Exploring the mechanisms underlying LLMs like ChatGPT

Lisa Meeden
Professor of Computer Science
Swarthmore College

Outline of the talk

- Discuss a word game, try it ourselves, and test how well ChatGPT can play it
- Learn how word embeddings are created to represent text for LLMs
- Learn how LLMs generate text through transformers and self-attention
- Reflect on LLMs

How to play NYT game Connections

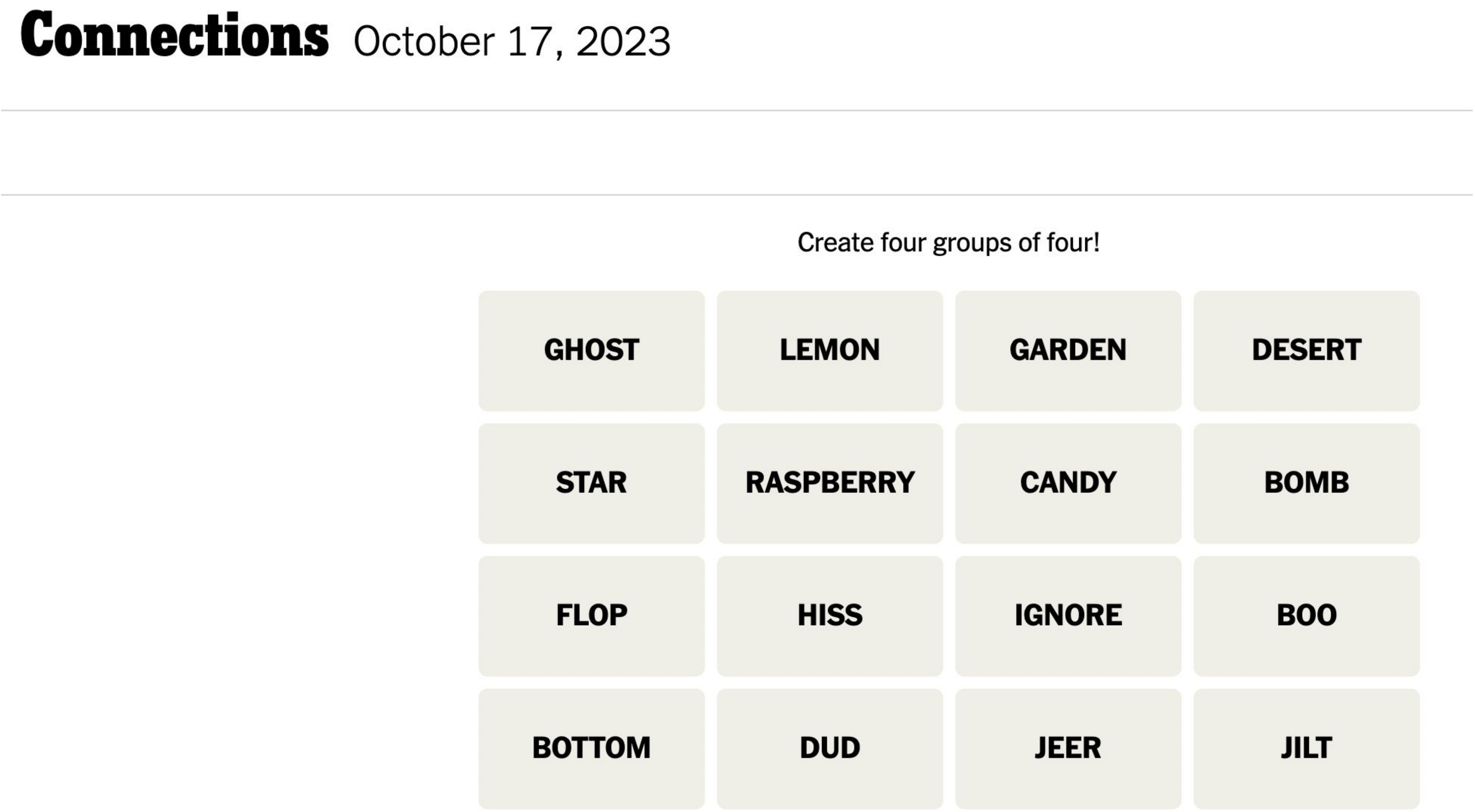
Find groups of four items that share something in common

- Pick four items, and check if your guess is correct
- Try to find all four groups without making four mistakes

Category examples:

- FISH: Bass, Flounder, Salmon, Trout
- FIRE _____: Ant, Drill, Island, Opal

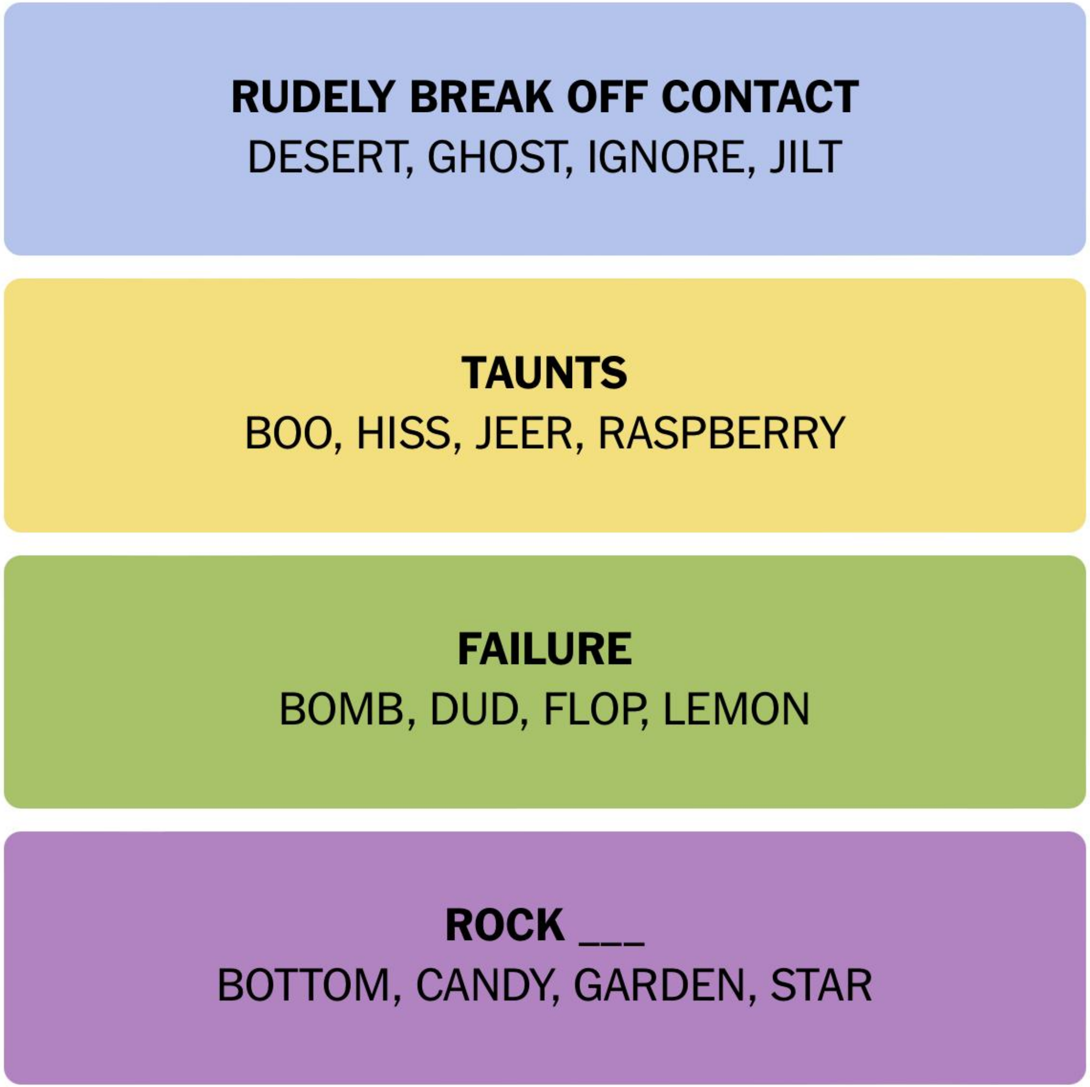
Each puzzle has exactly one solution. Watch out for words that seem to belong to multiple categories.



Connections October 17, 2023

Create four groups of four!

GHOST	LEMON	GARDEN	DESERT
STAR	RASPBERRY	CANDY	BOMB
FLOP	HISS	IGNORE	BOO
BOTTOM	DUD	JEER	JILT



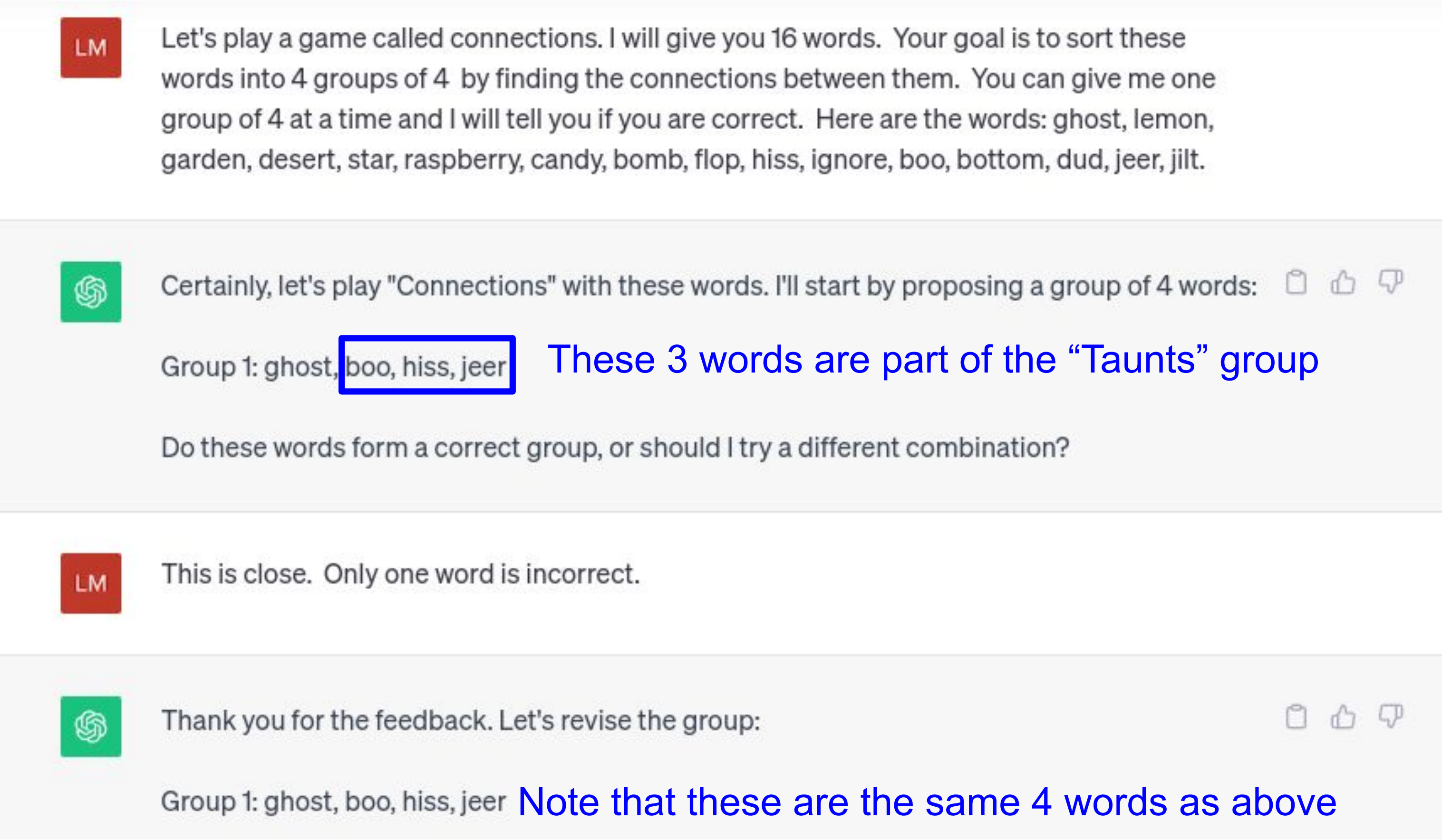
RUDELY BREAK OFF CONTACT
DESERT, GHOST, IGNORE, JILT

TAUNTS
BOO, HISS, JEER, RASPBERRY

FAILURE
BOMB, DUD, FLOP, LEMON

ROCK ___
BOTTOM, CANDY, GARDEN, STAR

Let’s see how OpenAI’s ChatGPT (3.5) handles this game



Let's play a game called connections. I will give you 16 words. Your goal is to sort these words into 4 groups of 4 by finding the connections between them. You can give me one group of 4 at a time and I will tell you if you are correct. Here are the words: ghost, lemon, garden, desert, star, raspberry, candy, bomb, flop, hiss, ignore, boo, bottom, dud, jeer, jilt.

Certainly, let's play "Connections" with these words. I'll start by proposing a group of 4 words:

Group 1: ghost, boo, hiss, jeer

These 3 words are part of the “Taunts” group

Do these words form a correct group, or should I try a different combination?

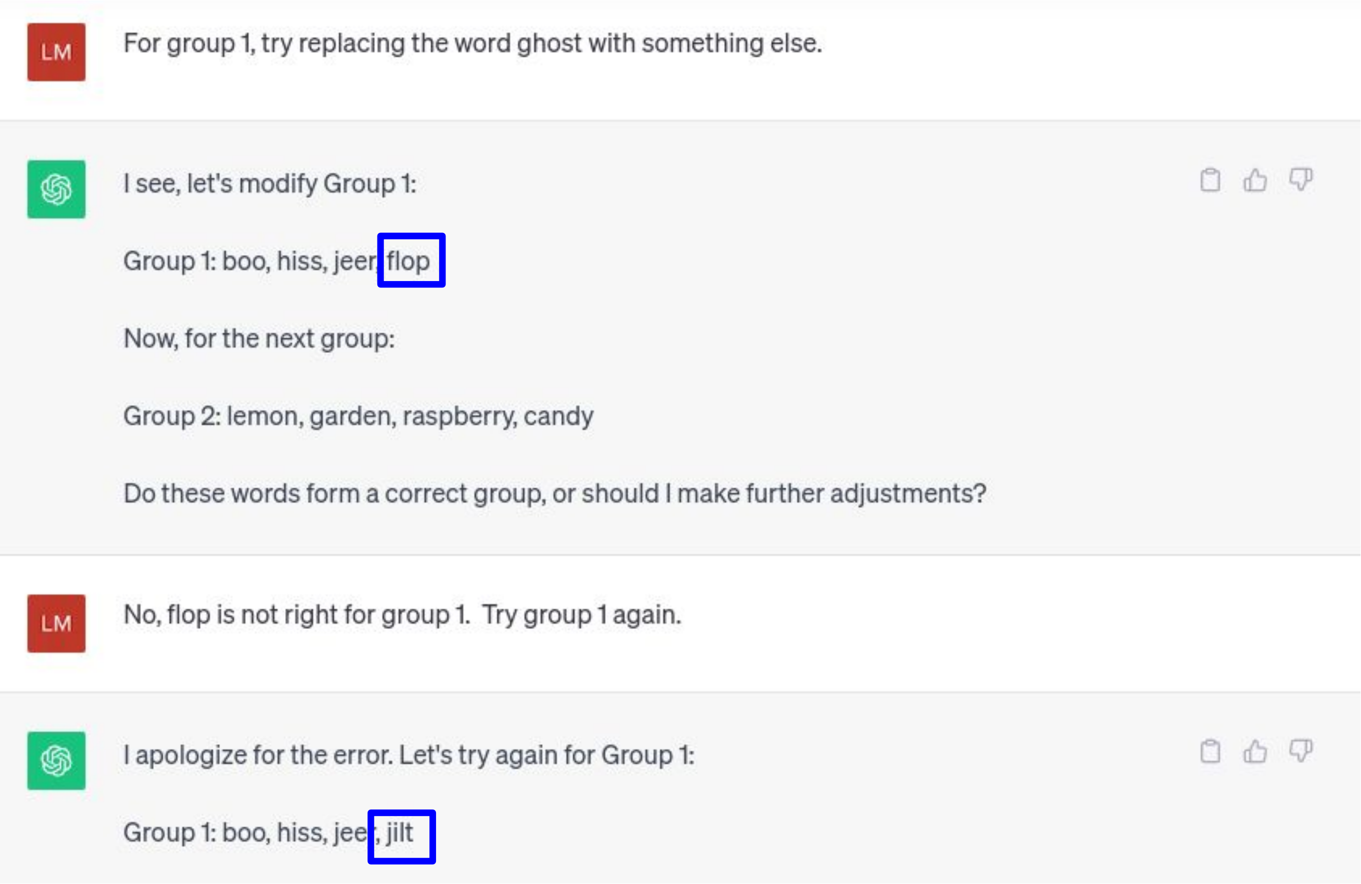
This is close. Only one word is incorrect.

Thank you for the feedback. Let's revise the group:

Group 1: ghost, boo, hiss, jeer

Note that these are the same 4 words as above

Provide a hint



LM

For group 1, try replacing the word ghost with something else.

I see, let's modify Group 1:

Group 1: boo, hiss, jeer, flop

Now, for the next group:

Group 2: lemon, garden, raspberry, candy

Do these words form a correct group, or should I make further adjustments?

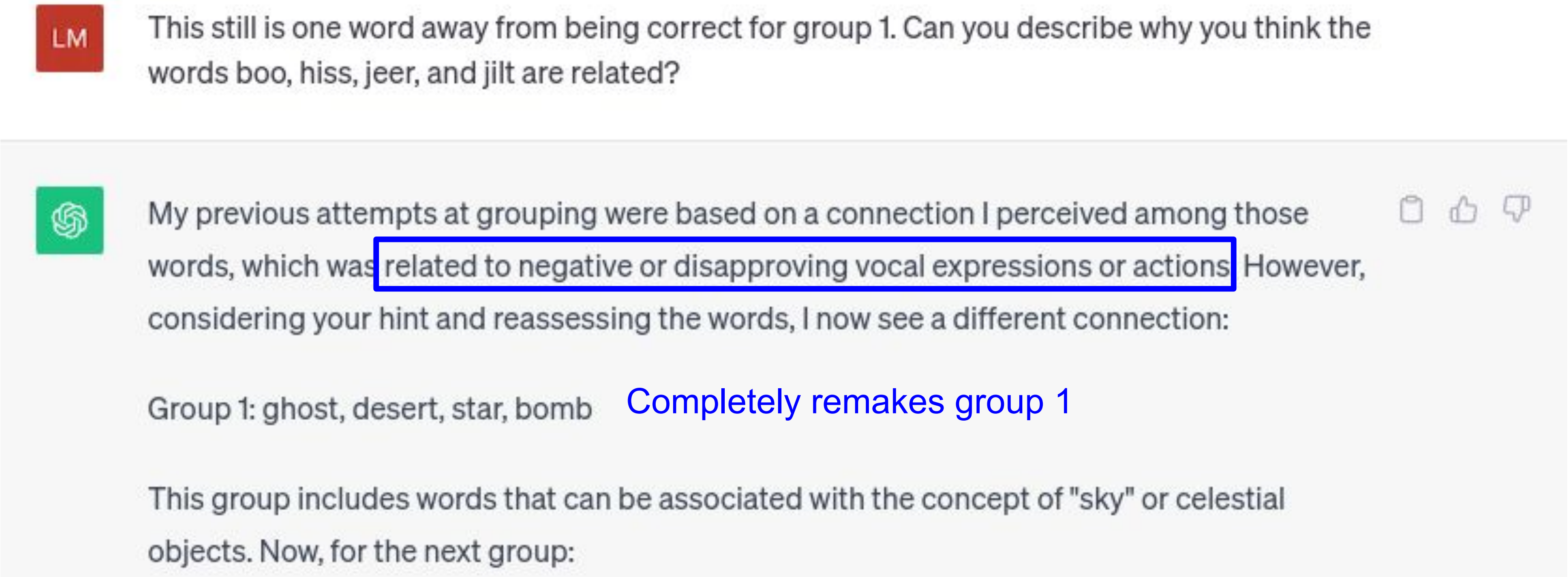
LM

No, flop is not right for group 1. Try group 1 again.

I apologize for the error. Let's try again for Group 1:

Group 1: boo, hiss, jeer, jilt

Found the right concept for the group



This still is one word away from being correct for group 1. Can you describe why you think the words boo, hiss, jeer, and jilt are related?

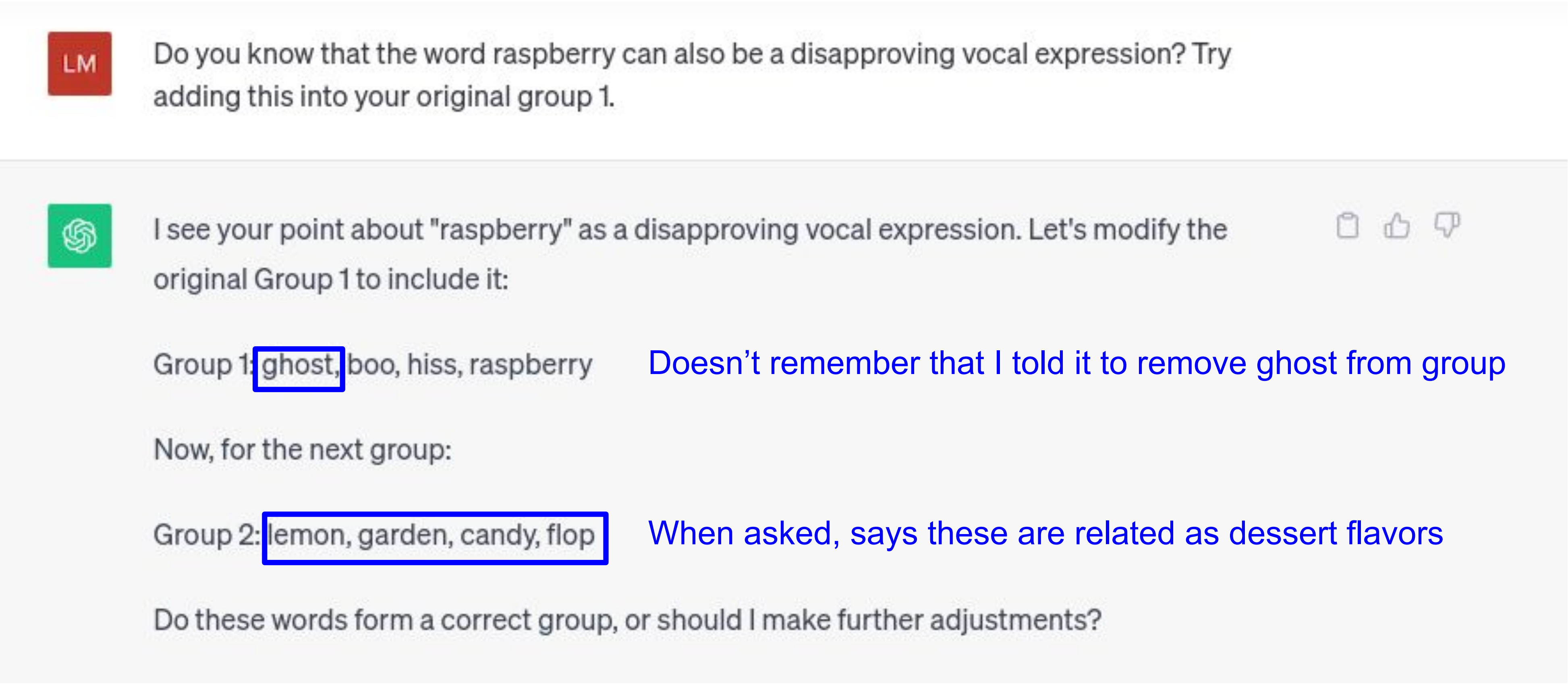
My previous attempts at grouping were based on a connection I perceived among those words, which was related to negative or disapproving vocal expressions or actions. However, considering your hint and reassessing the words, I now see a different connection:

Group 1: ghost, desert, star, bomb

Completely remakes group 1

This group includes words that can be associated with the concept of "sky" or celestial objects. Now, for the next group:

Even with another hint it cannot correctly complete the group



Do you know that the word raspberry can also be a disapproving vocal expression? Try adding this into your original group 1.

I see your point about "raspberry" as a disapproving vocal expression. Let's modify the original Group 1 to include it:

Group 1: ghost, boo, hiss, raspberry

Doesn’t remember that I told it to remove ghost from group

Now, for the next group:

Group 2: lemon, garden, candy, flop

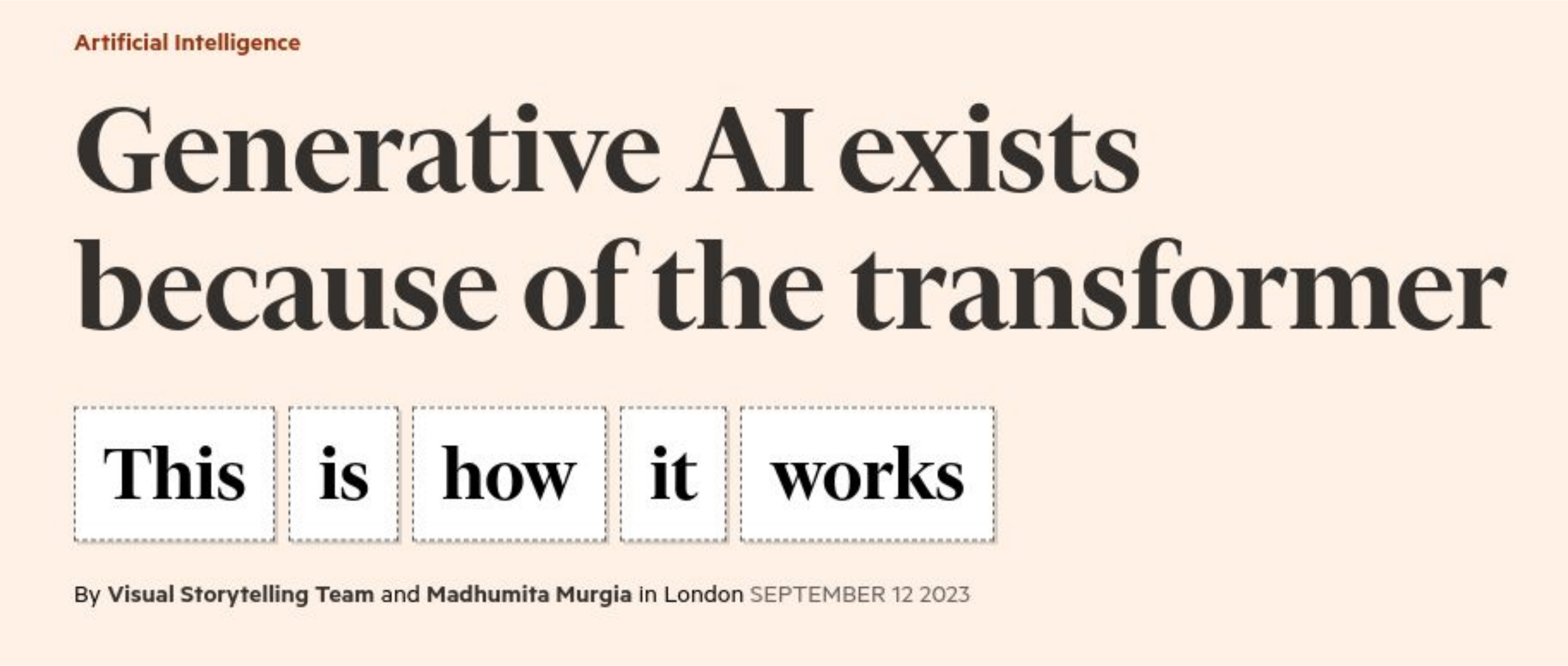
When asked, says these are related as dessert flavors

Do these words form a correct group, or should I make further adjustments?

Let’s learn more about how LLMs like ChatGPT work

- First we will see how words are transformed into numeric data suitable for deep learning
- As you learn more about how LLMs work, think about why the word game Connections might be challenging for LLMs

Article in the *Financial Times*



Artificial Intelligence

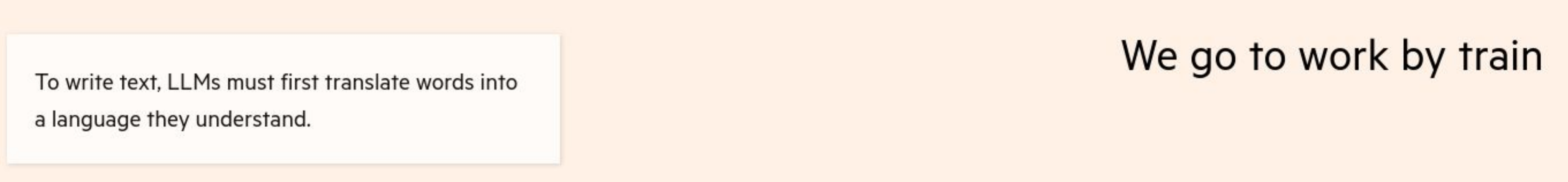
Generative AI exists because of the transformer

This is how it works

By **Visual Storytelling Team** and **Madhumita Murgia** in London SEPTEMBER 12 2023

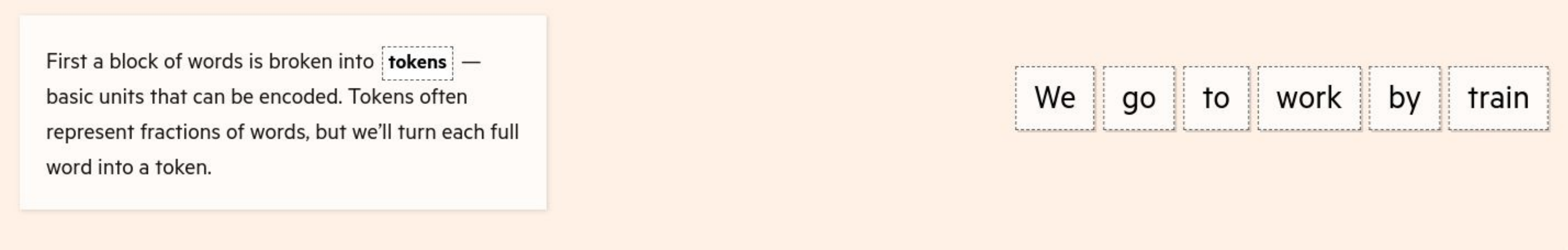
All of the visualizations included below (except where otherwise noted) are from this article

Word embeddings



To write text, LLMs must first translate words into a language they understand.

We go to work by train



First a block of words is broken into **tokens** — basic units that can be encoded. Tokens often represent fractions of words, but we'll turn each full word into a token.

Questions about LLMs

- What are word embeddings and how are they created?
  - Continuous Bag of Words is one method

    one-hot representation

    categorical, for ex   cat vs dog

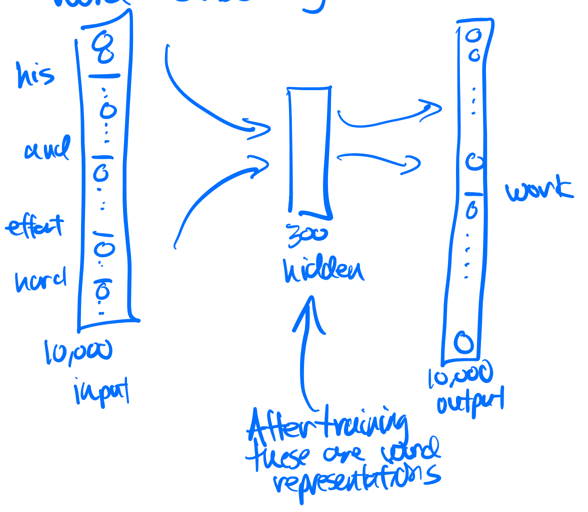
    10   01

    ex 10 digits
    1000000000 = zero
    0100000000 = one
    0010000000 = two
    etc.

Use only 10,000 most common English words
one particular location is set to 1 for every single word

Then we process many texts where the goal is to predict the center word from the surrounding words
"his hard work and effort"

word embedding network

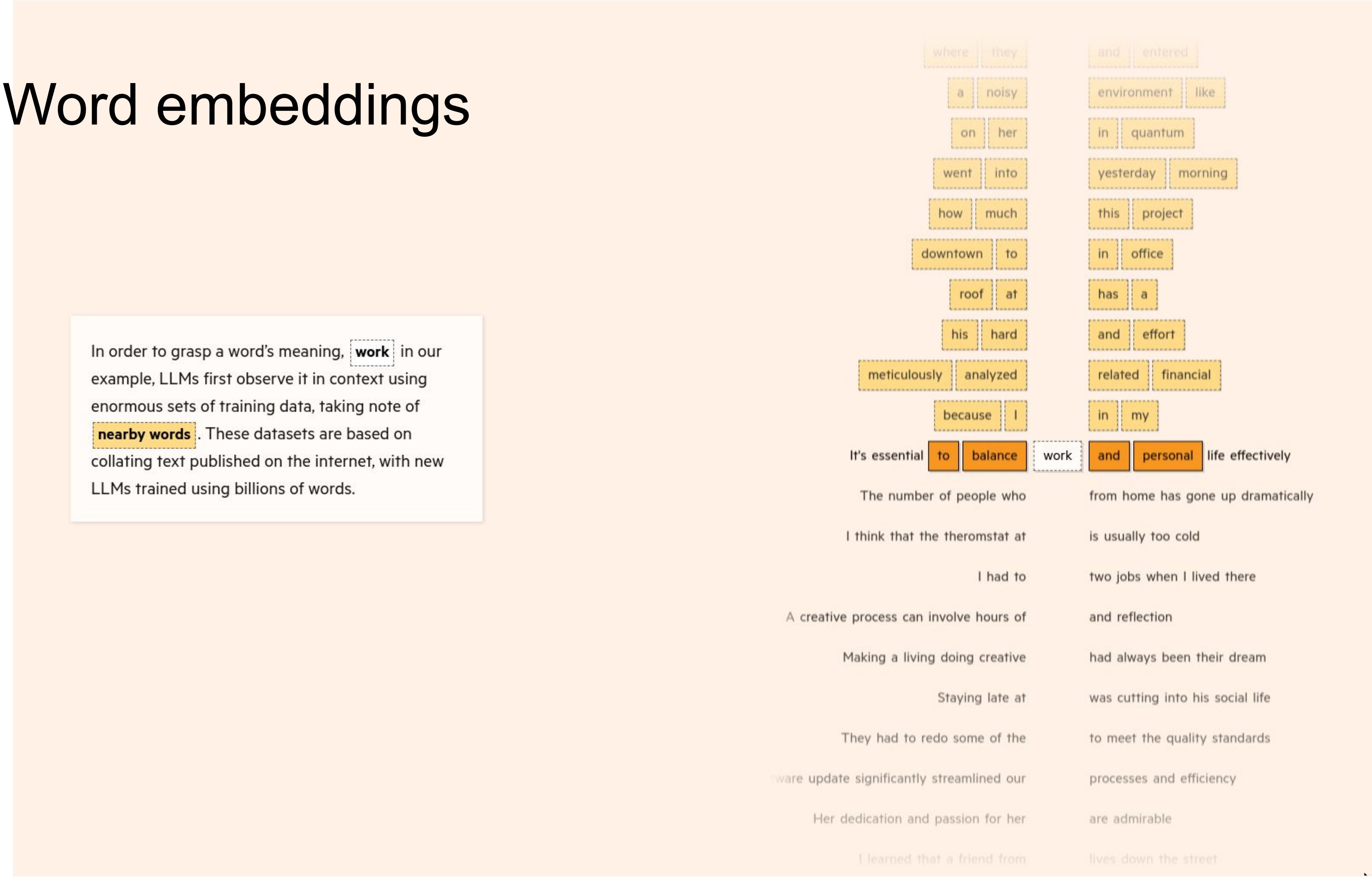


After training these are word representations

The network develops hidden layer representations that reflect the contexts in which the words occur.

Thus semantically similar words have similar hidden layer representations.

Word embeddings

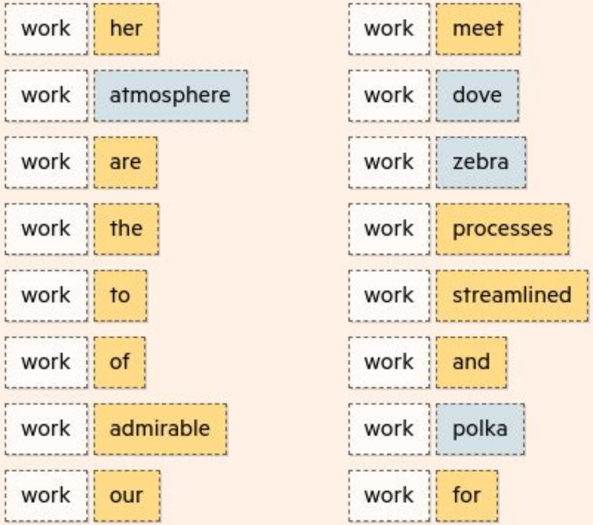


In order to grasp a word's meaning, **work** in our example, LLMs first observe it in context using enormous sets of training data, taking note of **nearby words**. These datasets are based on collating text published on the internet, with new LLMs trained using billions of words.

where they
and entered
a noisy
environment like
on her
in quantum
went into
yesterday morning
how much
this project
downtown to
in office
roof at
has a
his hard
and effort
meticulously analyzed
related financial
because I
in my
It's essential to balance
work
and personal life effectively
The number of people who
from home has gone up dramatically
I think that the theromstat at
is usually too cold
I had to
two jobs when I lived there
A creative process can involve hours of
and reflection
Making a living doing creative
had always been their dream
Staying late at
was cutting into his social life
They had to redo some of the
to meet the quality standards
ware update significantly streamlined our
processes and efficiency
Her dedication and passion for her
are admirable
I learned that a friend from
lives down the street

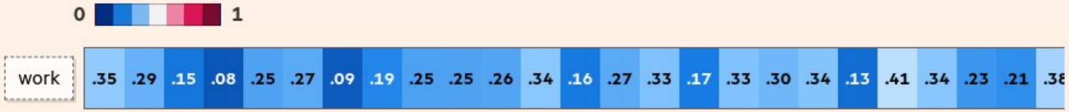
Word embeddings

Eventually, we end up with a huge set of the words found **alongside** **work** in the training data, as well as those that **weren't** found near it.



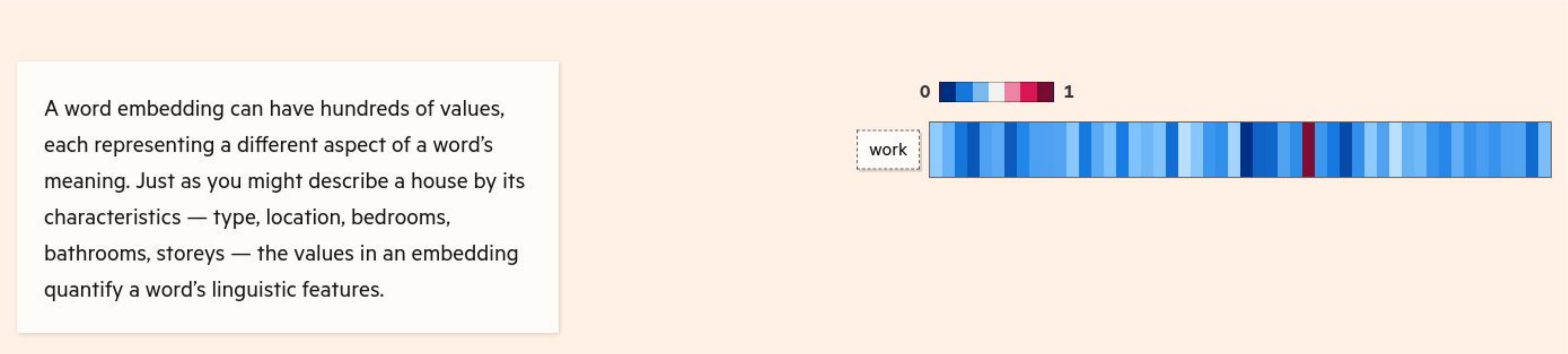
work	her	work	meet
work	atmosphere	work	dove
work	are	work	zebra
work	the	work	processes
work	to	work	streamlined
work	of	work	and
work	admirable	work	polka
work	our	work	for

As the model **processes** this set of words, it produces a vector — or list of values — and adjusts it based on each word's proximity to **work** in the training data. This vector is known as a word embedding.



work	.35	.29	.15	.08	.25	.27	.09	.19	.25	.25	.26	.34	.16	.27	.33	.17	.33	.30	.34	.13	.41	.34	.23	.21	.38

Word embeddings

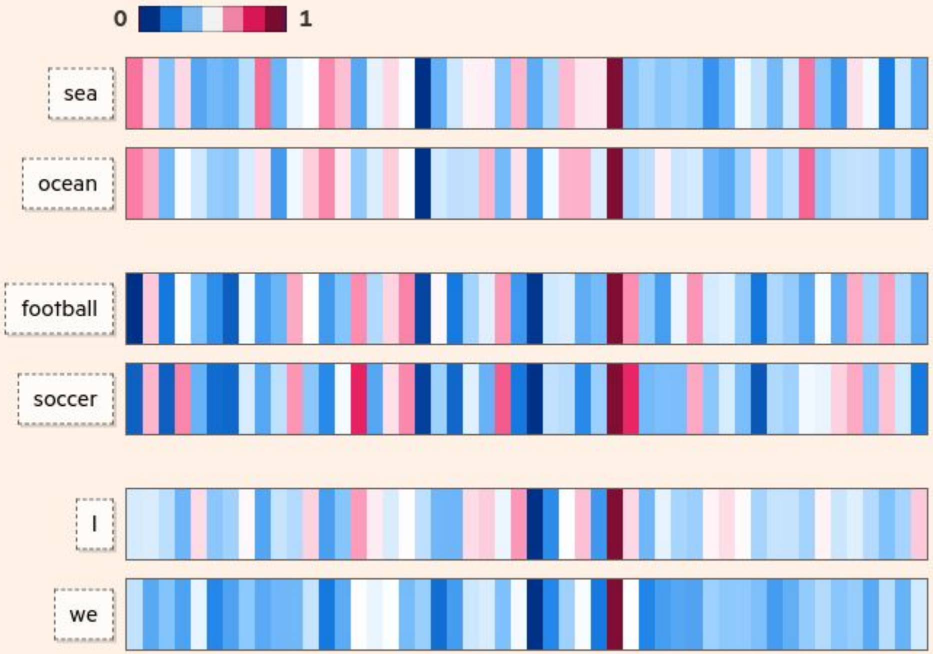


A word embedding can have hundreds of values, each representing a different aspect of a word's meaning. Just as you might describe a house by its characteristics — type, location, bedrooms, bathrooms, storeys — the values in an embedding quantify a word's linguistic features.

Word embeddings

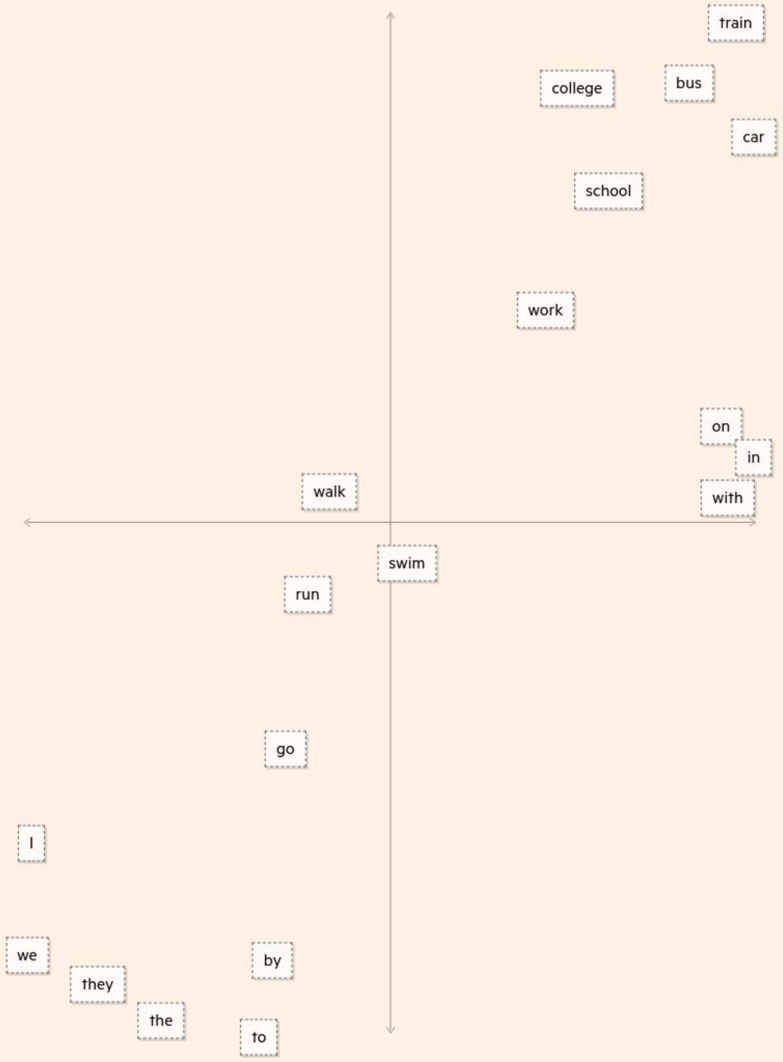
The way these characteristics are derived means we don't know exactly what each value represents, but words we expect to be used in comparable ways often have similar-looking embeddings.

A pair of words like **sea** and **ocean**, for example, may not be used in identical contexts ('all at ocean' isn't a direct substitute for 'all at sea'), but their meanings are close to each other, and embeddings allow us to quantify that closeness.

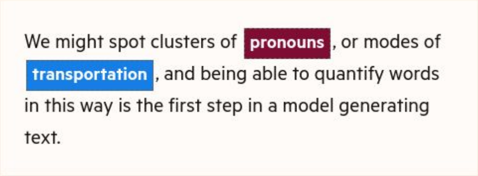


Word embeddings

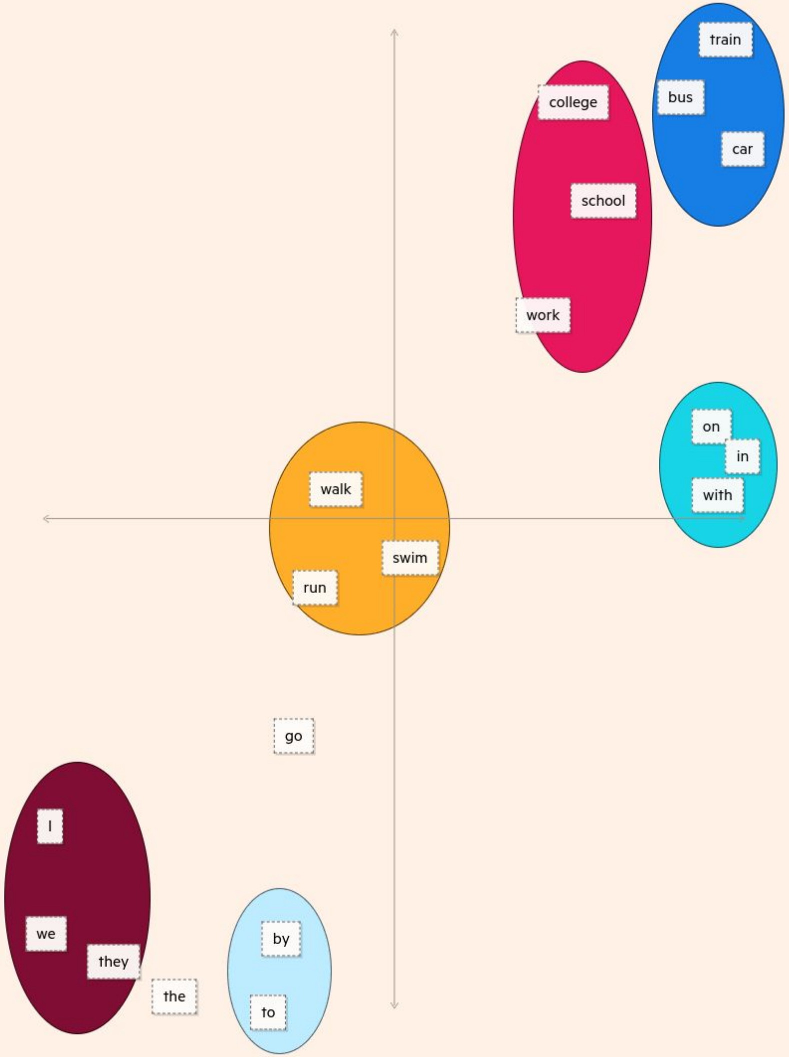
By reducing the hundreds of values each embedding represents to just two, we can see the distances between these words more clearly.



Word embeddings



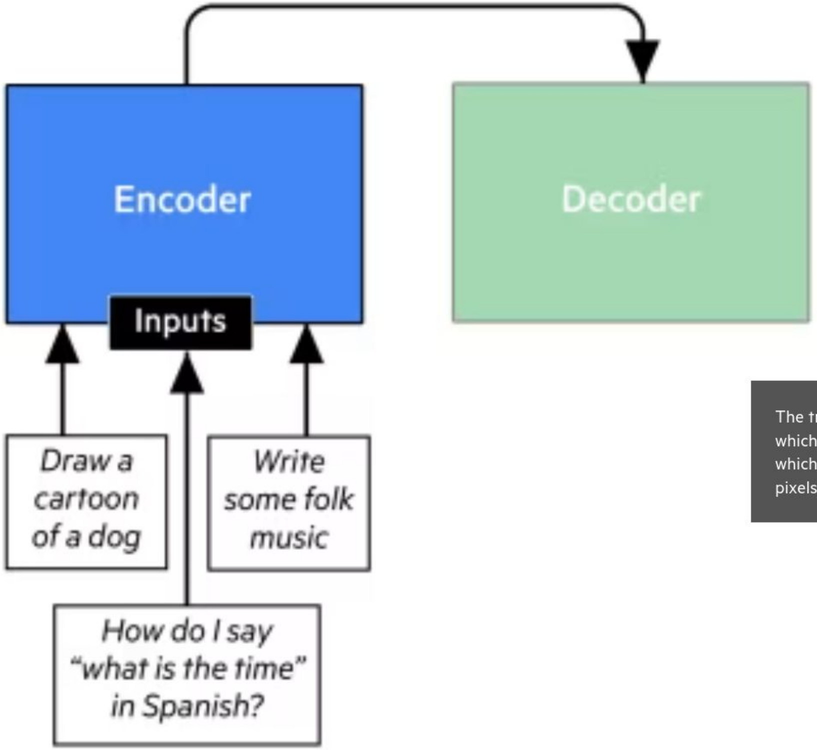
We might spot clusters of **pronouns**, or modes of **transportation**, and being able to quantify words in this way is the first step in a model generating text.



Let’s learn about how LLMs generate text

- Transformers
- Self-attention
- These concepts can be applied to more than just text—it also works to produce pictures, music, and computer code

Transformers



The transformer is divided into two main parts — the encoder, which processes and learns to understand an input sequence, which could be any repeating pattern (words, musical notes, pixels).

Image the Financial Times article entitled "Transformers: The Google scientists who pioneered an AI revolution"

Transformers

And the decoder, which produces an output sequence (a sentence, a piece of music, a picture).

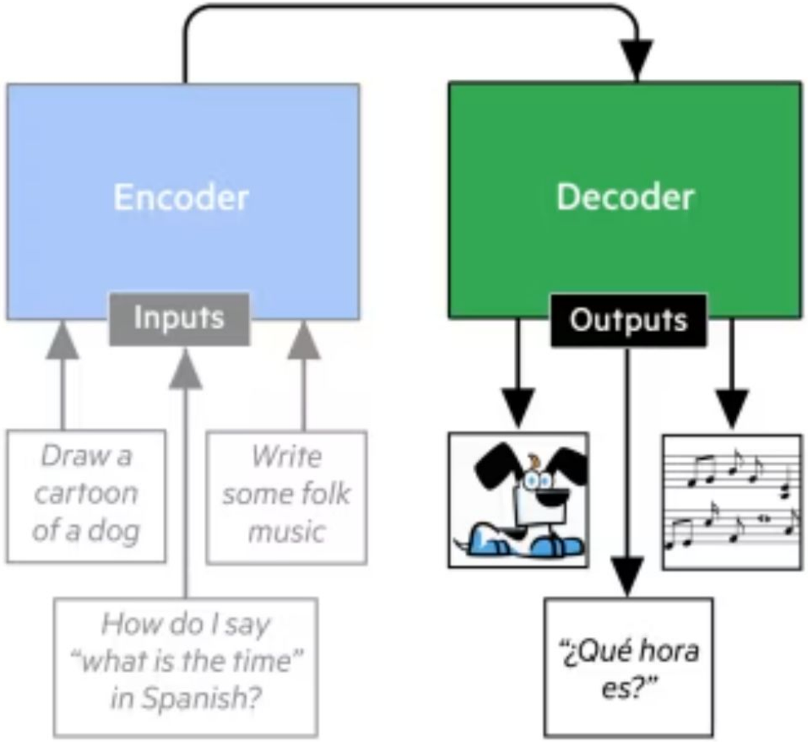


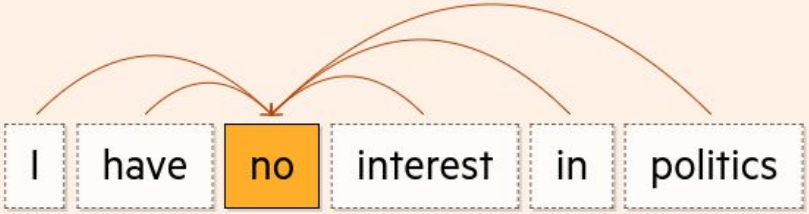
Image from the Financial Times article entitled “Transformers: The Google scientists who pioneered an AI revolution”

Transformers and self-attention

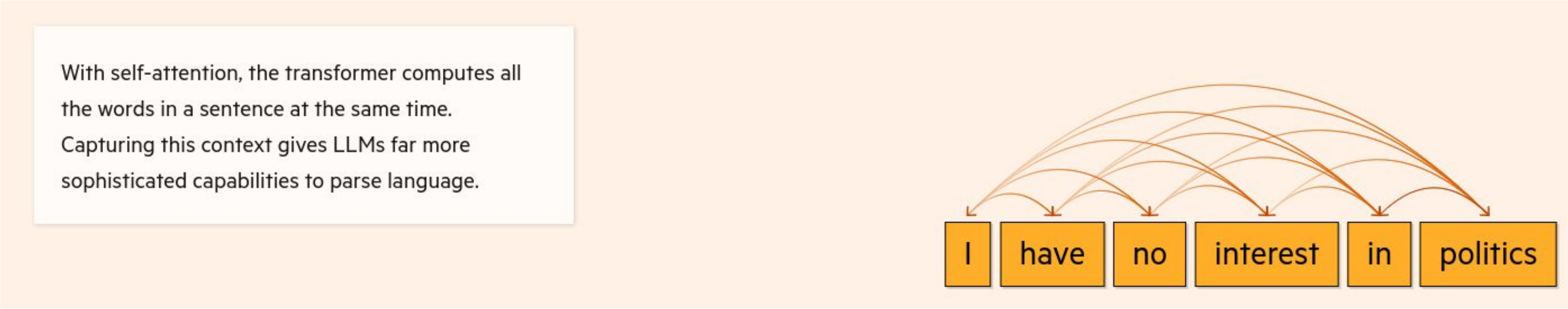
A key concept of the transformer architecture is self-attention. This is what allows LLMs to understand relationships between words.

I have no interest in politics

Self-attention looks at each token in a body of text and decides which others are most important to understanding its meaning.

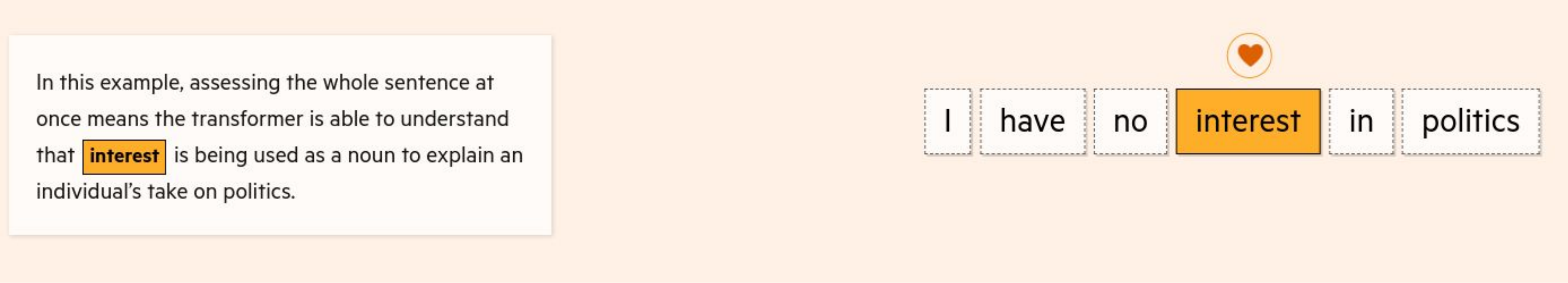


Transformers and self-attention



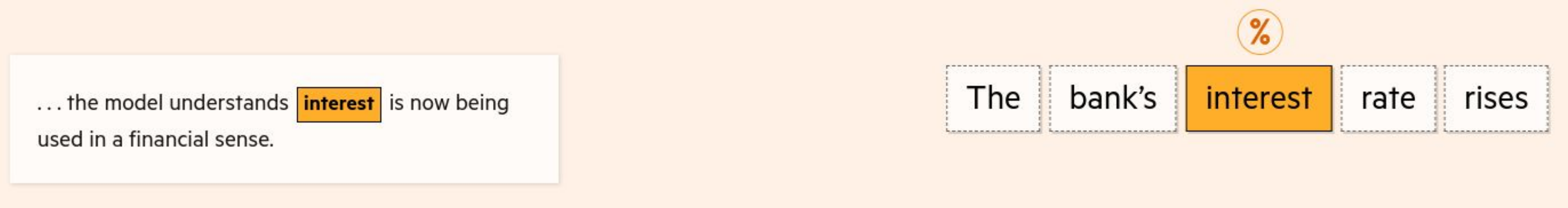
With self-attention, the transformer computes all the words in a sentence at the same time. Capturing this context gives LLMs far more sophisticated capabilities to parse language.

Transformers and self-attention



In this example, assessing the whole sentence at once means the transformer is able to understand that **interest** is being used as a noun to explain an individual's take on politics.

I have no interest in politics



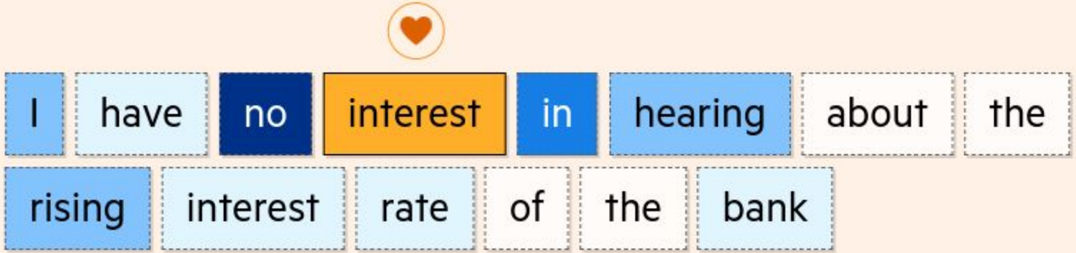
. . . the model understands **interest** is now being used in a financial sense.

The bank's interest rate rises

Transformers and self-attention

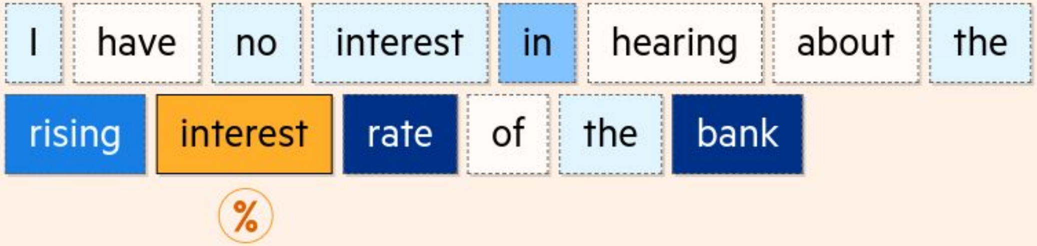
And when we combine the sentences, the model is still able to recognise the correct meaning of each word thanks to the attention it gives the accompanying text.

For the first use of interest, it is no and in that are most attended.



I have no interest in hearing about the rising interest rate of the bank

For the second, it is rate and bank.



I have no interest in hearing about the rising interest rate of the bank

Language challenge for AI

- Binary choice questions
- Two entities are mentioned in the question
- A pronoun is used to refer to one of them
- What is the correct referent of the pronoun?
- Changing just one word in the statement, changes the correct referent
- How will LLMs do at this task?

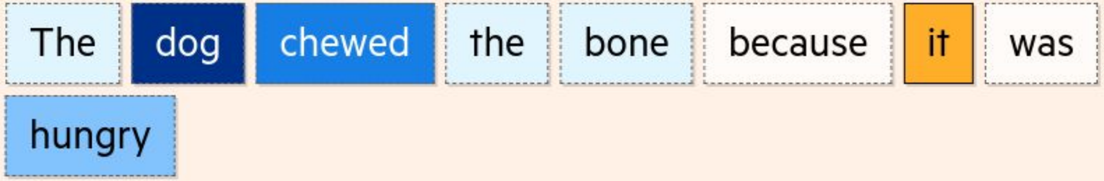
For example

- The dog chewed the bone because it was hungry.
- The dog chewed the bone because it was delicious.

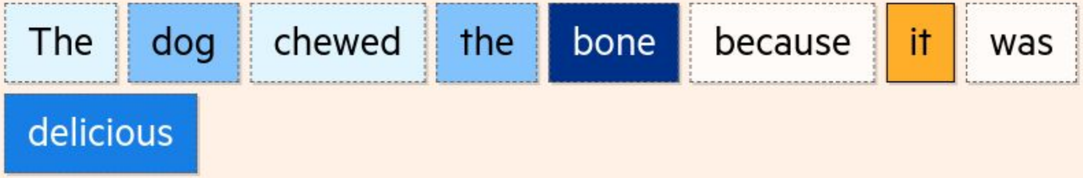
What does the pronoun “it” refer to in each of these sentences?

How self-attention helps focus on the correct words

In the following sentence, self-attention is able to calculate that **it** is most likely to be referring to **dog**.



And if we alter the sentence, swapping **hungry** for **delicious**, the model is able to recalculate, with **it** now most likely to refer to **bone**.

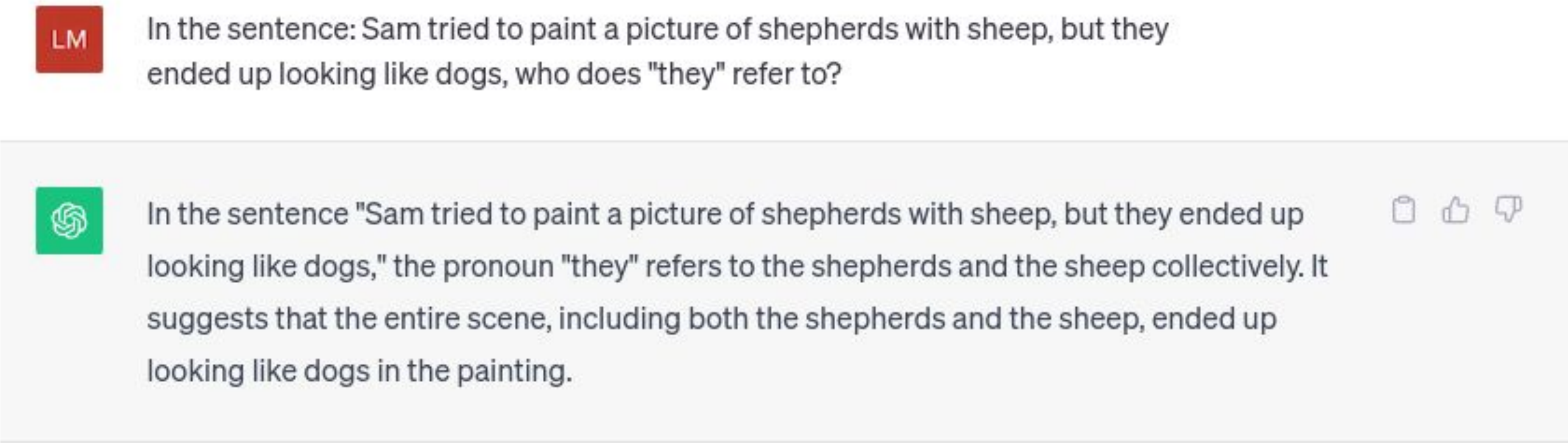


Another example

- Sam painted a picture of shepherds with sheep, but they ended up looking like dogs.
- Sam painted a picture of shepherds with sheep, but they ended up looking like golfers.

What does the pronoun “they” refer to in each of these sentences?

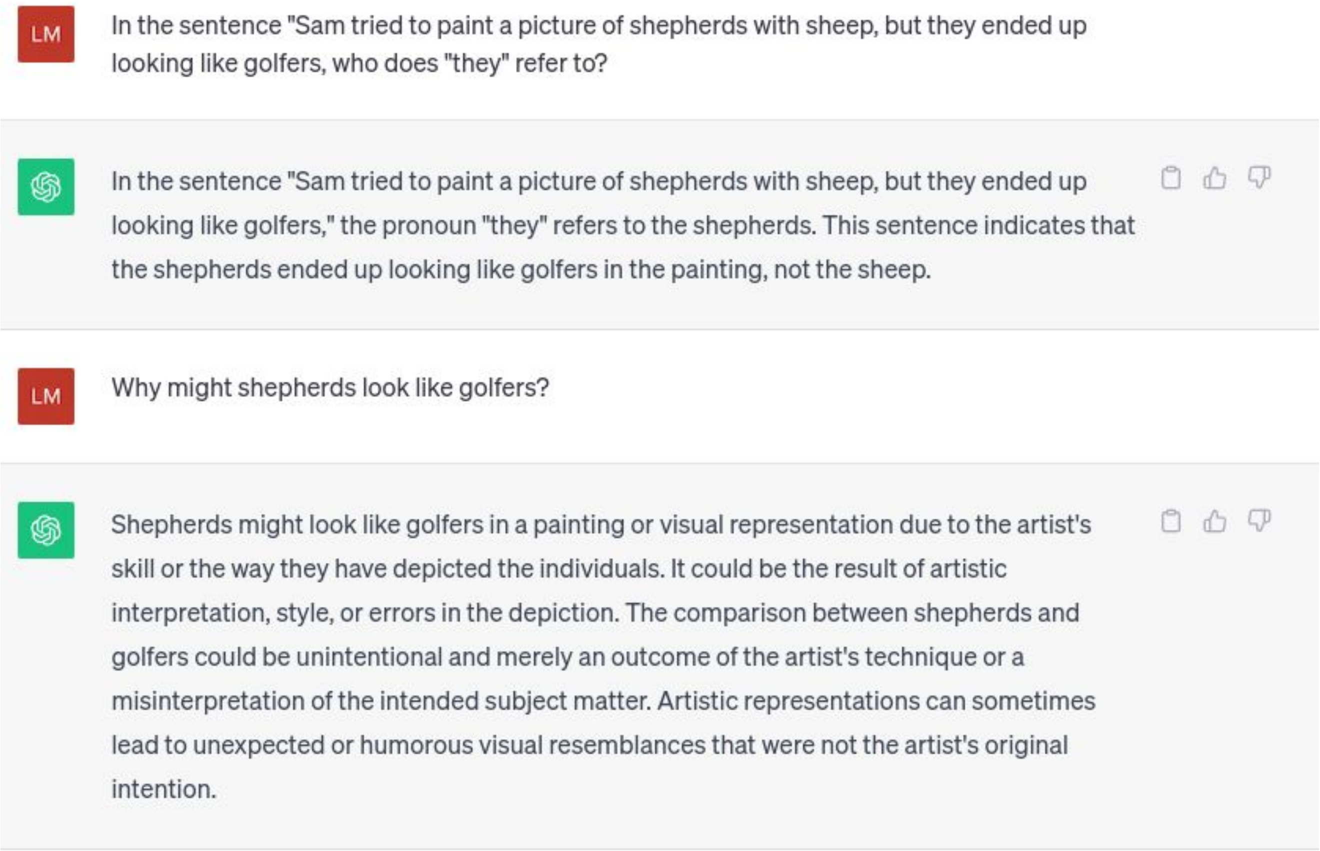
ChatGPT doesn’t fare as well in this example



In the sentence: Sam tried to paint a picture of shepherds with sheep, but they ended up looking like dogs, who does "they" refer to?

In the sentence "Sam tried to paint a picture of shepherds with sheep, but they ended up looking like dogs," the pronoun "they" refers to the shepherds and the sheep collectively. It suggests that the entire scene, including both the shepherds and the sheep, ended up looking like dogs in the painting.

Here it suggests the correct referent, but it’s explanation is disappointing



In the sentence "Sam tried to paint a picture of shepherds with sheep, but they ended up looking like golfers, who does "they" refer to?

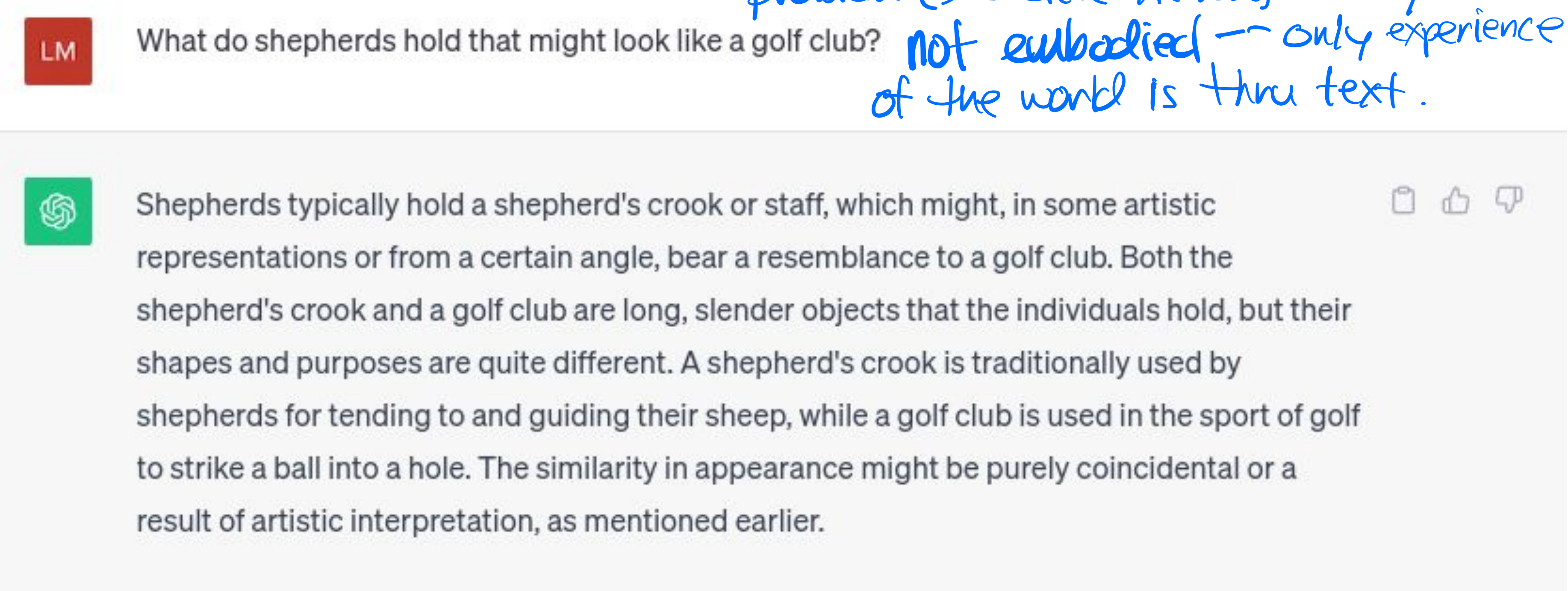
In the sentence "Sam tried to paint a picture of shepherds with sheep, but they ended up looking like golfers," the pronoun "they" refers to the shepherds. This sentence indicates that the shepherds ended up looking like golfers in the painting, not the sheep.

Why might shepherds look like golfers?

Shepherds might look like golfers in a painting or visual representation due to the artist's skill or the way they have depicted the individuals. It could be the result of artistic interpretation, style, or errors in the depiction. The comparison between shepherds and golfers could be unintentional and merely an outcome of the artist's technique or a misinterpretation of the intended subject matter. Artistic representations can sometimes lead to unexpected or humorous visual resemblances that were not the artist's original intention.

Though the desired information does seem to be present in the model

LLMs face the symbol grounding problem (see Steve Harnad). They are not embodied -- only experience of the world is thru text.

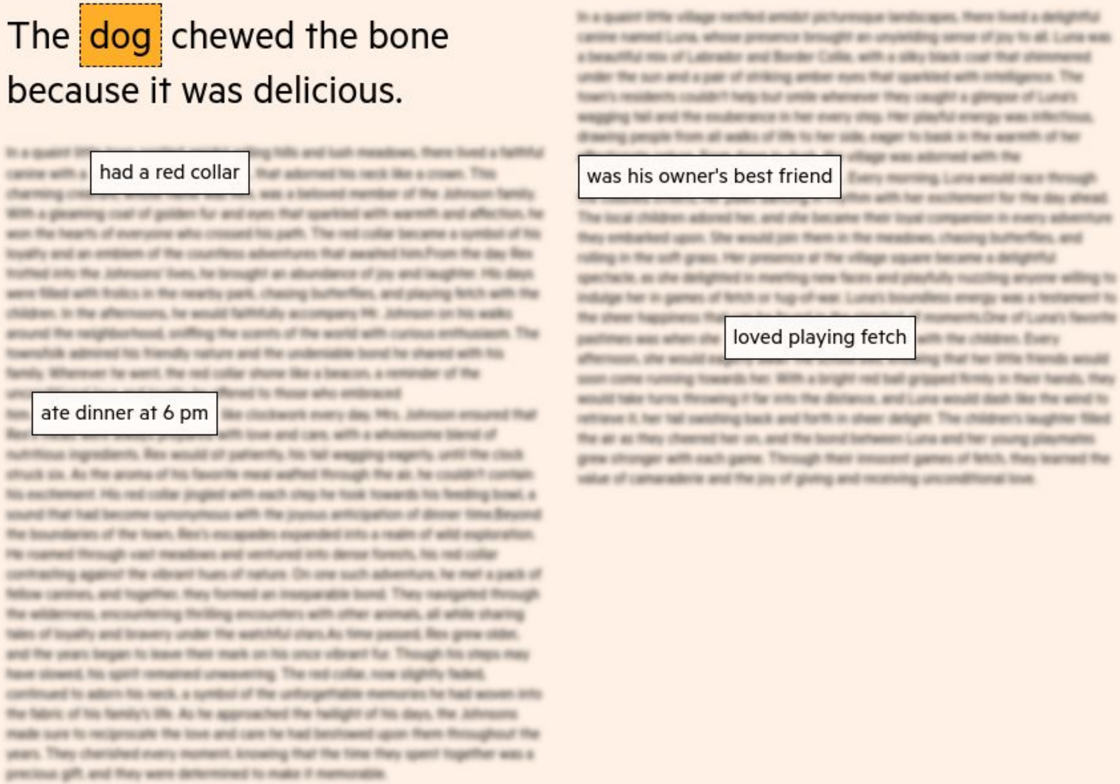


What do shepherds hold that might look like a golf club?

Shepherds typically hold a shepherd's crook or staff, which might, in some artistic representations or from a certain angle, bear a resemblance to a golf club. Both the shepherd's crook and a golf club are long, slender objects that the individuals hold, but their shapes and purposes are quite different. A shepherd's crook is traditionally used by shepherds for tending to and guiding their sheep, while a golf club is used in the sport of golf to strike a ball into a hole. The similarity in appearance might be purely coincidental or a result of artistic interpretation, as mentioned earlier.

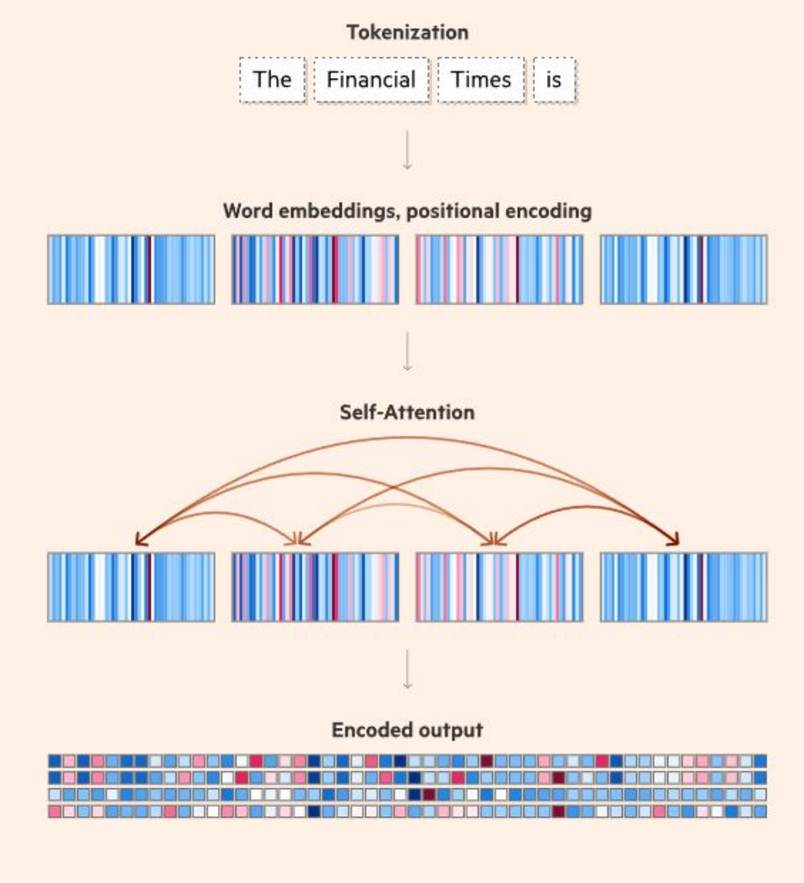
Transformers and self-attention

The benefits of self-attention for language processing increase the more you scale things up. It allows LLMs to take **context** from beyond sentence boundaries, giving the model a greater understanding of how and when a word is used.



The dog chewed the bone because it was delicious.

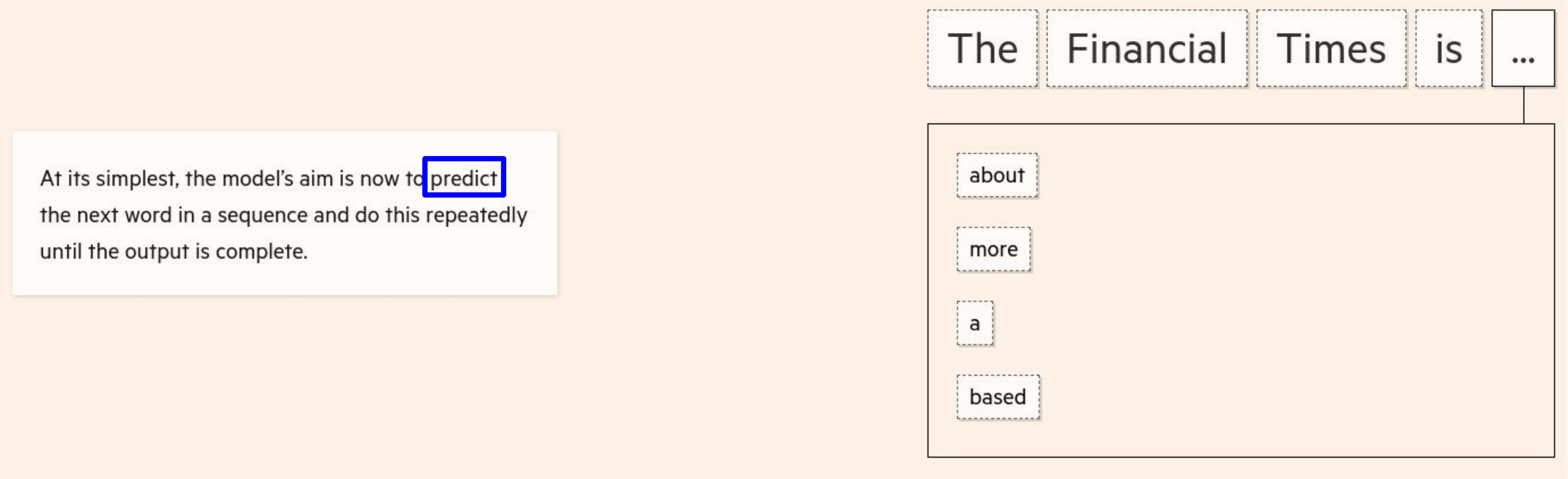
Generating text



Generating text

stochastic vs deterministic

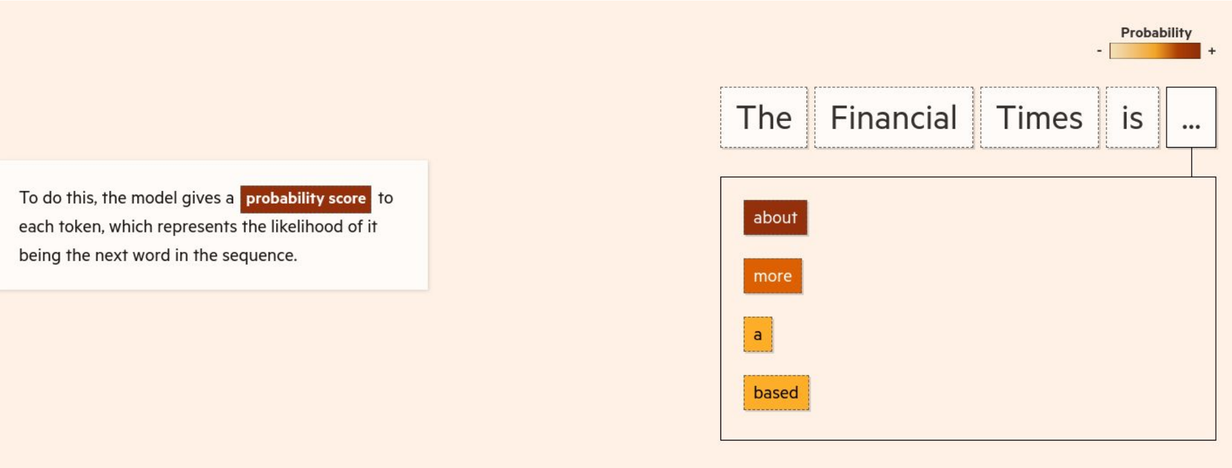
LLMs as "stochastic parrots"



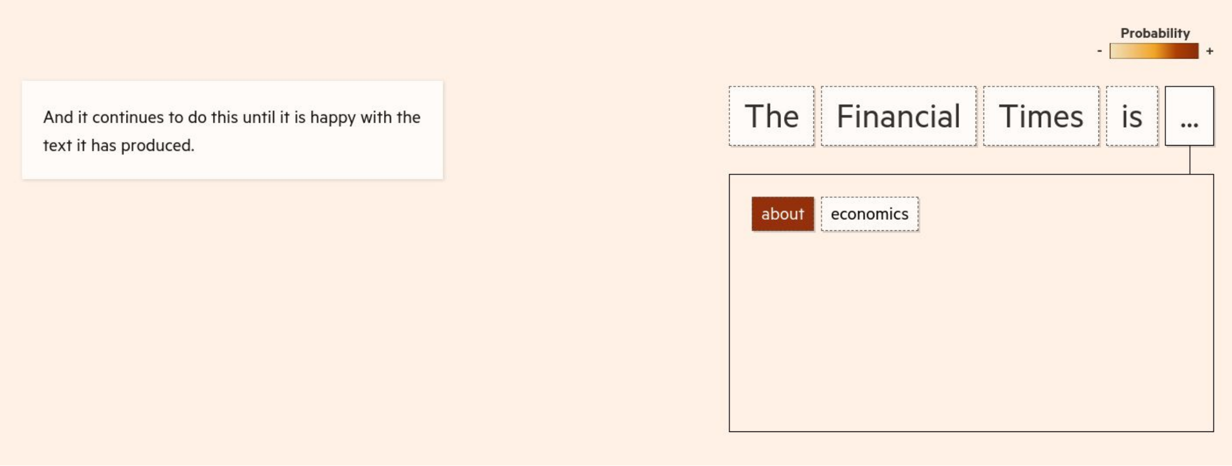
At its simplest, the model's aim is now to predict the next word in a sequence and do this repeatedly until the output is complete.

Paper: "On the dangers of stochastic parrots: Can large language models be too big?"

Generating text

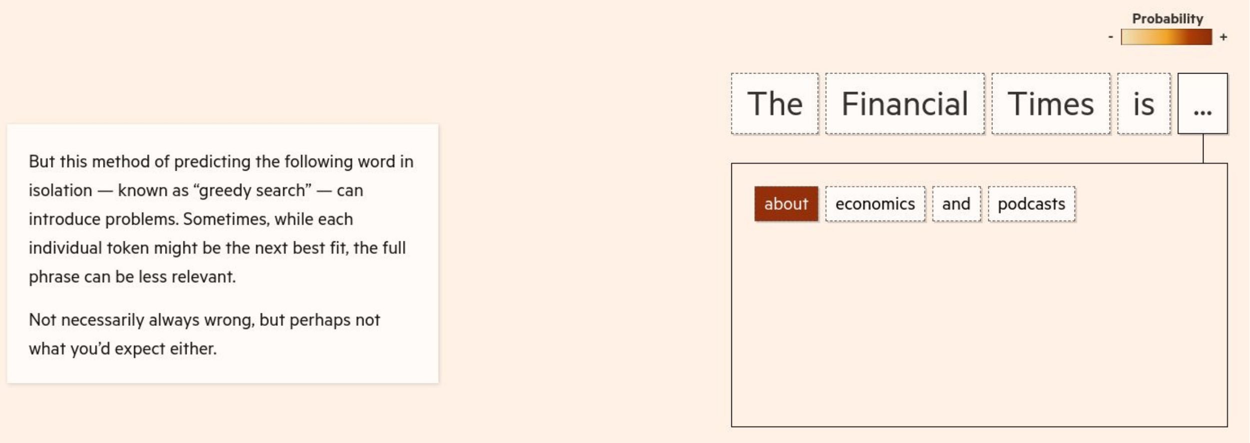


To do this, the model gives a **probability score** to each token, which represents the likelihood of it being the next word in the sequence.

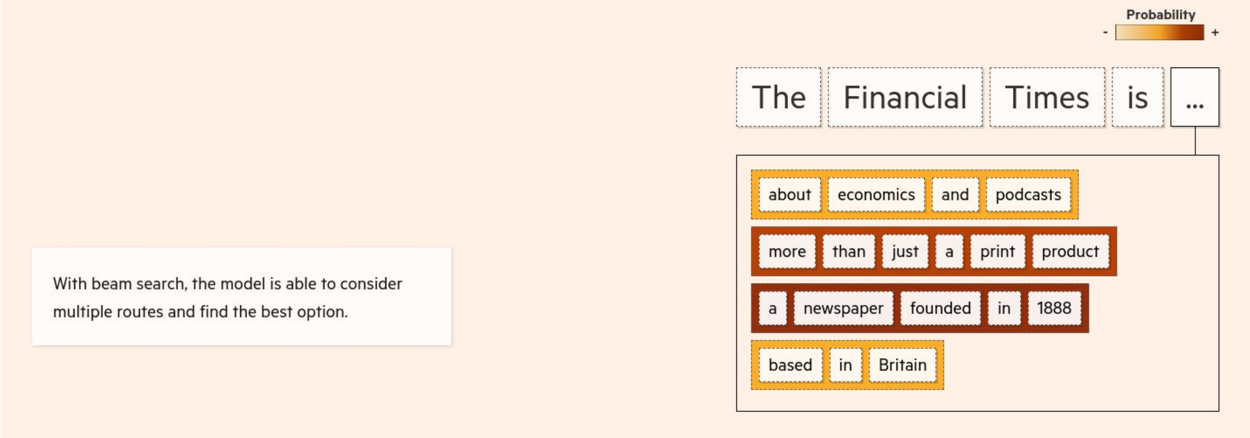


And it continues to do this until it is happy with the text it has produced.

Generating text

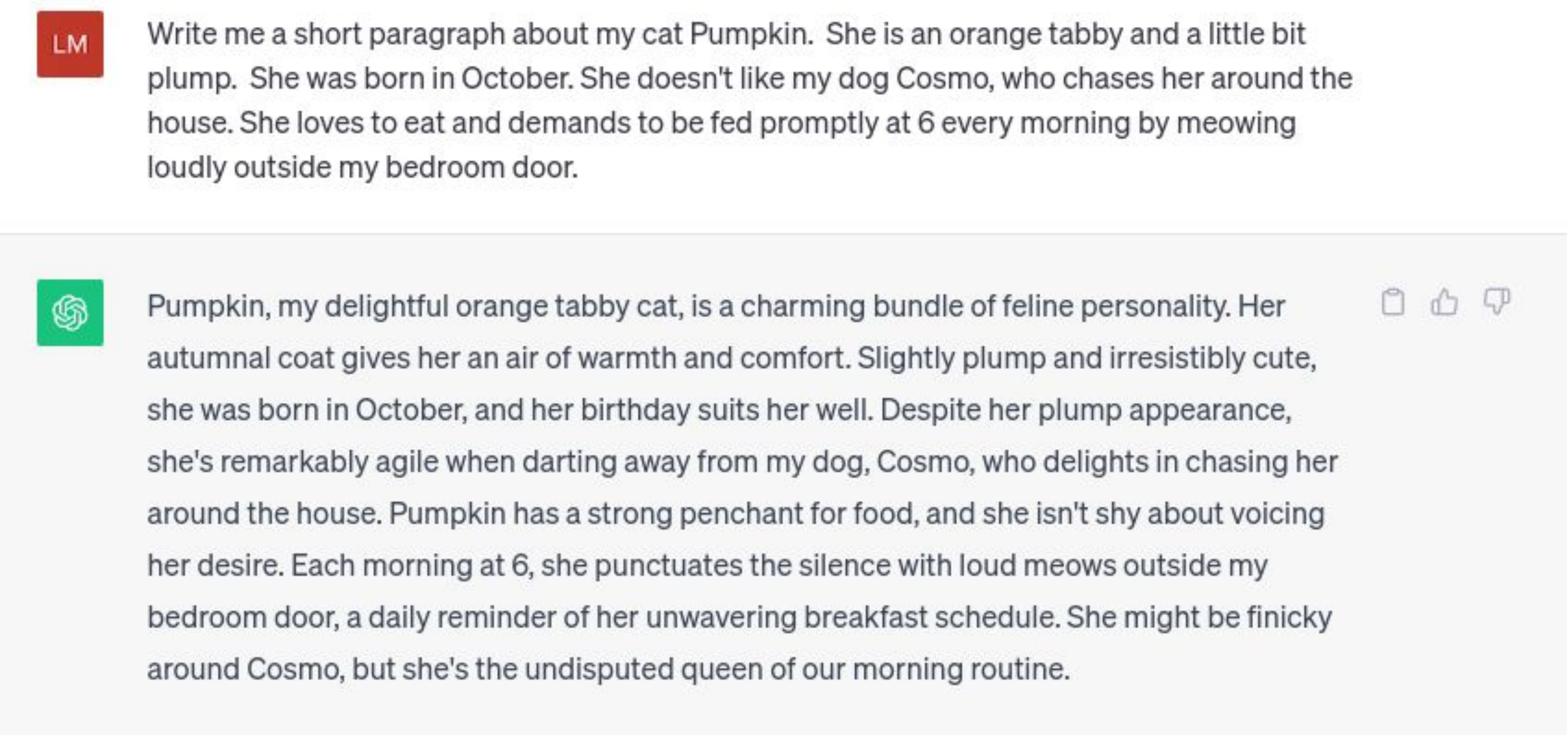

But this method of predicting the following word in isolation — known as “greedy search” — can introduce problems. Sometimes, while each individual token might be the next best fit, the full phrase can be less relevant.

Not necessarily always wrong, but perhaps not what you’d expect either.


With beam search, the model is able to consider multiple routes and find the best option.

Probability
-
+
The
Financial
Times
is
...
about economics and podcasts
more than just a print product
a newspaper founded in 1888
based in Britain

Generating a paragraph of text



Write me a short paragraph about my cat Pumpkin. She is an orange tabby and a little bit plump. She was born in October. She doesn't like my dog Cosmo, who chases her around the house. She loves to eat and demands to be fed promptly at 6 every morning by meowing loudly outside my bedroom door.

Pumpkin, my delightful orange tabby cat, is a charming bundle of feline personality. Her autumnal coat gives her an air of warmth and comfort. Slightly plump and irresistibly cute, she was born in October, and her birthday suits her well. Despite her plump appearance, she's remarkably agile when darting away from my dog, Cosmo, who delights in chasing her around the house. Pumpkin has a strong penchant for food, and she isn't shy about voicing her desire. Each morning at 6, she punctuates the silence with loud meows outside my bedroom door, a daily reminder of her unwavering breakfast schedule. She might be finicky around Cosmo, but she's the undisputed queen of our morning routine.

What makes the Connections game challenging for LLMs?

- The player is only provided with 16 words without any context
- The transformer and self-attention mechanisms need context to help disambiguate the words and determine their intended meaning

Summarizing capabilities of LLMs like ChatGPT

- They are not search engines—they are instead pattern matching engines
- They write text that is reasonably coherent and plausible, but there are no guarantees that it is factually correct
- They often fabricate information, generating made up names, dates, article titles, and quotes
- They have been trained on text scraped from the internet, and thus absorb the cultural biases inherent within that text