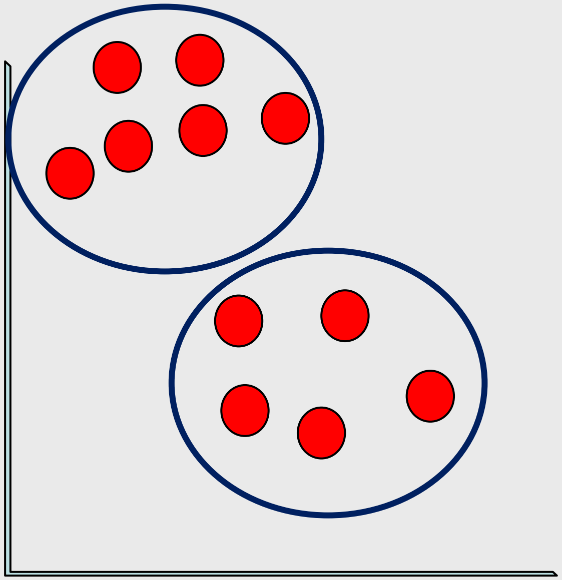
Transformers

Geometry of Data Points

- The geometry of these data points consists of two clusters



Understanding Text

- **The texts form two clusters**

Document
Time to tell the story of Tesla & SpaceX
Asteroid impact risk is well understood, but not comets. Those worry me.
Yesterday, I did the most important thing I can do to support @BarackObama - I voted
The President has been steady on the issues
A favorite Obama family recipe is up for a vote

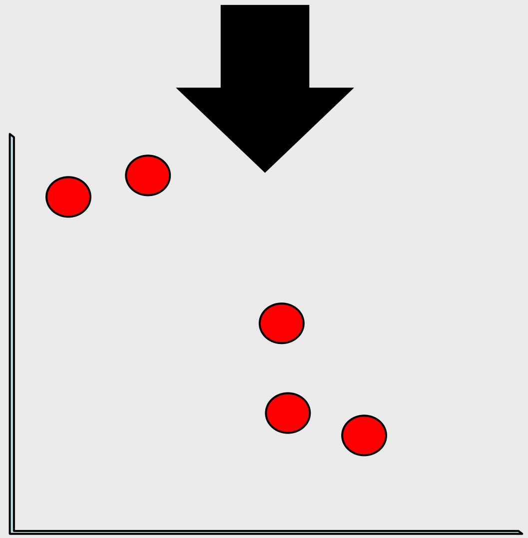
Geometry of Text

- Text documents in clusters have words about similar topics
- Data points in clusters are close to each other
- Can we turn the notion of similar into a notion of close for text?
- Can we give geometry to text?

Embeddings

- **If we embed text, we turn text into a numeric data point**
- **A good embedding will encode meaning of the text in the geometry of the data points**
- **With a good embedding, we can do a lot of useful things with text**

Document
Time to tell the story of Tesla & SpaceX
Asteroid impact risk is well understood, but not comets. Those worry me.
Yesterday, I did the most important thing I can do to support @BarackObama - I voted
The President has been steady on the issues
A favorite Obama family recipe is up for a vote



Embedding Text

- We are given a corpus of text documents

Document	Text
1	The mouse is in the yard
2	The garden snake is in The Secret Garden!
3	The cat in the hat
4	The cat in the cradle

- How can we embed this text?

Term Frequency (TF) Embedding

- Each document can be represented as a set of terms with different frequencies (counts)

- Ex) The garden snake is in The Secret Garden!
- tf embedding:

Term	Frequency
The	2
garden	1
snake	1
is	1
in	1
Secret	1
Garden	1
!	1

Cleaning Text

- We like to clean the text before doing any analysis

- Common cleaning steps
  - Remove stop words: common words like “and”, “the”, etc.
  - Remove punctuation
  - Make all words lowercase

TF Embedding on Cleaned Text

- **Ex) The garden snake is in The Secret Garden!**

tf embedding

Term	Frequency
The	2
garden	1
snake	1
is	1
in	1
Secret	1
Garden	1
!	1

Clean tf embedding

Term	Frequency
garden	2
snake	1
secret	1

TF Embedding on Cleaned Text

Document	Text
1	The mouse is in the yard
2	The garden snake is in The Secret Garden!
3	The cat in the hat
4	The cat in the cradle

TF Embedding on Cleaned Text

Word	Doc 1	Doc 2	Doc 3	Doc 4
mouse	1	0	0	0
yard	1	0	0	0
garden	0	2	0	0
snake	0	1	0	0
secret	0	1	0	0
cat	0	0	1	1
hat	0	0	1	0
cradle	0	0	0	1

TF Embedding Properties

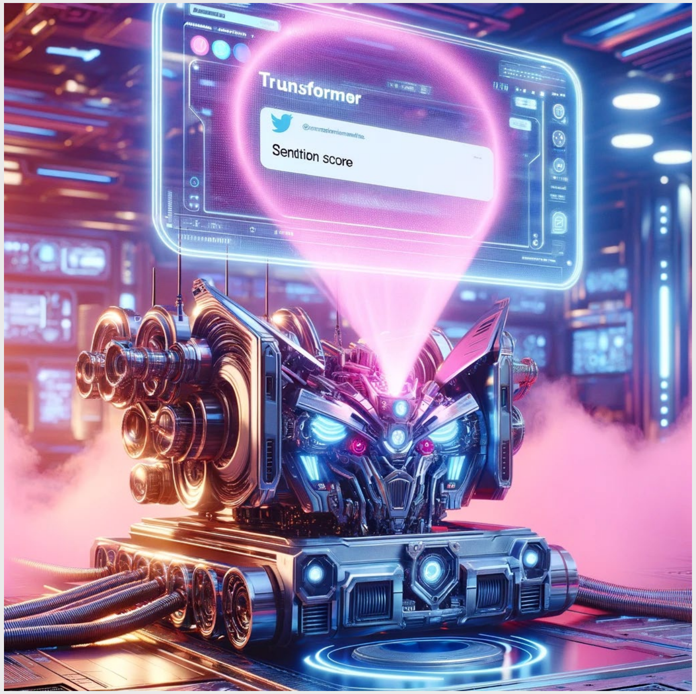
- Order of words doesn't matter
  - Tauhid eats fish = Fish eats Tauhid

- Word meaning is independent of its context
  - (baseball) bat = (vampire) bat

- Not the best embedding but a good start

Transformers

- Neural network developed in 2017 by Google
- Revolutionized natural language processing



Transformers

- **Neural network developed in 2017 by Google**
- **Revolutionized natural language processing**

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

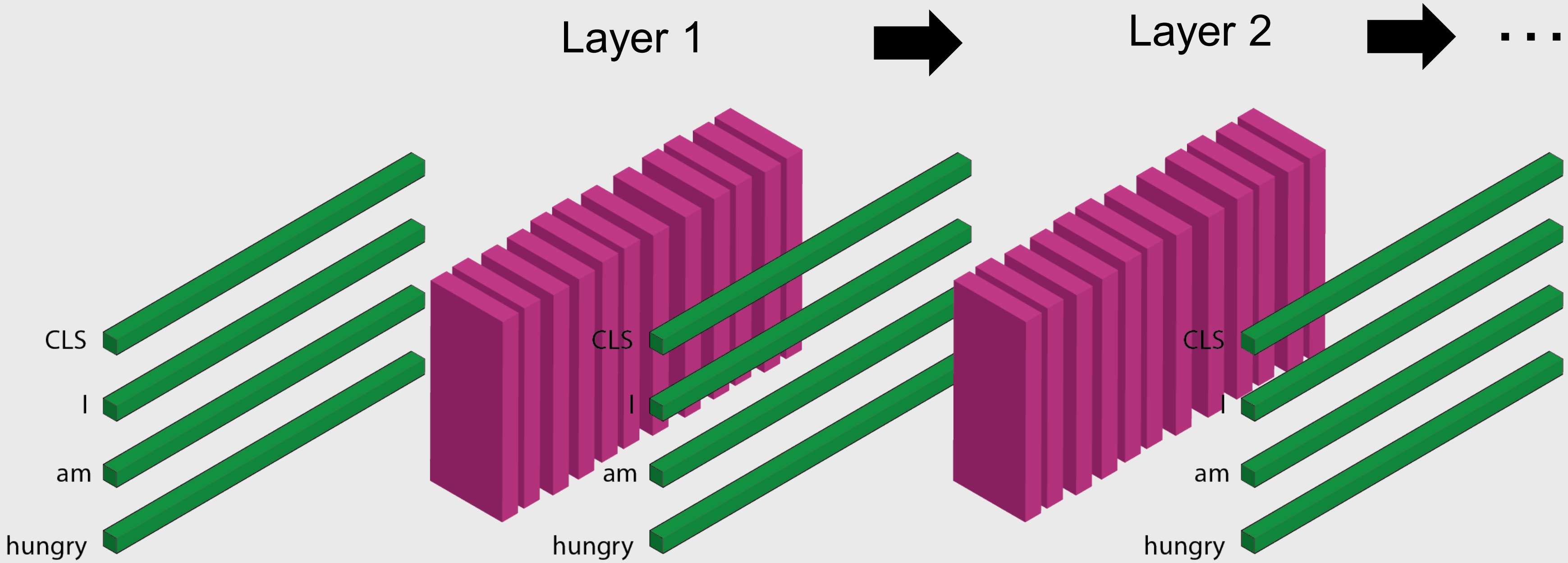
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention

What Can Transformers Do?

- Measure sentiment
- Translation
- Web search
- Text summarization
- Generate text
- Question answering
- Write Python code
- Be your friend ☺
- ANYTHING!!!!

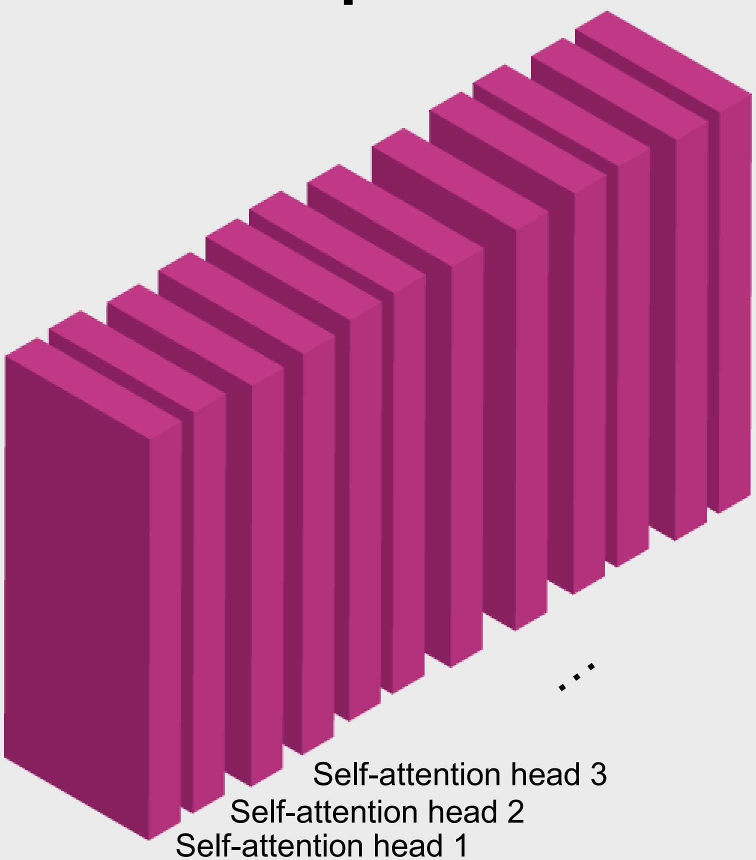
Transformer Architecture

- The transformer has many layers
- Each layer has an embedding vector for each word in the input text



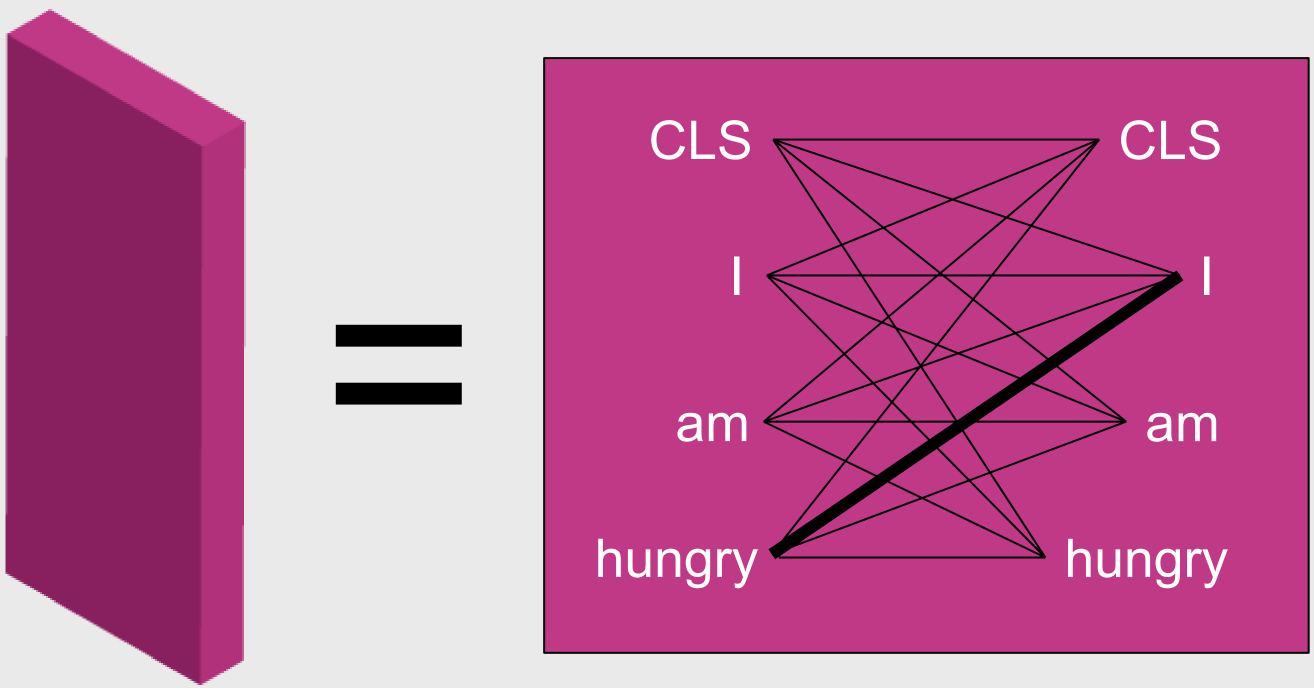
Transformer Layer

- Each layer has many self-attention heads
- Each attention head operates on a small chunk of the input vectors in parallel



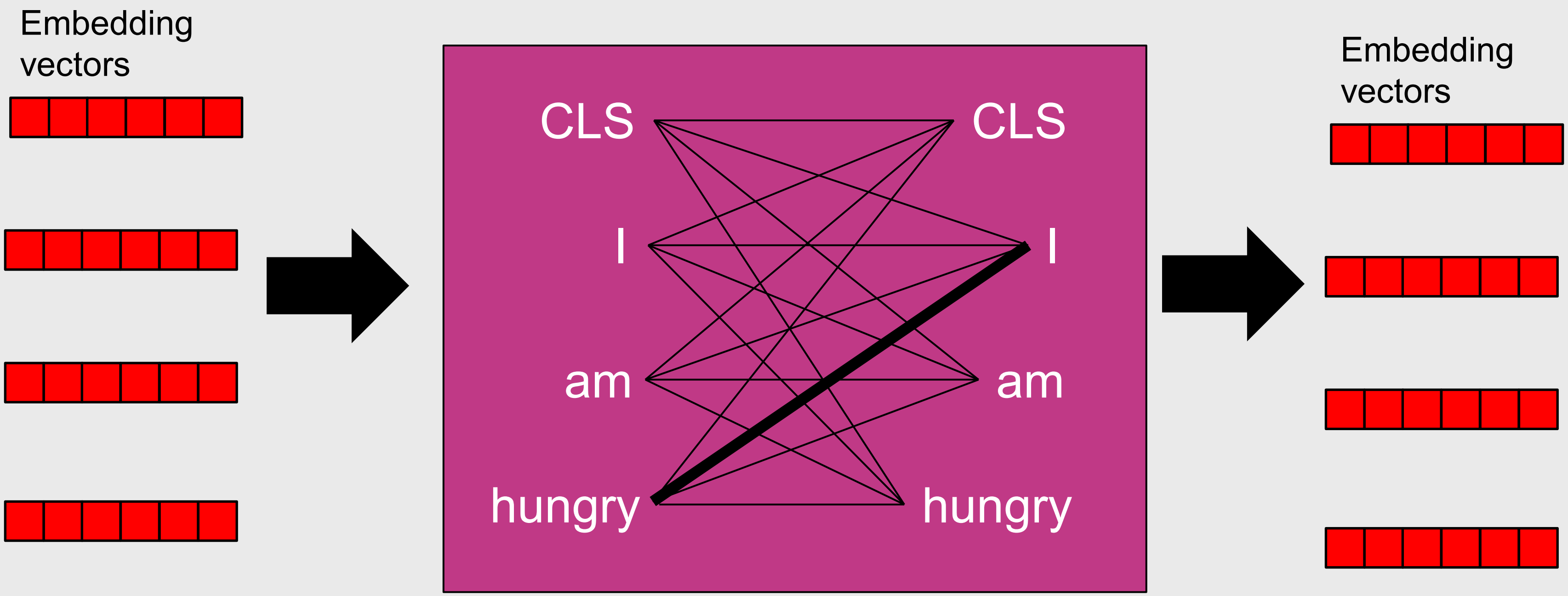
Self-Attention Head

- Each self-attention head computes weights from each word to each other word, creating an attention pattern
- This attention pattern represents some aspect of the language (subject, sentiment, etc)



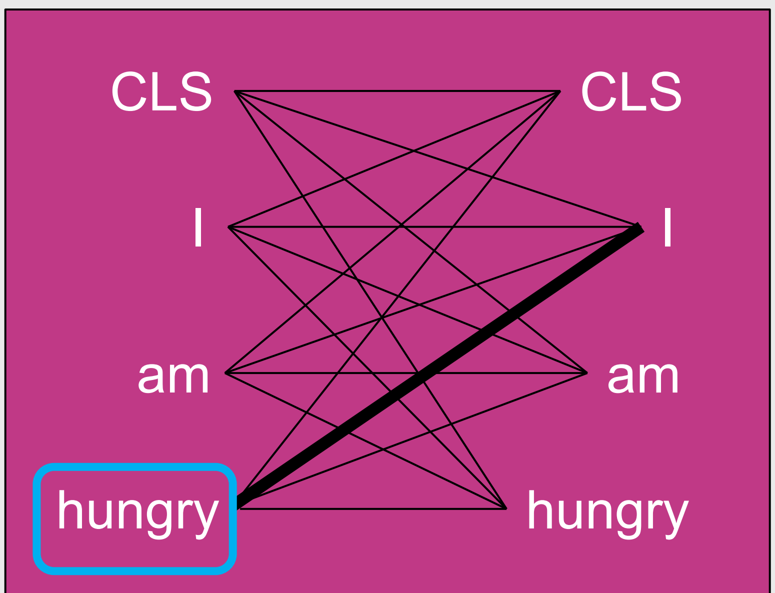
Self-Attention Head

- **Each self-attention head outputs a context dependent embedding vector for each word**



Self-Attention Head

- **Output embedding vector of a word is a weighted combination of input embedding vector of all words**
- **Weight is given by attention pattern**



hungry

Word	Vector	Weight	Weight x Vector
CLS		0.01	
I		0.9	
am		0.07	
hungry		0.02	
		Sum:	

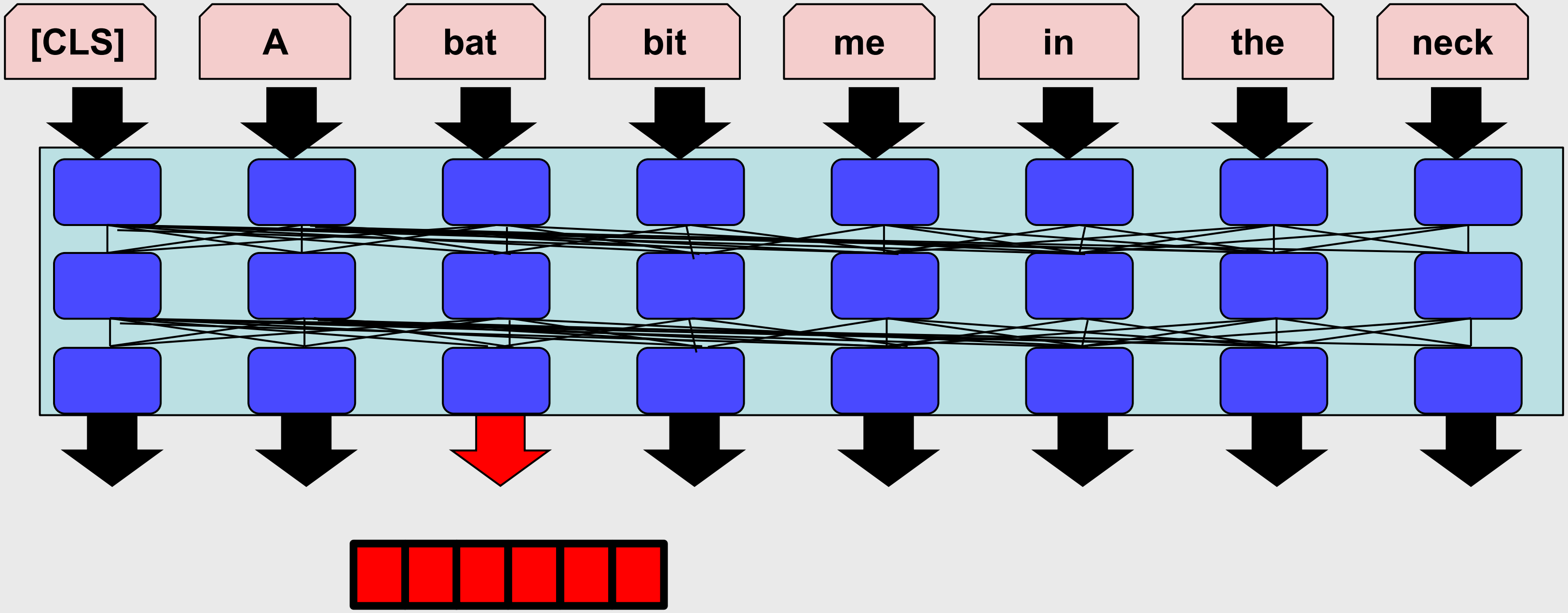
Context Dependent Embeddings

Consider these sentences that use the word bat

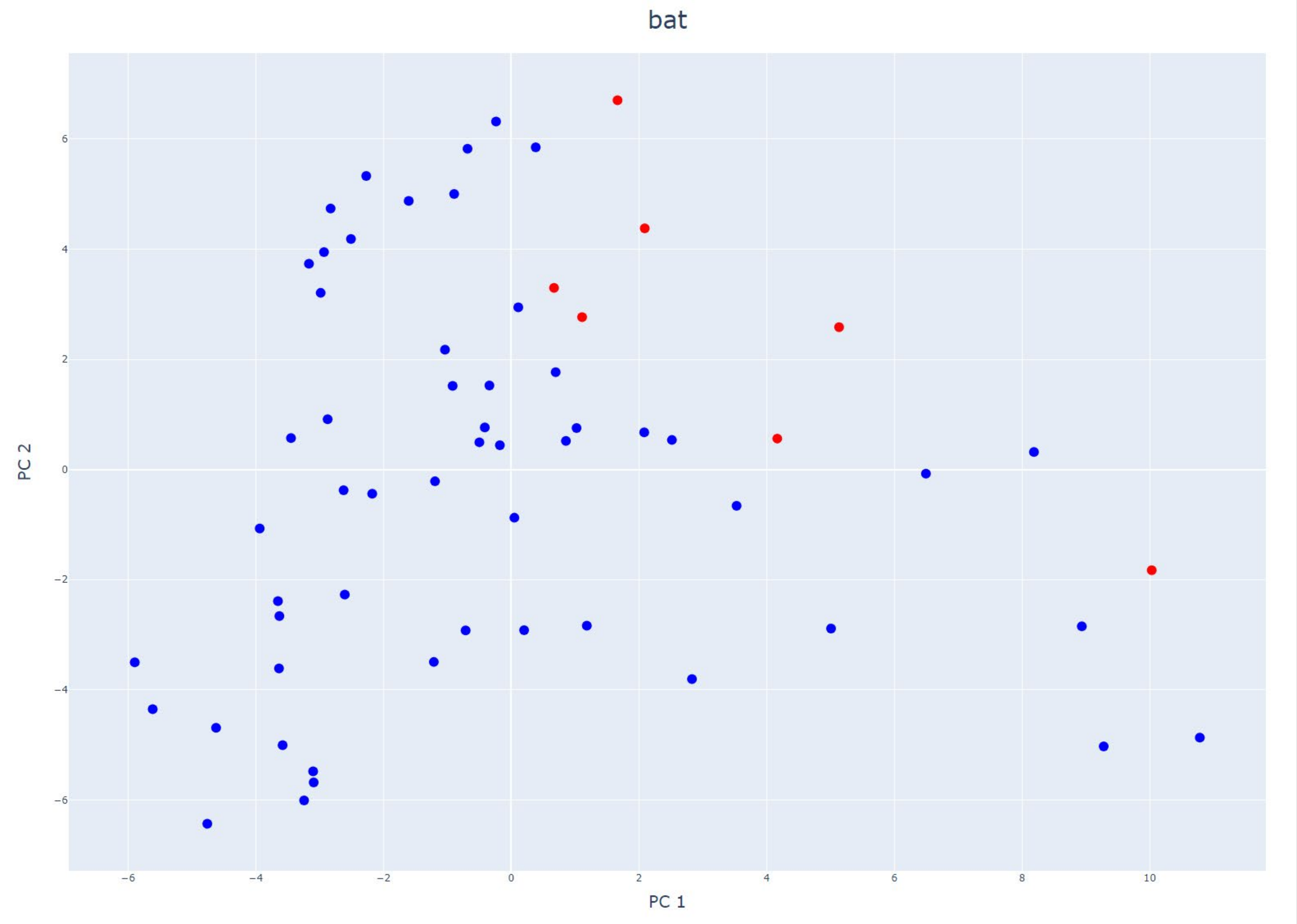
1. A bat flew out of the cave
2. The bat pooped on the ground
3. A bat bit me in the neck
4. Im afraid of a bat because it is like a rat with wings
5. A bat flew out of the baseball players hand
6. I hit a home run with the metal bat
7. No one swung a bat harder than Babe Ruth

Context Dependent Embeddings

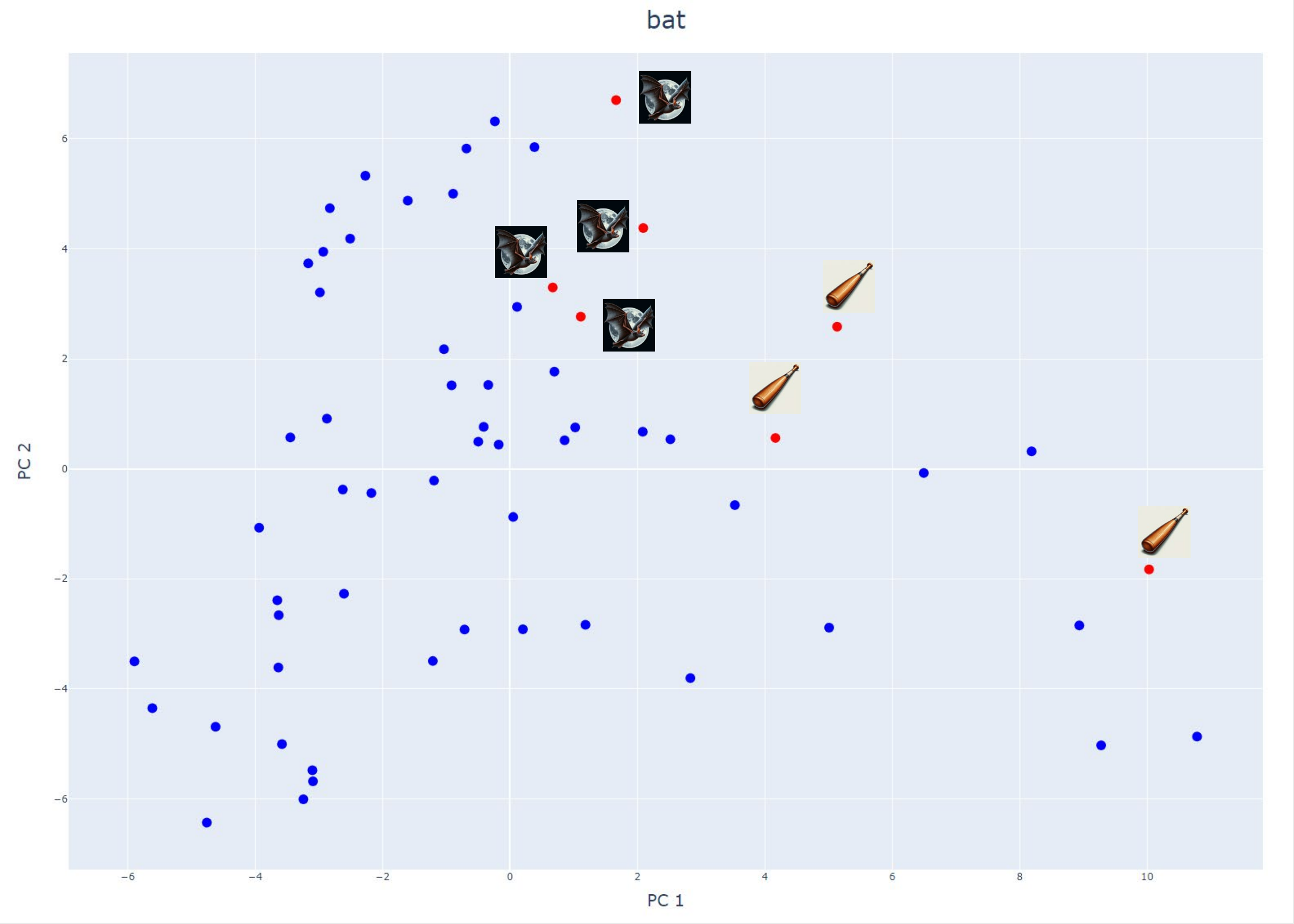
- The transformer outputs a context dependent embedding vector for each word in each sentence
- Let's see if the bat embeddings differ depending on the meaning of the word



Context Dependent Embeddings

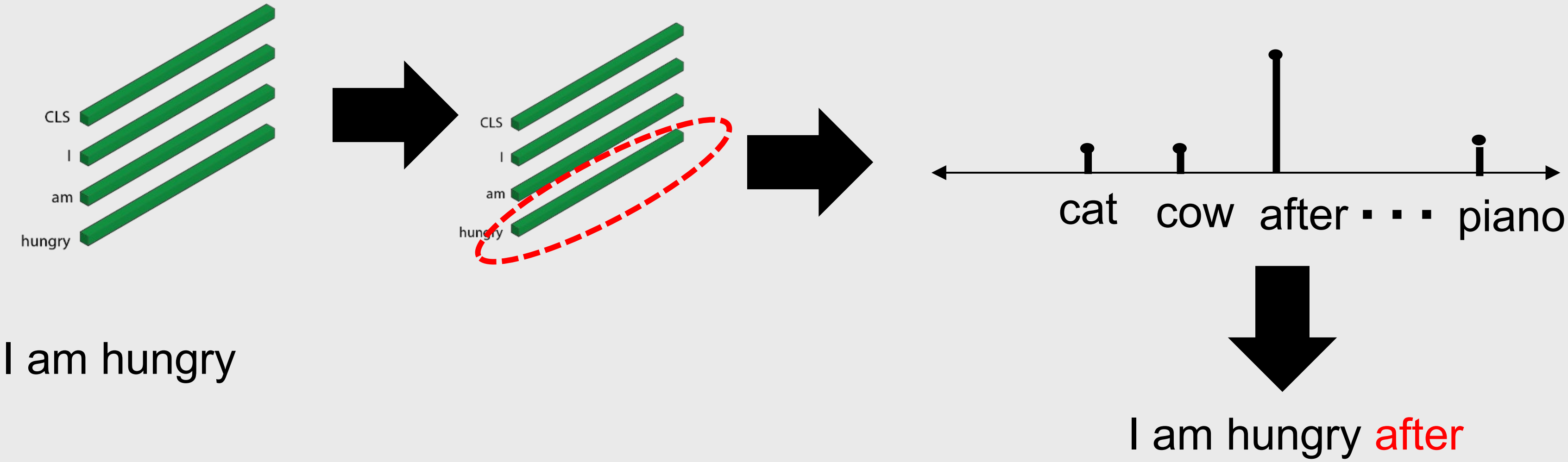


Context Dependent Embeddings



Generative Transformers

- A generative transformer turns the word embedding vector into a probability distribution over all words in the vocabulary
  - Ex) GPT, GPT-2, GPT-3, ChatGPT, GPT-4
- Text is generated by sampling from this distribution
- These types of transformers are also called large language models (LLMs)



ChatGPT

- GPT = Generative Pre-trained Transformer
- ChatGPT Released in 2023 by OpenAI
- Newest version of ChatGPT (GPT-4o) has over 1.8 trillion parameters
  - More than 120 layers
  - More than 96 attention heads per layer
  - At least 12,288 dimensional word embedding
- Trained on all text data in the world + human labeled data

Transformer Basic Training

- **A transformer is trained to complete the masked language task (MLT)**
  - **Fill in the masked word**
- **Unsupervised process – no labeled data needed**

Data	Prediction
I went to the [MASK]	[MASK] = store
I went to the store to buy [MASK]	[MASK] = eggs
I went to the store to buy eggs and they were [MASK]	[MASK] = expensive

ChatGPT (GPT-3.5)

- ChatGPT (GPT-3.5) was a major advancement in generative AI
- Trained using a clever technique: reinforcement learning from human feedback (RLHF)

ChatGPT Is a Tipping Point for AI

by Ethan Mollick

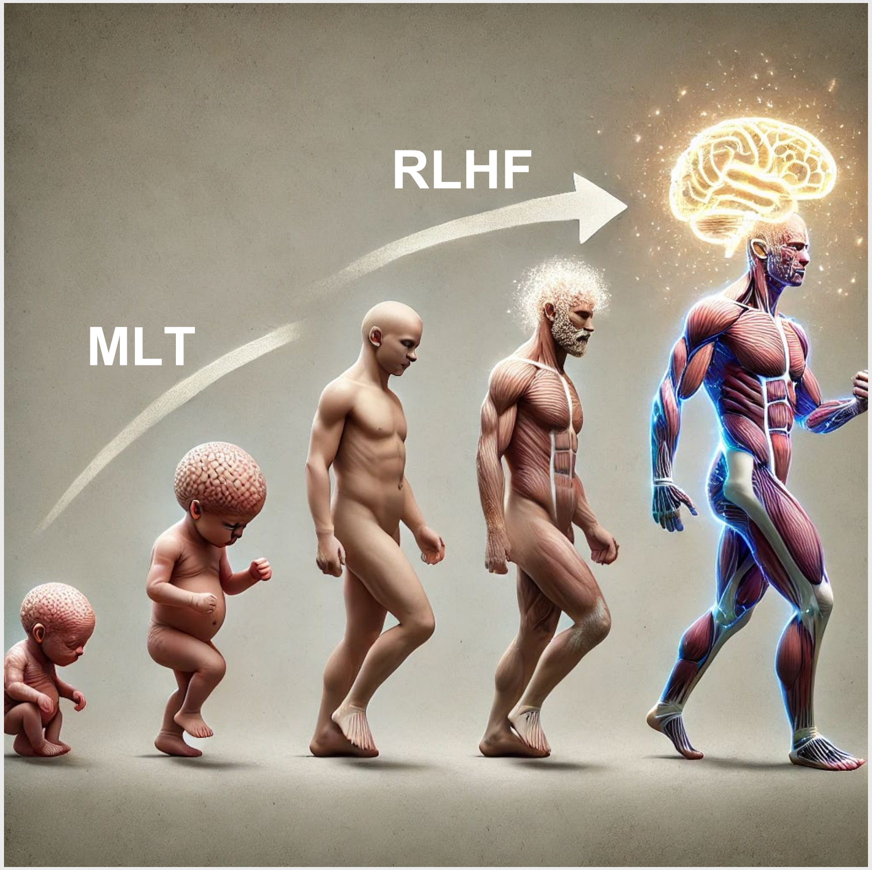
December 14, 2022

Tetiana Lazunova/Getty Images

Transformer Advanced Training

- Masked language task (MLT) takes a transformer from a baby to a child

- To go from an child to a superhuman a new training technique was needed: Reinforcement Learning from Human Feedback (RLHF)
  1. Fine-tune on human created data
  2. Train a reward model to score how good the transformer is
  3. Let the transformer try to beat its high score



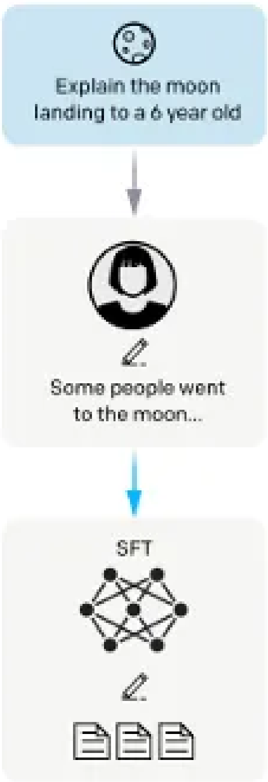
Fine Tuning

Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

Prompt dataset is a series of prompts previously submitted to the Open API



A labeler demonstrates the desired output behavior.

40 contractors hired to write responses to prompts

This data is used to fine-tune GPT-3 with supervised learning.

Input / output pairs are used to train a supervised model on appropriate responses to instructions.

https://towardsdatascience.com/how-chatgpt-works-the-models-behind-the-bot-1ce5fca96286

Reward Model

Step 2

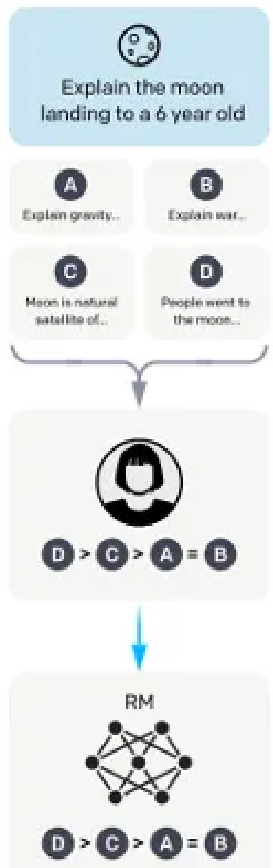
Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

Responses are generated by the SFT model

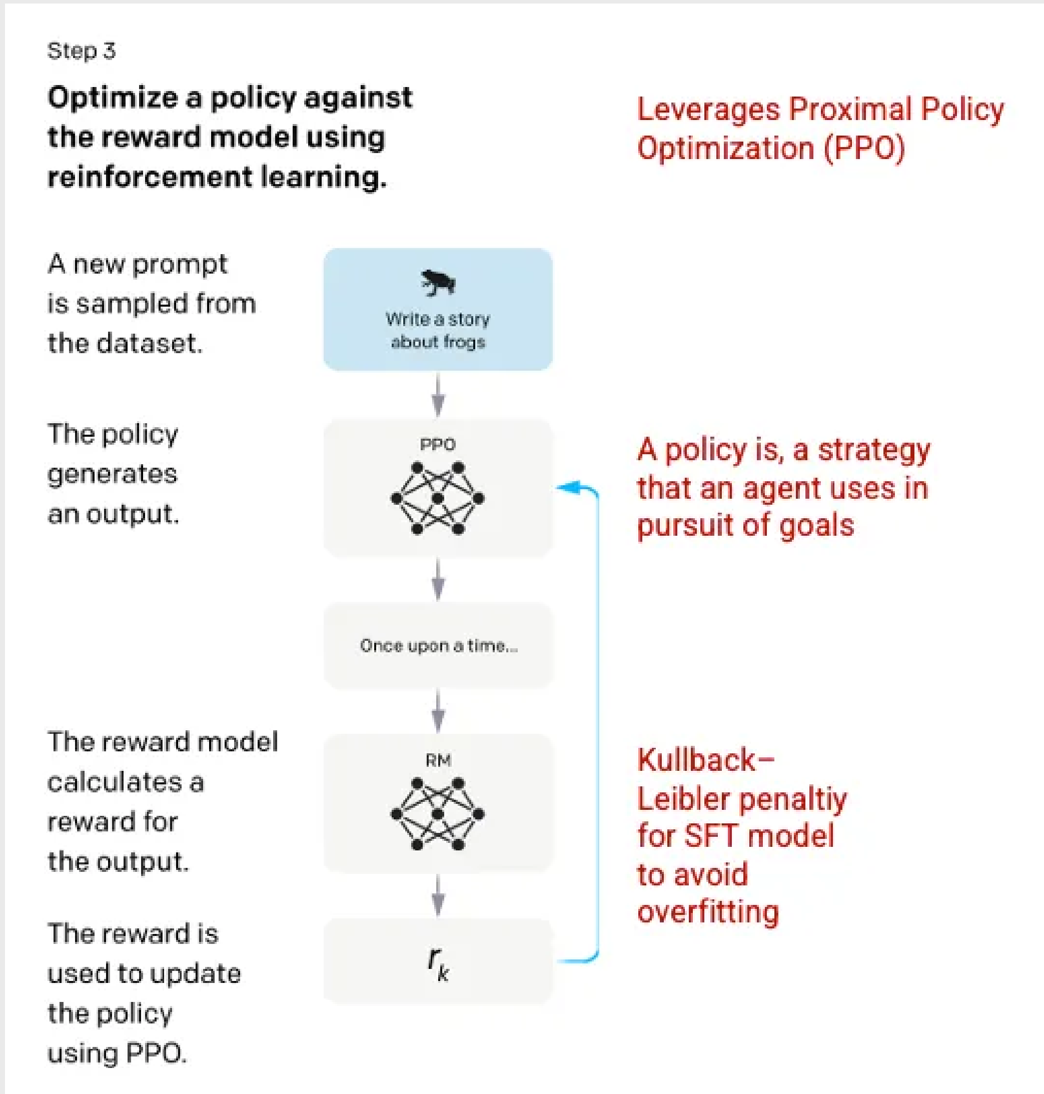
A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



$\binom{k}{2}$ combinations of rankings served to the model as a batch datapoint

Beat High Score with Reinforcement Learning



GPTs Need GPUs

Trump highlights partnership investing $500 billion in AI

1 of 11 | Hours after returning to the White House, President Donald Trump made a symbolic mark on the future of artificial intelligence by repealing former President Joe Biden's guardrails for the fast-developing technology.

TECHNOLOGY | ARTIFICIAL INTELLIGENCE

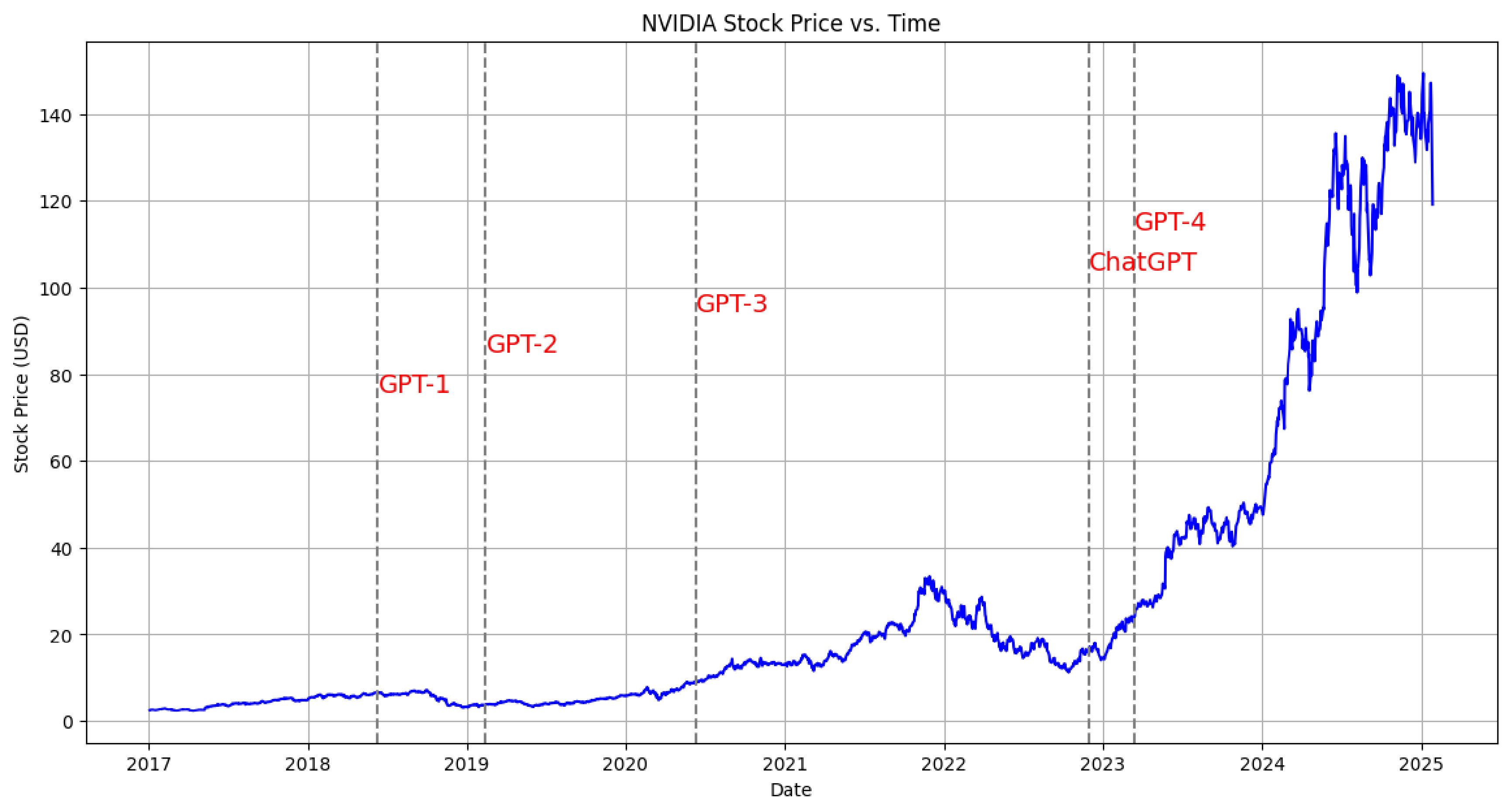
Sam Altman Seeks Trillions of Dollars to Reshape Business of Chips and AI

OpenAI chief pursues investors including the U.A.E. for a project possibly requiring up to $7 trillion

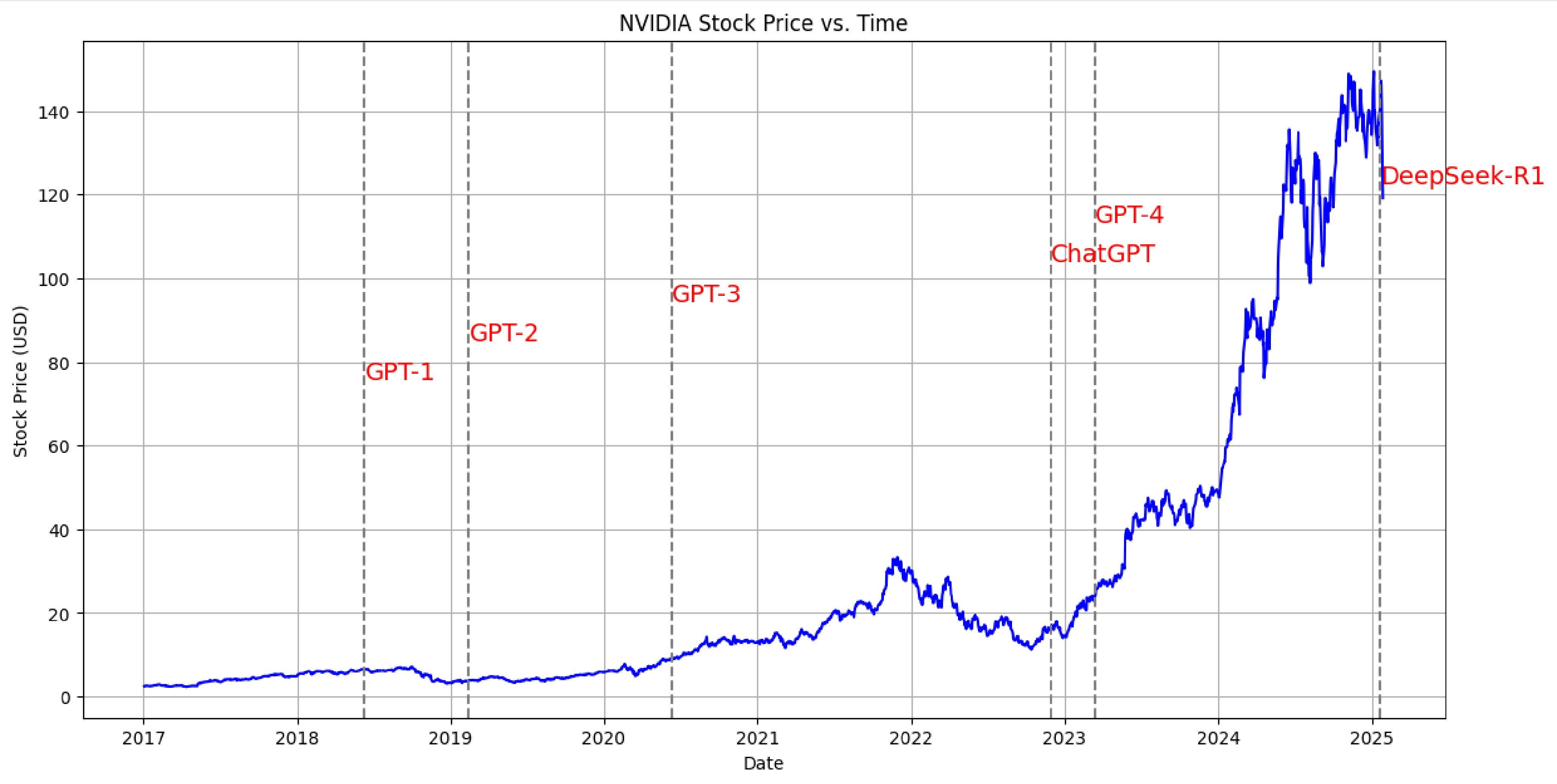
By Keach Hagey Follow *and Asa Fitch* Follow

Feb. 8, 2024 9:00 pm ET

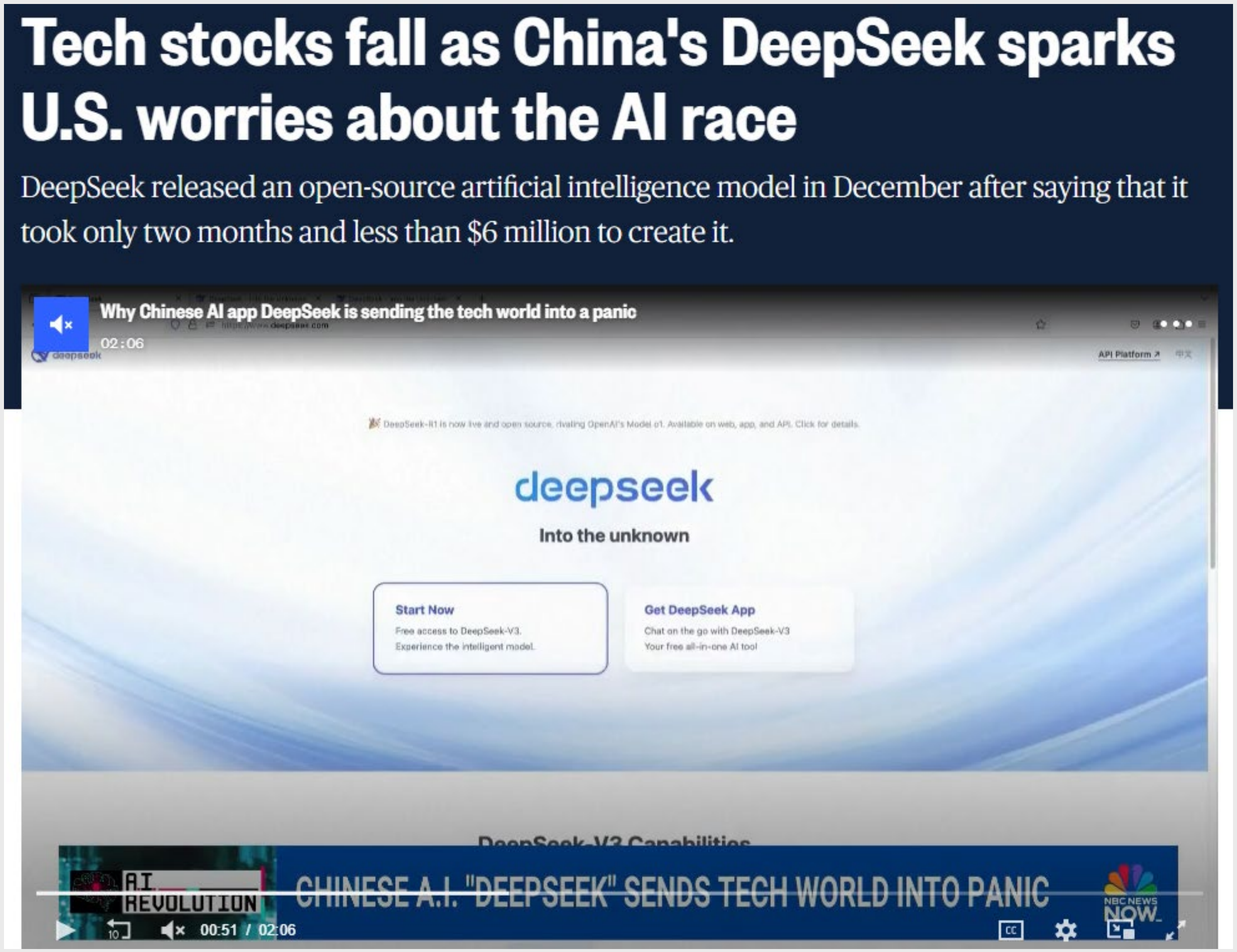
GPTs Need GPUs



GPTs Need GPUs?



GPTs Need Fewer GPUs



Tech stocks fall as China's DeepSeek sparks U.S. worries about the AI race

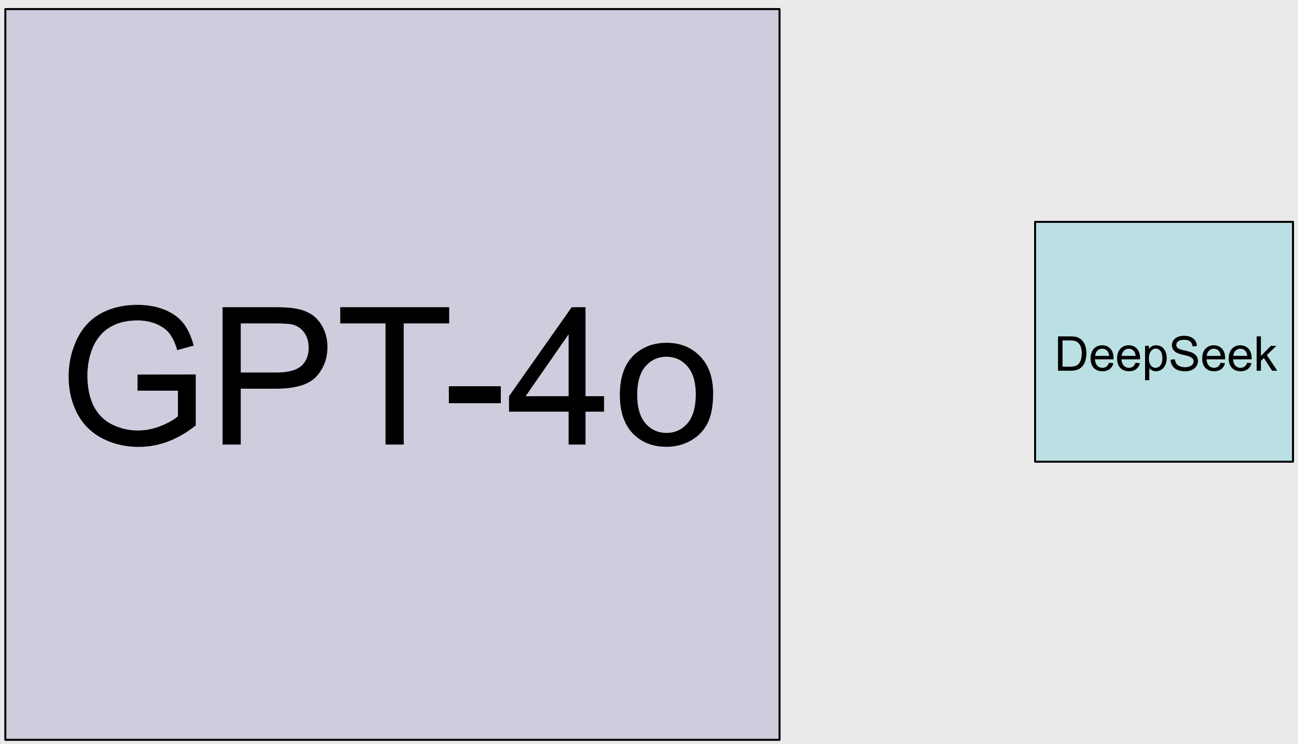
DeepSeek released an open-source artificial intelligence model in December after saying that it took only two months and less than $6 million to create it.

How DeepSeek Makes Transformers Cheaper

- **DeepSeek uses a few clever ideas to make it fast and cheap**
  - **Smaller model**
  - **Lower numerical precision**
  - **Mixture of Experts**
  - **Multi-headed Latent Attention**
  - **Reinforcement learning without human feedback**

Smaller Model

- **DeepSeek has 200 billion parameters vs ChatGPTs 1.8 trillion parameters**



Lower Numerical Precision

- **Lower numerical precision – 8 bits to represent a number vs 32 bits**

32 bits

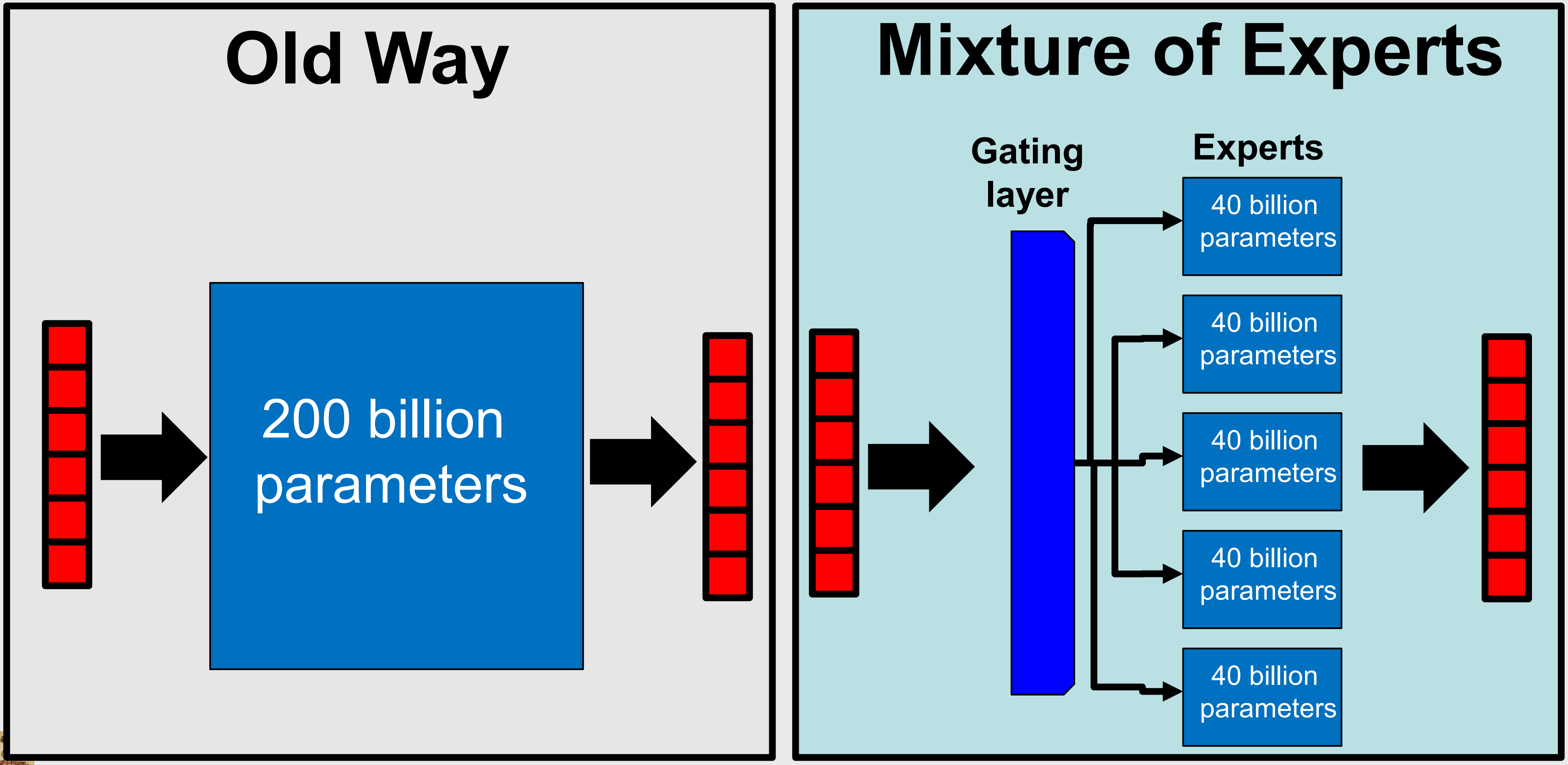
3.1415927

8 bits

3.14

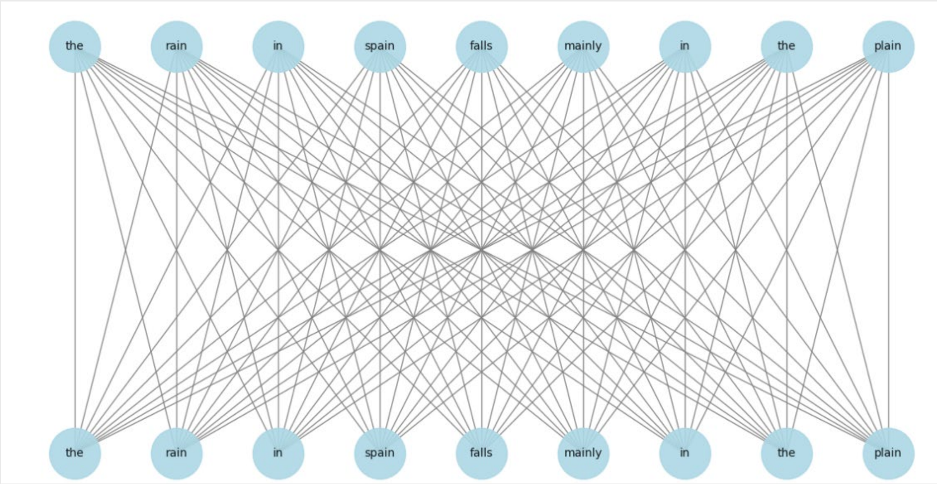
Mixture of Experts

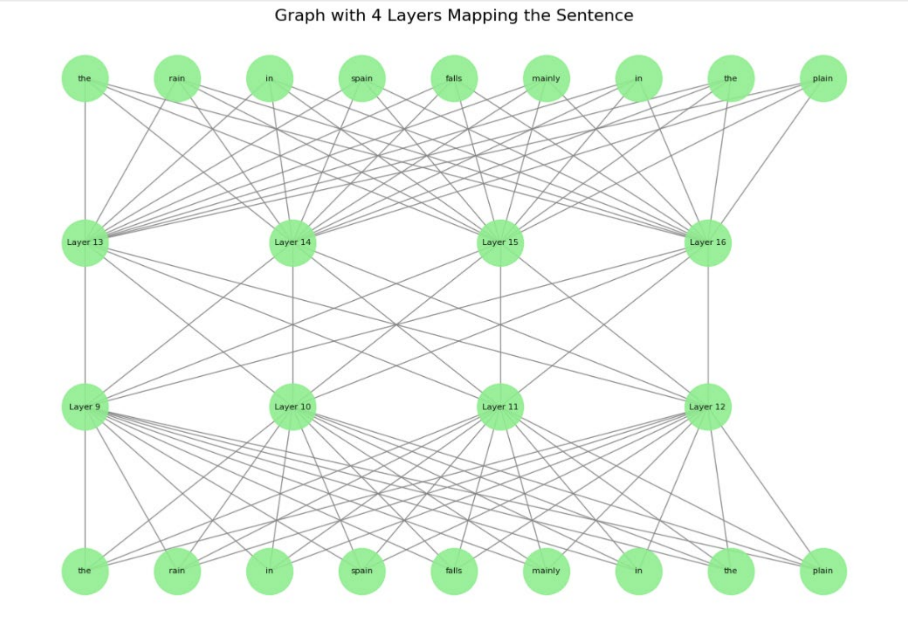
- **Mixture of Experts – Pass text to a subset of the parameters (an expert) instead of all the parameters**



Multi-Headed Latent Attention

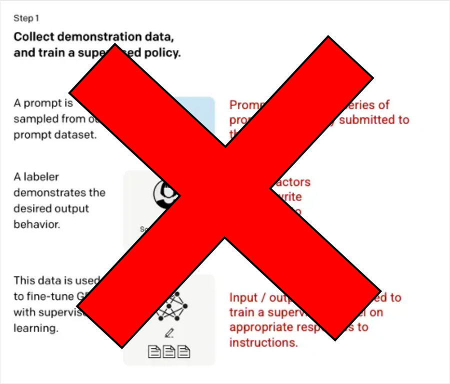
- Do attention in low dimensional latent space
- Need many fewer parameters in model

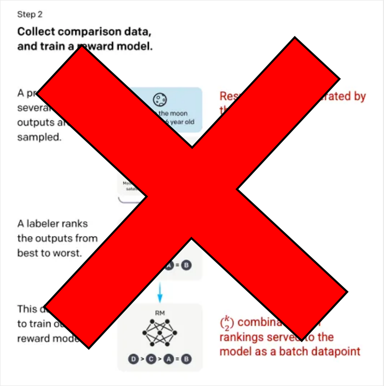


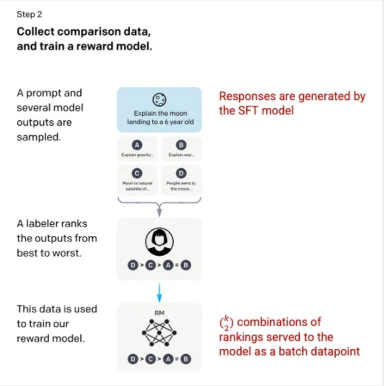


RL without Human Feedback

- DeepSeek does not do fine tuning
- DeepSeek does not train a reward model
  - Reward is whether or not response is correct
  - Math and coding problems
- Reduces training cost







DeepSeek vs GPT-4o

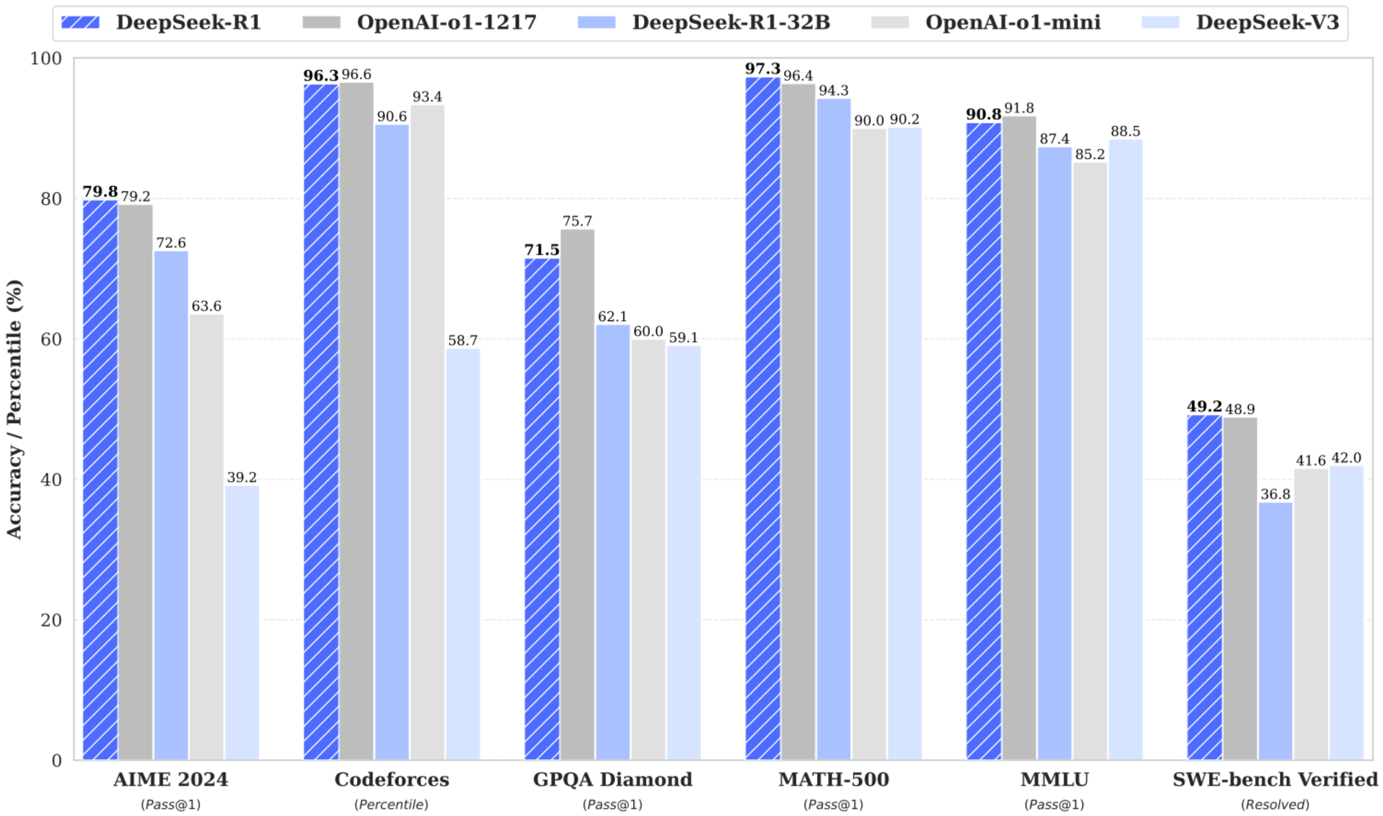
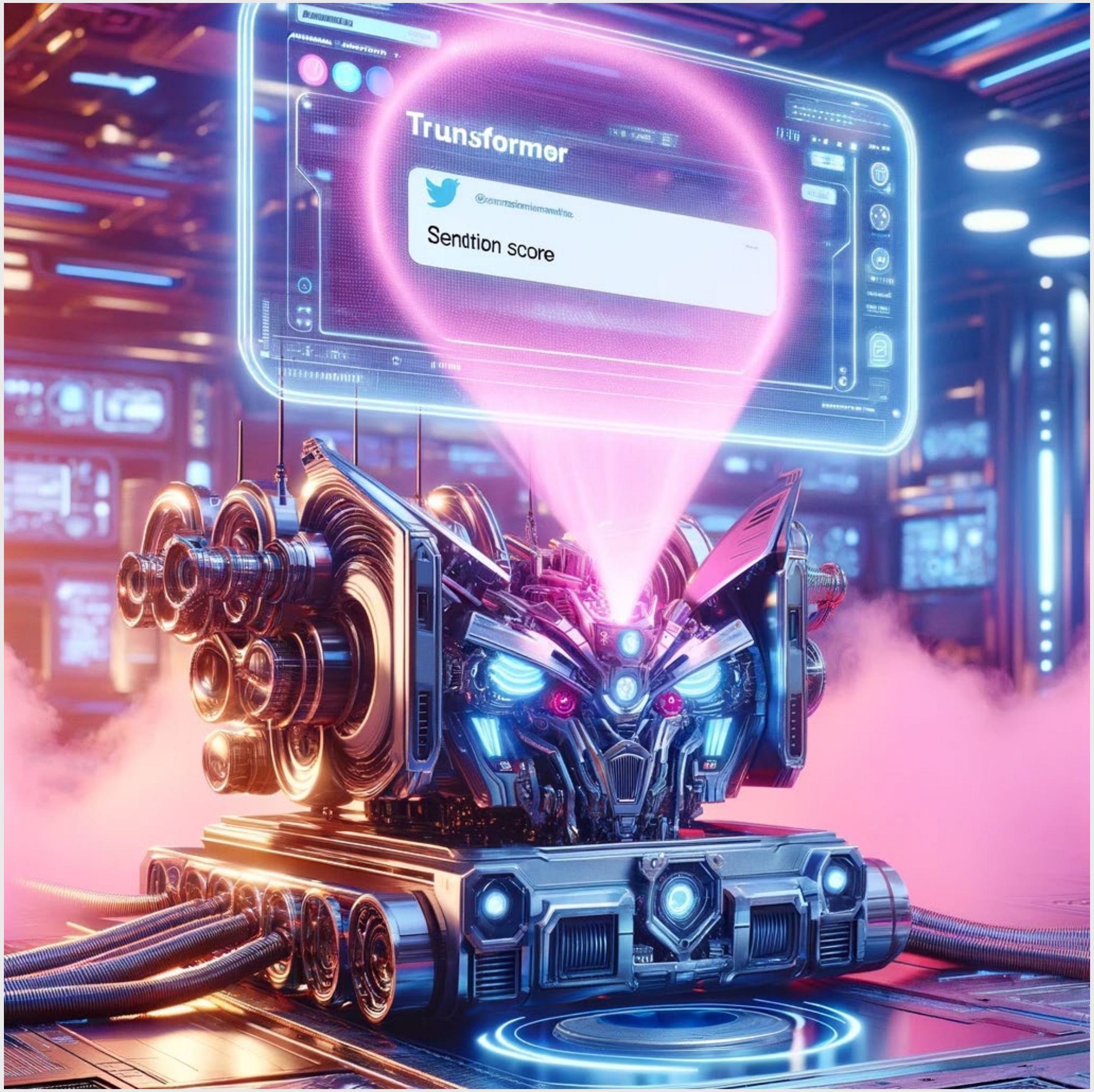


Figure 1 | Benchmark performance of DeepSeek-R1.

Sentiment Analysis with Transformers



Sentiment

- Tweet 1: My birthday cake was awful
- Tweet 2: My birthday cake was great

Sentiment and Keywords

- Sentiment is conveyed by specific words
- Maybe we could use a word frequency approach to measure sentiment
- Early sentiment classifiers did this
  - Naïve Bayes classifier

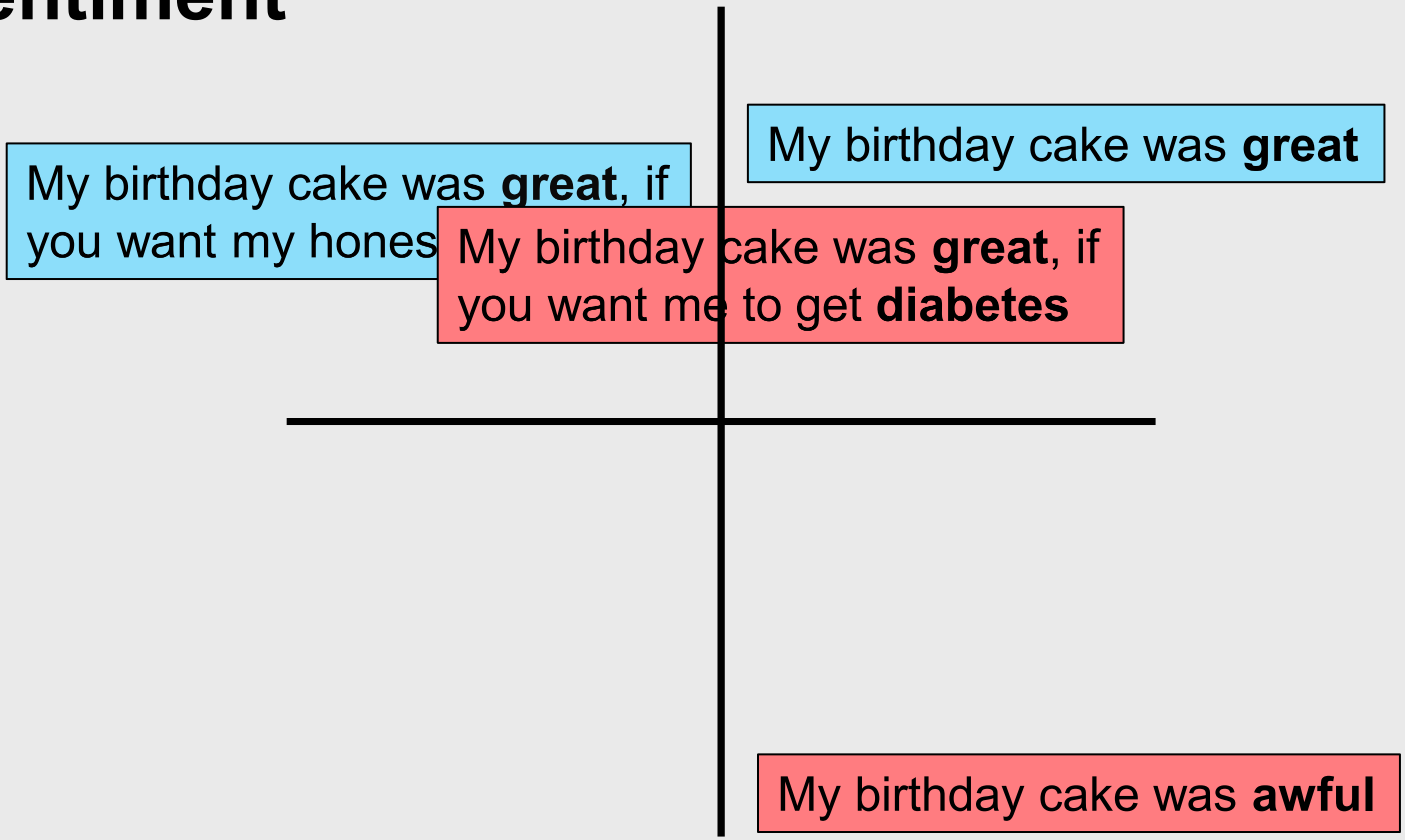
Sentiment and Context

- Tweet 1: My birthday cake was great, if you want my honest opinion

- Tweet 2: My birthday cake was great, if you want me to get diabetes

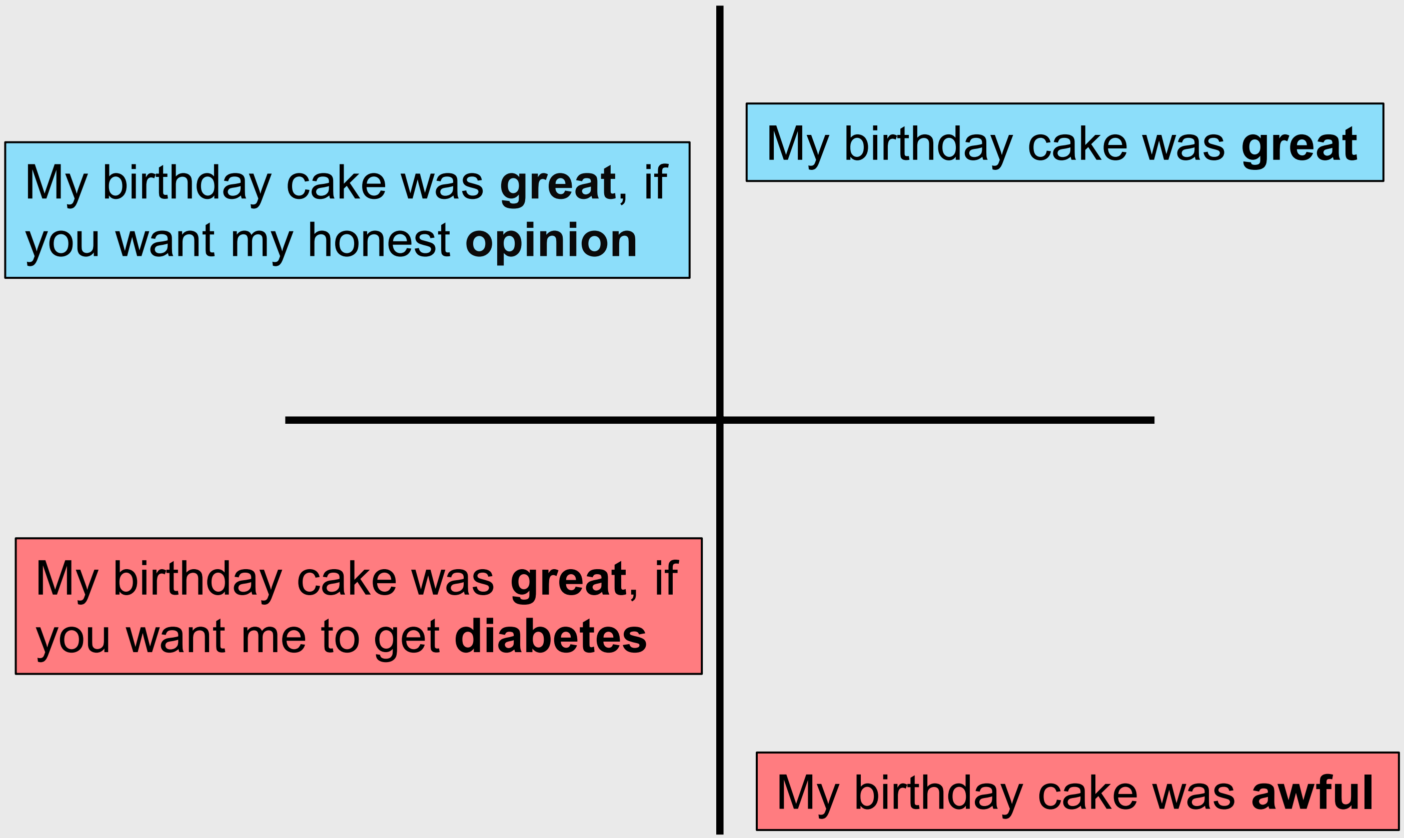
Term Frequency Embeddings

- **Term frequency based embeddings may cluster tweets with similar words, but different sentiment**



Context Dependent Embeddings

- **A context dependent embedding can cluster by sentiment**



Sentiment and Attention

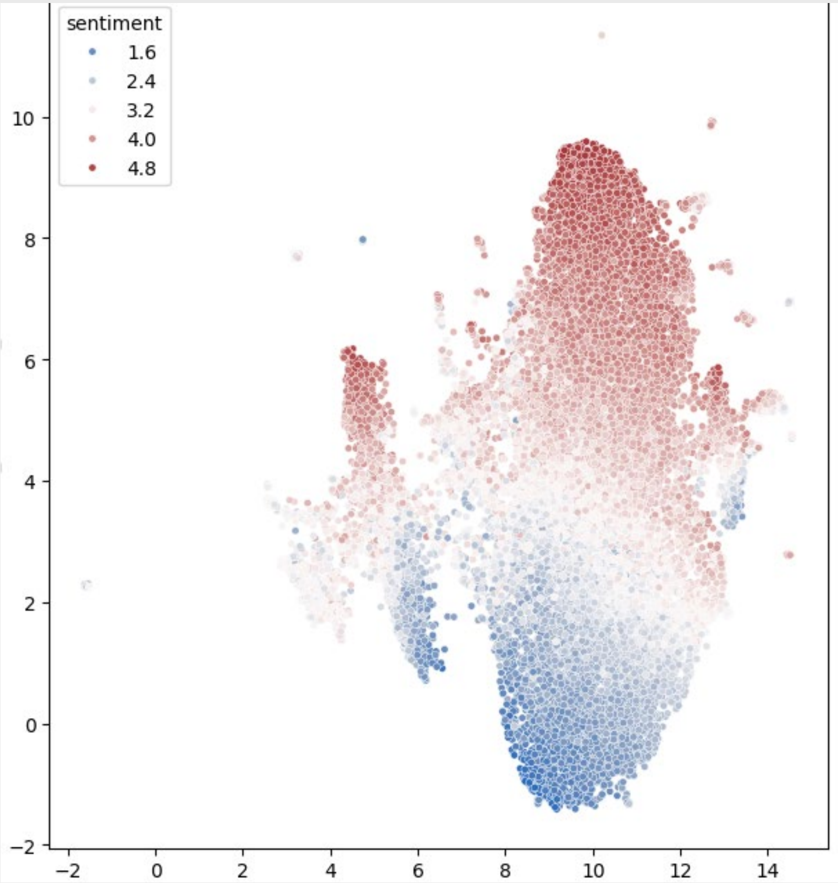
- Sentiment is conveyed by specific words
- We also need to know the context of the words
- Context = to which words does a word pay attention?

Attention

- We need a model that allows words in a sentence to pay "attention" to other words
- Words can pay attention in different ways
- We can choose the type of "attention" that captures sentiment
- Solution: Transformers

Transformer Embeddings and Sentiment

- **We already saw how the attention mechanism lets a transformer make context dependent embeddings**
- **Transformer embeddings are able to capture sentiment geometrically**
- **The transformer embedding "separates" tweets based on their sentiment**



Measuring Sentiment with Pre-Trained Transformers

- In the old days, we would have to train a transformer to measure sentiment
  - Collect and label data ☹
  - Train for hours on a GPU ☹
- Today, we have ChatGPT
  - No training data needed
  - No training needed
  - Sometimes you don’t even need any examples

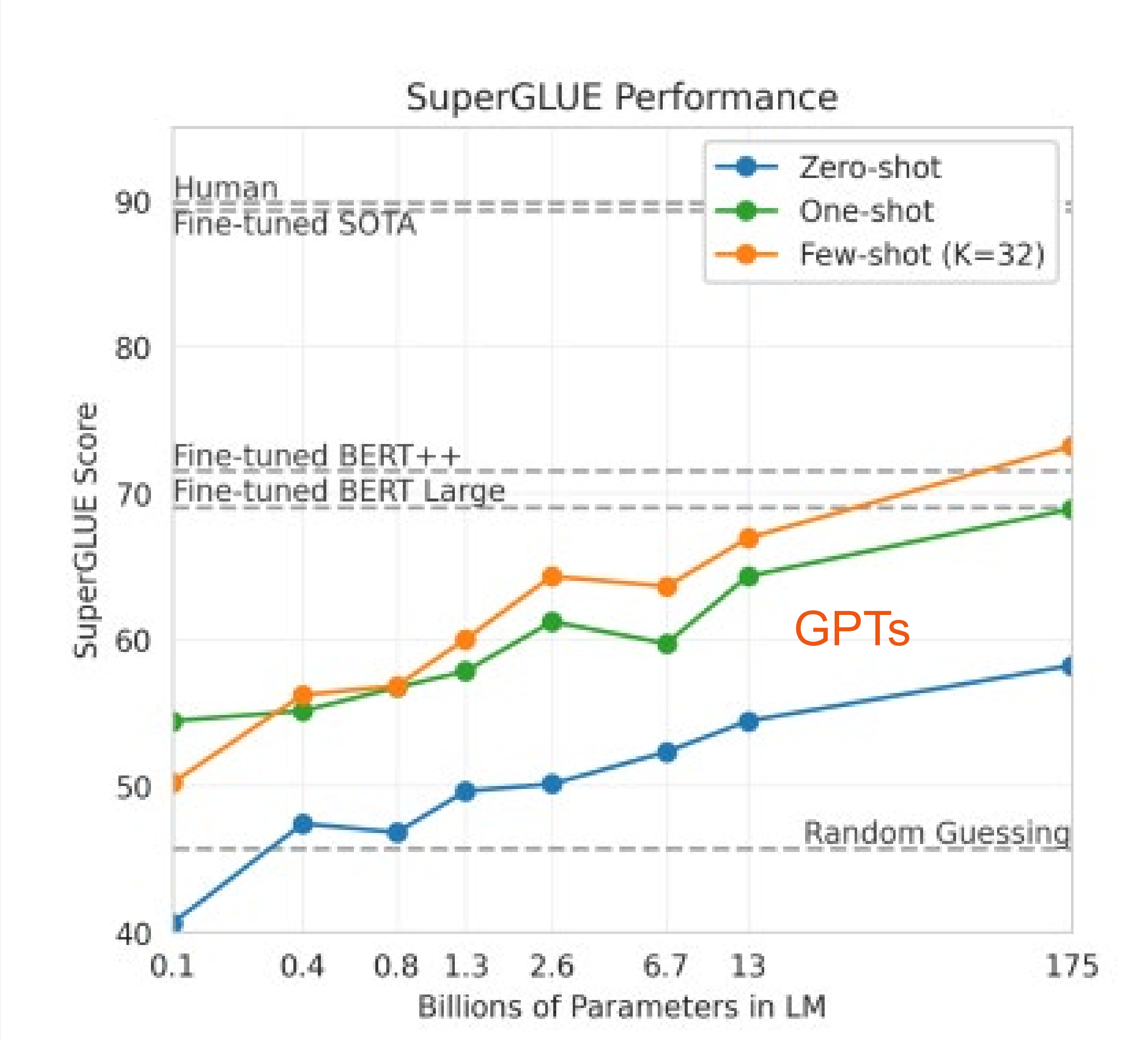
Few-Shot Learning

- How do we make the language model generate text for a specific task?

- Old way – fine tune on new set of data
  - Collect and label data ☹
  - Train for hours on a GPU ☹

- New way – few-shot learning
  - Put a few example texts in the input
  - No training needed
  - Works surprisingly well if the model is large enough

Few-Shot Learning

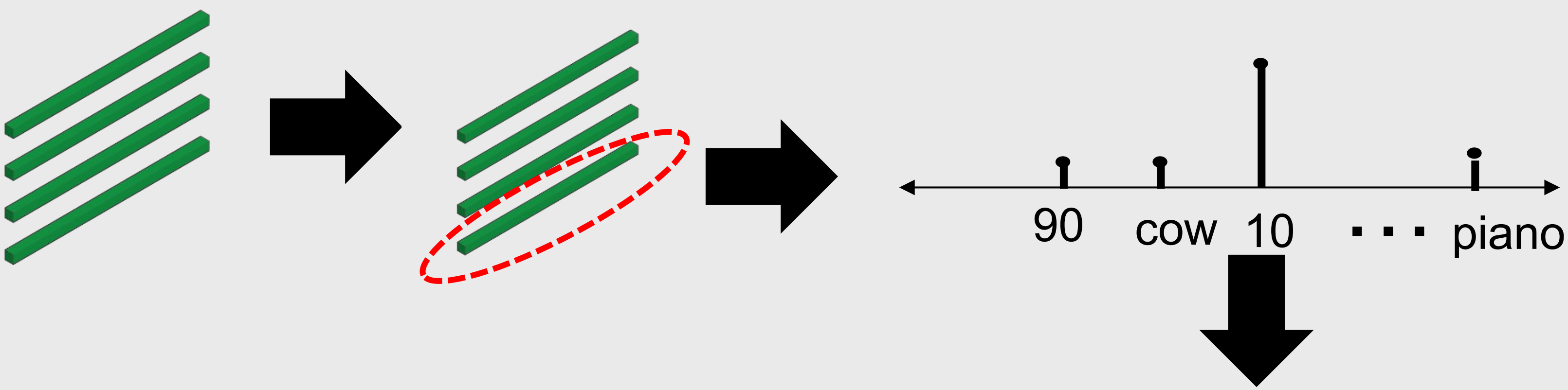


Emergent Behavior

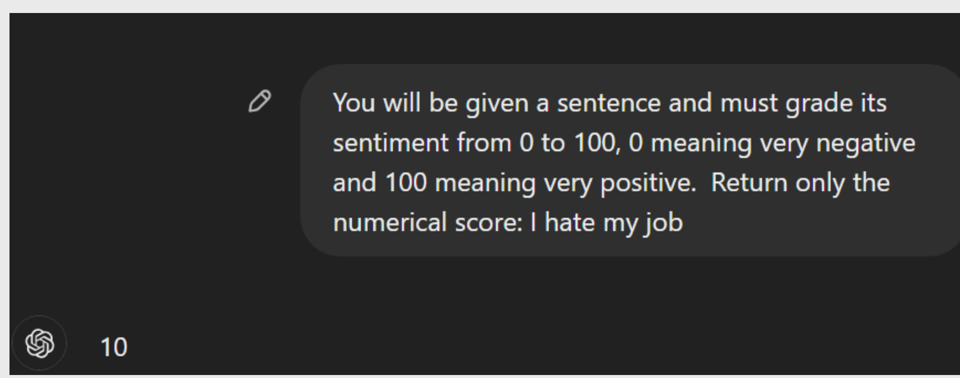
- The transformer was trained to complete sentences
- It has shown the ability to perform many behaviors it was not trained on
- These are emergent behaviors – but no one really understands why the emerge
- Emergent behaviors are the reason why transformers are so powerful

Measuring Sentiment with ChatGPT

- **ChatGPT can measure sentiment based on a prompt**
- **Ex) "You will be given a sentence and must grade its sentiment from 0 to 100, 0 meaning very negative and 100 meaning very positive. Return only the numerical score: I hate my job**

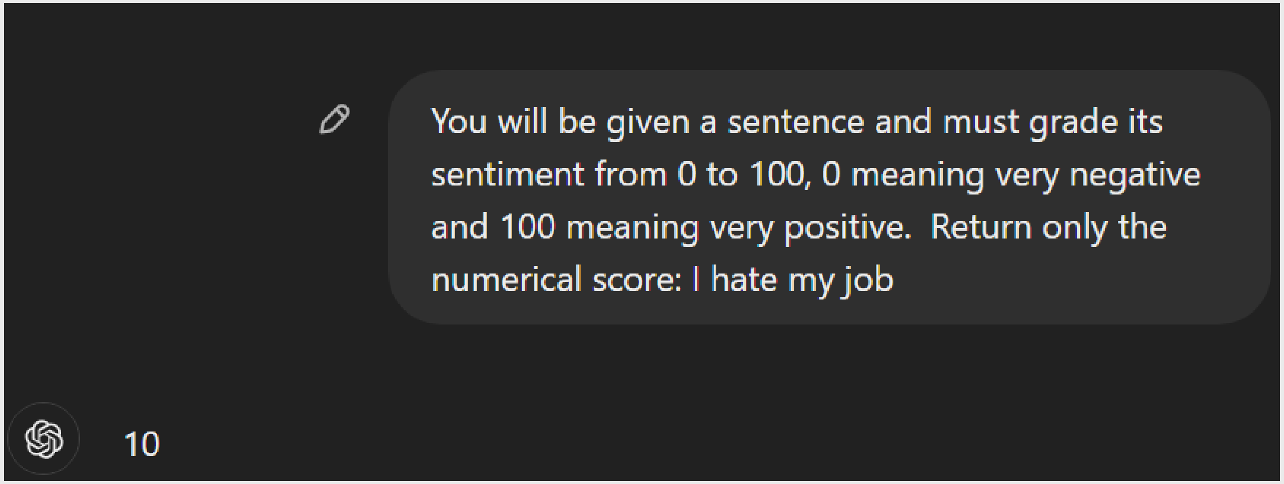


You will be given a sentence and must grade its sentiment from 0 to 100, 0 meaning very negative and 100 meaning very positive. Return only the numerical score: I hate my job

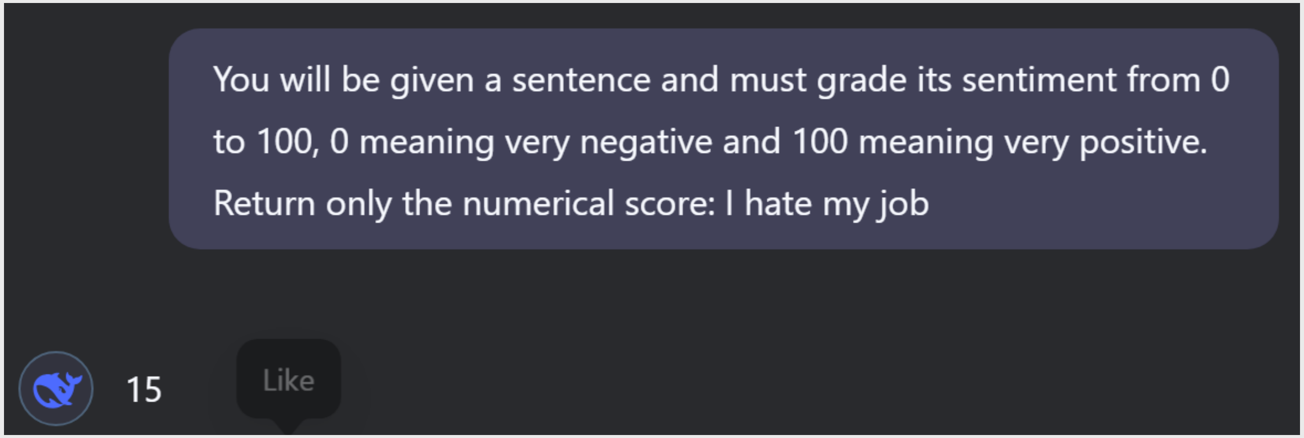


ChatGPT vs DeepSeek

ChatGPT



DeepSeek



Measuring Complex Sentiment with ChatGPT

- **ChatGPT can give us more nuanced sentiment and explain why**

You will be given a sentence and must grade its Ekman's Six Basic Emotions from 0 to 100: Return a table of the scores and explanation why: I hate my job

Here is an analysis of the sentence "I hate my job," graded on **Ekman's Six Basic Emotions** (Anger, Fear, Joy, Sadness, Disgust, and Surprise), along with explanations for each score:

Emotion	Score (0-100)	Explanation
Anger	85	The word "hate" conveys strong negative feelings and frustration, indicating anger.
Fear	40	The sentiment may imply fear of being stuck in an undesirable situation or uncertainty about the future.
Joy	0	The sentence expresses no happiness or positivity, as "hate" and "job" in this context are negative.
Sadness	75	The phrase suggests unhappiness, disappointment, or distress about the situation.
Disgust	90	The word "hate" evokes a sense of aversion or repulsion, strongly tied to disgust.
Surprise	10	There is little indication of surprise in the sentence, as it reflects ongoing dissatisfaction.

Next Time: Coding Session

- We will learn to measure tweet sentiment with ChatGPT and visualize tweet embeddings

You will be given a sentence and must grade its Ekman's Six Basic Emotions from 0 to 100: Return a table of the scores and explanation why: I hate my job

Here is an analysis of the sentence "I hate my job," graded on **Ekman's Six Basic Emotions** (Anger, Fear, Joy, Sadness, Disgust, and Surprise), along with explanations for each score:

Emotion	Score (0-100)	Explanation
Anger	85	The word "hate" conveys strong negative feelings and frustration, indicating anger.
Fear	40	The sentiment may imply fear of being stuck in an undesirable situation or uncertainty about the future.
Joy	0	The sentence expresses no happiness or positivity, as "hate" and "job" in this context are negative.
Sadness	75	The phrase suggests unhappiness, disappointment, or distress about the situation.
Disgust	90	The word "hate" evokes a sense of aversion or repulsion, strongly tied to disgust.
Surprise	10	There is little indication of surprise in the sentence, as it reflects ongoing dissatisfaction.