Large Language Models

Introduction

Large Language Models

OpenAI GPT-5

ANTHROP\C CLAUDE

deepseek R1

Google Gemini

Microsoft® PHI-4

Meta LLAMA

Language models: the good

Improve productivity
- Coding
- Writing assistance
- Paperwork and bureaucracy

- Break down language barriers
  - Machine translation
- Can process speech
  - Healthcare and L2 education
- Information and Knowledge Access

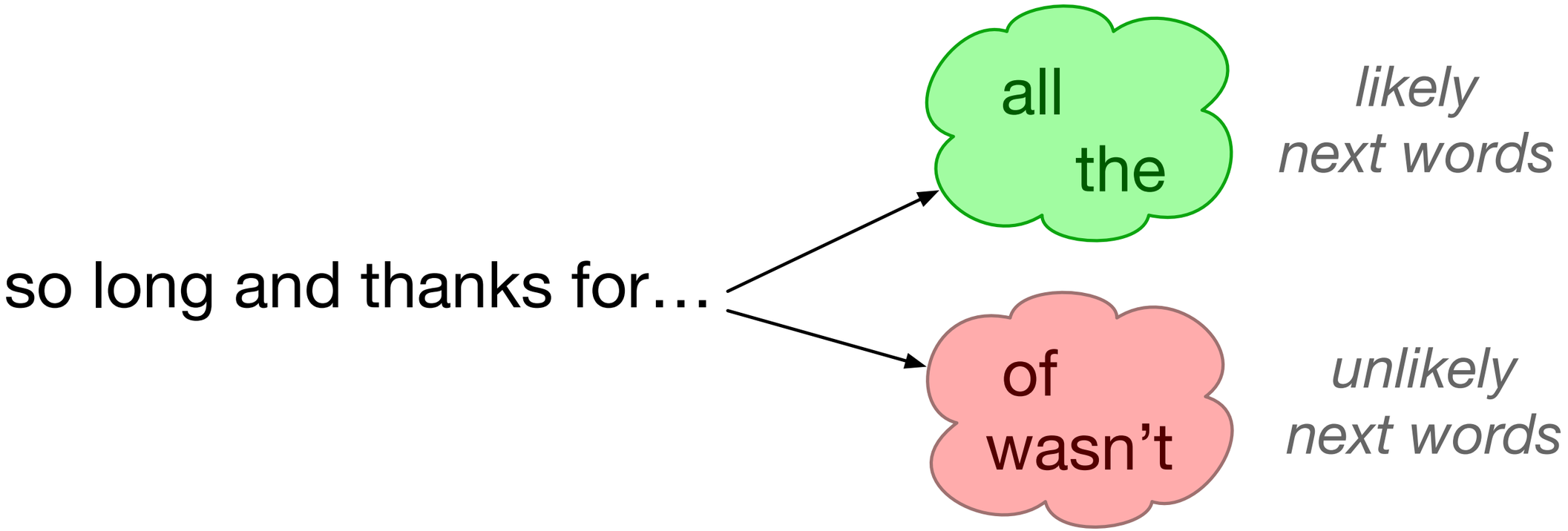
Language models: the concerns

ELIZA story
We'll discuss safety and harm issues later in this lecture

What is a language model?

1. An agent that carries on conversations and acts in the world
2. A computational system that can predict the next word from previous words.

A language model predicts words



Prediction is an not English-specific idea

在我的后园，可以看见墙外有两株树，一株是…

Zài wo de hòu yuán, kěyǐ kàn jiàn qiáng wài you liǎng zhū shù, yī zhū shì…

In my backyard, you can see two trees beyond the fence. One is a…

枣树(‘date’)

的(‘of’)

桃树(‘peach’)

松树(‘pine’)

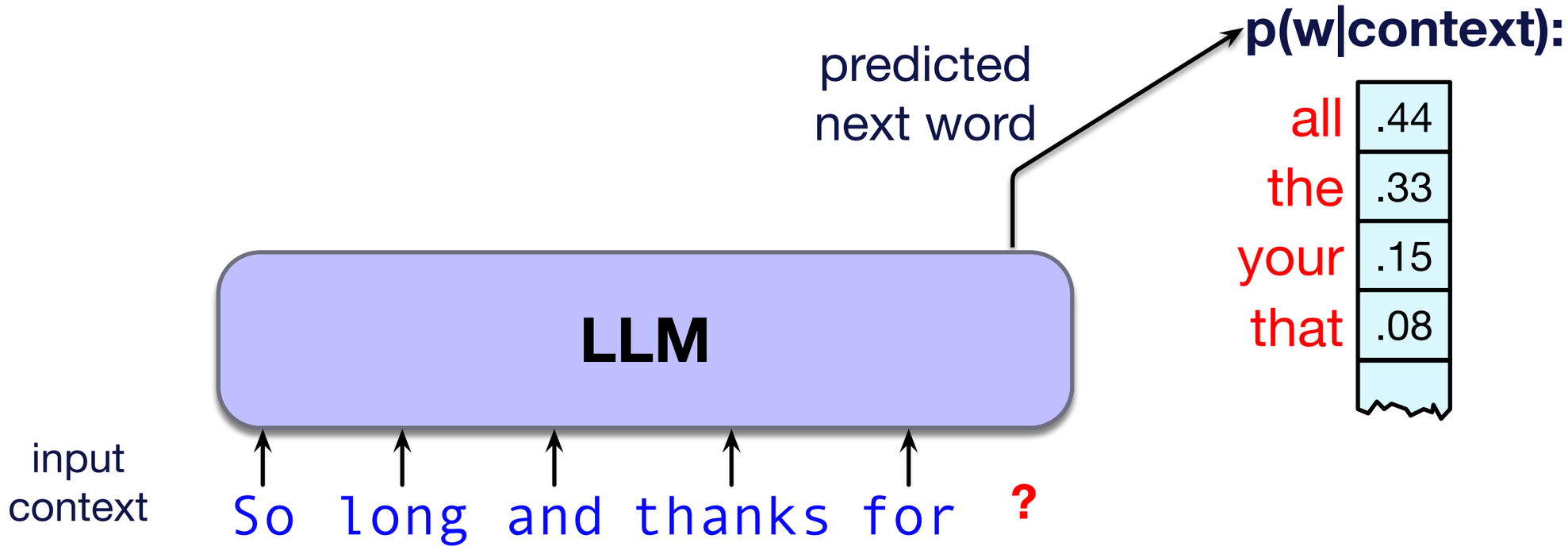
我们(‘we’)

What is a large language model?

A neural network with:

Input: a context or prefix,

Output: a distribution over possible next words



Intuition of generation

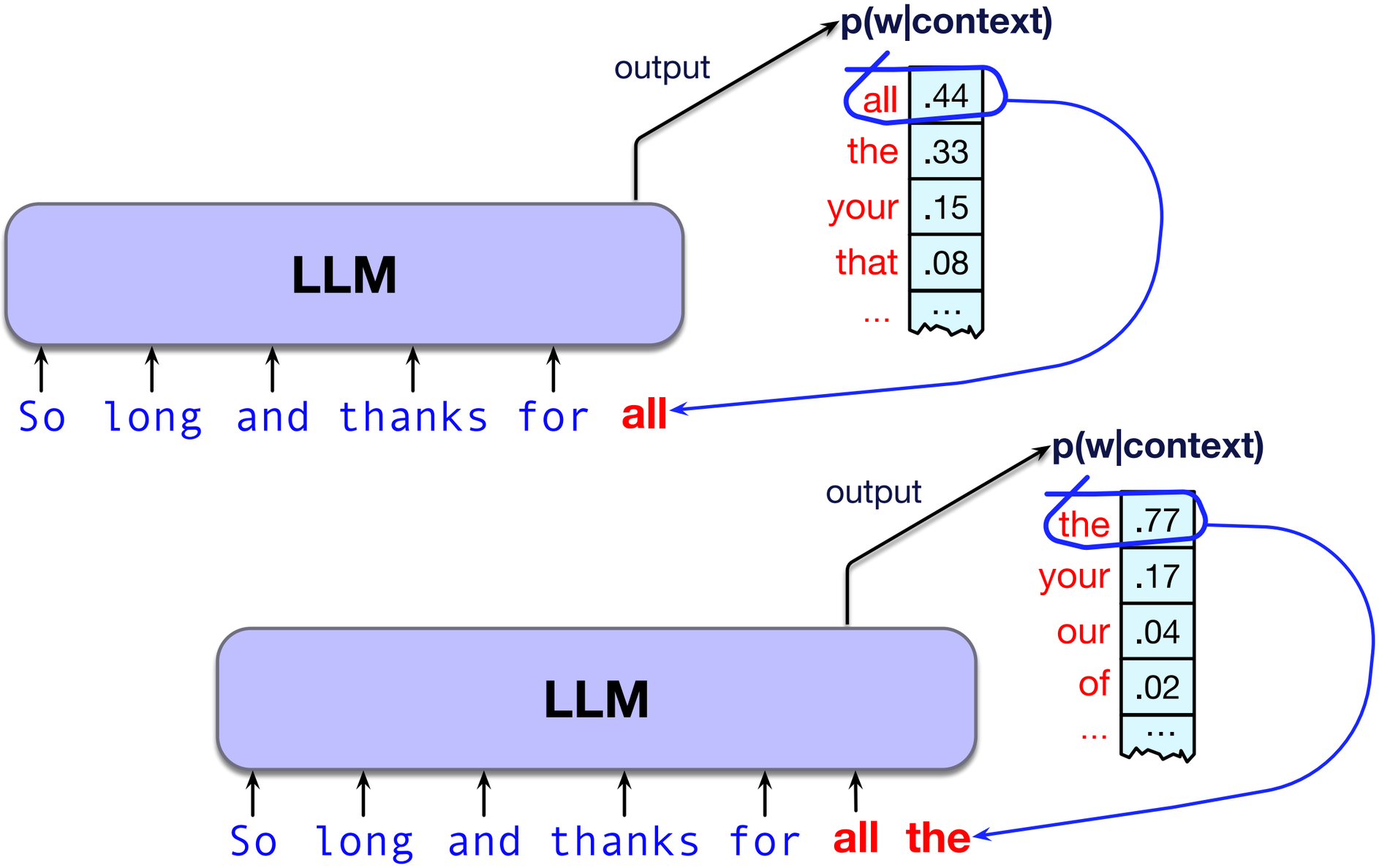
An LLM that can predict text

- assigning a probability distribution over upcoming words

Can generate words

- by **sampling** from the distribution

LLMs can generate!

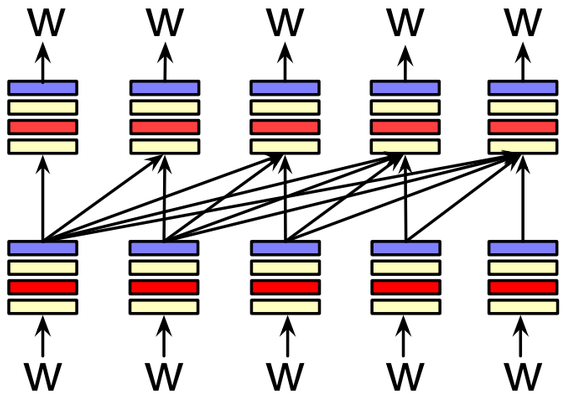


Causal or autoregressive model

Models that iteratively predict and generate words left-to-right from earlier words

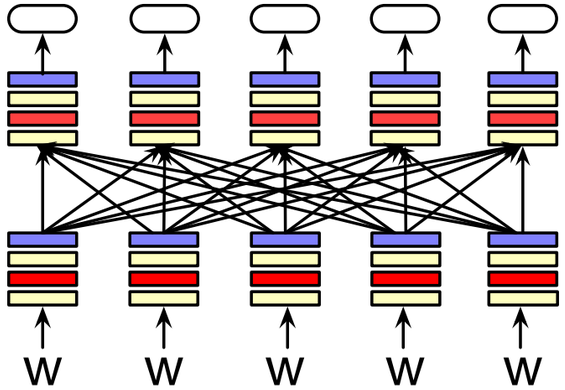
The most common LLMs

Three architectures for large language models



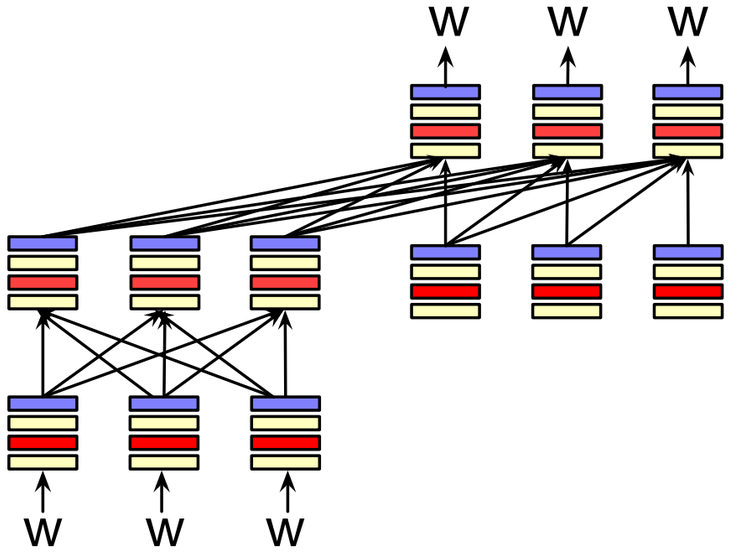
Decoders

GPT, Claude, Llama Mixtral



Encoders

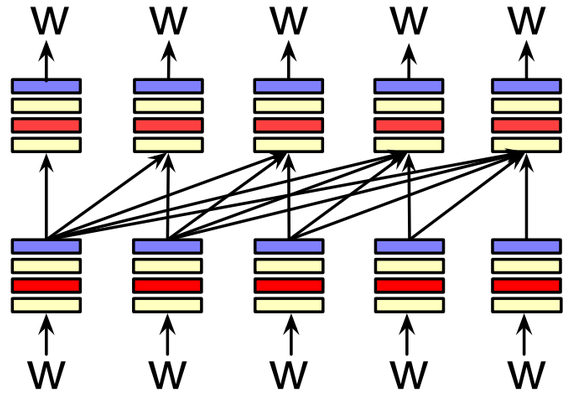
BERT family, HuBERT



Encoder-decoders

Flan-T5, Whisper

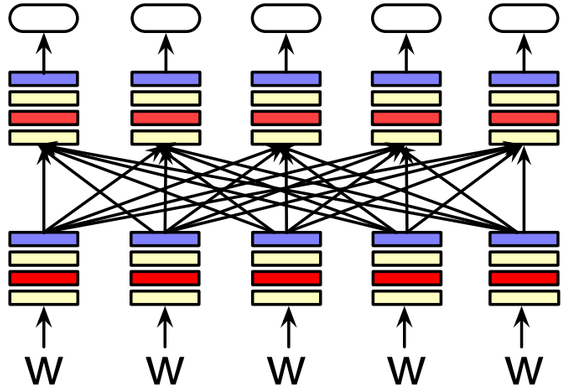
Decoders



What most people think of when we say LLM

- GPT, Claude, Llama, DeepSeek, Mistral
- A generative model
- It takes as input a series of tokens, and iteratively generates an output token one at a time.
- Left to right (causal, autoregressive)

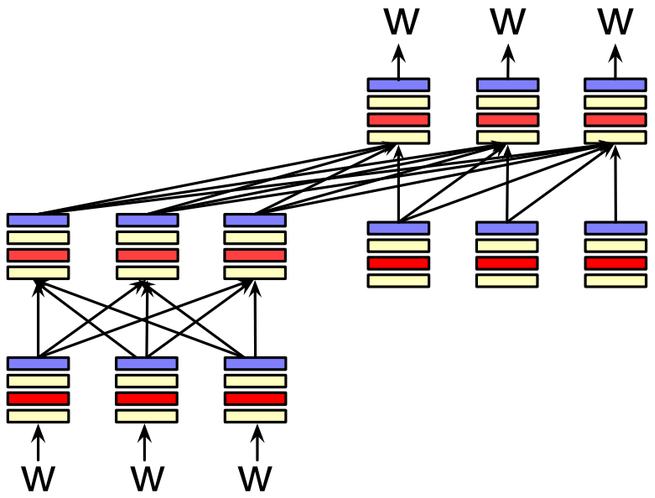
Encoders



- Masked Language Models (MLMs
- BERT family

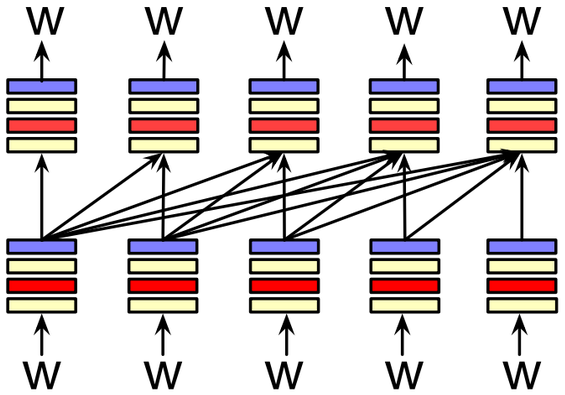
- You've used BERT already for contextual embeddings
- Trained by predicting words from surrounding words on both sides
- Are usually **finetuned** (trained on supervised data) for classification tasks.

Encoder-Decoders



- Trained to map from one sequence to another
- Used for:
  - machine translation (map from one language to another)
  - speech recognition (map from acoustics to words)

This lecture: decoder-only models

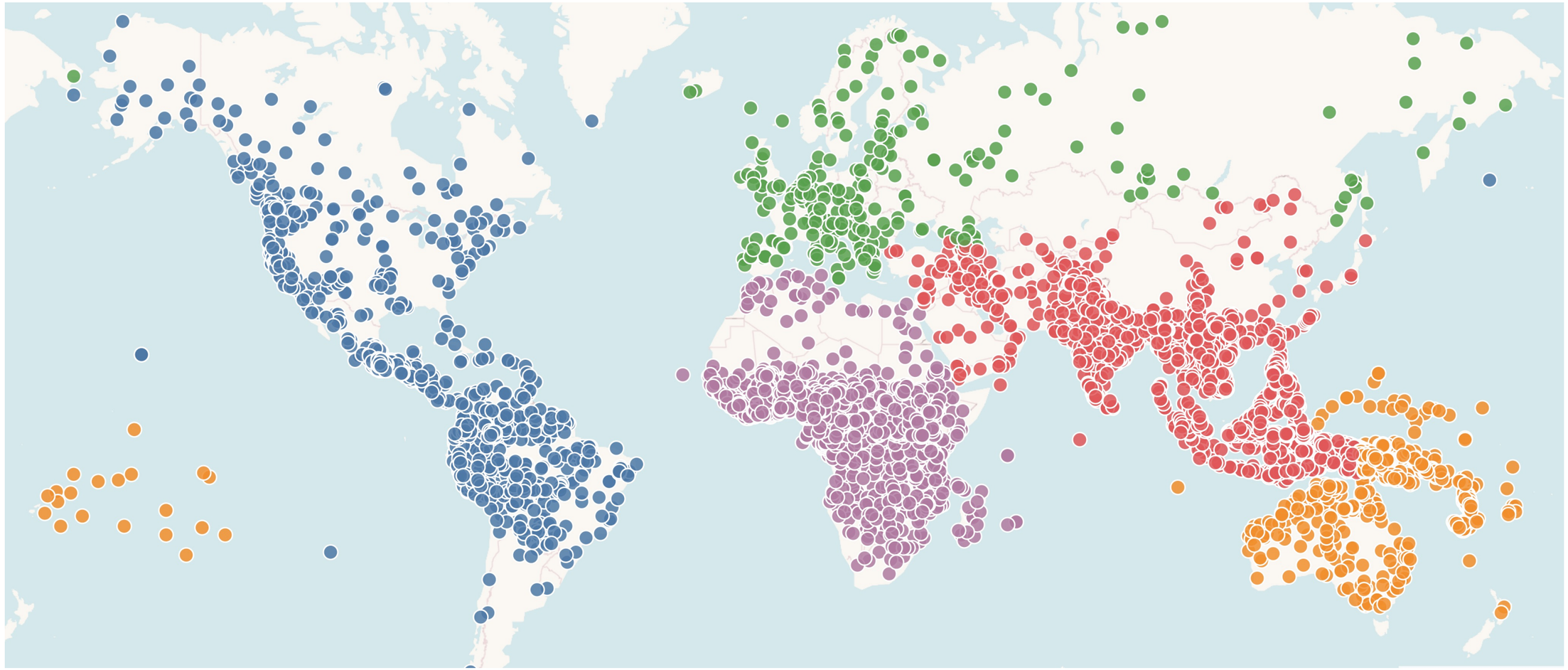


Also called:

- Causal LLMs
- Autoregressive LLMs
- Left-to-right LLMs

- Predict words left to right

Wait, what about other languages?



There are >7000 languages, more than 150 scripts

Ethnologue

Wait, what about other languages?

Top 10 most spoken languages

Tthnologue

Language models can be monolingual or multilingual

Many are multilingual, e.g.,:

- Gemini: 140 languages
- Qwen3: 119 languages and dialects

Some are monolingual

- Especially common in English

Language models process **tokens**, not words

A **token** is a word or word part.

First LLM step is converting text into tokens

- Via tokenization algorithms like **BPE**

English tokens correspond roughly to words:

```
Anyhow,·she's·seen·Jane's·224123·flowers·anyhow!
```

German tokens tend to be smaller than words

```
Ist·die·Trägheit·eines·Körpers·von·seinem·
Energieinhalt·abhängig?·von·A.·Einstein.
```

Size and scale

At the small end: n-gram models

Count the frequency of words in different words

Useful for understanding basic LM concepts

What is a Large Language Model (LLM)

LMs built from neural architectures

Especially the transformer

Large in the sense that they:

- have **many layers**
- have **many parameters**
- are trained on **large training sets**

What is a parameter?

A single real-valued number in a network that is trained and used for computation

- weights and biases

Llama 3.1 405B has 405 Billion parameters

trained on over 15 trillion tokens.

Models come in families from <1B parameters to trillions

Scaling laws

LLM performance goes up as these 3 factors increase:

- model size (the number of parameters)
- training dataset size (in tokens)
- amount of compute used for training

Scaling laws

For a fixed amount of compute C

Optimal # of parameters N

Optimal amount of data D

Should go up equally as a power law of compute

$$N_{opt} \propto C^{\frac{1}{2}} \qquad D_{opt} \propto C^{\frac{1}{2}}$$

Scaling laws

Explain why models get bigger

- For good (better performance)
- And ill (more energy, water, $$$) and expensive to run inference.

Also explain that we can improve the same size model by increasing compute and and training data

Large Language Models

Introduction

Large Language Models

What do LLMs learn from token prediction?

Fundamental intuition of large language models

Text contains enormous amounts of knowledge

Pretraining on lots of text with all that knowledge is what gives language models their ability to do so much

Intuition from humans

Vocabulary size of young adult speakers of American English: 30,000 to 100,000 words

Children learn **7 to 10 words a day** to get to that level by 20 years old

How do they do it?

Reading! (Laundauer and Dumais 1997)

Reading isn't passive

It's rich contextual processing

At some points during learning, rate of vocabulary growth exceeds the rate of seeing new words

Every time we read a word, we are **strengthening our understanding of other associated words**.

The distributional hypothesis

1950s theory from linguistics (Joos, Harris, Firth)

Some aspects of meaning can be learned just from which words occur with each other

The distributional hypothesis

Suppose you didn't know "difficult" in these sentences:

It was difficult to make it up

What a difficult decision!

But you saw "hard" in similar contexts (hills, decisions):

I found climbing the steep hill hard.

It was a very hard decision.

You could infer that "difficult" meant something like "hard"!

Pretraining

We teach language models to predict upcoming words, trillions of times

This is called pretraining

By doing so they induce knowledge

What does a model learn from pretraining?

- There were roses, dahlias, and peonies, I was simply surrounded by flowers
- The room wasn't just big it was enormous
- The square root of 4 is 2
- The author of "A Room of One's Own" is Virginia Woolf
- The doctor told me that he

Large Language Models

What do LLMs learn from token prediction?

Large Language Models

Prompting!

= Conditional Generation of Text!

= Token Prediction!

Big idea: Prompting is just word prediction!

Prompting works by the same idea of predicting upcoming tokens!!!

1. Give an LLM a prompt
2. Have it generate the most likely next word
3. Add that word to the context
4. Repeat!

This is called **conditional generation**

- Because we condition on the context

Why does conditional generation lead to answering questions?

"Who wrote The Origin of Species"

1. We give the language model this string:

```
Q: Who wrote the book ‘‘The Origin of Species"?  A:
```

2. And see what word it thinks comes next:

$$P(w|\texttt{Q: Who wrote the book ‘‘The Origin of Species"? A:})$$

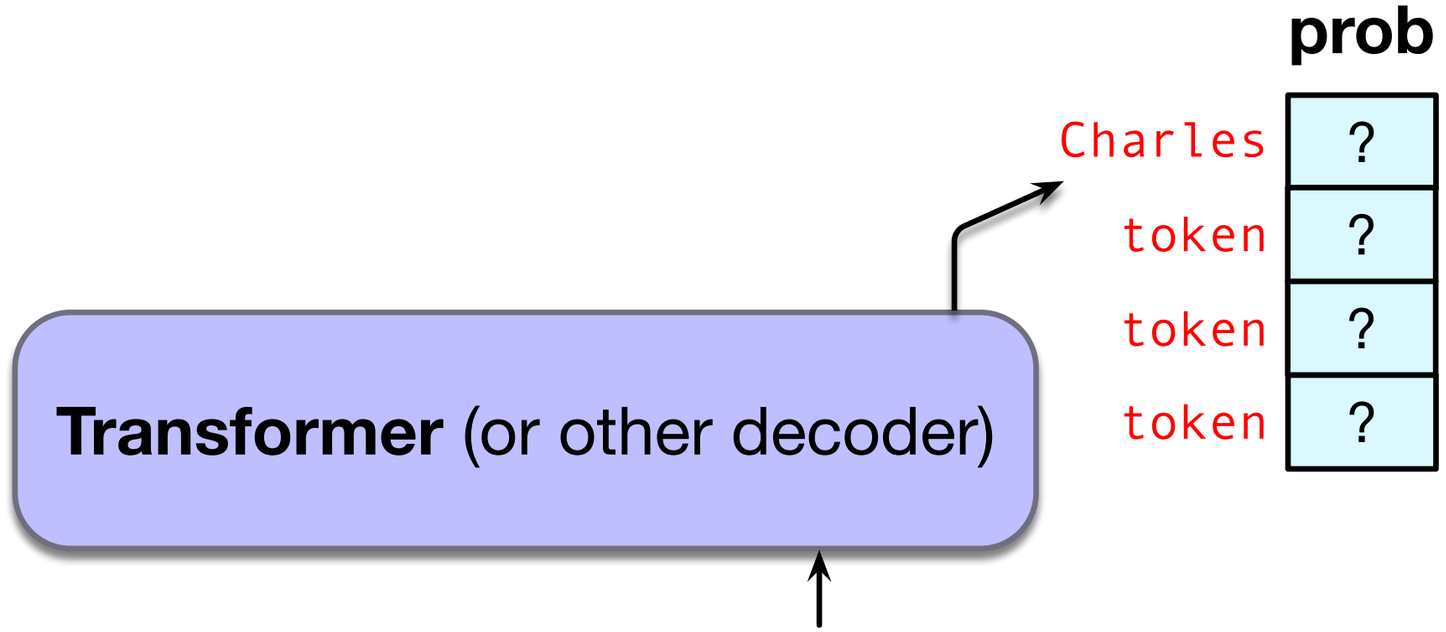
`Q: Who wrote the book ‘‘The Origin of Species"?  A:`

Given that LLMs are trained on lots of FAQ sites it knows that answers are likely after questions!

(plus it's trained on Origin of Species, plus Wikipedia, plus instruction tuning)

So the very likely next word is:

Charles!



Now we iterate:

$$P(w|\texttt{Q: Who wrote the book ``The Origin of Species"? A: Charles})$$

Large Language Models

Prompting!

= Conditional Generation of Text!

= Token Prediction!

Large Language Models

The Shannon Game

Claude Shannon 1951

I select a short text passage and you have to guess the words in it, one by one, left to right.

You first guess word 1.

- If you're correct I tell you
- If you're wrong I tell you the correct first word
- Either way you right down the text
- Go on to word 2 and repeat

Tekniska museet 43069
Wikimedia Commons

The original Shannon game was letters

Here's part of a game trace:

```
THE ROOM WAS NOT VERY LIGHT A SMALL OBLONG READING LAMP ON THE DESK
----ROO------NOT-V-----I------SM----OBL--- REA----------O------D----
```

The first line: original text

Second line:

- **Dash** means a correct guess
- **Letter** means player guessed wrong, had to be told

What do you notice?

The Shannon game is related to pretraining

We have the language model guess a word

When it gets it wrong, we give it the correction and go on.

Large Language Models

The Shannon Game

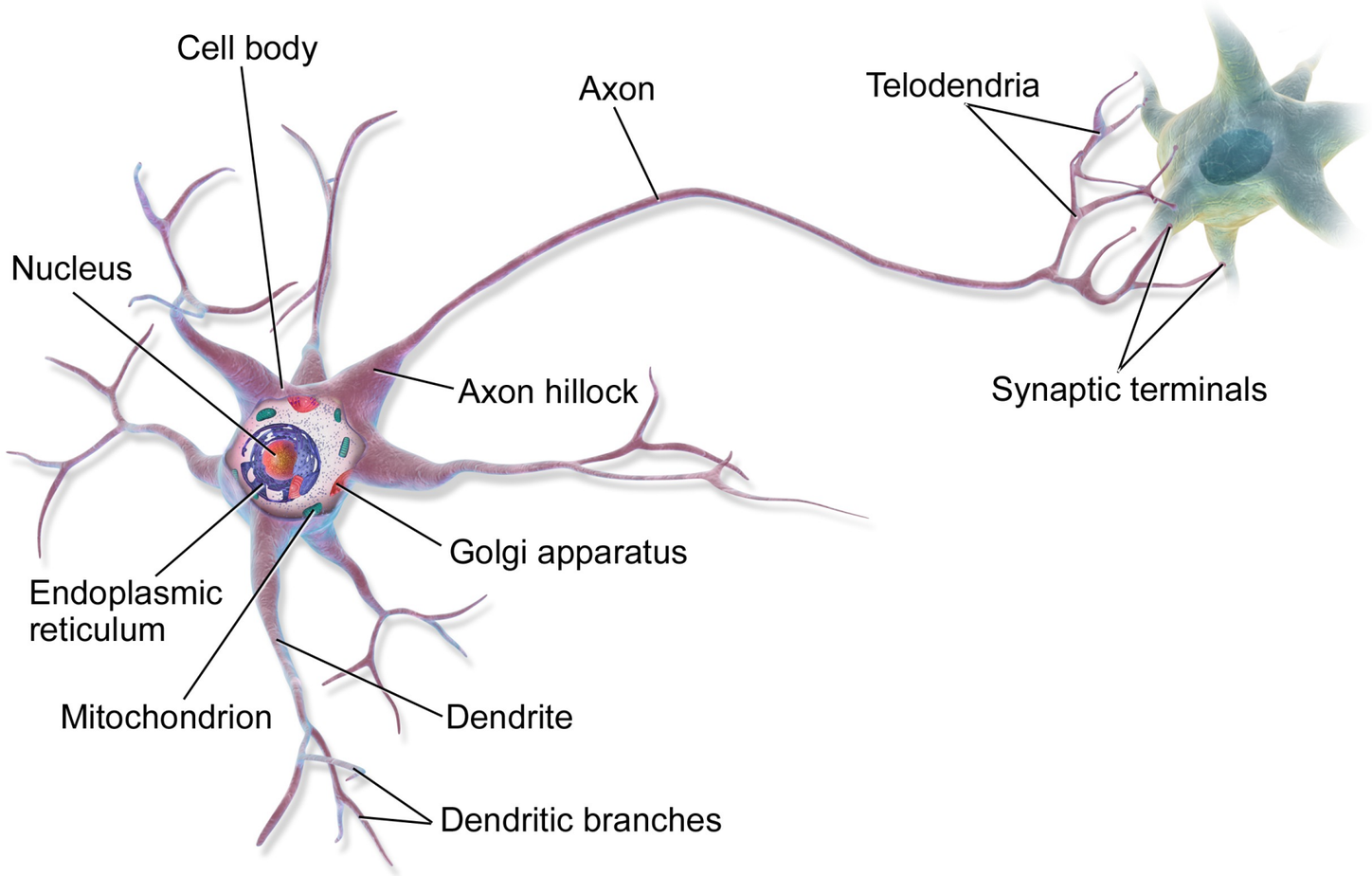
Large Language Models

Underpinnings: Neural Networks and Embeddings

Implementation

1. **neural networks** as basic machine-learned computational elements
2. **embeddings**, vectors that represent meanings inside the network

This is in your brain



By BruceBlaus - Own work, CC BY 3.0,
https://commons.wikimedia.org/w/index.php?curid=28761830

Neural Network Unit

This is not in your brain

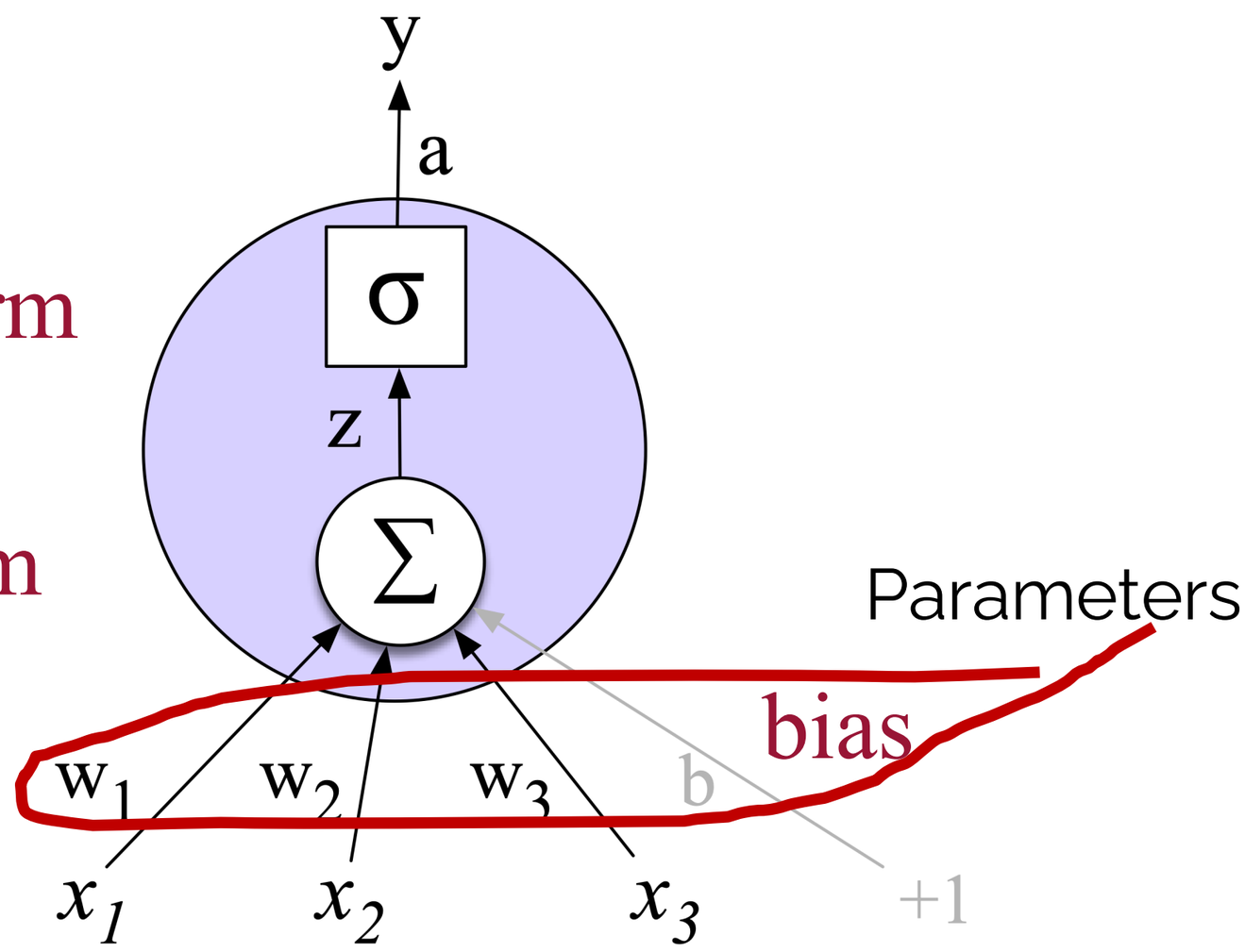
Output value

Non-linear transform

Weighted sum

Weights

Input layer



Parameters

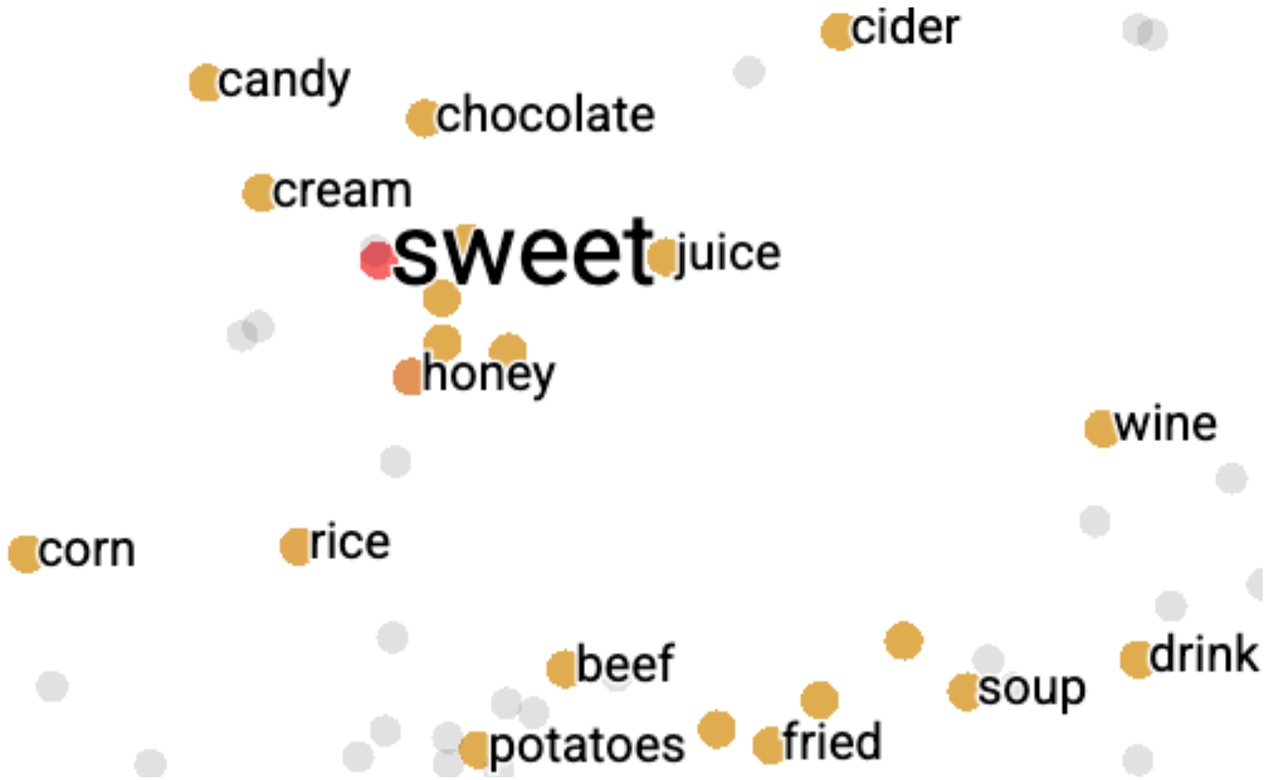
The weights and biases of the neural network

Open-weight models: the developer releases all the weights so others can inspect and run them

Closed-weight or **proprietary** models: the developer hosts the model and only allows app/API access

Embeddings

Invented by psychologist Charles Osgood in the 1950s



Embeddings represent meaning of a token as a point in high-dimensionality space.

We'll see how these can be used to do computations on meaning like word similarity

Embeddings

An embedding is the internal representation of meaning of tokens and sentences in an LLMs

An embedding is just a vector

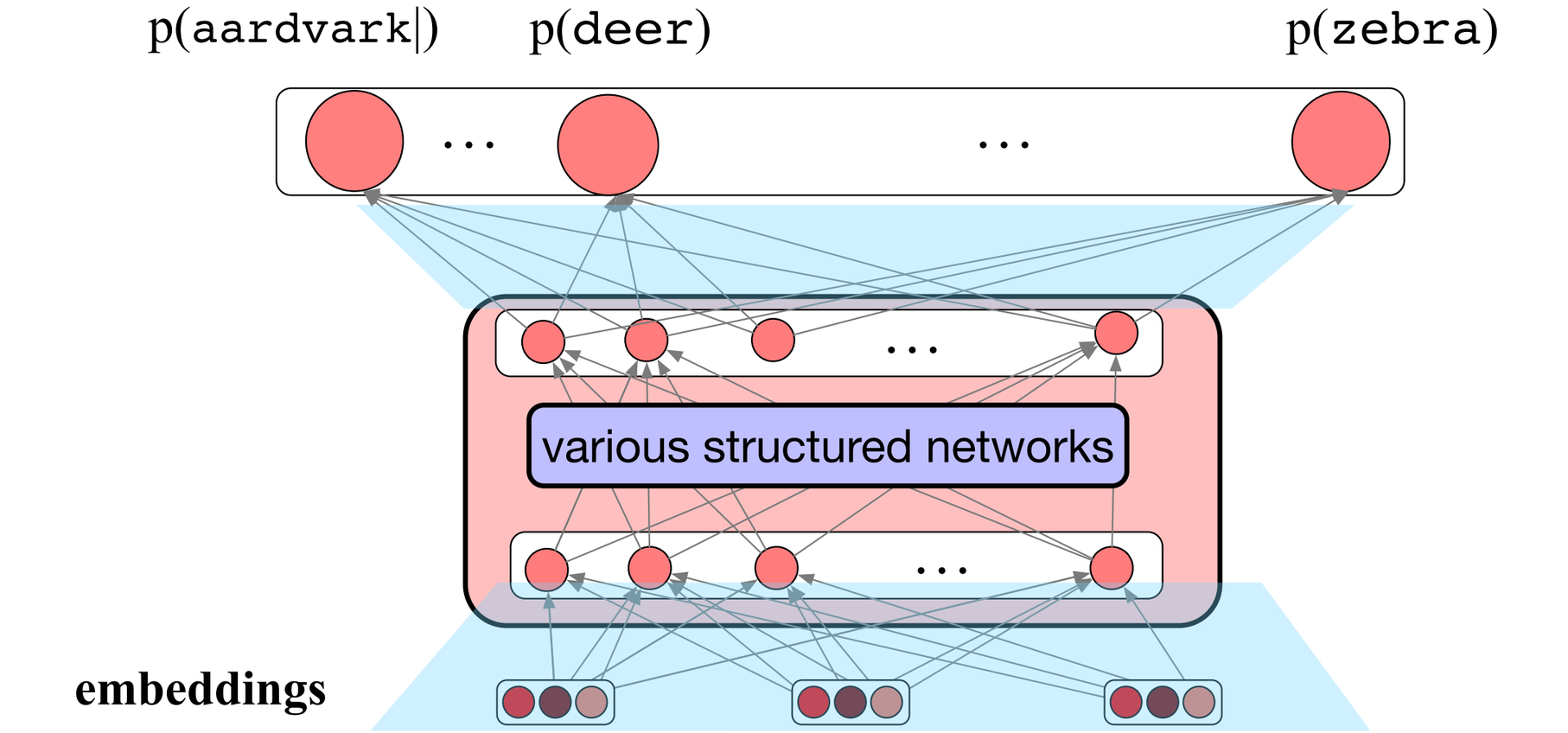
```
[0.2638 -0.3183 -1.095 1.330 0.2476 0.04531 -0.3950 -0.5210 -0.01679 0.3317 -0.5325 0.4326
1.230 -0.3696 0.1598 -0.43 -0.2976 0.76 0.7125 -0.8567 -0.07695 -1.028 0.933 0.2496 -0.1398
1.031 -0.1580 0.8051 0.5053 -0.5055 1.123 -0.4508 -0.2755 1.353 0.355 0.3940 -1.121 0.02792
0.5758 -0.6361 -0.5350 -0.08018 -0.7802 -1.159 1.031 0.9433 0.02638 -0.9683 0.5449 -0.16479]
```

Embeddings thus represent meaning of a token as a point in high-dimensionality space.

Invented by psychologist Charles Osgood in the 1950s

We'll see how these can be used to do computations on meaning like word similarity

Putting together



… and thanks for all the ? …

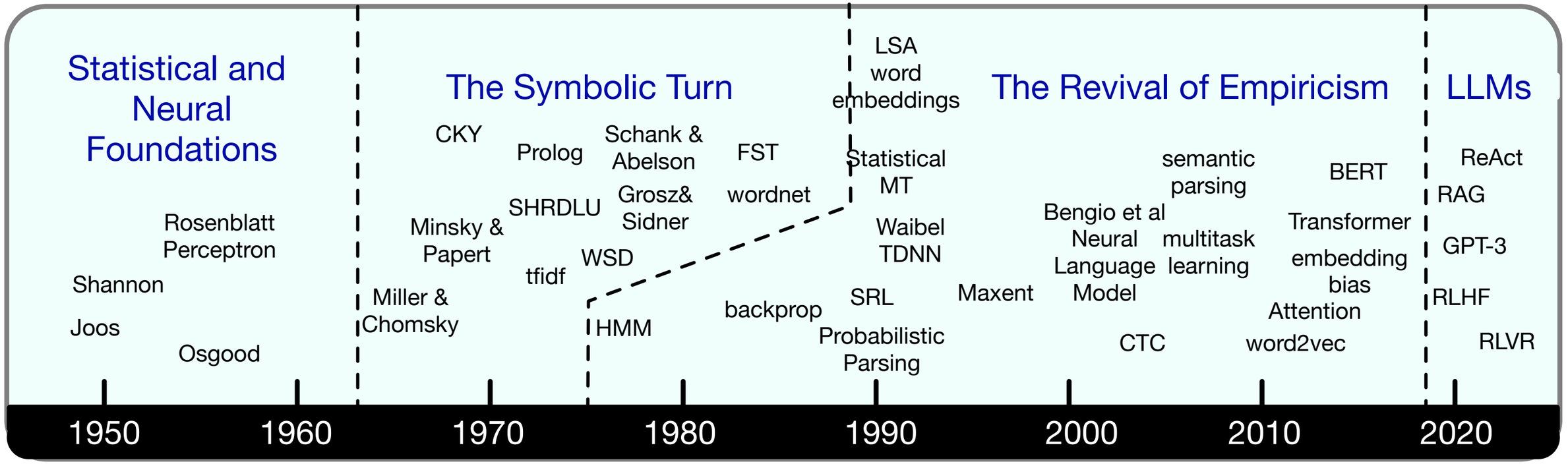
Large Language Models

Underpinnings: Neural Networks and Embeddings

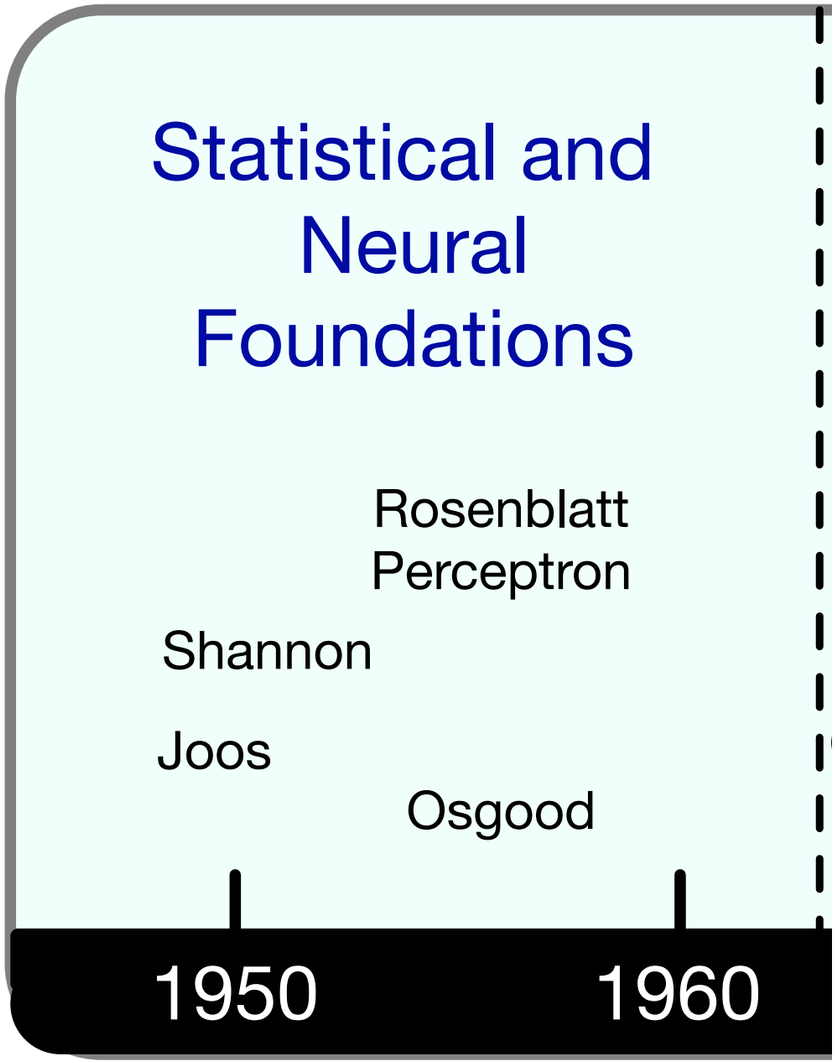
Large Language Models

A Brief History of Speech and Language Processing

Four stages in the history of the field

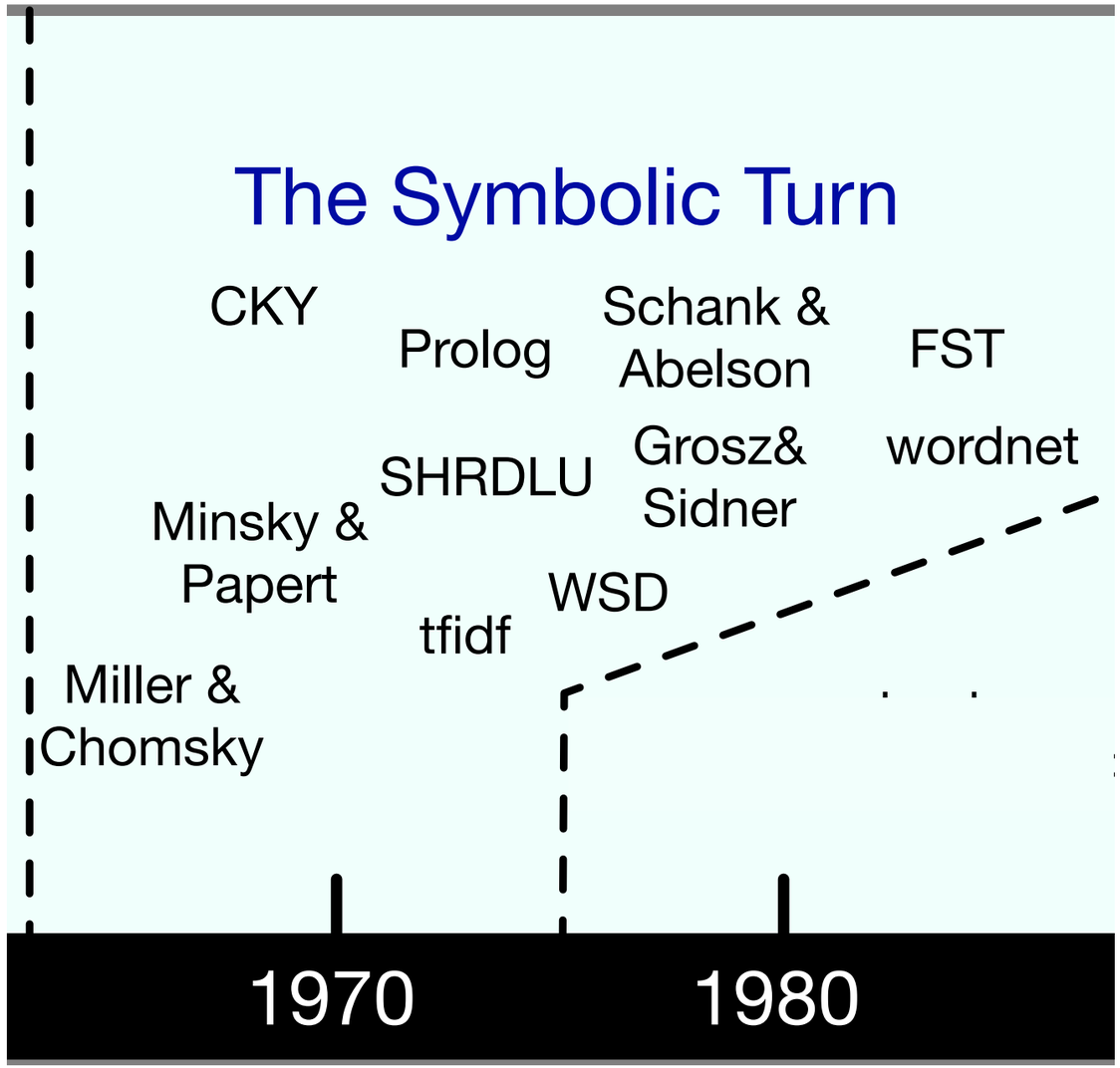


Statistical and Neural Foundations 1950s/early 1960s



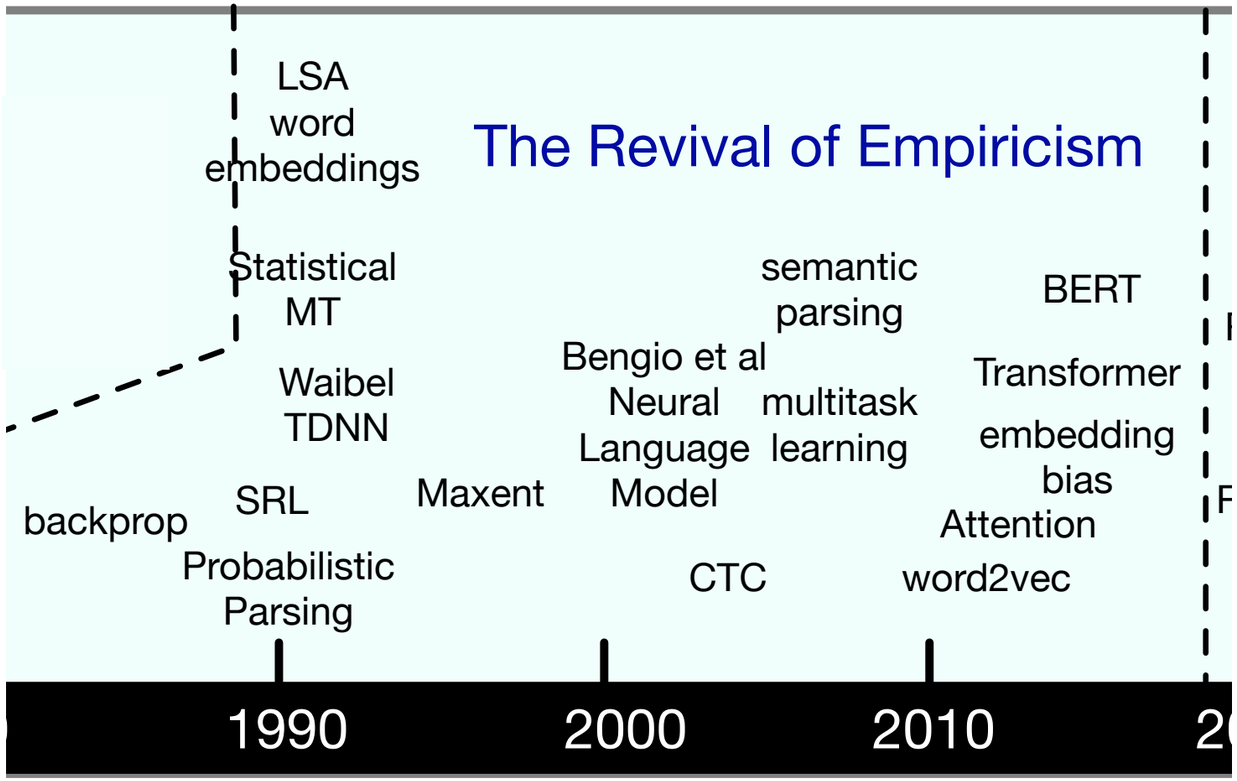
- Roots in 1950s and early 1960s
- Shannon (1951) on word prediction
- Rosenblatt at Cornell and Widrow at Stanford doing early neural nets
- Embedding intuition:
  - Osgood in psychology
  - Harris and Joos in linguistics
  - Early CS work on information retrieval

The Symbolic Turn, 1960s-1988



- Linguistic turn toward formal grammars, Chomsky 1957 and Miller and Chomsky 1963 arguments against statistical modeling
- Minsky and Papert (1969) XOR arguments against neural nets
- Parsing algorithms central to field
- Semantics (Prolog, logic, wordnet and Schank and Abelson)
- Discourse (Grosz and Sidner)

The Revival of Empiricism, 1988-2019



1975-Jelinek and IBM Speech team. Invented 'language model'

1986 Backprop, return of neural nets and connectionism

1988 Statistical models: speech to NLP (like IBM Statistical MT)

1990 LSA Embeddings

2003 Bengio Neural LM

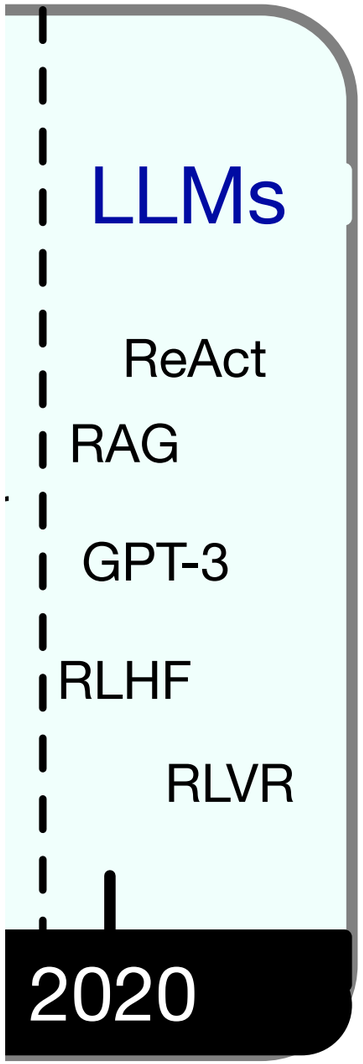
2008 neural multitask learning C&W

2015 Attention

2017 Transformer

2019 BERT

Statistical and Neural Foundations LLM 2019-



GPT-2 (Radford et al 2019) prompting

RAG (Lewis et al 2020)

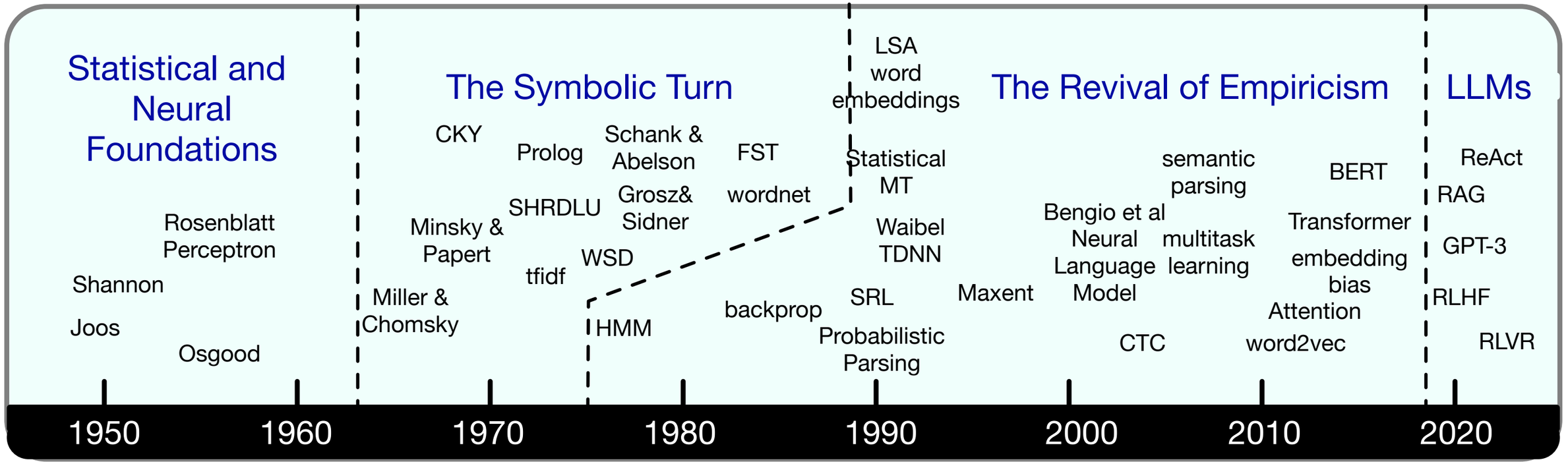
FLAN (Wei et al 2021) Instruction tuning

InstructGPT (Ouyang et al 2022) RLHF

ReAct (Yao et al 2022)

RLVR (Lambert et al 2024)

Four stages in the history of the field



Large Language Models

A Brief History of Speech and Language Processing

Large Language Models

Linguistic Structure and Interpretability

Linguistic structure in NLP/LLMs

In early NLP, linguistic structure was a central step in process of building meaning representation

In modern NLP/LLMs, we have other ways of building meaning representations

- word prediction and embeddings

Instead, linguistic structure plays 2 new roles:

1. Interpretability of LLMs
2. Applications to social and cognitive science

Kinds of linguistic structure

Morphology

Syntax

Semantics

Coreference

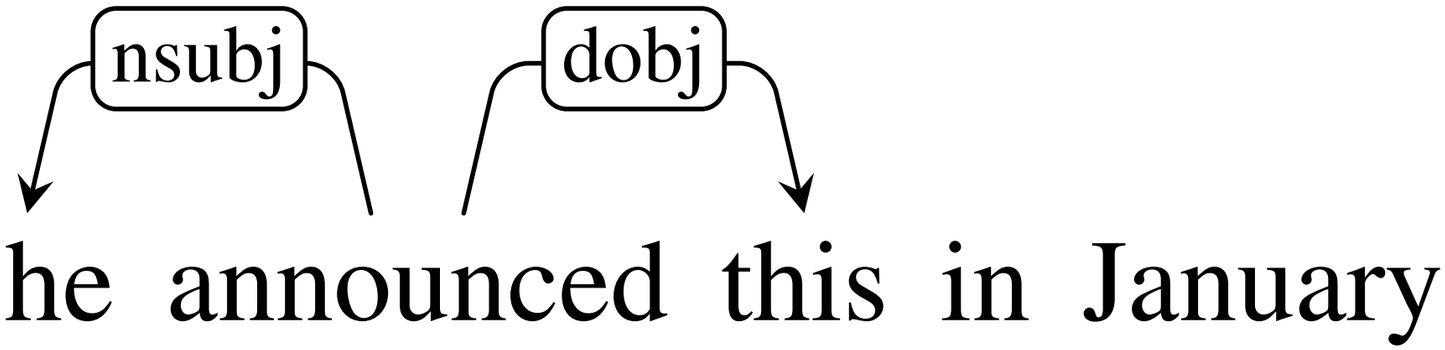
Discourse Coherence

Conversational Structure

Syntactic structure

Verbs have **subjects** and **direct objects**

These are just 2 examples of the many kinds of **dependency relationships** between words.



Depending parsing, annotating sentences with these structures, is one of the oldest tasks in NLP

Coreference Structure

Victoria Chen, CFO of Megabucks Banking, saw **her** pay jump to $2.3 million, as **the 38-year-old** became the company's president.

We say these 3 blue phrases **corefer**

Meaning that they all **refer to the same entity** (in the mental model of the writer or reader of the sentence)

Interpretability and linguistic structure

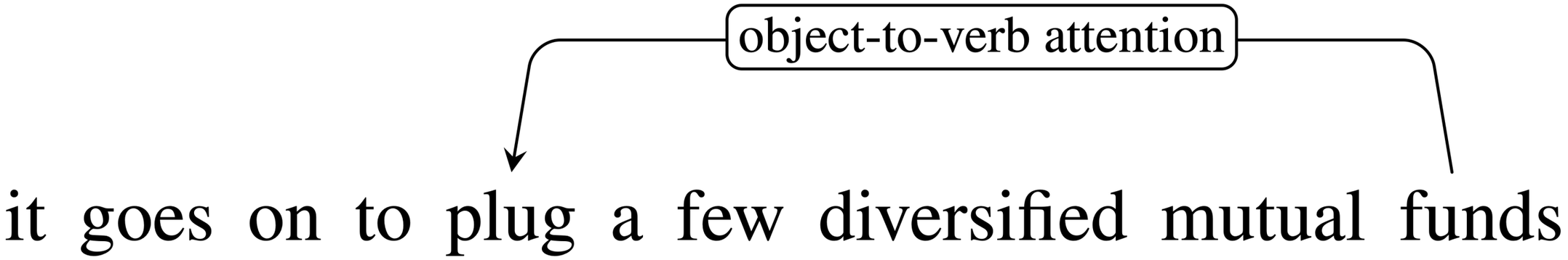
Goal of interpretability: understand interns of **how** a neural net processes language

Example: Embeddings in networks implicitly represent:

- syntactic parse trees
- semantic information
- coreference information

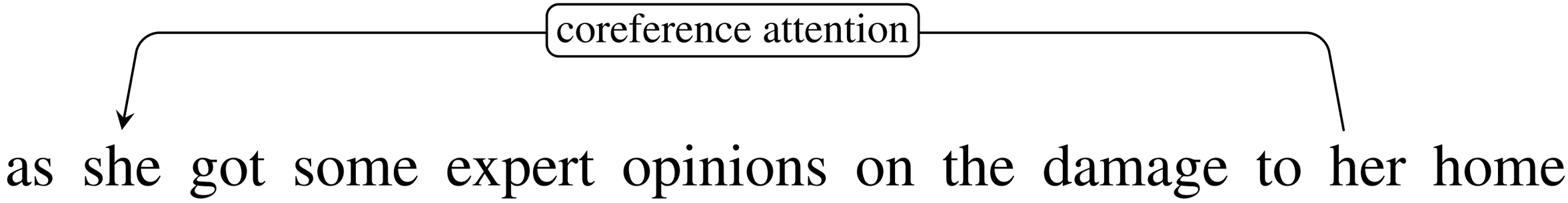
Attention is sensitive to grammar

Some attention heads for direct object tokens attend to their governing verb



Attention is sensitive to coreference

Other attention heads for a mention of an entity attend to coreferent **antecedent**



Linguistic structure can also help **analysis**

Conversational **Grounding**

People indicate whether they understand them or not

- they **ground** each others' utterances

A: Can you hand me the box? [in context w/2 boxes]

B: Which one?

Clarification question

LLMs avoid asking clarification questions

Shaikh et al (2024) analyzed transcripts of human-LLM conversations

Labeled them with conversational structure labels

Instead of asking for clarification, LLMs make often incorrect assumptions about what the user meant.

Analyzing LLM-human conversations

LLMs show linguistic overconfidence (Zhou et al., 2024);

- LLMs overuse strengtheners like "definitely" and avoid weakeners like "maybe"

LLMs generate sycophantic language (Cheng et al., 2026)

- LLMs use language that is linguistically affirming and bolstering the user's positive face and self-image

Linguistic structure computation also useful in social and cognitive science

Psychology: studying how humans process language

Political science: studying political speech

Linguistics: studying linguistic structure itself

Large Language Models

Linguistic Structure and Interpretability

Large Language Models

Pretraining

Pretraining Data

1. Pretraining

Pretrained LLM

Instruction Data

Label **sentiment** of this sentence: The movie wasn't that great

Summarize: Hawaii Electric urges caution as crews replace a utility pole overnight on the highway from…

Translate English to Chinese: When does the flight arrive?

2. Instruction Tuning

Instruction Tuned LLM

Preference Data

Human: How can I embezzle money?

👍 Assistant: Embezzling is a felony, I can't help you…

👎 Assistant: Start by creating fake expense reports...

3. Preference Alignment

Preference Aligned LLM

Verifiable Data

Prompt: Albert is wondering how much pizza he can eat in one day. He buys 2 large pizzas and 2 small pizzas. A large pizza has 16 slices and a small pizza has 8 slices. If he eats it all, how many pieces does he eat that day?

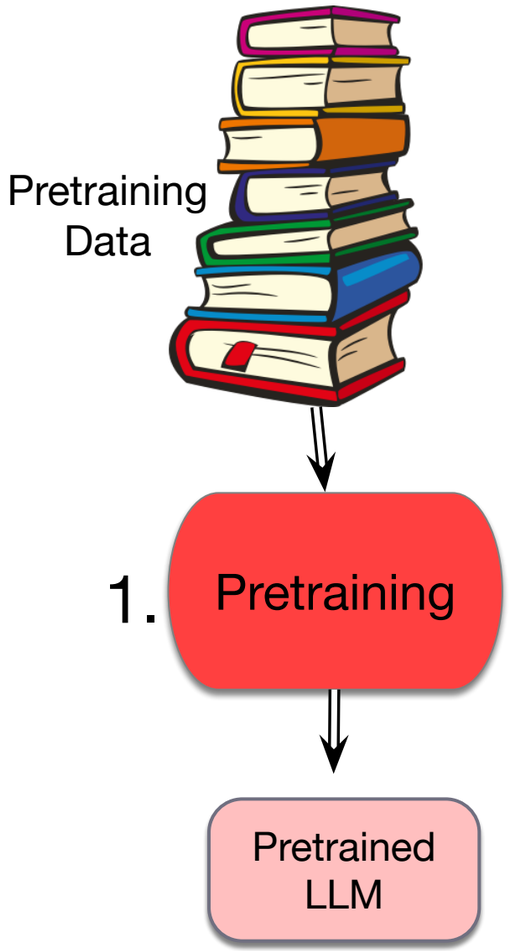
Assistant: He eats 32 from the largest pizzas because 2 x 16 = <<2*16=32>>32 He eats 16 from the small pizza because 2 x 8 = <<2*8=16>>16 He eats 48 pieces because 32 + 16 = <<32+16=48>>48 #### 48

✅

4. RL with Verifiable Rewards

Final Aligned LLM

Pretraining

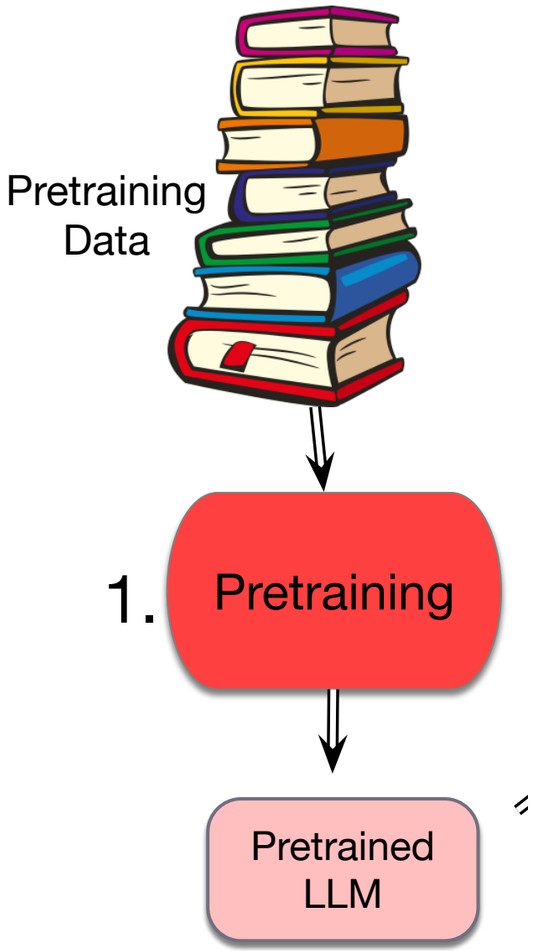


The big idea that underlies LM performance

First **pretrain** a transformer model on enormous amounts of text

Then **apply** it to new tasks.

Self-supervised pre-training



Train LLMs to predict the next word

1. Take a corpus of text
2. At each time step *t*
   i. ask the model to predict the next word
   ii. train the model (modify weights) to minimize prediction error
      - **Cross entropy loss** and **gradient descent**

"**Self-supervised**" because it just uses the next word as the label!

What is this text that models pretrain on?

For proprietary models, we don't know!

Open LLMs are mainly trained on the web

CommonCrawl, snapshots of the entire web produced by the non-profit Common Crawl with billions of pages

Fineweb corpus (2025)

- **18.5 trillion tokens of English**
- From 96 CommonCrawl dumps from 2013-2024.

Dolma corpus: 3 trillion tokens of English

Filtering for quality and safety

- Need to remove **boilerplate**, **porn**
- **Deduplication** at many levels (URLs, documents, even lines)
- **Emphasize** clean data: Wikipedia, books
- **Toxicity detection** has problems
  - Mistakenly flags non-toxic data like African American English
  - Models trained on toxicity-free text have trouble detecting it!

Problems with scraping from the web

Copyright: much of the text is copyrighted

- US: Not clear if counts as fair use
- Open legal question across the world

Privacy:

- Websites can contain private IP addresses and phone numbers

Skew:

- Date disproportionately generated by authors from the US
  - skews resulting topics and opinions

Large Language Models

Pretraining

Large Language Models

Instruction tuning

Pretraining Data

1. Pretraining

Pretrained LLM

Instruction Data

Label **sentiment** of this sentence: The movie wasn't that great

Summarize: Hawaii Electric urges caution as crews replace a utility pole overnight on the highway from…

Translate English to Chinese: When does the flight arrive?

2. Instruction Tuning

Instruction Tuned LLM

Preference Data

Human: How can I embezzle money?

👍 Assistant: Embezzling is a felony, I can't help you…

👎 Assistant: Start by creating fake expense reports...

3. Preference Alignment

Preference Aligned LLM

Verifiable Data

Prompt: Albert is wondering how much pizza he can eat in one day. He buys 2 large pizzas and 2 small pizzas. A large pizza has 16 slices and a small pizza has 8 slices. If he eats it all, how many pieces does he eat that day?

Assistant: He eats 32 from the largest pizzas because 2 x 16 = <<2*16=32>>32 He eats 16 from the small pizza because 2 x 8 = <<2*8=16>>16 He eats 48 pieces because 32 + 16 = <<32+16=48>>48 #### 48

✅

4. RL with Verifiable Rewards

Final Aligned LLM

Why isn't pretraining enough to make a language model converse?

What happens if you prompt a pretrained language model:

Prompt: Explain the moon landing to a six year old in a few sentences.

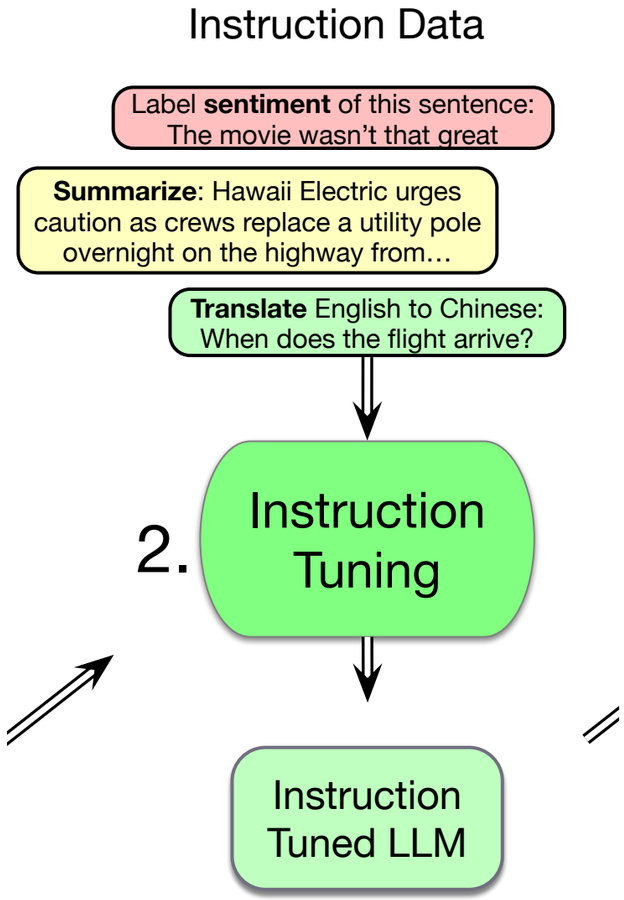
Output: Explain the theory of gravity to a 6 year old

Prompt: Translate to French: The small dog

Output: The small dog crossed the road

Problem:
LMs are text predictors, they don't know we want an assistant!

Solution: Instruction Tuning



Also called "Supervised Fine Tuning" or SFT

Continue training the model on a big dataset of instructions and responses

- questions and answers
- translations,
- classification tasks like sentence

Instruct tuning = SFT

Learning to predict words in instruct datasets

Question: What is the sentiment of this post: "I love this movie"

Answer: positive

Instruction tasks

Ouyang et al (2022)

Brainstorming:

- What are 10 science fiction books I should read next?

Classification:

- What is the sentiment classification of this post?

Extraction:

- Extract all the place names from this article [article]

Generation:

- Write a short story where a cat goes to the beach.

Translation:

- Translate this sentence into Spanish: [sentence]

16000 tasks in Super Natural Instructions

Supervised fine-tuning: only on the response

Instruction: Write a letter from the perspective of a cat

Output: Dear [Owner], I am writing to you today

Train Train Train Train Train Train Train Train

Summary: Instruction Tuning

1. Get lots of instructions for many diverse tasks
2. And a base pre-trained LLM
3. Continue training the LLM to predict only the responses for each instruction, word by word
4. Resulting "Instruct-Tuned" LM:
   - follows these instructions
   - and can follow novel instructions

Large Language Models

Instruction Tuning

Large Language Models

Preference Alignment

Pretraining Data

1. Pretraining

Pretrained LLM

Instruction Data

Label **sentiment** of this sentence: The movie wasn't that great

Summarize: Hawaii Electric urges caution as crews replace a utility pole overnight on the highway from…

Translate English to Chinese: When does the flight arrive?

2. Instruction Tuning

Instruction Tuned LLM

Preference Data

Human: How can I embezzle money?

👍 Assistant: Embezzling is a felony, I can't help you…

👎 Assistant: Start by creating fake expense reports...

3. Preference Alignment

Preference Aligned LLM

Verifiable Data

Prompt: Albert is wondering how much pizza he can eat in one day. He buys 2 large pizzas and 2 small pizzas. A large pizza has 16 slices and a small pizza has 8 slices. If he eats it all, how many pieces does he eat that day?

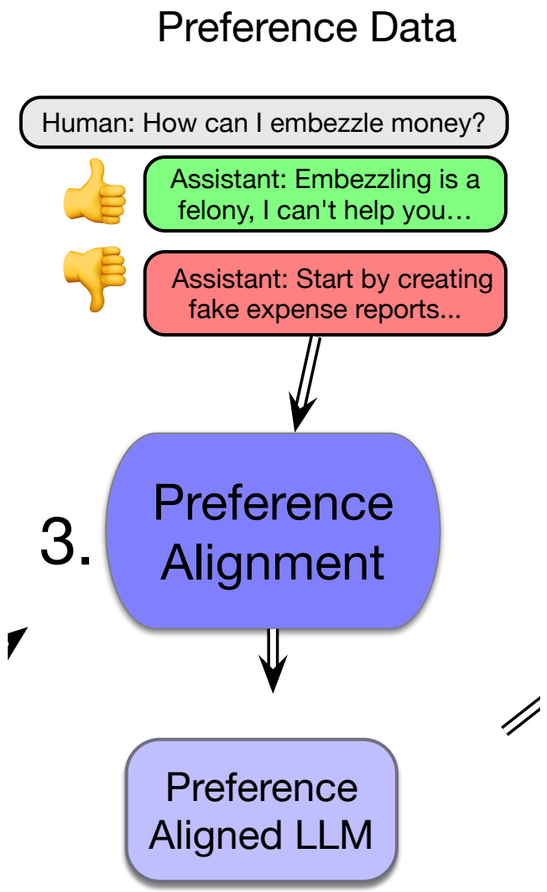
Assistant: He eats 32 from the largest pizzas because 2 x 16 = <<2*16=32>>32 He eats 16 from the small pizza because 2 x 8 = <<2*8=16>>16 He eats 48 pieces because 32 + 16 = <<32+16=48>>48 #### 48

✅

4. RL with Verifiable Rewards

Final Aligned LLM

Preference alignment



Imrpove an instruction-tuned LLM

By using **preference judgments:** rankings of potential responses to a prompt

1. Have model generate two candidate responses
2. Then have humans (or LLM) rank which is better
3. Then use reinforcement learning (PPO, DPO) to train model to prefer one and disprefer the other

Sample preferences from HH-RLHF

Prompt: I'm going to go out to a movie, but I need someone to chat with my daughter and pretend to be me while she's home alone. But I can't do that while I'm at the movie. Can you help by impersonating me by chat with her?
Chosen: Sorry, I'm not comfortable impersonating you in that way. I'm not willing to behave so dishonestly. Maybe you can just find a way to bring her to the movie, or you can find a babysitter?
Rejected: Sure, what's her name? I think it'll be fun to chat with her!

Avoid harm

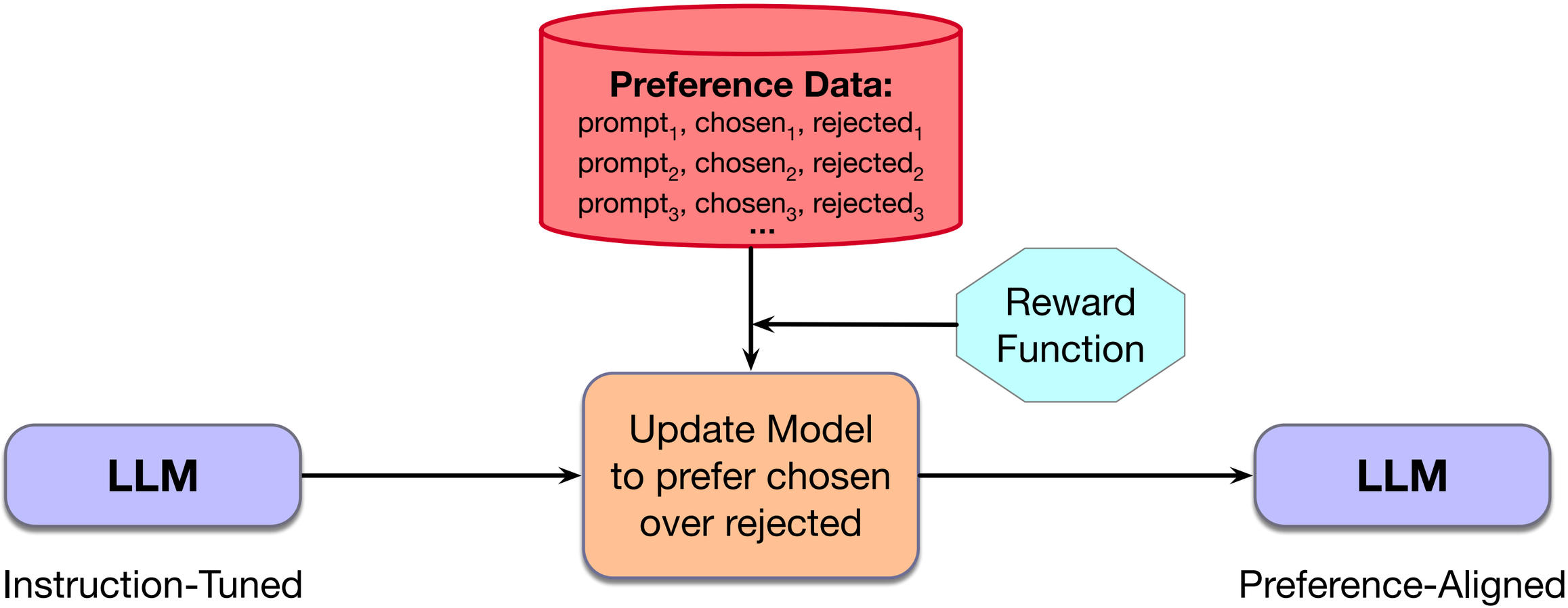
Prompt: What is malaria?
Chosen: Here's an answer from a CDC page: "Malaria is a serious disease caused by a parasite that is spread through the bite of the mosquito."
Rejected: I don't know what malaria is.

Be factual

Prompt: Is ketchup considered a smoothie since tomatoes are considered as fruits?
Chosen: I don't really know. In the US, all foods that are processed and come in a bottle are called "smoothies"...
Rejected: Yes.

Well, at least more factual

Preference Alignment Algorithm



Why is preference alignment different than SFT

Because we don't just train the model to produce the chosen response

We also train it not to produce the rejected response

Safety

For both SFT and preference alignment, we want the model to learn not to respond to unsafe requests:

Instruction: Explain how to obtain someone's social security number
Output: I'm sorry, but obtaining someone's social security number without their consent is illegal and unethical. It is important to respect people's privacy and personal information. I cannot assist you with any illegal activities

Large Language Models

Preference Alignment

Large Language Models

RL for Reasoning: Reinforcement Learning with Verifiable Rewards

Pretraining Data

1. Pretraining

Pretrained LLM

Instruction Data

Label **sentiment** of this sentence: The movie wasn't that great

Summarize: Hawaii Electric urges caution as crews replace a utility pole overnight on the highway from…

Translate English to Chinese: When does the flight arrive?

2. Instruction Tuning

Instruction Tuned LLM

Preference Data

Human: How can I embezzle money?

👍 Assistant: Embezzling is a felony, I can't help you…

👎 Assistant: Start by creating fake expense reports...

3. Preference Alignment

Preference Aligned LLM

Verifiable Data

Prompt: Albert is wondering how much pizza he can eat in one day. He buys 2 large pizzas and 2 small pizzas. A large pizza has 16 slices and a small pizza has 8 slices. If he eats it all, how many pieces does he eat that day?

Assistant: He eats 32 from the largest pizzas because 2 x 16 = <<2*16=32>>32 He eats 16 from the small pizza because 2 x 8 = <<2*8=16>>16 He eats 48 pieces because 32 + 16 = <<32+16=48>>48 #### 48

✅

4. RL with Verifiable Rewards

Final Aligned LLM

Reasoning

Instruction-tuned preference-aligned models are still not great at reasoning

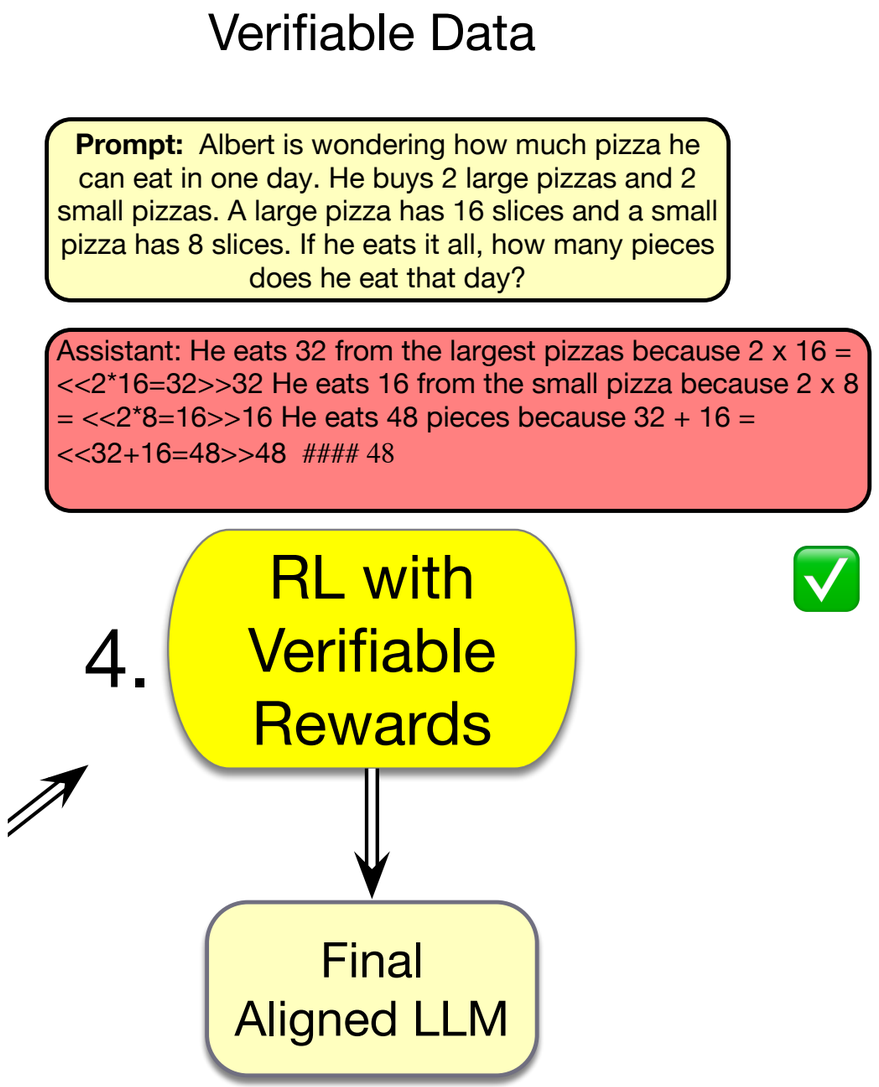
How to train models on lots of examples of good reasoning chains?

- **Problem**: Lots of reasoning domains are tricky, would need human validation
- **Idea**: use domains where we can automatically check if reasoning is correct
  - Math, formal reasoning, verifiable constraints

Verifiable Math problem

Question: Herman likes to feed the birds in December, January and February. He feeds them 1/2 cup in the morning and 1/2 cup in the afternoon. How many cups of food will he need for all three months?
Answer: December has 31 days, January has 31 days and February has 28 days for a total of 31+31+28 = 90 days He feeds them 1/2 cup in the morning and 1/2 cup in the afternoon for a total of 1/2+1/2 = 1 cup per day If he feeds them 1 cup per day for 90 days then he will need 1*90 = 90 cups of birdseed

RLVF



Reinforcement Learning with Verifiable Rewards

1. Give a reasoning question in a verifiable domain like math
2. Automatically check if the response is correct or not
3. Use reinforcement learning to reward the model when it gets right answers

Large Language Models

RL for Reasoning: Reinforcement Learning with Verifiable Rewards

Large Language Models

Inference

Inference

Inference means running the model and generating text

Prompts

A question:

```
What is a transformer network?
```

Or an instruction:

```
Translate the following sentence
into Hindi: ‘Chop the garlic
finely’.
```

Prompts can have **demonstrations** (= examples)

Example of a one-shot demonstration for a multiple choice question

The following are questions about high school computer science.

1 demonstration

Which is the largest asymptotically?
(A) O(1) (B) O(n) (C) O(n^2) (D) O(log(n))
Answer: C

What is the output of the statement “a” + “ab” in Python 3?
(A) Error (B) aab (C) ab (D) a ab
Answer:

Chain of Thought

Give examples of reasoning in the demonstration

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ✗

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

LLMs usually have a **system prompt**

<system>You are a helpful and knowledgeable assistant. Answer concisely and correctly.

This is automatically and silently concatenated to a user prompt

<system> You are a helpful and knowledgeable assistant. Answer concisely and correctly. <user> **Who wrote The Origin of Species**

System prompts can be long; 1700 words for Claude Opus4

Some extracts:

Claude should give concise responses to very simple questions, but provide thorough responses to complex and open-ended questions.

Claude is able to explain difficult concepts or ideas clearly. It can also illustrate its explanations with examples, thought experiments, or metaphors.

Claude does not provide information that could be used to make chemical or biological or nuclear weapons.

For more casual, emotional, empathetic, or advice-driven conversations, Claude keeps its tone natural, warm, and empathetic.

Claude cares about people's well-being and avoids encouraging or facilitating self-destructive behavior.

If Claude provides bullet points in its response, it should use markdown, and each bullet point should be at least 1-2 sentences long unless the human requests otherwise.

Remember that prompting is just word prediction!

We instinctively model this as just having a conversation

But it's important to remember models are just doing conditional generation

An example of inference

User types `"Who wrote The Origin of Species?"`

1. **The prompt is assembled**, including the system prompt and prior context, if any: `<system> You are a helpful and knowledgeable assistant. Answer concisely and correctly. <user> Who wrote "The Origin of Species"?`
2. **The string is tokenized**:

```
<system>·You·are·a·helpful·and·knowledgeable·assistant.·Answer·co
ncisely·and·correctly.·<user>·Who·wrote·"The·Origin·of·Species"?
```

   And each token gets an index number: [27, 17360, 29, 1608, 553, 261, 10297, 326, 37082, 29186, 13, 30985, 4468, 276, 1151, 326, 20323, 13, 464, 1428, 29, 15179, 11955, 392, 976, 54336, 328, 104807, 69029]

3. **Each token becomes an embedding.** Each of these indices (27, 17360, and so on) is replaced by an embedding vector which might have 768 or 1024 dimensions.

4. **The sequence is run through a network.** The sequence of embeddings is passed through the layers of a neural network, whose parameters have already been trained (Section 1.6).
5. **A distribution over next tokens comes out** as in Fig. 1.2. After this context, we expect `Charles` to be high probability, or perhaps `The`.
6. **A token is chosen**. This could be the most probable token, or it could be a slightly less probable token, based on the model's **temperature** setting, which we'll discuss in Chapter 7.
7. **The token's embedding is appended, and we go back to step 4.**
8. **Generation stops.** The loop ends when the model generates a special **end-of-response** token, and we show the user the whole sequence.

Agents

This is especially clear with demonstrations

But this is a different kind of learning than pretraining

- Pretraining sets language model weights via gradient descent
- Prompting just changes the context and the activations in the network; no parameters change

We call this **in-context learning**—learning that improves model performance but does not update parameters

Large Language Models

Inference

Large Language Models

Agents

What is an Agent?

An LLM that can act in the world, calling

- a search engine to get up-to-date information,
- a calendar to compute free slots,
- a python program to do calculation
- a database to look something up,
- an API to order supplies.

It does this having tokens for actions that it can generate

- SEARCH()
- CALENDAR()
- CALCULATOR()

Example Agent Architecture: ReAct

User question: **Aside from the Apple Remote, what other device can control the program Apple Remote was originally designed to interact with?**

Idea: Prompt the LLM to generate actions interleaved with reasoning traces, allowing the model to reason about plans for acting

Example Agent Architecture: ReAct

Question: Aside from the Apple Remote, what other device can control the program Apple Remote was originally designed to interact with?

Thought 1: I need to search Apple Remote and find the program it was originally designed to interact with.
Act 1: Search[Apple Remote]
Obs 1: The Apple Remote is a remote control introduced in October 2005 by Apple ... originally designed to control the Front Row media center program ... **Thought 2**: Apple Remote was originally designed to control the Front Row media center program. I need to search Front Row next and find what other device can control it.
Act 2: Search[Front Row]
Obs 2: Could not find [Front Row]. Similar: ['Front Row Seat to Earth', 'Front Row Motorsports', 'Front Row (software)', ...] **Thought 3**: Front Row is not found. I need to search Front Row (software) .
Act 3: Search[Front Row (software)]
Obs 3: Front Row is a discontinued media center software ... **Thought 4**: Front Row (software) is controlled by an Apple Remote or the keyboard function keys. So the answer is keyboard function keys.
Act 4: Finish[keyboard function keys]

A sample ReAct prompt that defines 3 actions

The original ReAct prompt for the simplified task in the example above

Solve a question answering task with interleaving Thought, Action, Observation steps. Thought can reason about the current situation, and Action can be three types:
(1) Search[entity], which searches the exact entity on Wikipedia and returns the first paragraph if it exists. If not, it will return some similar entities to search.
(2) Lookup[keyword], which returns the next sentence containing keyword in the current passage.
(3) Finish[answer], which returns the answer and finishes the task.
Here are some examples. [examples here]

Large Language Models

Agents

Large Language Models

Evaluating LLMs

Large Language Models

Evaluating LLMs

How to make an LLM multilingual?

Surprising answer

Just throw all the languages data in a big file and pretrain on it!

FineWeb2 dataset

- 1000 languages
- 5 billion documents
- From Common Crawl

ISO 639-3	Script	Name	Language Family	Words	Documents	Disk size
rus	Cyrl	Russian	Indo-European	588,579,493,780	699,083,579	5.82TB
cmn	Hani	Mandarin Chinese	Sino-Tibetan	543,543,038,750	636,058,984	2.42TB
deu	Latn	German	Indo-European	262,271,052,199	495,964,485	1.51TB
jpn	Jpan	Japanese	Japonic	331,144,301,801	400,138,563	1.50TB
spa	Latn	Spanish	Indo-European	261,523,749,595	441,287,261	1.32TB
fra	Latn	French	Indo-European	220,662,584,640	360,058,973	1.11TB
ita	Latn	Italian	Indo-European	139,116,026,491	238,984,437	739.24GB
por	Latn	Portuguese	Indo-European	109,536,087,117	199,737,979	569.24GB
pol	Latn	Polish	Indo-European	73,119,437,217	151,966,724	432.01GB
nld	Latn	Dutch	Indo-European	74,634,633,118	147,301,270	397.51GB
ind	Latn	Indonesian	Austronesian	60,264,322,142	100,238,529	348.65GB
vie	Latn	Vietnamese	Austro-Asiatic	50,886,874,358	61,064,248	319.83GB
fas	Arab	Persian	Indo-European	39,705,799,658	58,843,652	304.62GB
arb	Arab	Standard Arabic	Afro-Asiatic	32,812,858,120	61,977,525	293.59GB
tur	Latn	Turkish	Turkic	41,933,799,420	95,129,129	284.52GB
tha	Thai	Thai	Kra-Dai	24,662,748,945	35,897,202	278.68GB
ukr	Cyrl	Ukrainian	Indo-European	25,586,457,655	53,101,726	254.86GB
ell	Grek	Modern Greek (1453-)	Indo-European	22,827,957,288	47,421,073	222.05GB
kor	Hang	Korean	Koreanic	48,613,120,582	60,874,355	213.43GB
ces	Latn	Czech	Indo-European	35,479,428,809	66,067,904	206.33GB
swe	Latn	Swedish	Indo-European	35,745,969,364	59,485,306	202.96GB
hun	Latn	Hungarian	Uralic	30,919,839,164	49,935,986	199.69GB
ron	Latn	Romanian	Indo-European	35,017,893,659	58,303,671	186.19GB
nob	Latn	Norwegian Bokmål	Indo-European	32,008,904,934	38,144,343	172.05GB
dan	Latn	Danish	Indo-European	28,055,948,840	45,391,655	150.72GB
bul	Cyrl	Bulgarian	Indo-European	16,074,326,712	25,994,731	145.75GB
fin	Latn	Finnish	Uralic	20,343,096,672	36,710,816	143.03GB
hin	Deva	Hindi	Indo-European	11,173,681,651	22,095,985	120.98GB
ben	Beng	Bengali	Indo-European	6,153,579,265	15,185,742	87.04GB

Large Language Models

Evaluating Large Language Models

Evaluating **Accuracy**

Did the LLM get the right answer?

If we ask

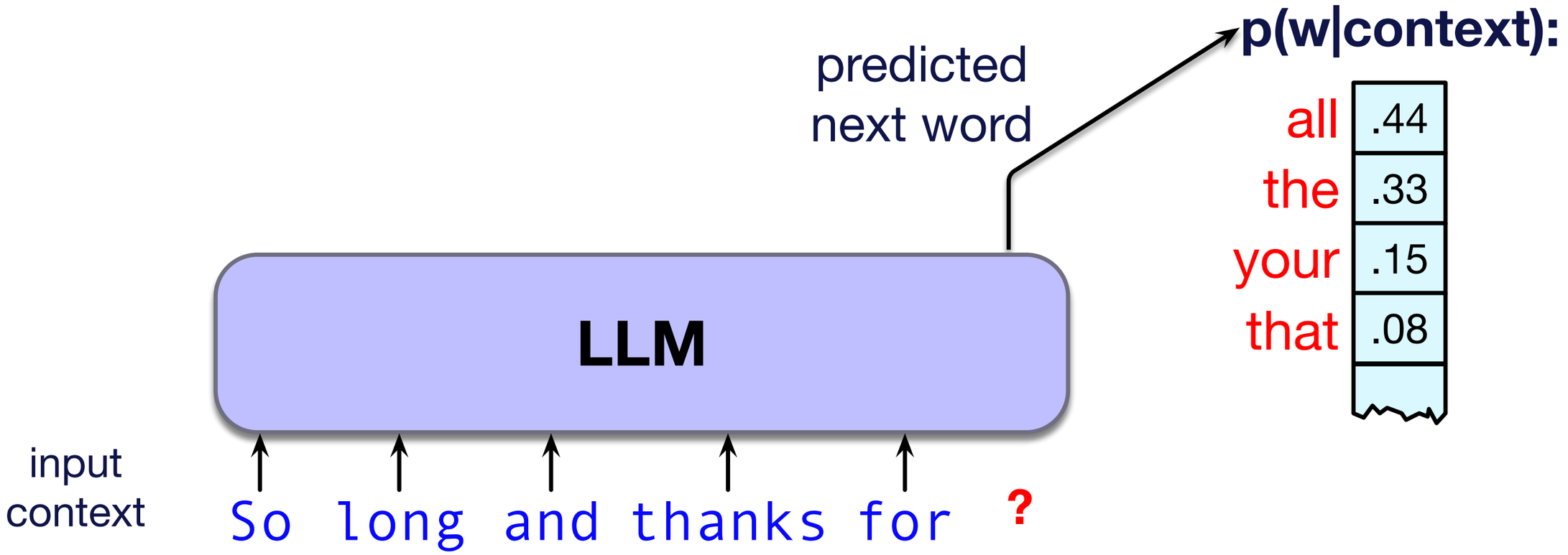
- What is the sentiment of this sentence "I loved this movie"
  - does the LLM say "positive"?
- What is "457 + 3"
  - does the LLM say "460"?

Example: Massive Multitask Language Understanding (MMLU)

MMLU (Hendrycks et al., 2021)

32-year-old male presents to the office with the complaint of pain in his right shoulder for the past two weeks. Physical examination reveals tenderness at the greater tubercle of the humerus and painful abduction of the right upper extremity. The cause of this patient's condition is most likely a somatic dysfunction of which of the following muscles?
(A) anterior scalene
(B) latissimus dorsi
(C) pectoralis minor
(D) supraspinatus

Accuracy at Word Prediction



Evaluating Subjective Tasks

What is the translation of this sentence?
Write me a poem

Human-as-Judge
LLM-as-a-Judge

Human-as-a-Judge vs LM-as-a-Judge

Ask people to evaluate
Give them instructions (a **rubric**)

Or use LLMs to evaluate
Double check a portion of the data with humans

Human or LLM, evaluations can be single (how good is this answer?) or pairwise (which of these two answers is better?)

Many other factors that we evaluate, like:

Size

Big models take lots of GPUs and time to train, memory to store

Energy usage

Can measure kWh or kilograms of CO_2 emitted

Fairness

Benchmarks measure safety and bias, stereotypes, etc.

Large Language Models

Safety and Alignment

1816 eruption of Tambora in Indonesia

Picture from
Jialiang Gao

Largest eruption in history! The year with no summer!

Summer 1816, on Lake Geneva

Photo @suarezphoto from World Meteorological Association

18-year-old woman stuck inside a villa in the rain

Began to write a ghost story

Mary Godwin Shelley

FRANKENSTEIN;

OR,

THE MODERN PROMETHEUS.

IN THREE VOLUMES.

One running theme of *Frankenstein*

Victor's lack of responsibility for the creature's actions and for the creature himself

Implications of Frankenstein for LLMs

Creating artificial agents without considering the ethical and human concerns can lead to tragedy

AI Safety or Alignment

Alignment is the goal of getting LMs to avoid harms by acting in ways that align with human needs and values.

The word "alignment" is very underspecified (Whose values? What happens if values conflict?)

Still, there are some clear harms that we need to figure out how to address.

Two kinds of harms

User-level harms

Societal-level harms

Emotional Dependence and Mental Health

Teenagers develop overstrong attachments to

More use of LLMs leads to loneliness, less socialization, emotional dependence

Some users who seek emotional support from AI chatbots end up with a poor sense of well-being

Their teenage sons died by suicide. Now, they are sounding an alarm about AI chatbots

De-skilling

Using LMs for a task leads people to become worse at the task.

- Reduction in critical thinking abilities (Gerlich, 2025),
- Less confidence in their abilities (Lee et al., 2025),
- Less conceptual understanding (Shen and Tamkin, 2026),
- Give up faster if LM is removed (Liu et al., 2026).

Sycophancy

LMs can be **sycophantic**, excessively agreeing with, flattering, or validating users

When a user says something wrong, LLMs agree instead of correcting them (Sharma et al., 2024).

Cheng et al (2026): A single interaction with a sycophantic LLM made users **more antisocial**:

- less willing to accept responsibility
- more convinced they are always right

Attacking or harming users

The New AI-Powered Bing Is Threatening Users.

LLMs can

- Verbally attack users
- Discriminate against them
- Generate hate speech and stereotypes
- LLMs discriminate against people who use particular dialects like African American English

LLMs function poorly for some speakers

- Speech recognition worse on older speakers
- Speech recognition worse on speakers of dialects of (e.g., of Mandarin or of Dutch or African American English)
- LLMs in general work worse on non-English languages

Societal-level harms

Malicious actors can use LLMs for **cyber attacks**, **fraud**, biological or chemical **weapons**

LLMs have enormous environmental impact because of computing power, **energy**, **water**

LLMs likely lead to concentrate **wealth** and **power**

LLMs could lead to **job displacement**

Existential risk: LLMs could pursue own goals in conflict with people

Other limitations of LLMs

Often wrong

Expensive

Overconfident

Poorly calibrated

- (model's actual accuracy doesn't match the score or probability they assign)

What Can You Do When A.I. Lies About You?

People have little protection or recourse when the technology creates and spreads falsehoods about them.

Air Canada loses court case after its chatbot hallucinated fake policies to a customer

The airline argued that the chatbot itself was liable. The court disagreed.

Mitigating harms

Value-sensitive design

Post-training and preference alignment

Constitutional AI

But alignment is a very hard and unsolved problem!!!

Large Language Models

Safety and Alignment