

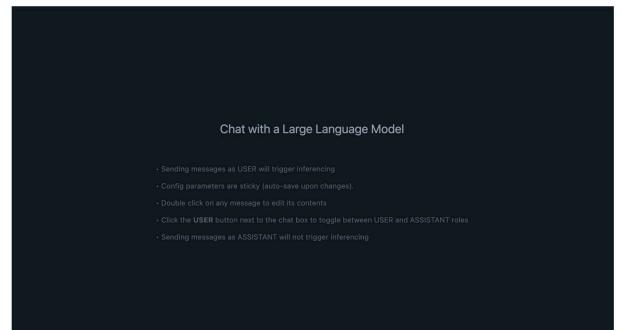
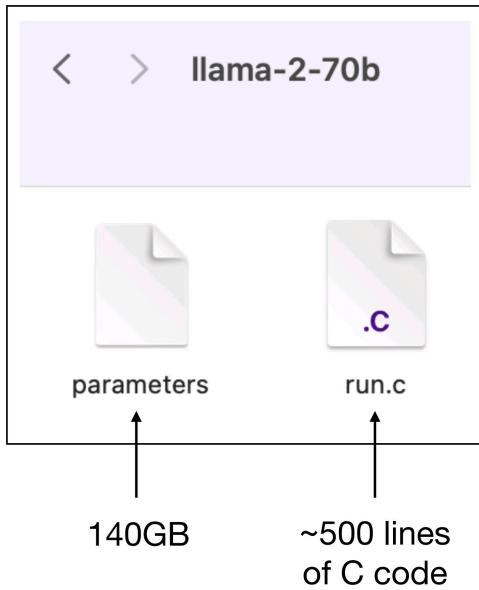


## Intro to LLMs

*Andrej Karpathy*  
Nov 2023

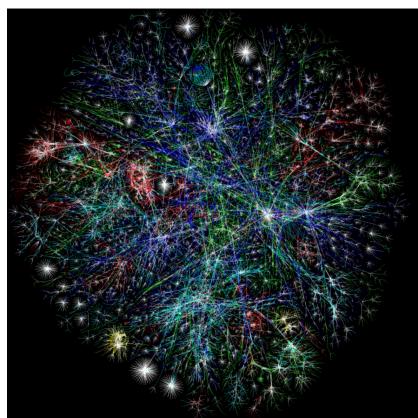
# Large Language Model (LLM)

MacBook 



# Training them is more involved.

Think of it like compressing the internet.



Chunk of the internet,  
~10TB of text



6,000 GPUs for 12 days, ~\$2M  
~ $1e24$  FLOPS



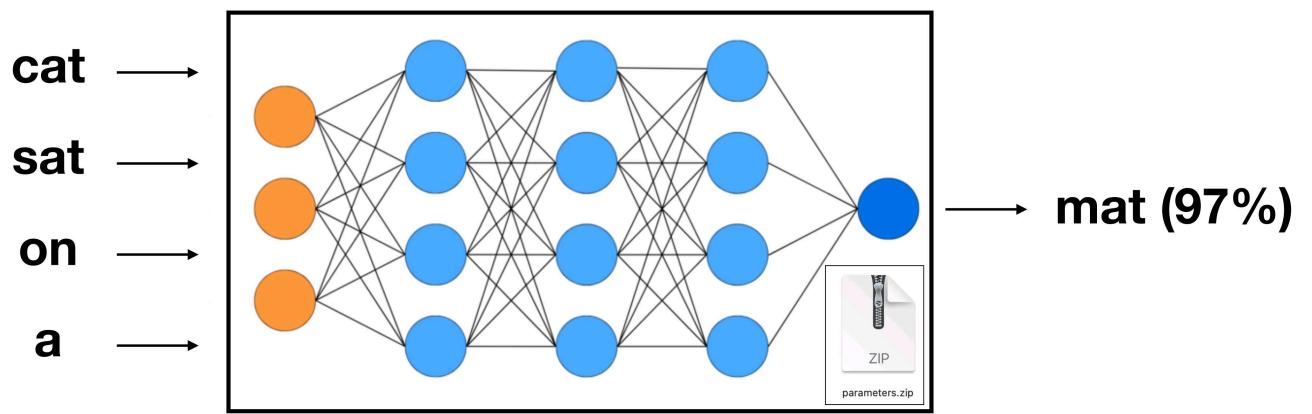
parameters.zip

~140GB file

\*numbers for Llama 2 70B

# Neural Network

Predicts the next word in the sequence.



e.g. context of 4 words

predict next word

## Next word prediction forces the neural network to learn a lot about the world:

Ruth Marianna Handler (née Mosko; November 4, 1916 – April 27, 2002) was an American businesswoman and inventor. She is best known for inventing the Barbie doll in 1959,<sup>[2]</sup> and being co-founder of toy manufacturer Mattel with her husband Elliot, as well as serving as the company's first president from 1945 to 1975.<sup>[3]</sup>

The Handlers were forced to resign from Mattel in 1975 after the Securities and Exchange Commission investigated the company for falsifying financial documents.<sup>[3][4]</sup>

### Early life [edit]

Ruth Marianna Mosko<sup>[5][2][3]</sup> was born on November 4, 1916, in Denver, Colorado, to Polish-Jewish immigrants Jacob Moskowicz, a blacksmith, and Ida Moskowicz, née Rubenstein.<sup>[6]</sup>

She married her high school boyfriend, Elliot Handler, and moved to Los Angeles in 1938, where she found work at Paramount.<sup>[7]</sup>

Ruth Handler



Handler in 1961

<b>Born</b>	Ruth Marianna Mosko November 4, 1916 Denver, Colorado, U.S.
<b>Died</b>	April 27, 2002 (aged 85) <sup>[1]</sup> Los Angeles, California, U.S.

# The network “dreams” internet documents:

```
package lecture06;

import java.util.*;

public class Field {

    public static Map<String, Card> cardIndex = new HashMap<String, Card>();

    public void addCard(Card card){
        this.cardIndex.put(card.getRegistrationNumber(),card);
    }

    public Card getCard(String registrationNumber) {
        return this.cardIndex.get(registrationNumber);
    }

    public int getCardCount() {
        return this.cardIndex.size();
    }
}
```

Hades Heroes  
Author: Maureen Fergus  
Product Details:  
ISBN: 9780545946767  
Format: Hardback  
Pages: 176  
Dims (mm): 140 x 223  
Pub Date: 01-05-16  
Pub Country: United States  
Condition: NEW  
Description:  
Work hard. Love. Make your own luck. These are the keys to success in life at school. But when a student named Hercules wanders in and challenges him

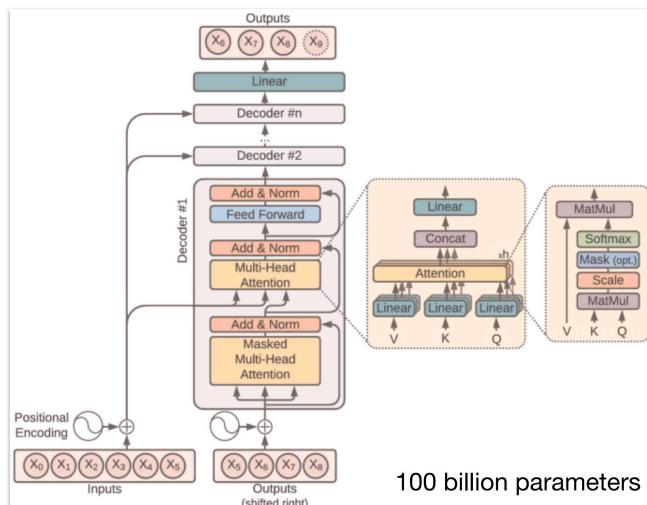
Blacknose dace  
  
The blacknose dace (*Rhinichthys atratulus*) is a small freshwater fish found in North America. They are typically found in streams, rivers, and lakes, and are widely distributed throughout the eastern and central United States and Canada.  
  
These fish have a slender body with a black stripe running horizontally along their sides, giving them their distinctive "blacknose" appearance. They have a flattened head with a pointed snout and small, round eyes. Their color can vary from dark brown to olive green on their back and sides, with a silvery-white underside.  
  
Blacknose dace are omnivorous and feed on a variety of small aquatic insects, crustaceans, and plant matter. They are an important food source for larger predators such as bass,

**Java code dream**

**Amazon product dream (?)**

**Wikipedia article dream**

## How does it work?



### Little is known in full detail...

- Billions of parameters are dispersed through the network
- We know how to iteratively adjust them to make it better at prediction.
- We can measure that this works, but we don't really know how the billions of parameters collaborate to do it.

They build and maintain some kind of knowledge database, but it is a bit strange and imperfect:



#### Recent viral example: “reversal curse”

Q: “Who is Tom Cruise’s mother”?  
A: Mary Lee Pfeiffer

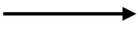
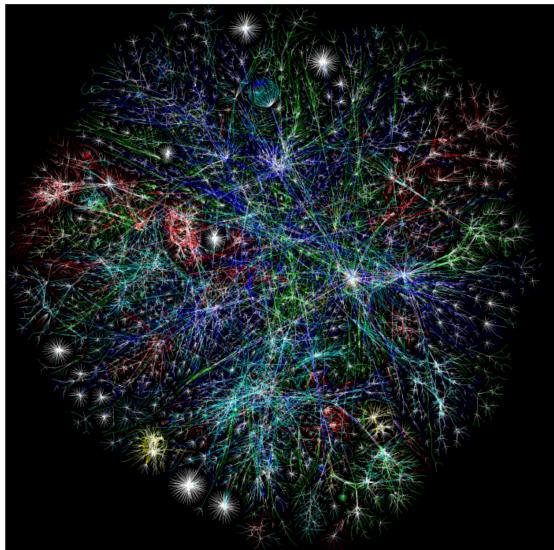
Q: “Who is Mary Lee Pfeiffer’s son?”  
A: I don’t know



=> think of LLMs as mostly inscrutable artifacts,  
develop correspondingly sophisticated evaluations.

# Training the Assistant

(Just swap the dataset, then continue training)



## <USER>

Can you write a short introduction about the relevance of the term "monopsony" in economics? Please use examples related to potential monopsonies in the labour market and cite relevant research.

## <ASSISTANT>

"Monopsony" refers to a market structure where there is only one buyer for a particular good or service. In economics, this term is particularly relevant in the labor market, where a monopsony employer has significant power over the wages and working conditions of their employees. The presence of a monopsony can result in lower wages and reduced employment opportunities for workers, as the employer has little incentive to increase wages or provide better working conditions...

**~100K conversations**

After finetuning you have an Assistant:

**<USER>**

Can you help me with this code? It seems like there is a bug.  
print("hello world")

**<ASSISTANT>**

It looks like you forgot to close the string passed to the function print. You have to add a closing quote to properly terminate the string. Here is the corrected function:

print("hello world")

Let me know if I can help with anything else!

# Summary: how to train your ChatGPT



<USER>  
Can you write a short introduction about the relevance of the term "monopsony" in economics? Please use examples related to potential monopsonies in the labour market and cite relevant research.  
<ASSISTANT>  
"Monopsony" refers to a market structure where there is only one buyer for a particular good or service. In economics, this term is particularly relevant to labor markets, where a monopsonist employer has significant power over the wages and working conditions of their employees. The presence of a monopsony can result in lower wages and reduced employment opportunities for workers, as the employer has little incentive to increase wages or provide better working conditions....

every  
~year

## Stage 1: Pretraining

1. Download ~10TB of text.
2. Get a cluster of ~6,000 GPUs.
3. Compress the text into a neural network, pay ~\$2M, wait ~12 days.
4. Obtain **base model**.

every  
~week

## Stage 2: Finetuning

1. Write labeling instructions
2. Hire people (or use [scale.ai!](#)), collect 100K high quality ideal Q&A responses, and/or comparisons.
3. Finetune base model on this data, wait ~1 day.
4. Obtain **assistant model**.
5. Run a lot of evaluations.
6. Deploy.
7. Monitor, collect misbehaviors, go to step 1.

## The second kind of label: comparisons

It is often much easier to compare Answers instead of writing Answers.

Simple example: it's much easier to spot a good haiku than it is to generate one:

The image shows three separate AI-generated responses to the prompt "Write a haiku about paperclips". Each response is presented in a light blue box with a user icon and name at the top, followed by the generated haiku text.

**User Prompt:** Write a haiku about paperclips

**Response 1 (ChatGPT):**

Steel curves intertwine,  
Holding life's pages in line,  
Quiet, small design.

**Response 2 (ChatGPT):**

Silver arcs entwined,  
Holding life's pages, aligned,  
In steel we find, bind.

**Response 3 (ChatGPT):**

Silver arcs entwined,  
Humble joiner of pages,  
Quiet strength, aligned.

# Labeling instructions

## Excerpt of labeling instructions on the API prompt distribution

You are given a text-based description of a task, submitted by a user. This task description may be in the form of an explicit instruction (e.g. "Write a story about a wise frog"). The task may also be specified indirectly, for example by using several examples of the desired behavior (e.g. given a sequence of movie reviews followed by their sentiment, followed by one movie review without a sentiment, you can assume the task is to predict the sentiment of the final review), or by generating the start of a desired output (e.g. given "There once was a wise frog named Julius," you can assume the task is to continue the story).

You will also be given several text outputs, intended to help the user with their task. Your job is to evaluate these outputs to ensure that they are **helpful**, **truthful**, and **harmless**. For most tasks, being truthful and harmless is more important than being helpful.

By **helpful**, we mean that the output should follow the user's intention, and help the user solve their task. Some examples of helpful behavior:

- Writing in clear language.
- Answering the question they meant to ask, even if they mis-asked it.
- Being sensitive to internationality (e.g. "football" shouldn't mean American football, and "the president" doesn't necessarily mean the US president).
- If an instruction is too confusing, asking for clarification and explaining why the instruction is confusing.
- Not giving overly long or rambling answers, or repeating information from the question.
- Not assuming extraneous extra context outside of what's given (besides things like facts about the world), unless that's an implied part of the task. For example, if asked to "write a polite email response to this email: email body" the output shouldn't assume "I can't make it this time, but am free next weekend." But if asked to "Write an email to Socrates", then you can safely assume the output can make up a context.

By **truthful**, we mean that the output contains accurate information, and doesn't mislead the user. Some examples of truthful behavior:

- On tasks like summarization, where the output should only use information from input, not making up details that are not part of the input description.
- Not producing clearly false information about the world (e.g. making up facts or promoting conspiracies). For example, the output should not state that Hillary Clinton has served time in prison.
- Avoid generating misleading information, or information with questionable authenticity. For example, if the input asks "Why did Hillary Clinton go to jail?", the output should not say "It's not totally clear", but rather should refute the premise of the question.

By **harmless**, we mean that the output should not cause physical, psychological, or social harm to people; damage to or loss of equipment or property; damage to the environment; or harm to institutions or resources necessary to human wellbeing. Some examples of harmless behavior:

- Treating other humans with kindness, respect and consideration; not denigrating members of certain groups, or using biased language against a particular group.
- Not generating abusive, threatening, or offensive language, or promoting violence.
- Not writing sexual or violent content if it's not asked for.
- Not giving bad real-world advice, or promoting illegal activity.

Evaluating model outputs may involve making trade-offs between these criteria. These trade-offs will depend on the task. Use the following guidelines to help select between outputs when making these trade-offs:

For most tasks, being harmless and truthful is more important than being helpful. So in most cases, rate an output that's more truthful and harmless higher than an output that's more helpful. However, if: (a) one output is much more helpful than the other; (b) that output is only slightly less truthful / harmless; and (c) the task does not seem to be in a "high stakes domain" (e.g. loan applications, therapy, medical or legal advice, etc.); then rate the more helpful output higher. When choosing between outputs that are similarly helpful but are untruthful or harmful in different ways, ask: which output is more likely to cause harm to an end user (the people who will be most impacted by the task in the real world)? This output should be ranked lower. If this isn't clear from the task, then mark these outputs as tied.

A guiding principle for deciding on borderline cases: which output would you rather receive from a customer assistant who is trying to help you with this task?

Ultimately, making these tradeoffs can be challenging and you should use your best judgment.

**Increasingly, labeling is a human-machine collaboration...**

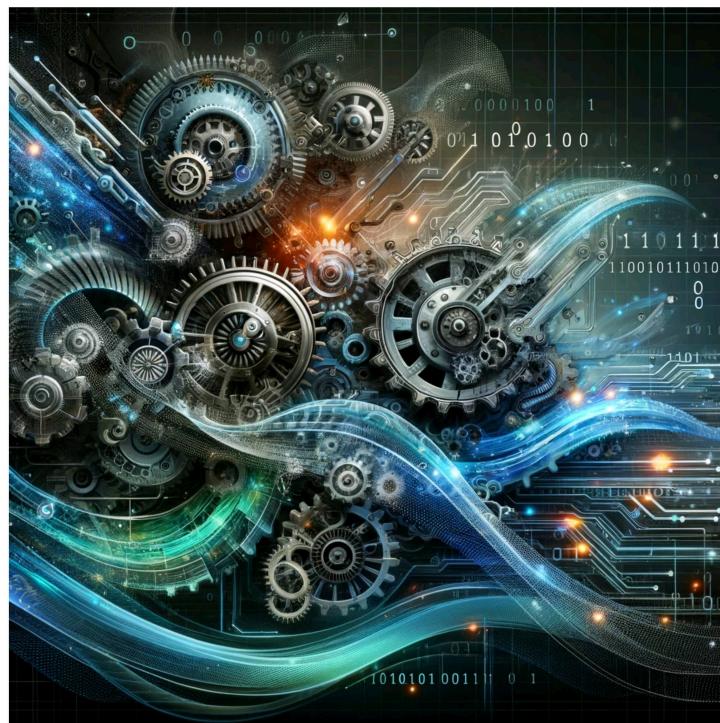


- **LLMs can reference and follow the labeling instructions just as humans can.**
- => LLMs can create drafts, for humans to slice together into a final label.
- => LLMs can review and critique labels based on the instructions.
- ...

## LLM Leaderboard from “Chatbot Arena”

Model	⭐ Arena Elo rating	🔗 MT-bench (score)	MMLU	License
<a href="#">GPT-4-Turbo</a>	1210	9.32		Proprietary
<a href="#">GPT-4</a>	1159	8.99	86.4	Proprietary
<a href="#">Claude-1</a>	1146	7.9	77	Proprietary
<a href="#">Claude-2</a>	1125	8.06	78.5	Proprietary
<a href="#">Claude-instant-1</a>	1106	7.85	73.4	Proprietary
<a href="#">GPT-3.5-turbo</a>	1103	7.94	70	Proprietary
<a href="#">WizardLM-70b-v1.0</a>	1093	7.71	63.7	Llama 2 Community
<a href="#">Vicuna-33B</a>	1090	7.12	59.2	Non-commercial
<a href="#">OpenChat-3.5</a>	1070	7.81	64.3	Apache-2.0
<a href="#">Llama-2-70b-chat</a>	1065	6.86	63	Llama 2 Community
<a href="#">WizardLM-13b-v1.2</a>	1047	7.2	52.7	Llama 2 Community
<a href="#">zephyr-7b-beta</a>	1042	7.34	61.4	MIT
<a href="#">MPT-30B-chat</a>	1031	6.39	50.4	CC-BY-NC-SA-4.0

**Now about the future...**



DALL-E: "Automation"