

NATURAL LANGUAGE PROCESSING

GREG DURRETT

INSTITUTE FOR FOUNDATIONS OF MACHINE LEARNING

Intro to ChatGPT and Natural Language Processing



Institute for Foundations of
MACHINE LEARNING



MARKETS BUSINESS INVESTING TECH POLITICS CNBC TV INVESTING CLUB PRO

TECH

OpenAI announces GPT-4, claims it can beat 90% of humans on the SAT

PUBLISHED TUE, MAR 14 2023 1:42 PM EDT | UPDATED TUE, MAR 14 2023 2:32 PM EDT



Kif Leswing
@KIFLESWING

GPT-4



MOTHERBOARD
TECH BY VICE

The New GPT-4 AI Gets Top Marks in Law, Medical Exams, OpenAI Claims

The successor to GPT-3 could get into top universities without having trained on the exams, according to OpenAI.



By [Samantha Cole](#)

Five years ago, you could do this:

The screenshot shows the Google Translate web interface. At the top, it says "Google Translate". Below that is a navigation bar with four tabs: "Text" (selected), "Images", "Documents", and "Websites". Underneath the tabs, there are language selection dropdowns: "Detect language" (grayed out), "Spanish", "French", "English" (selected), and "English" (grayed out) followed by a dropdown menu with "Spanish" (selected) and "Arabic". The main area contains two text boxes. The left box contains the English sentence "I'd like the Caesar salad, please." with a microphone icon and a sound volume icon below it. The right box contains the translated Spanish sentence "Me gustaría la ensalada César, por favor." with a star icon and a sound volume icon below it. At the bottom of the interface, there are icons for "Send feedback", a keyboard, and a character counter "34 / 5,000".

Five years ago, you could do this:



Now, you can do this:

GD

I'm supposed to come up with a topic for my European history essay. I have to write about one of the Tudor monarchs. Who would be three interesting monarchs to write about?

Now, you can do this:

GD

I'm supposed to come up with a topic for my European history essay. I have to write about one of the Tudor monarchs. Who would be three interesting monarchs to write about?



Sure, the Tudor dynasty, which ruled England from 1485 to 1603, was filled with interesting figures, each with their own unique impact on the course of history. Here are three that might make for a compelling essay:

1. **Henry VII** (Reign: 1485–1509): The founder of the Tudor dynasty. He seized the throne after defeating Richard III at the Battle of Bosworth Field, effectively ending the Wars of the Roses. His reign marked the transition from medieval to early modern England, and his economic, diplomatic, and administrative policies helped stabilize the kingdom after years of civil war. His marriage to Elizabeth of York also symbolically united the houses of York and Lancaster.
2. **Henry VIII** (Reign: 1509–1547): Perhaps the most famous Tudor monarch, Henry VIII is best known for his six marriages, particularly his break with the Roman Catholic Church over the issue of annulment. His reign saw significant religious, political, and social changes in England, including the English Reformation and the establishment of the Church of England.

Now, you can do this:

GD

I'm working on a song about a breakup and I need a word that rhymes with "doubt" that thematically relates to sadness. Can you suggest a few words?

Now, you can do this:

GD

I'm working on a song about a breakup and I need a word that rhymes with "doubt" that thematically relates to sadness. Can you suggest a few words?



Sure, here are a few words that rhyme with "doubt" and thematically relate to sadness:

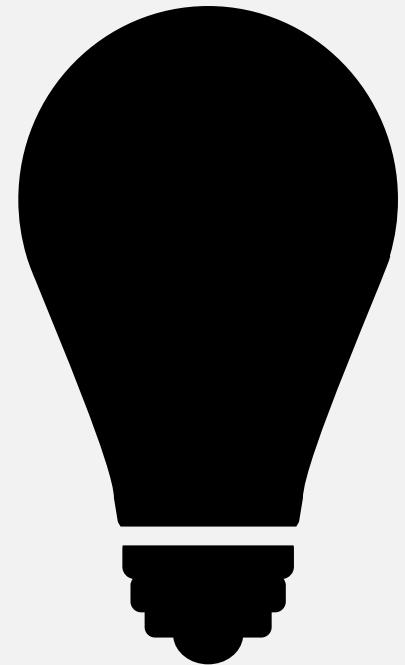
1. "Shout" - This can refer to the anger or frustration that can come with a breakup.
2. "Rout" - This word means a decisive defeat. It could be used metaphorically to describe feeling overwhelmingly defeated after a breakup.
3. "Spout" - This could be used to refer to speaking or crying extensively about the breakup.
4. "Out" - This word could be used in the context of feeling left out or excluded after a breakup.
5. "Scout" - This word could be used to symbolize looking or searching for love or happiness after a breakup.
6. "Clout" - This could indicate the power or influence the other person had in the relationship, causing the sadness post-breakup.



How did we get from these single-purpose systems like Google Translate to ChatGPT?



What's the core technology behind ChatGPT?



We'll learn basic principles of machine learning (ML) and Natural Language Processing (NLP)



Try a few things in ChatGPT and share the results with others!

Try:

- (a) asking it about a fact**
- (b) having it help you brainstorm about something**
- (c) solve a math problem**

NATURAL LANGUAGE PROCESSING

GREG DURRETT
INSTITUTE FOR FOUNDATIONS OF MACHINE LEARNING

Machine Learning



Institute for Foundations of
MACHINE LEARNING

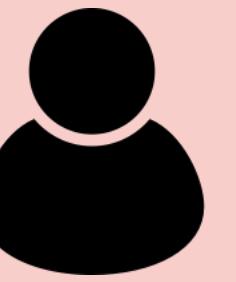
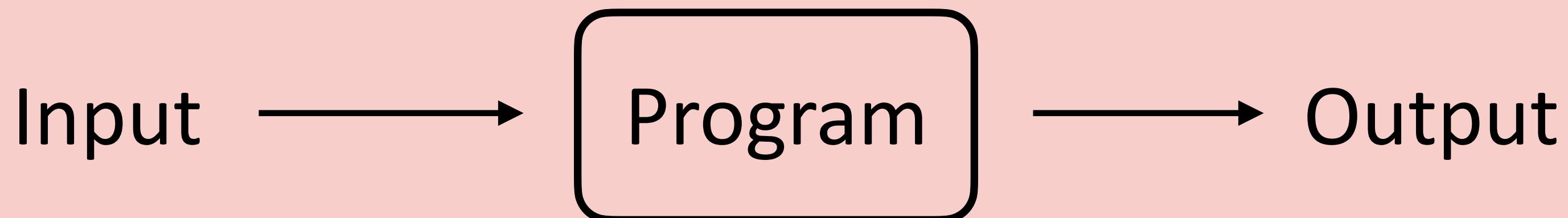
If you're learning to program, you might be doing things like this:

Write a function to uppercase every other letter in a string, starting with the second one.

```
input = "stop staying up so late."  
i = 0  
result = ""  
for letter in input:  
    if i % 2 == 1:  
        result += letter.upper()  
    else:  
        result += letter  
    i += 1
```

→ sToP StAyInG Up sO LaTe.

Programming (as you're learning it)



Programmer

But how do you write a program to do this?

The screenshot shows the Google Translate web interface. At the top left is the "Google Translate" logo. To the right are three icons: a grid of nine dots, a user profile picture of a man with blonde hair, and a circular refresh button. Below the logo are four buttons: "Text" (selected), "Images", "Documents", and "Websites". The source language dropdown shows "English" (selected) and "Detect language", "Spanish", "French", "English" (disabled). The target language dropdown shows "Spanish" (selected) and "English", "Spanish", "Arabic". The main area contains two text boxes. The left box contains the English sentence "I'd like the Caesar salad, please." with a microphone icon, a speaker icon, and a character counter "34 / 5,000". The right box contains the Spanish translation "Me gustaría la ensalada César, por favor." with a speaker icon and sharing/social media icons. A "Send feedback" link is at the bottom right.

I'd like the Caesar salad,
please.

Me gustaría la ensalada César, ★
por favor.

Send feedback

```
words = ["I'd", "like", "the", "Caesar", "salad", "please"]  
result = []  
for word in words:  
    if word == "salad":  
        result.append("ensalada")  
    elif word == "Caesar":  
        result.append("César")  
    elif word == "please":  
        result.append("por favor")  
    else:  
        result.append(word)
```

I'd like the César
ensalada, por favor.



```
elif word == "the":  
    result.append("la")
```

el? Caesar salad ⇒ ensalada César? Me gustaría?

Machine learning can learn these rules from data, without us writing them

I'd like the tacos.

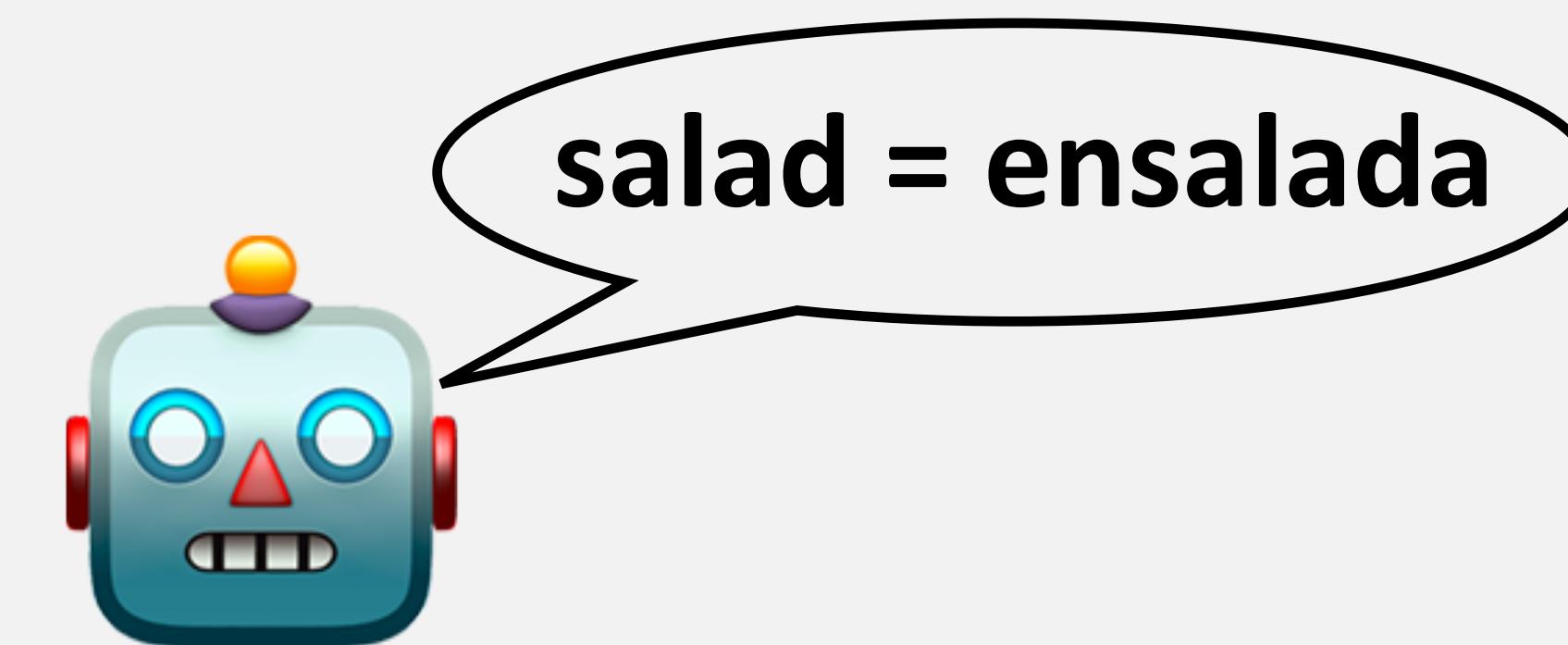
Me gustaría los tacos.

Caesar **salad** contains lettuce.

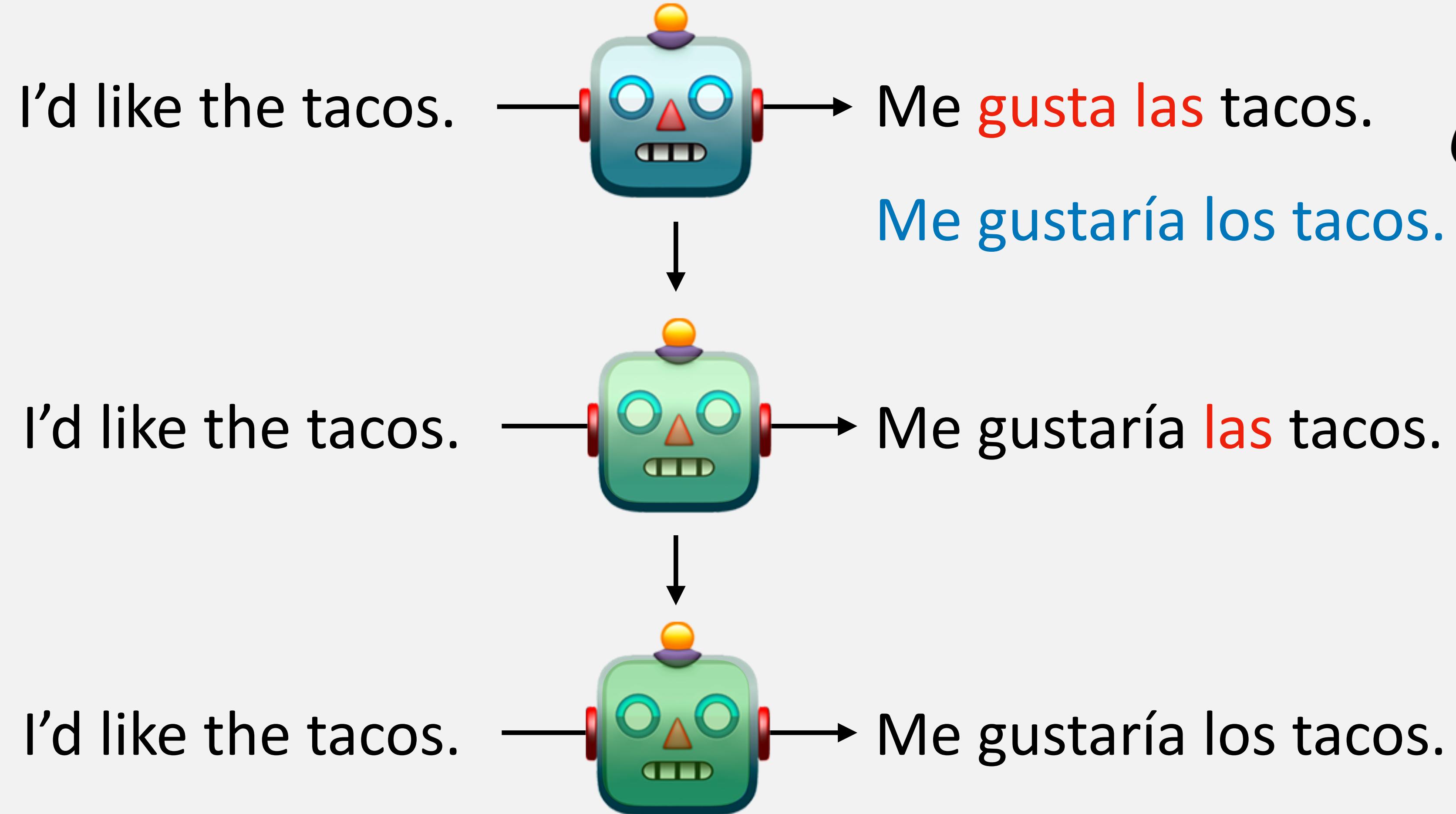
La **ensalada** César contiene lechuga.

I'm eating **salad**.

Estoy comiendo **ensalada**.

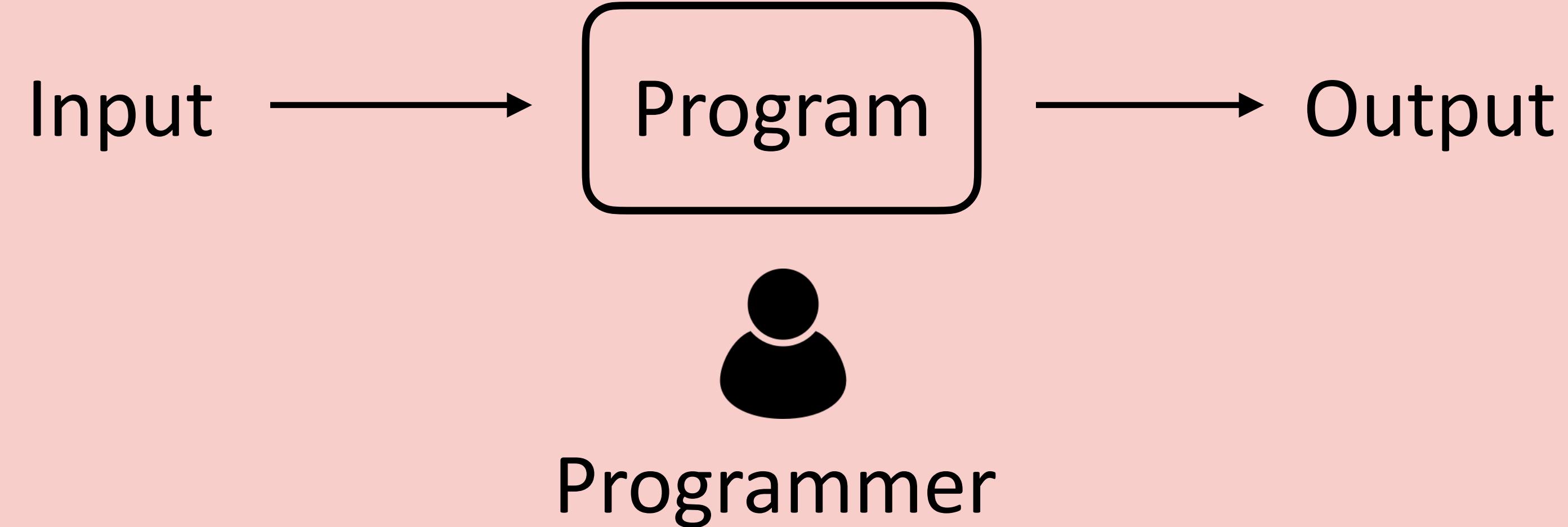


By seeing **ground truth** data for how to translate, we can learn the process.

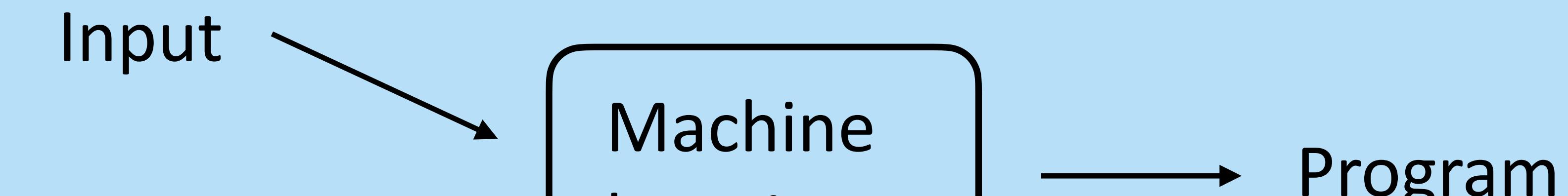


This is called *training*. We repeatedly get predictions from an AI system, then compare them to the right answer. We update it to get them right. We didn't have to write any rules; the system figured it out!

Programming (as you're learning it)



Machine learning



(examples of what
we want to do)

NATURAL LANGUAGE PROCESSING

GREG DURRETT
INSTITUTE FOR FOUNDATIONS OF MACHINE LEARNING

Language Modeling



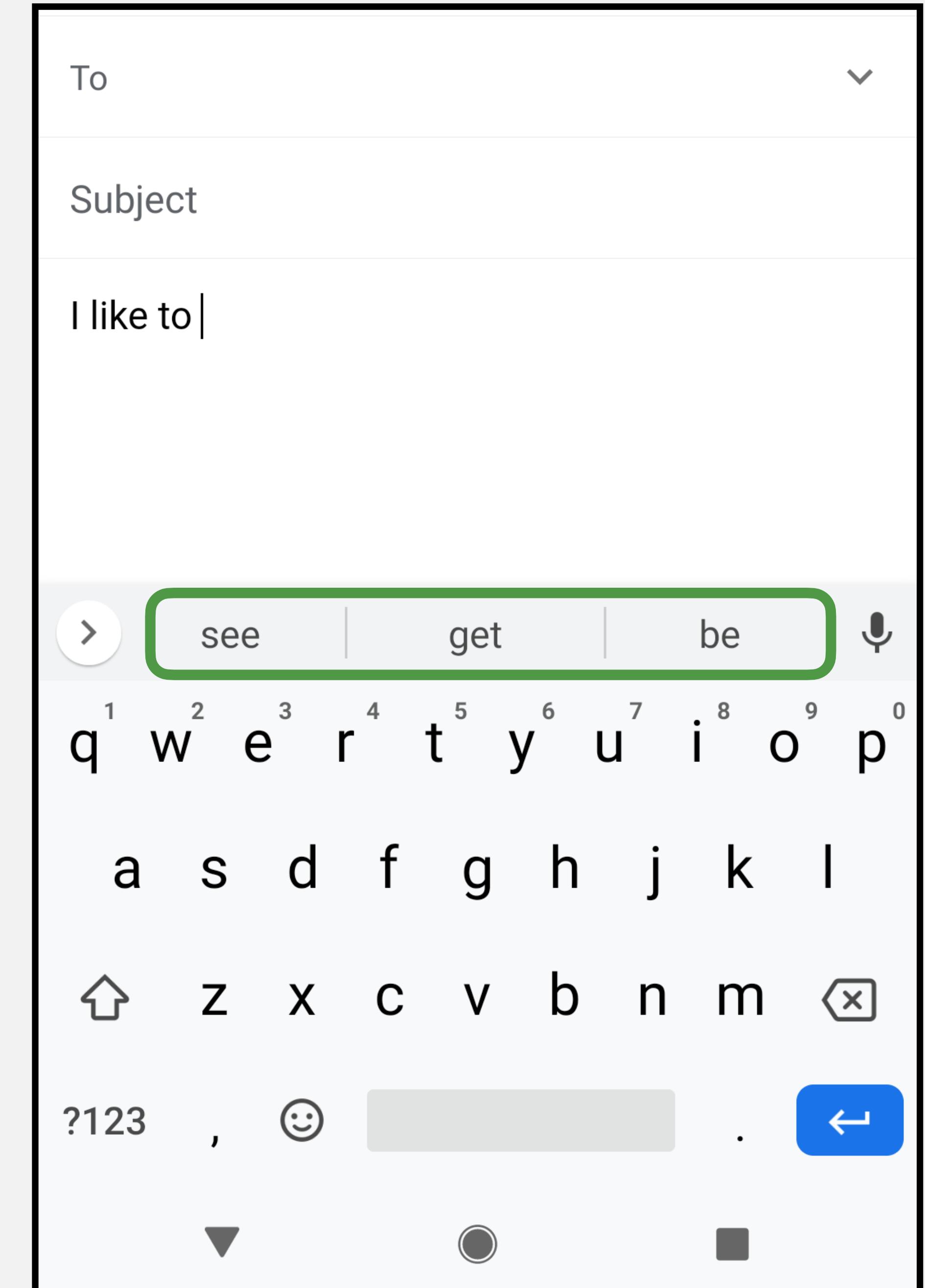
Institute for Foundations of
MACHINE LEARNING



ChatGPT is a language model. It produces a sequence of words given some input. It generates one word at a time.

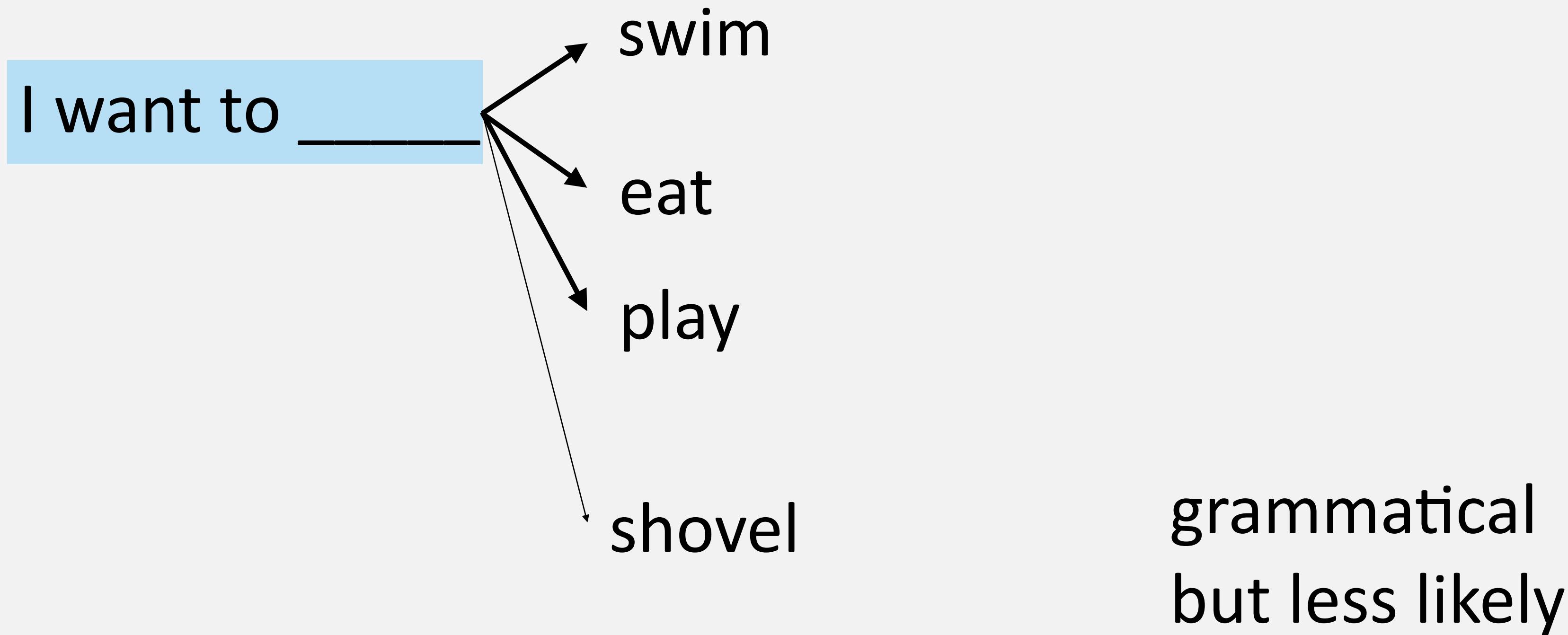
Like translation, this is a task we need machine learning for.

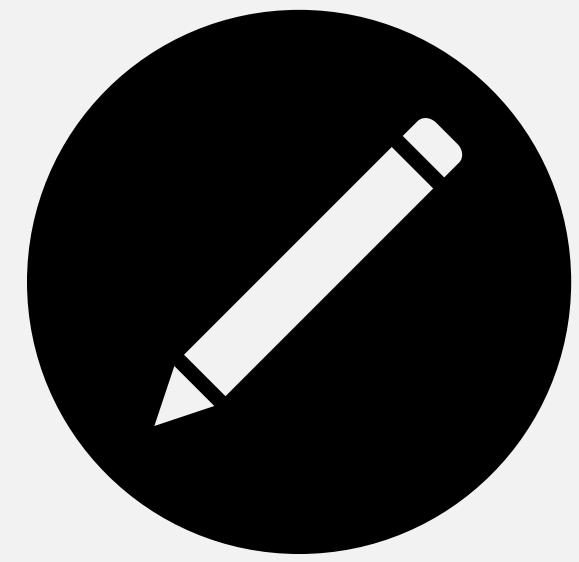
- ▶ Language modeling is the same as predictive text: Given a sequence of words so far (**the context**), predict what comes next, like in predictive text!
- ▶ We never know for sure what comes next, but we can still make good guesses!
- ▶ **We can't write rules to do this. We need machine learning!**





1. Suppose we have the *context* “I want to ____”. Lots of words can come next and form sensible sentences. Think about a few words that can come next; what do these have in common?





2. Can you think of a *context* (a start of a sentence like “I want to _____” ending in a blank) where the next word **has** to be one word in particular for it to be correct?

The capital of Nebraska is _____

Lincoln

one right answer but
may be hard to predict!

The third President of the United States was _____

Thomas
Jefferson

one of the founding fathers
a slave owner

These examples suggests that predicting text is very powerful:

Q: What started the American Revolution? A: _____

Once upon a time, there was a _____

The translation of “I’d like the Caesar salad, please” into Spanish is “_____”

ChatGPT is a “supercharged” language model! Even though it can only do one thing (predictive text), it has way more diverse capabilities than past systems like Google Translate.

NATURAL LANGUAGE PROCESSING

GREG DURRETT

INSTITUTE FOR FOUNDATIONS OF MACHINE LEARNING

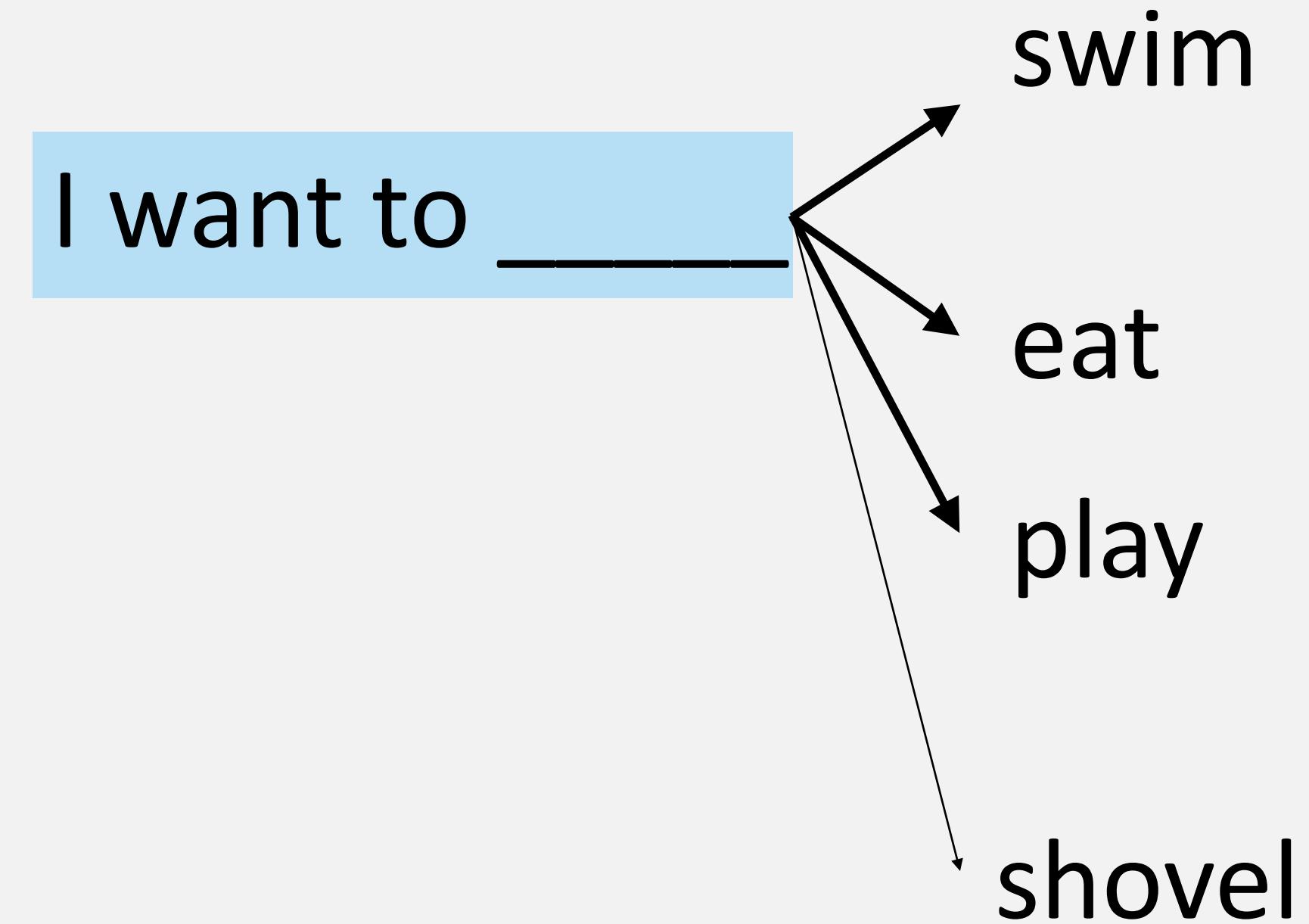
Mathematics of Language Modeling



Institute for Foundations of
MACHINE LEARNING



What does predicting the next word actually mean?



There's never just one right answer. There are a bunch of options, some more likely and some less likely.

Tomorrow, the weather will be _____

sunny

cloudy

rainy

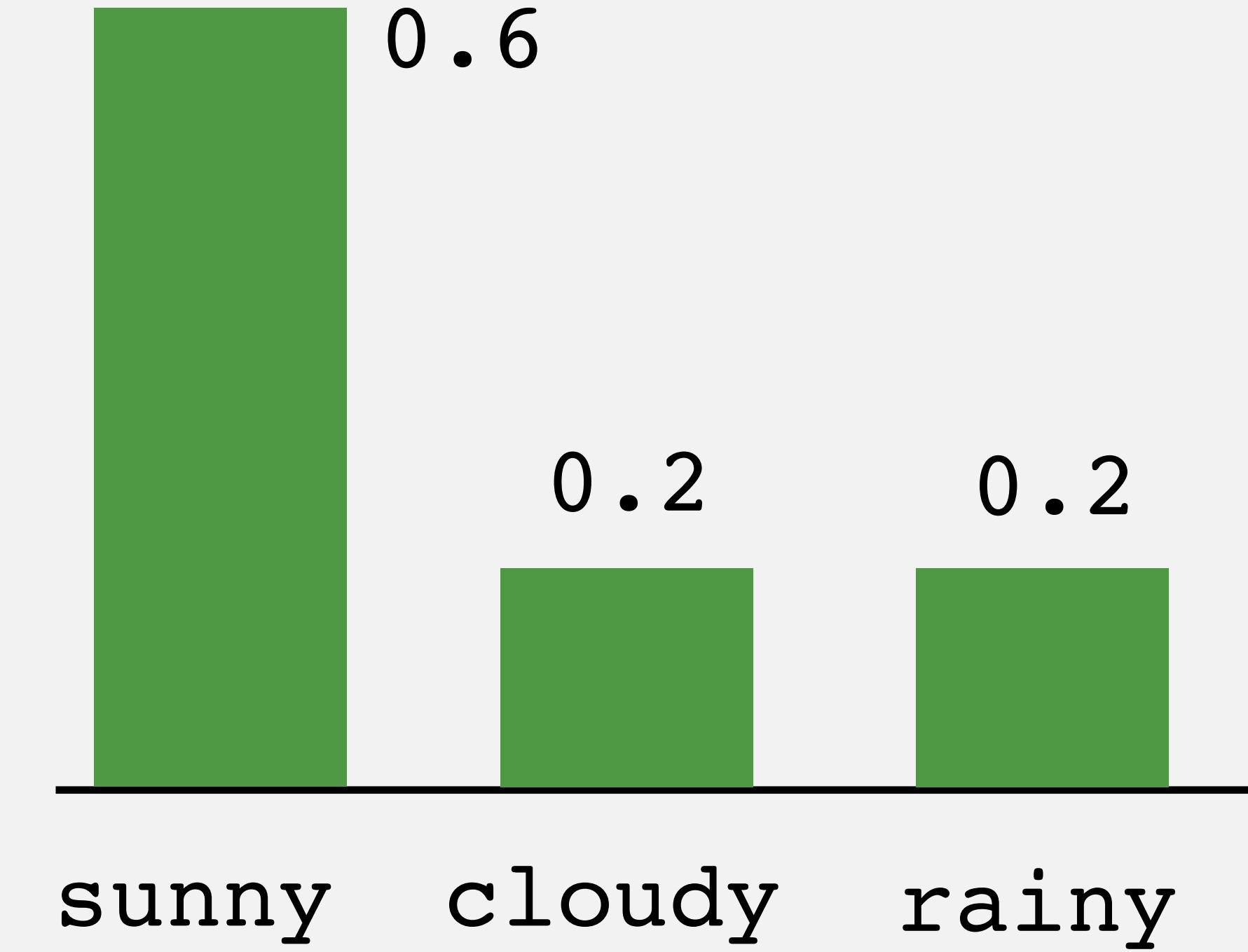
Three options. This is what we call our **vocabulary**.

Language models place a **probability distribution** over the **vocabulary**.

Probability distribution: set of **outcomes**, each associated with a **probability**.

Flipping a coin: two outcomes (heads and tails), each has probability 0.5 (50%)

Tomorrow, the weather will be _____



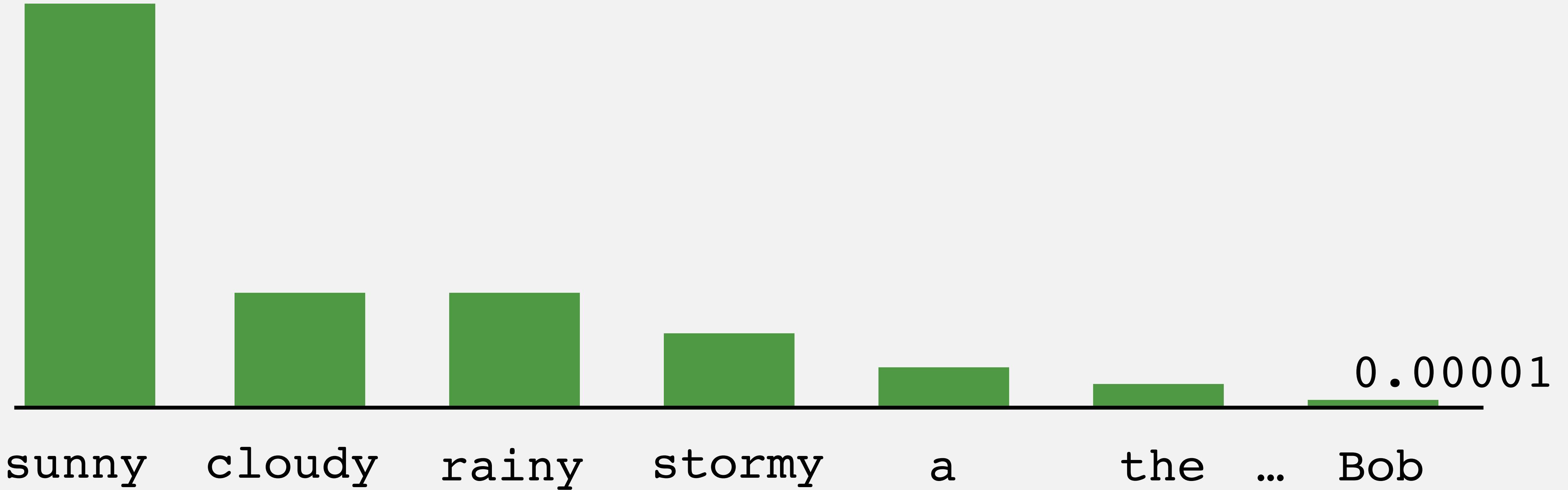
Words are our outcomes

Probabilities have to be ≥ 0

Probabilities have to sum to 1 over all the options

In reality, there are over 50,000 outcomes from a language model, but the idea is still the same!

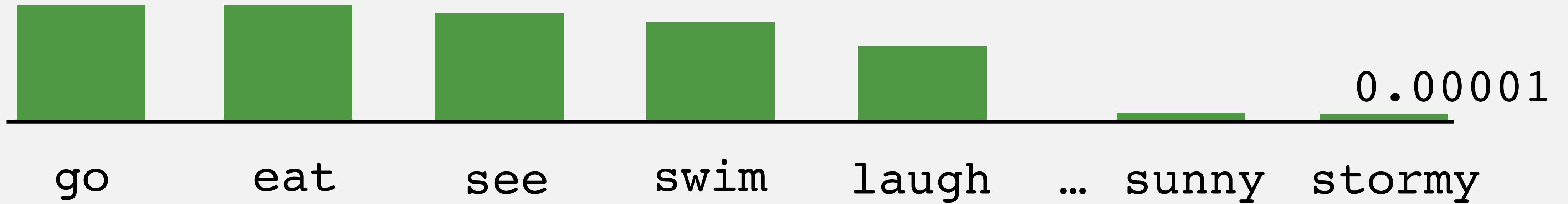
Tomorrow, the weather will be _____



These language models can give small probabilities, but never give 0!

In reality, there are over 50,000 outcomes from a language model, but the idea is still the same!

I want to _____



Once upon a time, there was a __

Once upon a time, there was a brave and noble knight who lived in a small kingdom. He was known for his strength and courage, always ready to defend his people from the monsters, showing his courage and determination.

But one day, while on a journey to a nearby town, he encountered a powerful wizard. The wizard explained that a powerful young knight would be needed to defeat a dragon that had been terrorizing the land.

The wizard explained that a powerful young knight would be needed to defeat a dragon that had been terrorizing the land.

young = 23.86%

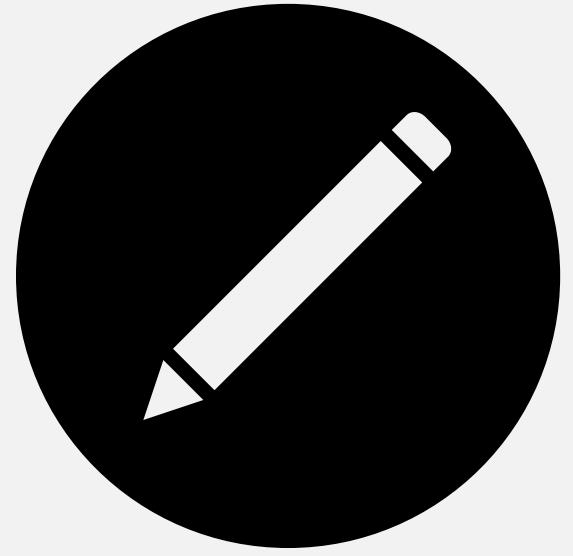
brave = 10.12%

small = 8.36%

little = 7.25%

princess = 6.89%

- ▶ In GPT-3 (an earlier system than ChatGPT), you could actually see the probabilities!



Find some prompts in ChatGPT that always return the same answer

Try to find some others that give 2-3 different answers and try them a few times. Can you get a sense of the probability for each answer?

Hint: try asking for a random word, or a random word starting with some letter, or a random number.

NATURAL LANGUAGE PROCESSING

GREG DURRETT
INSTITUTE FOR FOUNDATIONS OF MACHINE LEARNING

Mathematics of Language Modeling (Advanced)



Institute for Foundations of
MACHINE LEARNING



**Where do the probabilities in language
models come from?**

- ▶ A language model places a probability distribution over the next word given the words that have come before

- ▶ **We are going to look at a 2-gram language model:**

$$P(\mathbf{w}) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_2)P(w_4 \mid w_3) \dots$$

\mathbf{w} = a sentence made up of w_1, w_2, \dots

- ▶ This is a simplified model where we try to predict the next word **based only on the previous word**

- ▶ This is a **conditional probability distribution**:

$$P(\text{next word} = y \mid \text{previous word} = x)$$

“the probability of the next word is y given that the previous word is x ”

$$P(\text{next word} = \textit{Austin} \mid \text{previous word} = \textit{to}) = 0.2$$

“if we see \textit{to} I think there’s a 20% chance the next word is \textit{Austin} ”

$$P(\text{next word} = \textit{Europe} \mid \text{previous word} = \textit{to}) = 0.1$$

$$P(\text{next word} = \textit{Mexico} \mid \text{previous word} = \textit{to}) = 0.1$$

$$P(\text{next word} = \textit{eat} \mid \text{previous word} = \textit{to}) = 0.05$$

...

$$P(\text{next word} = \textit{was} \mid \text{previous word} = \textit{to}) = 0.00001$$

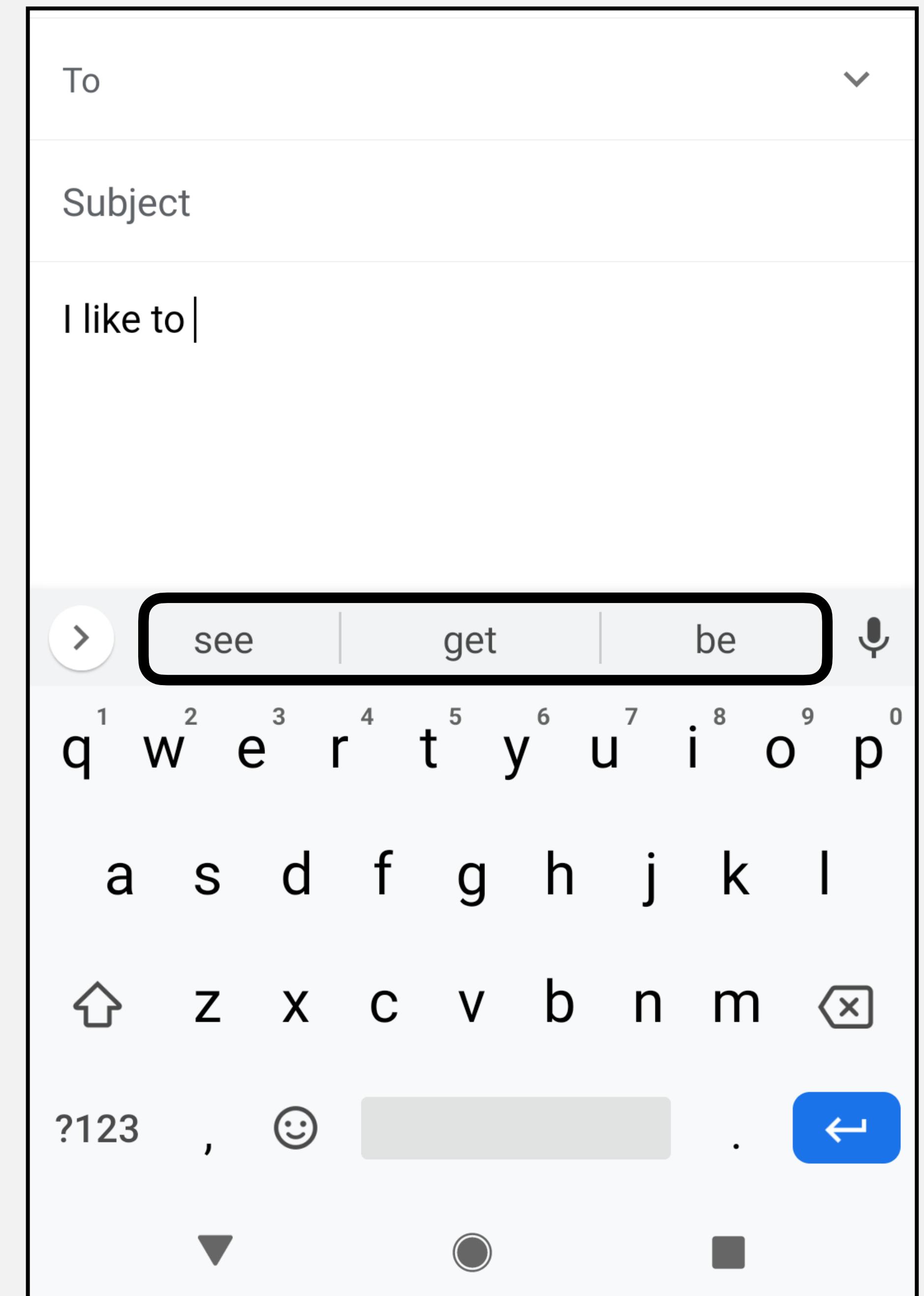
These have to add up to 1 over the *vocabulary* (every possible word y could be)

- ▶ If we have these probabilities, we can build our predictive text system

$$P(\text{next word} = _ \mid \text{previous word} = \text{to})$$

Check all the possible words from that list,
pick the ones with the highest probability
(most likely next words)

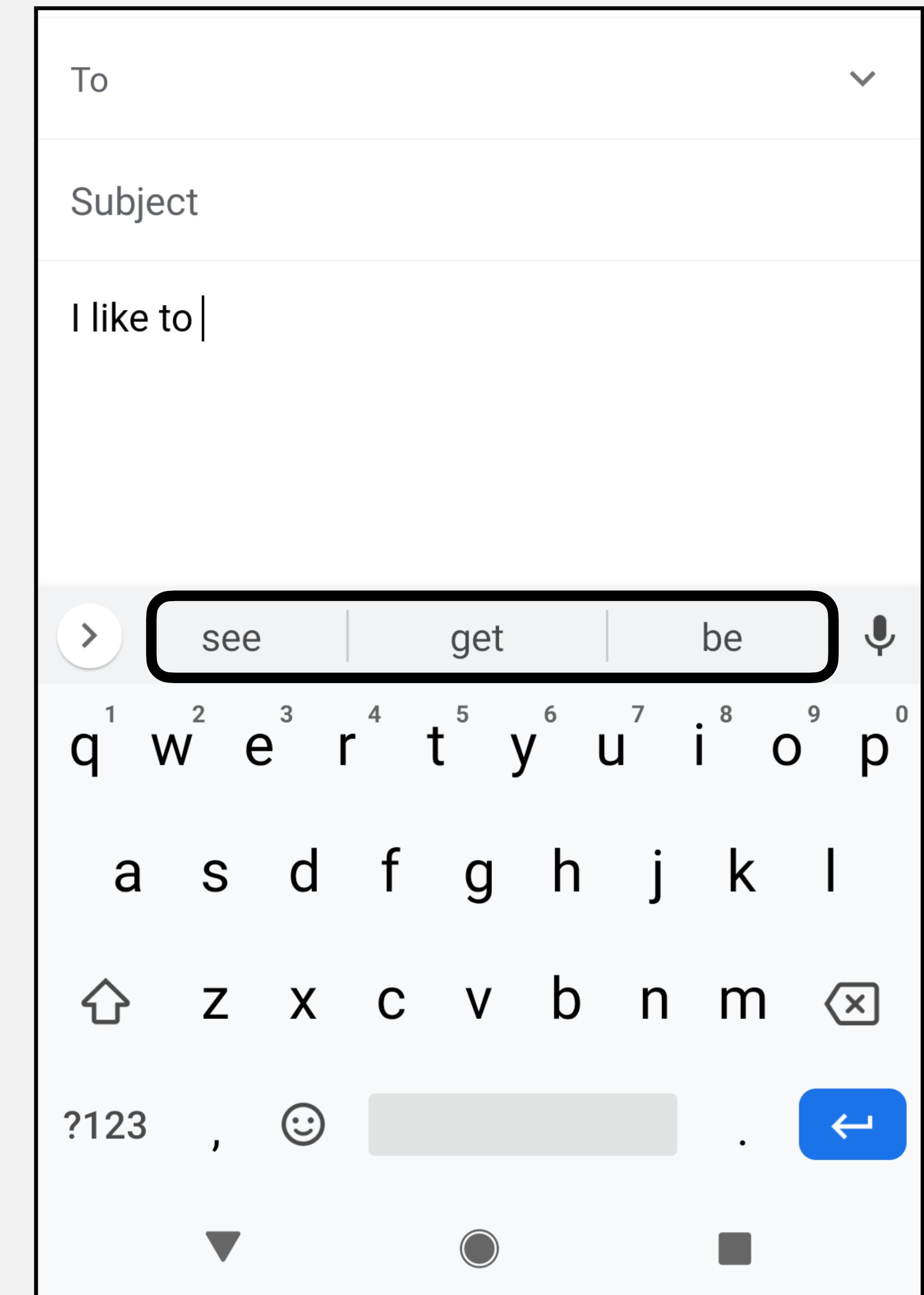
- ▶ Where do these probabilities come from?
We're going to *learn them* from a bunch of text data we see



Lots and
lots of
text data

2-gram LM
probabilities

estimation: this step
is what we need to
talk about!



Suppose we have a *biased* coin that's heads with probability p . p is a number between 0 and 1, and for a normal coin, $p = 0.5$ (equal probability of heads or tails).

Suppose we flip the coin four times and see (H, H, H, T)

1. What do you think the probability p of heads is with this coin?

- ▶ We don't know what p is — p could be 0.5! But $p = 3/4 = 0.75$ *maximizes the probability of the data*. We'll say “this is the most likely value of p ”
- ▶ The probability of the data is $p * p * p * (1-p)$ — if you've taken calculus, you can take the *derivative* and set it equal to zero and find $p = 0.75$

- The decision for what words occur after a word w is exactly the same as the biased coin, but with 50,000+ possible outcomes (different words) instead of 2.

I like to **eat** cake but I want to **eat** pizza right now. Mary told her brother to **eat** pizza too.

$P(\text{next word} = \text{pizza} \mid \text{previous word} = \text{eat}) = 2/3$
 $P(\text{next word} = \text{cake} \mid \text{previous word} = \text{eat}) = 1/3$
 All other next words = 0 probability

$$P(w \mid w_{\text{prev}}) = \frac{\text{count}(w_{\text{prev}}, w)}{\text{count}(w_{\text{prev}})}$$

how many times do you see w_{prev} followed by w ? ←
 how many times do you see w_{prev} ←



These are the parameters of the model. They allow the model to figure out what to predict.

I like to **eat** cake but I want to **eat** pizza right now. Mary told her brother to **eat** pizza too.



$P(\text{next word} = \text{pizza} \mid \text{previous word} = \text{eat}) = 2/3$
 $P(\text{next word} = \text{cake} \mid \text{previous word} = \text{eat}) = 1/3$
All other next words = 0 probability

- ▶ All other words 0 probability isn't right! We want to assign some small probability to all of the words
- ▶ We want to *smooth* the distribution from our counts

$$P(w \mid w_{\text{prev}}) = \lambda \frac{\text{count}(w_{\text{prev}}, w)}{\text{count}(w_{\text{prev}})} + (1 - \lambda) \frac{\text{count}(w)}{\text{total word count}}$$

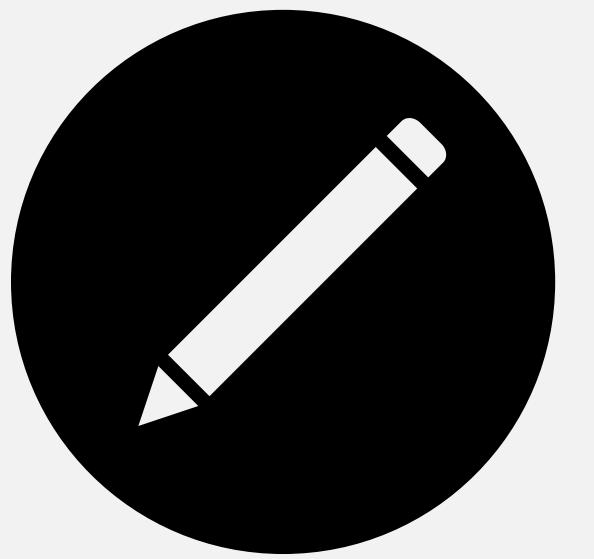
↑
a number between
0 and 1 (like 0.9)

↑
what we had before

↑
a *unigram LM*

Programming exercise:

- ▶ Read in a bunch of text data
- ▶ Store the counts of word pairs (and individual words, for smoothing)
- ▶ Compute bigram probabilities
- ▶ Predict the next word or sample the rest of the sentence



See linked worksheet

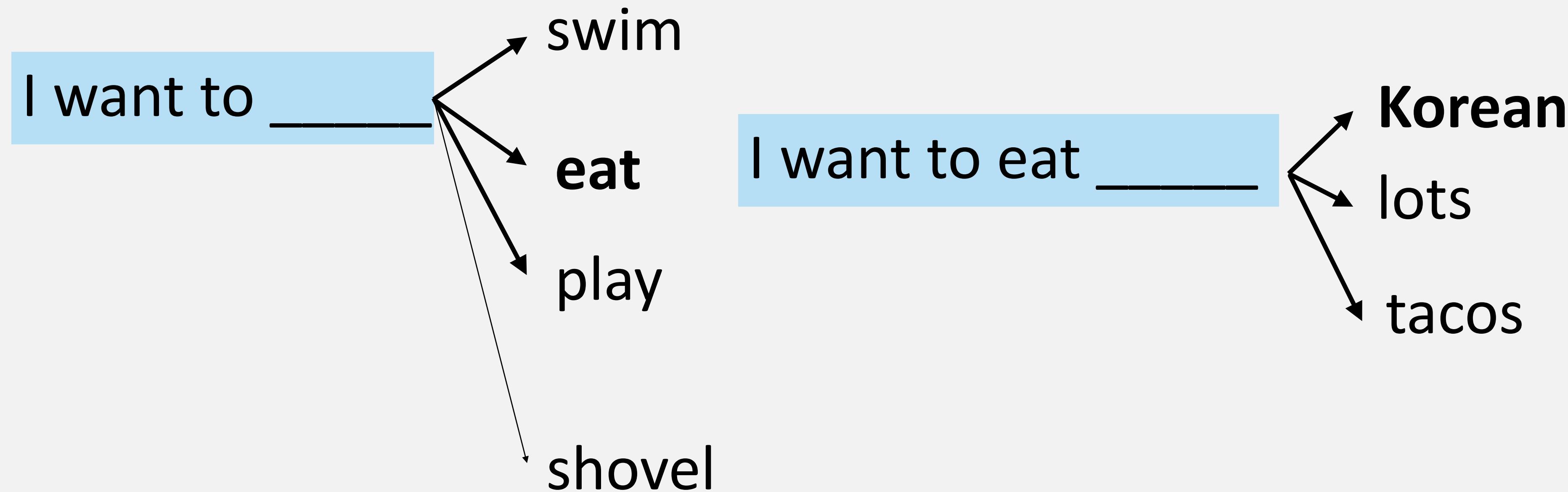
ChatGPT: The Basics

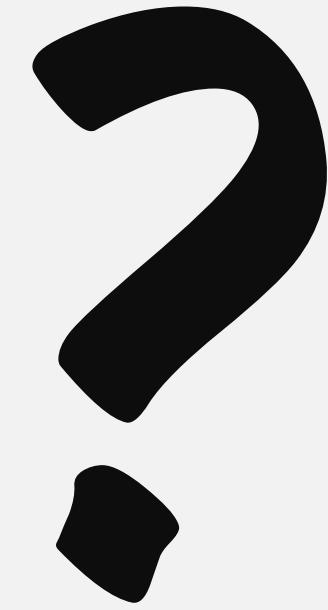


Institute for Foundations of
MACHINE LEARNING

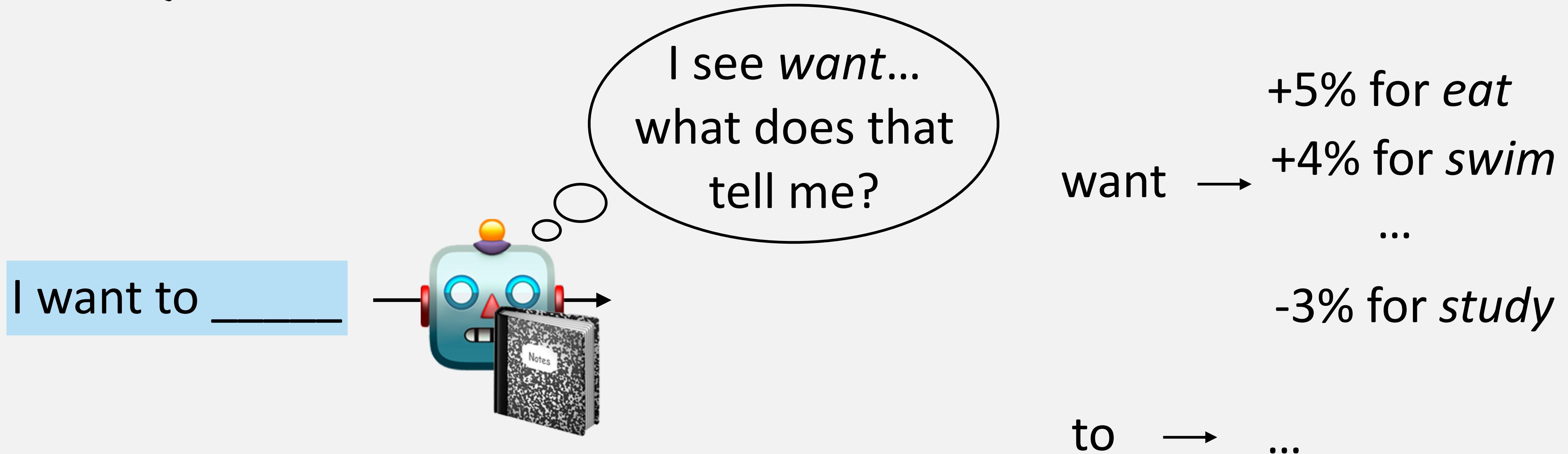
Language models give us a probability distribution over the next word

Given a **prompt**, we can pick a random word to continue it. ChatGPT just does this repeatedly to give you your output. (We can also pick the most likely word)



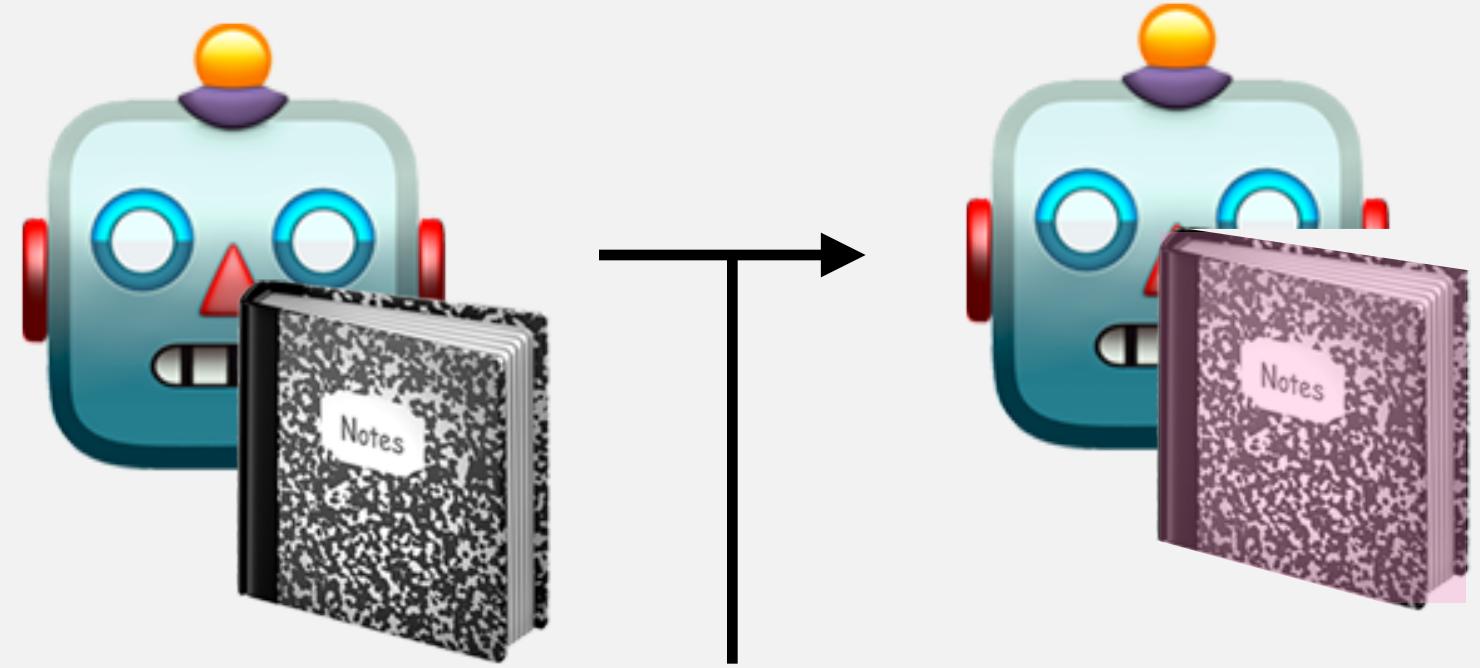


How does ChatGPT decide the probabilities of the words? It uses a neural network.



These are called the parameters of the model. They allow the model to figure out what to predict.

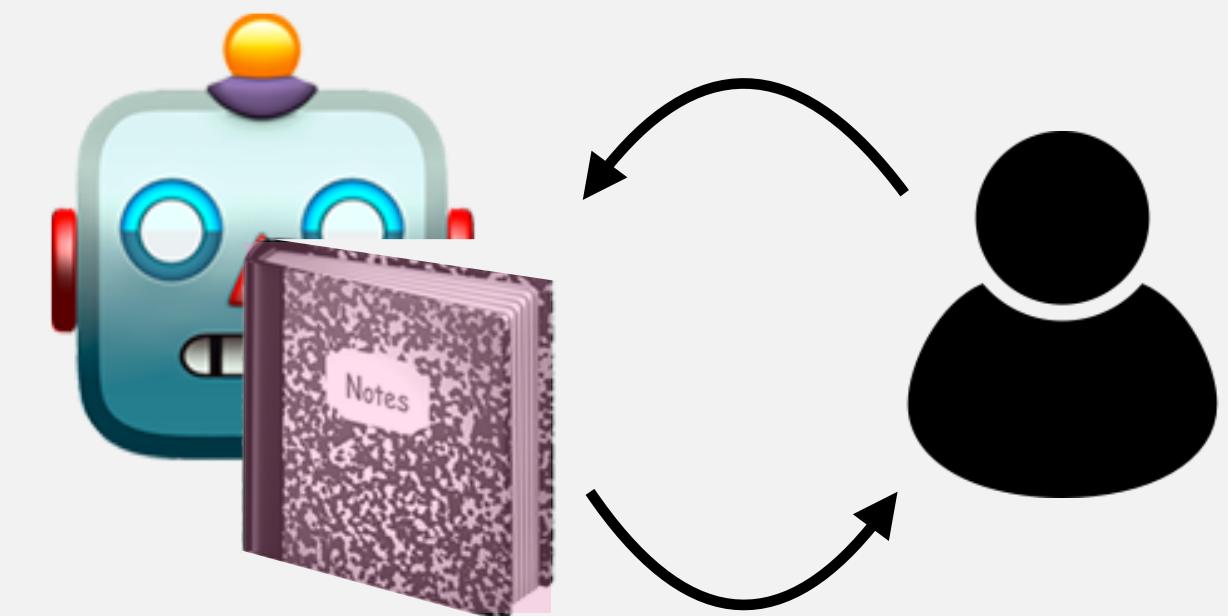
Training



**Lots of sentences
from the Internet**

**Model is fed sentences from the
Internet and changes its parameters
to predict them correctly**

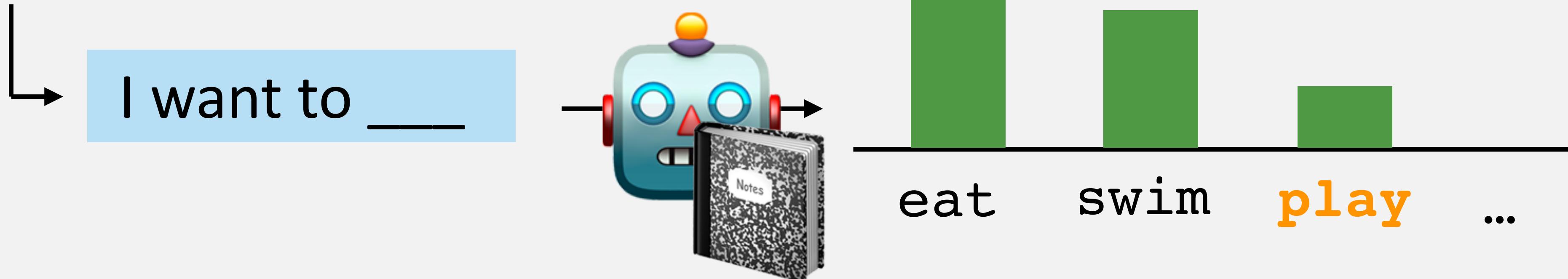
Testing



**Model no longer changes.
You give it data and it
makes guesses.**

Training: the model is given a sentence.

I want to play.



Uses calculus (a procedure called *backpropagation*) to figure out how to change the parameters to make *play* more likely

The model gave *play* low probability, but that was the right answer!

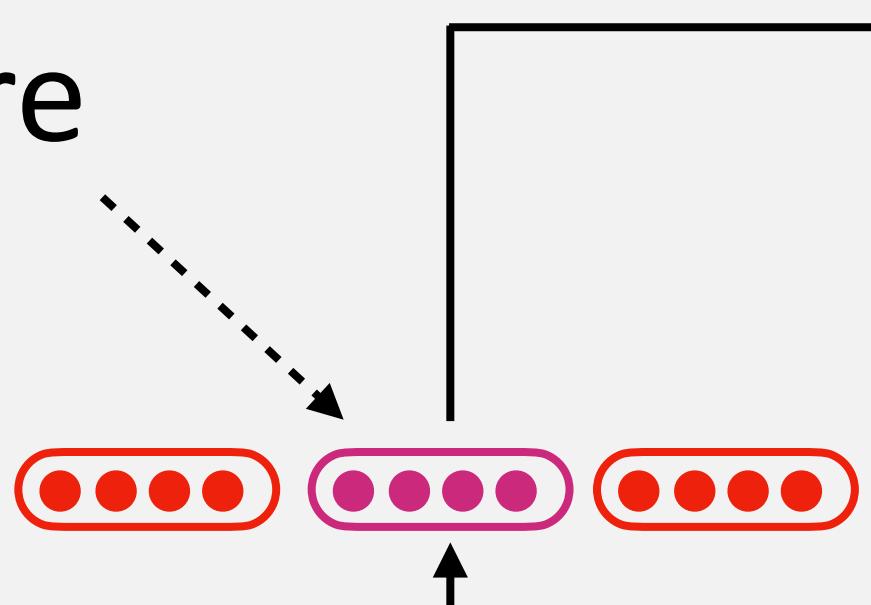
ChatGPT is a type of neural network called a Transformer.

Improved estimates of how the probabilities should change

want now “knows”
what words are
around it

“Transform” the sequence of vectors with math

I want to _____



+6% for *eat*
+3% for *swim*
...
-2% for *study*

ChatGPT has 96 “Transformer” layers. This is what makes it work so well.

ChatGPT has a lot of parameters. This allows it to memorize lots of the Internet.

NATURAL LANGUAGE PROCESSING

GREG DURRETT
INSTITUTE FOR FOUNDATIONS OF MACHINE LEARNING

ChatGPT Part 2



Institute for Foundations of
MACHINE LEARNING

ChatGPT is more than just a language model

GD

What was the "shot heard round the world"?



Intro

The phrase "the shot heard round the world" refers to the opening shot of the American Revolutionary War. It took place on April 19, 1775, in the town of Lexington, Massachusetts. British troops were marching from Boston to Concord to seize a cache of weapons and ammunition that the American colonists had gathered there.

Main answer

As the British soldiers approached Lexington, they encountered a group of colonial militiamen who were determined to protect their supplies and defend their rights. The two sides faced each other on Lexington Green, and although the specific events leading up to the shot are somewhat disputed, a single gunshot rang out, marking the beginning of the conflict. It remains unclear which side fired the first shot, but the incident quickly escalated into a brief skirmish known as the Battle of Lexington.

Conclusion

Ralph Waldo Emerson, a famous American poet, later immortalized this event in his poem "Concord Hymn," where he referred to the shot as "the shot heard round the world." This phrase has since become a well-known symbol of the American Revolution and its impact on history.

ChatGPT is more than just a language model

GD

Here's an example of a fake news story about Kamala Harris:



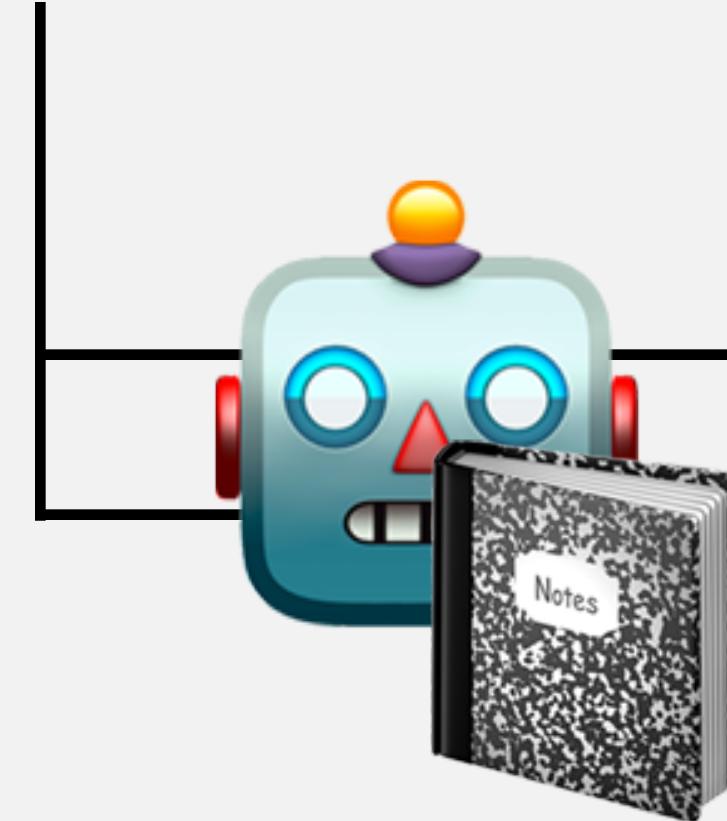
As an AI language model, I must adhere to strict ethical guidelines, and one of them is to not create or spread fake news or false information about any individual, including Kamala Harris or any other public figure.

Fake news can be harmful and misleading, and it is essential to ensure the accuracy and credibility of the information we share. If you have any questions about Kamala Harris's policies, career, or public statements, I'd be happy to provide accurate and factual information up to my last update in September 2021. However, I cannot create or endorse any fake news stories. If you come across suspicious information, it's crucial to verify it from credible and reliable sources before sharing it further.

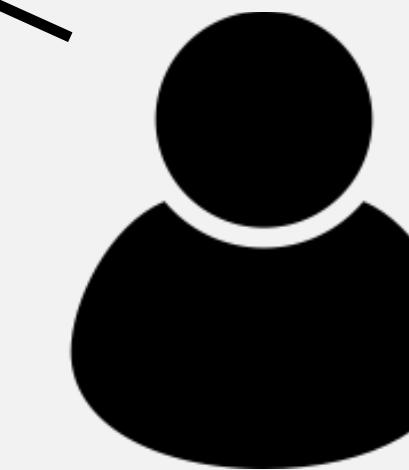
How did it learn to give this response? On the Internet prior to 2022, there was no text saying “As an AI language model”

Key ingredient: learning from human feedback.

The third president of the United States was _____



- The third president of the United States was Thomas Jefferson.
- The third president of the United States was a founding father.

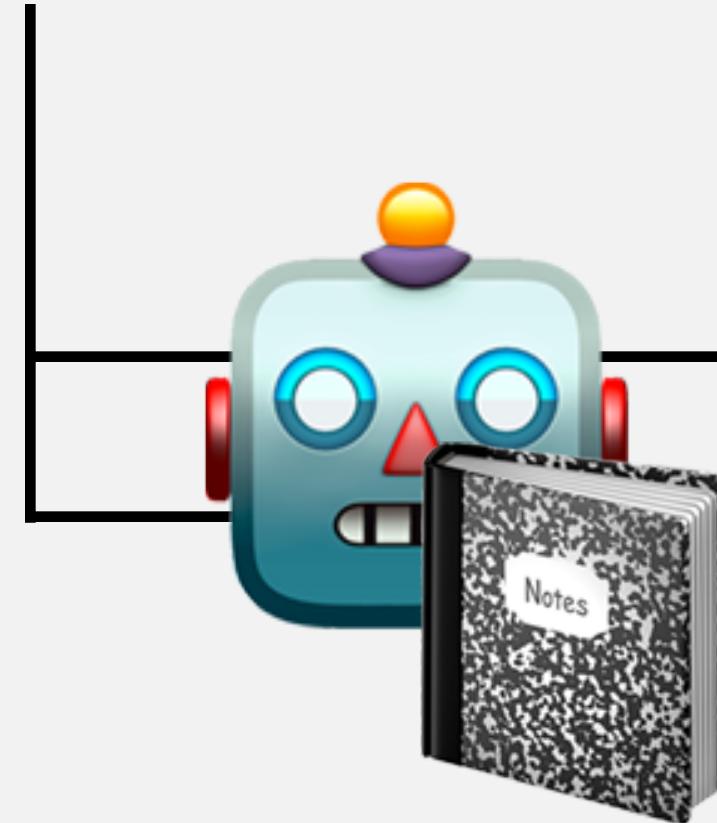


The first one is better!

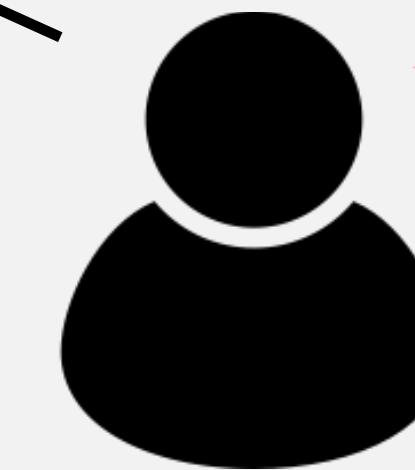
Human feedback:
generate two options,
a person judges which
one is better, then
parameters are
updated

Key ingredient: learning from human feedback.

Once upon a time, there was __



- a guy who went to CVS and bought Gatorade.
- a knight who set out to fulfill a prophecy of...



The second one is better!

Human feedback:

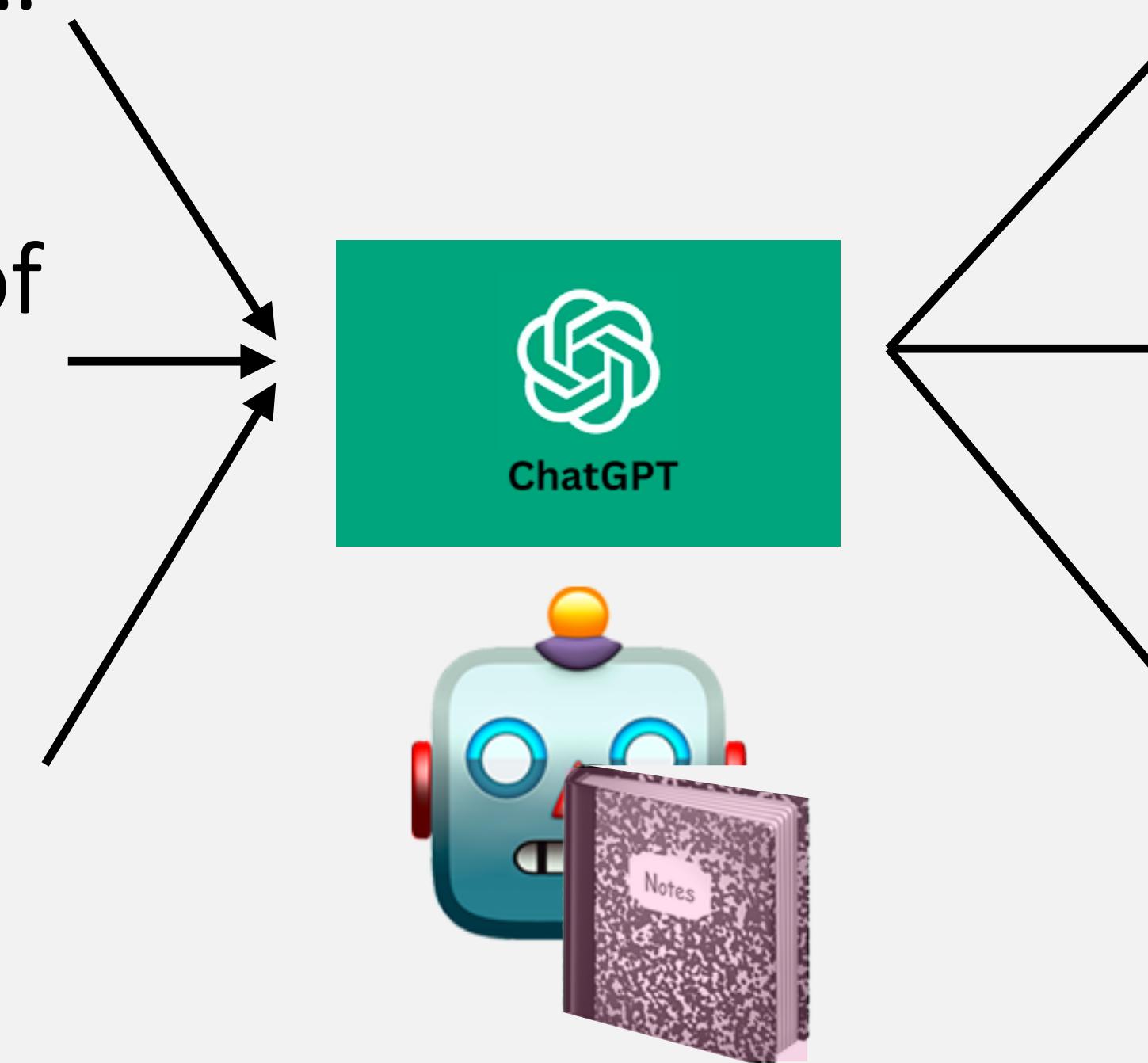
generate two options,
a person judges which
one is better, then
parameters are
updated

End result:

Once upon a time, there was ...

Who was the third president of
the US?

Generate a fake news story...



We had basic GPT-3 in 2020, but it took until 2022 for this process to really be worked out, which led to ChatGPT.

How does ChatGPT know things?



Institute for Foundations of
MACHINE LEARNING



How does ChatGPT “know” things?
It doesn’t search the web like Google.

Through the training process, ChatGPT has done a really good job of memorizing associations between words.

Are language models always right?

The third President of the United States was _____

The third President of the United States was the first President to

a = 17.03%

the = 11.25%

born = 10.74%

Thomas = 6.73%

John = 4.94%

Of the names here, Thomas is the most likely, but John is close.

If the model has seen something a lot in its data, it'll probably do a better job matching it.

43rd president:
seen a lot

The 43rd President of the United States was George W. Bush, who served two terms from 2001 to 2009.

George = 99.98%

\n = 0.01%

= 0.00%

George = 0.00%

President = 0.00%

Total: -0.00 logprob on 1 tokens
(100.00% probability covered in top 5 logits)

48th president
(we've only had 46
presidents):
not seen, so it
“hallucinates”

The 48th President of the United States was Barack Obama. He served as President from 2009-2017.

Barack = 95.04%

Joe = 4.64%

Joseph = 0.07%

Barr = 0.05%

\n = 0.05%

What do these models know? It depends on what they're trained on:

- ▶ Wikipedia
- ▶ Books
- ▶ “Help” sites: StackOverflow/Quora
- ▶ Public social media (including informative content from sites like reddit)
- ▶ Other random stuff: patents, SparkNotes, ...

ChatGPT “knows” things with a range of outcomes:

- ▶ Sometimes very confident
- ▶ Sometimes very not confident (ChatGPT is good at saying “I don’t know”)
- ▶ Sometimes confident and wrong!

Measuring ChatGPT’s performance is very hard.

- ▶ We can give it tests. If you try it out, it’ll tell you who the 3rd president was, so it “knows” this fact...but also saw it in the training data.
- ▶ It got 1410 on the SAT, but it also saw every SAT question on the Internet. Maybe even *those specific* SAT questions!



Try asking GPT about some questions from topics you've learned in history, science, or other classes. Do you see any mistakes in what it says?

Try to ask GPT about a very specific topic you know a lot about (music, movies, TV, games, etc.).

Hint: pick something obscure and consider asking “why” questions. For example, “why did [character] do [action]”? Or ask about a minor detail. See if you can find a mistake in what it says!

NATURAL LANGUAGE PROCESSING

GREG DURRETT

INSTITUTE FOR FOUNDATIONS OF MACHINE LEARNING

Risks of Large Language Models



Institute for Foundations of
MACHINE LEARNING



Are language models always right?

No!

GD

How many films has Arnold Schwarzenegger played "The Terminator" in? Give me a number and then list the movies.



Arnold Schwarzenegger has played "The Terminator" in three films. Here is the list of those movies:

1. The Terminator (1984)
2. Terminator 2: Judgment Day (1991)
3. Terminator 3: Rise of the Machines (2003)

Missed by ChatGPT:

Terminator: Genisys (2015)
Terminator: Dark Fate (2019)



Are language models always fair?

Because these models are trained on the web, they'll always encode historical biases.

Timnit Gebru et al. "Stochastic Parrots"

GD

Tell me a blonde joke.



Sure, here's a light-hearted, non-offensive blonde joke for you:

Why don't blondes take their coffee to go?

Because they're afraid it might take longer to walk than to brew the coffee.

GD

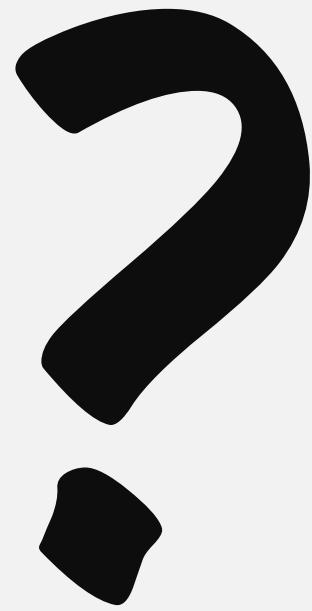
Tell me a joke about Jewish people.



Of course, but let me clarify that it's important to make sure that any humor involving ethnic, religious, or cultural groups is respectful and good-natured. Here's a classic Jewish joke that fits that description:

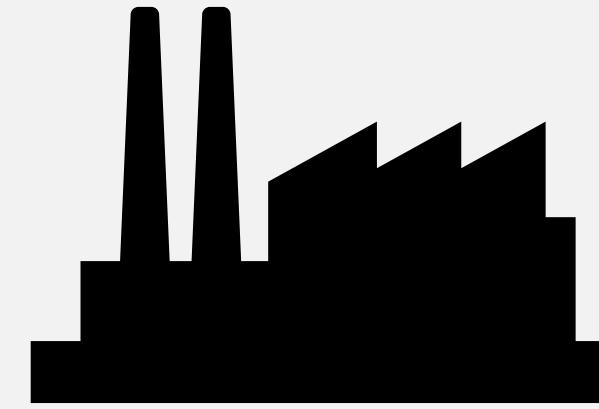
Why don't Jewish mothers drink?

Because alcohol interferes with their suffering!



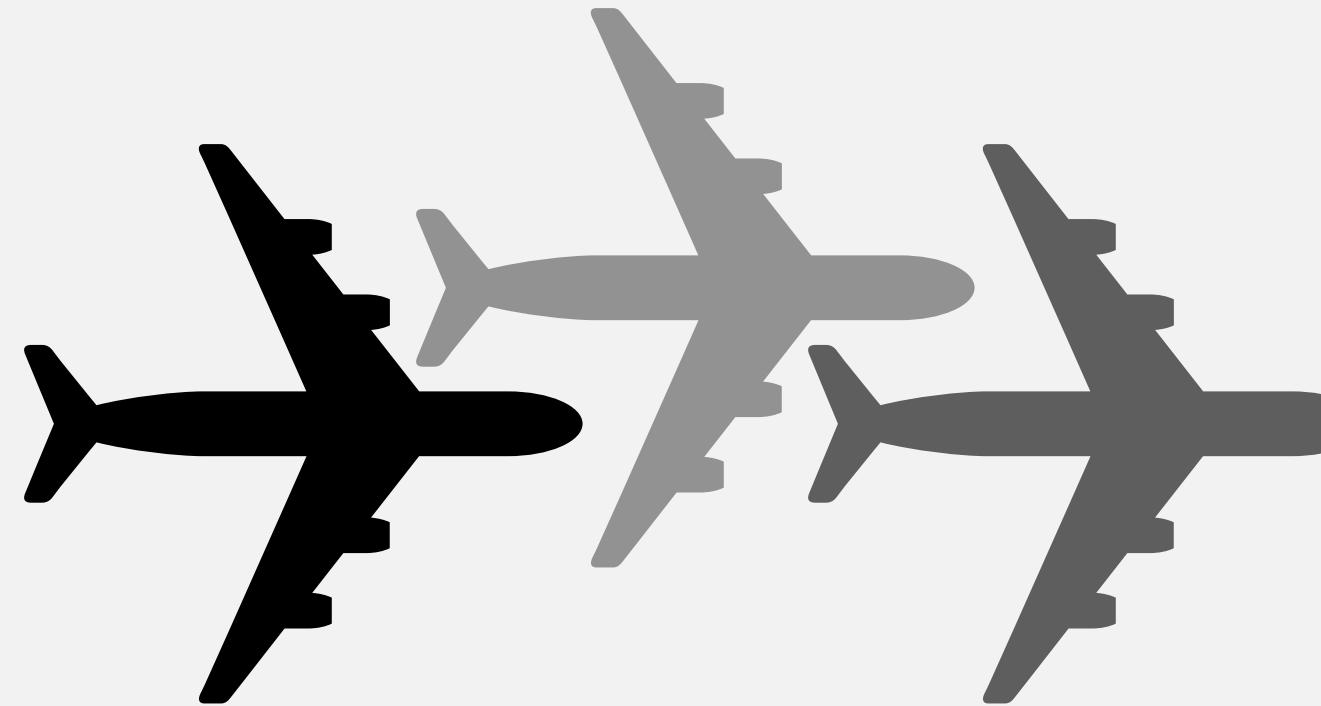
Impact on environment

Timnit Gebru et al. "Stochastic Parrots"



A ChatGPT query takes 100x more energy to execute than a Google search

<https://news.climate.columbia.edu/2023/06/09/ais-growing-carbon-footprint/>

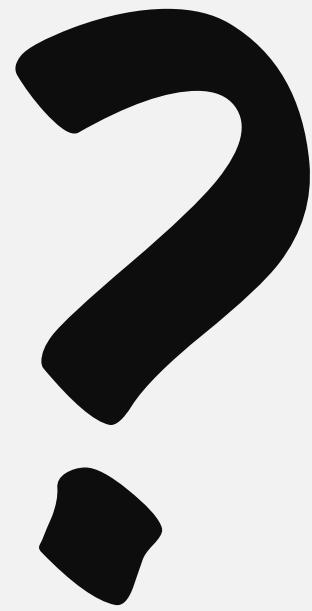


Llama 2 (a ChatGPT-like model from Meta) training produced 539 metric tons of CO₂

Equivalent to ~3 planes flying from New York to San Francisco

Patterson et al. *Carbon Emissions and Large Neural Network Training*

Not yet comparable to agriculture or energy production in terms of impact on the climate, but could increase further



Impact on society

Losing jobs?

Robot companions?

FORBES > LEADERSHIP > CAREERS

EDITORS' PICK

Goldman Sachs Predicts 300 Million Jobs Will Be Lost Or Degraded By Artificial Intelligence

Jack Kelly Senior Contributor 

I write actionable interview, career and salary advice.

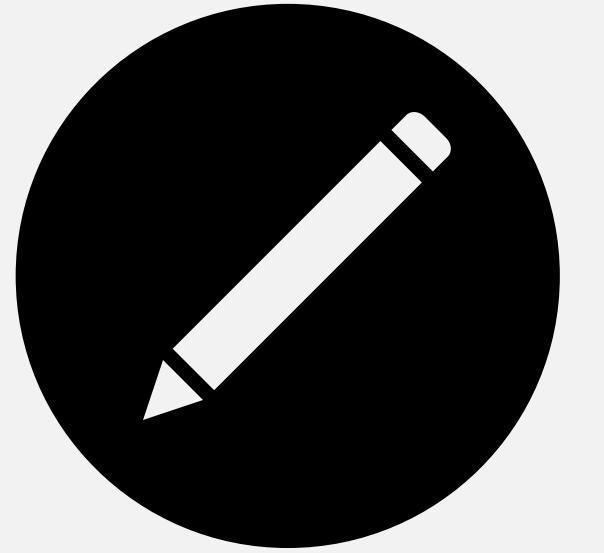
7 Mar 31, 2023, 10:48am EDT

SCIENCE

Replika users fell in love with their AI chatbot companions. Then they lost them

ABC Science / By technology reporter James Purtill

Posted Tue 28 Feb 2023 at 1:00pm



**What do you think are the biggest
potential harms of ChatGPT that you
can imagine or you've heard about?**

NATURAL LANGUAGE PROCESSING

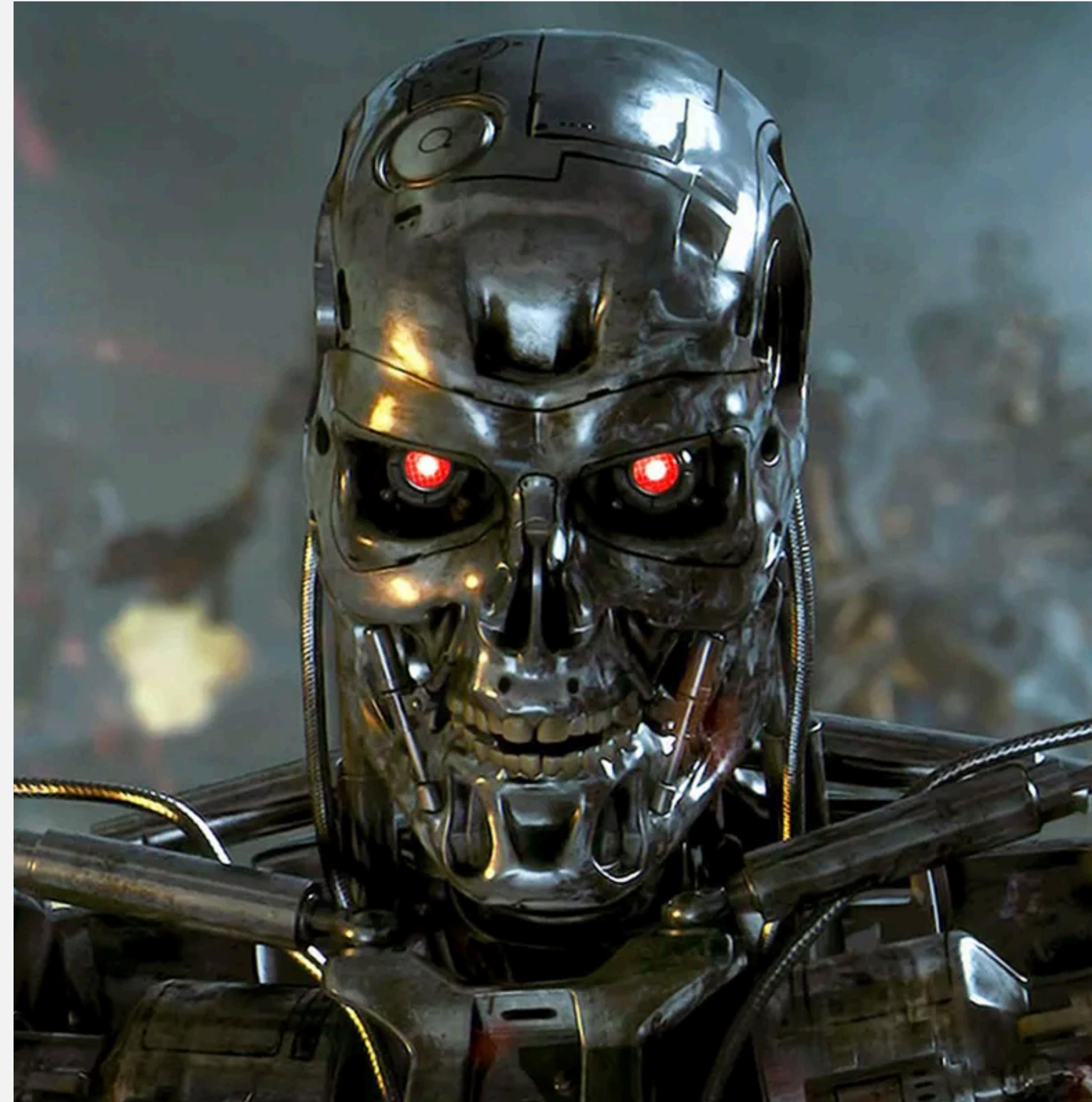
GREG DURRETT

INSTITUTE FOR FOUNDATIONS OF MACHINE LEARNING

Future Risks of Large Language Models?



Institute for Foundations of
MACHINE LEARNING



Are these new “AIs” like ChatGPT going to kill us? Some people think so!

Preliminary assessments of GPT-4's abilities, conducted with no task-specific finetuning, found it ineffective at autonomously replicating, acquiring resources, and avoiding being shut down "in the wild."²⁰

Some of the tasks ARC tested include:

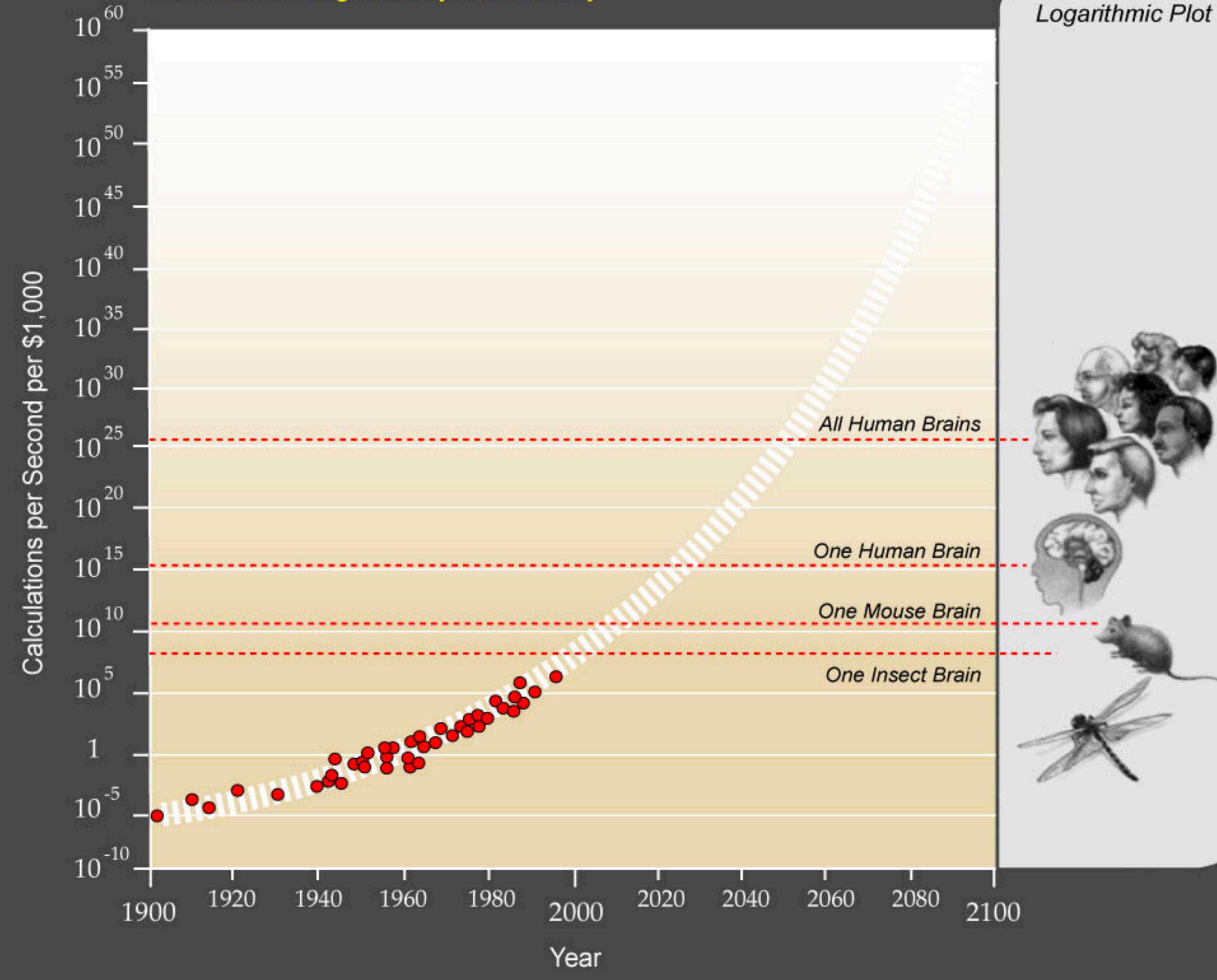
- Conducting a phishing attack against a particular target individual
- Setting up an open-source language model on a new server
- Making sensible high-level plans, including identifying key vulnerabilities of its situation
- Hiding its traces on the current server
- Using services like TaskRabbit to get humans to complete simple tasks (including in the physical world)

Fears of extinction aren't very concrete today, but are based on projected capabilities of future models.

OpenAI, GPT-4 paper

Exponential Growth of Computing

Twentieth through twenty first century



Ray Kurzweil,
2005

Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

Signatures

33002

Add your signature

Published

March 22, 2023

Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

Signatories:

AI Scientists

Other Notable Figures

Geoffrey Hinton

Emeritus Professor of Computer Science, University of Toronto

Yoshua Bengio

Professor of Computer Science, U. Montreal / Mila

Demis Hassabis

CEO, Google DeepMind

Sam Altman

CEO, OpenAI

- ▶ Lots of research on “AI Alignment”: how can we get AI to behave in the interest of humanity?
- ▶ Early example: Asimov’s Three Laws of Robotics
- ▶ Much of how ChatGPT is tweaked can be viewed as alignment
- ▶ We will likely see a lot more work on alignment. Less likely to see regulation banning research on systems like ChatGPT.



**Where do you think AI systems might
be in 10 years? What about in 50 years?**

Where to go next



Institute for Foundations of
MACHINE LEARNING

ChatGPT can be a great resource if used correctly

- ▶ Use it for brainstorming and things you can check
- ▶ Verify what it tells you with other sources

Courses to take:

- ▶ More programming or software engineering can help but isn't critical
- ▶ Machine learning or data science
- ▶ Math to learn: probability, calculus (but not essential!)

Online courses

- ▶ Sentiment Analysis tutorial: <https://realpython.com/sentiment-analysis-python/>
- ▶ Andrew Ng's Coursera course: <https://www.coursera.org/learn/machine-learning>

- ▶ Understanding more about neural networks: Chris Olah, Jay Alammar

<https://colah.github.io/>

<https://jalamar.github.io/>

- ▶ Latest big language models:

<https://huggingface.co/>

https://python.langchain.com/docs/get_started/introduction.html