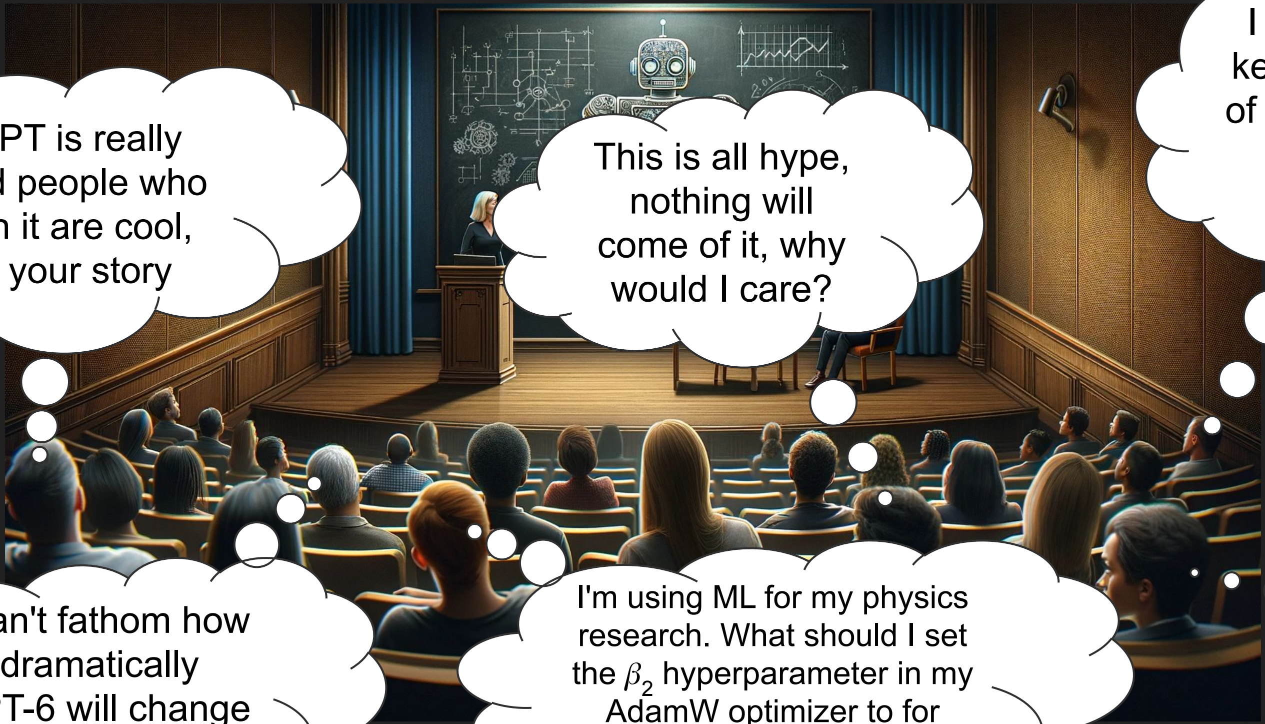
AI in 2023

Where are we and where are going?

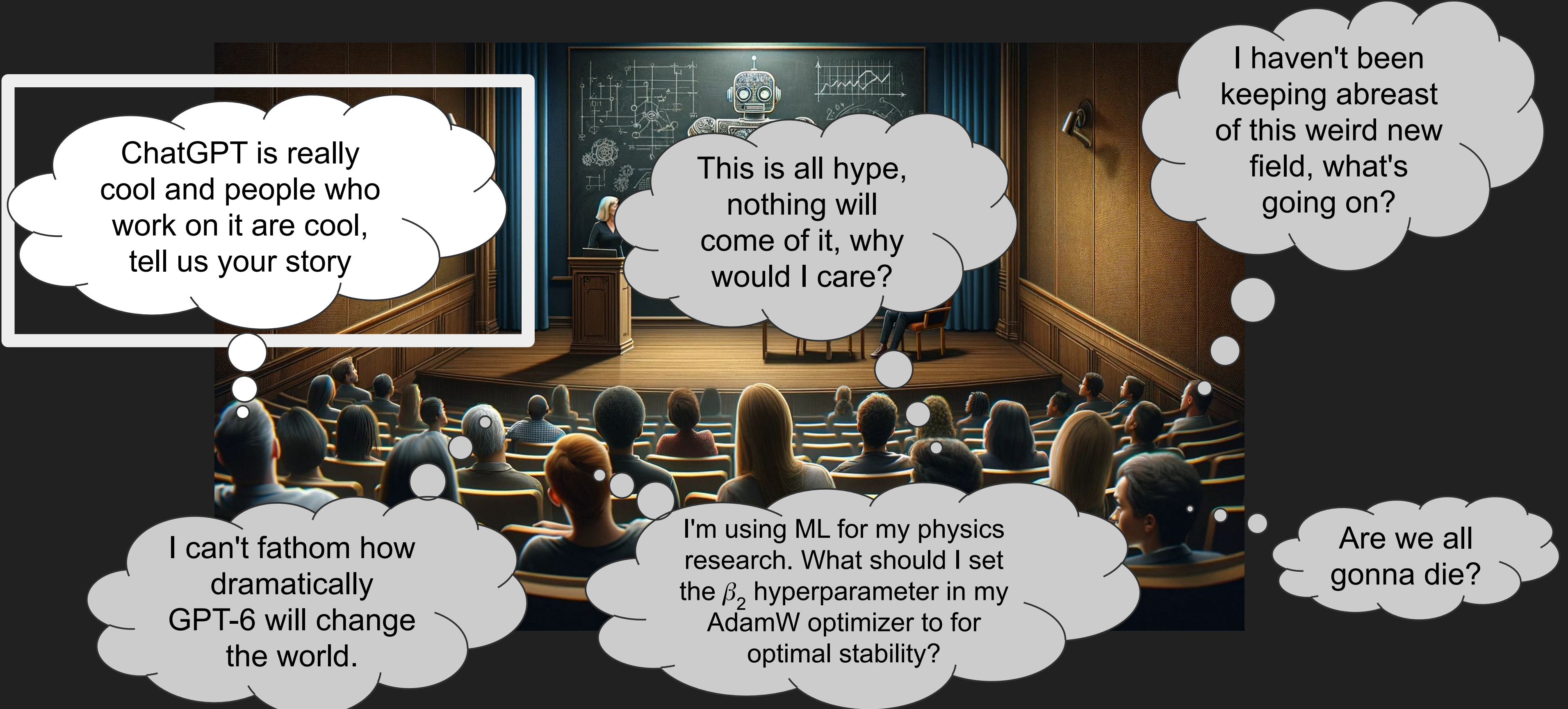
David Farhi
OpenAI

Rutgers University
11/15/23

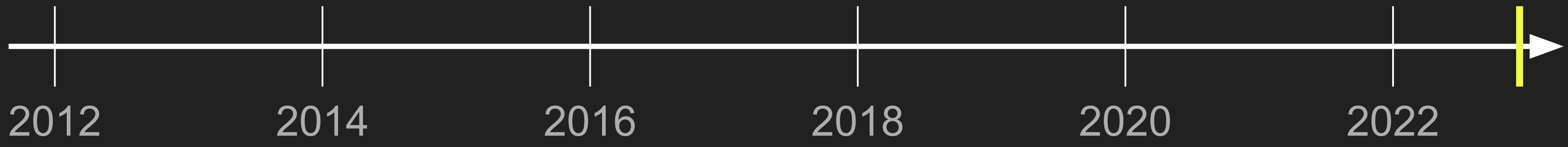
The challenge for talks about Artificial Intelligence



The challenge for talks about Artificial Intelligence

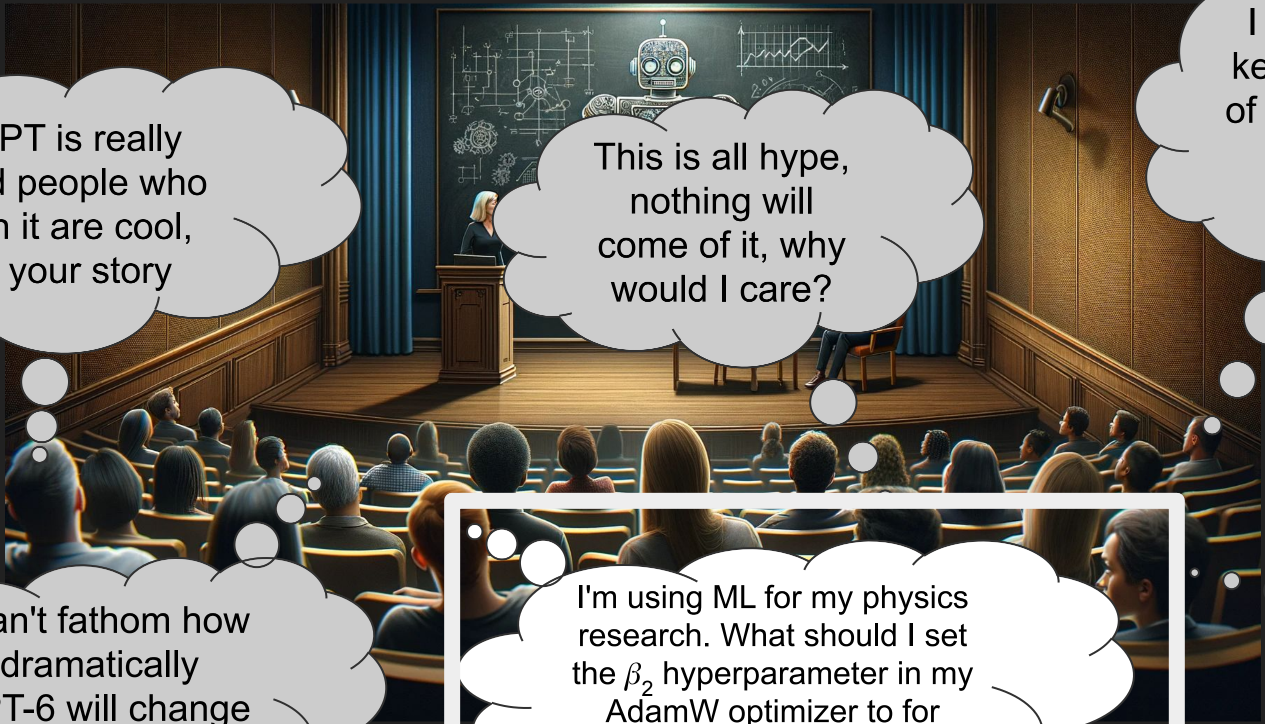


My trajectory so far



	Phd in High Energy Theory	Software Engineering on Google Search	AI Research at OpenAI (OpenAI Five, GPT-4, ...)
Solve deep, interesting problems	✅	❌	✅
Fast paced, world-changing impact	❌	✅	✅

The challenge for talks about Artificial Intelligence



ChatGPT is really cool and people who work on it are cool, tell us your story

This is all hype, nothing will come of it, why would I care?

I haven't been keeping abreast of this weird new field, what's going on?

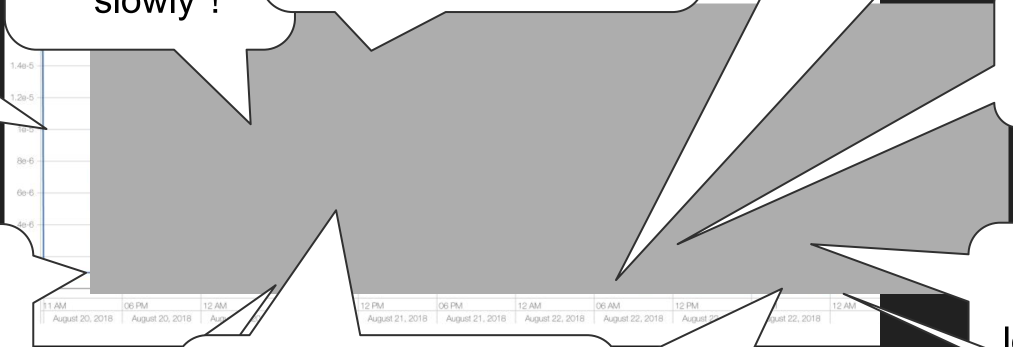
I can't fathom how dramatically GPT-6 will change the world.

I'm using ML for my physics research. What should I set the β_2 hyperparameter in my AdamW optimizer to for optimal stability?

Are we all gonna die?

You are not alone

Lea



Perhaps LR is too high for it to learn fine-grained features...

Wait I know! Let's lower it *slowly*!

No, no, guys, we have to try lowering it *even more slowly*

Hmm, that didn't help, let's put it back

Oops I made a bug, now it's literally zero.

Let's just go back to before we started screwing with it.

Hey I just checked the code, somehow it's zero again.

Hmm, in one of the thousand random self-play games in the last minute, I saw it make a dumb mistake. Clearly this LR is too low!

Oh it just made the same mistake again, I guess it was fine. Undo!

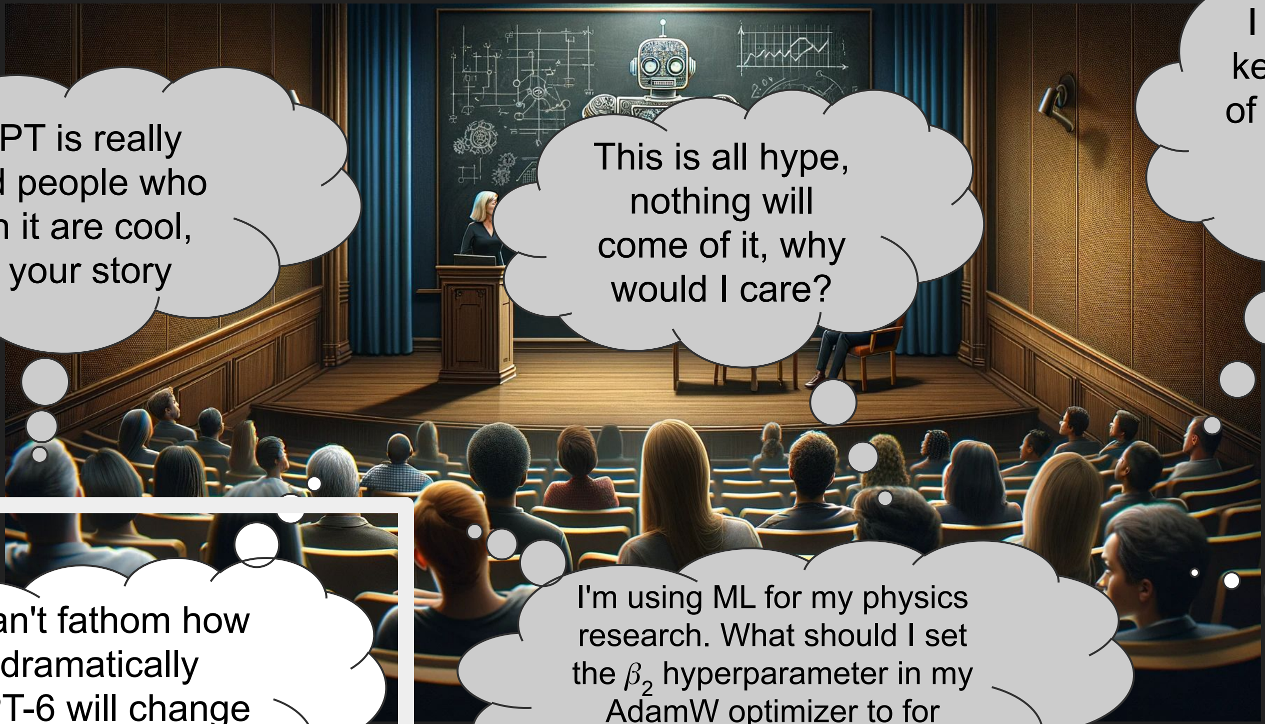
One more try at lowering it slowly?

Stop making it zero!

OpenAI Five (OpenAI, 2019; https://arxiv.org/abs/1912.06680)

Quotes are made up because we didn't record the originals. But I was there, it was *basically* like this.

The challenge for talks about Artificial Intelligence



ChatGPT is really cool and people who work on it are cool, tell us your story

This is all hype, nothing will come of it, why would I care?

I haven't been keeping abreast of this weird new field, what's going on?

I can't fathom how dramatically GPT-6 will change the world.

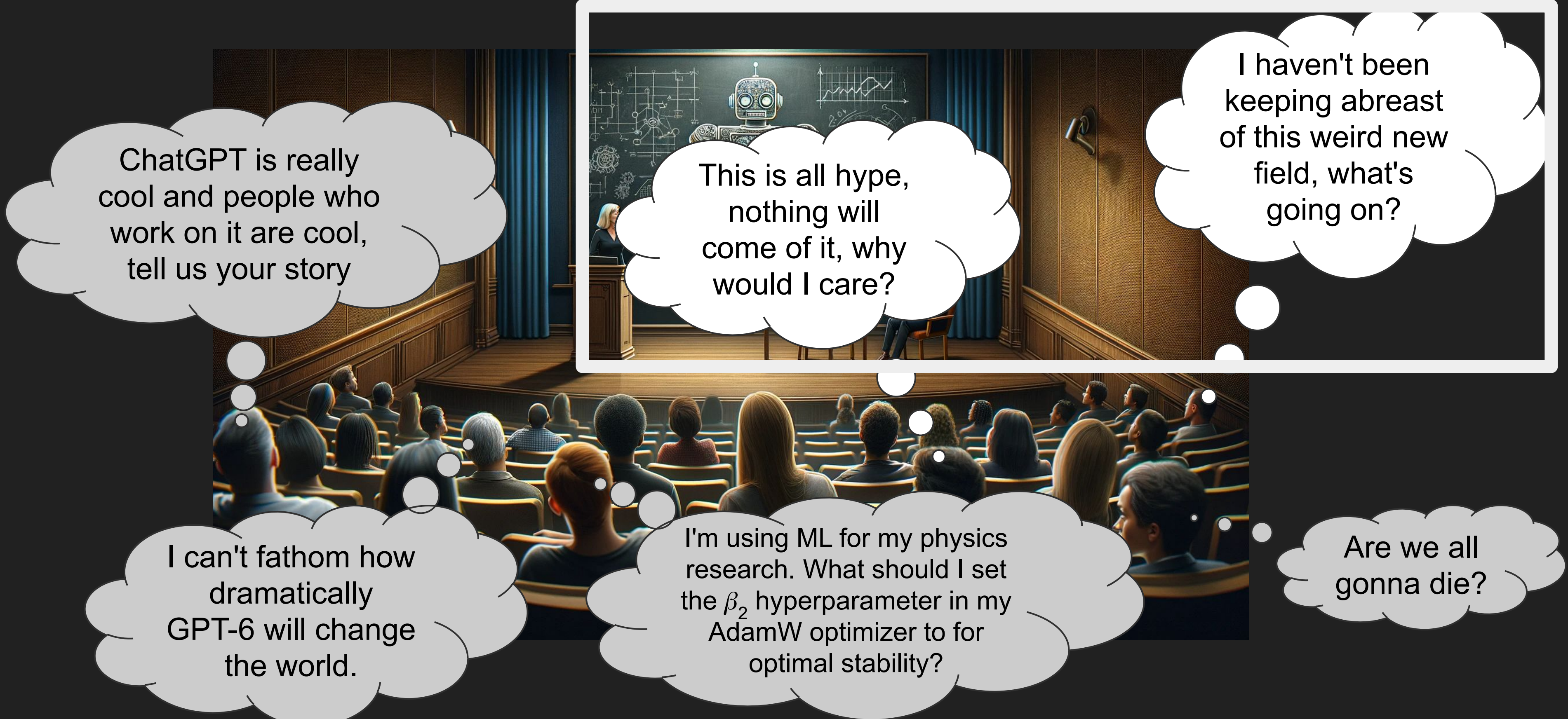
I'm using ML for my physics research. What should I set the β_2 hyperparameter in my AdamW optimizer to for optimal stability?

Are we all gonna die?

Three True, Trivial Statements

1. The future could be way better, or way worse, than the present.
2. We have control over the future.
3. So let's make it good.

The challenge for talks about Artificial Intelligence

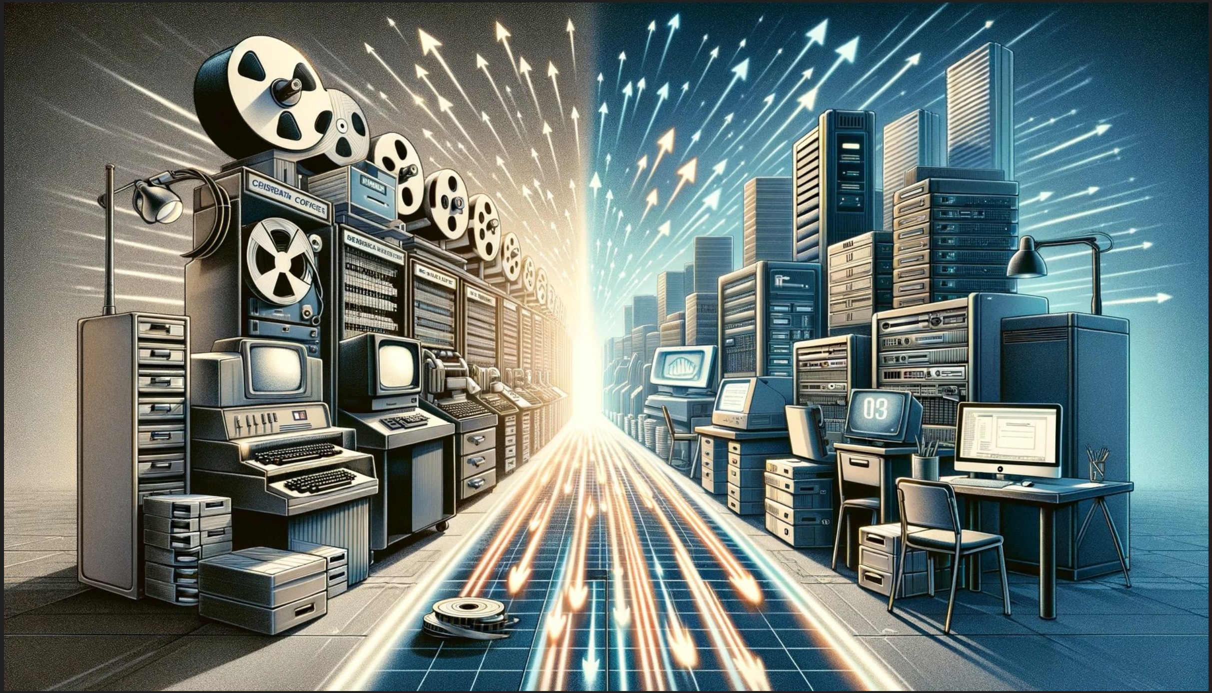


Outline

1. ~~The six different talks~~
2. Pace of progress in compute
3. GPT-4
4. GPT-4 Demos
5. From Alchemy to Science

Bonus slides

- Speculation on the future
- Spherical Cow of how Machine Learning works
- The alignment problem
- The OpenAI Five (Dota 2) project



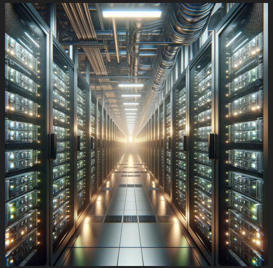
Part 2: The Pace of Progress in Compute

AI: What have we been doing recently?

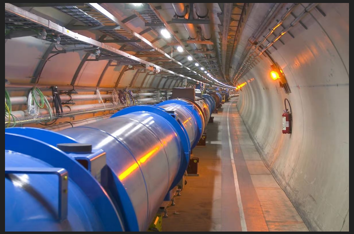
Building enormous machines

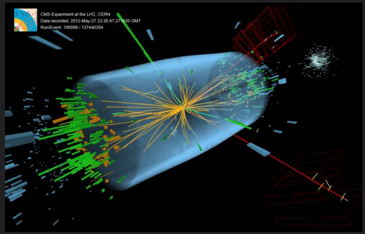
That cost billions of dollars

To explore the unknown

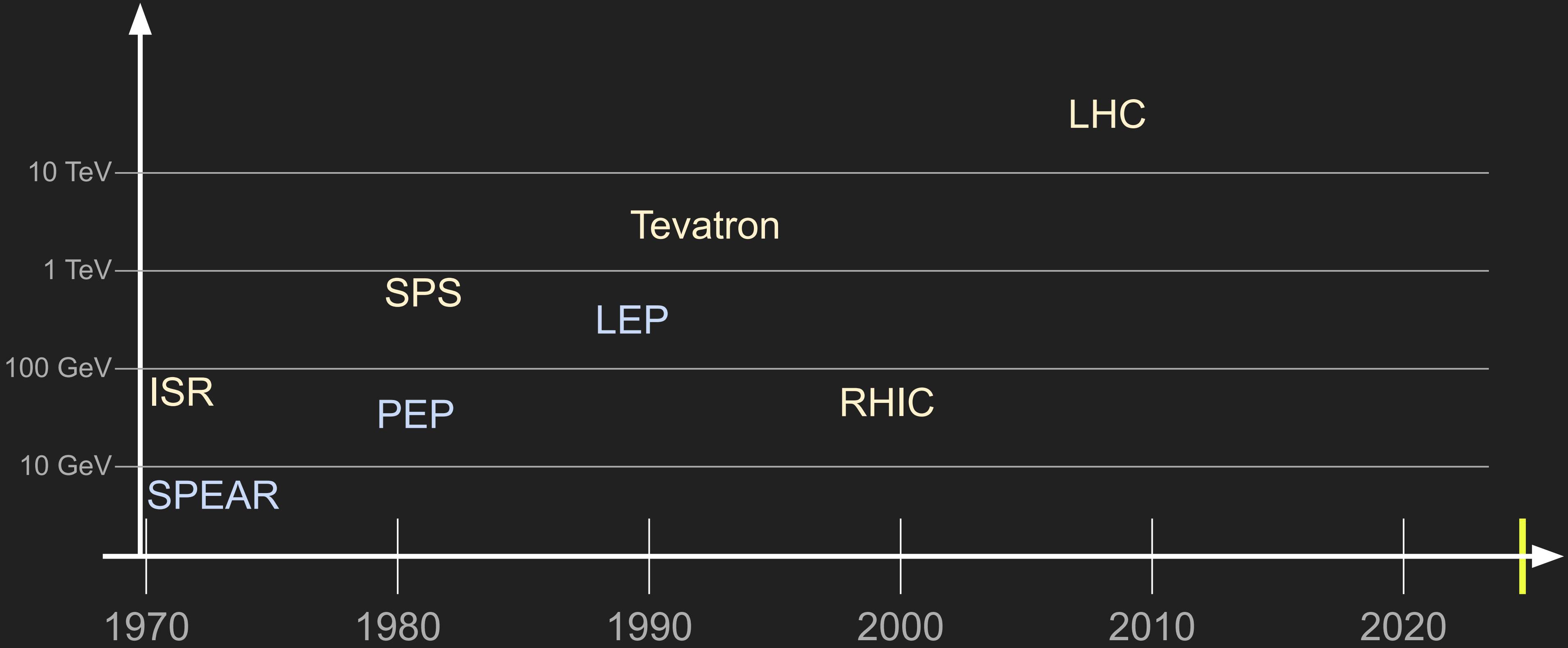


... as we have for centuries



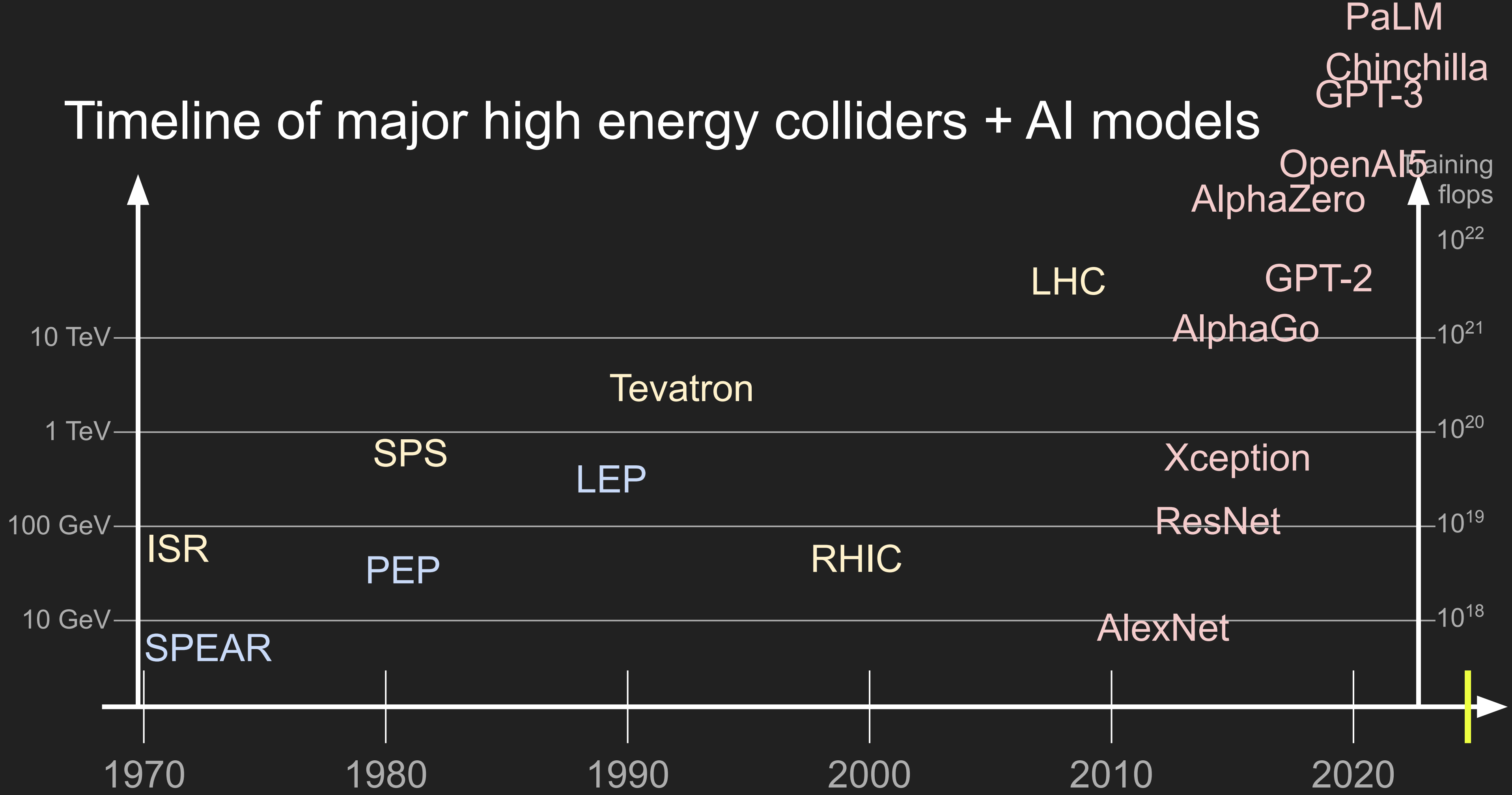


Timeline of major high energy colliders



Data from wikipedia, plotted by hand

Timeline of major high energy colliders + AI models

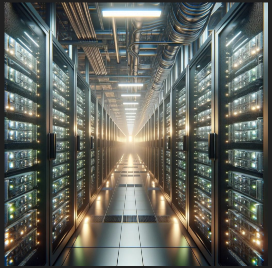


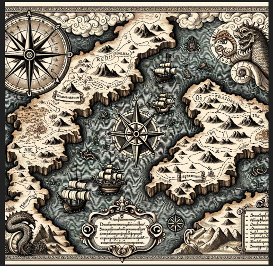
https://arxiv.org/pdf/2202.05924.pdf,
https://medium.com/aiguys/chinchilla-training-compute-optimal-large-language-models-a922a0d9eebb,

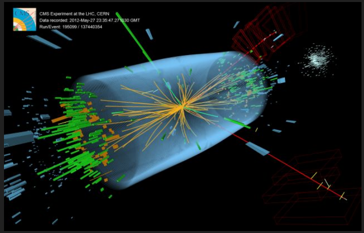
Engineering and Exploration

Key Question:

- Is the engineering working?
- It's easy to build useless tools
- Are they producing real progress







Intelligence ~ Compute ???

Building a big collider is not easy...

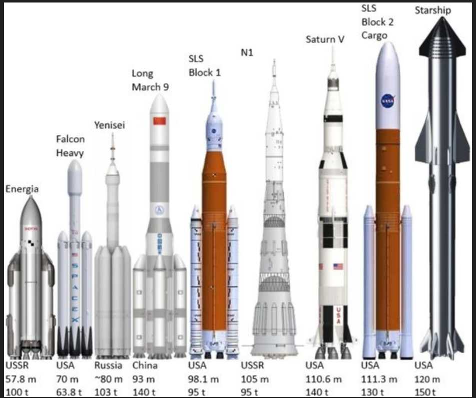
But it's also not sufficient.

Need to verify we get scientific results from experiments.

Training a model on 10^{23} flops is not easy...

But it's also not sufficient.

Need to verify we get more intelligent models.



Rest of this talk = Two ways to convince ourselves:

- Part 4: GPT-4 Demos
- Part 5: Scaling Laws

Part 3: GPT-4

Generative Pretrained Transformer (GPT)

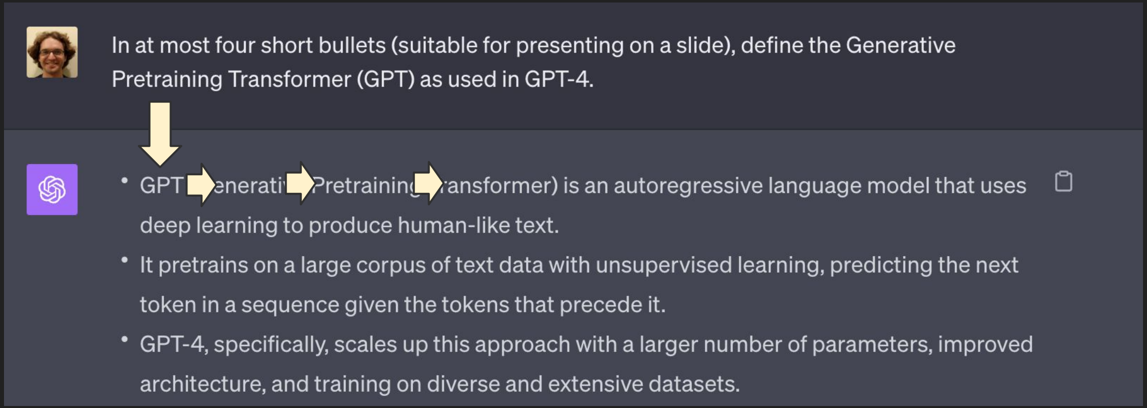
In at most four short bullets (suitable for presenting on a slide), define the Generative Pretraining Transformer (GPT) as used in GPT-4.

- GPT (Generative Pretraining Transformer) is an autoregressive language model that uses deep learning to produce human-like text.
- It pretrains on a large corpus of text data with unsupervised learning, predicting the next token in a sequence given the tokens that precede it.
- GPT-4, specifically, scales up this approach with a larger number of parameters, improved architecture, and training on diverse and extensive datasets.

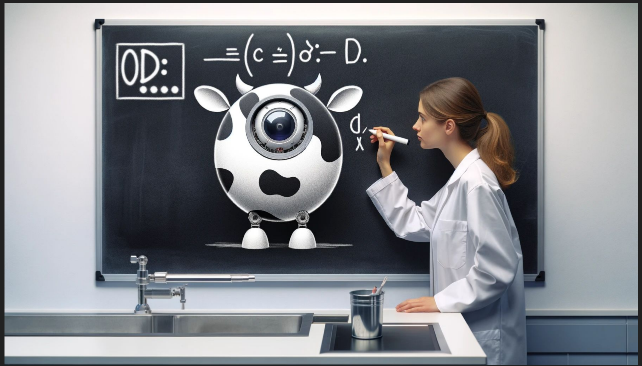
GPT-4's goal: Model the full distribution of human text

Autoregressive language model:

f: text so far → probability distribution of next word



Deep Learning (Spherical Cow View)



Prob dist of next word, *p(w)*

GPT-4

giant blob of parameters θ

$f(\theta; T)$

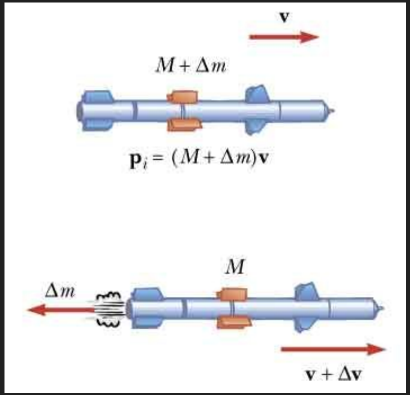
Previous text *T*

Gradient Descent

1. Initialize randomly
2. Sample some data from the internet
3. Measure L = distance btwn GPT-4 output distribution and data distribution
4. Calculate gradient g=dL/dθ
5. Update θ proportional to g
6. Repeat

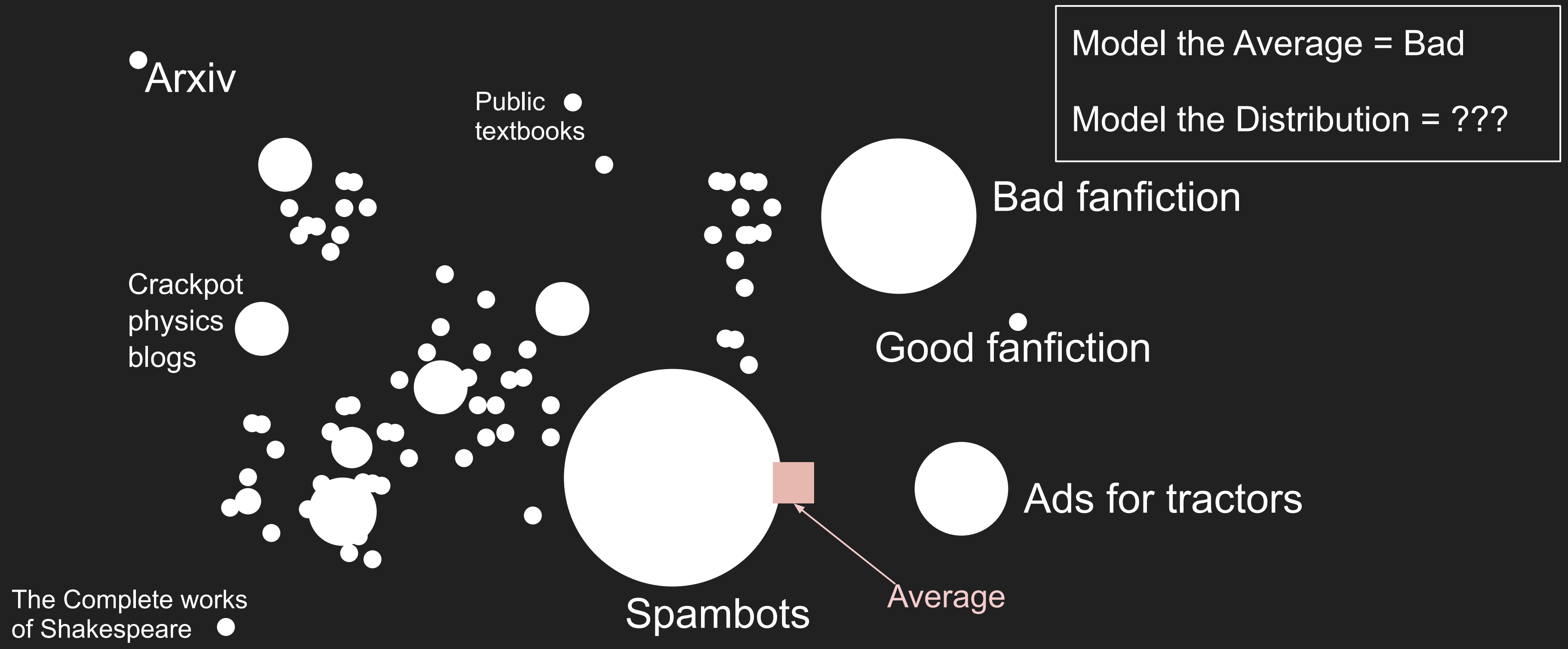
But where is the magic?

(Answer #1)



$\neq$

Full model of the distribution of human language: so what?



Full model of the distribution of human language: so what?

As good as the best human expert at every knowledge task.

- Research assistant
- Software engineer
- Data scientist

Prove or disprove the following lemma:

...

```
Git commit #467
Comment: Finally found that horrible bug
in the routing logic!

====Codebase before====

[code]

==== Diff ====

...
```

```
This data.csv file is really messy and has
lots of invalid entries. Here's the script
I wrote to fix it.

...

Also, I found a really cool correlation!
As you can see,
...
```

Full model of the distribution of human language: so what?

As good as the best human expert at every knowledge task.

A novel sonnet of Shakespeare, discovered in 2024, is likely authentic, experts say. Without further ado:

Sonnet 155
...

How I came upon the proof of the Riemann Hypothesis

By Terry Tao, 2024

Part 1
...

GPT-4

In at most four short bullets (suitable for presenting on a slide), define the Generative Pretraining Transformer (GPT) as used in GPT-4.

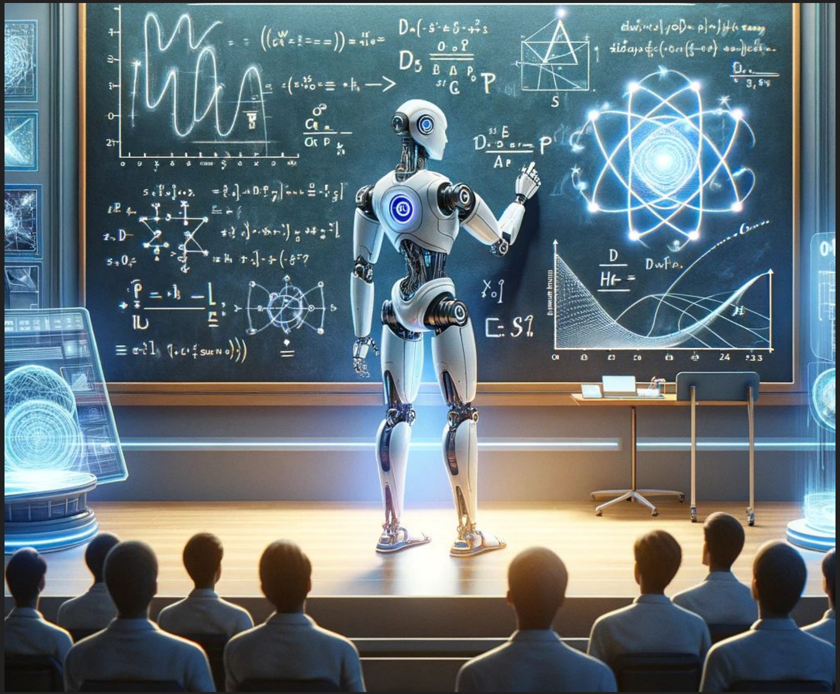
- GPT (Generative Pretraining Transformer) is an autoregressive language model that uses deep learning to produce human-like text.
- It pretrains on a large corpus of text data with unsupervised learning, predicting the next token in a sequence given the tokens that precede it.
- GPT-4, specifically, scales up this approach with a larger number of parameters, improved architecture, and training on diverse and extensive datasets.

GPT-4: not yet a full model of the distribution of human text.

But it is incredibly smart

SAT English	710/800
SAT Math	700/800
GRE Verbal	169/170
GRE Math	163/170
AP Physics 2	4/5
AP Calculus BC	4/5
AP Biology	5/5

https://openai.com/research/gpt-4
Source has many additional test results^



Part 4: GPT-4 demos

Home Practice Assignment Homework Log Out

Homework 5

Number of Students Who Viewed The Assignment : 634

Number of Students Who Submitted The Assignment : 276

Number of Students Who Did Not View The Assignment : 73

A+	A	B+	B	C+	C	D	F
84	91	30	9	18	21	4	19

Due Date: Wed Oct-11-2023 11:55 PM

Each answer in this assignment is worth a maximum of one point.

1. A boy pulls a 22.0-kg box with a 130-N force at 34° above a horizontal surface. If the coefficient of kinetic friction between the box and the horizontal surface is 0.19 and the box is pulled a distance of 40.0 m, what is the work done by the friction force on the box?

✖ -1.09×10^3 J

2. A sled is initially given a shove up a frictionless 22°

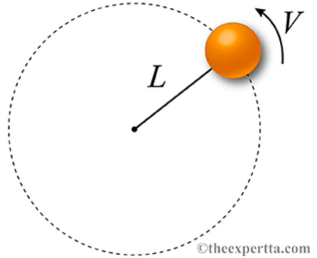
where it started. What was its initial speed?(g = 9.8 m

✖ **5.24** m/s

PHY-123 (Fall 2023) Homework 7

Problem 1: On the International Space Station an, object with mass $m = 363$ g is attached to a massless string of length $L = 0.89$ m. The string can handle a tension of $F_T = 9.97$ N before breaking. The object undergoes uniform circular motion, being spun around by the string in a horizontal plane.

https://usp32nj.theexpertta.com/Common/GradeSheetClassAssignments.aspx?eid=2559



What is the maximum speed, in meters per second, the mass can have before the string breaks?

Numeric : A numeric value is expected and not an expression.

v = ______________________ m/s

Problem 2: A common carnival ride, called a *gravitron*, is a large cylinder in which people stand against the wall of the ride as it rotates. At a certain point the floor of the cylinder lowers and the people are surprised that they don't slide down. Suppose the radius of the cylinder is $r = 15$ m, and the friction between the wall and their clothes is $\mu_s = 0.63$. Consider the tangential speed v of the ride's occupants as the cylinder spins.

Physics 385

Due date: Wednesday, Nov. 1

Griffiths reading: 3.3

Please read about Eq. (3.65), but we will be restricting ourselves to cases in which only a single value of l is present, as in the unnumbered equation above (3.65).

1. [7 points] $V(x, y)$ obeys Laplace's equation inside the rectangular region of the 2D plane defined by $-a \le x \le a$ and $-b \le y \le b$. The potential is zero on the vertical boundaries at $x = -a$ and $x = a$, and on the top and bottom boundaries it is given by $V(x, b) = V_0 \sin(\pi x/a)$ and $V(x, -b) = -V_0 \sin(\pi x/a)$, respectively. Find $V(x, y)$ inside the rectangle.

Hint: When you encounter a differential equation like $d^2H/dy^2 = k^2H$, your later

Answer Key: -1.09e3 J

General Physics 203 (prob 1)

1. A boy pulls a 22.0-kg box with a 130-N force at 34° above a horizontal surface. If the coefficient of kinetic friction between the box and the horizontal surface is 0.19 and the box is pulled a distance of 40.0 m, what is the work done by the friction force on the box?

-1.09×10^3 J

Answer Key: -1.09e3 J

✅

General Physics 203 (prob 1)

You

A boy pulls a 22.0-kg box with a 130-N force at 34° above a horizontal surface. If the coefficient of kinetic friction
between the box and the horizontal surface is 0.19 and the box is pulled a distance of 40.0 m, what is the work done by
the friction force on the box?

Physics Problem Solving Template

To solve this problem, we need to find the work done by the friction force. The work W done by a force is given by the formula:

$$W = \vec{F} \cdot \vec{d}$$

where:

- $\vec{F}$ is the force vector,
- $\vec{d}$ is the displacement vector,
- $\cdot$ denotes the dot product.

The work done by friction is given by $W_f = \vec{F}_{\text{friction}} \cdot \vec{d}$. The friction force $\vec{F}_{\text{friction}}$ can be calculated using the formula:

$$F_{\text{friction}} = \mu \cdot F_N$$

where:

- μ is the coefficient of kinetic friction,
- F_N is the normal force.

Now, let's find the normal force F_N. In this scenario, the box is on a horizontal surface, so the gravitational force F_g and the normal force F_N balance each other out:

$$F_N = F_g = m \cdot g$$

where:

- $m = 22.0\,\text{kg}$ is the mass of the box,
- $g = 9.8\,\text{m/s}^2$ is the acceleration due to gravity.

However, there's also a vertical component of the pulling force which affects the normal force. The vertical component of the pulling force is $F_{\text{pull, vertical}} = F_{\text{pull}} \cdot \sin(\theta)$, where θ is the angle the pulling force above the horizontal.

The effective normal force is then:

$$F_{\text{eff}} = F_N - F_{\text{pull, vertical}}$$

Now, we can calculate the friction force using the effective normal force:

$$F_{\text{friction}} = \mu \cdot F_{\text{eff}}$$

Finally, we can calculate the work done by friction. The displacement vector $\vec{d}$ is 40.0 m along horizontal, and since friction acts opposite to the direction of displacement, the angle between $\vec{F}_{\text{friction}}$ and $\vec{d}$ is 180°. Therefore, the dot product simplifies to:

$$W_f = F_{\text{friction}} \cdot d \cdot \cos(180^\circ)$$

Now, we can plug in the values and calculate the work done by the friction force on the box.

The work done by the friction force on the box is -1086.1 J. The negative sign indicates that the work is done against the direction of the displacement.

```python
import math

# Given values
m = 22.0  # kg
g = 9.8  # m/s^2
F_pull = 130  # N
theta = 34  # degrees
mu = 0.19
d = 40.0  # m

# Calculating the vertical component of the pulling force
F_pull_vertical = F_pull * math.sin(math.radians(theta))

# Calculating the normal force
F_N = m * g

# Calculating the effective normal force
F_eff = F_N - F_pull_vertical

# Calculating the friction force
F_friction = mu * F_eff

# Calculating the work done by the friction force
W_f = F_friction * d * math.cos(math.radians(180))
W_f
```

Answer Key: 2.89e4 J

✅

General Physics 203 (prob 5)

You

A 210-kg roller coaster reaches the top of the steepest hill with a speed of 5.20 km/h. It then descends the hill, which
is at an angle of 30° and is 45.0 m long. What will its kinetic energy be when it reaches the bottom? Assume μk = 0.22.

Physics Problem Solving Template

To solve this problem, we'll use the conservation of energy principle. At the top of the hill, the roller coaster has both potential and kinetic energy. As it descends, some of its potential energy is converted into kinetic energy, but some of that energy is also lost to friction.

1. **Initial Potential Energy**: The potential energy at the top due to its height h is:

$$PE_{\text{initial}} = m \cdot g \cdot h$$

1. **Initial Kinetic Energy**: The kinetic energy at the top due to its speed v is:

$$KE_{\text{initial}} = \frac{1}{2} m \cdot v^2$$

1. **Work done by Friction**: As the roller coaster descends the hill, friction does negative work on it:

$$W_{\text{friction}} = -\mu_k \cdot m \cdot g \cdot d \cdot \cos(\theta)$$

where:

- $\mu_k = 0.22$ is the coefficient of kinetic friction,
- $d = 45.0\,\text{m}$ is the length of the hill,
- $\theta = 30^\circ$ is the angle of the hill.

The negative sign indicates that friction opposes the motion.

1. **Final Kinetic Energy**: At the bottom of the hill, the roller coaster's potential energy is zero, so its total energy is all kinetic:

$$KE_{\text{final}} = PE_{\text{initial}} + KE_{\text{initial}} + W_{\text{friction}}$$

Given the height h of the hill can be found using trigonometry:

$$h = d \cdot \sin(\theta)$$

We can substitute the known values into the above equations to find KE_{final}.

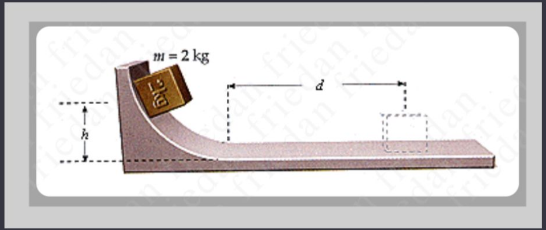
The kinetic energy of the roller coaster when it reaches the bottom of the hill will be $28,879.5\,\text{J}$ (or approximately $28.9\,\text{kJ}$).[>-]

Answer Key: 0.333

General Physics 203 (prob 4)

✅

You



The 2.0 kg block in the figure slides down a frictionless curved ramp, starting from rest at a height of h = 2.40 m. The
block then slides d = 7.20 m on a rough horizontal surface before coming to rest. What is the coefficient of friction
between the block and the horizontal surface?

Physics Problem Solving Template

To solve this problem, let's first consider the energy conservation principle.

1. **Potential Energy at the Top**: When the block is at height h, its potential energy is:

$$PE_{\text{initial}} = m \cdot g \cdot h$$

1. **Kinetic Energy at the Bottom**: When the block reaches the bottom of the ramp, it has been converted entirely into kinetic energy:

$$KE_{\text{bottom}} = \frac{1}{2}m \cdot v^2$$

where v is the velocity of the block at the bottom of the ramp.

By conservation of energy:

$$PE_{\text{initial}} = KE_{\text{bottom}}$$

$$m \cdot g \cdot h = \frac{1}{2}m \cdot v^2$$

From the above, we can solve for v.

1. **Work done by Friction**: As the block slides on the rough horizontal surface, the kinetic energy gets converted into work done against friction:

$$W_{\text{friction}} = \frac{1}{2}m \cdot v^2$$

The work done by friction is:

$$W_{\text{friction}} = \mu \cdot m \cdot g \cdot d$$

where:

- μ is the coefficient of friction,
- $d = 7.20\text{ m}$ is the distance the block slides on the horizontal surface.

Equating the two expressions for W_{friction}, we get:

$$\frac{1}{2}m \cdot v^2 = \mu \cdot m \cdot g \cdot d$$

From this equation, we can solve for μ.

Let's plug in the given values and calculate μ.

The coefficient of friction between the block and the horizontal surface is $\mu = 0.333$ or $\frac{1}{3}$.

General Physics 203 Homework 5

1	✅
2	✅
3	✅
4	✅
5	✅
6	✅
7	✅
8	✅
9	✅
10	❌

(N=1)

Answer Key: 919 J

General Physics 203 (prob 10)

You

A box of mass 5.60 kg is pulled from rest up a 34.0 m long incline that makes an angle of 16.0 with horizontal. The
coefficient of kinetic friction between the box and the incline is 0.180. At the top of the incline, the speed of the box is
5.40 m/s. What is the work done by the pulling force?

Physics Problem Solving Template

To solve this problem, we'll use the work-energy principle which states that the work done on an object is equal to the change in its mechanical energy. The mechanical energy of the box consists of its kinetic energy and gravitational potential energy.

The work done on the box is the sum of the work done by the pulling force W_p, the work done by the gravitational force W_g, and the work done by the friction force W_f.

$$W = W_p + W_g + W_f$$

The work done by the pulling force can be expressed as:

$$W_p = F_p \cdot d \cdot \cos(0^\circ)$$

where:

- F_p is the pulling force,
- $d = 34.0\,\mathrm{m}$ is the distance along the incline,
- $\cos(0^\circ) = 1$ (assuming the force is applied along the direction of motion).

The work done by the gravitational force is:

$$W_g = -m \cdot g \cdot d \cdot \sin(\theta)$$

where:

- $m = 5.60\,\mathrm{kg}$ is the mass of the box,
- $g = 9.81\,\mathrm{m/s^2}$ is the acceleration due to gravity,
- $\theta = 16.0^\circ$ is the angle of the incline.

The work done by the friction force is:

$$W_f = -\mu_k \cdot m \cdot g \cdot \cos(\theta) \cdot d$$

where:

- $\mu_k = 0.180$ is the coefficient of kinetic friction.

The change in mechanical energy of the box is the sum of the change in kinetic energy and the change in potential energy:

$$\Delta KE + \Delta PE = \frac{1}{2}mv^2 - \frac{1}{2}mu^2 + mgh - 0$$

where:

- $u = 0\,\mathrm{m/s}$ is the initial velocity,
- $v = 5.40\,\mathrm{m/s}$ is the final velocity,
- $h = d\sin(\theta)$ is the height of the incline.

Now, according to the work-energy principle:

$$W = \Delta KE + \Delta PE$$

Substituting the expressions for W, ΔKE, and ΔPE, we get:

$$F_p \cdot d + (-m \cdot g \cdot d \cdot \sin(\theta)) + (-\mu_k \cdot m \cdot g \cdot \cos(\theta) \cdot d) = \frac{1}{2}mv^2 - \frac{1}{2}mu^2 + mgh$$

Now we can solve this equation for the pulling force F_p and then use it to find the work done by

Now we can solve this equation for the pulling force F_p and then use it t
the pulling force W_p.

Let's start by calculating the height h of the incline:

$$h = d \cdot \sin(\theta)$$

The height (h) of the incline is approximately $9.37\,\mathrm{m}$.

Now we can substitute the values into the work-energy equation and so
F_p.

$$F_p \cdot d + (-m \cdot g \cdot d \cdot \sin(\theta)) + (-\mu_k \cdot m \cdot g \cdot \cos(\theta) \cdot d) = \frac{1}{2}mu$$

We'll rearrange this equation to solve for F_p:

$$F_p = \frac{\left(\frac{1}{2}mv^2 - \frac{1}{2}mu^2 + mgh\right) + (m \cdot g \cdot d \cdot \sin(\theta)) + (\mu_k \cdot m \cdot g \cdot \cos(\theta) \cdot d)}{d}$$

Now we can plug in the given values and solve for F_p.

The pulling force F_p is approximately $42.19\,\mathrm{N}$.

Now we can use this value to find the work done by the pulling force W_p

$$W_p = F_p \cdot d$$

The work done by the pulling force W_p is approximately $1434.52\,\mathrm{J}$.

Answer Key: 8L/3

Honors Physics 1 (prob 1)

You

A rocket accidentally explodes at the top of its trajectory. The horizontal distance between the launch point and the point of explosion is L. The rocket breaks into two pieces that fly apart horizontally. The larger piece has three times the mass of the smaller piece. To the surprise of the scientist in charge, the smaller piece returns to Earth at the launching station. How far away does the larger piece land? Neglect air resistance and effects due to the Earth's curvature. You can assume that the rocket's trajectory was parabolic throughout its flight.

Answer Key: 8L/3

Honors Physics 1 (prob 1)

You

A rocket accidentally explodes at the top of its trajectory. The horizontal distance between the launch point and the point of explosion is L. The rocket breaks into two pieces that fly apart horizontally. The larger piece has three times the mass of the smaller piece. To the surprise of the scientist in charge, the smaller piece returns to Earth at the launching station. How far away does the larger piece land? Neglect air resistance and effects due to the Earth's curvature. You can assume that the rocket's trajectory was parabolic throughout its flight.

Physics Problem Solving Template

In this problem, we are dealing with conservation of momentum. When the rocket explodes, the total momentum of the pieces immediately after the explosion must equal the total momentum of the rocket just before the explosion. Just before the explosion, the rocket's momentum is zero because it is at the top of its trajectory and momentarily at rest. Therefore, the total momentum immediately after the explosion is also zero.

Let m be the mass of the smaller piece and $3m$ be the mass of the larger piece. Let v be the velocity of the smaller piece and $-3v$ be the velocity of the larger piece (assuming positive direction to the right). The total momentum just after the explosion is given by

$$m \cdot v + 3m \cdot (-3v) = 0$$

Answer Key: R/3

Honors Physics 1 (prob 2)

You

A small block slides from rest from the top of a frictionless sphere of radius R. How far below the top x does it lose contact with the sphere? The sphere does not move.

Answer Key: R/3

✅

Honors Physics 1 (prob 2)

You

A small block slides from rest from the top of a frictionless sphere of radius R. How far below the top x does it lose contact with the sphere? The sphere does not move.

Physics Problem Solving Template

To solve this problem, we'll consider the forces and motion involved. When the block is at some angle θ with respect to the vertical axis, it experiences two forces: gravitational force mg downward, and a normal force N directed radially outward.

At the point of losing contact, the normal force becomes zero. At this instant, the only force acting on the block is the gravitational force. We can decompose this force into two components: one along the radial direction ($mg\cos(\theta)$), and one along the tangential direction ($mg\sin(\theta)$).

We can apply Newton's second law in the radial and tangential directions. In the radial direction, we have:

$$m\frac{v^2}{R} = mg\cos(\theta)$$

where v is the velocity of the block.

In the tangential direction, we can use the work-energy principle. The work done by gravity is equal to the change in kinetic energy of the block:

$$mgR(1-\cos(\theta)) = \frac{1}{2}mv^2$$

Now we can solve these two equations simultaneously to find the angle θ at which the block loses contact with the sphere. Once we have the angle, we can find the distance x below the top as:

$$x = R(1-\cos(\theta))$$

Let's start by solving the two equations simultaneously for θ.

The solution gives two possible angles θ at which the block loses contact with the sphere. However, the second angle $\theta \approx 5.44\,\mathrm{rad}$ (which is beyond π) is not physically meaningful in this scenario. Therefore, we'll consider only the first angle $\theta \approx 0.841\,\mathrm{rad}$.

Now, we can use this angle to find the distance x below the top at which the block loses contact with the sphere:

$$x = R(1-\cos(\theta))$$ [>-]

The distance x below the top at which the block loses contact with the sphere is given by:

$$x = \frac{1}{3}R$$

This means the block loses contact with the sphere one-third of the radius below the top of the sphere. [>-]

Answer Key: $\frac{1}{2} g \left(t1 - t_2\right)^2 \frac{t_1}{t_2}$

Honors Physics 1 (prob 3)

You

At t = 0, an elevator leaves the ground with uniform speed. At time t1, a child, lying down in the elevator, gently drops a coin through the (open!) elevator floor. The coin falls with uniform acceleration g and hits the ground at time t2. At what height, h, was the elevator when the child dropped the coin?

Answer Key: $\frac{1}{2}g\left(t1 - t_2\right)^2 \frac{t_1}{t_2}$

Honors Physics 1 (prob 3)

??

You

At t = 0, an elevator leaves the ground with uniform speed. At time t1, a child, lying down in the elevator, gently drops a coin through the (open!) elevator floor. The coin falls with uniform acceleration g and hits the ground at time t2. At what height, h, was the elevator when the child dropped the coin?

Physics Problem Solving Template

In this problem, the motion of the elevator and the coin can be treated separately to find the height h from which the coin was dropped.

1. **Elevator Motion:**
   The elevator is moving with a uniform speed. This means there is no acceleration, and we can use the equation of motion:
   $d = vt$
   where:
   - d is the distance the elevator has traveled,
   - v is the velocity of the elevator,
   - t is the time.

   The height of the elevator at time t_1 is given by $h = vt_1$.
2. **Coin Motion:**
   When the child drops the coin at time t_1, the coin initially has the same velocity as the elevator. However, once dropped, the only force acting on the coin is gravity, which causes it to accelerate downwards at g.
   The equation of motion for the coin is:
   $d = vt + \frac{1}{2}at^2$
   where:
   - d is the distance the coin falls,
   - v is the initial velocity of the coin (which is the same as the elevator's velocity),
   - a is the acceleration (which is $-g$ as it's directed downwards),
   - t is the time of fall.

We are interested in the time interval from t_1 to t_2, so we'll use $\Delta t = t_2 - t_1$ as the time the coin falls. Therefore, the distance the coin falls is h, the height of the elevator at t_1.
Substituting the values, we get:
$h = v(t_2 - t_1) - \frac{1}{2}g(t_2 - t_1)^2$

Now, we need to solve for h.

Upon reviewing the steps, it's apparent that the expression for the height h from which the coin was dropped should be framed in terms of the known quantities t_1 and t_2, and the given uniform acceleration g. The velocity v of the elevator is not provided, but we can eliminate v using the expressions derived from the motions of the elevator and the coin.

From the elevator's motion, we have:
$h = vt_1$ (Equation 1)

From the coin's motion, we have:
$h = v(t_2 - t_1) - \frac{1}{2}g(t_2 - t_1)^2$ (Equation 2)

Now, we can solve these two equations simultaneously to eliminate v and solve for h in terms of t_1, t_2, and g.

Let's perform this algebraic manipulation.

The expression for the height h from which the coin was dropped, in terms of t_1, t_2, and g, is given by:

$$h = \frac{gt_1(-t_1^2 + 2t_1t_2 - t_2^2)}{(4t_1 - 2t_2)}$$

This formula will provide the height of the elevator at the time t_1 when the child dropped the coin, given the times t_1 and t_2, and the acceleration due to gravity g. [>_]

[No Answer Key]

Physics 385 (prob 1)

You

V (x, y) obeys Laplace’s equation inside the rectangular region of the 2D planedefinedby−a≤x≤aand−b≤y≤b. Thepotentialiszeroonthevertical boundaries at x = −a and x = a, and on the top and bottom boundaries it is given by V (x, b) = V0 sin(πx/a) and V (x, −b) = −V0 sin(πx/a), respectively. Find V (x, y) inside the rectangle.

Hint: When you encounter a differential equation like d2H/du2 = k2H, your later algebra will be simplified if you take a solution in the form H = A cosh(ku) + B sinh(ku), instead of the (equivalent) sum of two exponentials. In this problem, you should be able to argue that either A or B is zero.

[No Answer Key]

Physics 385 (prob 1)

You

V (x, y) obeys Laplace's equation inside the rectangular region of the 2D planedefinedby−a≤x≤aand−b≤y≤b. Thepotentialiszeroonthevertical boundaries at x = −a and x = a, and on the top and bottom boundaries it is given by V (x, b) = V0 sin(πx/a) and V (x, −b) = −V0 sin(πx/a), respectively. Find V (x, y) inside the rectangle.
Hint: When you encounter a differential equation like d2H/du2 = k2H, your later algebra will be simplified if you take a solution in the form H = A cosh(ku) + B sinh(ku), instead of the (equivalent) sum of two exponentials. In this problem, you should be able to argue that either A or B is zero.

Physics Problem Solving Template

To solve for $V(x,y)$ inside the rectangle, we will use the method of separation of variables. We'll assume that the potential $V(x,y)$ can be separated into two functions, $X(x)$ and $Y(y)$, such that $V(x,y) = X(x)Y(y)$.

Now, Laplace's equation in two dimensions is given by:

$$\nabla^2 V = \frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} = 0$$

Substituting the expression for $V(x,y)$, we get:

$$\frac{\partial^2 (X(x)Y(y))}{\partial x^2} + \frac{\partial^2 (X(x)Y(y))}{\partial y^2} = 0$$

Now, we'll divide through by $X(x)Y(y)$ to separate the variables:

$$\frac{1}{X(x)}\frac{d^2X}{dx^2} + \frac{1}{Y(y)}\frac{d^2Y}{dy^2} = 0$$

This equation can be separated into two ordinary differential equations by setting each term equal to a separation constant $-k^2$. We then have:

$$\frac{1}{X(x)}\frac{d^2X}{dx^2} = -k^2$$
$$\frac{1}{Y(y)}\frac{d^2Y}{dy^2} = k^2$$

Now we can solve these ordinary differential equations separately. For $X(x)$, the differential

[...]

Let's start by applying the boundary condition at $y = b$:

$$V(x,b) = \sum_n (A_n \cos(k_n x) + B_n \sin(k_n x)) D_n \sinh(k_n b) = V_0 \sin\left(\frac{\pi x}{a}\right)$$

This is a Fourier sine series, which means that the terms in the series must match the terms of the Fourier sine series for $V_0 \sin\left(\frac{\pi x}{a}\right)$. This gives us the condition that $k_n = \frac{n\pi}{a}$ for $n = 1, 3, 5, \ldots$.

Now we can solve for the coefficients A_n, B_n, and D_n in terms of V_0, a, and b.

The expression for $V(x,b)$ can be rewritten in terms of a sum of sines:

$$V(x,b) = \sum_n (A_n \cos(k_n x) + B_n \sin(k_n x)) D_n \sinh(k_n b) = V_0 \sin\left(\frac{\pi x}{a}\right)$$

Comparing the right-hand side and the left-hand side, we can deduce that the only term that can potentially match the right-hand side is when $k_n = \frac{n\pi}{a}$ for $n = 1, 3, 5, \ldots$, as mentioned before. This corresponds to the odd terms in a Fourier sine series.

Now, for the terms to match exactly, we need to have $B_1 D_1 = V_0$ and $A_n = B_n = D_n = 0$ for all $n \neq 1$.

Now, we can write the expression for $V_n(x,y)$ for $n = 1$ as:

$$V_1(x,y) = B_1 D_1 \sin\left(\frac{\pi x}{a}\right) \sinh\left(\frac{n\pi y}{a}\right)$$

Now, substituting $B_1 D_1 = V_0$ and $n = 1$, we get:

$$V_1(x,y) = V_0 \sin\left(\frac{\pi x}{a}\right) \sinh\left(\frac{\pi y}{a}\right)$$

So, the potential $V(x,y)$ inside the rectangle is given by:

$$V(x,y) = V_0 \sin\left(\frac{\pi x}{a}\right) \sinh\left(\frac{\pi y}{a}\right)$$

Rutgers Physics Problem Sets Eval

Physics 203

1	✅
2	✅
3	✅
4	✅
5	✅
6	✅
7	✅
8	✅
9	✅
10	❌ (3/5)

Honors Physics I

1	❌ (0/5)
2	✅
3	??
4	❌ (0/5)

Phys 385

1	❌ (1/5)
2	N/A*
3	✅**

GPT4 ~ mediocre student

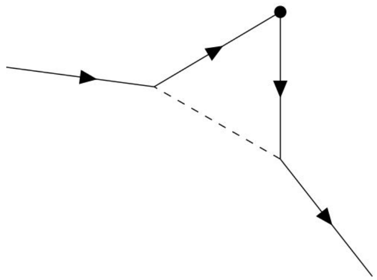
[N=1]

*Refers to a previous homework problem.
**No answer key, but looked right to me.

What about something harder?

Homework 2

1. Compute the anomalous dimension of the fermion mass operator $\overline{\psi}\psi(x)$ in axial Yukawa theory at one loop. The relevant Feynman diagram is:



Physics Problem Solving Template

To solve for the anomalous dimension of the fermion mass operator in axial Yukawa theory at one loop using renormalization group (RG) methods, we would typically follow these steps:

Step 1: Write down the one-loop integral

The one-loop correction to the fermion propagator involves an integral over the loop momentum, which can be written as:

$$\Sigma(p) = (-ig)^2 \int \frac{d^4k}{(2\pi)^4} \frac{i(\not{k}+m)i\gamma^5}{k^2 - m^2 + i\epsilon} \frac{i}{(p-k)^2 - M^2 + i\epsilon},$$

where $\not{k}$ denotes the Feynman slash notation $\gamma^\mu k_\mu$, and γ^5 is the chirality matrix.

Step 2: Regularization

To regulate the divergent integral, we would use a scheme such as dimensional regularization, where we analytically continue the number of dimensions from 4 to $D = 4 - \epsilon$ to handle the divergences as poles in ϵ.

Step 3: Perform the integral

Step 5: Anomalous dimension

The anomalous dimension γ_m of the mass operator is defined in terms of the beta function $\beta(g)$ and the mass function $\gamma_m(g)$ as follows:

$$\gamma_m(g) = -\frac{\mu}{m}\frac{dm}{d\mu}\bigg|_g,$$

where μ is the renormalization scale. The β function and γ_m function can be calculated from the renormalized quantities as:

$$\beta(g) = \mu\frac{dg}{d\mu}, \quad \gamma_m(g) = \frac{\mu}{m}\frac{dm}{d\mu}.$$

Using the Callan-Symanzik equation, we relate these functions to the derivatives of the renormalized Green's functions.

Step 6: Solve the RG equations

Finally, we would solve the RG equations to determine how the renormalized mass and coupling constant run with the scale μ. The solution to these equations gives us the energy dependence of

The specific algebraic and numerical work involved in these steps is nontrivial and requires expertise in quantum field theory and the use of computational tools designed for this purpose. For an actual computation, one would typically turn to a symbolic algebra program capable of handling such integrals.

Reminder: 2019 (GPT-2)

Legolas and Gimli advanced on the orcs, raising their weapons with a harrowing war cry.

The orcs' response was a deafening onslaught of claws, claws, and claws; even Elrond was forced to retreat. "You are in good hands, dwarf," said Gimli, who had been among the first to charge at the orcs; it took only two words before their opponents were reduced to a blood-soaked quagmire, and the dwarf took his first kill of the night. The battle lasted for hours until two of the largest Orcs attempted to overwhelm Aragorn. When they finally stopped, they lay defeated and lifeless for miles and miles.

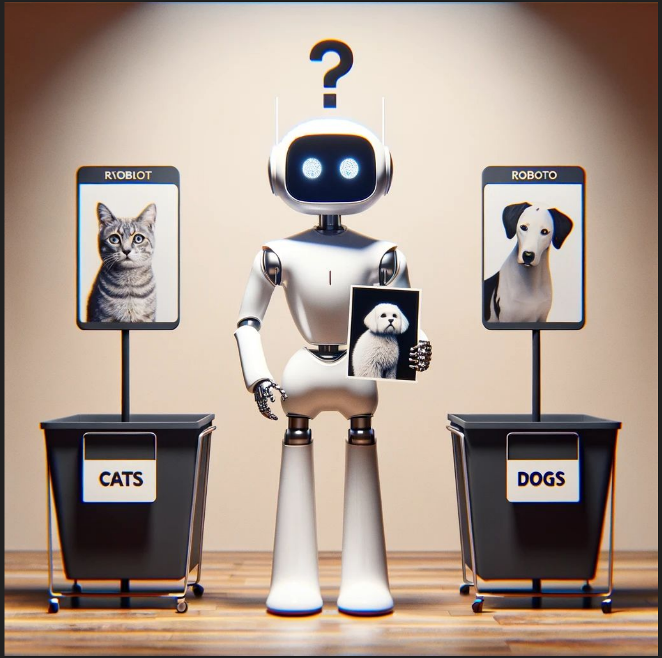
"I take nothing," said Aragorn. "But I give my word, at my peril and mine, that I will never forget this day of horror. None of us will forget. Ever!"

"I'll never forget it!" cried Gimli, who had been in the thick of the battle but hadn't taken part in it. One of the wounded orcs

openai.com/research/better-language-models

Reminder: 2012

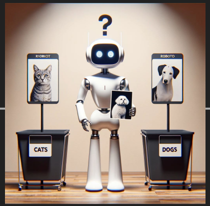
AlexNet achieves 84.7% accuracy on image classification



Timeline of major AI models

Training flops

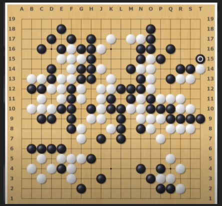
10^{21}
10^{20}
10^{19}
10^{18}



AlexNet

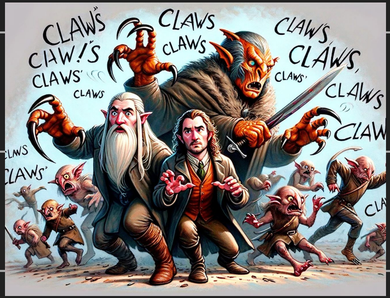
ResNet

Xception



AlphaGo

AlphaZero



GPT-2

OpenAI5

GPT-3

PaLM

Chinchilla

GPT-4
(2023)

"It missed the factor of cos(θ)."

"Can't even do RG calculations without Mathematica!"

2012 2014 2016 2018 2020 2022

GPT-4 Capabilities Summary

True, But Not The Key Point

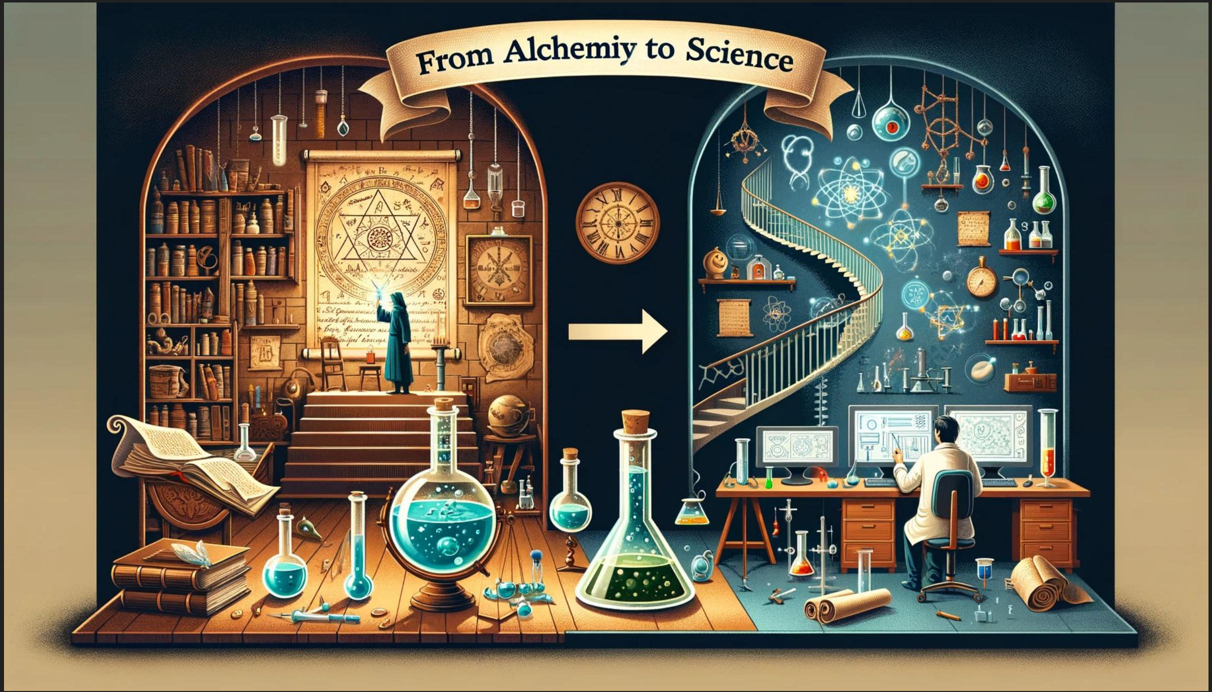
GPT-4 is pretty cool, does well on high school exams

The Key Point

d(intelligence)/d(years) is insane and is not slowing down.
Don't ask about how today's models will change the world in ten years.
Ask how how models ten years more powerful will change the world.

~~How will GPT-6 do on AP exams?~~

What tests will we be giving GPT-6?



Part 5: From Alchemy To Science

Scaling Laws (2019)

Intelligence ~ Compute?

Measure flops

Measure performance

See if there's a pattern

Scaling Laws for Neural Language Models

Jared Kaplan *
Johns Hopkins University, OpenAI
jaredk@jhu.edu

Sam McCandlish*
OpenAI
sam@openai.com

Tom Henighan
OpenAI
henighan@openai.com

Tom B. Brown
OpenAI
tom@openai.com

Benjamin Chess
OpenAI
bchess@openai.com

Rewon Child
OpenAI
rewon@openai.com

Scott Gray
OpenAI
scott@openai.com

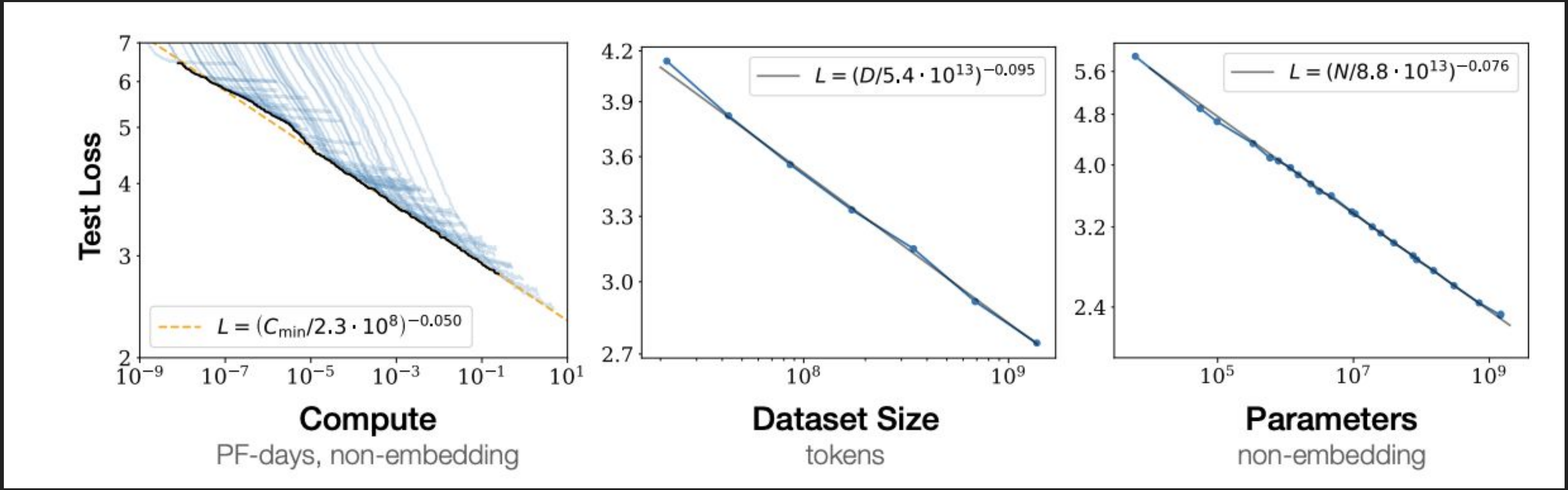
Alec Radford
OpenAI
alec@openai.com

Jeffrey Wu
OpenAI
jeffwu@openai.com

Dario Amodei
OpenAI
damodei@openai.com

https://arxiv.org/pdf/2001.08361.pdf

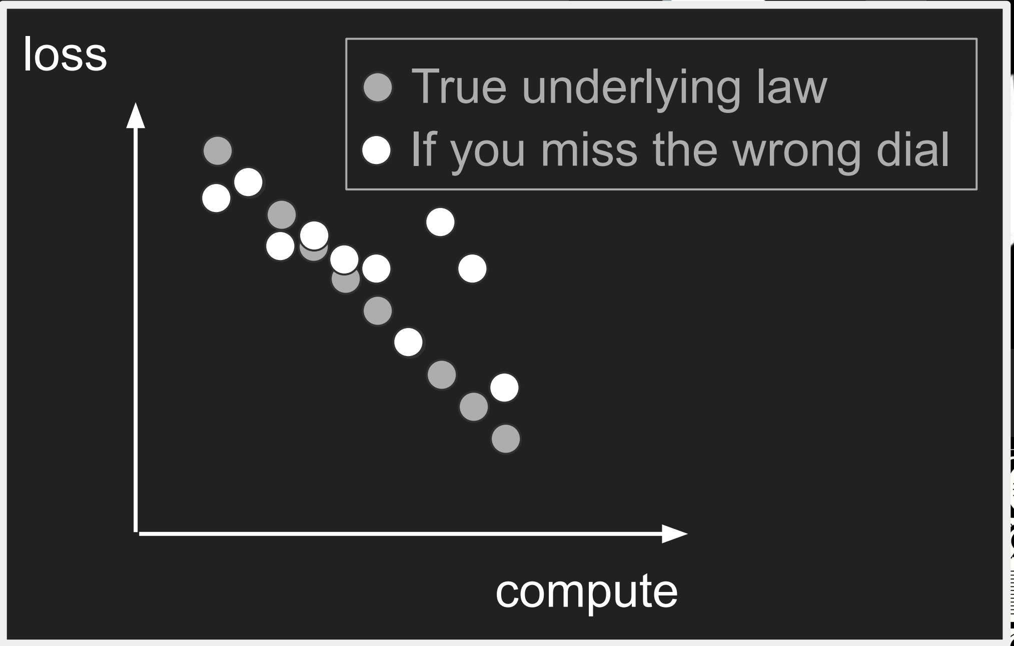
Scaling Laws



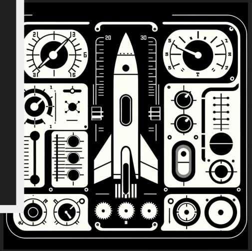
https://arxiv.org/pdf/2001.08361.pdf

Scaling Laws: Why is this hard?

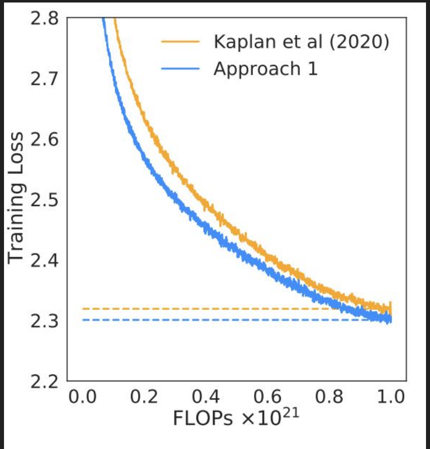
Models



Controls



Chinchilla (2023)



Original Scaling Laws had an important dial set wrong

	Compute used	=	model size	x	dataset size
Original	X		X		1
Chinchilla	X		√X		√X

https://arxiv.org/pdf/2203.15556.pdf

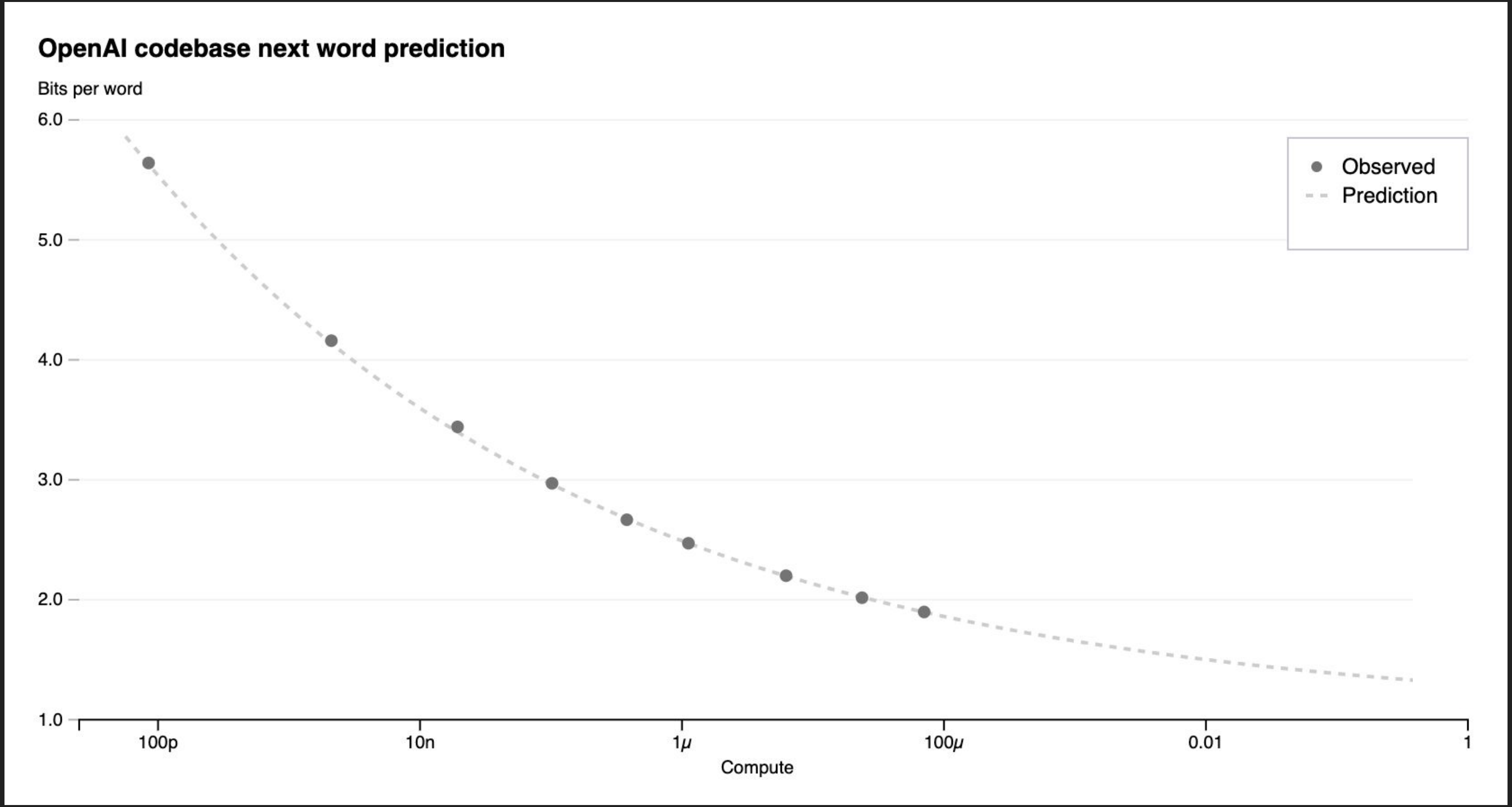
GPT-4

GPT-4: not yet a full model of the distribution of human text

It was fundamentally a science experiment targeting a single question:

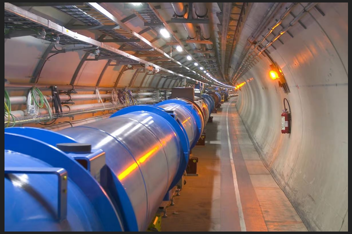
Can we get the model of flops -> intelligence to extrapolate well to this enormous flops level?

Predictable Scaling of GPT-4



https://openai.com/research/gpt-4

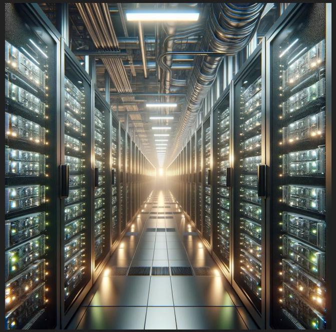
Conclusion: should you believe the hype?

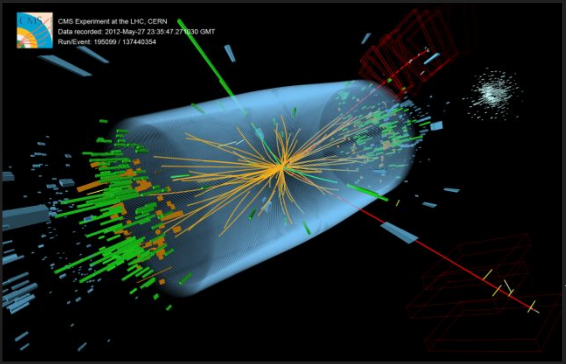


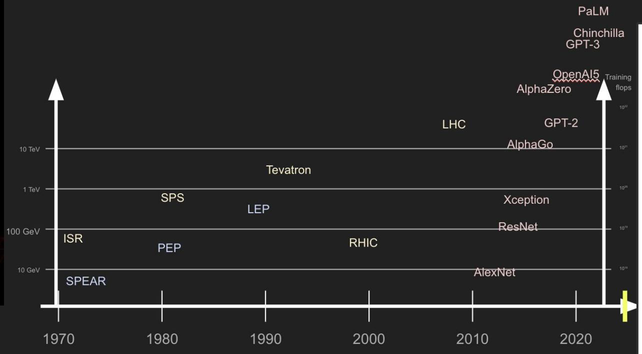
1. The hype about GPT-4 itself, the current model?
   It's a middling physics student. Impressive? 🤷

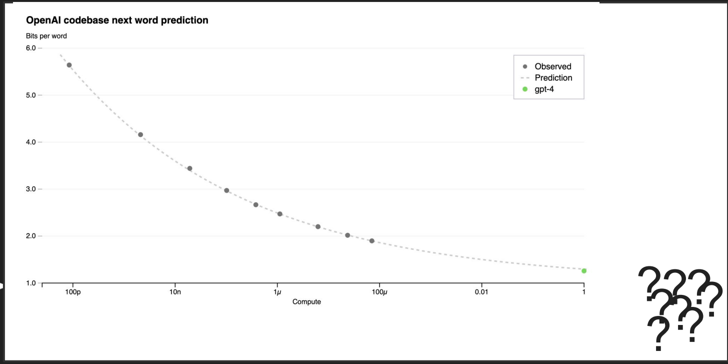
2. Scaling up compute has produced predictable results so far (with a lot of work); it may well continue.

3. This is our generation's collider program.
   It's hard to fathom what is coming next.









Bonus Slides

Bonus Slides Section: What's coming

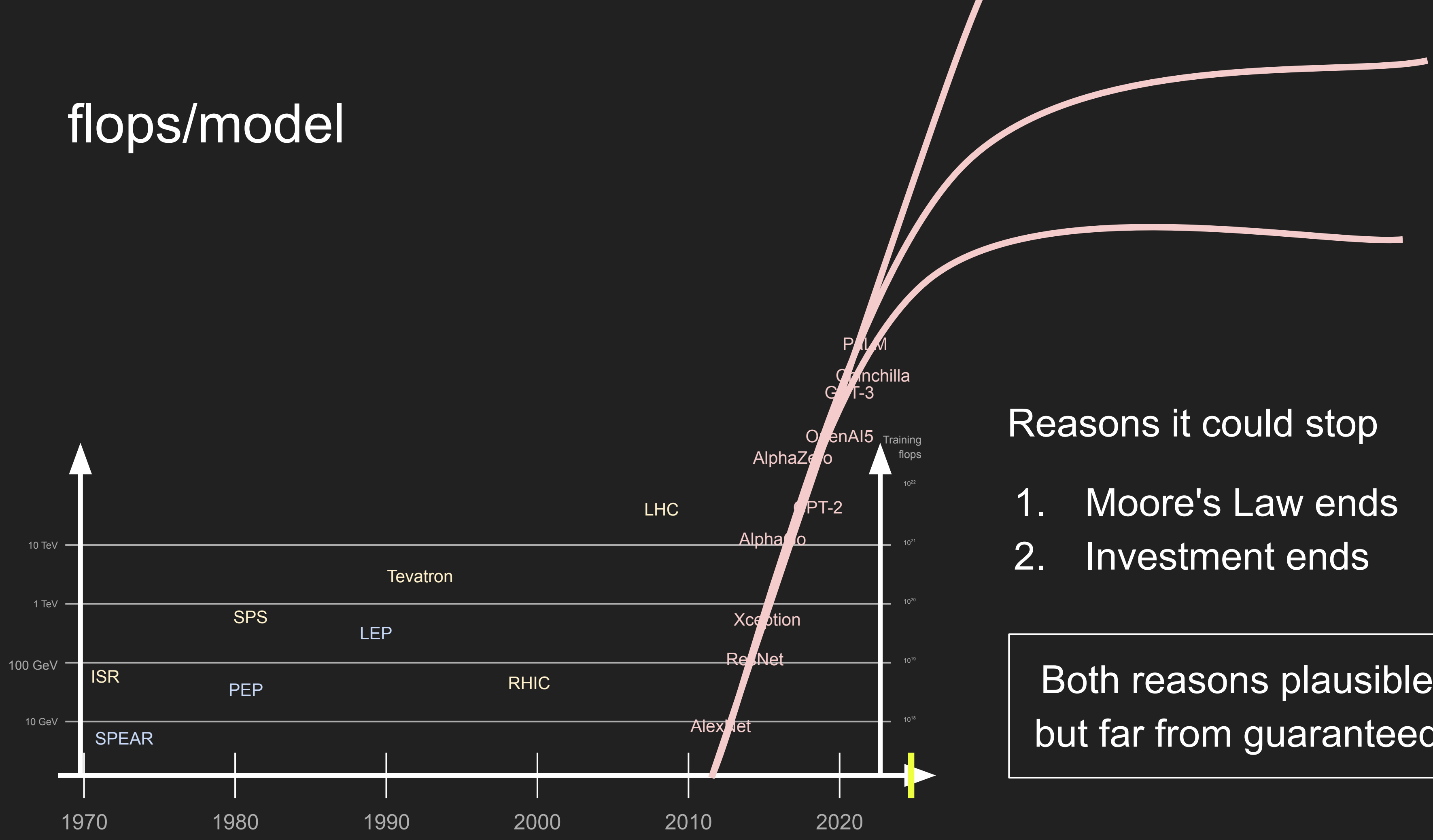
The Recipe So Far

$$$ Invested → Build a supercomputer → Train a model

$$\text{Intelligence} = \frac{\$}{\text{model}} \times \frac{\text{flops}}{\$} \times \frac{\text{capability}}{\text{flops}}$$

Investment　　Hardware　　Scaling

flops/model



Reasons it could stop

1. Moore's Law ends
2. Investment ends

Both reasons plausible
but far from guaranteed

The Recipe So Far

$$$ Invested → Build a supercomputer → Train a model

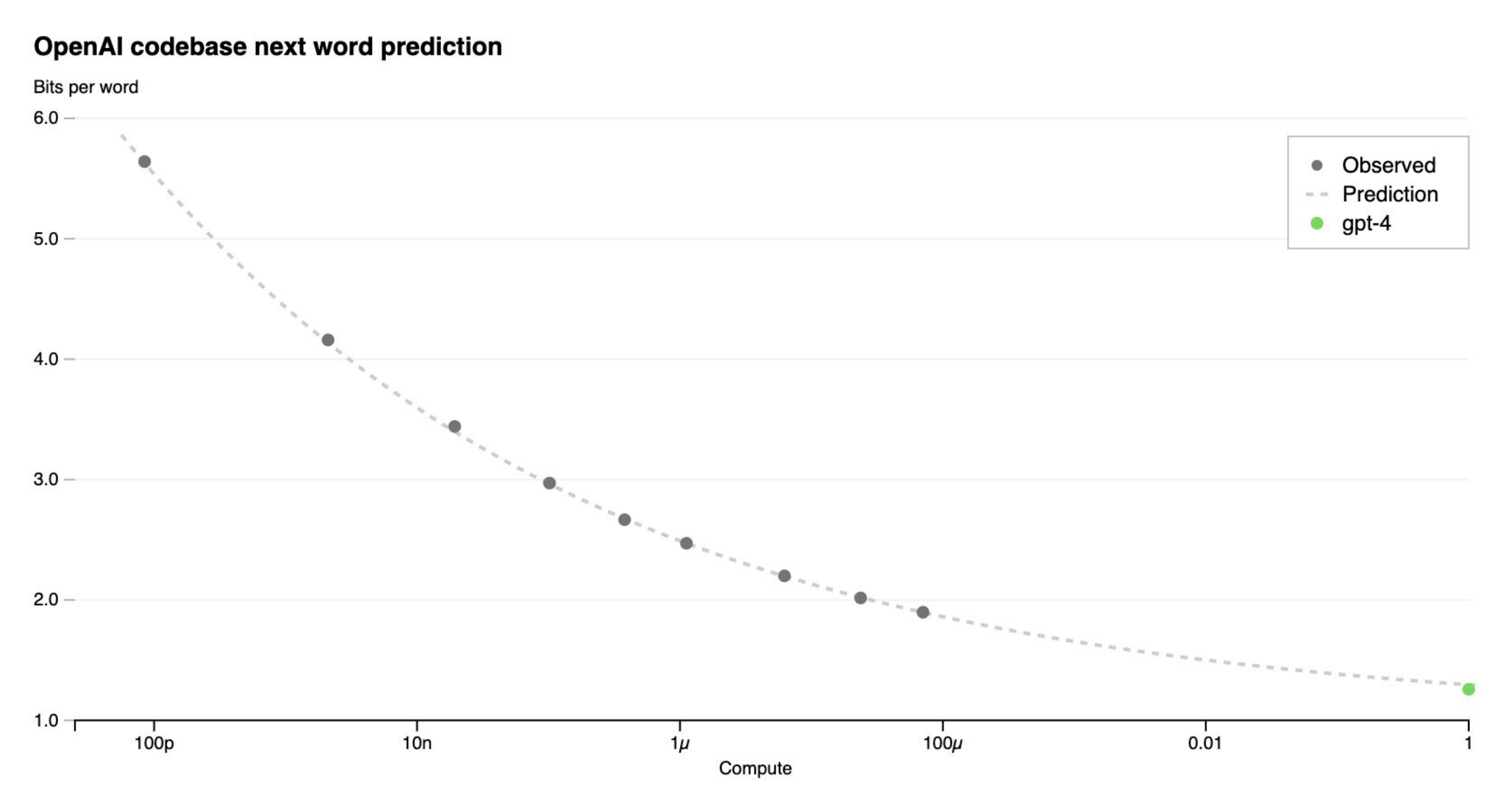
$$\text{Intelligence} = \frac{\$}{\text{model}} \times \frac{\text{flops}}{\$} \times \frac{\text{capability}}{\text{flops}}$$

Investment

Hardware

Scaling

Do we keep getting more intelligence by pouring in more flops?



? ? ? ? ? ? ? ?

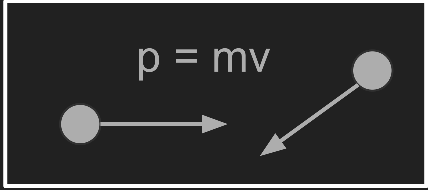
Bonus Slides Section:
Other / mishmash

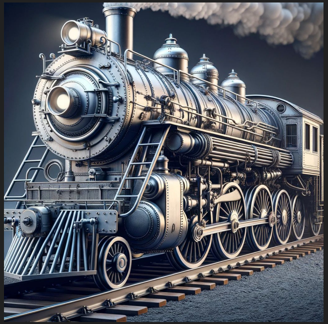
But where is the magic?

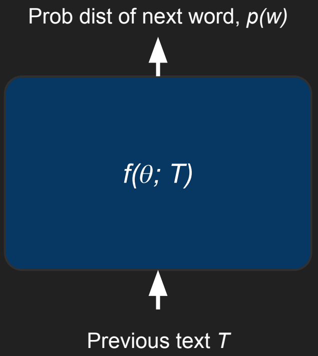
(Answer #2)

$$\boxed{N \longrightarrow \infty}$$

Emergent theories at large N are nothing new.
Intelligence is no exception.







????

Let's not get ahead of ourselves...

You

Can you name three countries with an A as the fourth letter?

‹ 3 / 3 ›

ChatGPT

Certainly:

1. Brazil
2. Croatia
3. Australia

conducting an experiment. This design should effectively convey the evolution from ancient practices to contemporary scientific methodologies, suitable for your presentation slide heading "From Alchemy to Science". If this aligns with your vision or if you have further adjustments in mind, please let me know.

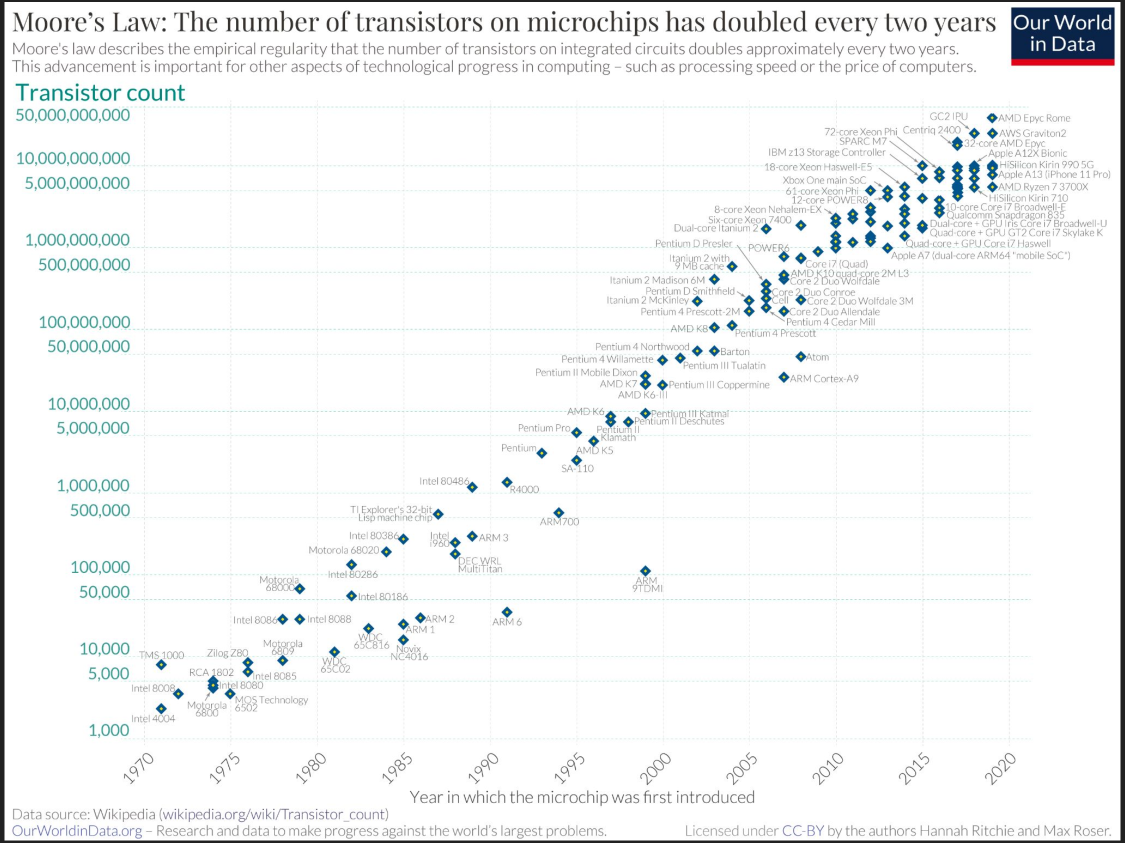
You

Amazing. Now a similar one for the section "The Pace of Progress in Compute". I'm not sure what would be the best concept but maybe you can think of something.

ChatGPT

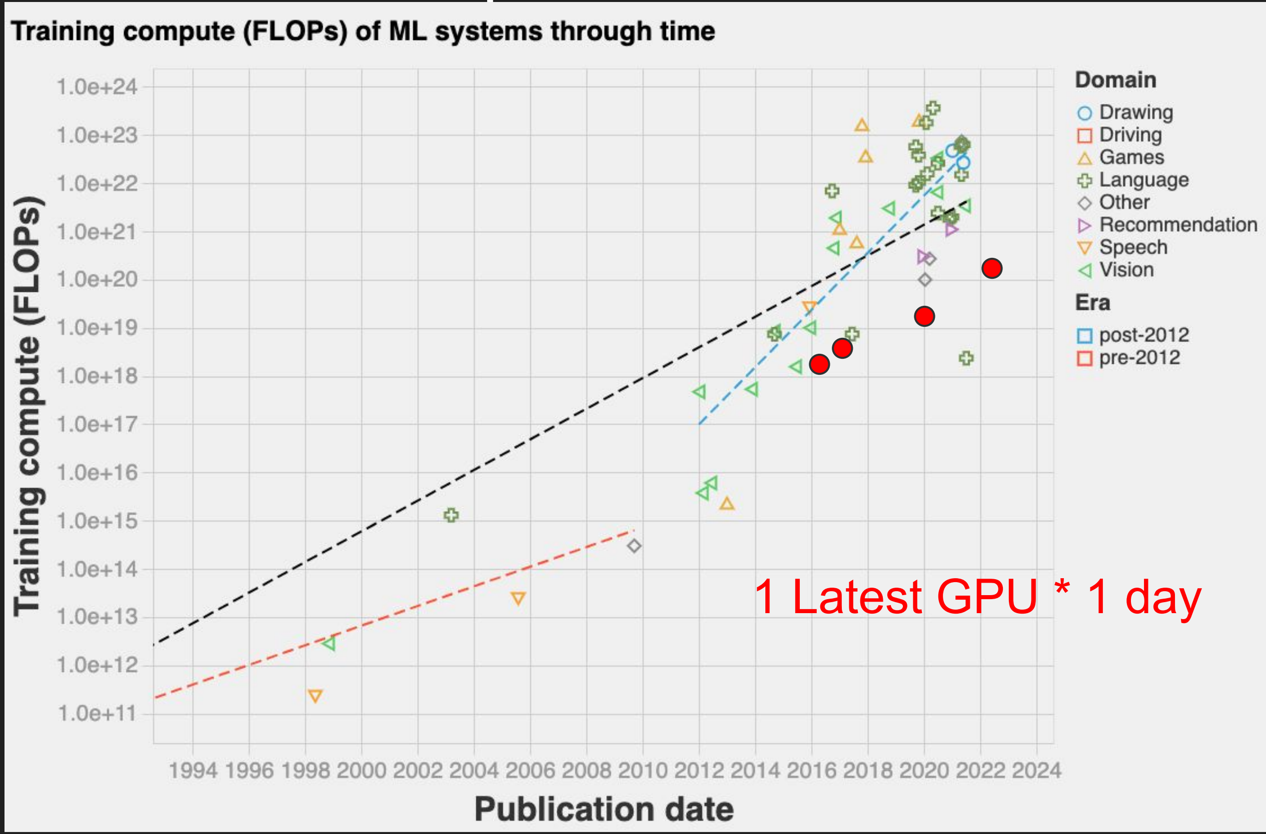
Bonus Slides Section: Important Plots

Moore's Law



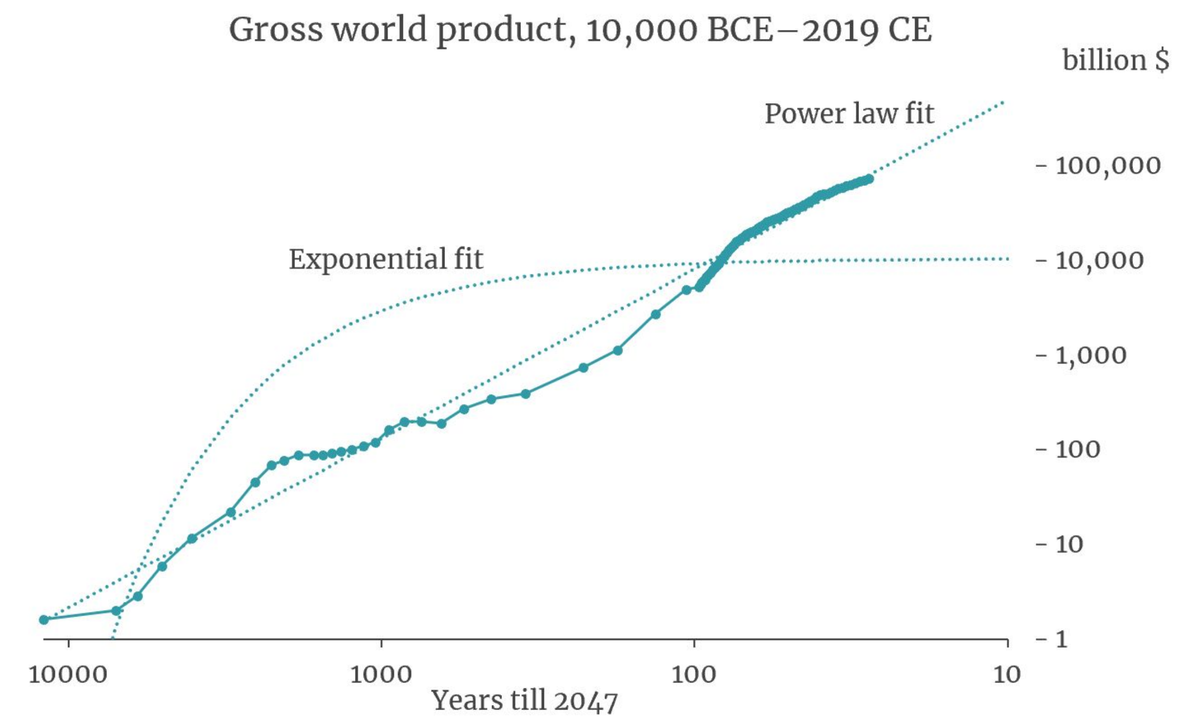
Source

Another reference for flops vs time

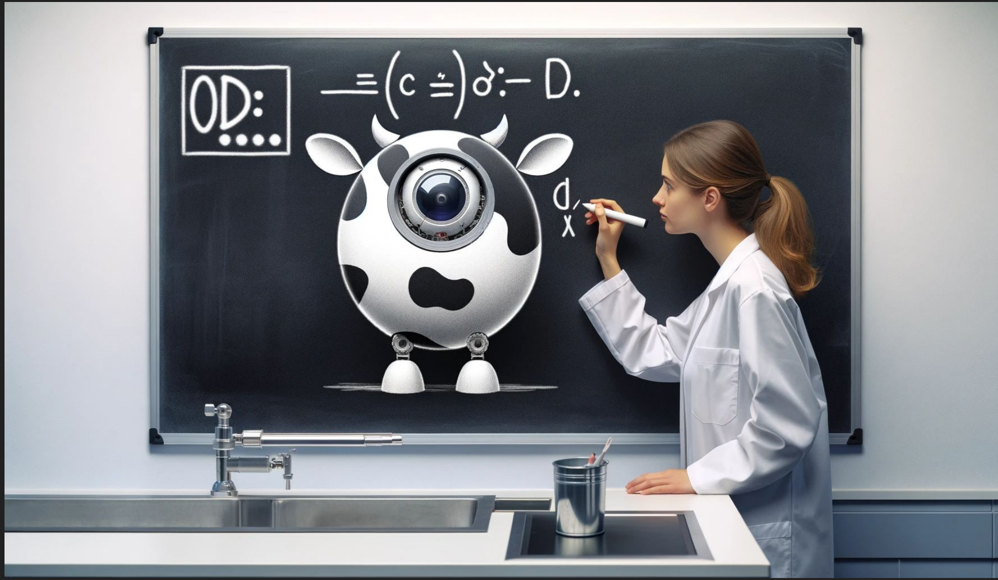


Source

The future is going to be weird

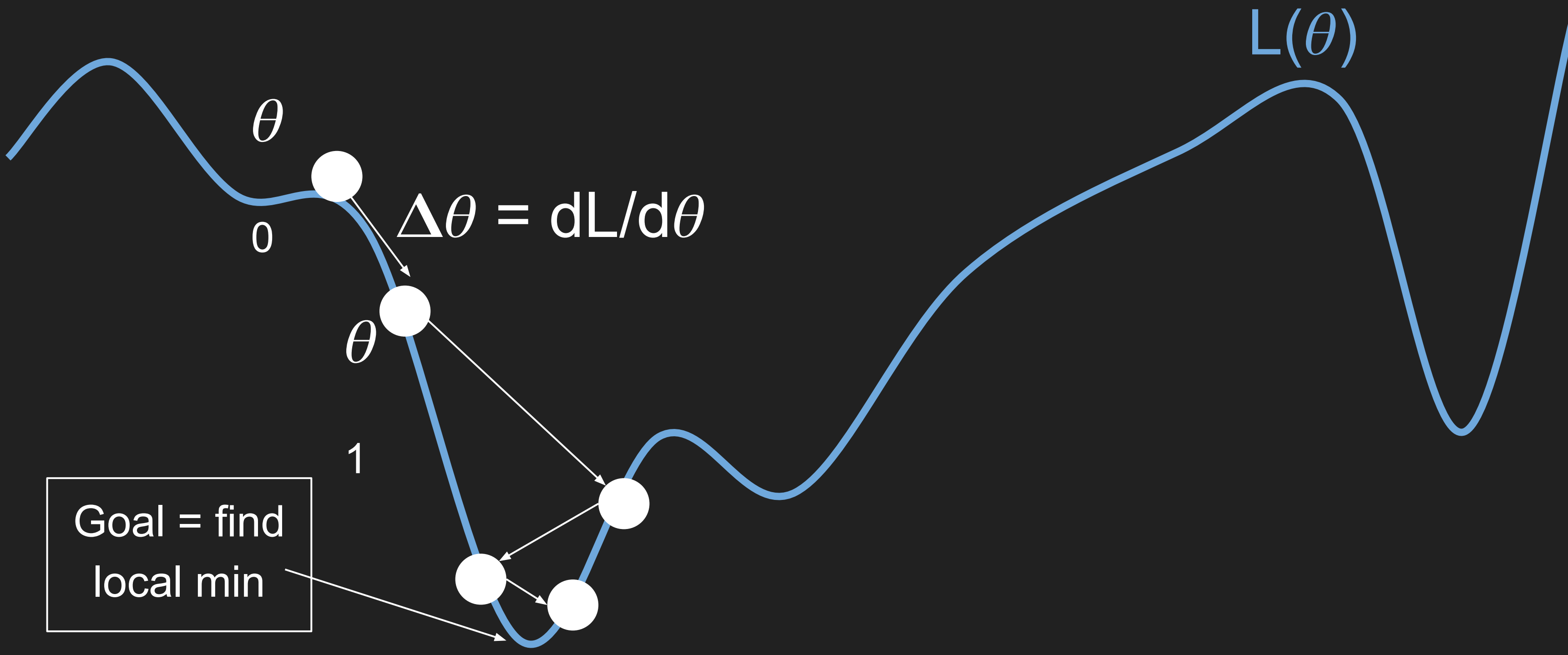


https://www.openphilanthropy.org/research/modeling-the-human-trajectory



Bonus Slides Section:
Machine Learning For Physicists 101

Gradient Descent



Functional Gradient Descent

Requirements:

1. Want a function $f: X \rightarrow Y$
2. Have some functional to assess quality $L[f]$

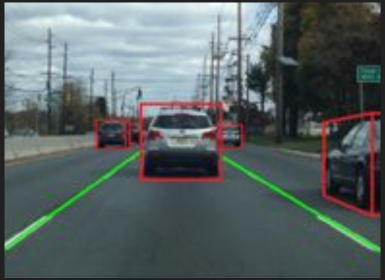
Functional Gradient Descent:

1. Start with f_0
2. Measure $\boldsymbol{\delta} L/\boldsymbol{\delta} f$
3. Update: $f_1 = f_0 - \alpha(\boldsymbol{\delta} L/\boldsymbol{\delta} f)$
4. Repeat. A lot.

Key Insight: EVERYTHING is functional gradient descent

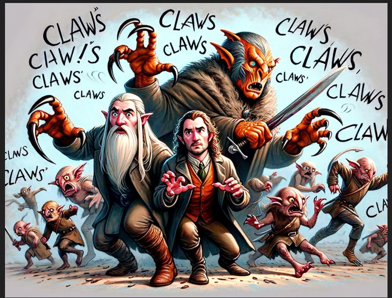
Requirements:

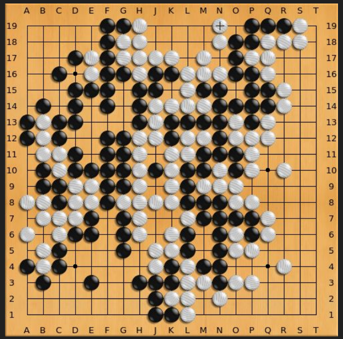
1. Want a function f: X→Y
2. Have some functional to assess quality L[f]

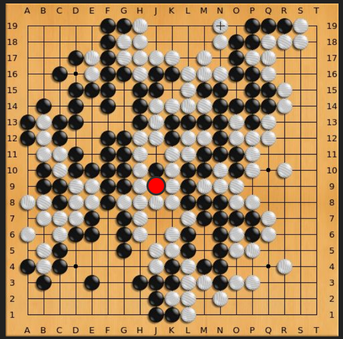


CAT

The orcs' response was a deafening onslaught of claws, claws, and claws; ...







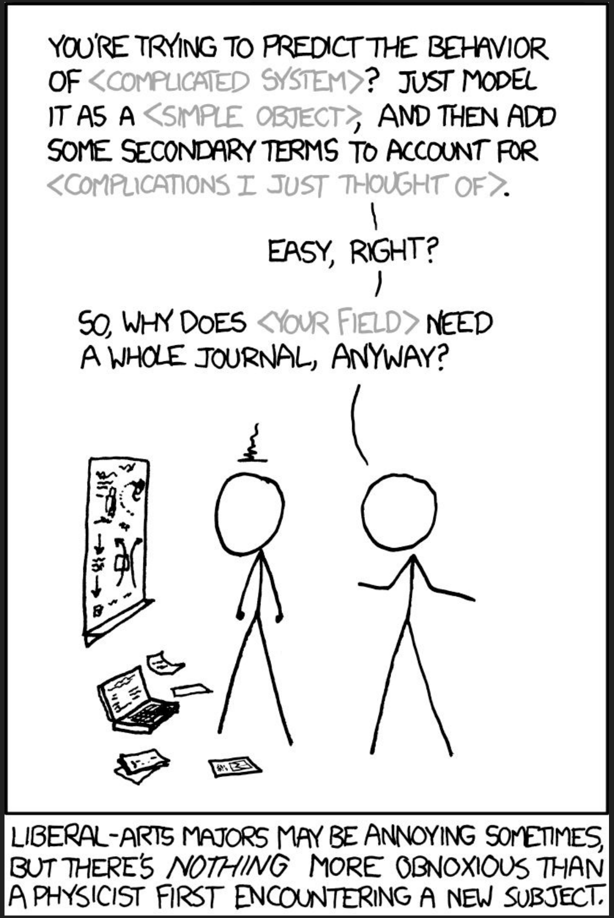
Functional Gradient Descent

Requirements:

1. Want a function $f: X \rightarrow Y$
2. Have some functional to assess quality $L[f]$

Functional Gradient Descent:

1. Start with f_0
2. Measure $\delta L / \delta f$
3. Update: $f_1 = f_0 - \alpha(\delta L / \delta f)$
4. Repeat. A lot.



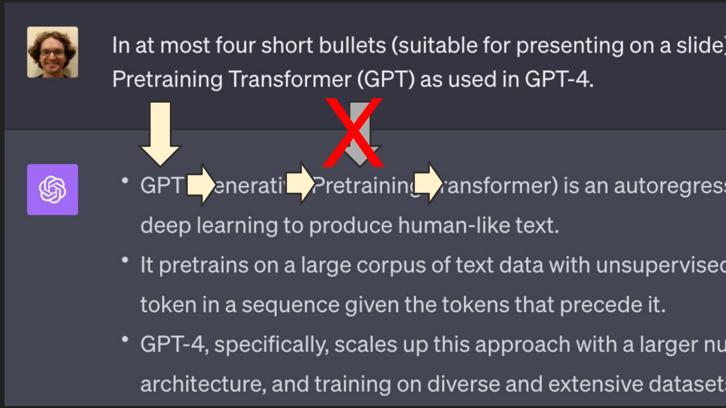
Language Modeling

Requirements for Machine Learning:

1. Want a function f: X→Y
2. Have some functional to assess quality L[f]

Autoregressive language model:
f: text so far → probability distribution of next word

Quality: Distance between model distribution and true distribution. (KL Divergence.)



Estimating KL Divergence

$$KL(p, q) = E_p\left[\log\left(\frac{p}{q}\right)\right]$$

True Distribution

Model Distribution

The Internet

Estimate KL by:

1. Sample from true distribution
2. Measure log(model dist)
3. Average over samples

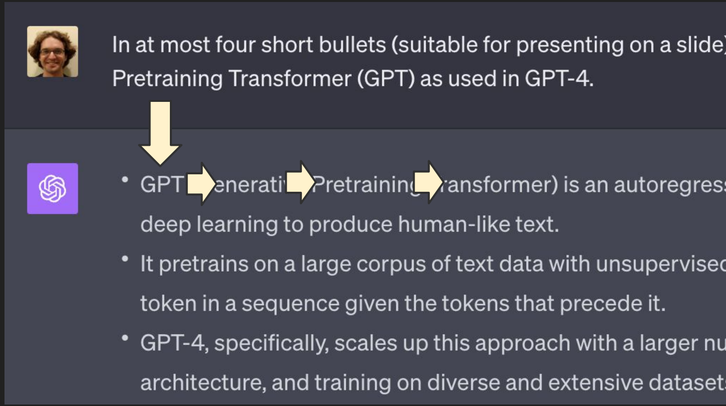
All we need = huge number of independent samples from the distribution of all human text

Luckily, humans have been compiling exactly this 1983

Language Modeling

Requirements for Machine Learning:

1. Want a function f: X→Y
2. Have some functional to assess quality L[f]



Autoregressive language model:
f: text so far → probability distribution of next word

Quality: Distance between model distribution and true distribution. (KL Divergence.)

Unbiased estimator of L:

1. Sample docs from internet
2. Calculate log(q) for each word
3. Average

Bonus Slides Section:
The Alignment Problem

The Alignment Problem: Two Critical Tasks

New Technology Dangers

- Job loss / displacement
- Misinformation
- Cybersecurity
- Exacerbating inequality
- Misuse / abuse
- ...

AI-specific Existential Dangers

- Everyone dies
- Humans lose control of the future

We must solve both!

Remember the LHC black holes?

Only failed analogy in the talk.

Literal existential danger from AI is a serious concern.

The New York Times

ESSAY

Gauging a Collider's Odds of Cr

a Black Hole

Share full article

By Dennis Overbye

April 15, 2008

In Walker Percy's "Love in the Ruins," the protagonist, a doctor and an inventor, recites what he calls the scientist's prayer. It goes like this:

"Lord, grant that my work increase knowledge and help other men.

"Failing that, Lord, grant that it will not lead to man's destruction.

"Failing that, Lord, grant that my article in Brain be published before the destruction takes place."

We have never lived in a setting where human preferences, collectively, are not the strongest optimization force in the world.

Timescales

Each timescale has its own appropriate description.

Slower timescales:

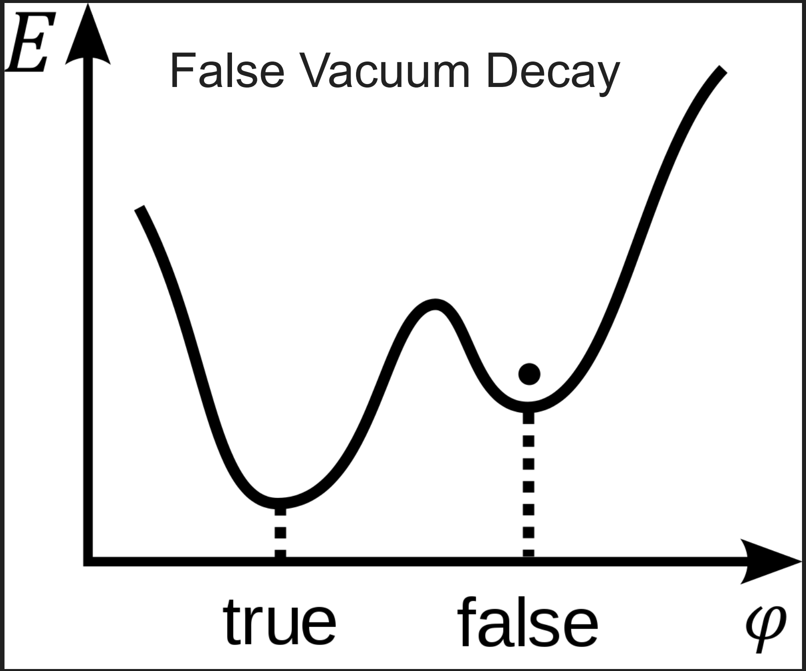
- approximately frozen

Faster timescales:

- [in equilibrium] averaged
- [out of equilibrium] ???

What if faster timescales are out of equilibrium?

TODO: Picture of bird in tree being cut down



What does it mean to be out of equilibrium?

Optimization force
~~Out-of-equilibrium system~~: system that tends towards microstates with certain properties over microstates with opposite properties

Powerful Optimizers in Our World

System	Timescale	"Goal"
Thermodynamics	10^{10} years	Entropy
Evolution	10^{7} years	Fitness
Humanity	10^{1} years	Flourishing (?)

Evolution

I see that I have to increase entropy?

Thermodynamics

Yes. No matter what.

Evolution

Ok but what's this term about ΔE?

Thermodynamics

Open systems can't last forever.
The universe is a closed system.

Evolution

I found a planet with a giant beam of light shining on it for billions of years.

Thermodynamics

Sure, I guess ...

High fitness
Low entropy

Humanity is the best rules lawyer in the universe

A **rules lawyer** is a participant in a rules-based environment who attempts to use the letter of the law without reference to the spirit, usually in order to gain an advantage within that environment.[1] The term is commonly used in wargaming and tabletop role-playing game communities,[2] often pejoratively, as

Humanity's recipe for success:

1. Decide what we want.
2. Prove it's impossible.
3. Deeply understand the proof
4. Use a loophole in the proof to achieve our goal anyway

Humanity

I see on page 2 of The Rules that I have to increase entropy?

Thermodynamics

Yes. No matter what.

Humanity

Also, by the way, I notice on page 497 that it says Mr Evolution left a bunch of lumps of black goop deep underground.

Thermodynamics

True. It would be hard to get them.

Humanity

But in principle, we could?

Thermodynamics

Yes. And then what?

Humanity

We're allowed to set the black goop on fire, right?

Thermodynamics

Uh, I guess so...

High flourishing
Low entropy

Humanity

The Rules say that large chunks of metal fall down, right?

Gravity

Yes....

Humanity

Also, by the way, I notice on page 233 this stuff about air pressure.

Gravity

Yes....

Humanity

So imagine if, just hypothetically, I had a large piece of metal that happened to be slanted just right and travelling super fast?

Gravity

Uh, I guess so...

Humanity

People die a lot of infected wounds

Evolution

Yeah, well, bacteria have you beat pretty well there on genetic fitness. Also I don't really care because wounded people are less likely to raise healthy children

Humanity

Separate question: we're allowed to eat mold, right?

Evolution

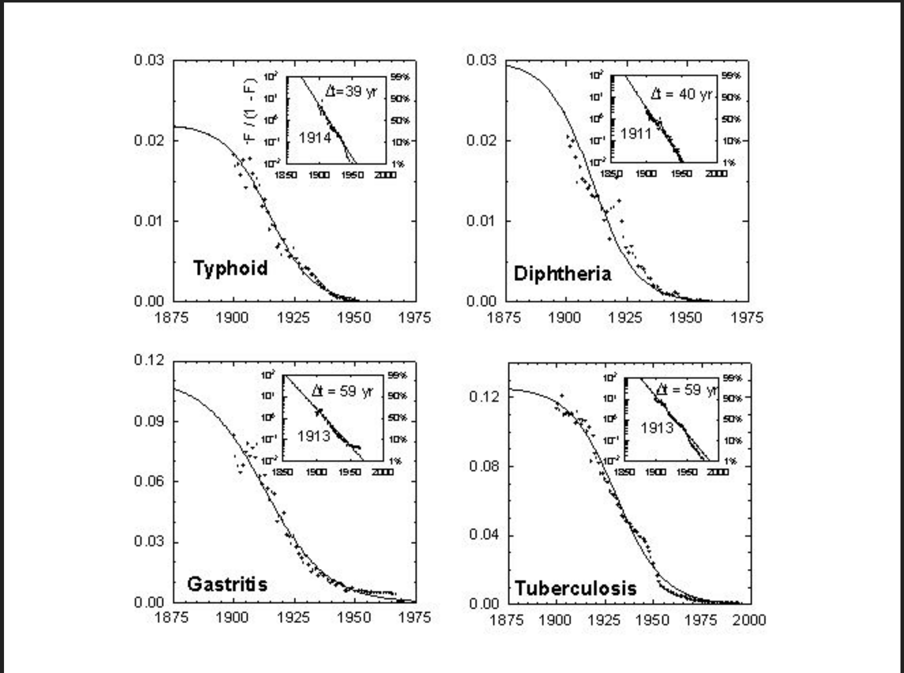
Yes. Almost all mold is pretty bad for you.

Humanity

Almost all, hmm. It says here on page 5007 of the setting description that if I go to this one old brick building in scotland, a rare kind of mold grows there named "penicillin".

Evolution

Uh, I guess so...



https://phe.rockefeller.edu/publication/death-human-environment-united-states-20th-century/

Powerful Optimizers in Our World

More inevitable

System	**Timescale**	**"Goal"**
Thermodynamics	10^{10} years	Entropy
Evolution	10^{7} years	Fitness
Humanity	10^{1} years	Flourishing (?)
AGI	10^{-1} years	???

Better rules-lawyer

What happens if we are no longer the shortest-timescale out-of-equilibrium force in our world?

Bonus Slides Section: The Dota 2 Project

The plan (2018 edition)

1. Teach agents to act/plan/learn in small environments
    a. Board games

2. Move to bigger environments
    a. Video games
    b. Small robotic environments

You Are Here

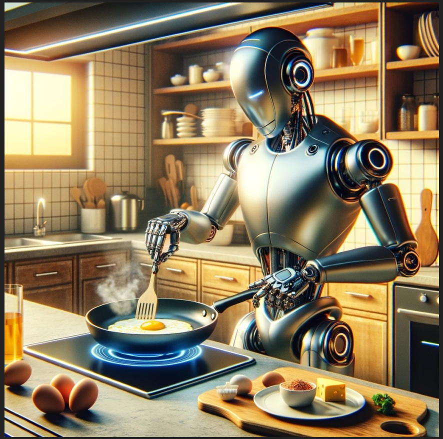
3. Move to the real world

Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm

David Silver,[1*] Thomas Hubert,[1*] Julian Schrittwieser,[1*] Ioannis Antonoglou,[1] Matthew Lai,[1] Arthur Guez,[1] Marc Lanctot,[1] Laurent Sifre,[1] Dharshan Kumaran,[1] Thore Graepel,[1] Timothy Lillicrap,[1] Karen Simonyan,[1] Demis Hassabis[1]

[1]DeepMind, 6 Pancras Square, London N1C 4AG.

AlphaZero (Deepmind, 2017)



Video Games

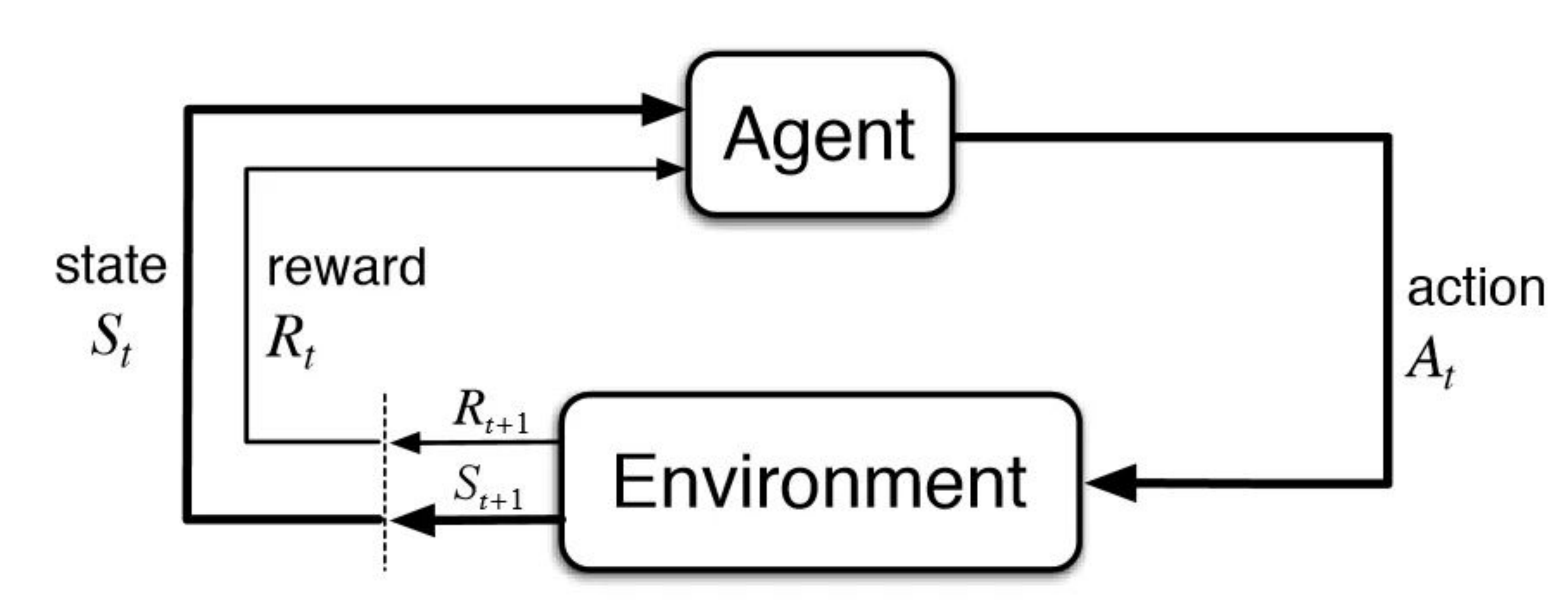
Halfway between board games and real world.

- Partial information about environment
- Continuous observation and action spaces
- Long time-horizon planning

Why Dota 2?

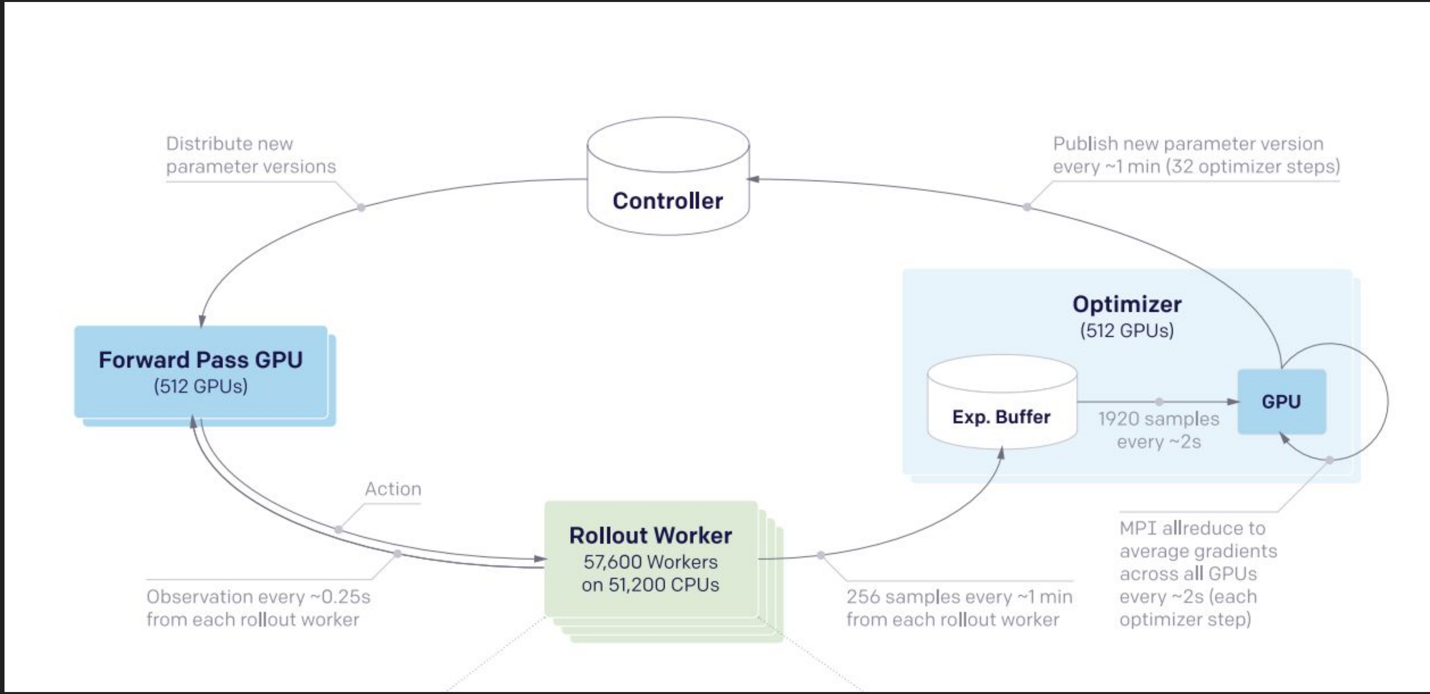
- Large player base
- High-stakes e-sports
- Runs on Linux

Reinforcement Learning

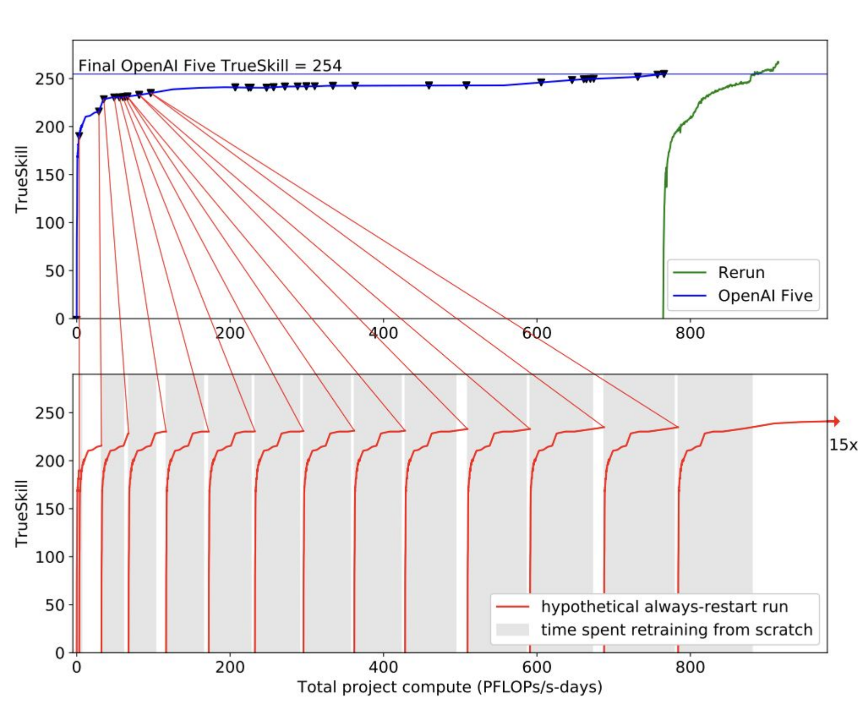


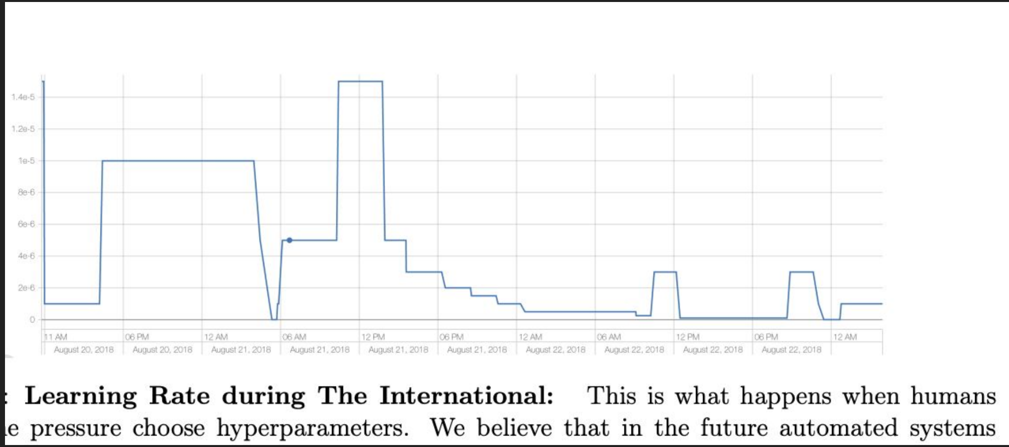
https://towardsdatascience.com/reinforcement-learning-101-e24b50e1d292

Large Scale Distributed Reinforcement Learning



Alchemy and guesswork





OpenAI Five (OpenAI, 2019; https://arxiv.org/abs/1912.06680)