AI in 2023

Where are we and where are we going?
The challenge for talks about Artificial Intelligence

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

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

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 $\beta_2$ hyperparameter in my AdamW optimizer to for optimal stability?

Are we all gonna die?

I haven't been keeping abreast of this weird new field, what's going on?
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 can't fathom how dramatically GPT-6 will change the world.

I'm using ML for my physics research. What should I set the $\beta_2$ hyperparameter in my AdamW optimizer to for optimal stability?

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

Are we all gonna die?
My trajectory so far

2012
- Phd in High Energy Theory

2014
- Solve deep, interesting problems

2016
- Fast paced, world-changing impact

2018
- Software Engineering on Google Search

2020
- AI Research at OpenAI (OpenAI Five, GPT-4, ...)

2022
The challenge for talks about Artificial Intelligence

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

Are we all gonna die?

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

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

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 $\beta_2$ hyperparameter in my AdamW optimizer to for optimal stability?

Are we all gonna die?
You are not alone
Learning Rate of OpenAI Five over one weekend

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

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

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

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

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

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!

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

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

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

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

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!

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

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

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

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

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!

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

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

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

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

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!

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

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

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

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

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!

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

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

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

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

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!

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

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

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

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

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!

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

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

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

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

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!

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

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

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

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

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!

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

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

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

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

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!

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

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

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

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

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!

Let's just go back to before we started screwing with it.
The challenge for talks about Artificial Intelligence

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?

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

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 $\beta_2$ 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

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

I'm using ML for my physics research. What should I set the $\beta_2$ hyperparameter in my AdamW optimizer to for optimal stability?

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

Are we all gonna die?
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

- **1970**: SPEAR
- **1980**: ISR, PEP
- **1990**: SPS, LEP
- **2000**: Tevatron
- **2010**: LHC
- **2020**: RHIC
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... Training a model on $10^{23}$ flops is not easy...

But it's also not sufficient. But it's also not sufficient.

Need to verify we get scientific results from experiments. 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:
\[ f: \text{text so far} \rightarrow \text{probability distribution of next word} \]
Deep Learning (Spherical Cow View)

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

GPT-4

Giant blob of parameters $\theta$

$f(\theta; T)$

Previous text $T$

Gradient Descent

1. Initialize randomly
2. Sample some data from the internet
3. Measure $L =$ distance between GPT-4 output distribution and data distribution
4. Calculate gradient $g = \frac{dL}{d\theta}$
5. Update $\theta$ proportional to $g$
6. Repeat
But where is the magic?
(Answer #1)
The Complete works of Shakespeare

Spambots

Arxiv

Crackpot physics blogs

Public textbooks

Bad fanfiction

Good fanfiction

 Ads for tractors

Model the Average = Bad

Model the Distribution = ???
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 ====
...
A novel sonnet of Shakespeare, discovered in 2024, is likely authentic, experts say. Without further ado:

*Sonnet 155*

...
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

Source has many additional test results
Part 4: GPT-4 demos
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 \(-a \leq x \leq a\) and \(-b \leq y \leq b\). 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 \(d^2 H / dv^2 = k^2 H\), your later experience should lead you to guess \(H = \exp(\pm kv)\).
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?

Answer Key: $-1.09 \times 10^3 \text{ J}$
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?

\[ W = \vec{F} \cdot \vec{d} \]

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 \( W_f = \vec{F}_{friction} \cdot \vec{d} \). The friction force \( \vec{F}_{friction} \) can be calculated using the formula:

\[ F_{friction} = \mu \cdot F_N \]

where:

- \( \mu \) 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 \( F_{pull, \text{vertical}} = F_{pull} \cdot \sin(\theta) \), where \( \theta \) is the angle the pulling force forms above the horizontal.

The effective normal force is then:

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

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

\[ F_{friction} = \mu \cdot F_{eff} \]

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

\[ W_f = F_{friction} \cdot d \cdot \cos(180°) \]

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 \(-1086.1 \, \text{J}\). The negative sign indicates that the work is done against the direction of the displacement.
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 $\mu k = 0.22$.

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:
   \[ PE_{initial} = m \cdot g \cdot h \]

2. **Initial Kinetic Energy**: The kinetic energy at the top due to its speed $v$ is:
   \[ KE_{initial} = \frac{1}{2} m \cdot v^2 \]

3. **Work done by Friction**: As the roller coaster descends the hill, friction does negative work on it:
   \[ W_{friction} = -\mu_k \cdot m \cdot g \cdot d \cdot \cos(\theta) \]
   
   where:
   
   * $\mu_k = 0.22$ is the coefficient of kinetic friction,
   * $d = 45.0$ 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:

   \[ KE_{final} = PE_{initial} + KE_{initial} + W_{friction} \]

   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 J (or approximately 28.9 kJ).✅
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:

   \[ P E_{\text{initial}} = m \cdot g \cdot h \]

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

   \[ K E_{\text{bottom}} = \frac{1}{2} m \cdot v^2 \]

   where \( v \) is the velocity of the block at the bottom of the ramp.

By conservation of energy:

\[ P E_{\text{initial}} = K E_{\text{bottom}} \]

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

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:

   - \( \mu \) is the coefficient of friction,
   - \( d = 7.20 \) m is the distance the block slides on the horizontal surface.

   Equating the two expressions for \( W_{\text{friction}} \), we get:

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

   From this equation, we can solve for \( \mu \).

   Let’s plug in the given values and calculate \( \mu \).

The coefficient of friction between the block and the horizontal surface is \( \mu = 0.333 \) or \( 0.63 \).
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✓</td>
</tr>
<tr>
<td>2</td>
<td>✓</td>
</tr>
<tr>
<td>3</td>
<td>✓</td>
</tr>
<tr>
<td>4</td>
<td>✓</td>
</tr>
<tr>
<td>5</td>
<td>✓</td>
</tr>
<tr>
<td>6</td>
<td>✓</td>
</tr>
<tr>
<td>7</td>
<td>✓</td>
</tr>
<tr>
<td>8</td>
<td>✓</td>
</tr>
<tr>
<td>9</td>
<td>✓</td>
</tr>
<tr>
<td>10</td>
<td>❌</td>
</tr>
</tbody>
</table>

(N=1)
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 frictional force $W_f$.

$W = W_p + W_g + W_f$

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

$W_p = F_p \cdot d \cdot \cos(0°)$

where:

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

The work done by the gravitational force is:

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

where:

- $m = 5.60$ kg is the mass of the box,
- $g = 9.81$ m/s² is the acceleration due to gravity,
- $θ = 16.0°$ is the angle of the incline.

The work done by the frictional force is:

$W_f = -μ_k \cdot m \cdot g \cdot \cos(θ) \cdot d$

where:

- $μ_k = 0.180$ 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:

$ΔKE + ΔPE = \frac{1}{2}mv_f^2 - \frac{1}{2}mv_i^2 + mgh = 0$

where:

- $v_i = 0$ m/s is the initial velocity,
- $v_f = 5.40$ m/s is the final velocity,
- $h = d \sin(θ)$ is the height of the incline.

Now, according to the work-energy principle:

$W = ΔKE + ΔPE$

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

$F_p \cdot d + (-m \cdot g \cdot d \cdot \sin(θ)) + (-μ_k \cdot m \cdot g \cdot \cos(θ) \cdot d) = \frac{1}{2}mv_f^2 - \frac{1}{2}mv_i^2 + mgh$

Now we can rearrange this equation to solve for $F_p$:

$F_p = \frac{(\frac{1}{2}mv_f^2 - \frac{1}{2}mv_i^2 + mgh) + (m \cdot g \cdot d \cdot \sin(θ)) + (μ_k \cdot m \cdot g \cdot \cos(θ) \cdot d)}{d}$

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

$F_p = \text{approximately 42.19 N}$

The pulling force $F_p$ is approximately 42.19 N.

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

$W_p = F_p \cdot d = \text{approximately 1434.52 J}$

The work done by the pulling force $W_p$ is approximately 1434.52 J.
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.
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.

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: $8L/3$
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.
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 \( \theta \) 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 \( \theta \) 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 \( \theta \).

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

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

\[
x = R(1 - \cos(\theta))\quad [\text{Ans}]
\]

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.\([\text{Ans}]\)
You

At \( t = 0 \), an elevator leaves the ground with uniform speed. At time \( t_1 \), 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 \( t_2 \). At what height, \( h \), was the elevator when the child dropped the coin?
Honors Physics 1 (prob 3)

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 $\Delta h$, the height of the elevator at $t_1$.

Substituting the values, we get:
\[ \Delta h = v(t_2 - t_1) - \frac{1}{2}g(t_2 - t_1)^2 \]

Now, we need to solve for $\Delta h$.

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 $t_1$, $t_2$, 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:
\[ h = vt_1 \quad \text{(Equation 1)} \]

From the coin’s motion, we have:
\[ h = v(t_2 - t_1) - \frac{1}{2}g(t_2 - t_1)^2 \quad \text{(Equation 2)} \]

Now, we can solve these two equations simultaneously to eliminate $v$ and solve for $\Delta h$ 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 $g$. [Equation]
V(x, y) obeys Laplace's equation inside the rectangular region of the 2D plane defined by \(-a \leq x \leq a\) and \(-b \leq y \leq b\). 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 \(d^2H/du^2 = k^2H\), 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 385 (prob 1)

You

V(0, y) obeys Laplace's equation inside the rectangular region of the 2D plane defined by -a < x < a and 0 < y < b. The potential is zero on the horizontal boundaries at x = -a and x = a, and on the top and bottom boundaries it is given by V(x, b) = V0 sin(ka/b) and V(x, -b) = -V0 sin(ka/b), respectively. Find V(x, y) inside the rectangle.

Hint: When you encounter a differential equation like \(d^2w/dx^2 = k^2w\), your later algebra will be simplified if you take a solution in the form H = A \cos(ku) + B \sin(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 will 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:

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

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

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

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

\[
\frac{1}{X(x)} \frac{d^2 X}{dx^2} = -k^2
\]

\[
\frac{1}{Y(y)} \frac{d^2 Y}{dy^2} = k^2
\]

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

\[
\frac{d^2 X}{dx^2} + k^2 X = 0
\]

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, y) can be rewritten in terms of a sum of sines:

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

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

<table>
<thead>
<tr>
<th>Physics 203</th>
<th>Honors Physics I</th>
<th>Phys 385</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✅</td>
<td>(0/5)</td>
</tr>
<tr>
<td>2</td>
<td>✅</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>✅</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>✅</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>✅</td>
<td>(0/5)</td>
</tr>
<tr>
<td>6</td>
<td>✅</td>
<td>(1/5)</td>
</tr>
<tr>
<td>7</td>
<td>✅</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>✅</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>✅</td>
<td>(0/5)</td>
</tr>
<tr>
<td>10</td>
<td>❌</td>
<td>[N=1]</td>
</tr>
</tbody>
</table>

- **GPT4 ~ mediocre student**

*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:

\[
\Sigma(p) = (-ig)^2 \int \frac{d^4k}{(2\pi)^4} \frac{i(k + m)i\gamma^5}{k^2 - m^2 + i\epsilon} \frac{i}{(p - k)^2 - M^2 + i\epsilon},
\]

where \(\not{k}\) denotes the Feynman slash notation \(\gamma^\mu k_\mu\), and \(\gamma^5\) 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 \(\epsilon\).

Step 3: Perform the integral

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.

Step 5: Anomalous dimension

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

\[
\gamma_m(g) = \left. -\frac{\mu}{m} \frac{dm}{d\mu} \right|_g,
\]

where \(\mu\) is the renormalization scale. The \(\beta\) function and \(\gamma_m\) 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 \(\mu\). The solution to these equations gives us the energy dependence of...
Remind: 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 had thrown a knife and one of the wounds was...
Reminder: 2012

AlexNet achieves 84.7% accuracy on image classification
“It missed the factor of $\cos(\theta)$.”

“Can't even do RG calculations without Mathematica!”
GPT-4 Capabilities Summary

True, But Not The Key Point
GPT-4 is pretty cool, does well on high school exams

The Key Point
\( \frac{d(\text{intelligence})}{d(\text{years})} \) 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
tenighan@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

Scaling Laws

- **Compute (PF-days, non-embedding):**
  \[ L = \left( C_{\text{min}} / 2.3 \cdot 10^8 \right)^{-0.050} \]

- **Dataset Size (tokens):**
  \[ L = \left( D / 5.4 \cdot 10^{13} \right)^{-0.095} \]

- **Parameters (non-embedding):**
  \[ L = \left( N / 8.8 \cdot 10^{13} \right)^{-0.076} \]
Scaling Laws: Why is this hard?

- Models
- Controls

![Graph showing loss vs. compute with labeled data points: True underlying law and If you miss the wrong dial]
Chinchilla (2023)

Original Scaling Laws had an important dial set wrong

<table>
<thead>
<tr>
<th></th>
<th>Compute used</th>
<th>model size</th>
<th>dataset size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>X</td>
<td>X</td>
<td>1</td>
</tr>
<tr>
<td>Chinchilla</td>
<td>X</td>
<td>√X</td>
<td>√X</td>
</tr>
</tbody>
</table>

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

OpenAI codebase next word prediction

Bits per word

Observed
Prediction

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

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

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

Intelligence = \frac{\text{Investment}}{\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?

OpenAI codebase next word prediction

Bits per word

- Observed
- Prediction
- gpt-4

Compute

1.0  100p  10n  1µ  100µ  0.01  1
Bonus Slides Section: Other / mishmash
But where is the magic?

(Answer #2)

Emergent theories at large $N \rightarrow \infty$ are nothing new. Intelligence is no exception.

$P = mv$

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

$f(\theta; T)$

Previous text $T$
ChatGPT is going to change the world!

Let's not get ahead of ourselves...

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

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

Moore's Law: The number of transistors on microchips has doubled every two years. Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important for other aspects of technological progress in computing—such as processing speed or the price of computers.

Source: Our World in Data

Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)
OurWorldInData.org - Research and data to make progress against the world’s largest problems. Licensed under CC BY by the authors Hannah Ritchie and Max Roser.
Another reference for flops vs time

1 Latest GPU * 1 day
The future is going to be weird

https://www.openphilanthropy.org/research/modeling-the-human-trajectory
Bonus Slides Section:
Machine Learning For Physicists 101
Goal = find local min

\[ \Delta \theta = \frac{dL}{d\theta} \]
**Functional Gradient Descent**

Requirements:

1. Want a function $f : X \to Y$
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.
Key Insight: EVERYTHING is functional gradient descent

Requirements:

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

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

**Autoregressive language model:**
$f$: text so far $\rightarrow$ 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

Estimate KL by:
1. Sample from true distribution
2. Measure \( \log(\text{model dist}) \)
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

The Internet
Language Modeling

Requirements for Machine Learning:

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

**Autoregressive language model:**

$f$: text so far $\rightarrow$ 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?

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

Only failed analogy in the talk.

Literal existential danger from AI is a serious concern.

Gauging a Collider’s Odds of Creating a Black Hole

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.”
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?

Out-of-equilibrium system: system that tends towards microstates with certain properties over microstates with opposite properties.

Optimization force
### Powerful Optimizers in Our World

<table>
<thead>
<tr>
<th>System</th>
<th>Timescale</th>
<th>&quot;Goal&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thermodynamics</td>
<td>$10^{10}$ years</td>
<td>Entropy</td>
</tr>
<tr>
<td>Evolution</td>
<td>$10^7$ years</td>
<td>Fitness</td>
</tr>
<tr>
<td>Humanity</td>
<td>$10^1$ years</td>
<td>Flourishing (?)</td>
</tr>
</tbody>
</table>
I see that I have to increase entropy?

Evolution

Yes. No matter what.

Thermodynamics

Ok but what's this term about $\Delta E$?

Evolution

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

Thermodynamics

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

Evolution

Sure, I guess ...

Thermodynamics

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

Yes. No matter what.

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

True. It would be hard to get them.

But in principle, we could?

Yes. And then what?

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

Uh, I guess so...
The Rules say that large chunks of metal fall down, right?

Yes....

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

Yes....

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

Uh, I guess so....
People die a lot of infected wounds

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.

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

Yes. Almost all mold is pretty bad for you.

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".

Uh, I guess so...

### Powerful Optimizers in Our World

<table>
<thead>
<tr>
<th>System</th>
<th>Timescale</th>
<th>&quot;Goal&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thermodynamics</td>
<td>$10^{10}$ years</td>
<td>Entropy</td>
</tr>
<tr>
<td>Evolution</td>
<td>$10^7$ years</td>
<td>Fitness</td>
</tr>
<tr>
<td>Humanity</td>
<td>$10^1$ years</td>
<td>Flourishing (?)</td>
</tr>
<tr>
<td>AGI</td>
<td>$10^{-1}$ years</td>
<td>???</td>
</tr>
</tbody>
</table>

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

3. Move to the real world
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

Learning Rate during The International: This is what happens when humans pressure choose hyperparameters. We believe that in the future automated systems