



# LLM Workshop session 6: High-Level View of LLMs

Mahdi Samiei

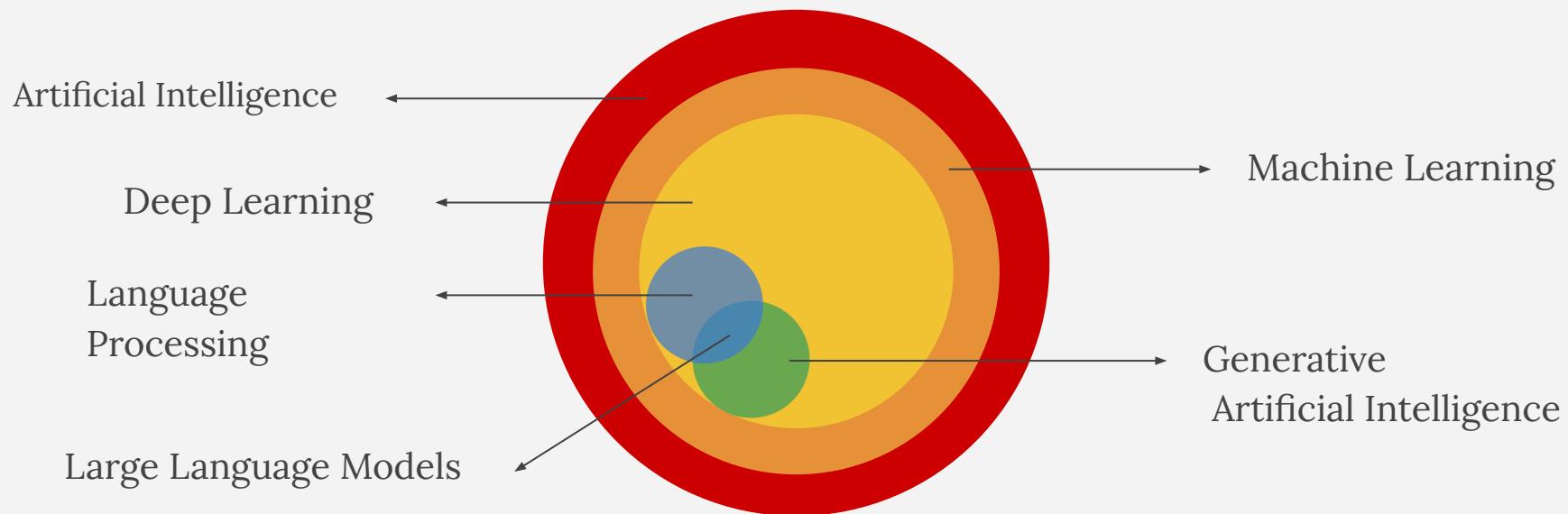
| 26 Feb 2024

# What we will discuss ...

- ▶ AI, ML, DL, GenAI, LLM, ...
- ▶ Is it easy to make LLMs?
- ▶ How to use LLMs better?
- ▶ Future of LLMs



# AI, ML, DL, Gen AI, NLP, LLM?



# What is AI?

**AI Effect:** It's part of the history of the field of artificial intelligence that every time somebody figured out how to make a computer do something—play good checkers, solve simple but relatively informal problems—there was a chorus of critics to say, 'that's not thinking'

“AI is whatever hasn't been done yet”



# What is Machine Learning?

Learning From Experience

Elements of Learning:

Data

Architecture and Parameters

Objective

Training

# What is Deep Learning?

Neural Networks: A special type of architecture in machine learning models

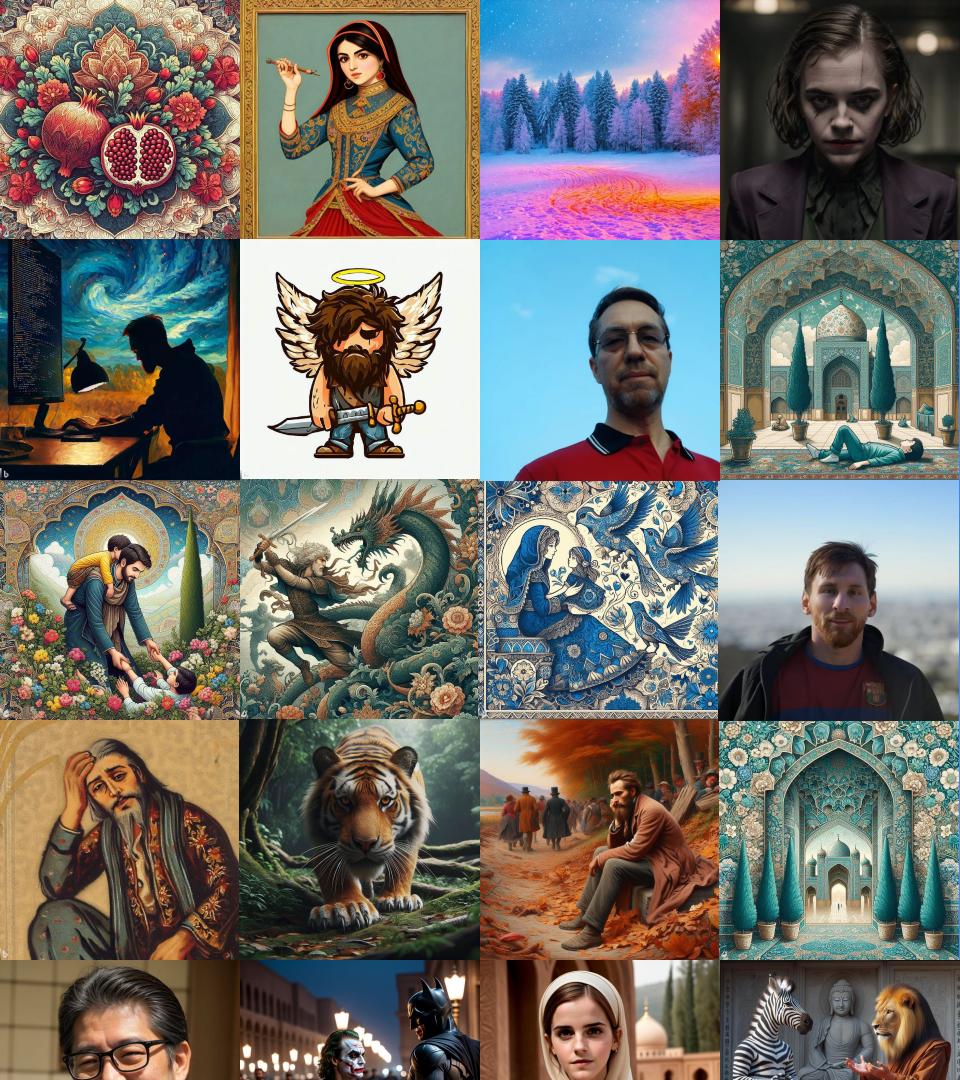
Not New Idea, jump after 2012 due to the advancement of hardware and data

Their own issues: Deep Learning

Different Modals: Image, Text, Speech, Graph, ...

# What is Gen AI?

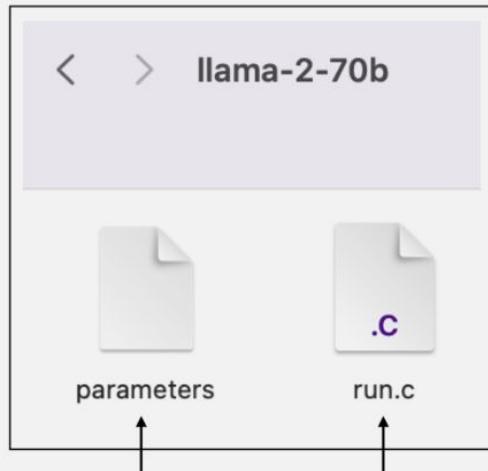
creating new contents



# LLM?

- ▼ What is LLM?
- ▼ How LLM is Created?
- ▼ What LLM Does?

# What is LLM?



~500 lines  
of C code

140GB

ChatGPT 3.5 ▾

SA You  
Hi! I am mahdi!

ChatGPT  
Hi Mahdi! Nice to meet you! How's your day going?  
🔗 ↻ 🔍

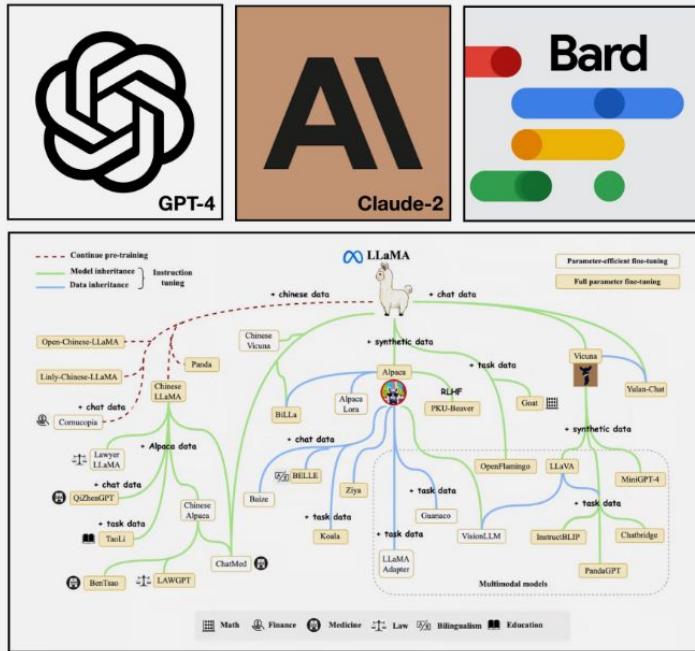
Message ChatGPT...



ChatGPT can make mistakes. Consider checking important information.



# Close-Source LLM vs Open-Source LLM



How is a  
Large Language Model  
built?



# A Game!

Baba alitoa maji.

Baba

Simba ni mwindaji wa kulungu.

Simba

Nguruwe, kulungu na simba ni wanyama.

Wanyama

Simba ni mfalme wa msituni.

Maji

Wanyama kama nguruwe wanaishi msituni.

Mwindaji

Simba msituni hakuweza kuniwinda, lakini kulungu aliniwinda.

Msituni

Ole kwa mfungwa aliyesahaulika, mwindaji amekwenda lakini bado yuko kwenye mtego.

Mfungwa

Kulungu anaishi .....

# A Game!

Baba alitoa maji.

Simba ni mwindaji wa kulungu.

Nguruwe, kulungu na simba ni wanyama.

Simba ni mfalme wa msituni.

Wanyama kama nguruwe wanaishi msituni.

Simba msituni hakuweza kuniwinda, lakini kulungu aliniwinda.

Ole kwa mfungwa aliyesahaulika, mwindaji amekwenda lakini bado yuko kwenye mtego.

Kulungu anaishi .....

Wanyama

Mwindaji

Msituni

Mfungwa

# Next Token Prediction Game!

Baba alitoa maji.

Simba ni mwindaji wa kulungu.

Nguruwe, kulungu na simba ni wanyama.

Simba ni mfalme wa msituni.

Wanyama kama nguruwe wanaishi msituni.

Simba msituni hakuweza kuniwinda, lakini kulungu aliniwinda.

Ole kwa mfungwa aliyesahaulika, mwindaji amekwenda lakini bado yuko kwenye mtego.

Kulungu anaishi .....

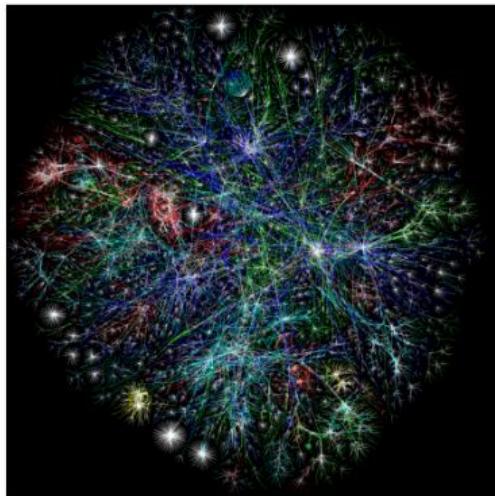
Wanyama

Mwindaji

Msituni

Mfungwa

# Phase 1: Train LLM on Lots of text by NTP



ZIP

parameters.zip

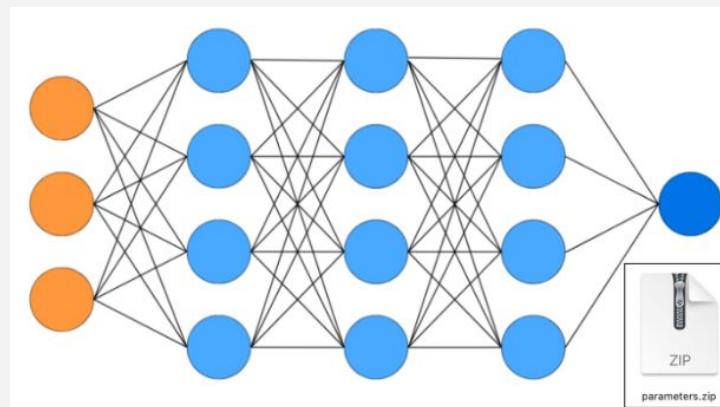
~140GB file

Chunk of the internet,  
~10TB of text

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

# What you get:

<شروع>



محمد رضا، ۱۳ درصد

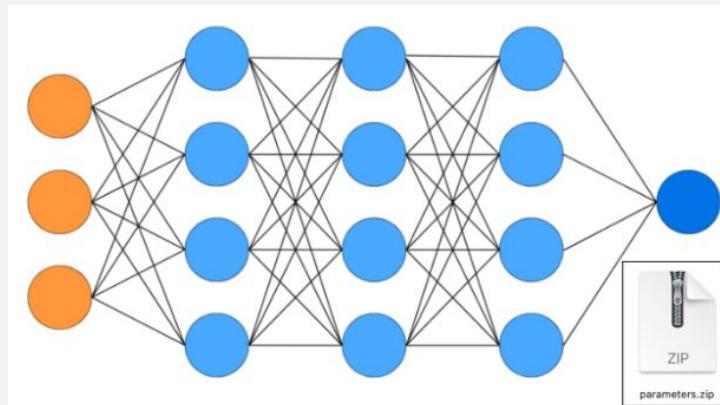
حسین، ۴ درصد

...

بود، ۰ درصد

# What you get:

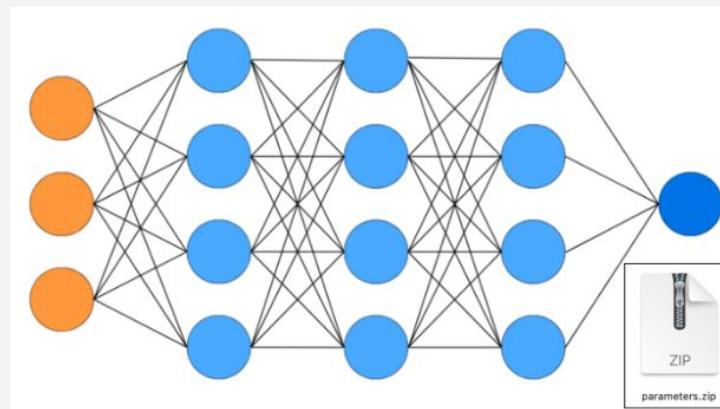
مشروع  
محمد رضا



- پسر، ۹۷ درصد
- فریدونی، ۲ درصد
- ...
- دختر، ۰ درصد

# What you get:

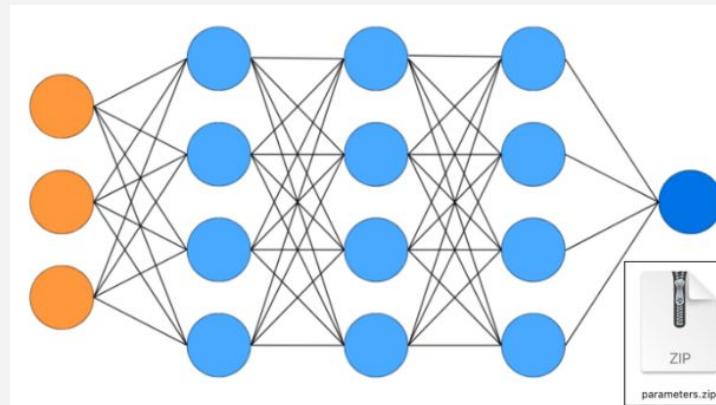
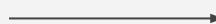
مشروع <  
محمد رضا پسر



- خوبی، ۹۰ درصد
- باهوشی، ۵ درصد
- ...
- اوگاندا، ۰ درصد

# What you get:

مشروع <  
محمد رضا پسر  
خوبی>

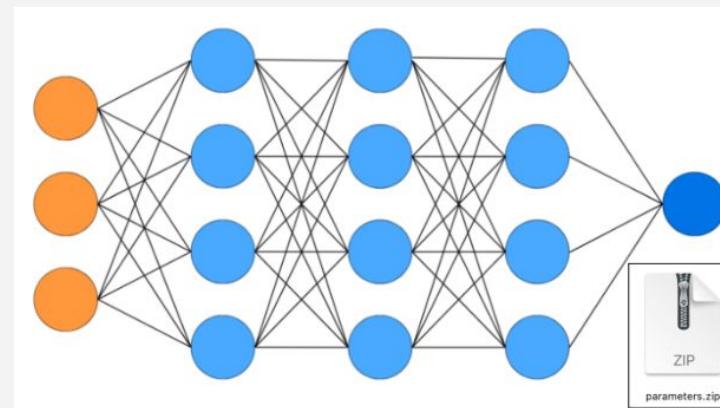


- نیست، ۷۲ درصد
- هست، ۱۳ درصد
- ...
- کلنگ، ۰ درصد

...

# What you get:

<شروع>  
محمد رضا پسر  
خوبی نیست



۱۴۳ درصد <پایان>,  
اما، ۵ درصد زرافه، ۰ درصد ...

So By Training NSP  
on Vast Amount of  
Text, Knowledge  
Emerges!

So By Training NSP on Vasts  
Amount of Text, Knowledge  
Emerges!

But it's not Enough!

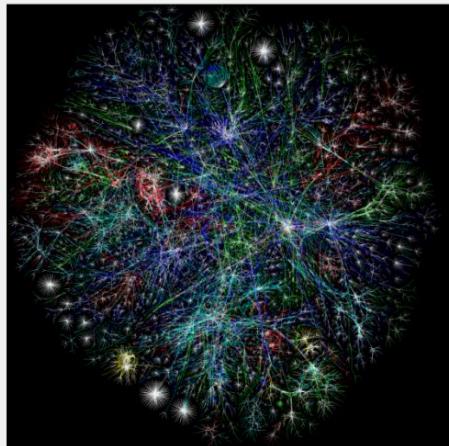
# We Want Assistant: Collecting Instruction Based Dataset

## Training the Assistant

(Just swap the dataset, then continue training)



written by  
people:



### <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

# We Want Assistant: Finetune Parameters on Instruction Dataset

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!

# What is missing?

- ▼ Misalignment: When the training objective does not capture the desiderata we want from models
- ? Helpful: should help the user solve their task according to the instructions.
- ? Honest: should express uncertainty when the model doesn't know the answer, instead of hallucinating a wrong answer.
- ? Harmless: should not cause physical, psychological, or social harm to people or the environment.

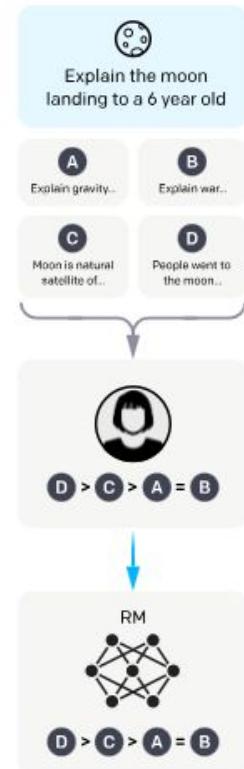
# RLHF: Change the Loss

- ▶ Use reward signal and Reinforcement Learning to fine tune the model.
- ▶ Need scored data: (instruction, answer, score)

Step 2

**Collect comparison data, and train a reward model.**

A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

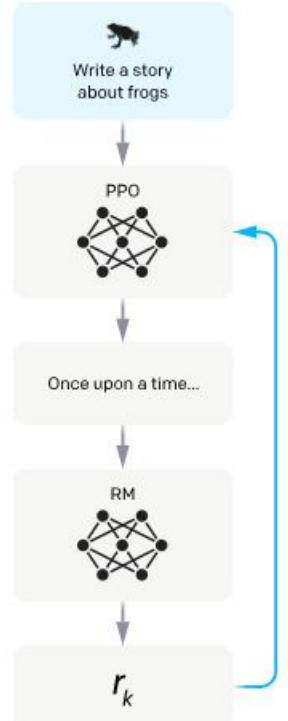
# RLHF: Change the Loss

- ▶ Use reward signal and Reinforcement Learning to fine tune the model.
- ▶ Need scored data: (instruction, answer, score)
- ▶ Fine-Tune LLM using Proximal Policy Optimization on Reward Function

Step 3

**Optimize a policy against the reward model using reinforcement learning.**

A new prompt is sampled from the dataset.



The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

What LLM Does?

How to behave with your lIrm?

Where LLM goes?

# LLM Leaderboards

Arena Elo | Full Leaderboard

Three benchmarks are displayed: **Arena Elo**, **MT-Bench** and **MMLU**.

- [Chatbot Arena](#) - a crowdsourced, randomized battle platform based on human preference votes.
- [MT-Bench](#): a set of challenging multi-turn questions. We use GPT-4 to grade the model responses.
- [MMLU](#) (5-shot): a test to measure a model's multitask accuracy on 57 tasks.

Code: The MT-bench scores (single-answer grading on a scale of 10) are computed by [fastchat.llm.judge](#). The MMLU scores are mostly computed by [InstructEval](#). Higher values are better for all benchmarks. Empty cells mean not available.

Model	Arena Elo	MT-bench	MMLU	Organization	License
<a href="#">GPT-4-1106-preview</a>	1254	9.32		OpenAI	Proprietary
<a href="#">GPT-4-0125-preview</a>	1253			OpenAI	Proprietary
<a href="#">Bard (Gemini Pro)</a>	1218			Google	Proprietary
<a href="#">GPT-4-0314</a>	1191	8.96	86.4	OpenAI	Proprietary
<a href="#">GPT-4-0613</a>	1164	9.18		OpenAI	Proprietary
<a href="#">Mistral Medium</a>	1152	8.61	75.3	Mistral	Proprietary
<a href="#">Claude-1</a>	1150	7.9	77	Anthropic	Proprietary
<a href="#">Owen1.5-72B-Chat</a>	1147	8.61	77.5	Alibaba	Qianwen LICENSE
<a href="#">Claude-2.0</a>	1132	8.06	78.5	Anthropic	Proprietary
<a href="#">Gemini Pro (Dev API)</a>	1122		71.8	Google	Proprietary
<a href="#">Claude-2.1</a>	1120	8.18		Anthropic	Proprietary
<a href="#">Mixtral-8x7b-Instruct-v0.1</a>	1120	8.3	70.6	Mistral	Apache 2.0
<a href="#">GPT-3.5-Turbo-0613</a>	1118	8.39		OpenAI	Proprietary
<a href="#">Gemini Pro</a>	1115		71.8	Google	Proprietary
<a href="#">Yi-34B-Chat</a>	1111		73.5	01 AI	Yi License
<a href="#">Claude-Instant-1</a>	1109	7.85	73.4	Anthropic	Proprietary
<a href="#">GPT-3.5-Turbo-0314</a>	1106	7.94	70	OpenAI	Proprietary
<a href="#">Midjourney V20B-14.0</a>	1106	7.74	62.7	MidJourney	CC-BY-NC-ND

# LLMs are not GOD yet

## Hallucination!

### 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 



SA

You

What is the name of wife of mostafa meshkati?



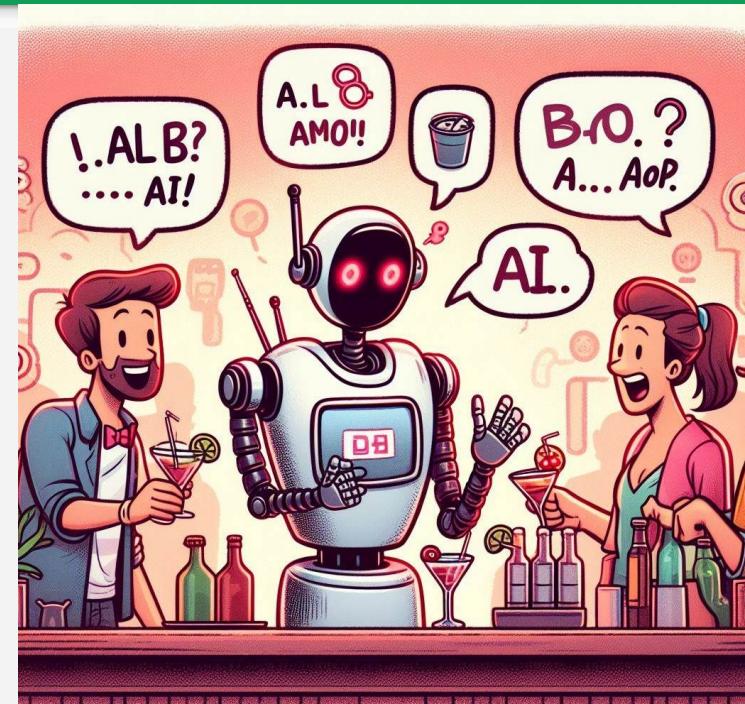
ChatGPT

Mostafa Meshkati, the Iranian-American engineer known for his work in human factors engineering and safety, is married to Dr. Sepideh Modrek.



# You must know how to talk to LLM: Prompt Engineering

**Prompt Engineering**, also known as **In-Context Prompting**, refers to methods for how to communicate with LLM to steer its behavior for desired outcomes without updating the model weights. It is an empirical science and the effect of prompt engineering methods can vary a lot among models, thus requiring heavy experimentation and heuristics.



# Prompt Engineering: Few-Shot Demonstration

## Zero-Shot Prompting

**Text:** i'll bet the video game is a lot more fun than the film.  
**Sentiment:**

**Text:** (lawrence bounces) all over the stage, dancing, running, sweating, mopping his face and g  
**Sentiment:** positive

**Text:** despite all evidence to the contrary, this clunker has somehow managed to pose as an actu  
**Sentiment:** negative

**Text:** for the first time in years, de niro digs deep emotionally, perhaps because he's been sti  
**Sentiment:** positive

**Text:** i'll bet the video game is a lot more fun than the film.  
**Sentiment:**

## few-Shot Prompting

# Prompt Engineering: Instruction Prompting

The purpose of presenting few-shot examples in the prompt is to explain our intent to the model; in other words, describe the task instruction to the model in the form of demonstrations. However, few-shot can be expensive in terms of token usage and restricts the input length due to limited context length. So, why not just give the instruction directly?

```
Definition: Determine the speaker of the dialogue, "agent" or "customer".
```

```
Input: I have successfully booked your tickets.
```

```
Output: agent
```

```
Definition: Determine which category the question asks for, "Quantity" or "Location".
```

```
Input: What's the oldest building in US?
```

```
Output: Location
```

```
Definition: Classify the sentiment of the given movie review, "positive" or "negative".
```

```
Input: i'll bet the video game is a lot more fun than the film.
```

```
Output:
```

# Prompt Engineering: CoT Prompting

Question: Tom and Elizabeth have a competition to climb a hill. Elizabeth takes 30 minutes.

copy

Answer: It takes Tom  $30 * 4 = <<30*4=120>>120$  minutes to climb the hill.

It takes Tom  $120 / 60 = <<120/60=2>>2$  hours to climb the hill.

So the answer is 2.

==

Question: Jack is a soccer player. He needs to buy two pairs of socks and a pair of soccer shoes.

Answer: The total cost of two pairs of socks is  $\$9.50 \times 2 = \$<<9.5*2=19>>19$ .

The total cost of the socks and the shoes is  $\$19 + \$92 = \$<<19+92=111>>111$ .

Jack need  $\$111 - \$40 = \$<<111-40=71>>71$  more.

So the answer is 71.

==

Question: Marty has 100 centimeters of ribbon that he must cut into 4 equal parts. Each of the

Answer:

# LLM are not enough? The world is changing

SA

You

When was the last Israeli-Palestinian war?



ChatGPT

The most recent major conflict between Israel and Palestinian militant groups was the 11-day war in May 2021. This conflict primarily involved Israel and Hamas, the militant group controlling the Gaza Strip. It resulted in significant casualties and destruction before a ceasefire was brokered by Egypt and other international mediators.



# LLM are not enough? LLM doesn't know everything

SA

You

Who is the founder of the Divar application?



ChatGPT

Divar, an online classifieds platform in Iran, was founded by Shahab Javanmardi.

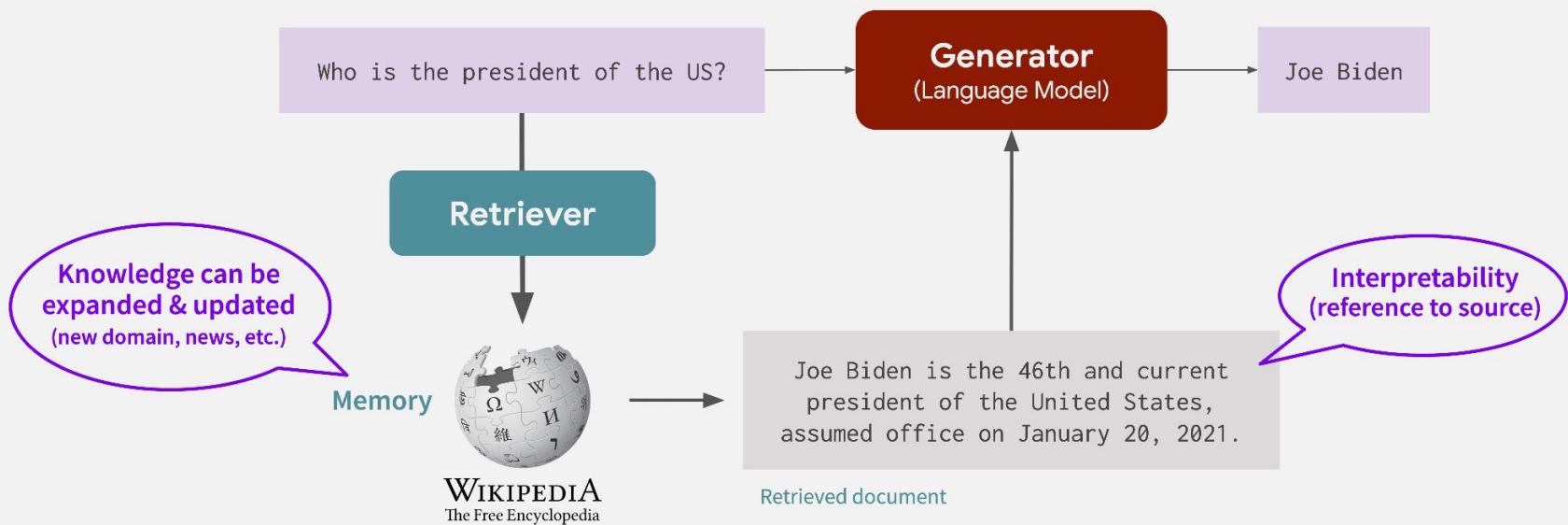


# What does human do?



# LLMs: Retrieval Augmented Generation

## Retrieval augmentation



# LLMs and External API

Question: In Fibonacci sequence, it follows the rule that each number is equal to the sum of the preceding two numbers. Assuming the first two numbers are 0 and 1, what is the 50th number in Fibonacci sequence?

The first number is 0, the second number is 1, therefore, the third number is  $0+1=1$ . The fourth number is  $1+1=2$ . The fifth number is  $1+2=3$ . The sixth number is  $2+3=5$ . The seventh number is  $3+5=8$ . The eighth number is  $5+8=13$ .

..... (Skip 1000 tokens)

The 50th number is 32,432,268,459.

CoT



32,432,268,459



```
length_of_fibonacci_sequence = 50  
fibonacci_sequence = np.zeros(length_of_)  
fibonacci_sequence[0] = 0  
fibonacci_sequence[1] = 1  
for i in range(3, length_of_fibonacci_sequence):  
    fibonacci_sequence[i] = fibonacci_sequence[i-1] +  
        fibonacci_sequence[i-2]  
ans = fibonacci_sequence[-1]
```

PoT



python



12,586,269,025

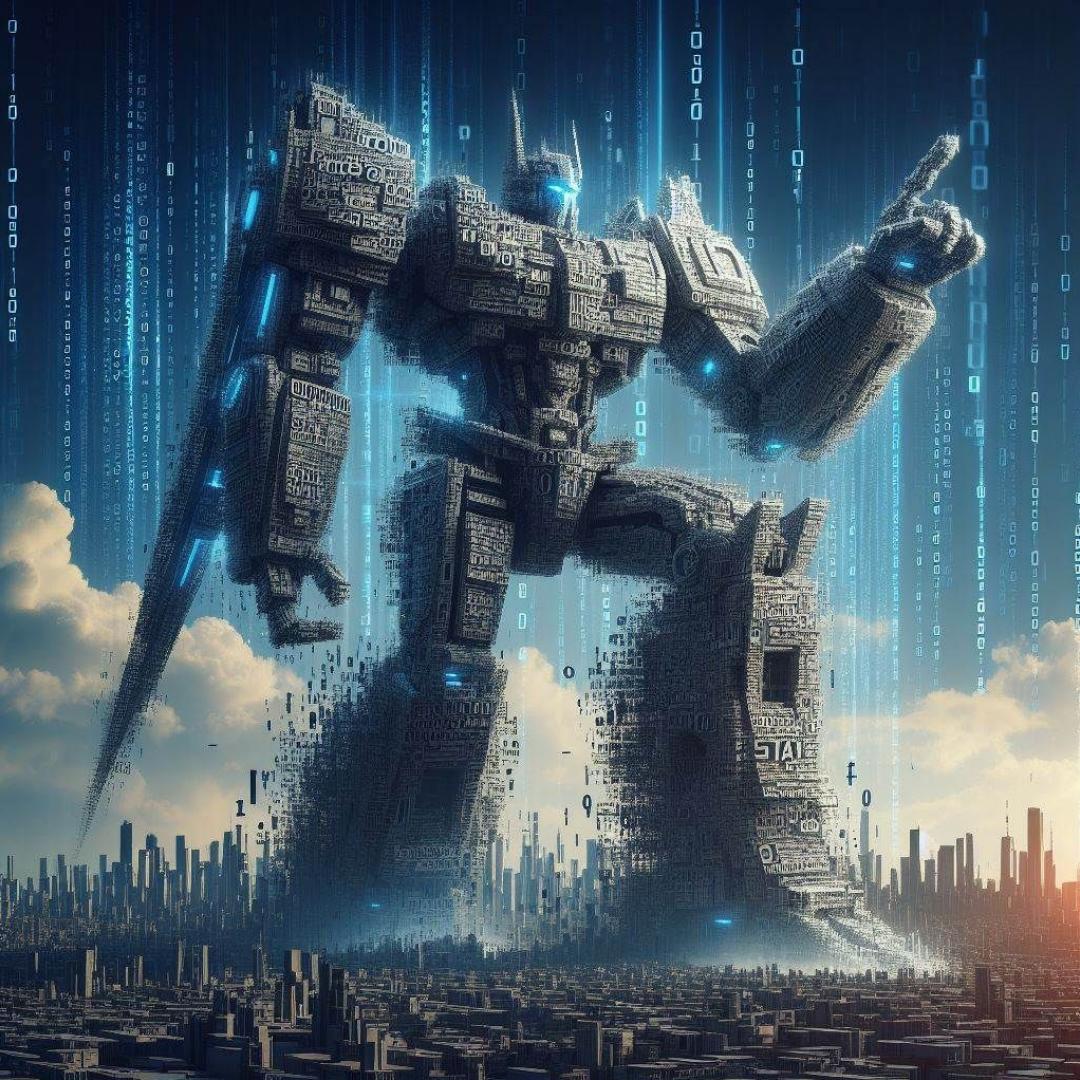


# LLM is a Modern KNN

It all depends on what data points  
you choose to input into LLM.

	SST-2	AGNews	R8	R52	MR	Average
<b>Supervised Methods</b>						
RoBERTa-Large (Liu et al., 2019)	95.99	95.55	97.76	96.42	91.16	95.38
DeBERTa (He et al., 2020)	94.75	95.32	98.33	96.32	90.19	94.99
RoBERTa-GCN (Lin et al., 2021)	95.80	<b>95.68*</b>	98.2	96.1	89.7	95.10
XLNet (Yang et al., 2019)	<b>96.10*</b>	95.55	-	-	-	-
VLAWE (Ionescu and Butnaru, 2019)	-	-	-	-	<b>93.3*</b>	-
GCN-SB (Zeng et al., 2022)	-	-	<b>98.53*</b>	<b>96.35*</b>	87.59	-
<b>Zero-shot Setting</b>						
Vanilla (Brown et al., 2020)	91.55	90.72	90.19	89.06	88.69	90.04
CoT (Kojima et al., 2022)	92.11	91.25	90.48	91.24	89.37	90.89
<b>CARP</b>	93.01	92.60	91.75	91.80	89.94	91.82
<b>Few-shot Setting (<math>k=16</math>)</b>						
<i>Random Sampler</i>						
Vanilla (Brown et al., 2020)	92.36	91.74	91.58	91.56	89.15	91.28
CoT (Kojima et al., 2022)	94.56	95.02	92.49	92.03	89.91	92.80
<b>CARP</b>	96.20	95.18	97.60	96.19	90.03	95.04
<i>SimCSE kNN-Sampler</i>						
Vanilla (Brown et al., 2020)	93.90	93.50	94.36	92.40	89.59	94.05
CoT (Kojima et al., 2022)	94.21	94.28	95.07	92.98	90.27	93.69
<b>CARP</b>	95.69	95.25	97.83	96.27	90.74	95.16
<i>FT kNN-Sampler</i>						
Vanilla (Brown et al., 2020)	94.01	94.14	95.57	95.79	90.90	94.08
CoT (Kojima et al., 2022)	95.48	94.89	95.59	95.89	90.17	94.40
<b>CARP</b>	96.80	95.99	98.29	96.82	91.90	95.97
<b>CARP</b> (WP Vote)	97.39	96.40	98.78	96.95	92.39	96.38

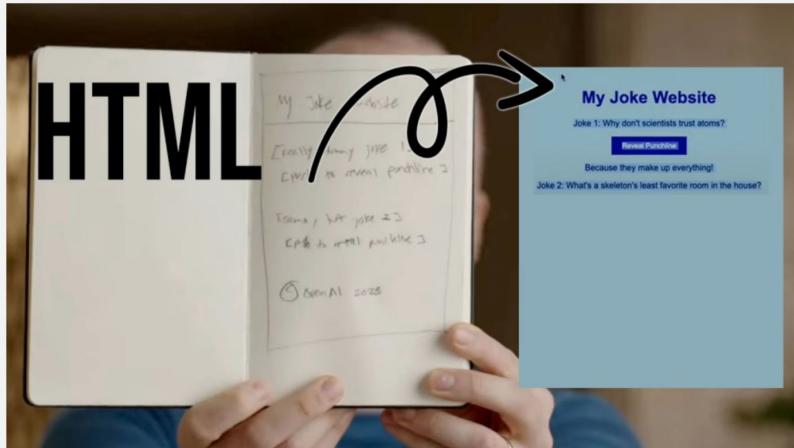
# Where LLM goes?



# Multi-Modal LLMs

## Vision

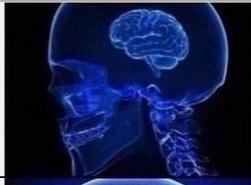
Can both see, and generate images



Example: Take a sketch of an idea and generate a working website.

# LLM as Backend

**WRITING  
A BACKEND**



**HIRING  
A BACKEND  
ENGINEER**



**ASKING  
CHAT-GPT  
FOR A BACKEND**

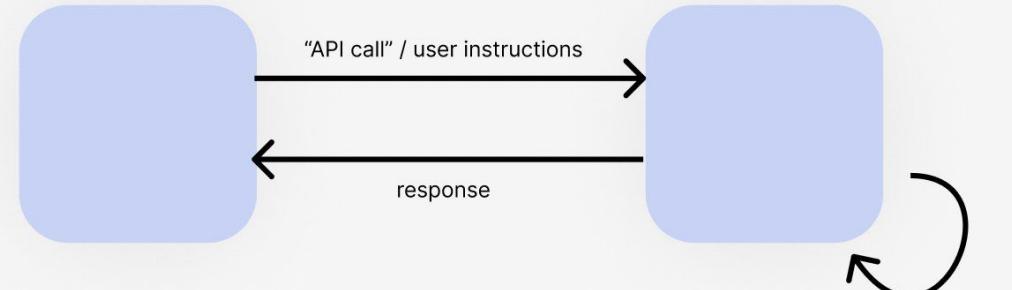


**THE LLM IS  
THE BACKEND**



**Front End**

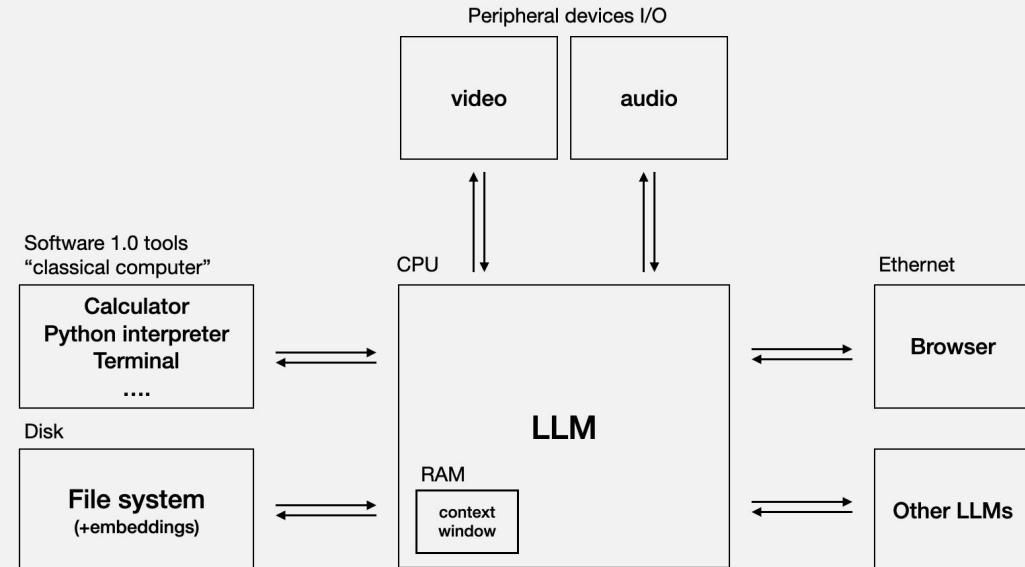
**LLM**



prompt, old state, user instructions

<https://github.com/RootbeerComputer/backend-GPT>

# LLM as OS



- It can read and generate text
- It has more knowledge than any single human about all subjects
- It can browse the internet
- It can use the existing software infrastructure (calculator, Python, mouse/keyboard)
- It can see and generate images and video
- It can hear and speak, and generate music
- It can think for a long time using a System 2
- It can “self-improve” in domains that offer a reward function
- It can be customized and finetuned for specific tasks, many versions exist in app stores
- It can communicate with other LLMs



**Andrej Karpathy**  
@karpathy

The hottest new programming  
language is English

# Thank You!