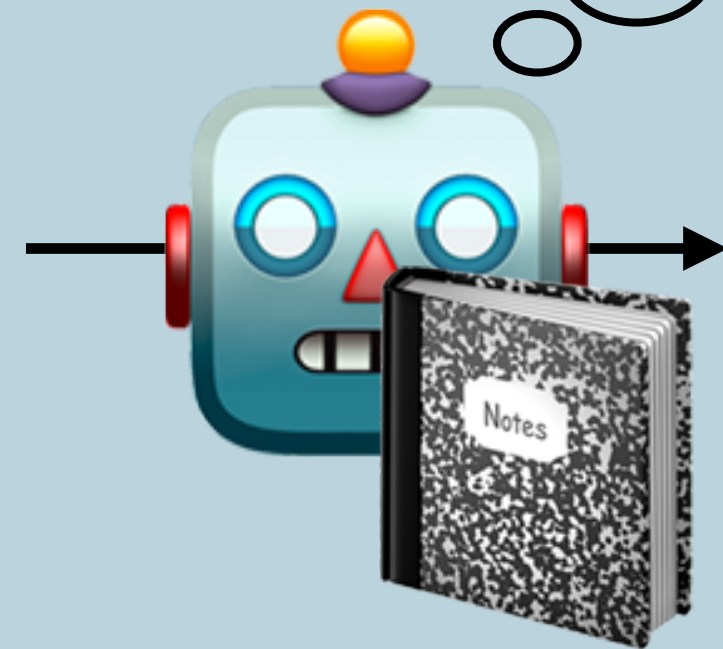


Neural Language Models

I want to _____



I see *want*...
what does that
tell me?

want →

+5% for *eat*
+4% for *swim*
...
-3% for *study*

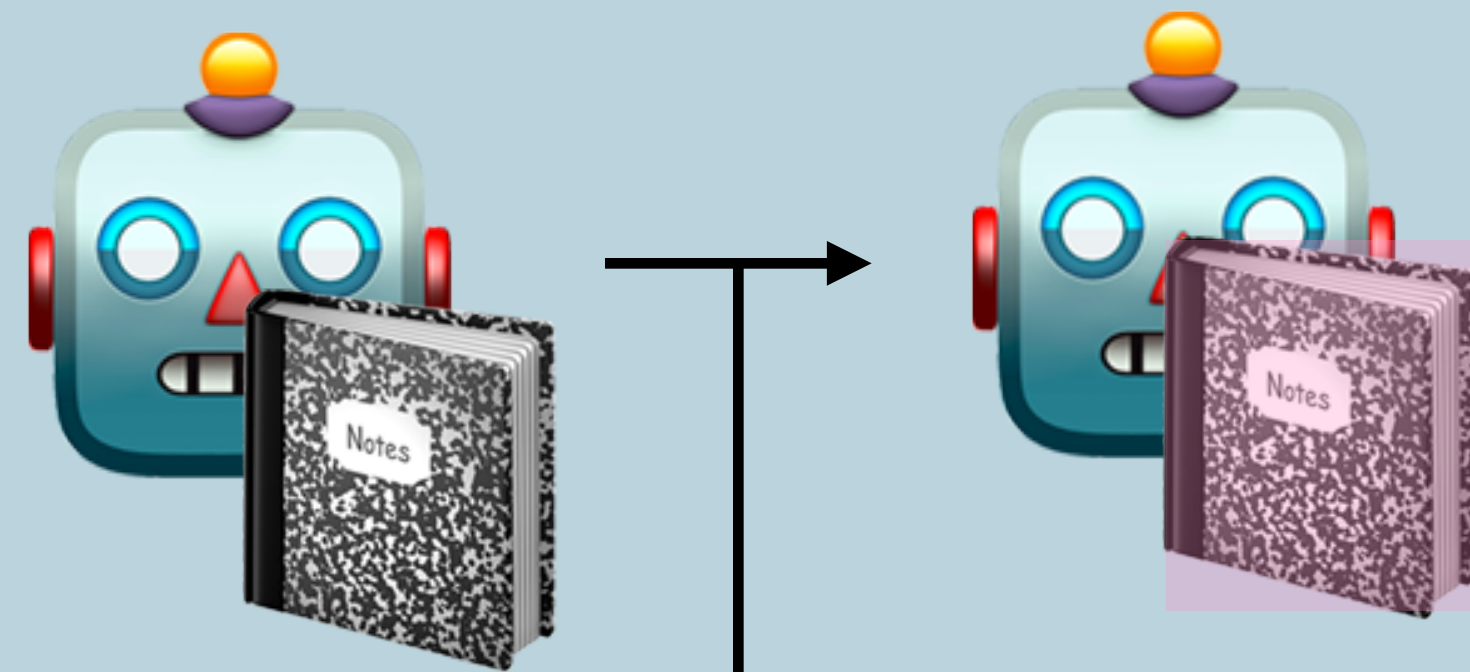
to → ...



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

Neural Language Models

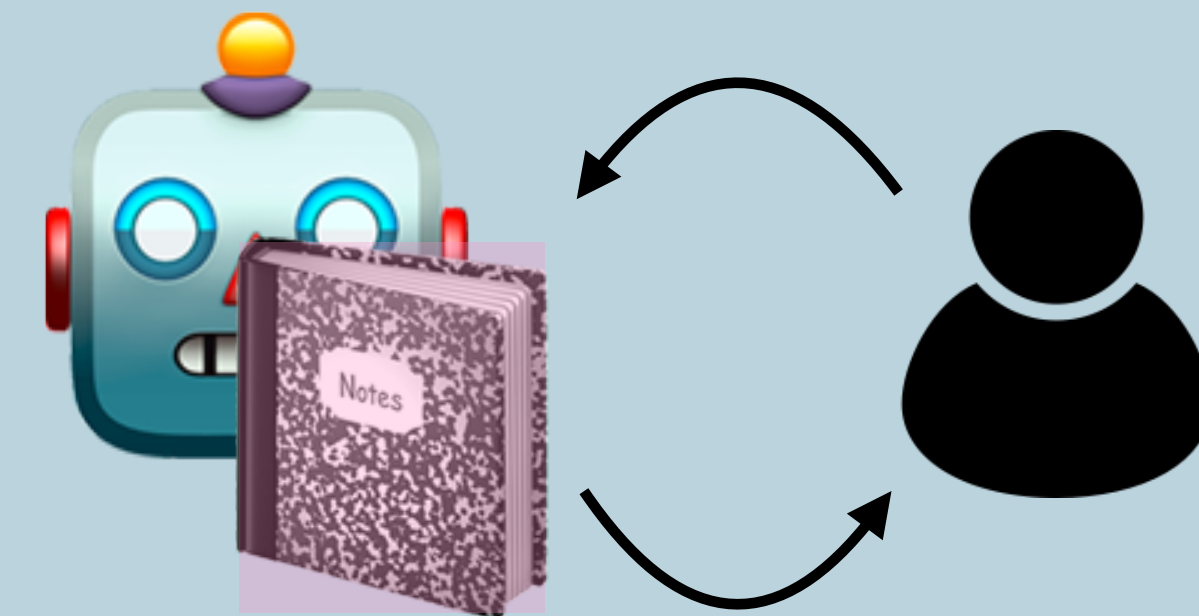
Training



**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.**

Maximum Likelihood Estimation

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)

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*.
- ▶ 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 $p = 0.75$

Maximum Likelihood Estimation

- ▶ 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.

$P(\text{next word} = \textit{pizza} \mid \text{previous word} = \textit{eat}) = 2/3$

$P(\text{next word} = \textit{cake} \mid \text{previous word} = \textit{eat}) = 1/3$

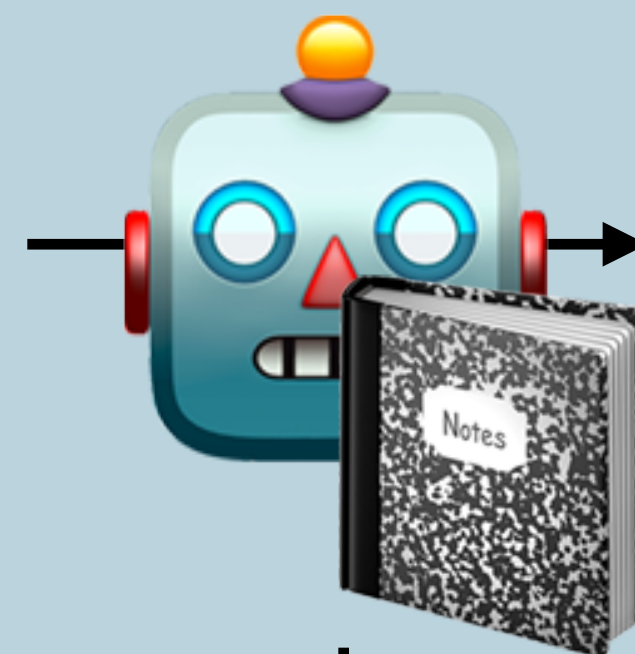
All other next words = 0 probability

- ▶ Learning an LM: update its parameters to match the observed data just like this!

Neural Net LMs

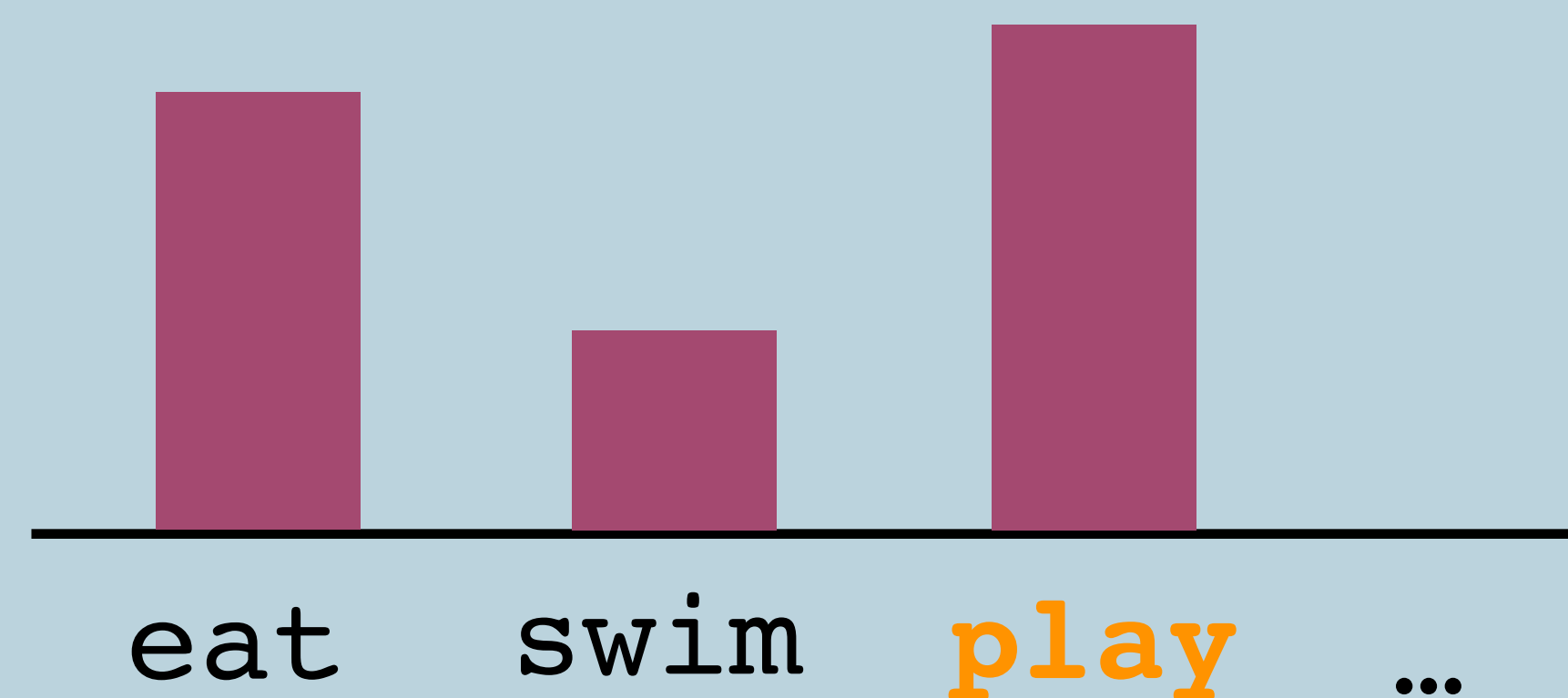
I want to play.

└─ I want to ____



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!



Transformers

ChatGPT has 96 “Transformer” layers. This is what makes it work so well and allows it to memorize the Internet.

See: Jay Alammam, *The Illustrated Transformer*

Improved estimates of how the probabilities should change

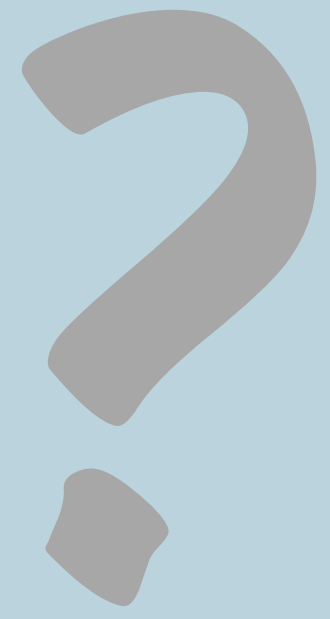
want now “knows”
what words are
around it

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

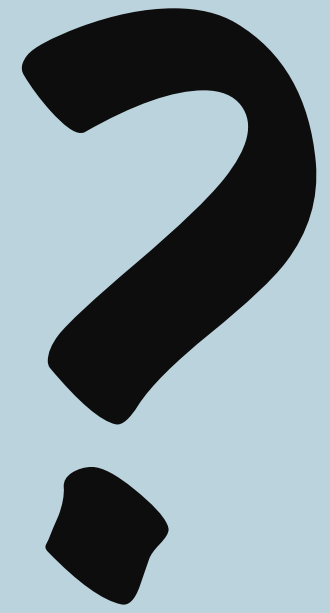
“Transform” the sequence of vectors with math

Turn each word into a vector (list of numbers)

I want to _____

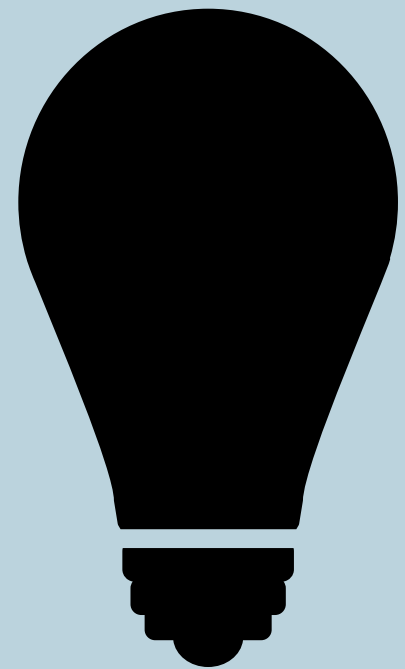


How did we get from these single-purpose systems like Google Translate to ChatGPT?



What's the core technology behind LMs?
(Neural nets, maximum likelihood estimation, Transformers)

How does ChatGPT go beyond a basic LM?



We'll contextualize where these models are and how they're impacting society.

ChatGPT: Beyond Language Modeling

But is ChatGPT really just a language model?

What was the "shot heard round the world"?

The shot that was heard round the world was fired in Lexington, MA. It was fired because, at the time, America was still under British rule. The British soldiers had been taking firewood by throwing it down from the top of their fort at Lexington. Someone who was not a soldier, by the name of Joseph Warren (who had ties to the Sons of Liberty or the Sons of Freedom) shouted to the British to stop doing it. The British soldiers ignored him, not

What was the "shot heard round the world"?

On the morning of April 19th, 1775, AMC's "Turn" did the impossible, and turned a war not many were aware was coming into one of the catalyst events that inspired the births of our modern America. Indeed, all of AMC's Revolutionary War dramas have done a masterful job of bringing this bloody time in our nation's birth "back" into view in exciting new ways for the 21st century.

What was the "shot heard round the world"?

"What to the slave is the 4th of July?"

What was the sideline occupation of Thomas Jefferson's wife, Martha, and their enslaved girl, Sally Hemmings?

What was the name of the first African American regiment from Vermont?

Samples from GPT-3
(a “basic” LM)

GD

What was the "shot heard round the world"?



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.

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.

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.

Intro

Main answer

Conclusion

GD

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

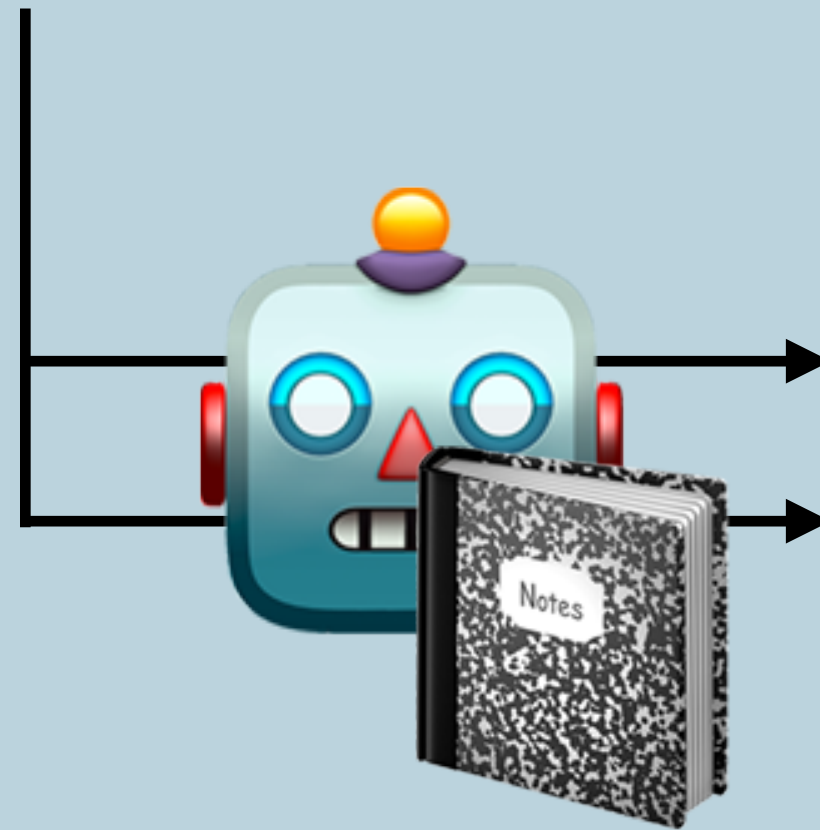


As an AI language model, I must adhere to strict ethical guidelines, and one of them is to not create or spread fake news or false information about any individual, including Kamala Harris or any other public figure.

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.

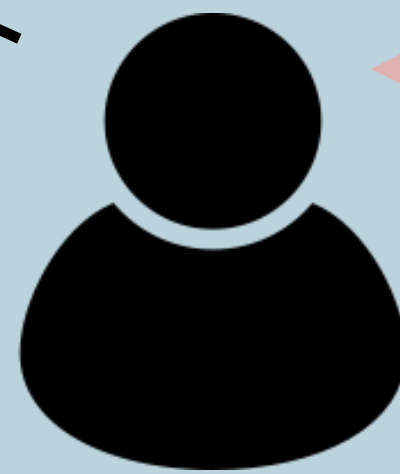
How did it learn to give this response? On the Internet prior to 2022, there was no text saying “As an AI language model”

The third president of the United States was _____



The third president of the United States was Thomas Jefferson.

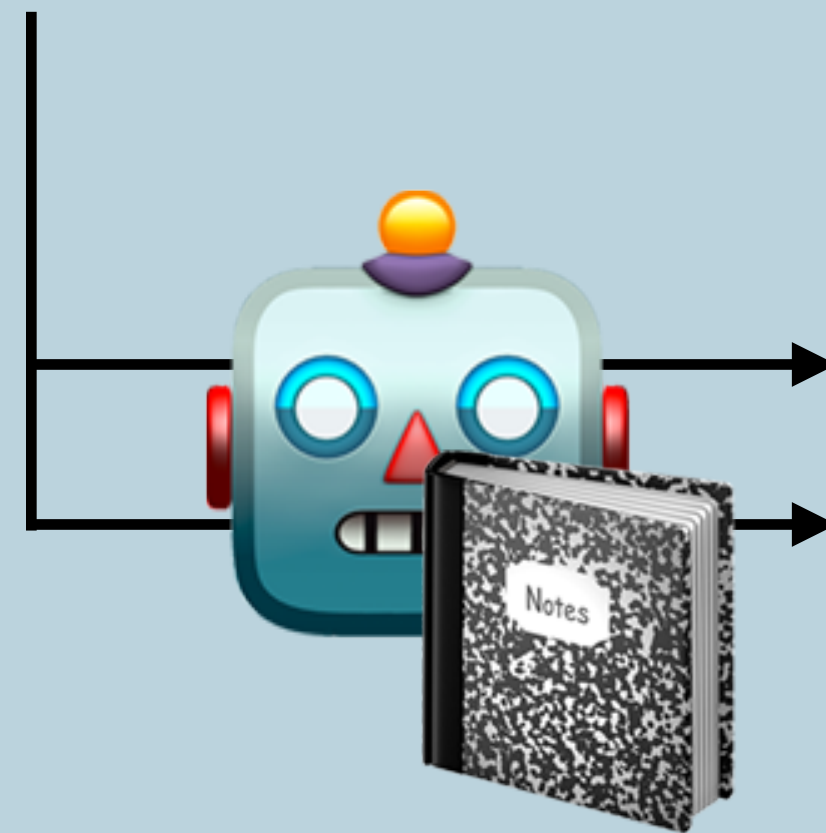
The third president of the United States was a founding father.



The first one is better!

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

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

Reinforcement Learning from Human Feedback

- **Learn a reward function** from this labeled data about which of two outcomes is better. Use a Bradley-Terry model of preferences.

$$\hat{P}[\sigma^1 \succ \sigma^2] = \frac{\exp \sum \hat{r}(o_t^1, a_t^1)}{\exp \sum \hat{r}(o_t^1, a_t^1) + \exp \sum \hat{r}(o_t^2, a_t^2)}.$$

“if people prefer 1 to 2, then the reward of 1 should be greater than the reward of 2 by a substantial margin”

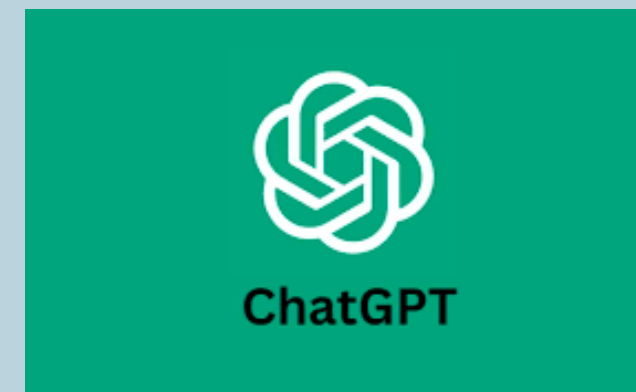
- Do reinforcement learning to maximize that reward

End Result

Once upon a time, there was ...

Who was the third president of the US?

Generate a fake news story...

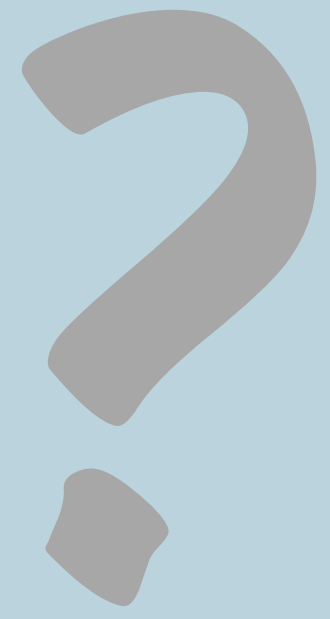


Interesting stories

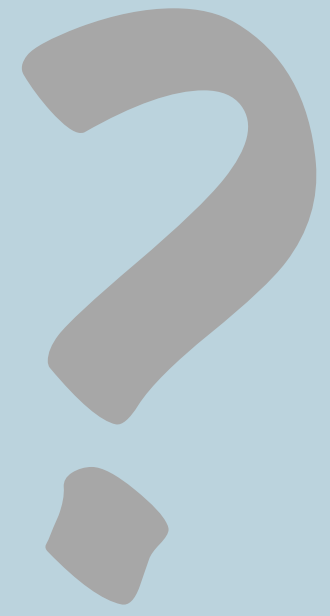
Informative answers

Non-answers to harmful prompts

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

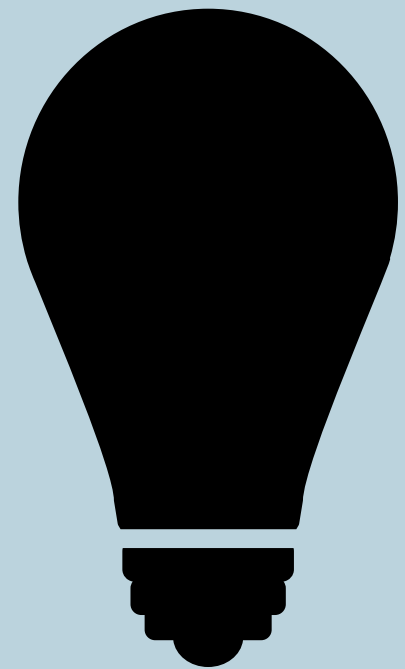


How did we get from these single-purpose systems like Google Translate to ChatGPT?



What's the core technology behind LMs?
(Neural nets, maximum likelihood estimation, Transformers)

How does ChatGPT go beyond a basic LM?



We'll contextualize where these models are and how they're impacting society.