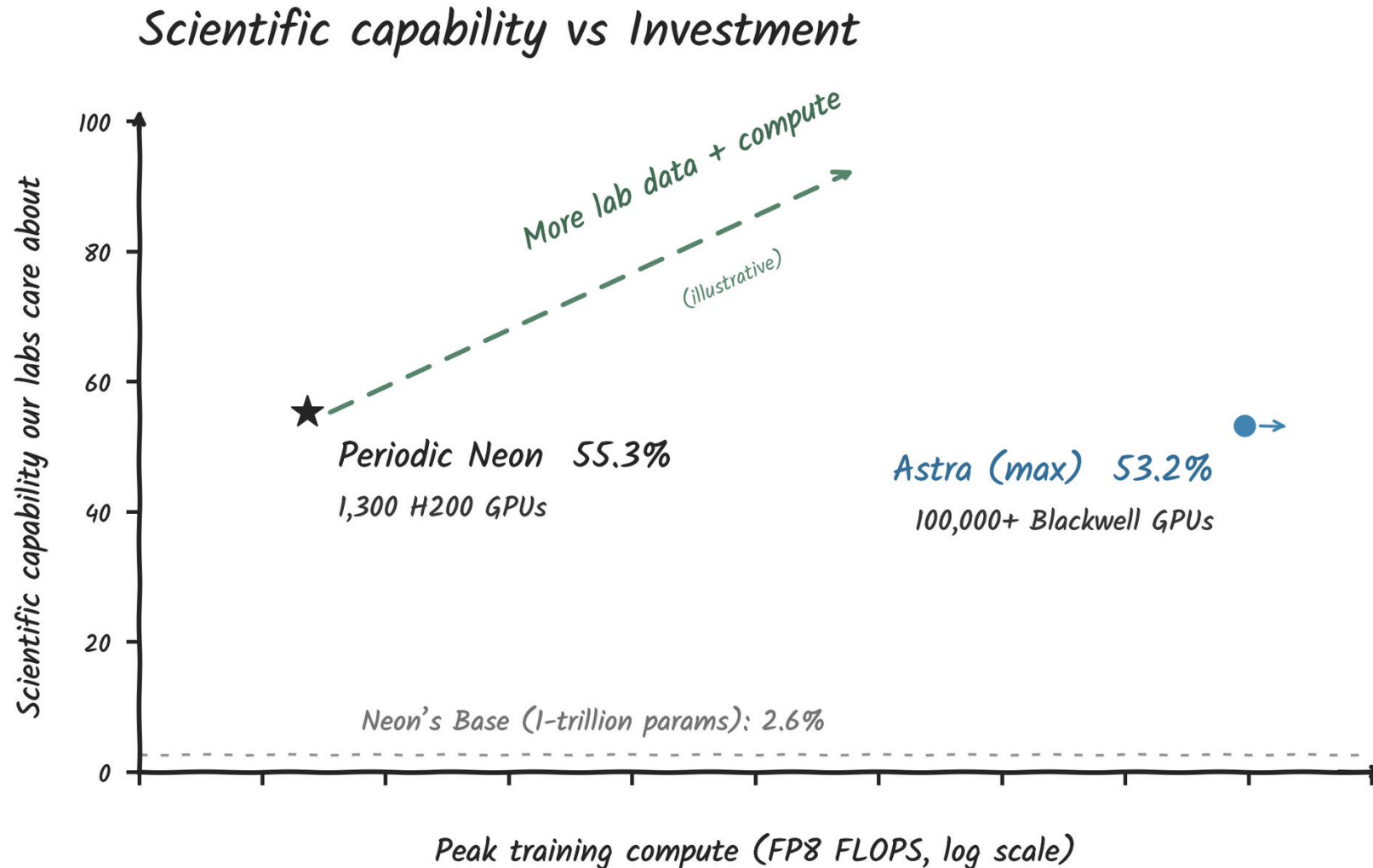
Periodic Neon Surpasses Frontier Models



Kimi K2.6 started at ~3% on our benchmark (unusable).

With midtraining and RL on our lab data, it surpasses GPT-6 Astra, at lower cost.

Compute Efficiency: Frontier-Level with $\sim 77\times$ Fewer GPUs



Watch Neon Analyze One Sample

The model begins by considering if the synthesis precursor and target phases might be candidates based on peak positions.

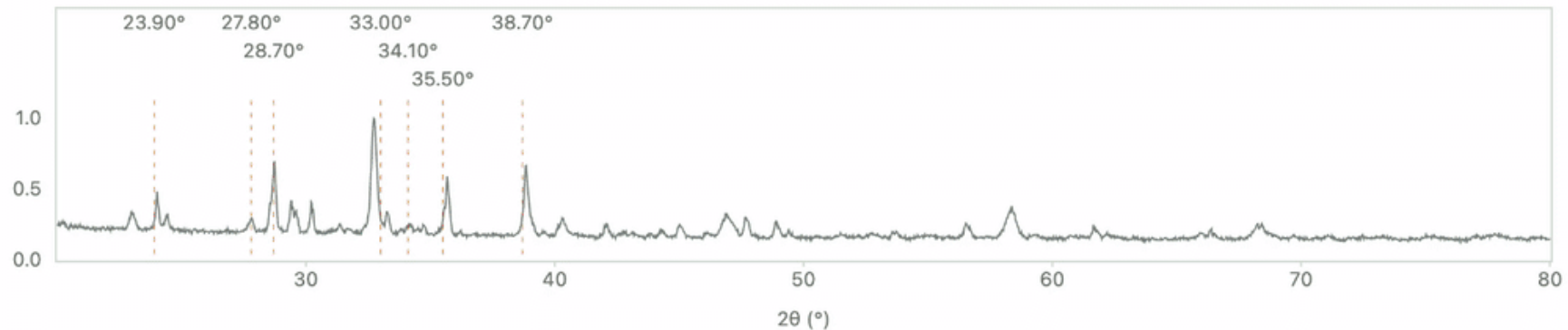


"The cubic form of Gd_2O_3 shows the strongest peak at 28.7° (222), followed by 33.0° (400), then 47.4° and 56.3° . Other peaks like BaCO_3 (witherite) appear at 23.9° , 27.8° , and 34.1° . CuO shows notable peaks too, like 35.5° and 38.7° . This peak pattern suggests incomplete reaction."

Intensity (normalized)

— Observed

— Model illustration

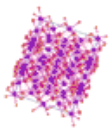


Previous

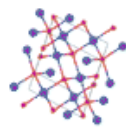
Play

Next

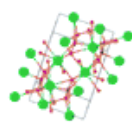
0:05 / 0:54



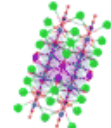
Gd_2O_3



CuO



BaCO_3



$\text{Ba}_2\text{GdCu}_3\text{O}_7$

Candidate phases

The Harness Helps 3.8× on the Same Model

XRD analysis success by harness

Harness-only comparison (Opus 5)

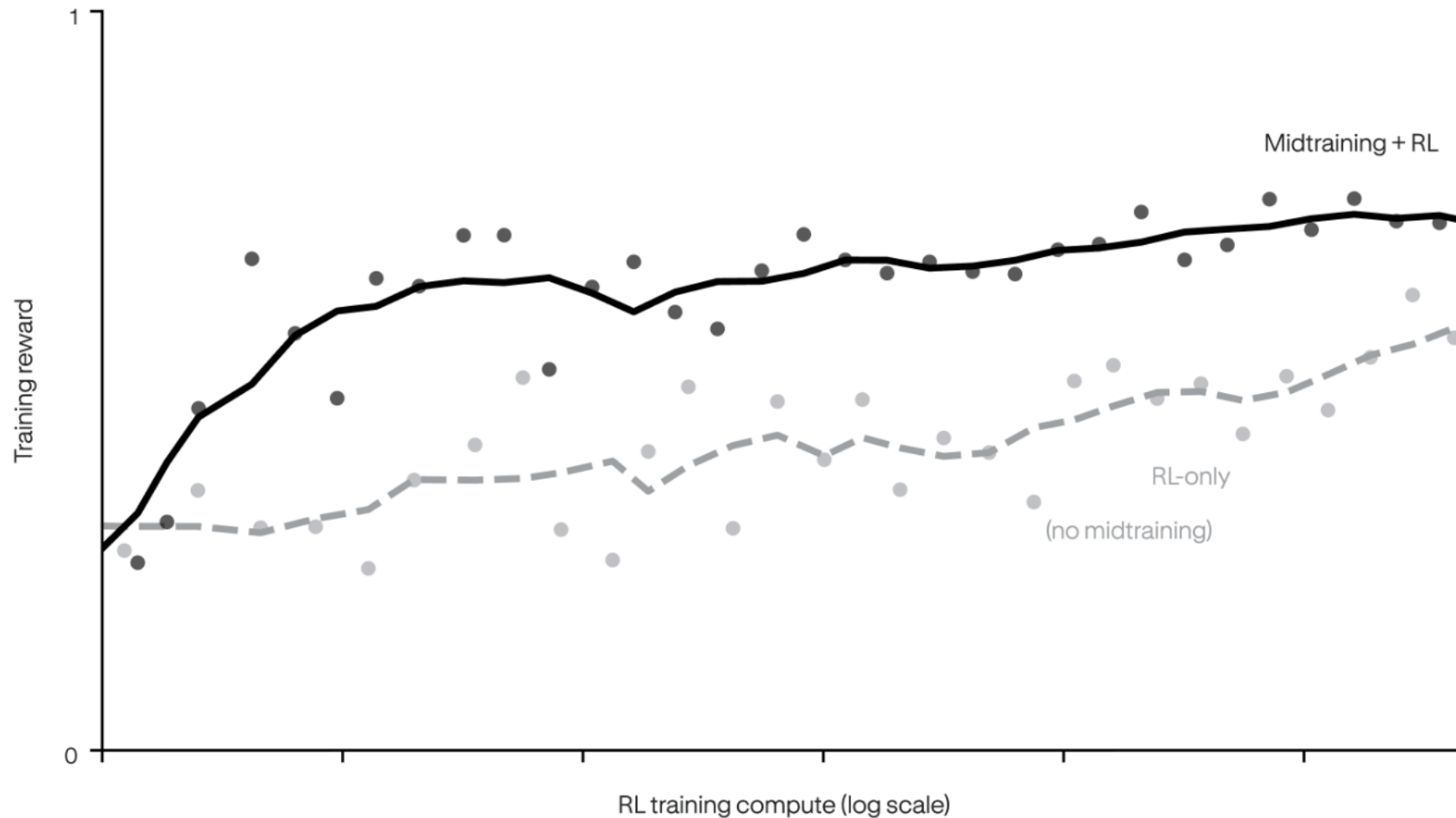


Main difference: in-house structure and simulation databases vs. open-source databases

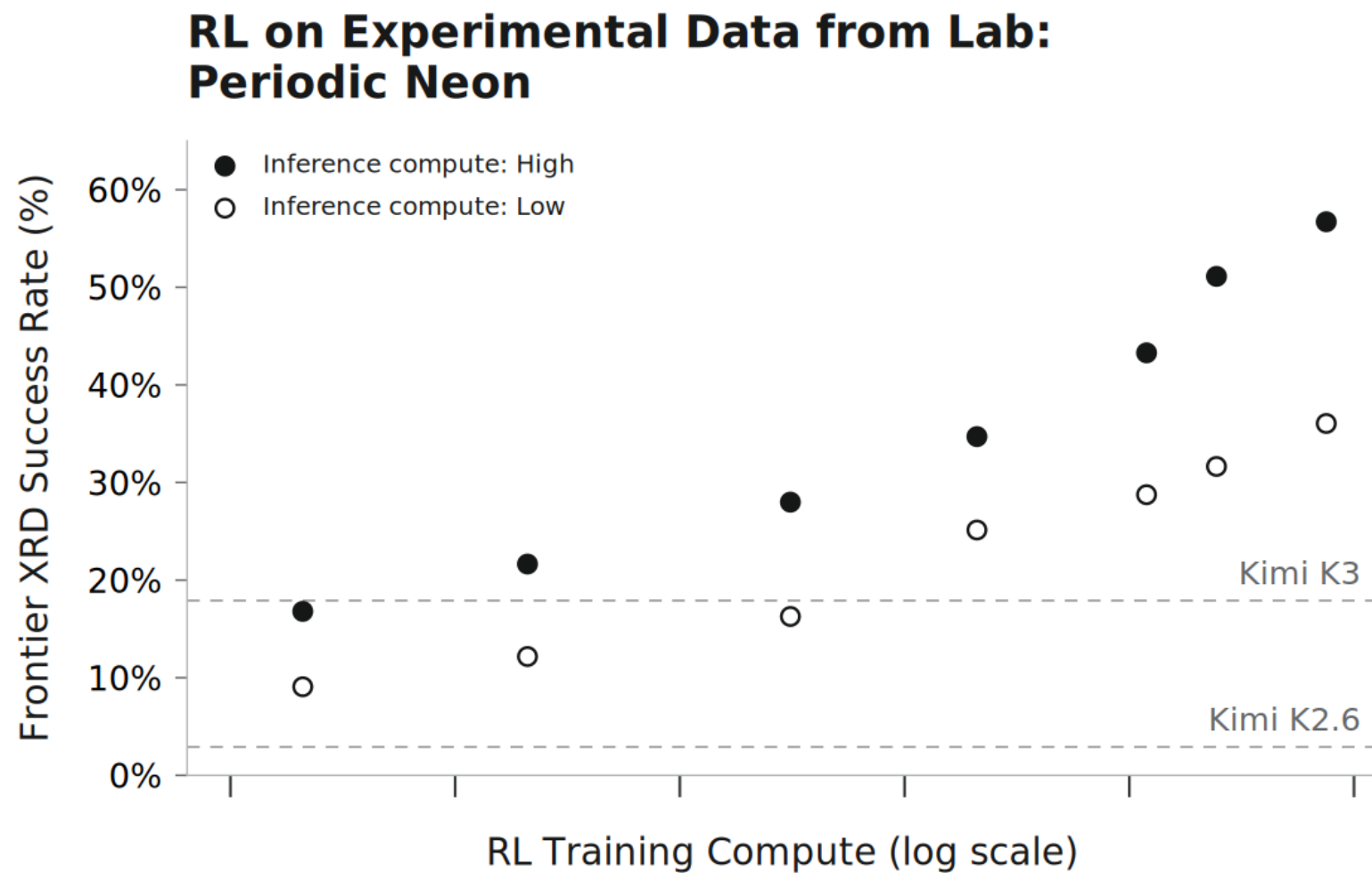
Every model from here on gets this harness

Knowledge in Weights Makes RL Learn Faster

Midtraining improves RL scaling for Periodic Neon

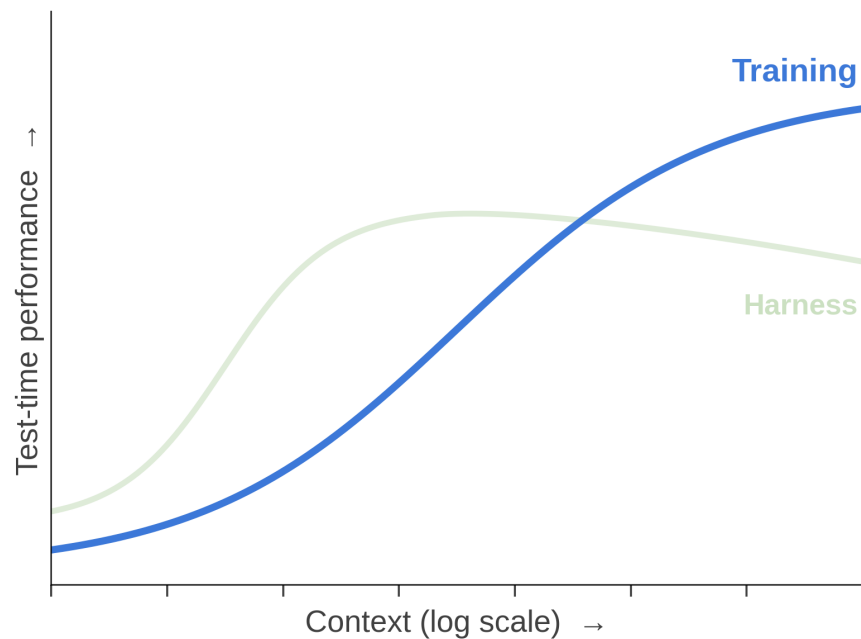


Training Scales with RL Compute

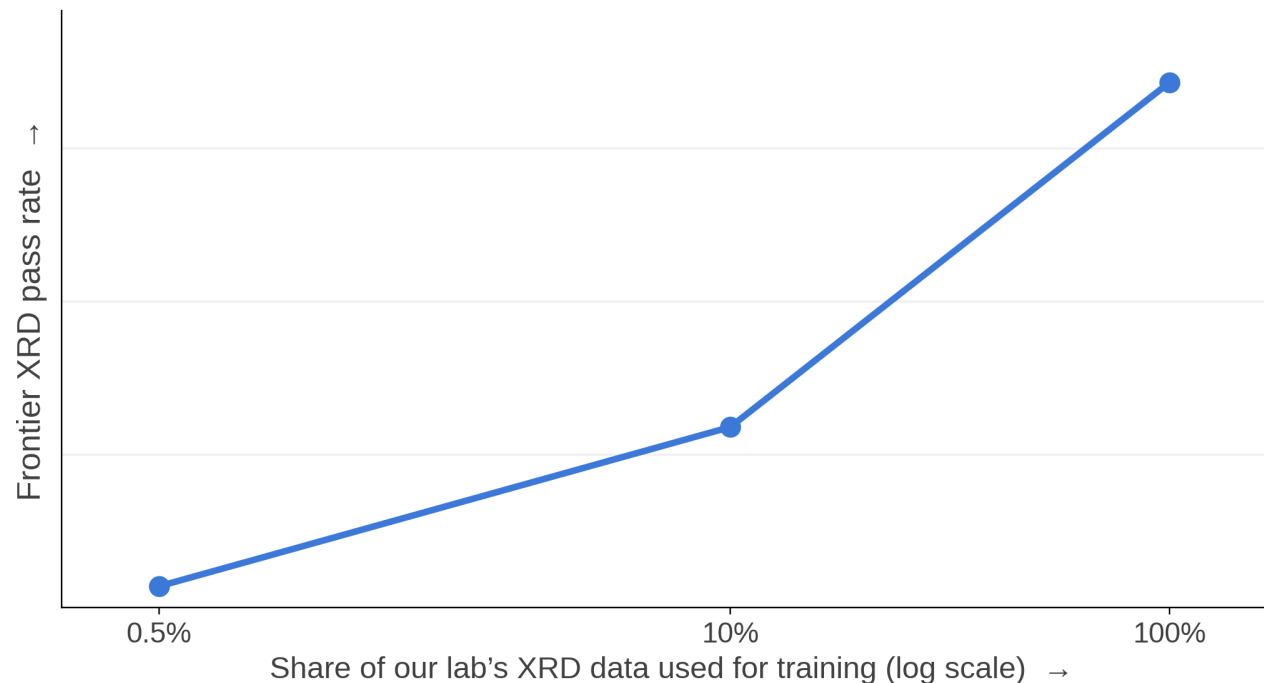


Training Scales with Lab Data

Expected

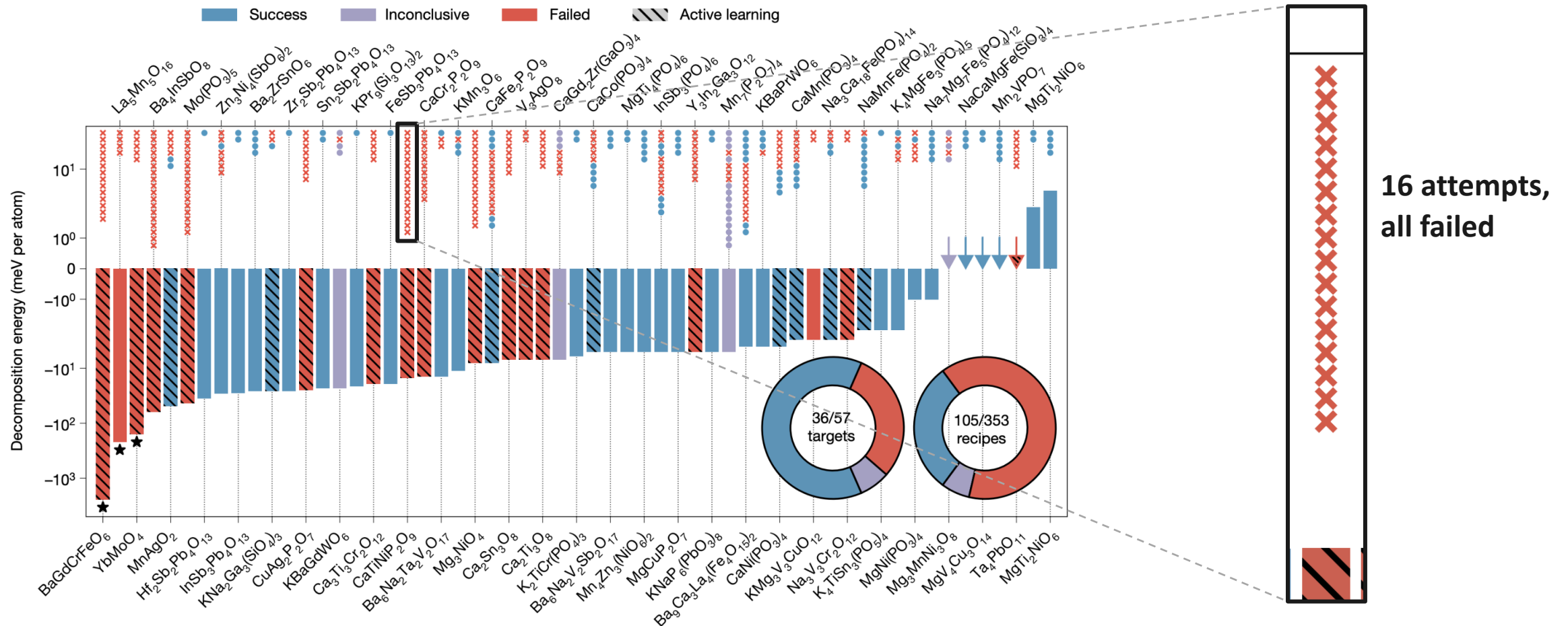


Measured on our lab data



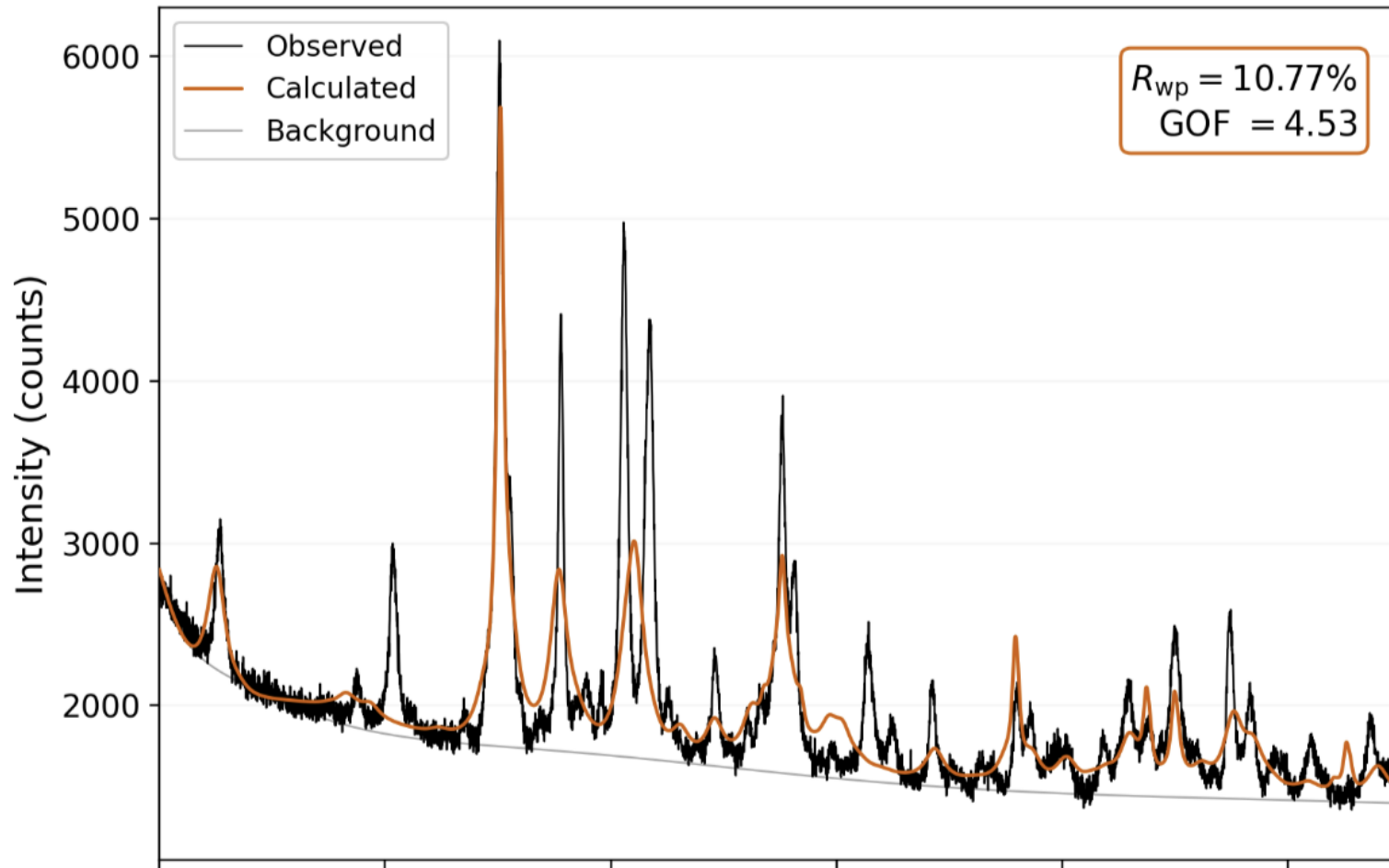
More lab data → **better model**

Closing the Loop on a Target That Failed 16 Times



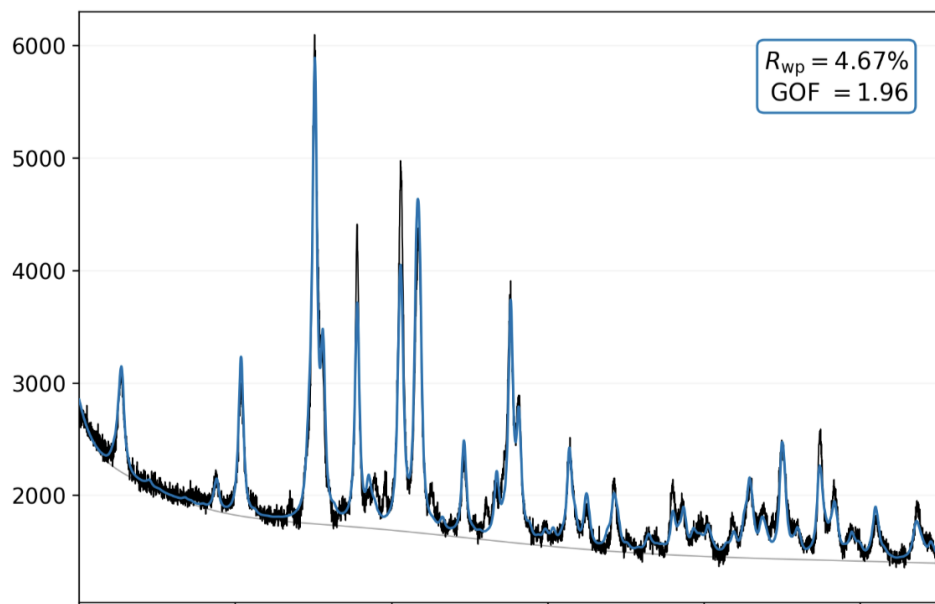
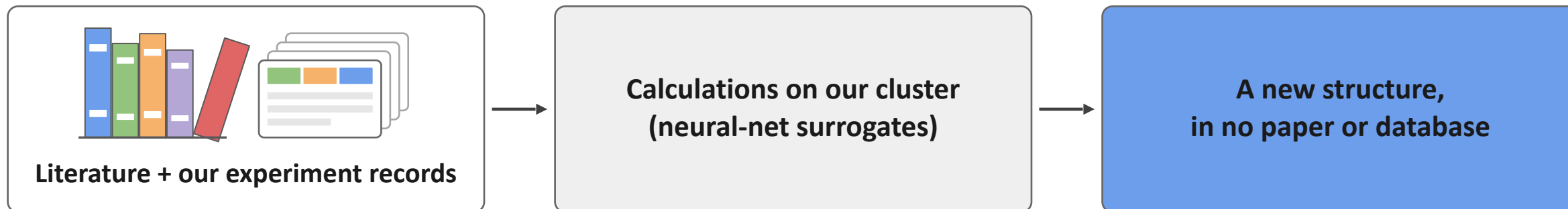
A-Lab, Nature 2023: 17 days of autonomous synthesis, 36 of 57 targets made

The Fit Doesn't Explain the Pattern



- Black: the measured pattern
- Orange: their best fit
- **Peaks missing**; two peaks merged into one
- Fit metrics agree (Rwp 10.8%)

Giving Our AI the Context

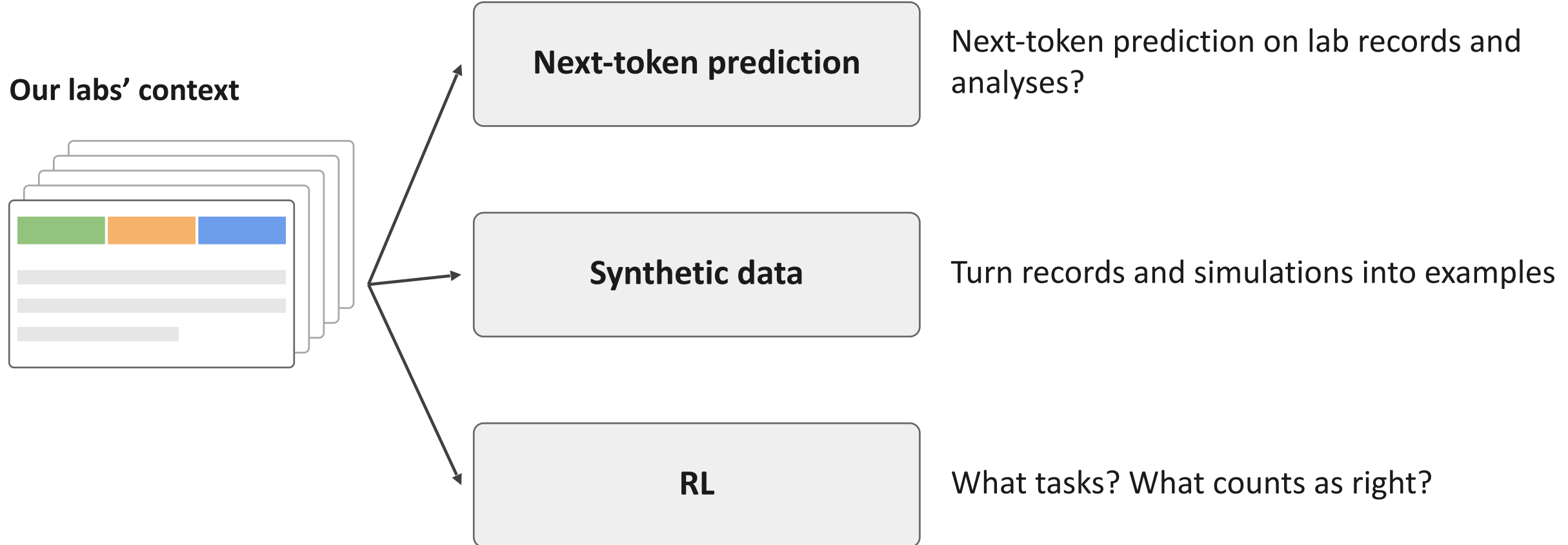


- With it, the fit **explains the data**
- $R_{wp} 10.8\% \rightarrow 4.7\%$
- Now we know “what was made”

It needed the **context**: the pattern, the failed attempts, the chemistry

Challenges

Challenge 1: Best Way to Internalize Our Labs' Context?



RL: Science is **falsifiable**, not easily **verifiable**

Agreement on XRD analysis success

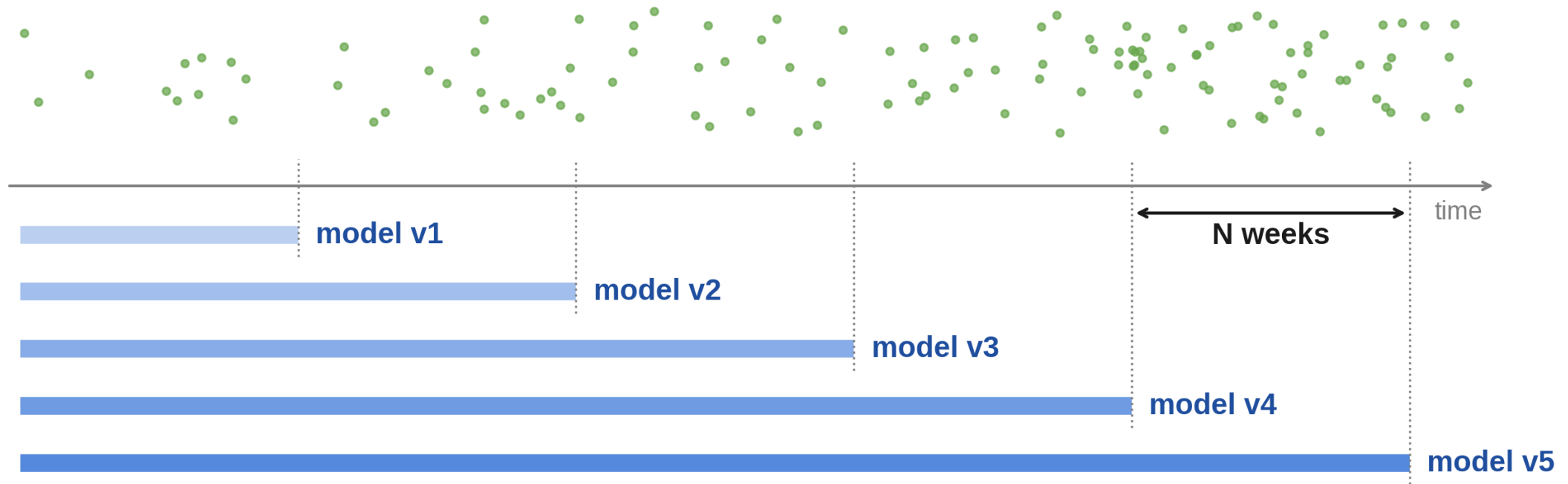


- XRD success can't be verified by pattern-fitting alone: it needs expert judgment
- Experts labeled thousands of patterns, three experts each
- An LLM-judge ensemble agrees with experts **almost as often as experts agree with each other**

Challenge 2: Our labs move right every day. How do the weights keep up?

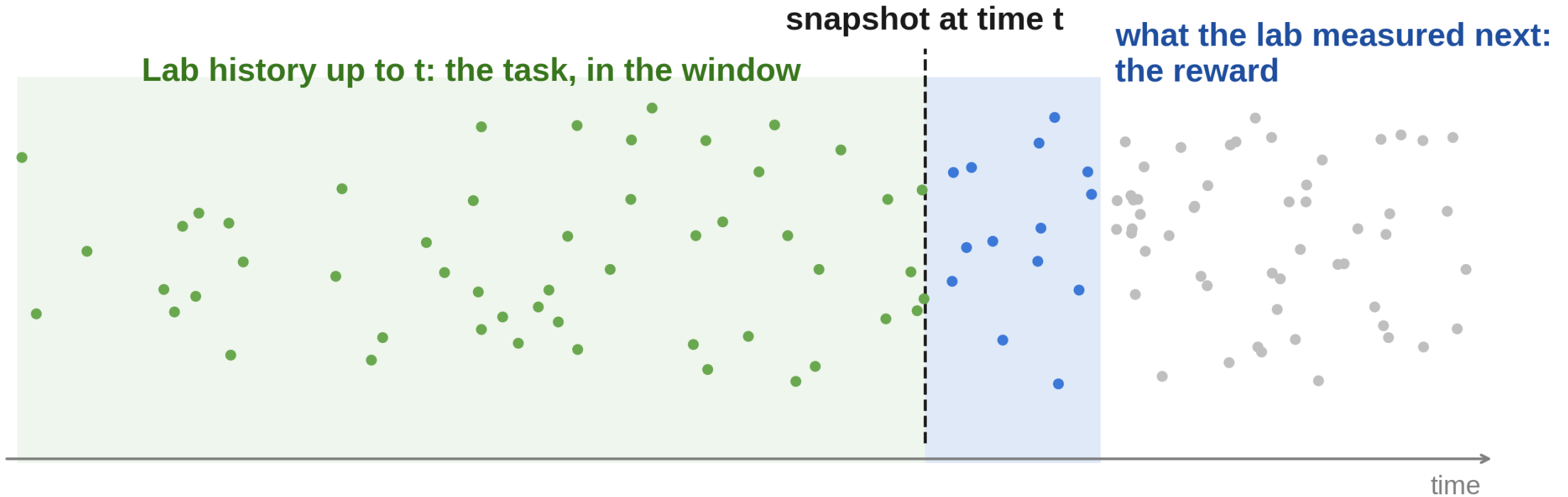
Continual Learning on Our Labs' History

lab data, growing every day



- Retrain **every N weeks** on everything measured so far
- Not at test time, but **one version of continual learning**

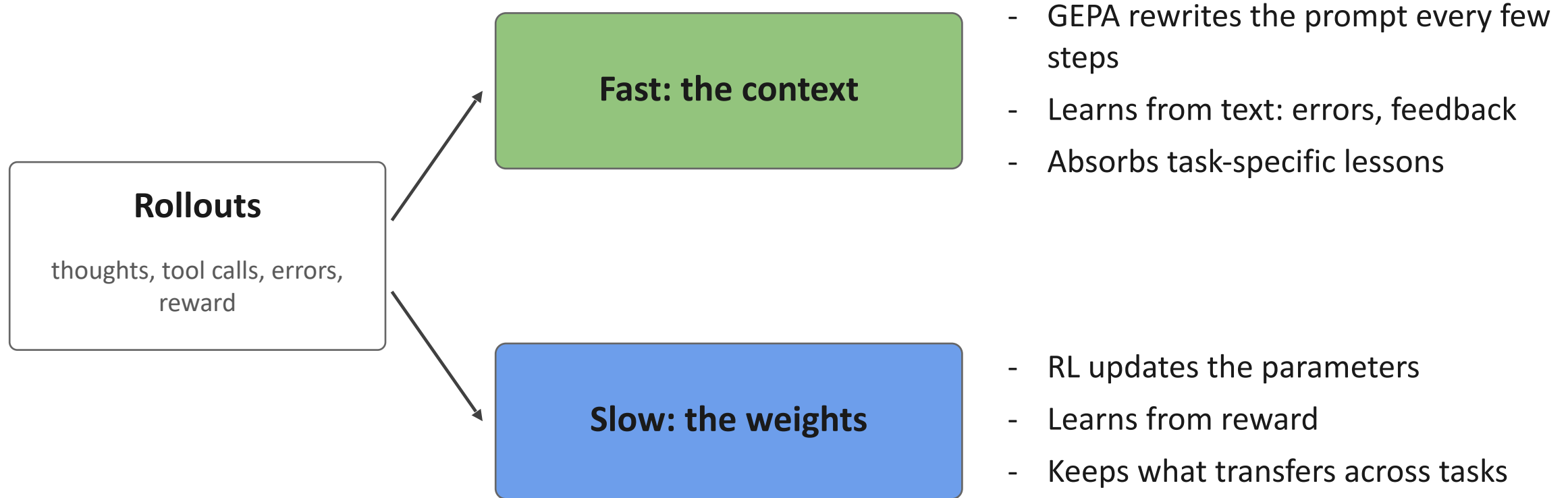
One idea: Predicting the Future from Our Labs' Past



Our labs' history becomes **RL environments**, future outcomes provide the learning signal

Challenge 3: How do we adapt both the context and the weights?

One Approach: Learning at Two Timescales



An old idea: complementary learning systems (fast hippocampus, slow neocortex)

The window holds the **task**.

The weights hold the **experience**.

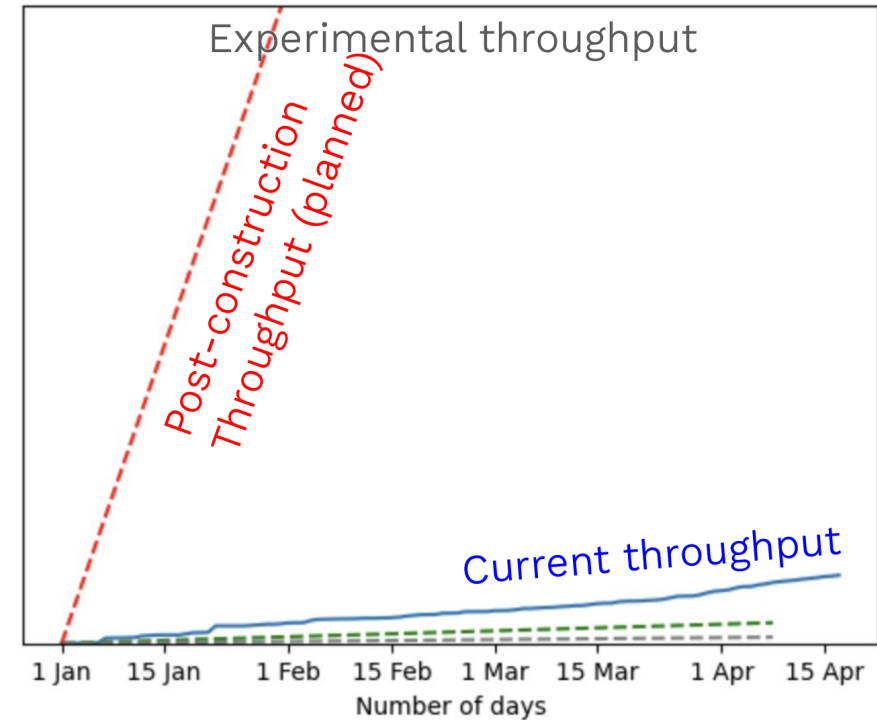
The Context Will Keep Growing, Much Faster



One of our labs, earlier this summer

Our high throughput labs are **still being built**

After construction (planned)



Thank You!

- Periodic Neon: the ML, infrastructure, and the science teams at Periodic Labs

