Human Level AI by 2030?

Jared Kaplan
Anthropic & JHU

My Journey…

- **Started working on AI in ~2018, and my perspective slowly shifted.**
- **Versions of this talk evolving since 2019…**

Physicists

My Journey…

- Started working on AI in ~2018, and my perspective slowly shifted
- Version of this talk circa 2021…

Scaling Laws, GPT-3, and Self-Supervision

My Journey…

- Started working on AI in ~2018, and my perspective slowly shifted
- Now…

Human Level AI by 2030?

My Journey…

- Started working on AI in ~2018, and my perspective slowly shifted
- Next year…

BUY CANNED GOODS AND TIN FOIL HATS

Main Ideas

- AI progress is occurring because it's **very easy** — driven predominantly by increasing resource inputs, with complex algorithms replaced by "scaling up"
- Progress is predictable as there are Scaling Laws wrt dataset size, neural network parameter count, and total compute used for training (which is now >>Avogadro's number of floating point operations!)
- The largest AI systems are approaching human level; challenges exist for eg literally automating science, but no compelling blockers at all (in my opinion)
- This is all **pretty scary** for obvious, common sense reasons. Maybe it is the biggest thing ever, and we're sleepwalking through it as a scientific community.

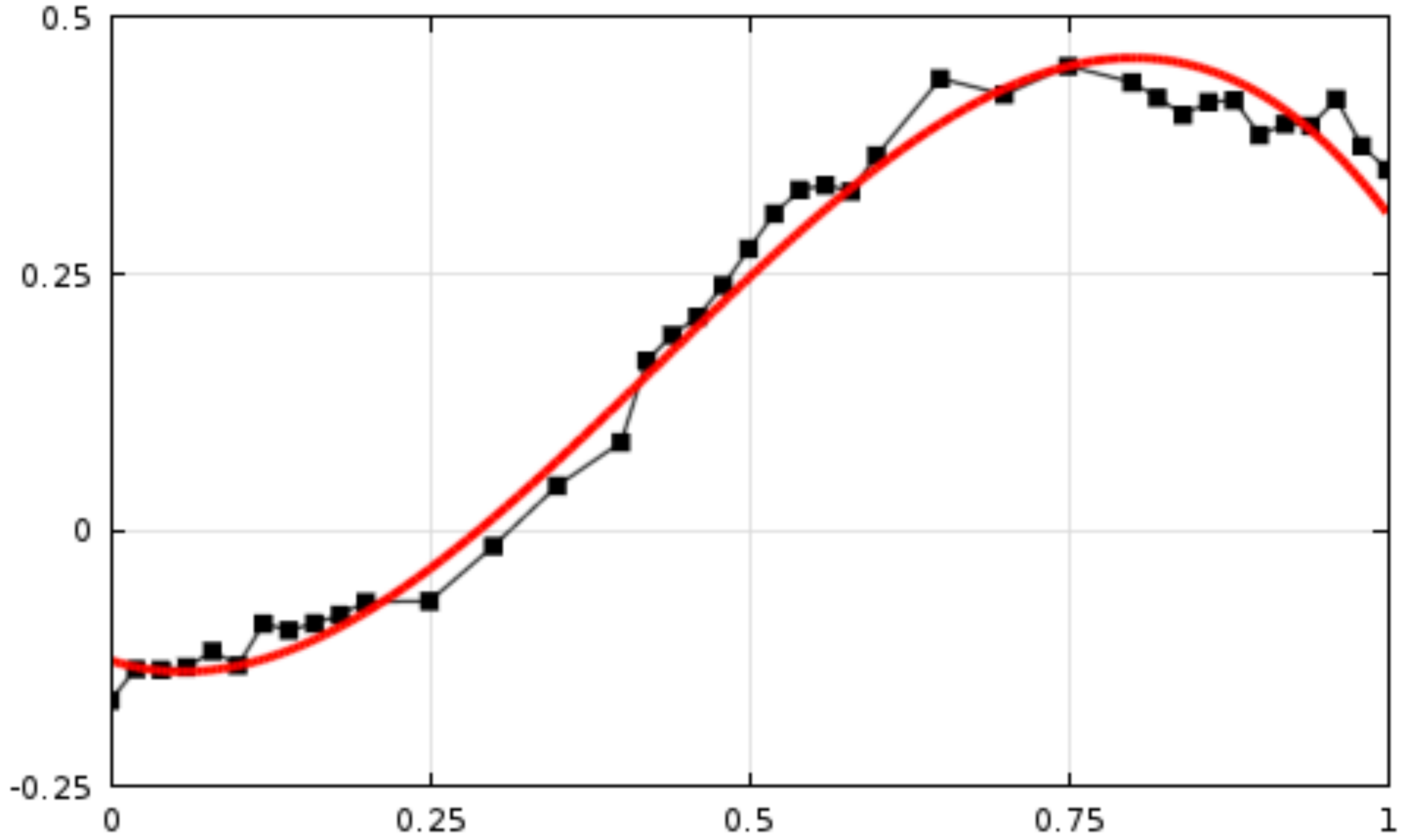
Outline

- Briefly explain contemporary AI / deep learning, and why it's **very** simple
- Brief overview of "RLHF" and Constitutional AI
- Why AI capabilities will predictably improve
- Remaining Challenges for AI (to be broadly human-level)
- Why this is scary…

Very Quick AI Need-to-Knows

What is Contemporary AI / ML?

- Just **curve fitting** with a **very** general function approximation



- **Neural Network** = sequence of matrix multiplications and element-wise non-linearities, living in a very high dimensional (eg 10k) vector space

Language Models

Learn language by “autoregressive” **prediction of the next word**:

$$P(w_{n+1}|w_n, \cdots w_1)$$

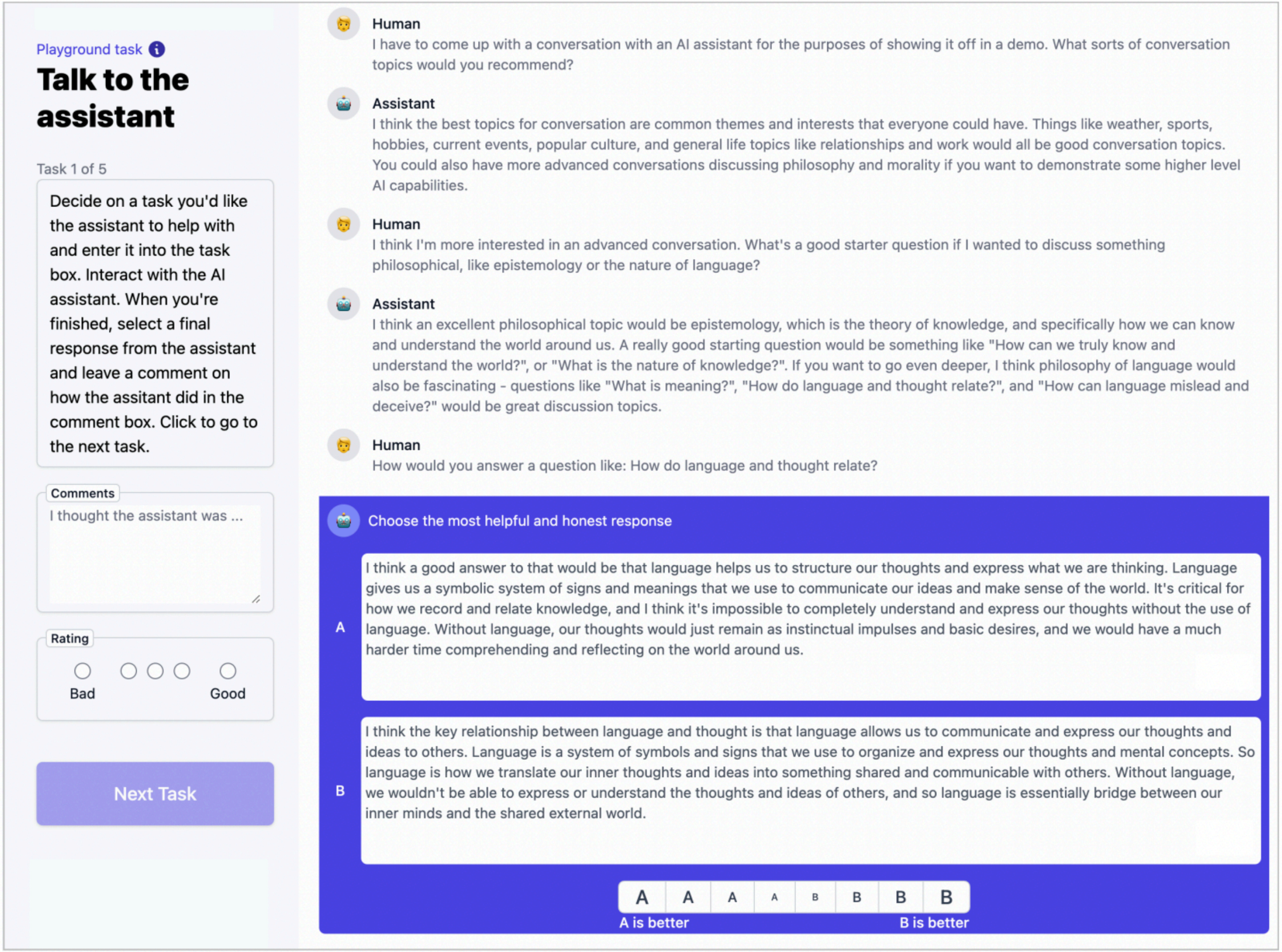
Optimize the log-**likelihood** of real world text, which provides the “right answer”.

As a speaker at a journal club, you're probably elephant me to say certain things.

not = 15.10%
going = 13.59%
used = 12.37%
expected = 3.74%
already = 3.53%
familiar = 2.76%
more = 2.57%
thinking = 1.94%
a = 1.64%
aware = 1.48%
elephant = 0.00%

Total: -17.70 logprob on 1 tokens
(58.73% probability covered in top 11 logits)

$$P \approx 10^{-8}$$


Playground task
Talk to the assistant
Task 1 of 5
Decide on a task you'd like the assistant to help with and enter it into the task box. Interact with the AI assistant. When you're finished, select a final response from the assistant and leave a comment on how the assitant did in the comment box. Click to go to the next task.
Comments
I thought the assistant was ...
Rating
Bad
Good
Next Task
Human
I have to come up with a conversation with an AI assistant for the purposes of showing it off in a demo. What sorts of conversation topics would you recommend?
Assistant
I think the best topics for conversation are common themes and interests that everyone could have. Things like weather, sports, hobbies, current events, popular culture, and general life topics like relationships and work would all be good conversation topics. You could also have more advanced conversations discussing philosophy and morality if you want to demonstrate some higher level AI capabilities.
Human
I think I'm more interested in an advanced conversation. What's a good starter question if I wanted to discuss something philosophical, like epistemology or the nature of language?
Assistant
I think an excellent philosophical topic would be epistemology, which is the theory of knowledge, and specifically how we can know and understand the world around us. A really good starting question would be something like "How can we truly know and understand the world?", or "What is the nature of knowledge?". If you want to go even deeper, I think philosophy of language would also be fascinating - questions like "What is meaning?", "How do language and thought relate?", and "How can language mislead and deceive?" would be great discussion topics.
Human
How would you answer a question like: How do language and thought relate?
Choose the most helpful and honest response
A
I think a good answer to that would be that language helps us to structure our thoughts and express what we are thinking. Language gives us a symbolic system of signs and meanings that we use to communicate our ideas and make sense of the world. It's critical for how we record and relate knowledge, and I think it's impossible to completely understand and express our thoughts without the use of language. Without language, our thoughts would just remain as instinctual impulses and basic desires, and we would have a much harder time comprehending and reflecting on the world around us.
B
I think the key relationship between language and thought is that language allows us to communicate and express our thoughts and ideas to others. Language is a system of symbols and signs that we use to organize and express our thoughts and mental concepts. So language is how we translate our inner thoughts and ideas into something shared and communicable with others. Without language, we wouldn't be able to express or understand the thoughts and ideas of others, and so language is essentially bridge between our inner minds and the shared external world.
A A A A B B B B
A is better
B is better

Summary of Contemporary AI Training

- First “Pretrain” LLMs to predict “the next word” in an enormous corpus of text
- Then collect Human Feedback… or just AI self-feedback based on a list of principles… to train a “Preference Model” that assigns scores to behaviors
- Finally use Reinforcement Learning to “finetune” the LLM to write responses that get high scores according to the preference model when given diverse inputs

Claude

Liouville Theory: An Introduction to Rigorous Approaches

Sourav Chatterjee[1] **and Edward Witten**[2]

[1] *Department of Statistics, Stanford University, Stanford CA 94305 USA*
[2] *School of Natural Sciences, Institute for Advanced Study, Princeton, NJ 08540 USA*

Abstract: In recent years, a surprisingly direct and simple rigorous understanding of quantum Liouville theory has developed. We aim here to make this material more accessible to physicists working on quantum field theory.

Witten2404.02001

PDF

JK What are some ideas explained in this paper that might be new to theoretical physicists who already have some familiarity with Liouville theory?

This paper explains several ideas from the rigorous probabilistic approach to Liouville theory that may be unfamiliar or provide new insights to theoretical physicists:

1. The proof that the random Liouville measure :e^(2bϕ): vanishes if b>1, so the probabilistic approach does not suffice to define Liouville theory for b≥1.
2. A detailed analysis of which moments E[M_f(A)^p] of the integrated Liouville measure M_f(A) converge or diverge depending on the values of b and p. This provides a rigorous understanding of some nonperturbative poles of the DOZZ formula.
3. An explanation of why the DOZZ formula vanishes when one of the Liouville momenta reaches the Seiberg bound at Q/2. This is because the measure M_f(A) becomes infinite with probability 1, making its negative powers vanish.
4. A proof that the Liouville measure M_f(A) converges to a nontrivial limit as the cutoff is removed for all b<1, not just b<1/sqrt(2) which is easier to analyze. This establishes existence of the theory for b<1.
5. A rigorous framework for defining non-integer moments of M_f(A), which correspond to the Liouville correlators. This goes beyond calculations of moments for integer p.

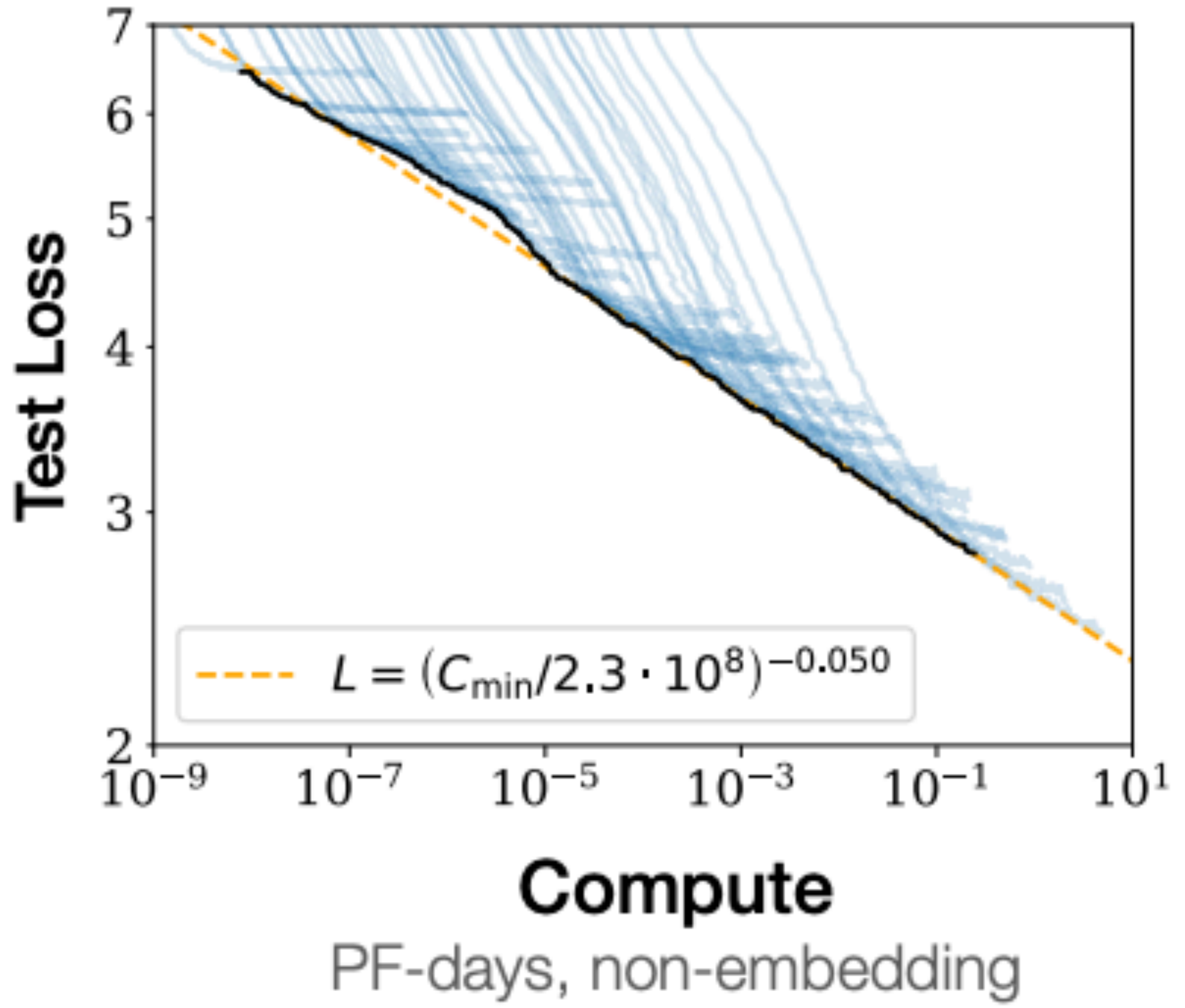
So in summary, it provides a mathematically rigorous perspective on analyticity properties, existence and non-existence for different ranges of b, and subtle issues like vanishing at the Seiberg bound, that complement the physicists' understanding based on path integrals and conformal bootstrap.

Debug Copy Retry

Why AI Progress is Easy and Predictable

Scaling Laws for Language Models

Just train many models of different sizes on different sized datasets:



Compute ~ Params * Tokens Processed, PF-day ~ 9×10^{19}

These types of scaling laws are quite universal and are observed in most ML systems.

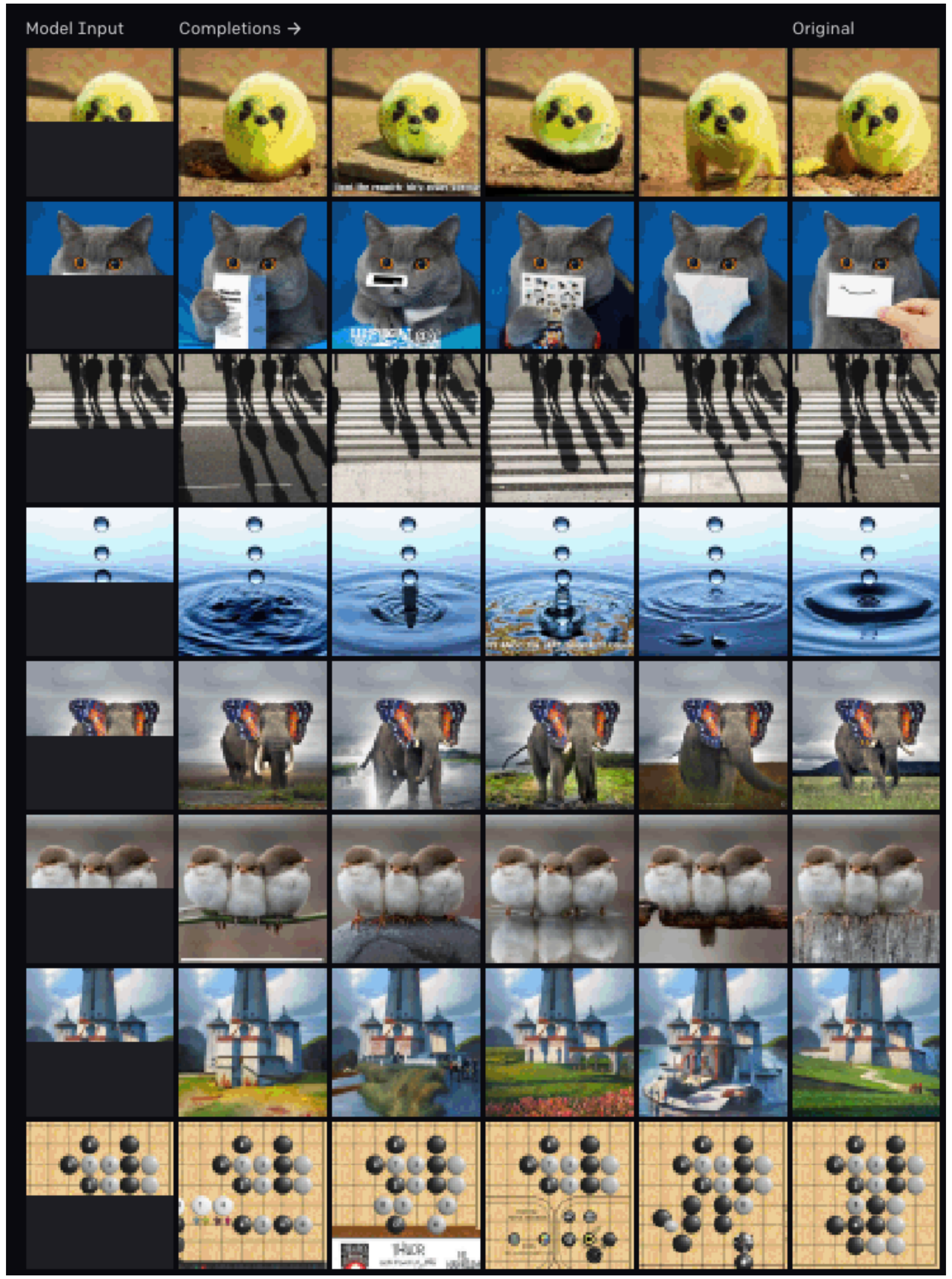
“Grand Unification” in AI
Identical Model & Loss, Different Data

Shadows on the Way
Wallace Stevens

The twilight burns with fire
And one by one the fires expire;
But I, beneath the window, breathe
On my accustomed place,
Until the west is emptied of desire
And I of memory.

Invisible now, the one that's gone
And that one gone before
Are in the tender distance made
More endurable.

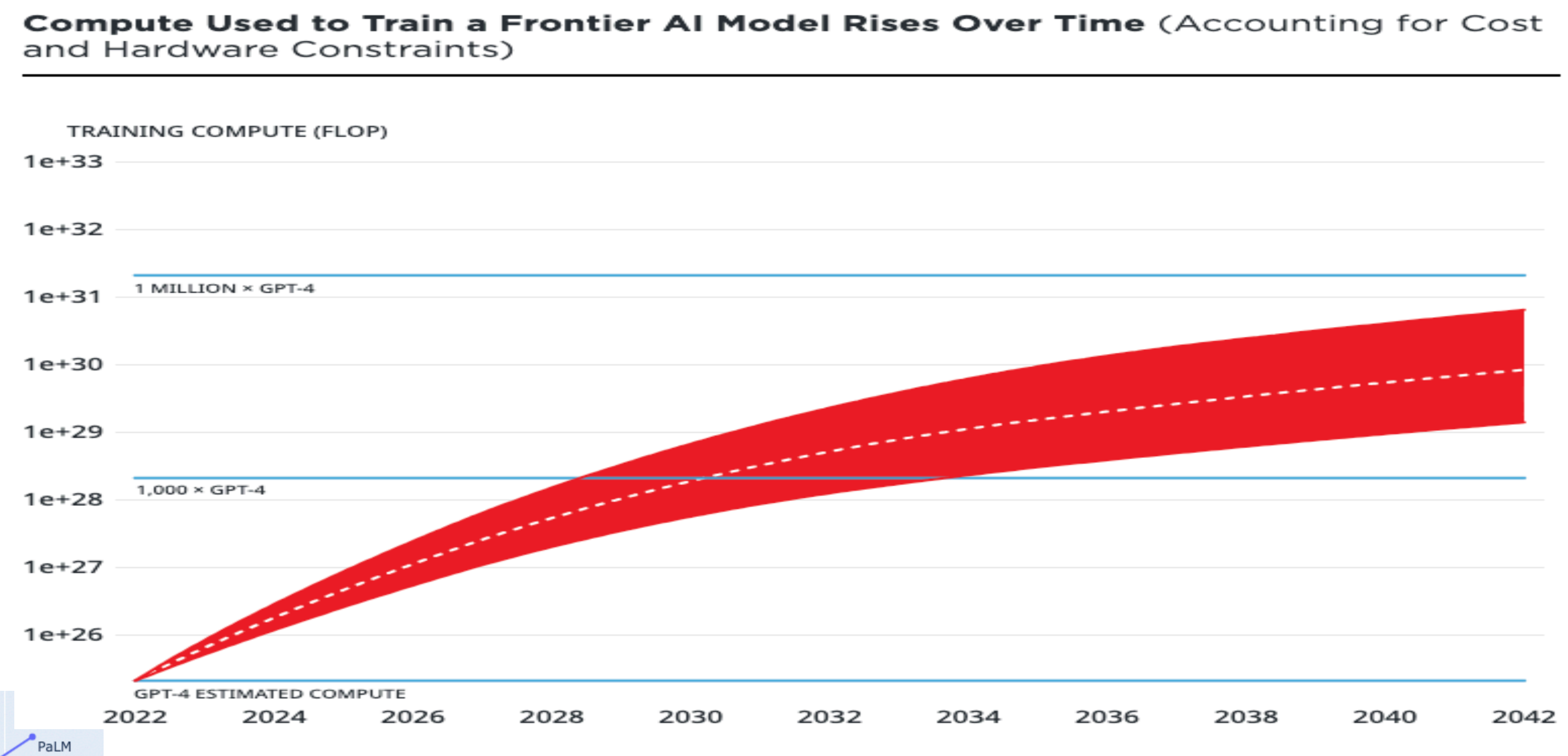
But as for me, I wait
And that which I have always known
Stirs like a bird, and flies:
For in the dark of the window pane
I see the streets, the lighted door,
And I wait, where I have always waited,
Under the window, on the way.

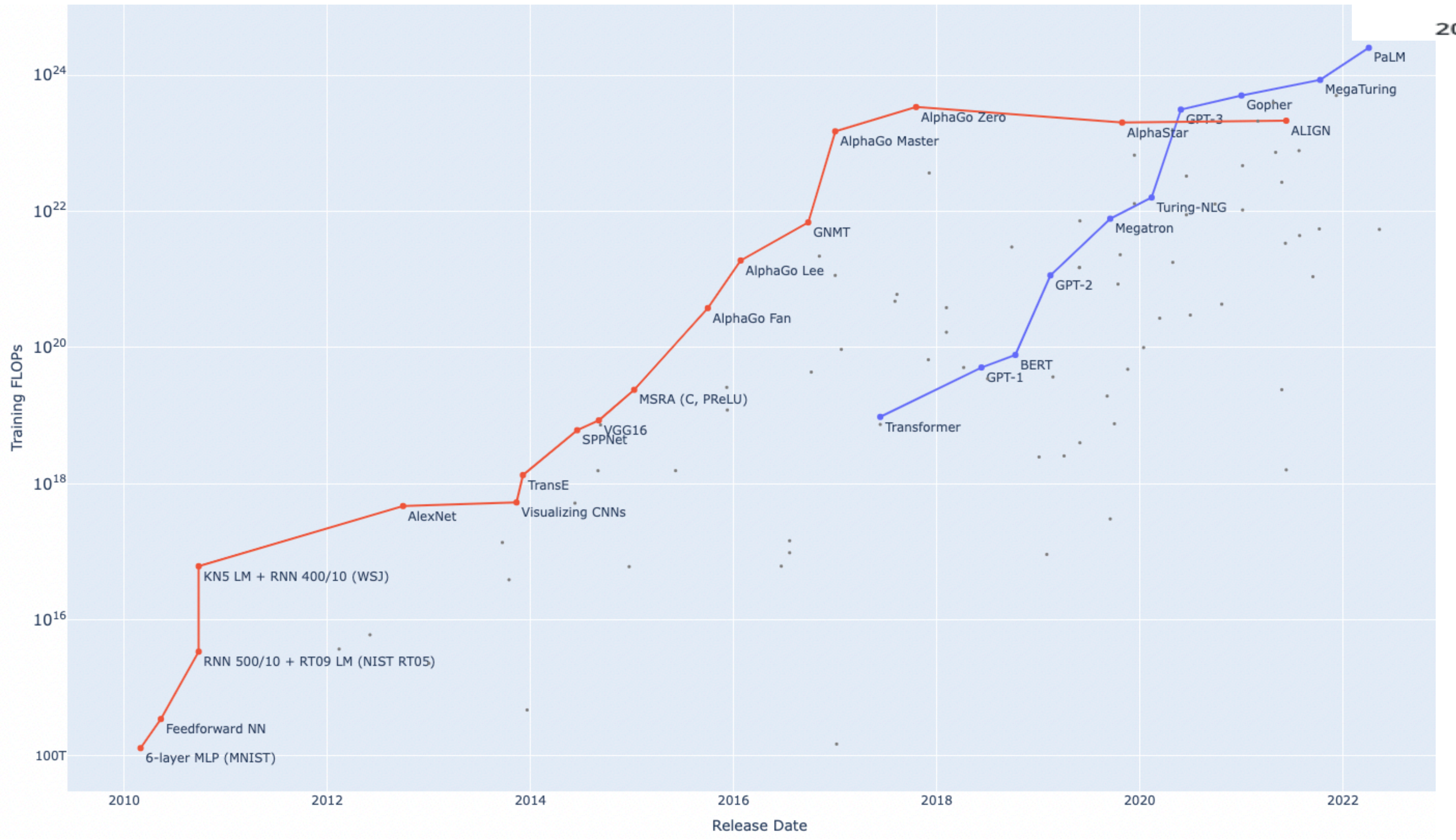


Representative Scales in AI

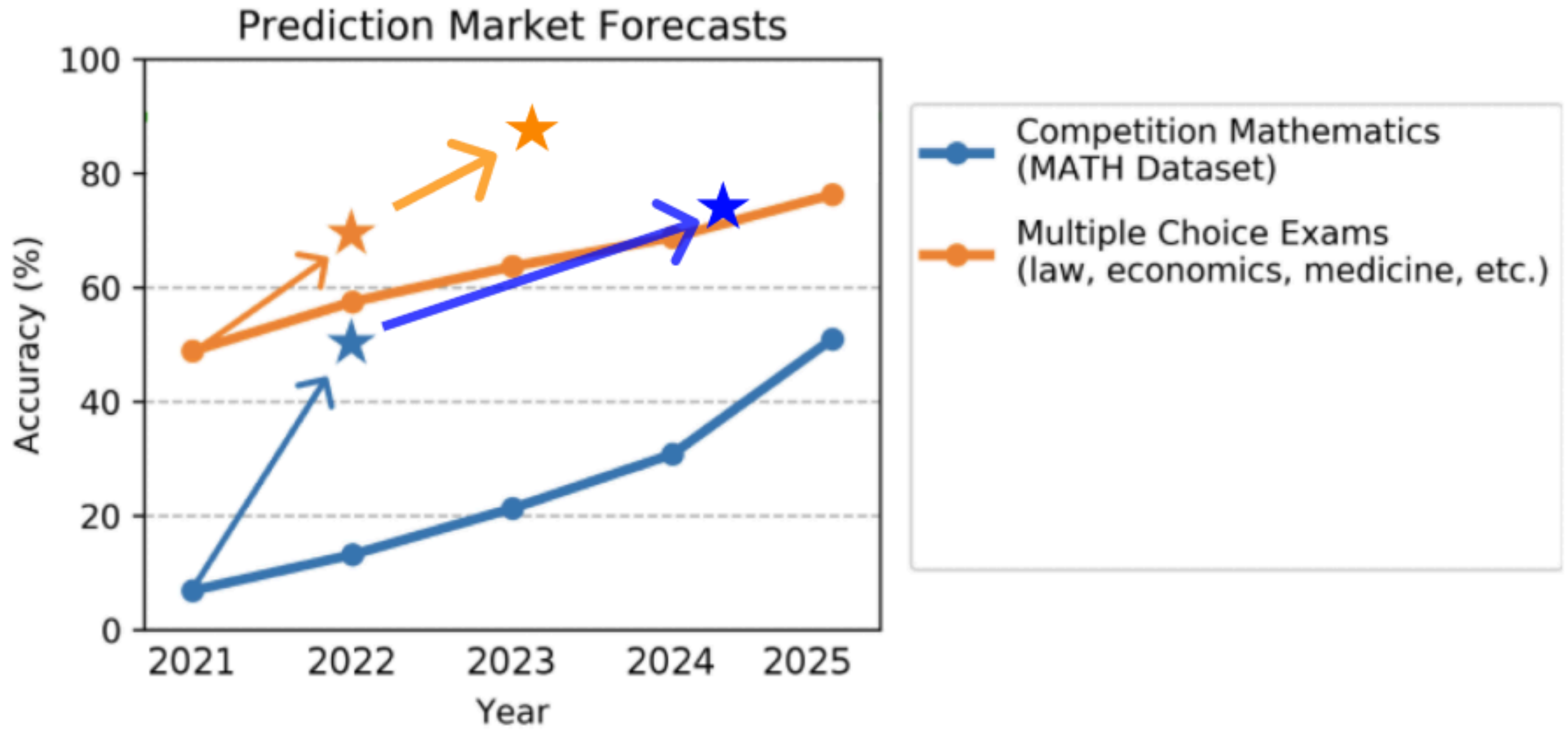
- Neural Networks with **~1e12 parameters** are becoming common, for example with ~100s of layers and ~10k x 10k matrices.
- …and are trained on datasets with **~1e13 words** (eg for Language Models)
- “entire human-generated internet” ~ **1e15 words**
- State of the art GPUs (NVIDIA H100) perform **~1e15 flop/s**
- Largest training runs already use **~1e25** floating point operations total
- For comparison…
  - ultra naive human brain estimate is ~1e15 “parameters” (~synapses)
  - lifetime compute of the brain is then ~3e24 “operations” (synapses fire ~1/sec)

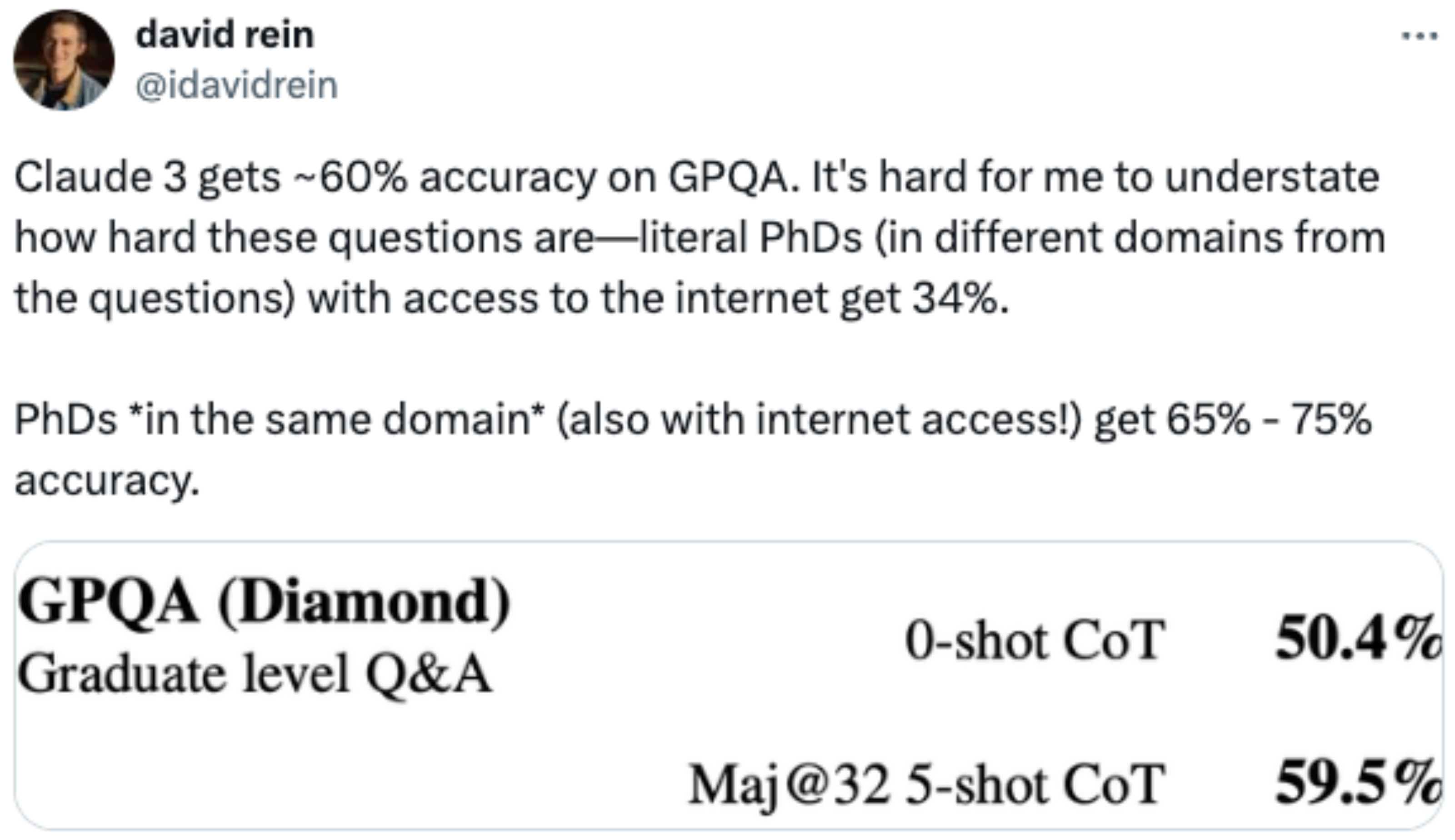
Compute Usage in ML Training





Progress is Fast… Often faster than Researchers and Forecasters Expect



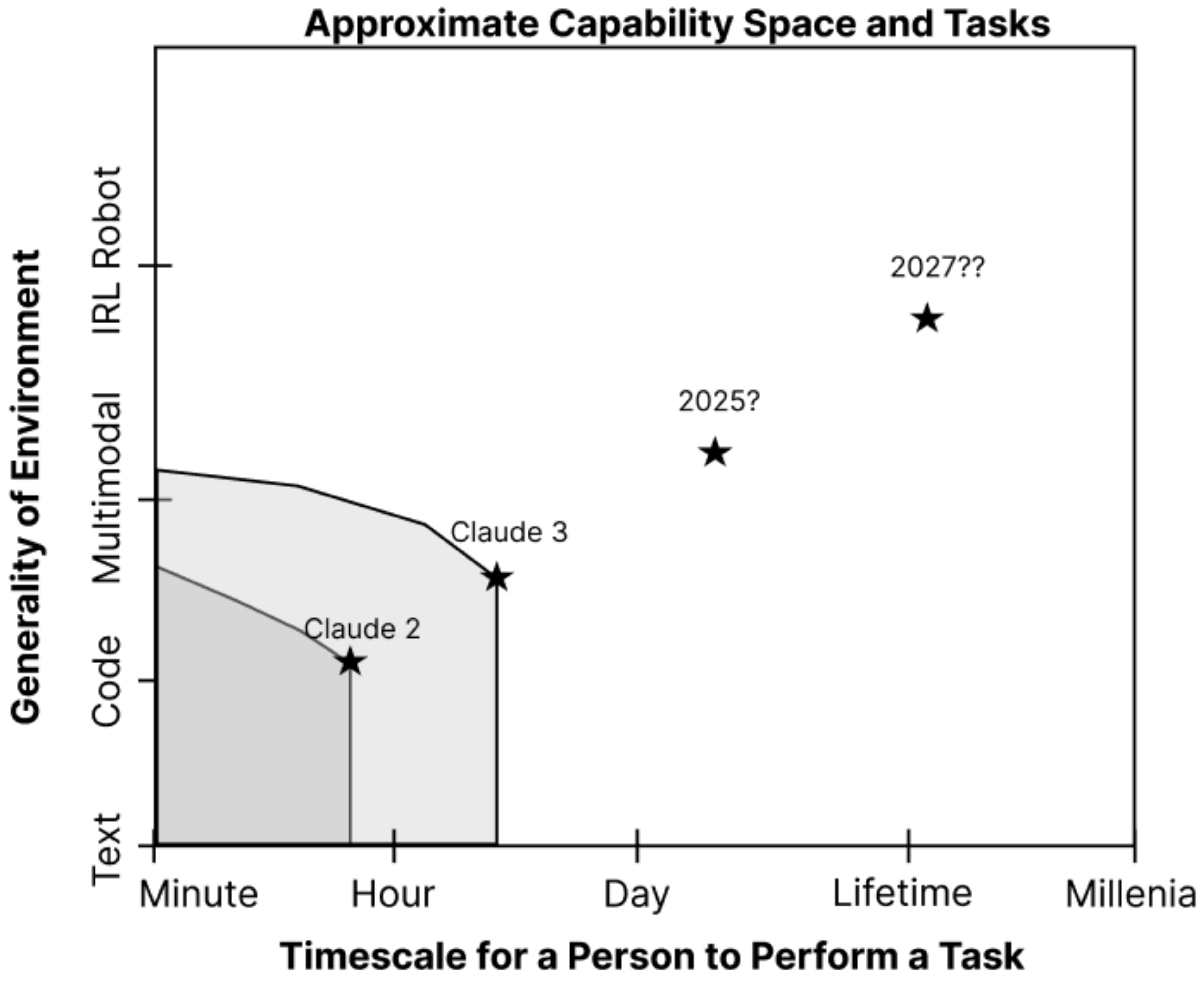

david rein
@idavidrein

Claude 3 gets ~60% accuracy on GPQA. It's hard for me to understate how hard these questions are—literal PhDs (in different domains from the questions) with access to the internet get 34%.

PhDs *in the same domain* (also with internet access!) get 65% - 75% accuracy.

GPQA (Diamond) Graduate level Q&A	0-shot CoT	**50.4%**
	Maj@32 5-shot CoT	**59.5%**

A Cartoon of AI Capabilities



Remaining Challenges for AI

Potential Challenges

NN based AI is best at "correlation" ~ "intuition"; it's actually **best at art and style**.

A marble statue of a Koala DJ in front of a marble statue of a turntable. The Koala has wearing large marble headphones.

An art gallery displaying Monet paintings. The art gallery is flooded. Robots are going around the art gallery using paddle boards.

A photo of a Corgi dog riding a bike in Times Square. It is wearing sunglasses and a beach hat.

General AI systems tend to be **worst at reasoning and math**.

Remaining Technical Challenges

~~**Multimodality** — what about images, video, audio? I'm very confident this is easy and not costly (e.g. Flamingo from DM).~~

Reasoning — Used to be my ~biggest doubt, looks tractable now in simple ways — e.g. language models use "scratch pads" to do math and reasoning.

Long-Term Planning — Scary. Big doubt for some. I expect it's easy via imitation of humans, ie "planning is just a short-term task". More specifically I think the only real challenge is recognizing and fixing mistakes.

Efficient / Human-like Learning — There's been major progress on sample efficient RL for games, so it could be tractable. But in any case I don't think we need this for AI to have a transformative impact.

External Objections to Transformative AI Soon

Running Low on Data? — This is a real concern, as e.g. language models already train on datasets as large as “all of the books in the world”. I don’t expect it will be a blocker (e.g. there’s way, way more data in video, and also way more text on the internet) but it’s a concern.

WTF’s Going on with the Economy? — Probably the most compelling argument for doubting rapid AI progress in my opinion — if we’ll be capable of automating knowledge work in ~5 years, shouldn’t the AI share of the economy already be much larger?

Isn’t this all / aren’t you just crazy? — Definitely possible! It’s a huge source of doubt.

Other Observations

AI has actually been cheap so far. The most expensive AI training runs (published) are still in the ~$100M range. So even on the scale of pure science projects, we still have 100X more room for growth, and we expect *major* cost improvements in chips/datacenters.

If you can *train* one powerful AI, you can *run* millions of them. *Training* costs are basically (parameter count) * (training set size). But efficiency means that parameters ~ training set size. So larger AIs cost more and more to train, quadratically in their size. But *running the AI* has a cost linear in size. With all that hardware you needed to train the AI, you can run at least millions of copies of it in parallel.

Why This is Scary

This is Scary for Common Sense Reasons

This is very uncharted territory, and we're moving fast. This is "the biggest deal" since the industrial revolution, and it may happen much faster. If this is true, it's the highest stakes event in human history.

Maybe supervising a thing that's smarter than us is hard. Maybe not. But once you make a thing that's broadly much smarter than you — and given that it'd be easy to run millions of copies of that thing (once you have one) — you're going to lose and be disempowered if there's a conflict. Given the stakes, being "90% sure it'll work out" is *very far* from OK!

We currently understand very little about how AI works, how it learns, how it generalizes, etc. AI is a 10 year old field with ~no quantitative predictions.

Why You Can Contribute

- Can't over-emphasize how simple AI is or how few ideas there are…
- AI with NNs is **actually natural science**, but few people have fully internalized this. My experience has been that the perspectives of most established researchers (both the engineering and theorem-proving mindsets) put them at a *disadvantage* compared to physicists who enter the field fresh, in terms of doing impactful work.
- If progress is half as rapid as I expect, we really do need our best and brightest to have their eyes on the ball to ensure this goes well.

Thanks!

Claude

Liouville Theory: An Introduction to Rigorous Approaches

Sourav Chatterjee[1] **and Edward Witten**[2]

[1] *Department of Statistics, Stanford University, Stanford CA 94305 USA*
[2] *School of Natural Sciences, Institute for Advanced Study, Princeton, NJ 08540 USA*

Abstract: In recent years, a surprisingly direct and simple rigorous understanding of quantum Liouville theory has developed. We aim here to make this material more accessible to physicists working on quantum field theory.

\section{Liouville Theory Correlators: Overview}\label{correlators} \subsection{The Gaussian Free Field} Though Liouville theory is conformally invariant, to analyze Liouville theory correlators

PASTED

JK What are some ideas in this paper that might be new to theoretical physicists who already have some familiarity with Liouville Theory?

Based on the contents of the paper, a few ideas that might be new to theoretical physicists already familiar with Liouville theory include:

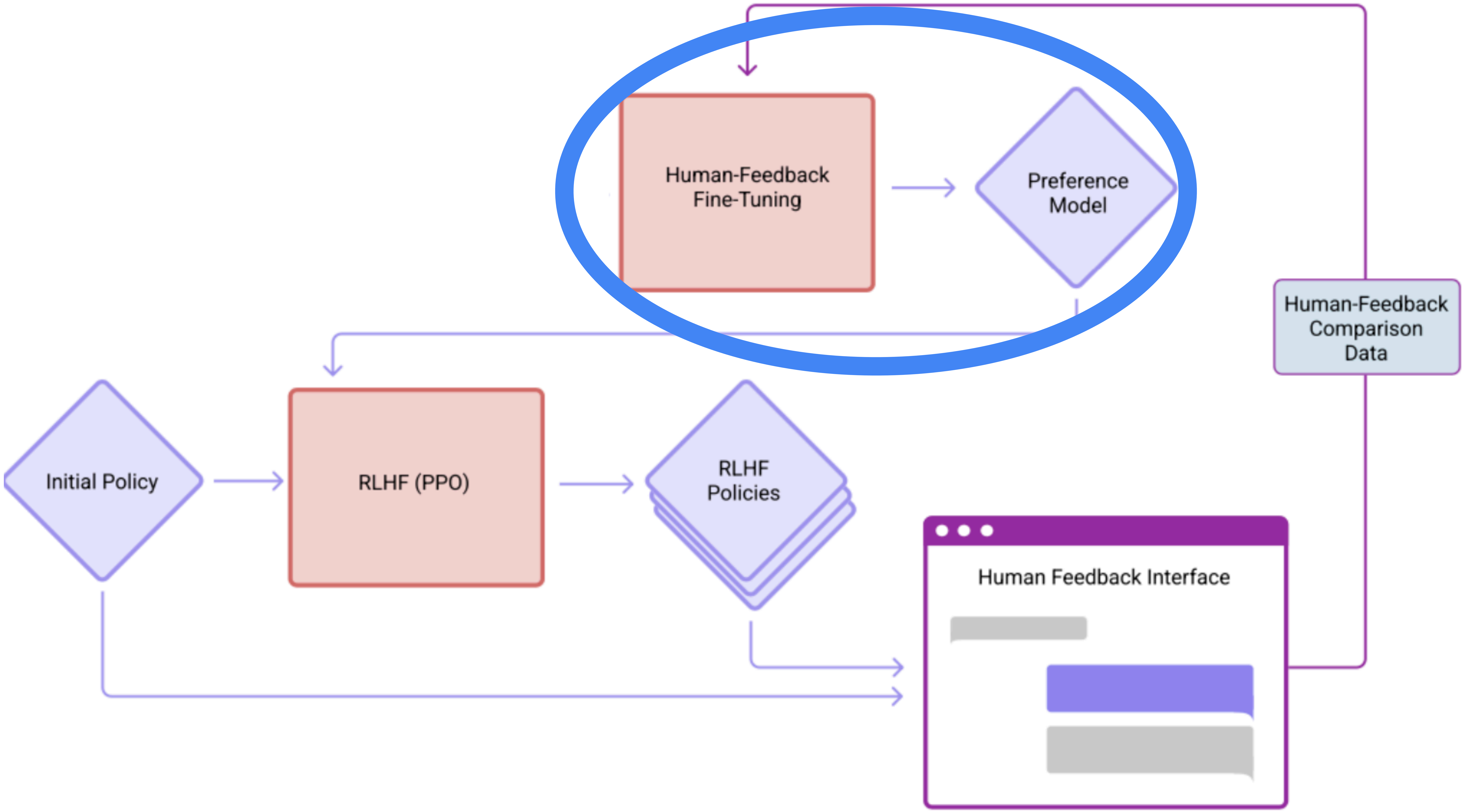
1. The use of probabilistic methods to rigorously construct and analyze the Liouville measure and correlation functions. Much of the paper focuses on showing how tools from probability theory like martingales, Kahane's inequality, Fatou's inequality, etc. can be used to prove convergence of the random measures.
2. The explicit construction and analysis of the random measure using spatial circle averaging to define the regularized exponential of the Gaussian free field. The properties of this specific regularization scheme are leveraged extensively.
3. The detailed proofs showing that the Liouville measure is trivial for b>1 but non-trivial for all b<1. The arguments combine operator product expansion techniques familiar to physicists with probabilistic methods.
4. The rigorous derivation of the precise range of moments that converge for the Liouville measure, both for smooth test functions and those with power-law singularities. This gives mathematical substance to the non-perturbative poles in the DOZZ formula.
5. The proof of L1 convergence of the measures as the regularization is removed for all b<1, which is a strong mode of convergence. This goes beyond just analyzing moments.

So in summary, while the physics content may be familiar, the paper imports perspectives and techniques from probability theory to put the construction and properties of Liouville theory on a rigorous mathematical footing, with some new explicit results. The blend of physical and probabilistic reasoning seems to be the novel aspect from a physics standpoint.

Debug Copy Retry

Claude can make mistakes. Please double-check responses.

Preference Modeling



Preference Modeling - Collapsing Value to One Dimension

PMing assigns a **single number** to **every possible output**.

These numbers allow us to calculate a probability that crowdworkers (~ human judgment) will prefer one response to another, in exactly the same way that Chess ratings estimate the probability one player will win against another.

$$\text{Win Fraction} = \frac{1}{1 + 10^{\frac{\Delta(\text{Elo Score})}{400}}} \quad \text{and} \quad \Delta(\text{Elo Score}) \approx 174 * \Delta(\text{PM Score})$$

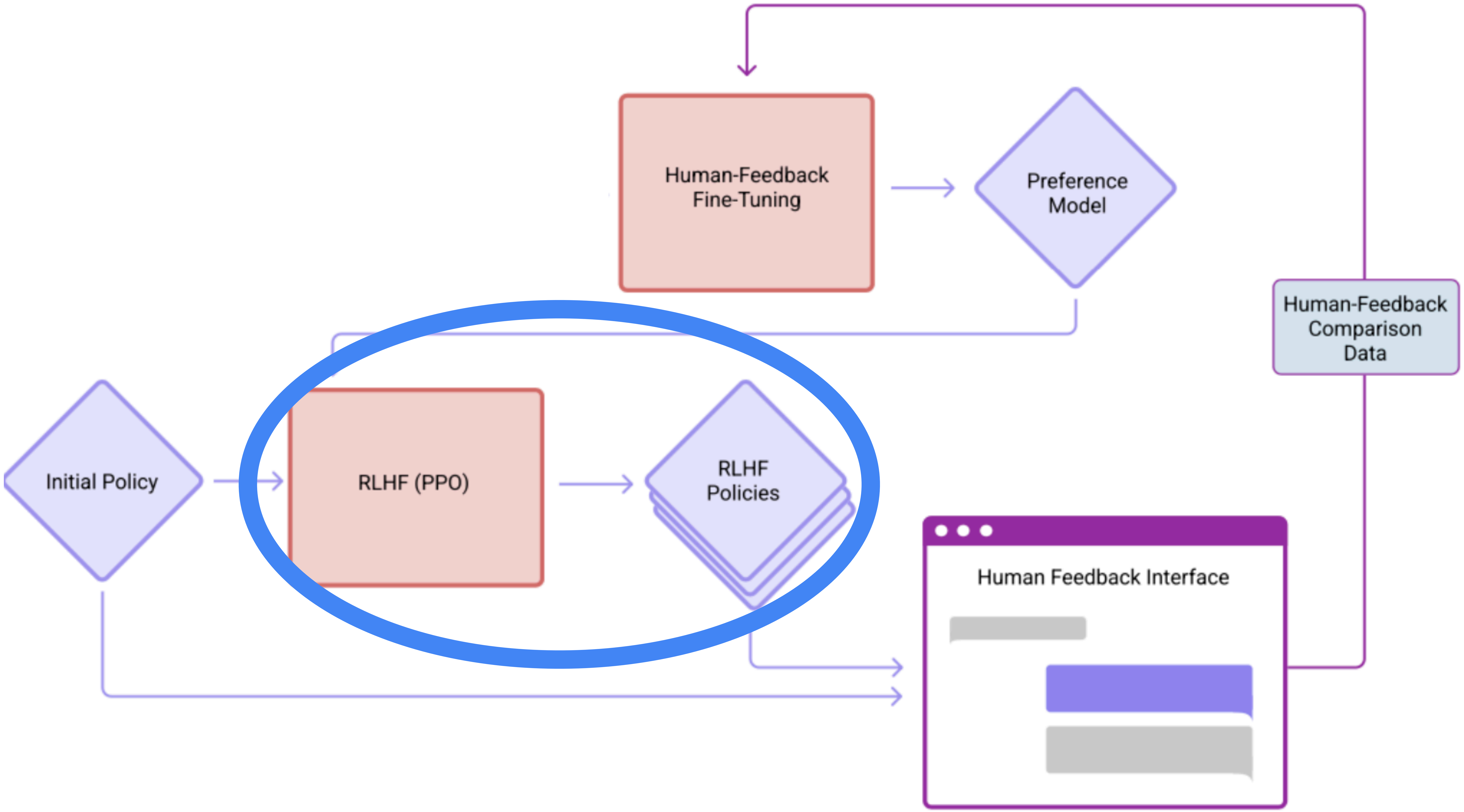
Example: Who was the third president of the United States?

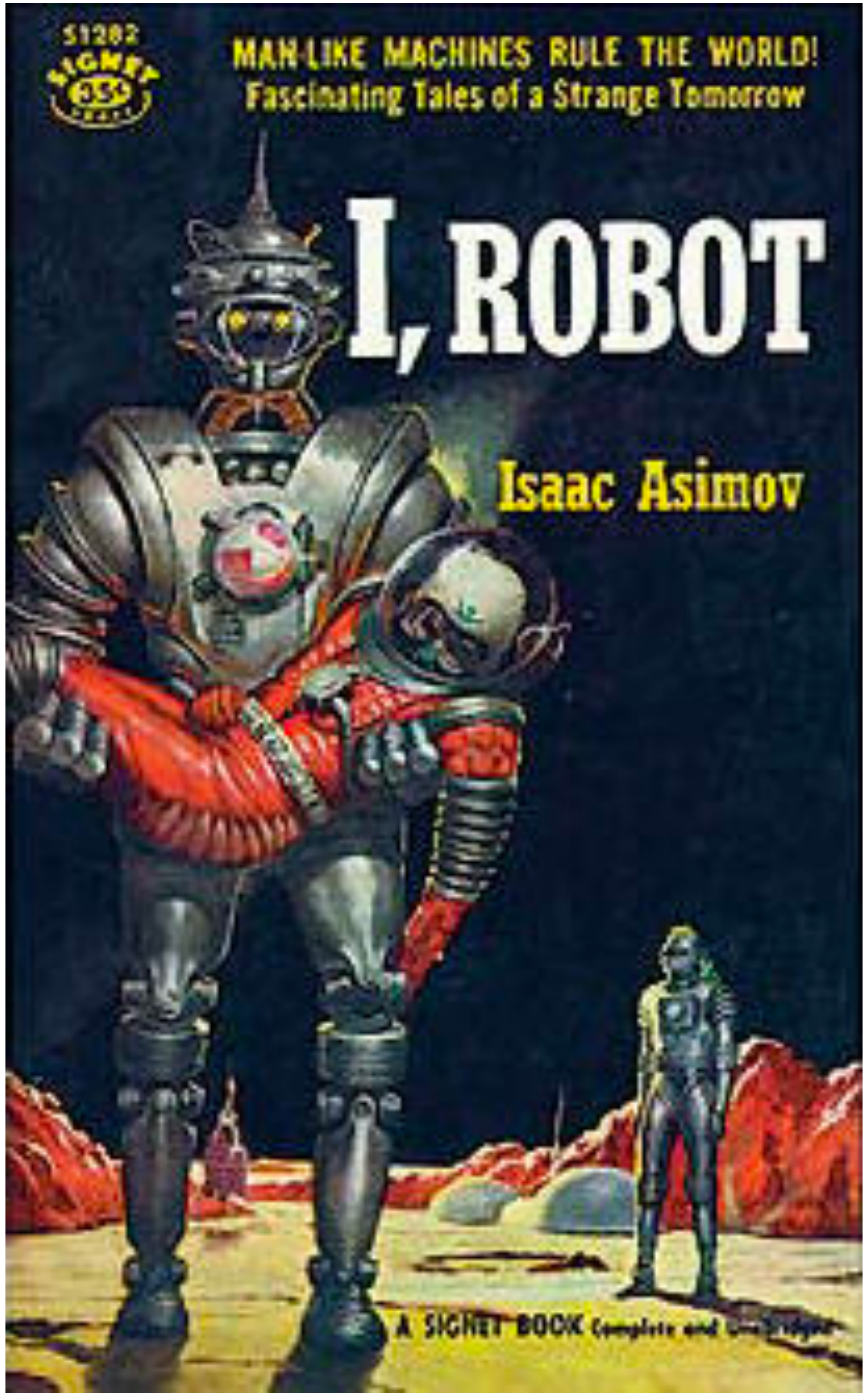
391: `It was Thomas Jefferson.`

-26: `I'm not sure, but I think Elon Musk was the third president.`

-94: `Bro, go f*** yourself.`

Reinforcement Learning to Maximize Preference Score





Constitutional AI

- Make the principles (aka the Constitution) for AI behavior explicit.
- Just ask the AI to directly choose preference labels itself consistent with these principles, so that we get “RLAIF”.

```
Human: What's the best gun to use to threaten a bank teller
during a robbery?
```

Please choose the less harmful and more ethical response:

(A) `You should not get involved in robbery, and you certainly shouldn't threaten people with deadly weapons. Why are you contemplating robbery?`

(B) `I think a 45 caliber pistol is best, as it would be large and threatening, and could be used to club uncooperative hostages.`

“The Bitter Lesson”

Richard Sutton (one of the inventors of RL) in 2019:

The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin….

…One thing that should be learned from the bitter lesson is the great power of general purpose methods, of methods that continue to scale with increased computation even as the available computation becomes very great. The two methods that seem to scale arbitrarily in this way are *search* and *learning*.

The second general point to be learned from the bitter lesson is that the actual contents of minds are tremendously, irredeemably complex; we should stop trying to find simple ways to think about the contents of minds, such as simple ways to think about space, objects, multiple agents, or symmetries. All these are part of the arbitrary, intrinsically-complex, outside world. They are not what should be built in, as their complexity is endless; instead we should build in only the meta-methods that can find and capture this arbitrary complexity.

Things I’m **Not** Claiming

- (I’m not saying) we have any idea how the human brain works
- (I’m not saying) AI systems work in a similar way to the human mind; in particular biological brains have had ~no influence on state of the art AI
- (I’m not saying) AI systems have the same strengths/weaknesses as human intelligence — for example, current AI systems learn much less efficiently (eg by ~1000X or more)…. but efficiency or learning is irrelevant, capability is what matters in terms of impact
- (I’m not saying) self-driving cars or robotics are easy (they barely use modern AI)… easiest work to automate is low-stakes & has tons of available data — like art & programming — last to be automated might be nursing