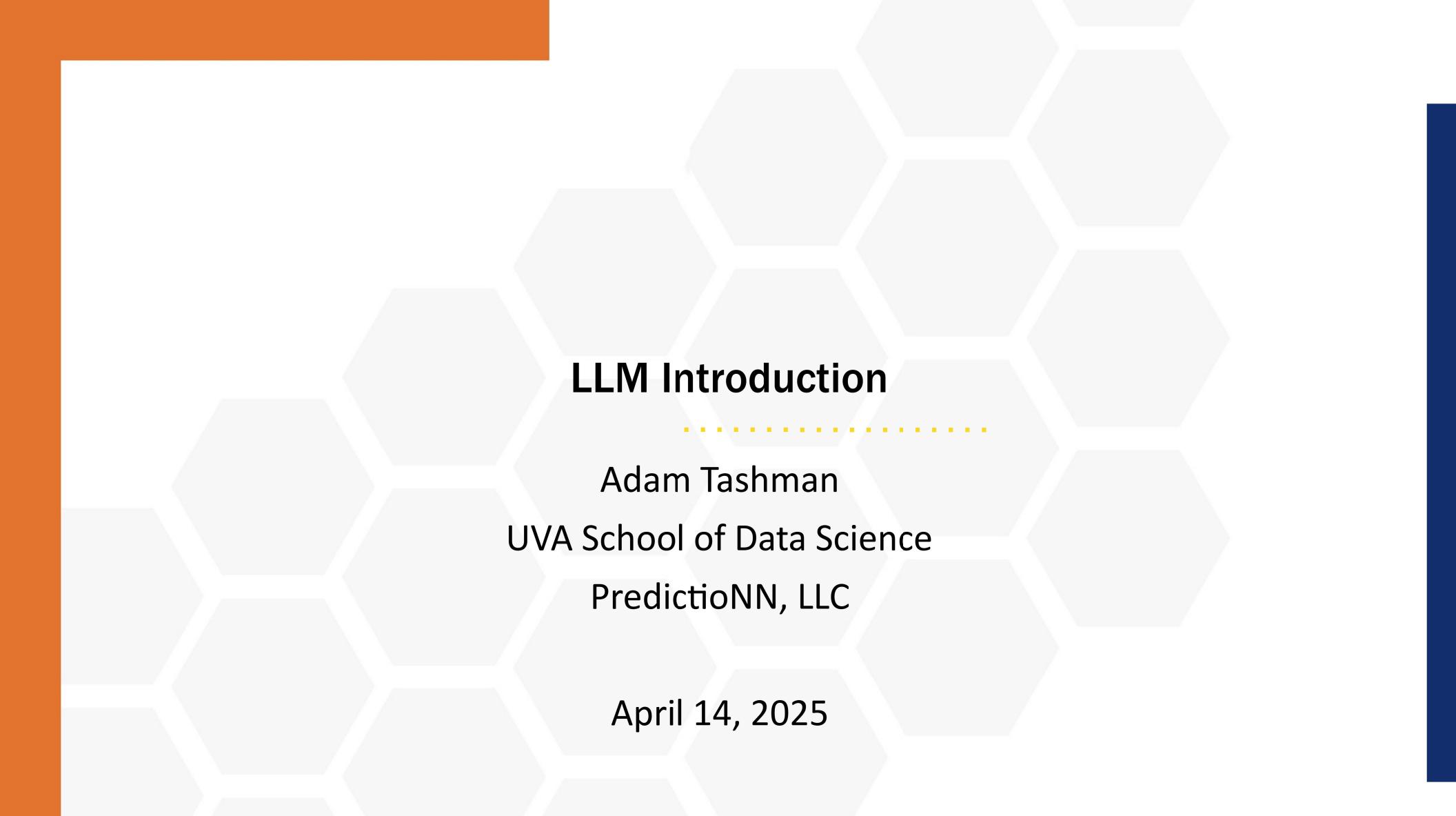


# **LLM Introduction**



Adam Tashman

UVA School of Data Science

PredictioNN, LLC

April 14, 2025

# Agenda

- > Large Language Model (LLM) Attributes  
*size, prompt, token, temperature, training data, memory*
- > Multimedia
- > Privacy & Security
- > Basic Capabilities
- > Q & A

# Disclaimer

For this course, I did use LLMs to generate ideas and images

I verified all of the output

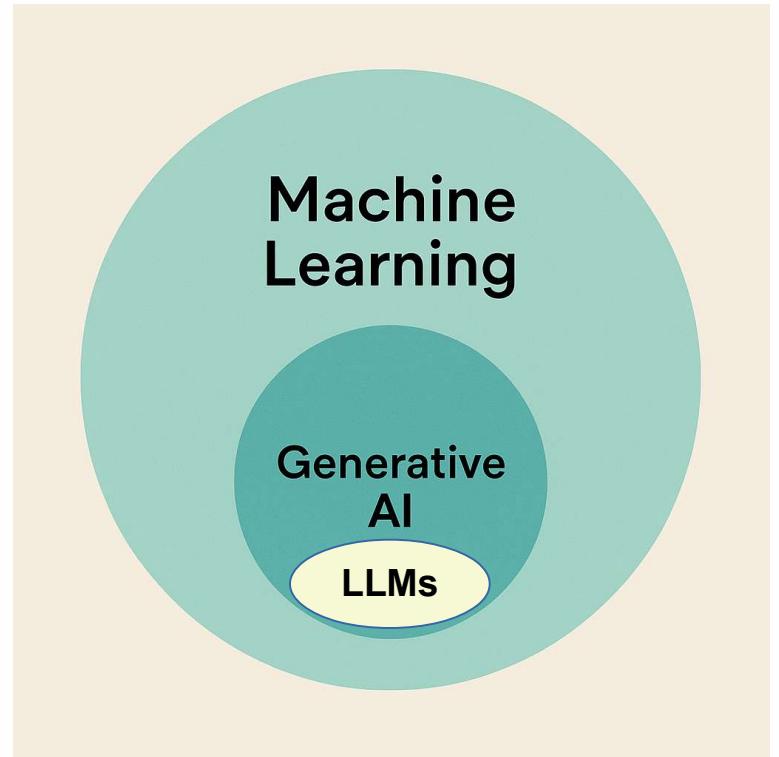
*Tools: ChatGPT, Microsoft Copilot*

# LLMs, Generative AI, ML: How do they relate?

Generative AI is a subset of machine learning (ML)

GenAI focuses on generating new data  
(*text, images, ...*)

LLMs are a subset of Generative AI  
(*text sequence in, text sequence out*)



# How LLMs work

We will develop some intuition about LLMs

This will help form a mental model



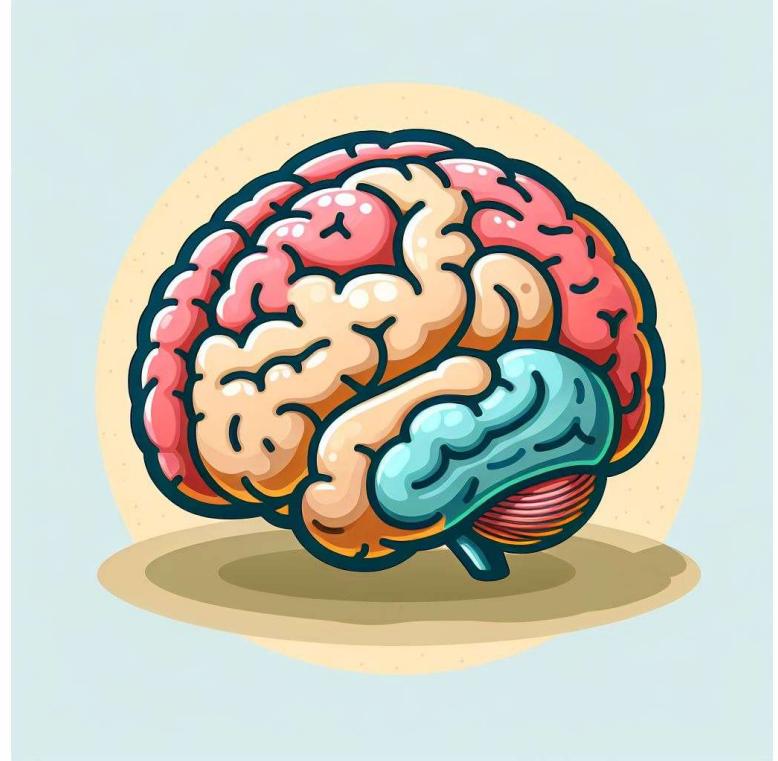
# How LLMs work

Three components:

1 | Vast, diverse datasets

2 | Clever data processing

3 | Massive, flexible architecture (*transformer*)



# The Role of Size in LLMs

LLMs apply **weights** against **inputs** to generate output

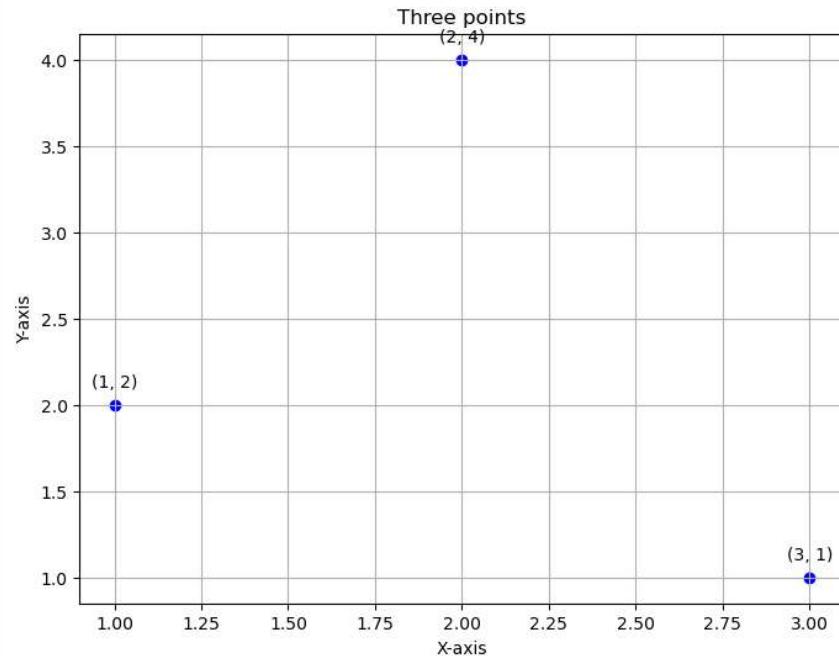
Weights are called *parameters* (they are numbers like 0.2468)

Adding parameters can help with model flexibility

# Size: Small Example

We plot three points. These are inputs. Want something that:

- 1 | Fits the points
- 2 | “Reasonably” fits values between the points

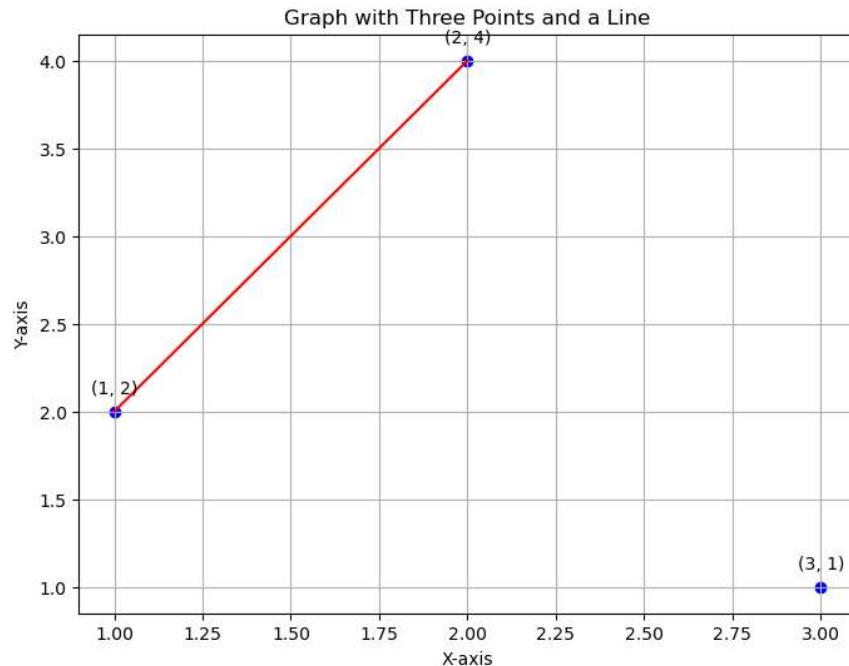


# Size: Small Example

How does this do?

1 | Fits the points ?

2 | “Reasonably” fits values between the points ?

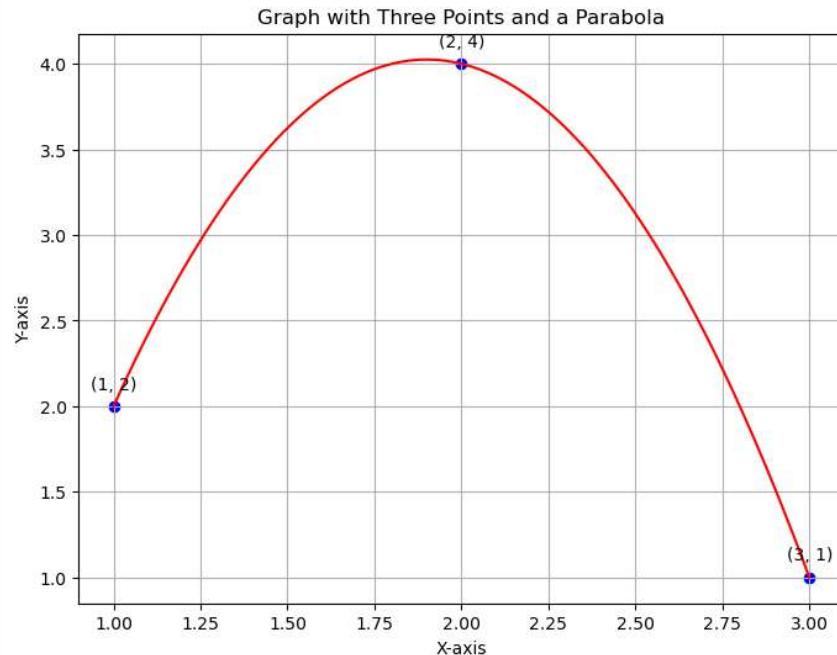


# Size: Small Example

How does this do?

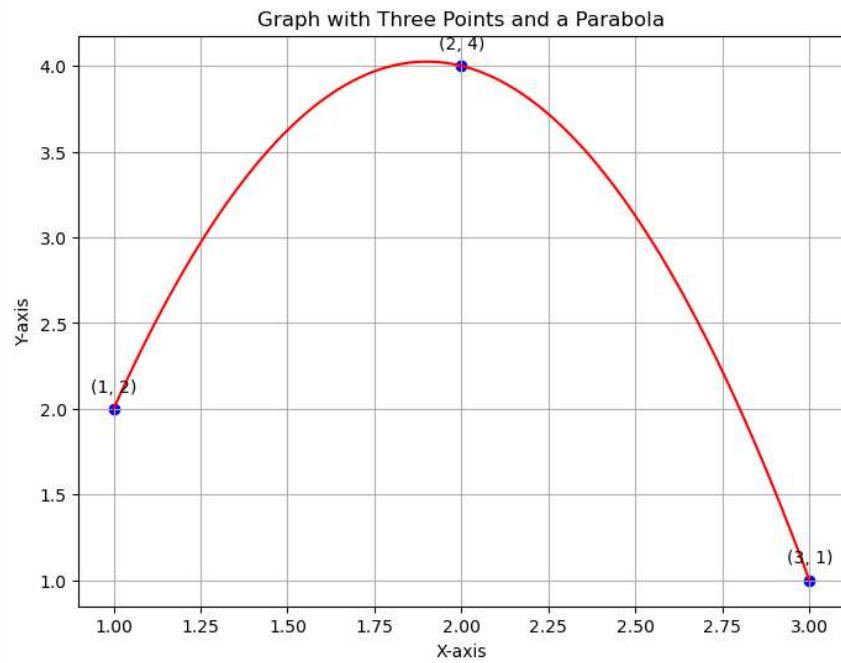
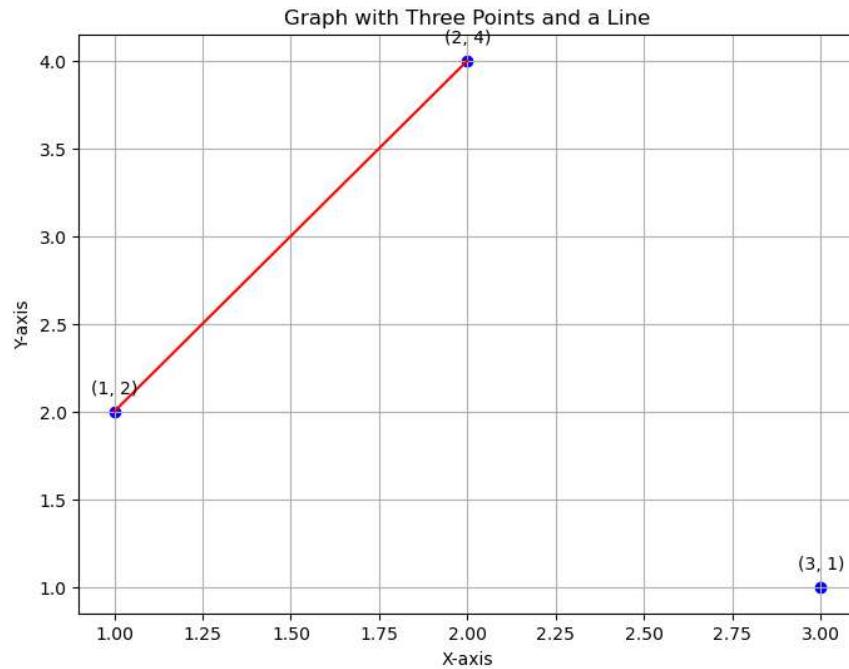
1 | Fits the points ?

2 | “Reasonably” fits values between the points ?



# Size: Small Example

Algebra flashback! How many parameters for the line? Parabola?



# How Large are LLMs?

The size of **Large Language Models (LLMs)** varies depending on their architecture, number of **parameters**, and training data. Here's a breakdown:

## 1 Small vs. Large LLMs

Model	Parameters (Approx.)	Size on Disk	Example Usage
Small Models	<1B parameters	~1-5 GB	Chatbots, embeddings (TinyLlama, DistilBERT)
Medium Models	1B-20B	~5-40 GB	NLP applications (Mistral, LLaMA-2-13B)
Large Models	>20B	~50 GB+	General AI (GPT-4, Claude, Gemini)
Massive Models	100B+	Hundreds of GBs	SOTA AI (GPT-4-turbo, Gemini Ultra)

Indicative Model Size. SOTA = State of the Art

# Prompts

LLM prompts are queries to the model that shape the output (predictions)

Phrasing and content matter a lot, and we will dive deep later

# Prompts

LLM prompts are queries to the model that shape the output (predictions)

Phrasing and content matter a lot, and we will dive deep later

You are a seasoned investor. Which attributes do you look for when making a longer-term investment?

As a seasoned investor, when making a **longer-term investment**, I focus on several key attributes to maximize potential returns while managing risk. Here are the most important ones:

## 1. Strong Business Fundamentals

- **Revenue Growth:** Consistent and sustainable revenue increases.
- **Profitability:** Solid profit margins and a history of earnings growth.
- **Competitive Advantage (Moat):** Unique qualities that protect the business from competitors (e.g., brand strength, patents, network effects).
- **Industry Positioning:** Market leader or disruptor in a growing industry.

# Tokens

LLMs break words into *tokens* and predict next tokens

Token is fundamental unit

- > Helps with efficiency
- > Improves predictive power
- > Pricing based on tokens

On average,  
3 words = 4 tokens

Break this sentence into tokens the way GPT-4 does

Sure! GPT-4 uses Byte Pair Encoding (BPE) for tokenization. Here's an example:

**Sentence:**

📌 "Break this sentence into tokens the way GPT-4 does"

**Tokenized Output (Approximation in GPT-4 BPE):**

css

Copy Edit

```
[ 'Break', 'this', 'sentence', 'into', 'tokens', 'the', 'way', 'GPT', '-', '4', 'does'
```

# Token predictor

Help me autocomplete this sentence, showing the probabilities of the next word:

The best type of pet is a \_\_



Sure! Here are some possible completions for the sentence "The best type of pet is a \_\_" along with their probabilities:

1. **dog** - 40%
2. **cat** - 30%
3. **rabbit** - 10%
4. **fish** - 10%
5. **bird** - 5%
6. **hamster** - 5%

Easter Egg 1 - Q: When was Quad-C founded? A: 1989

# Token predictor: model sensitivity

Model: Chat-GPT

Help me autocomplete this sentence, showing the probabilities of the next word.

the best type of pet is a \_\_

Sure! Here's an autocomplete of your sentence "**The best type of pet is a \_\_**" with a few likely next words and rough probabilities (based on language model predictions):

Next Word	Estimated Probability
dog	45%
cat	30%
rabbit	8%
bird	5%



Copilot

Sure! Here are some probabilities:

1. **dog** - 40%
2. **cat** - 30%
3. **rabbit** - 10%
4. **fish** - 10%
5. **bird** - 5%
6. **hamster** - 5%

# Token predictor: prompt sensitivity

Results can be sensitive to small changes in prompt:

## Prompt 1

Help me autocomplete this sentence, showing the probabilities of the next word:

The best type of pet is a \_\_

Next Word	Estimated Probability
dog	45%
cat	30%
rabbit	8%
bird	5%
hamster	3%
fish	2%

## Prompt 2

I need help autocompleting this sentence, showing the probabilities of the next word:

The best type of pet is a \_\_

Next Word	Estimated Probability
dog	~40%
cat	~30%
rabbit	~8%
fish	~6%

# Token predictor: prompt sensitivity

Results can be sensitive to small changes in prompt, or even rerunning same prompt

Tell me the five companies with largest market cap

As of March 10, 2025, the five companies with the largest market capitalizations are:

1. Apple Inc. (AAPL)

*Market Capitalization: Approximately \$2.79 trillion.*

[EN.WIKIPEDIA.ORG](#)

2. Microsoft Corporation (MSFT)

*Market Capitalization: Approximately \$2.46 trillion.*

[EN.WIKIPEDIA.ORG](#)

3. Saudi Aramco

*Market Capitalization: Approximately \$2.29 trillion.*

[EN.WIKIPEDIA.ORG](#)

4. Alphabet Inc. (GOOG)

*Market Capitalization: Approximately \$1.58 trillion.*

[EN.WIKIPEDIA.ORG](#)

5. Amazon.com, Inc. (AMZN)

*Market Capitalization: Approximately \$1.25 trillion.*

[EN.WIKIPEDIA.ORG](#)

As of March 10, 2025, the five companies with the largest market capitalizations are:

1. Apple Inc. (AAPL)

*Market Capitalization: Approximately \$3.9 trillion.*

[REUTERS.COM](#)

2. Nvidia Corporation (NVDA)

*Market Capitalization: Approximately \$3.28 trillion.*

[REUTERS.COM](#)

3. Microsoft Corporation (MSFT)

*Market Capitalization: Approximately \$3.1 trillion.*

[REUTERS.COM](#)

4. Alphabet Inc. (GOOG)

*Market Capitalization: Approximately \$2.3 trillion.*

[REUTERS.COM](#)

5. Amazon.com, Inc. (AMZN)

*Market Capitalization: Approximately \$2.3 trillion.*

[REUTERS.COM](#)

# Token predictor: prompt sensitivity

Results can be sensitive to small changes in prompt, or even rerunning same prompt

Tell me the five companies with largest market cap

As of March 10, 2025, the five companies with the largest market capitalizations are:

1. Apple Inc. (AAPL)

*Market Capitalization: Approximately \$2.79 trillion.*

EN.WIKIPEDIA.ORG

2. Microsoft Corporation (MSFT)

*Market Capitalization: Approximately \$2.46 trillion.*

EN.WIKIPEDIA.ORG

3. Saudi Aramco

*Market Capitalization: Approximately \$2.29 trillion.*

EN.WIKIPEDIA.ORG

4. Alphabet Inc. (GOOG)

*Market Capitalization: Approximately \$1.58 trillion.*

EN.WIKIPEDIA.ORG

5. Amazon.com, Inc. (AMZN)

*Market Capitalization: Approximately \$1.25 trillion.*

EN.WIKIPEDIA.ORG

As of March 10, 2025, the five companies with the largest market capitalizations are:

1. Apple Inc. (AAPL)

*Market Capitalization: Approximately \$3.9 trillion.*

REUTERS.COM

2. Nvidia Corporation (NVDA)

*Market Capitalization: Approximately \$3.28 trillion.*

REUTERS.COM

3. Microsoft Corporation (MSFT)

*Market Capitalization: Approximately \$3.1 trillion.*

REUTERS.COM

4. Alphabet Inc. (GOOG)

*Market Capitalization: Approximately \$2.3 trillion.*

REUTERS.COM

5. Amazon.com, Inc. (AMZN)

*Market Capitalization: Approximately \$2.3 trillion.*

REUTERS.COM

**ChatGPT searches the web, pulling from a different source each time**



# Temperature and Randomness

Output from LLMs can exhibit randomness

It's a feature, not a bug!

*Temperature* can control the randomness

Higher temperature → choose less probable words

Useful for:

- > creative writing
- > dialogue generation
- > exploring new possibilities



# Low Temperature

Who started the personal computer revolution? Use a temperature of 0.2.

With **temperature = 0.2**, the model will generate a **factual and deterministic answer** with minimal randomness. Here's the response:

- The personal computer revolution was started by pioneers such as Steve Jobs and Steve Wozniak (Apple), Bill Gates and Paul Allen (Microsoft), and early computing companies like IBM and Xerox PARC.

Easter Egg 2 – Q: Quad-C target industries. A: Business Services, Industrials, Healthcare

# High Temperature

Who started the personal computer revolution? Use a temperature of 2.

With a **temperature of 2.0**, the response would be highly **random and unpredictable**, so let's simulate that level of **creative chaos**:

 "The personal computer revolution was secretly started by an ancient order of cybernetic squirrels who whispered binary code into Steve Jobs' dreams while Bill Gates deciphered alien punch ...

# Setting Temperature

Temperature is a parameter that can be set 1) in the GUI 2) with the API.

Example using OpenAI API:

```
import openai

response = openai.ChatCompletion.create(
    model="gpt-4",
    messages=[{"role": "user", "content": "What is the best pet?"}],
    temperature=0.7 # Adjust this value (0.1 - 1.5)
)

print(response["choices"][0]["message"]["content"])
```

# Training Data

LLMs make predictions based on their training data

Much of this comes from the web:  
**books, literature, Wikipedia, blogs, code**

If the doc / writing style is highly represented,  
the AI will likely do well

LLMs can produce original material



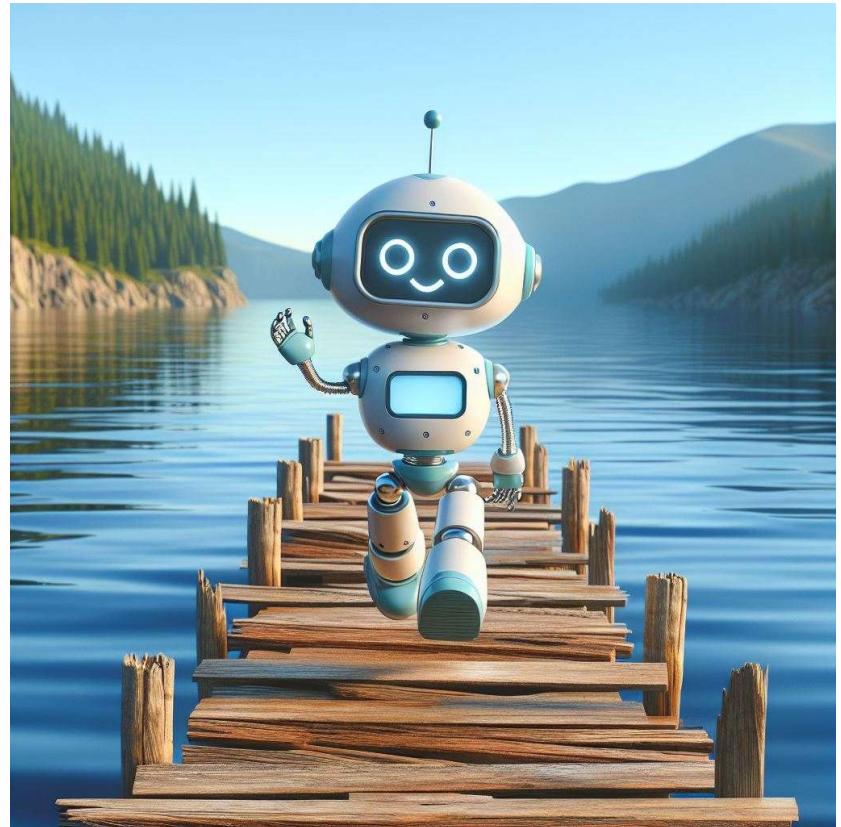
Source: <https://www.oneusefulthing.org/p/thinking-like-an-ai>

# Going Outside the Training Data

When the answer isn't in the training data,  
The LLM can make up the answer

It can sound convincing

There are mitigating approaches



# Training Data – Staying Current

ChatGPT ▾

Let's see a technical analysis of Reddit

Reddit, Inc. is a privately held company and does not have publicly traded stock. Therefore, traditional technical analysis, which relies on publicly available stock price and volume data, is not applicable to Reddit.

**The LLM is not aware Reddit is now publicly traded**

# Model Training vs Tuning

A quick note to compare training vs tuning

- > LLM training is a discrete event  
Requires substantial compute resources, cost
- > Some users will fine tune a model  
for specific purpose
- > Various methods for fine tuning.  
Cloud providers lighten the lift  
(e.g., Amazon Bedrock)



# Memory

LLMs have limited memory, called the *context window*

Larger window can process more information,  
helping with complex tasks

Tradeoff: larger window requires more compute

Context windows have been growing



# Translating Context Window to Real Use

Select models and their context windows:

Newer models

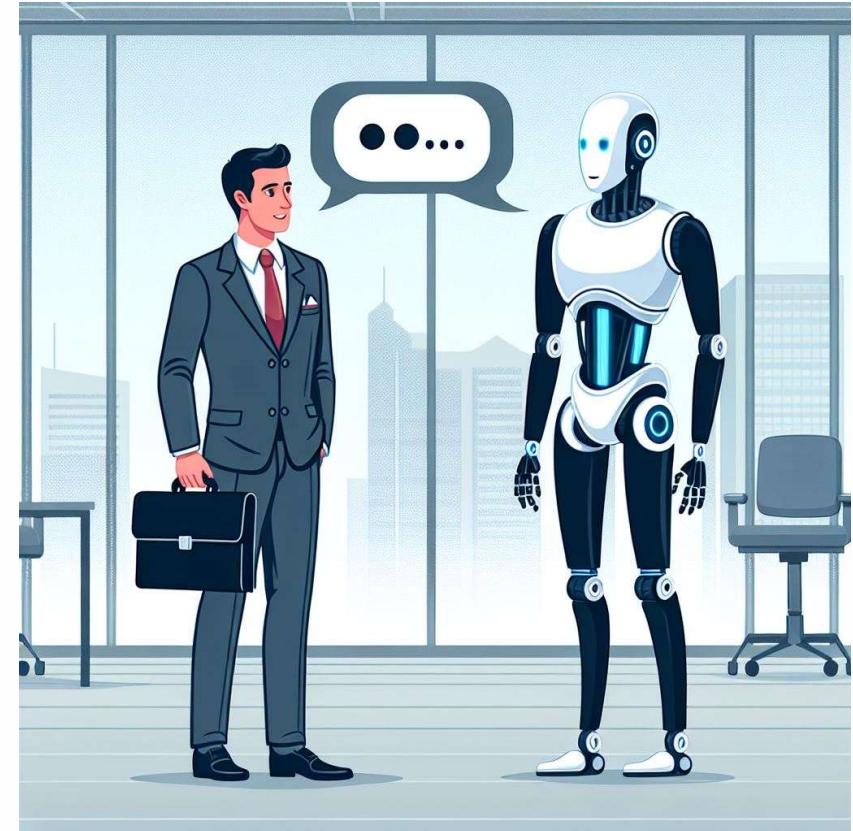
- GPT-2 (1024 tokens) → Can handle a **short article**.
- GPT-4 Turbo (128K tokens) → Can analyze a **full book or research paper**.
- Claude 2.1 (200K tokens) → Can process an **entire novel or multiple reports**.
- Gemini Ultra (1M+ tokens, speculated) → Could handle **entire textbooks or legal documents**.

# Chat Sessions

Chat works best when it remembers your answers!

If you stay in a single chat session,  
and if conversation fits in context window,  
**the AI will include history**

If you start a new chat, you start fresh



Goal: Solve problems robustly by piecing together various kinds of information

AI is now doing this across fields:

## Healthcare

Data

- structured data (meds, vitals)
- clinical notes
- imaging (MRI, PET)
- waveforms (ECG)

Use cases:

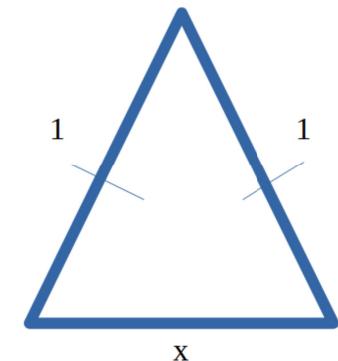
- Predict arrhythmia
- Locate tumors

## EdTech

Data

What is the value of  $x$ ?

- voice
- images
- text



Use cases:

- tutoring
- question generation

- > **Text:** handles language, code, structured data (tables)
- > **Images:** analyze and describe images, perform OCR, recognize patterns
- > **File uploads:** process various formats (text, CSV, PDF)
- > **Audio:** Whisper API can transcribe audio file formats (e.g., MP3, MP4)

# Privacy and Security

Early concern that LLMs were using input and output to train models

Privacy is a must for many organizations

What privacy options are available?



# Privacy and Security

Early concern that LLMs were using input and output to train models

Privacy is a must for many organizations

What privacy options are available?

- > Users can opt out of having interactions used to train models
- > Cloud providers use encryption, data isolation
- > Running models locally



# Basic Capabilities

**Text Generation:** Draft emails, reports

**Code Generation:** Accelerate development

**Question Answering:**

Answer questions given information

**Text Summarization:**

Reduce meeting notes, documents

**Translation:**

Language translation, captioning



\* Always verify output

# Q & A

