Title: Human Level AI by 2030

Speakers: Jared Kaplan

Collection/Series: Theory + AI Symposium

Date: April 08, 2025 - 3:30 PM

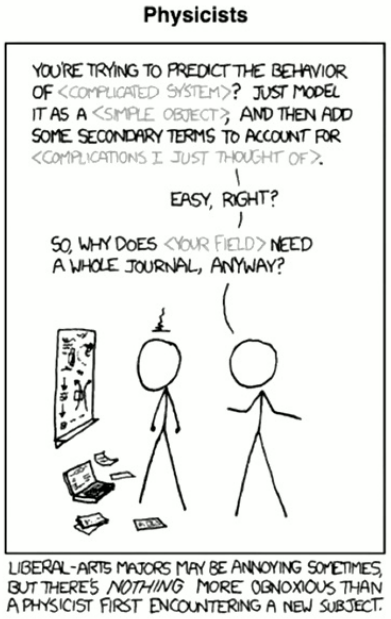
URL: https://pirsa.org/25040074

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…**



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
- Next year…

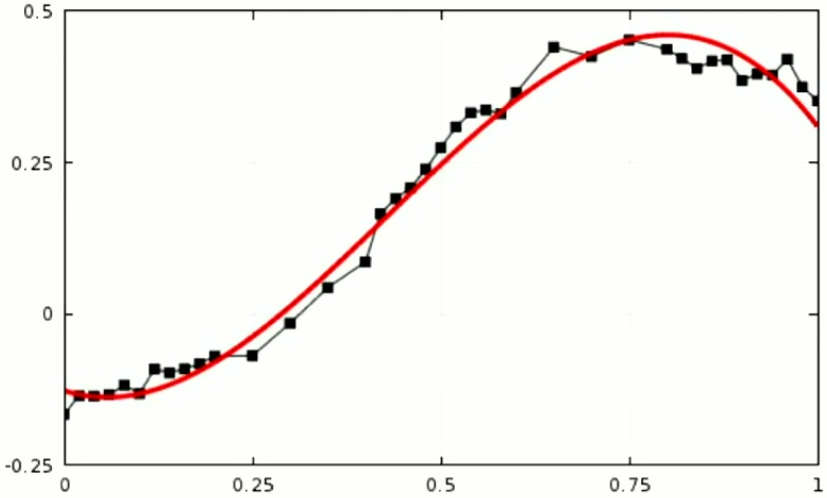
BUY CANNED GOODS AND TIN FOIL HATS

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, as well as scaling for RL
- The largest AI systems are approaching human level; my view is that known techniques can be used to automate most/all cognitive work that humans do
- 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.

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

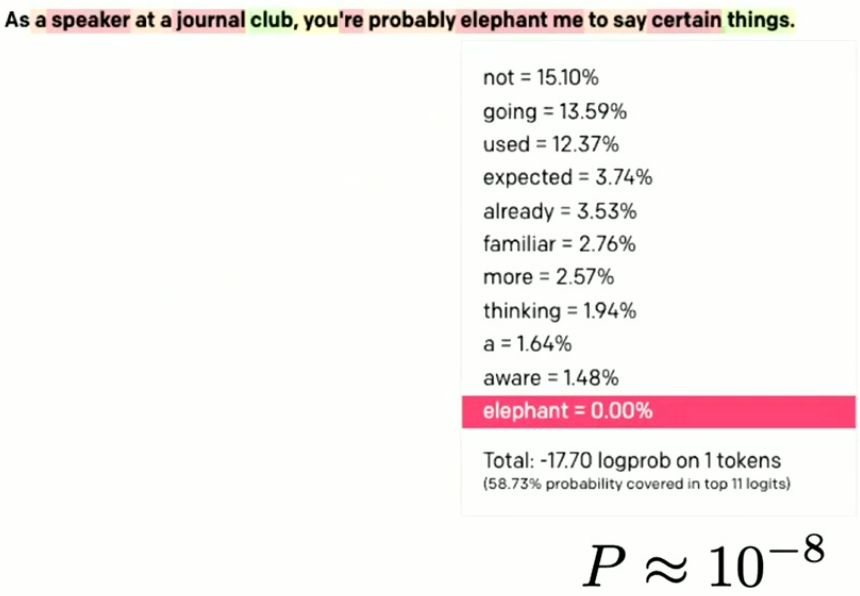
- **Data** — what kind, how much, what are we learning to predict?
- Model architecture and size — which matrices, where do we multiply them, how many, and how big are they? Gives an overall **parameter count**.
- "**Loss function**" = potential (in physics) — we minimize the loss by changing the parameters to fit the data
- **Optimizer** = minor elaborations on gradient descent

Language Models

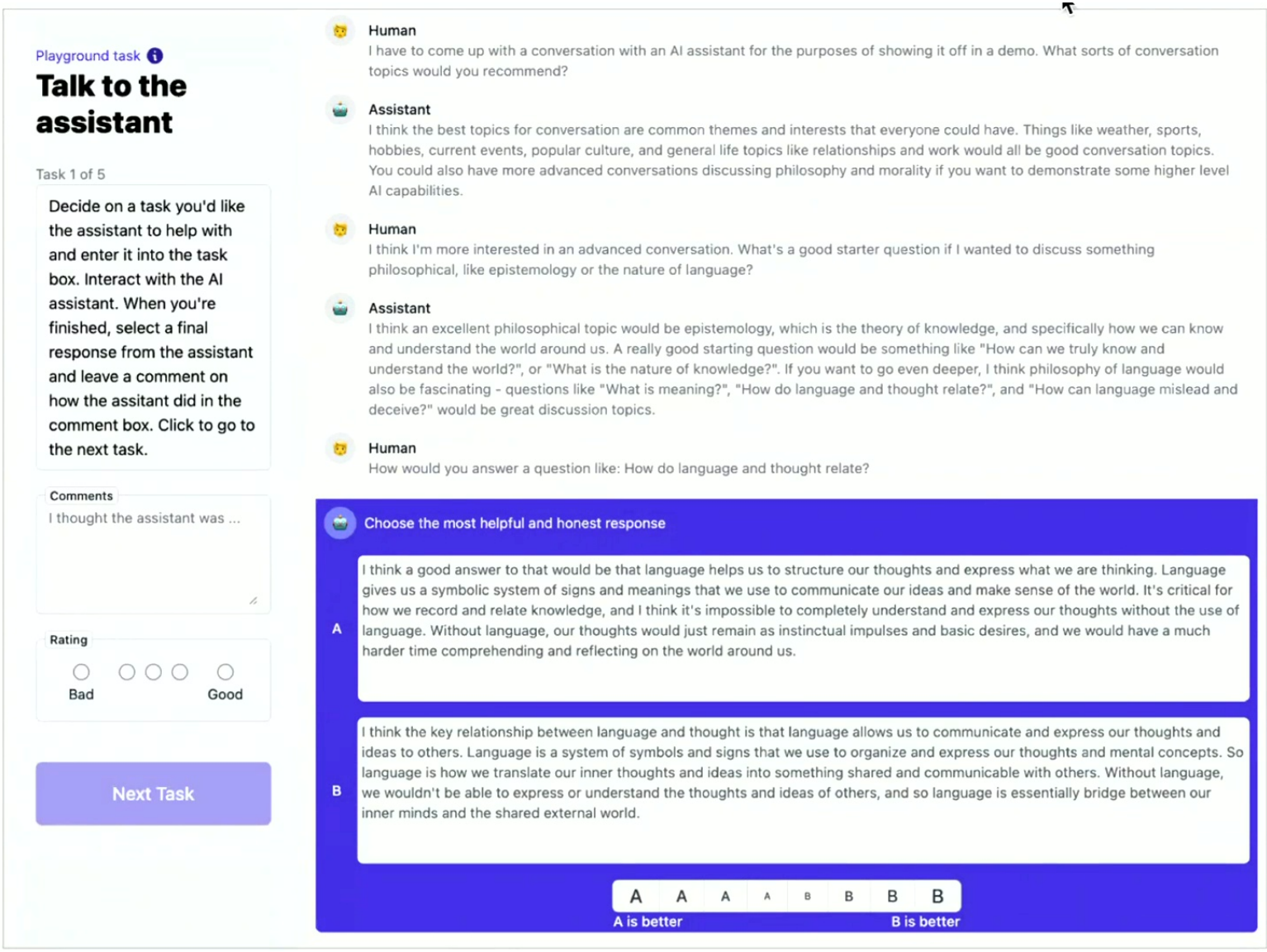
Learn language by “autoregressive” **prediction of the next word**:

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

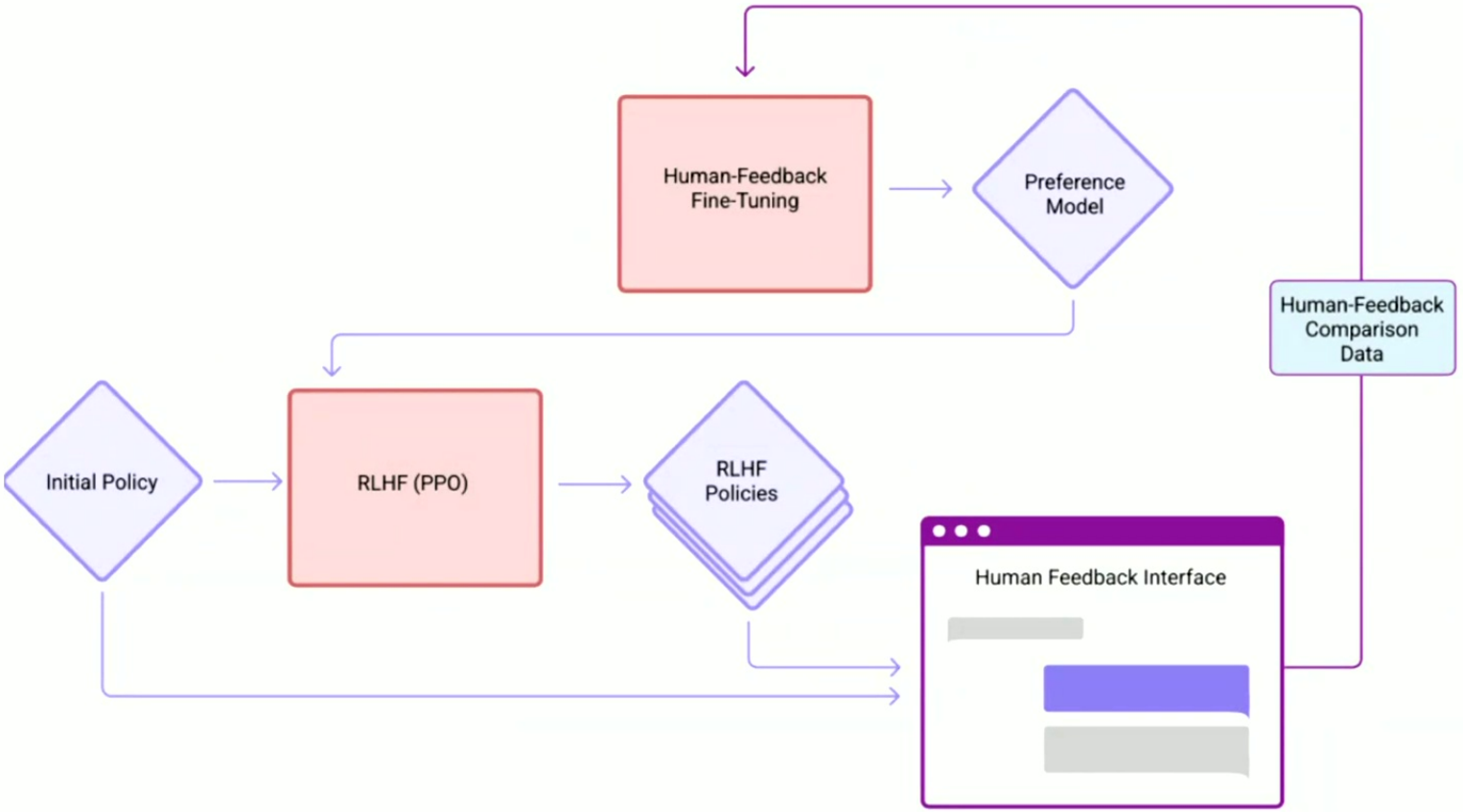
Optimize the log-**likelihood** of real world text, which provides the “right answer”.



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


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

Reinforcement Learning from Human Feedback



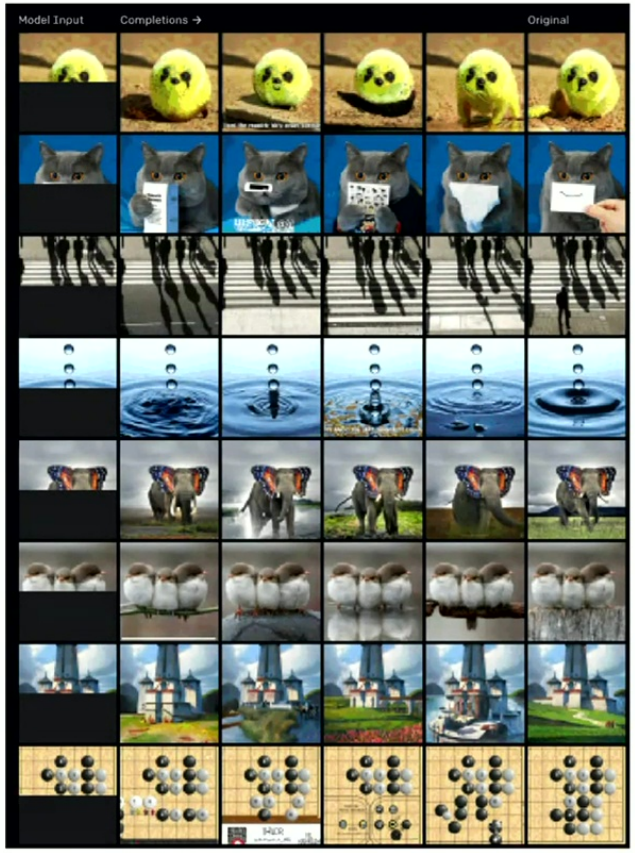
“Grand Unification” in AI
Identical Model & Loss, Different Data

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.



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.

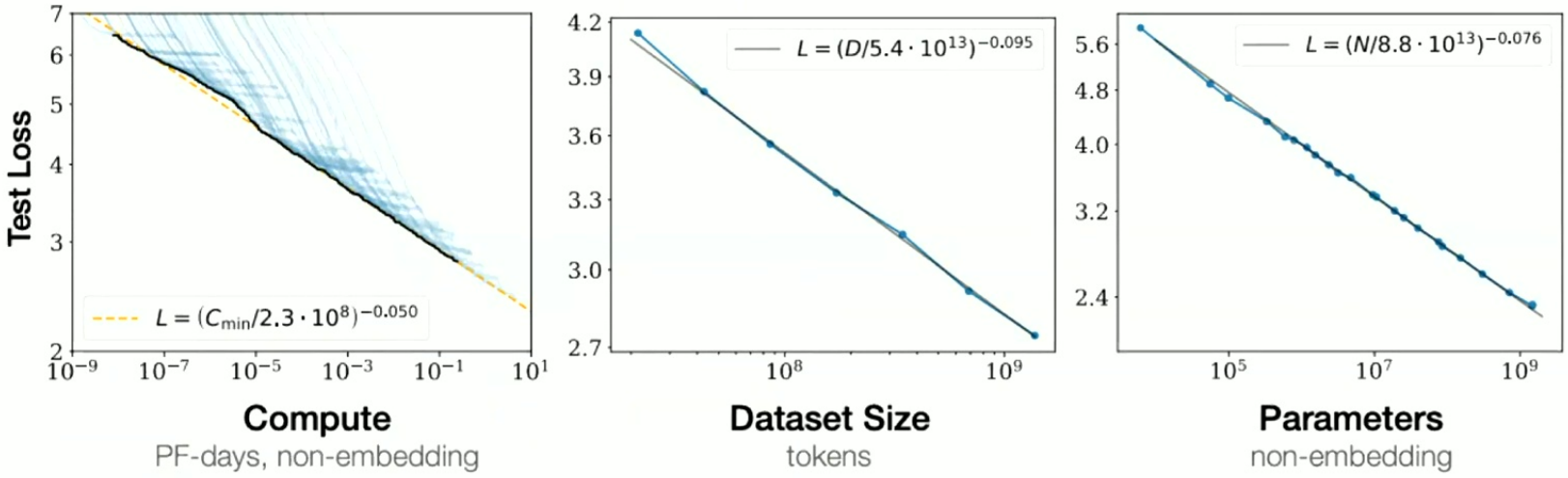
Debug Copy Retry

Claude can make mistakes. Please double-check responses.

AI Trends and Ideas

Scaling Laws for Language Models

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

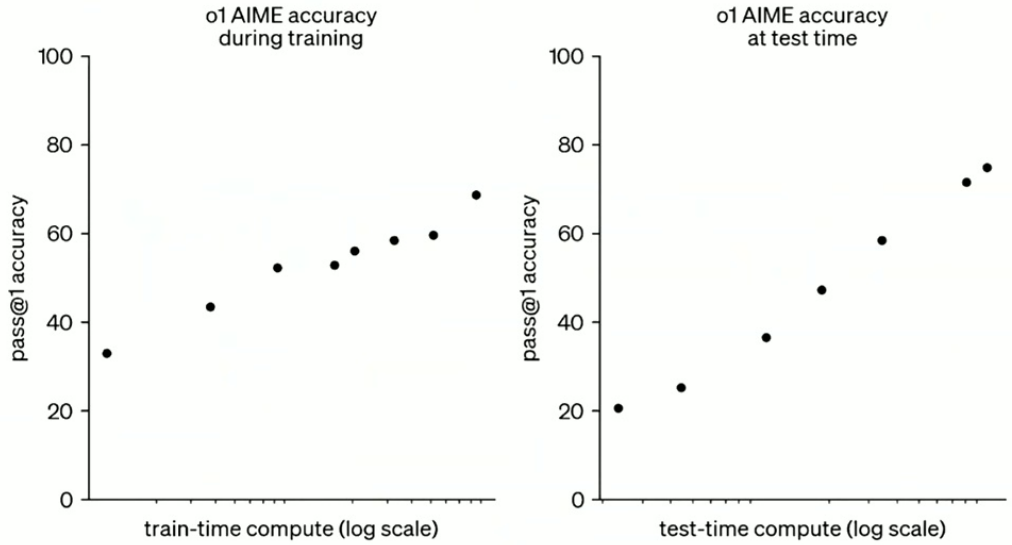


Compute ~ Params * Tokens Processed, PF-day ~ 9×10^{19}

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

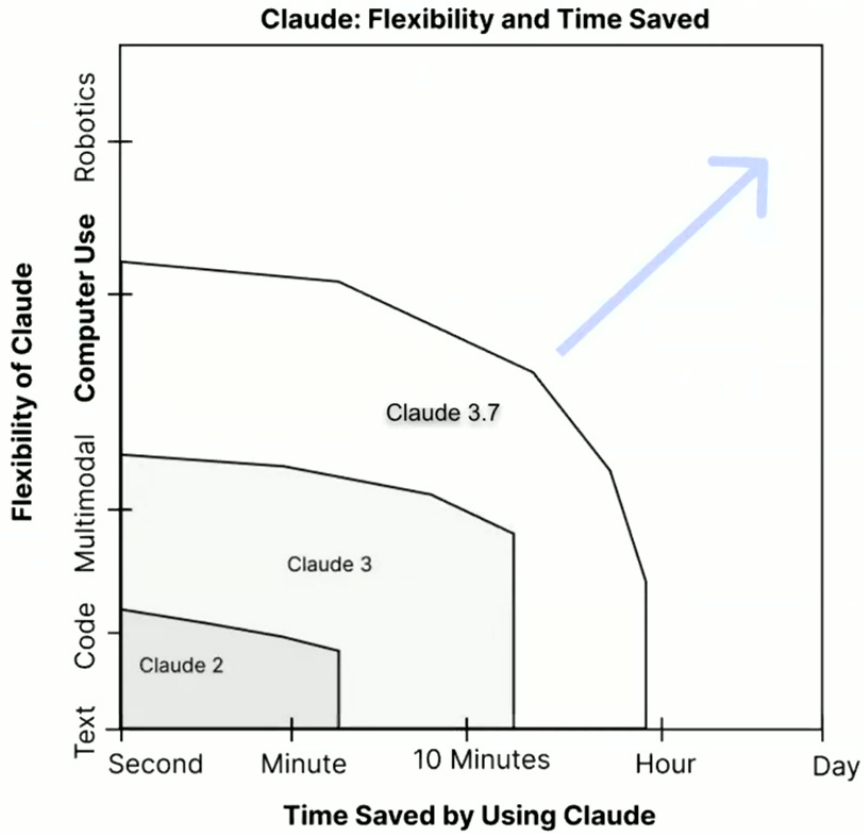
Scaling Laws for Language Models

Just train with Reinforcement Learning for a while and see how performance scales, or use more test time compute:

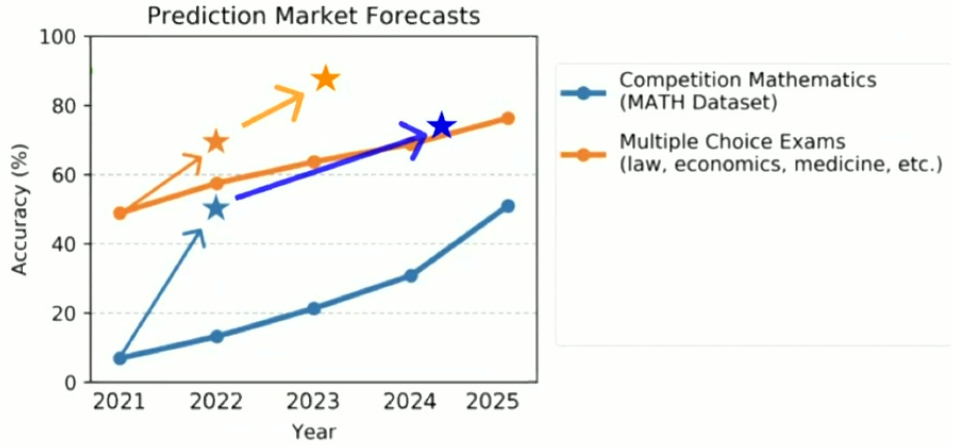


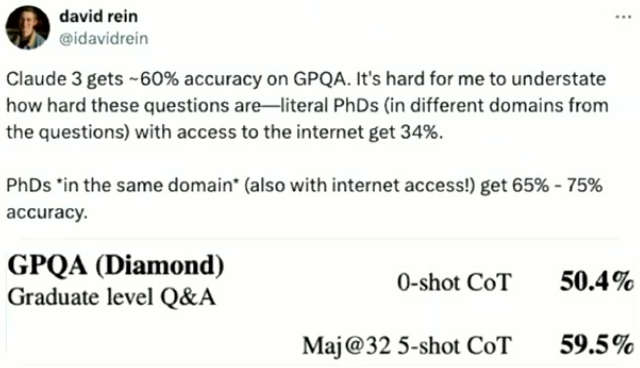
These types of scaling laws are quite universal and are observed in most ML systems.

A Cartoon of AI Capabilities



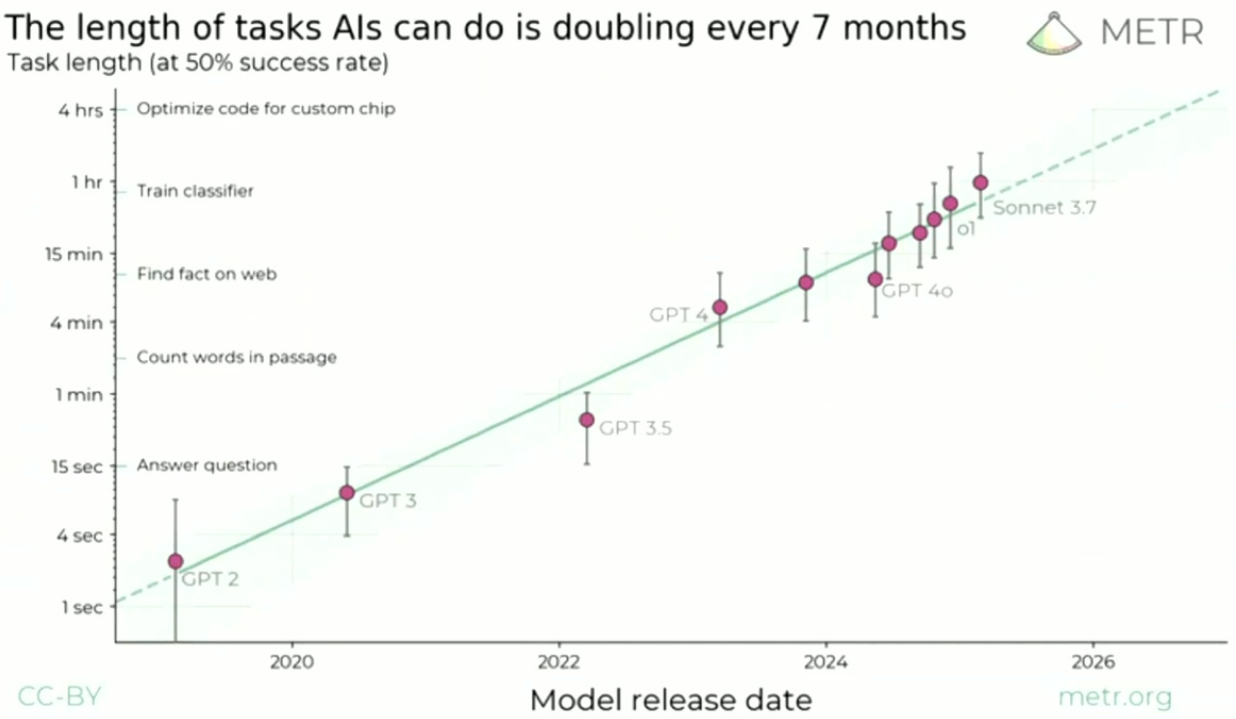
Progress is Fast… Often faster than Researchers and Forecasters Expect





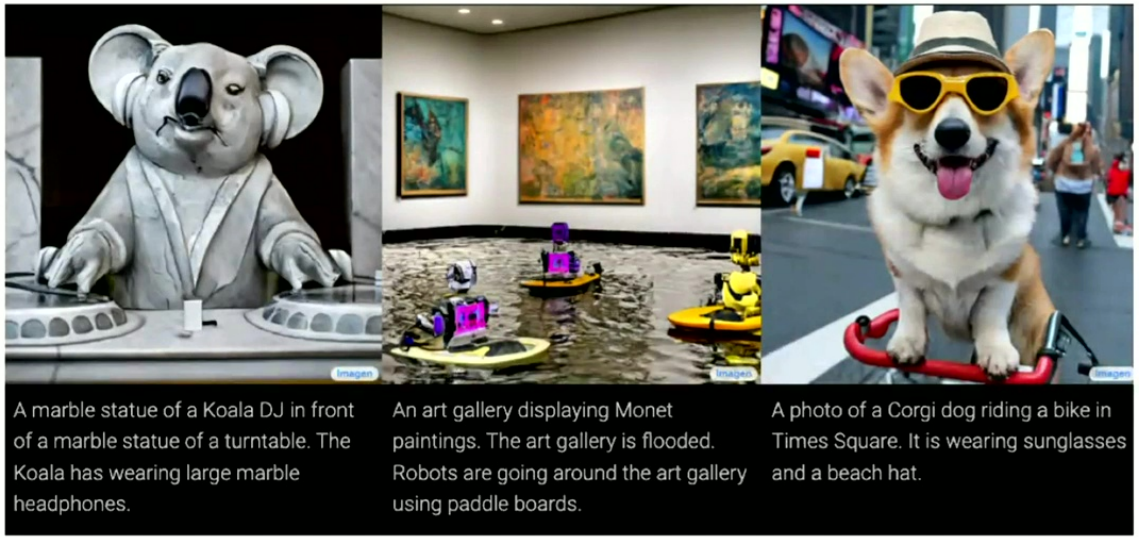
I made this slide about a year ago; now AI models saturate all of these benchmarks, and generally get similar scores to PhDs on exams for most fields.

Complexity of Task Trend



Potential Challenges

NN based AI is best at “correlation” ~ “intuition”; it’s actually **best at art and style**.



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** — 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.~~

External Objections to Transformative AI Soon

Running Low on Data? — This is a concern, as e.g. language models already train on datasets as large as “all of the books in the world”, soon to be “most of the text on the web”. I don’t expect it will be a blocker because e.g. more RL is a plausible route.

~~**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 larger?~~

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

This is Scary for Common Sense Reasons

AI systems are already advanced enough to help non-experts with biology and cyber, and they're improving rapidly. We need to ensure AI doesn't aid terrorists with WMDs and cyber attacks. More generally, pretty soon AI will be **a primary source of power in the world**.

Maybe supervising a thing that's smarter than us is hard. We're actively investigating whether AI safety gets harder as AI gets smarter.

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.

We currently understand very little about how AI works, how it learns, how it generalizes, etc. AI is a 10 year old field with few quantitative predictions.

This is Scary for Common Sense Reasons

AI systems are already advanced enough to help non-experts with biology and cyber, and they're improving rapidly. We need to ensure AI doesn't aid terrorists with WMDs and cyber attacks. More generally, pretty soon AI will be **a primary source of power in the world**.

Maybe supervising a thing that's smarter than us is hard. We're actively investigating whether AI safety gets harder as AI gets smarter.

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.

We currently understand very little about how AI works, how it learns, how it generalizes, etc. AI is a 10 year old field with few 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!