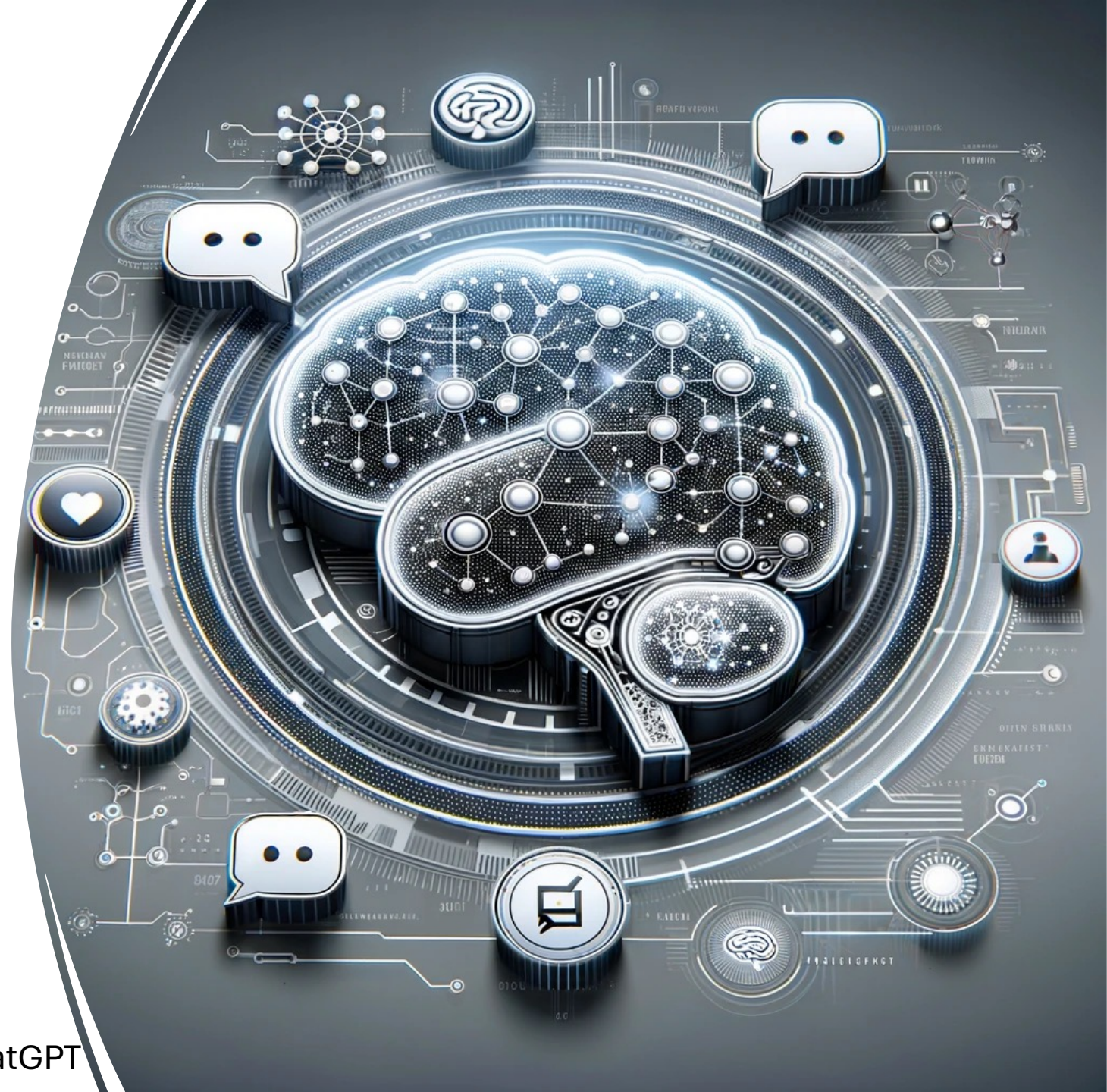


ChatGPT

What's it?

How to use it?

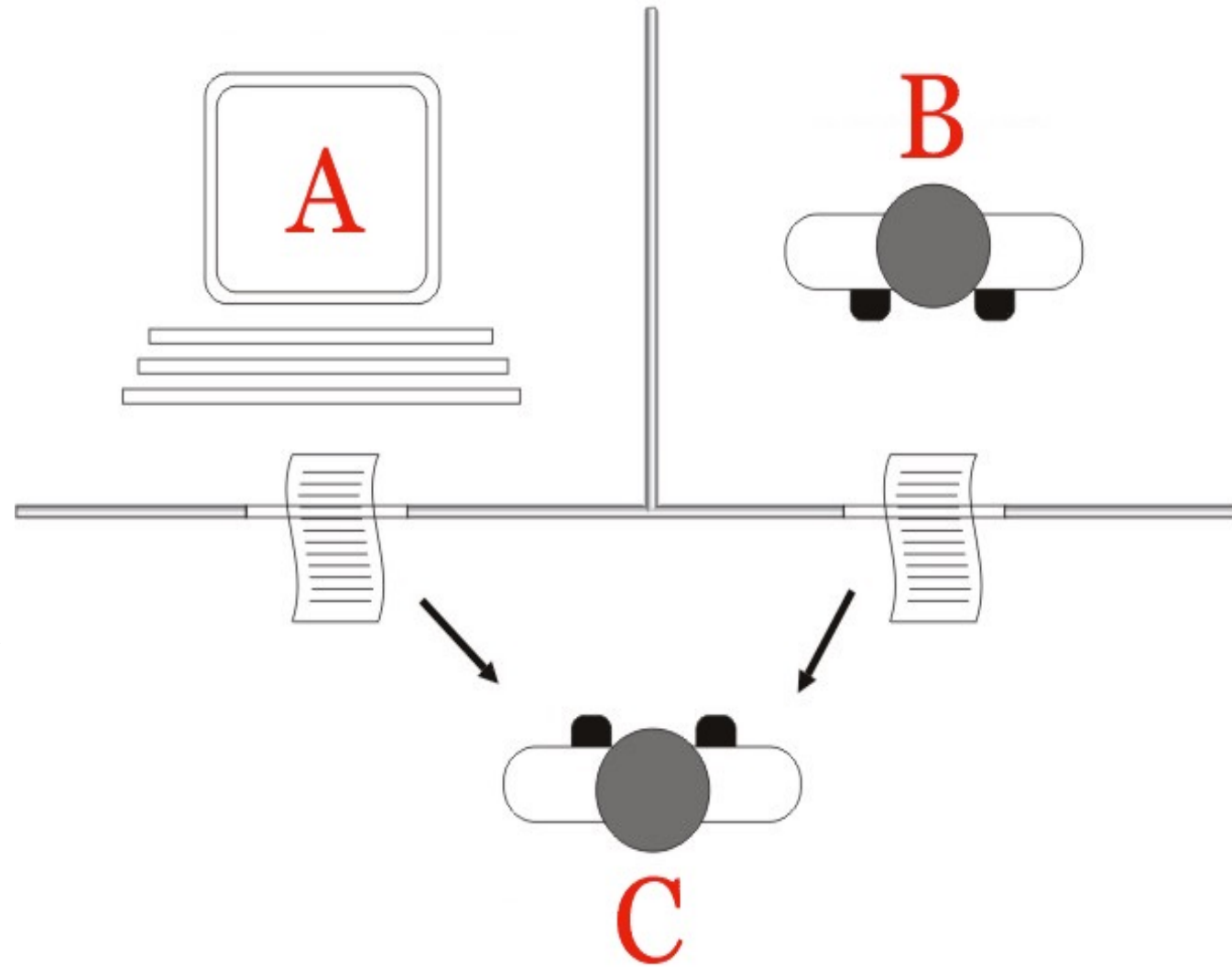


The image was generated by ChatGPT

Turing Test

- The **Turing test**, originally called the **imitation game** by [Alan Turing](#) in 1950, is a test of a machine's ability to exhibit intelligent behaviour equivalent to, or indistinguishable from, that of a human.
- The player C, an interrogator, is given the task of trying to determine which player – A or B – is a computer and which is a human. The interrogator is limited to using the responses to written questions to make the determination

Source: https://en.wikipedia.org/wiki/Turing_test



History

- ChatGPT is a product of OpenAI's ongoing advancements in natural language processing and artificial intelligence
- Versions
 - GPT-1 was first introduced in 2018. GPT-2 in 2019 showcases significant improvements in generating coherent and contextually relevant text.
 - GPT-3, released in June 2020, further expanded the model's capabilities with 175 billion parameters, making it one of the most powerful language models to date.
 - GPT-4, released in March 2023, has larger undisclosed parameter size.
 - GPT-4o, released on 05/13/2024, supports multimodal (image and voice) model. **Training data is up to Oct 2023.**

First Name

Job - a "conversationalist" to human customers

Chat GPT

Innate learned abilities - "innate potentials"

Family Name

G: Generative
P: Pre-Trained
T: Transformer

Source: <https://github.com/wenjing/Get-to-Know-ChatGPT>

What is GPT

- **Generative:** Refers to the model's ability to generate text, producing coherent and contextually relevant sentences from a given prompt.
- **Pre-trained:** Indicates that the model has been trained on a large dataset prior to fine-tuning, allowing it to learn language patterns, grammar, and facts from vast amounts of text data.
- **Transformer:** Describes the neural network architecture used, which relies on self-attention mechanisms to process and generate text efficiently, capturing long-range dependencies in language.

Source: ChatGPT

GPT, LLM, NLP, AI

- GPT is an example of a Large Language Model (LLM).
- LLM is used in natural language processing (NLP) and AI.
- Other LLM applications that may use other models
 - Text summarization
 - Sentiment analysis
 - Translation
 - Information retrieval
 - Text classification

LLM

- A large language model (LLM) is a language model notable for its ability to achieve general-purpose language understanding and generation.
- LLMs acquire these abilities by learning statistical relationships from text documents during a computationally intensive self-supervised and semi-supervised training process.
- LLMs are artificial neural networks following a transformer architecture.

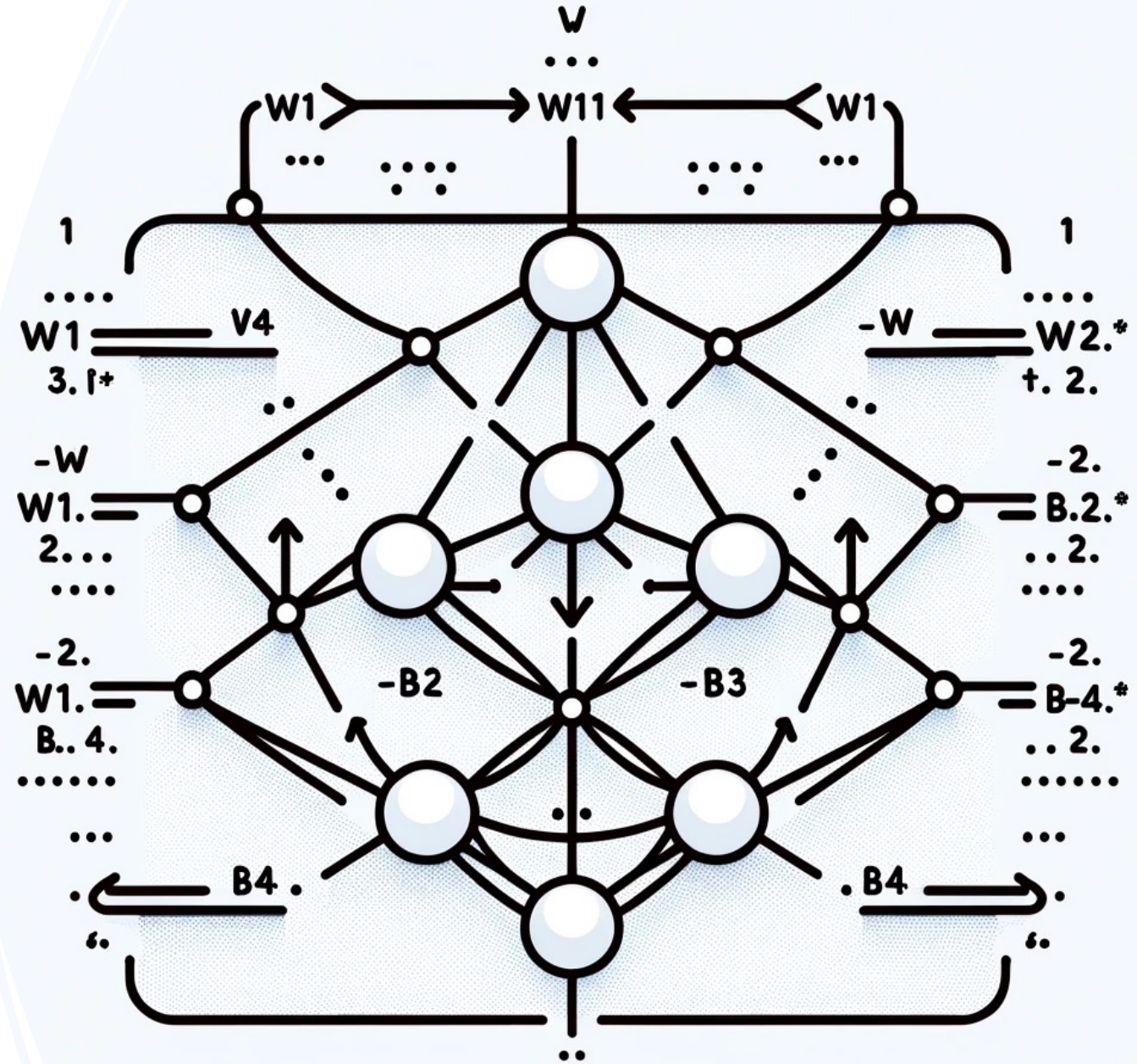
Source: https://en.wikipedia.org/wiki/Large_language_model

The Size Matters

- LLM model is basically a neural network with a lot of parameters. Roughly speaking, the more parameters, the better the model is.
- GPT-3 has 175 billion parameters, and GPT-4 **may** have over 1 trillion parameters.
- It costs millions of dollars to train a LLM.

What Are Parameters?

- Parameters in Large Language Models (LLMs) are the numerical values that the model learns during the training process.
- These parameters are the weights and biases within the neural network that determine how the model processes input data and generates output.



Why So Many

1. **Complexity of Human Language:** Human language is incredibly complex and nuanced, requiring models to understand a wide range of linguistic patterns, idioms, contexts, and relationships between words.
2. **Accuracy and Performance:** Larger models with more parameters can capture more details and subtleties in data, leading to higher accuracy and better performance on a variety of tasks, such as translation, summarization, and question answering.
3. **Generalization:** More parameters help the model generalize better from training data to unseen data, making it capable of handling a wider range of inputs and providing more coherent and contextually relevant outputs.

An Analogy

Think of a neural network like a very complex recipe:

- **Ingredients** (Inputs): The words or sentences you want the model to understand or generate.
- **Recipe Steps** (Neurons and Layers): The processes the model uses to transform the inputs into an output. Each step (neuron) combines ingredients in a certain way.
- **Adjustable Knobs** (Parameters): The knobs you can turn to tweak the recipe. More knobs (parameters) mean you can make finer adjustments to get the recipe just right.

When you have millions or billions of knobs (parameters), you can adjust the recipe in very detailed ways, allowing the model to handle a wide variety of texts and generate high-quality responses.

Source: ChatGPT

Meta's Llama-2-70b Model Training

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

Pre-Training



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**.



<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...

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.

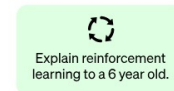
Fine Tune and RLHF (Open AI)

To make our models safer, more helpful, and more aligned, we use an existing technique called reinforcement learning from human feedback (RLHF).

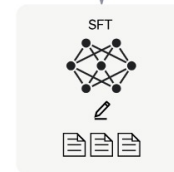
On prompts submitted by our customers to the API, our labelers provide demonstrations of the desired model behavior, and rank several outputs from our models. We then use this data to fine-tune GPT-3.

Step 1
Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.



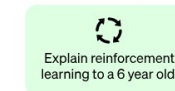
A labeler demonstrates the desired output behavior.



This data is used to fine-tune GPT-3.5 with supervised learning.

Step 2
Collect comparison data and train a reward model.

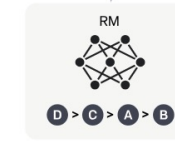
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.

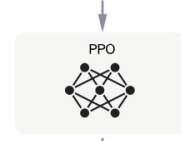


Step 3
Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

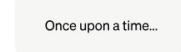
A new prompt is sampled from the dataset.



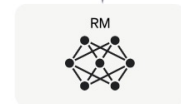
The PPO model is initialized from the supervised policy.



The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



Model Inference

- When the model is trained and ready, a user queries the model with a question, the question text is passed into that 140GB file and processed character-by-character then return the most relevant text as result outputs.
- By meaning the **most relevant**, it means the model will return the text that are most likely to be the next character of the input text. For example:

```
> Input: "I like to eat"
```

```
> Output: "apple"
```

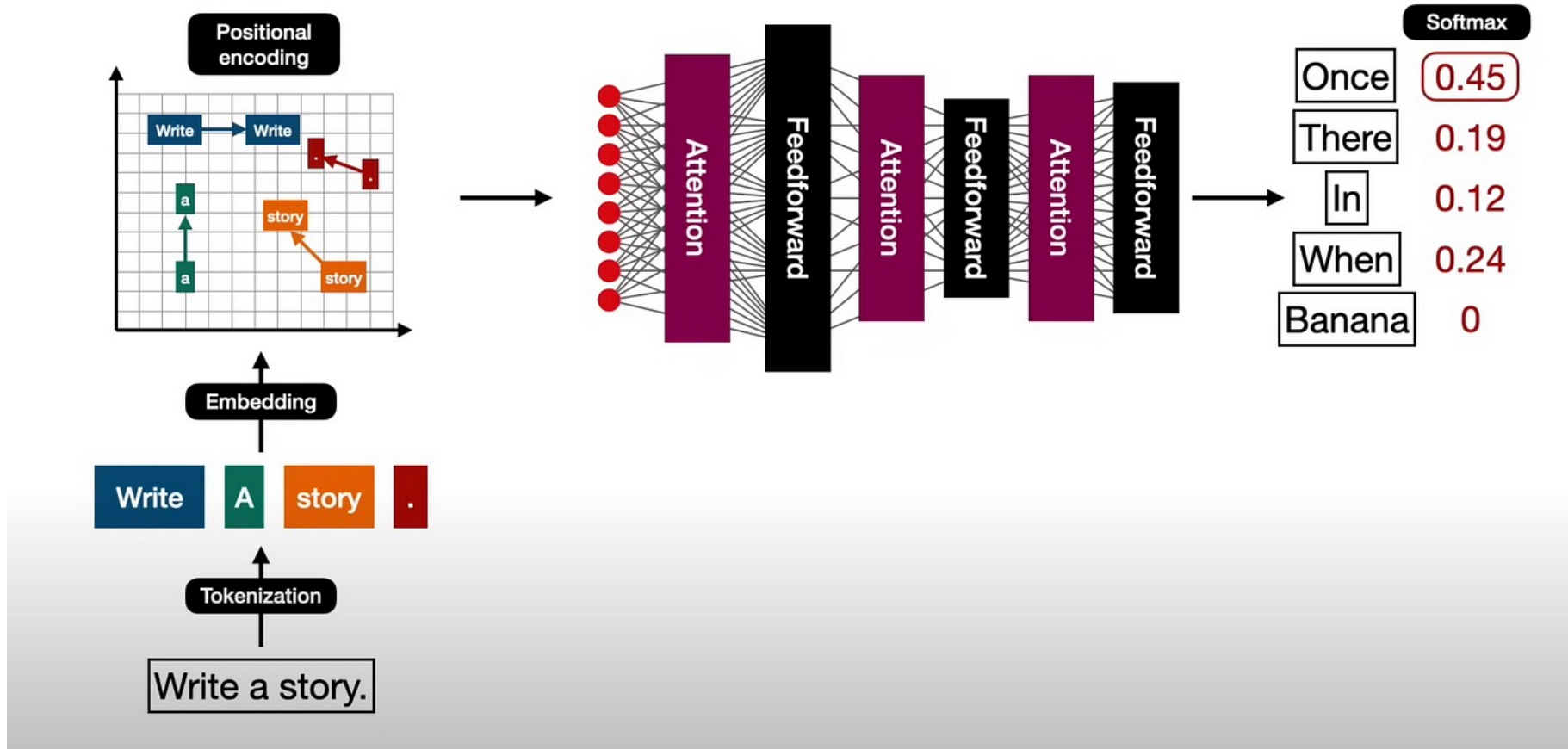

Probabilistic Database

- *I like to eat apple* is a very common sentence that the model *learned* for multiple times.
- *I like to eat banana* is also a common sentence but less common than the one above.

So basically, an LLM model is a probabilistic database that assigns a probability distribution to any given character and its relevant contextual characters.

This sounds impossible before. But since the paper [“Attention is all you need”](#) was published in 2017, the **transformer architecture** was introduced to enable such **large-scaled contextual** understanding possible by training a neural network on **a very large dataset**.

The Architecture (Implementation)



Successful Applications?

Source: ChatGPT

- Customer service chatbots: can help businesses save time and resources by automating customer service tasks.
- *Personalized* assistants: can be trained on the individual's preferences and history, allowing them to provide more personalized and relevant responses.
- Education tools: A TED talk - The Amazing AI super tutor for students and teachers (https://www.ted.com/talks/sal_khan_the_amazing_ai_super_tutor_for_students_and_teachers/c)
- Mental health chatbots: can offer a non-judgmental and confidential space for individuals to talk about their feelings and receive guidance and support.

For College Students

Source: ChatGPT

1. Research assistance: ChatGPT can help students find relevant sources and information for their research papers. Students can input a topic or question, and ChatGPT can generate responses with links to articles, books, and other resources.
2. Study assistance: ChatGPT can help students study by generating quiz questions or providing explanations for difficult concepts. Students can input a topic they are studying, and ChatGPT can provide questions and explanations to help them better understand the material.
3. Writing assistance: ChatGPT can help students improve their writing skills by providing suggestions for sentence structure, word choice, and overall clarity. Students can input a sentence or paragraph, and ChatGPT can generate suggestions for improvement.
4. Personal organization: ChatGPT can help students stay organized by providing reminders and to-do lists. Students can input tasks they need to complete, and ChatGPT can generate reminders and alerts to help them stay on track.

Principles of Using ChatGPT

- Source: <https://www.gamingdeputy.com/lecun-strongly-recommends-dr-harvard-shared-the-use-of-gpt-4-for-scientific-research-down-to-every-workflow/> Kareem Carr, A Ph.D. in biostatistics from Harvard University Shared The Use Of GPT-4 For Scientific Research, Down To Every Workflow
- The first principle
 - *Never ask a large language model (LLM) for information that you cannot verify yourself or ask it to perform a task that you cannot verify has been done correctly.* The only exception is that it's not a critical assignment, eg, asking LLM for apartment decorating ideas.
 - ChatGPT often gives wrong answers.

More Hints

- 2nd: Clearly tell LLM what information should be used, use terminology and symbols to orient the LLM to the correct contextual information.
- 3rd: literately, be more specific and define the output format (code, tutorial, article). ChatGPT remembers chat history (128K words, 8K for free users).
- 4th: ask one more time for different (might be better) answers.

For Paid API Users

- Appending “Be Concise” saves 40-90%
- LLMs operate on tokens. Words may have more than one tokens. For example, “eating” has two tokens: “eat” and “ing”. A 750-word document will be about 1000 tokens.

Source: <https://github.com/ray-project/llm-numbers>

Useful Resources

- Awesome ChatGPT Prompts

<https://github.com/f/awesome-chatgpt-prompts>

- Prompt Engineering Guide

<https://www.promptingguide.ai/introduction/basics>

Please Use It

THANK YOU!

