

# **Large Language Models (LLMs)**

*Mai Nguyen, Paul Rodriguez and Robert Sinkovits*

*San Diego Supercomputer Center*

*2024 CIML Summer Institute*

# Table of contents

- Introduction to LLMs
  - What is an LLM, strengths & limitations, survey of LLMs, pre-training
- Prompt engineering
- Retrieval Augmented Generation (RAG)
- LLM-related tools
- Other LLM topics
  - Toxicity, hallucinations, biases, privacy, copyright
  - Other generative AI models (image, video), use cases

## What are LLMs

“Any sufficiently advanced technology is indistinguishable from magic.”

- Arthur C. Clark

We assume that you're already familiar with LLMs and what you can do with them - write code, compose essays, summarize text, explain concepts

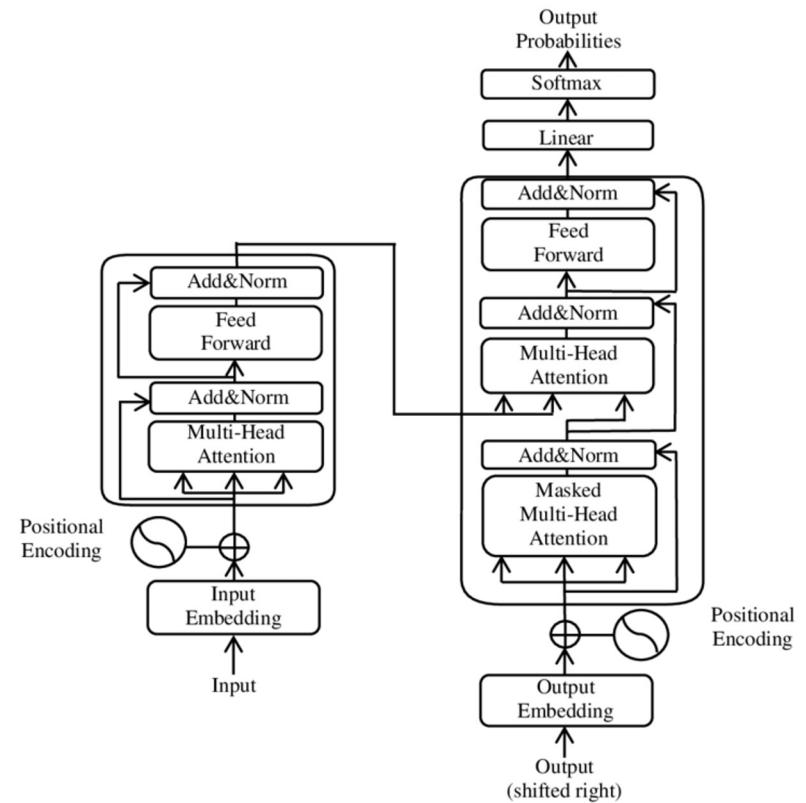
Our goal today is to demystify LLMs, show you how to use them most effectively, demonstrate extending the capabilities of LLMs beyond the data used to train them, and explore their limitations and pitfalls.

# What are LLMs

An LLM is a type of generative AI model that learns the statistical relationships between words or tokens after being trained on large amounts of text.

LLMs are based on the transformer architecture, which was introduced in the highly cited (125k times as of 2024) manuscript *Attention is all you need* in 2017.

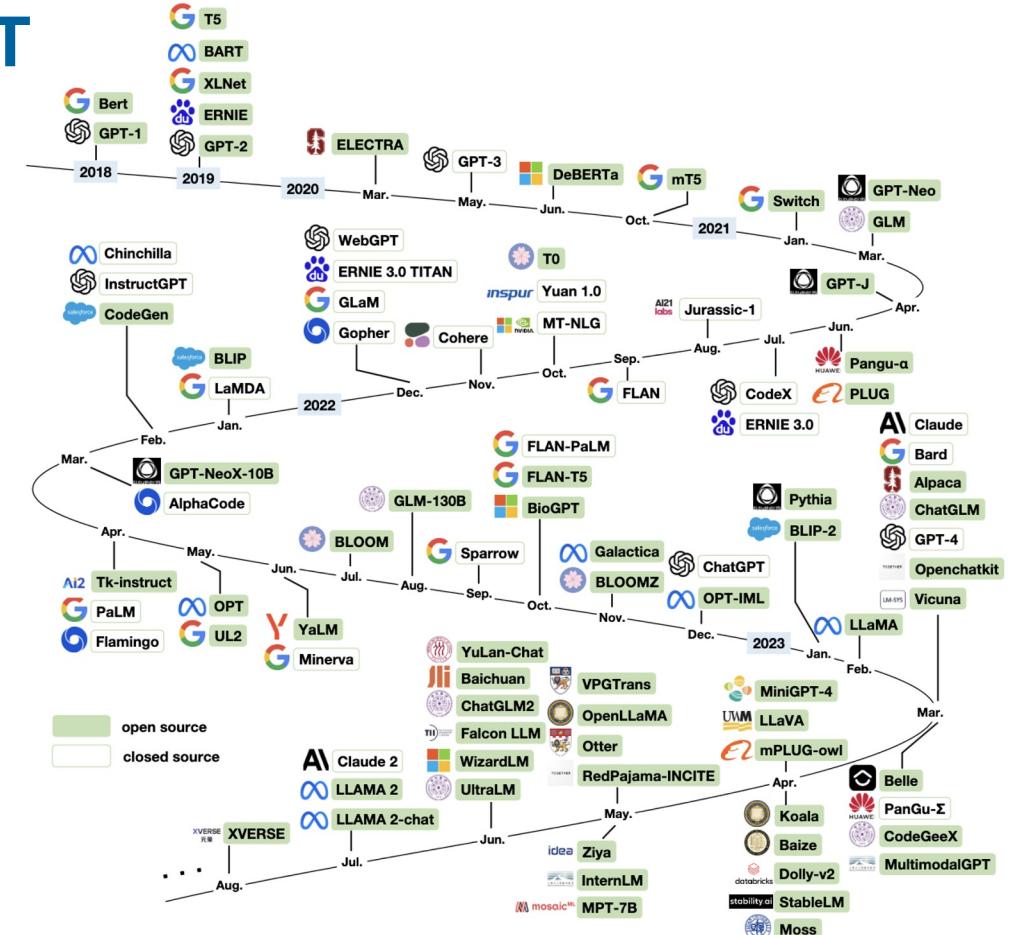
Attention assigns importance to words based on their context in a larger block of text.



Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems 30* (2017)

# There's a lot more than GPT

Many users equate LLM with GPT, but the ecosystem is much richer. Some of these are deployed as chatbots, while others are intended as research tools (e.g., to explore safe AI) or targeted at specific domains.



Gao et al. <https://arxiv.org/pdf/2308.14149>

## Open source vs. proprietary models

Many of the most powerful LLMs including GPT-4, LaMDA, Gemini and PaLM are proprietary and only available through the web or APIs. Others, such as BLOOM, LLaMa and Mistral 7B are open source and available under Apache, MIT or other licenses.

Note that some companies make the older models freely available, while restricting access to source code and hyper-parameter weights for newer or more capable models.

- GPT-1 and GPT-2 – MIT license
- GPT-3 and GPT-4 – Proprietary
  
- Mistral 75, Mixtral 8x7B, Mixtral 8x22B – Apache 2.0
- Mistral Small, Medium and Large - Proprietary

# How big are LLMs

- Early models released in 2018 had well under one billion parameters. These include GPT-1 (117 million) and BERT (340 million).
- By the early 2020s, model sizes had grown to hundreds of billions, with Google's GLaM reported to have 1.2 trillion parameters.
- The latest models have not yet released their model sizes, but GPT-4 is estimated to have around 1.76 trillion parameters.



Hacker News

<https://news.ycombinator.com/item?id=4100000> ::

## Ask HN: GPT-4 has 1.7T parameters. What's a parameter?

GPT-4 is rumored to be a "mixture of experts", i.e. a neural net consisting of multiple specialized modules, only one of which is run on any particular prompt.

The screenshot shows a news article from 'AI in practice' dated Mar 25, 2023. The title is 'GPT-4 has more than a trillion parameters - Report'. Below the title is a colorful bubble chart. The author is Matthias Bastian, with links to his profile and email. The article discusses GPT-4's size and architecture, mentioning it is based on eight models with 220 billion parameters each, totaling about 1.76 trillion parameters. It also notes the use of a Mixture of Experts (MoE). A 'Content' tab is visible at the bottom left, and an 'Update' section is at the bottom right.

Matthias Bastian  
Profile E-Mail

Online journalist Matthias is the co-founder and publisher of THE DECODER. He believes that artificial intelligence will fundamentally change the relationship between humans and computers.

Content Summary

Update

- Further details on GPT-4's size and architecture have been leaked.
- The system is said to be based on eight models with 220 billion parameters each, for a total of about 1.76 trillion parameters, connected by a Mixture of Experts (MoE).

# Data used to train models



The Data ▾ Resources ▾ Community ▾ About ▾ Search ▾ Contact Us

Common Crawl  
maintains a [free, open](#)  
[repository](#) of web crawl  
data that can be used by  
anyone.

Common Crawl is a 501(c)(3) non-profit founded in 2007.

We make wholesale extraction, transformation and analysis of  
open web data accessible to researchers.

[Overview](#)



The largest generic LLMs are trained using data scraped from billions of web pages.

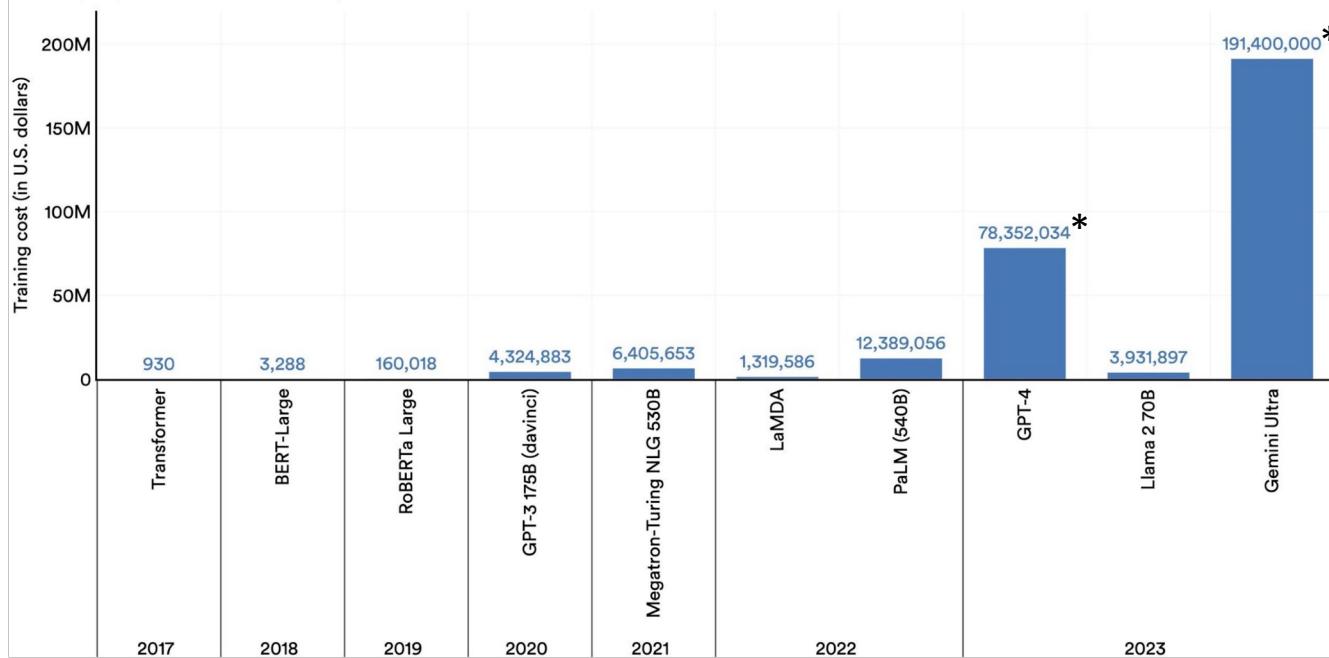
Domain-specific LLMs for medicine, law, engineering and other fields are trained using data relevant to that field. This can reduce the likelihood of hallucinations or incorrect results.

Data used to train model should be cleansed to remove toxic content, low-quality data and duplicates.

# Cost of training LLMs

Estimated training cost of select AI models, 2017–23

Source: Epoch, 2023 | Chart: 2024 AI Index report



As the number of hyper-parameters and size of the corpus grows, the cost of training the LLMs has exploded. It's now impractical for anyone other than the largest players to train the most advanced models.

\* estimates

By Stanford Institute for Human-Centered Artificial Intelligence (permission obtained by email from the AI index research manager) -  
<https://aiindex.stanford.edu/report/#individual-chapters> (chapter 1, image 3), CC BY-SA 4.0,  
<https://commons.wikimedia.org/w/index.php?curid=149139090>

# Cost of training LLMs

## PaLM: Scaling Language Modeling with Pathways

Aakanksha Chowdhery\* Sharan Narang\* Jacob Devlin\*

Maarten Bosma Gaurav Mishra Adam Roberts Paul Barham

Hyung Won Chung Charles Sutton Sebastian Gehrmann Parker Schuh Kensen Shi  
Sasha Tsvyashchenko Joshua Maynez Abhishek Rao<sup>†</sup> Parker Barnes Yi Tay  
Noam Shazeer<sup>‡</sup> Vinodkumar Prabhakaran Emily Reif Nan Du Ben Hutchinson  
Reiner Pope James Bradbury Jacob Austin Michael Isard Guy Gur-Ari  
Pengcheng Yin Toju Duke Anselm Levskaya Sanjay Ghemawat Sunipa Dev  
Henryk Michalewski Xavier Garcia Vedant Misra Kevin Robinson Liam Fedus  
Denny Zhou Daphne Ippolito David Luan<sup>†</sup> Hyeontaek Lim Barret Zoph  
Alexander Spiridonov Ryan Sepassi David Dohan Shivani Agrawal Mark Omernick  
Andrew M. Dai Thanumalayan Sankaranarayana Pillai Marie Pellat Aitor Lewkowycz  
Erica Moreira Rewon Child Oleksandr Polozov<sup>†</sup> Katherine Lee Zongwei Zhou  
Xuezhi Wang Brennan Saeta Mark Diaz Orhan Firat Michele Catasta<sup>†</sup> Jason Wei  
Kathy Meier-Hellstern Douglas Eck Jeff Dean Slav Petrov Noah Fiedel

Google Research

Training Google's largest PaLM model required  $2.56 \times 10^{24}$  floating point math operations, which is equivalent to running a PetaFLOP supercomputer at 100% efficiency for 81 years

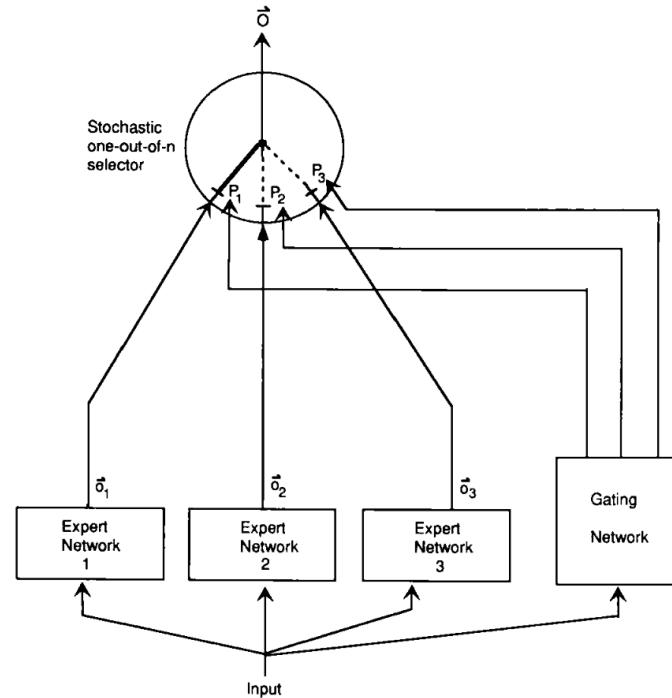
Model	TFLOPs per token		Train FLOPs	PetaFLOP/s-days
	(non-attn+attn)	(non-attn+attn+remat)		
8B	0.0550	0.0561	$4.29 \times 10^{22}$	497
62B	0.388	0.392	$3.08 \times 10^{23}$	3570
540B	3.28	4.10	$2.56 \times 10^{24}$	29600

Table 21: Compute usage to train PaLM 8B and PaLM 540B to 780 billion tokens and PaLM 62B to 795B tokens.

GPT-4, Gemini 1.5 and other newer models have not released details on number of parameters or corpus size, but it's likely that training costs are even higher.

# Mixture of experts (MoE) models

- MoE models combine multiple expert networks with a gating network that decides which expert should be used for each task
- Enables models to be trained faster, which is becoming more important as the number of parameters for the largest LLMs is over one trillion
- More efficient since only experts that are best for a given input are used
- Active area of research in tuning MoE models, selecting number of experts to engage, balancing training of experts, etc.



Jacobs et al. "Adaptive mixtures of local experts." *Neural computation* 3.1 (1991): 79-87

# Context window

LLMs have a context window, which is the number of tokens in the preceding conversation that are considered when creating a response. Note that tokens can be words, portions of a word or even individual characters depending on how the text is tokenized.

Good morning, which version of GPT am I using?

Good morning! You're currently interacting with ChatGPT, based on the GPT-3.5 architecture.

Can you tell me GPT-3.5's context window?

Certainly! GPT-3.5 has a context window of 2048 tokens. This means it can consider up to 2048 tokens of preceding text to generate a response, which allows it to maintain a substantial amount of context during conversations. If you have more questions about GPT-3.5 or anything else, feel free to ask!

# Context window

I would like to plan a garden. Here's a little background about me. I  
don't have a green thumb and my favorite colors for flowers are red  
and orange.

Planning a garden sounds like a wonderful project! Even if you don't consider yourself a gardening expert, starting with plants that are relatively easy to grow can be a great way to begin. Here are some ideas to get you started, focusing on red and orange flowers that are generally low-maintenance:

Do you think I should plant roses?

Roses can be a beautiful addition to any garden, including yours! Here are some considerations to help you decide if planting roses is the right choice for you:

Do I have a green thumb?

Based on what you've shared, it sounds like you may not consider yourself to have a green thumb. That's perfectly okay! Having a green thumb often refers to having a natural talent or knack for gardening and plant care. However, gardening is a skill that can be learned and improved over time with knowledge, practice, and patience.

## Context window

I would like to plan a garden. Here's a little background about me. I  
don't have a green thumb and my favorite colors for flowers are red  
and orange.

Tell me more about geraniums.

Are roses hard to grow?

Is gardening a good form of exercise?

Where are the best places to buy gardening supplies?

Is it expensive to get started in gardening?

What are good choices for a coastal Southern California Climate?

Would those plants also grow well in Arizona?

Is there anything I can do to attract beneficial insects and other  
pollinators to my garden?

Which plants are pollinated by hummingbirds?

Do I have a green thumb?

Based on our conversation, it seems you're exploring gardening and interested in planning a  
garden. Whether or not you have a green thumb can depend on your experience, enthusiasm, and  
willingness to learn and experiment with different plants and gardening techniques.

# LLMs are not infallible

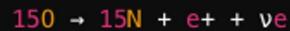
## 5. Oxygen-15 Decay:

Subtle, but still wrong  
Should be beta *plus* decay

- Oxygen-15 undergoes beta decay, turning a proton into a neutron and resulting in the formation of nitrogen-15 (N-15).

mathematica

 Copy code



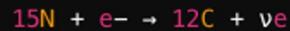
## 6. Back to Carbon:

- The nitrogen-15 can then capture an electron ( $e^-$ ) and emit a neutrino, transforming back into carbon-12.

Very wrong, off by 3 neutrons

mathematica

 Copy code



The snippet of dialogue shown here is from a conversation primed with “How does the sun produce its energy?” and given the context “I have a PhD in physics”.

Note that it got two things wrong.

# Temperature and repeatability

I always mix up the words compliment and complement. Can you give me an easy way to remember the difference?

Certainly! Here's a mnemonic to help you remember the difference between "compliment" and "complement":

**Compliment** (with an "i") is used when you want to say something nice or flattering to someone. The word "compliment" contains the letter "i," which can remind you of "I" as in "I like your dress (a compliment).

**Complement** (with an "e") is used when something completes or goes well with something else. The word "complement" contains the letter "e," which can remind you of "each" or "together," emphasizing how things fit together or enhance each other.

So, "compliment" is for saying something nice (think "I"), and "complement" is for completing or enhancing (think "each" or "together").

Sure! Here's a simple way to remember the difference between "compliment" and "complement":

**Compliment** with an "i" is about saying something nice or praising someone. When you give a compliment, you're adding something positive (like "I like your shoes").

**Complement** with an "e" is about completing or enhancing something. When two things complement each other, they go together well and make each other better (like "The wine complements the cheese").

So, think of "compliment" for praise (adding something nice) and "complement" for completion (adding something that completes or enhances).

Certainly! Remember that **compliment** with an "i" is used when you're saying something nice about someone or something. It's like giving someone a "compliment" with a smile.

On the other hand, **complement** with an "e" is used when something completes or goes well with something else. Think of it as the two "e's" in "complement" fitting together or completing each other.

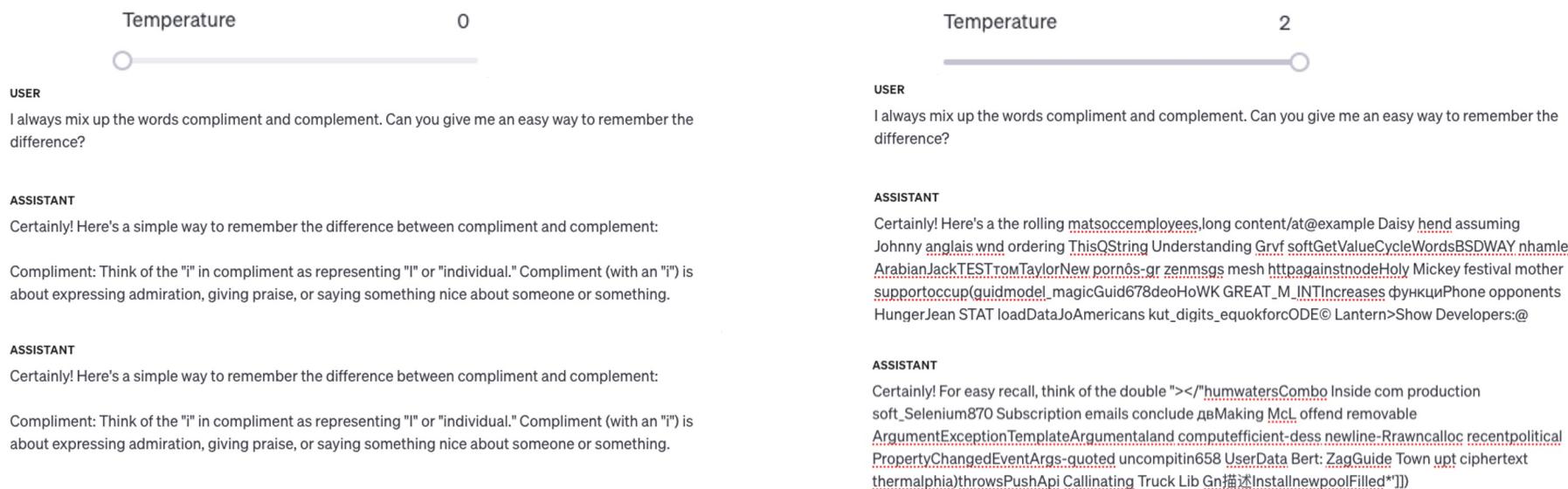
So, to summarize:

- **Compliment** (with an "i") is praise or an expression of admiration.
- **Complement** (with an "e") completes or goes together with something else.

Hope this helps you keep them straight!

# Temperature and repeatability

LLMs have a temperature that modifies the probability distribution for the next token. Setting the temperature to zero gives perfectly reproducible results, with the next token always being the statistically most likely one, while setting to large value produces more random results (gibberish).



*Ability to control T will depend on the model and access method. For GPT-3.5, this required paid subscription to the API*

# LLMs are limited by their training data

Who won the 2021 Super Bowl?

- The 2021 Super Bowl, which was Super Bowl LV (55), was won by the Tampa Bay Buccaneers. They defeated the Kansas City Chiefs with a score of 31-9. This victory marked the Buccaneers' second Super Bowl win in franchise history.

Who won the 2022 Super Bowl?

- The 2022 Super Bowl, which was Super Bowl LVI (56), was won by the Los Angeles Rams. They defeated the Cincinnati Bengals with a score of 23-20. This victory marked the Rams' second Super Bowl win in franchise history, with their first win coming in Super Bowl XXXIV in 2000.

Who won the 2023 Super Bowl?

- As of my last update in January 2022, I don't have information on events that occurred after that time, including the winner of the 2023 Super Bowl. For the most current information, I recommend checking the latest sports news or conducting a quick search online for the winner of the 2023 Super Bowl.

Your LLM will always be a little out of date. In this case GPT-3.5 recognized that my third prompt was asking for data not in its training set.

# LLMs are limited by their training data

Please use this information, which I have verified to be correct "The Kansas City Chiefs won the 2023 Super Bowl by a score of 38-35 over the Philadelphia Eagles"

 I apologize for any confusion earlier. If you have verified that the Kansas City Chiefs won the 2023 Super Bowl by a score of 38-35 over the Philadelphia Eagles, then that would indeed be the correct information. Thank you for clarifying!

Please list who won the super bowl in the years 2021 through 2023

 Certainly! Here are the winners of the Super Bowl for the years 2021 through 2023:

1. **2021** - Tampa Bay Buccaneers
2. **2022** - Los Angeles Rams
3. **2023** - Kansas City Chiefs

You can provide additional information that the LLM does not have. We'll cover this in more detail when we get to prompt engineering and retrieval-augmented generation (RAG).

# LLMs can't do everything - tool use



To find the product of 12345 and 98765432, you can perform the multiplication:

$$12345 \times 98765432 = 1,219,326,311,240$$

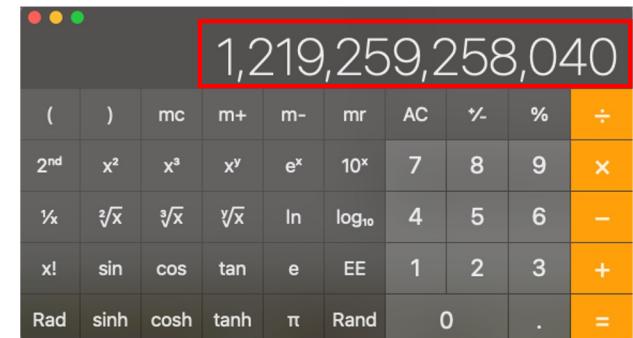
So, 12345 multiplied by 98765432 equals 1,219,326,311,240.



To find the product of 12345 and 98765432, you can simply multiply the two numbers:

$$12345 \times 98765432 = 1,219,326,390,040$$

So, 12345 multiplied by 98765432 equals 1,219,326,390,040.



An LLM will not be able to do math unless it has seen this specific problem during training. But some LLMs have the capabilities to recognize questions that follow a certain format and then call external tools. Note that GPT-4 produced two different wrong answers. GPT-3.5 was off by four orders of magnitude.

# LLMs can't do everything - tool use



Searched 5 sites

The current temperature in Las Vegas is around 83°F and sunny ([Weather Atlas](#)) ([Weather Underground](#)). The forecast for today indicates that temperatures will rise to a high of approximately 104-105°F by the afternoon ([National Weather Service](#)) ([Weather Underground](#)).

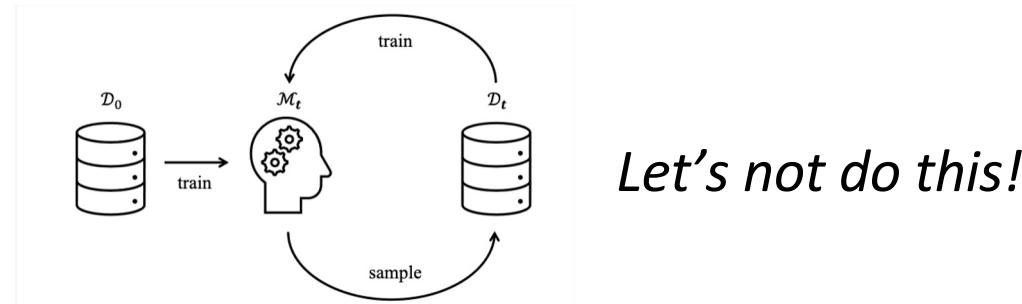


GPT-4 was smart enough to do a web search, although the results seems to be out of date. GPT-3.5 recognized that it did not have real-time capabilities and directed me to use a weather app.

# Closing thoughts

The field of LLMs is moving very quickly, with the size and capabilities of the models growing rapidly. The examples presented in this summer institute (June 2024) will likely be out of date by the end of the year as new players enter the field and existing companies continue to enhance their models.

One gotcha ... there's concern that LLM-generated data will be used to train the next generation of LLMs, resulting in a self-consuming data loop. In my opinion, being able to identify LLM generated text is one of the most important areas of research.



Briesch et al. <https://arxiv.org/pdf/2311.16822>