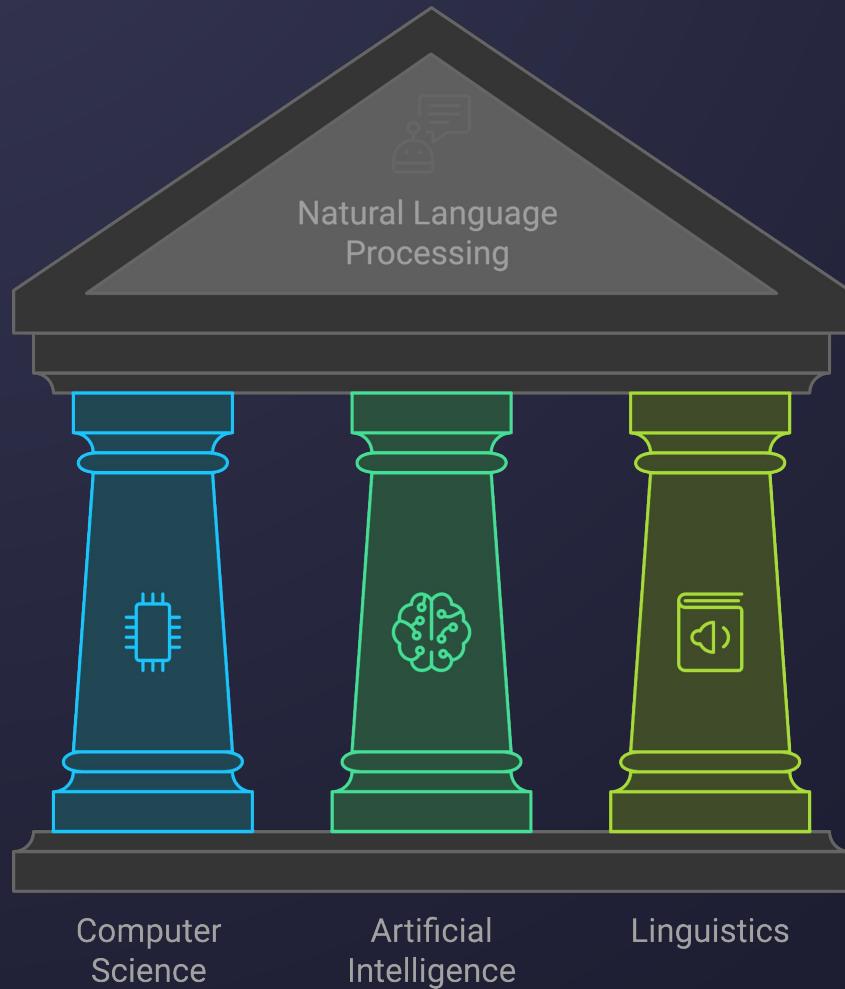


# NLP Basics and LLM Intro

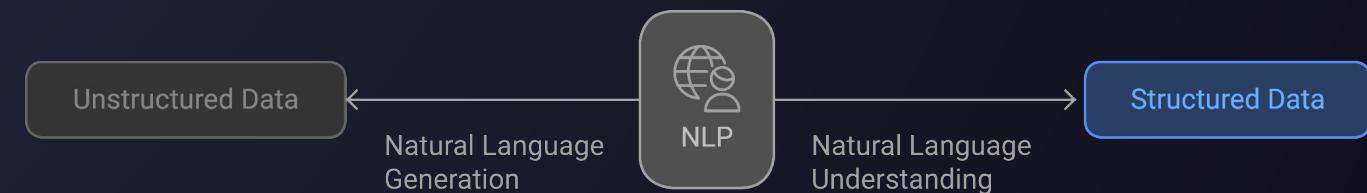


# Introduction to NLP

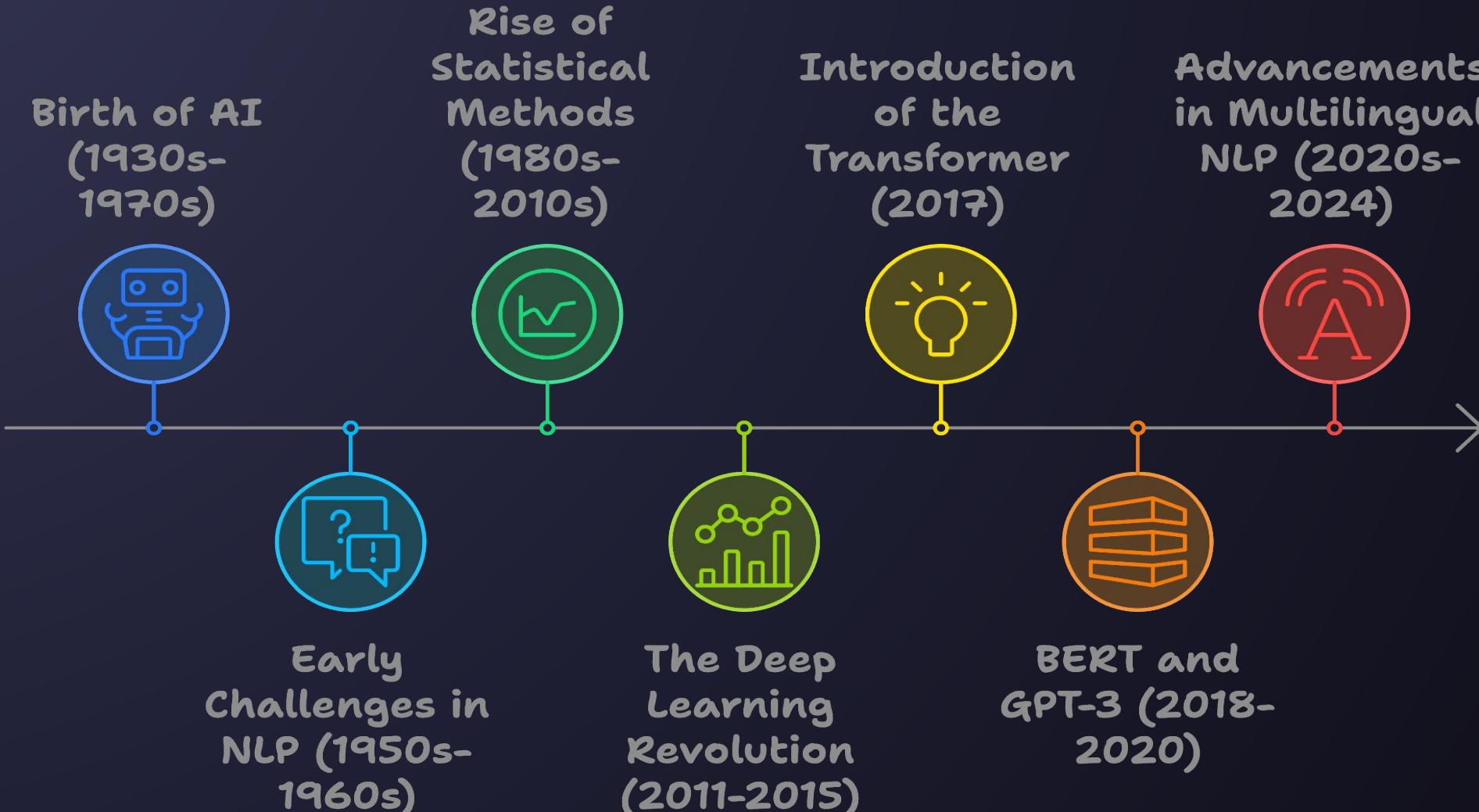
# What is NLP?



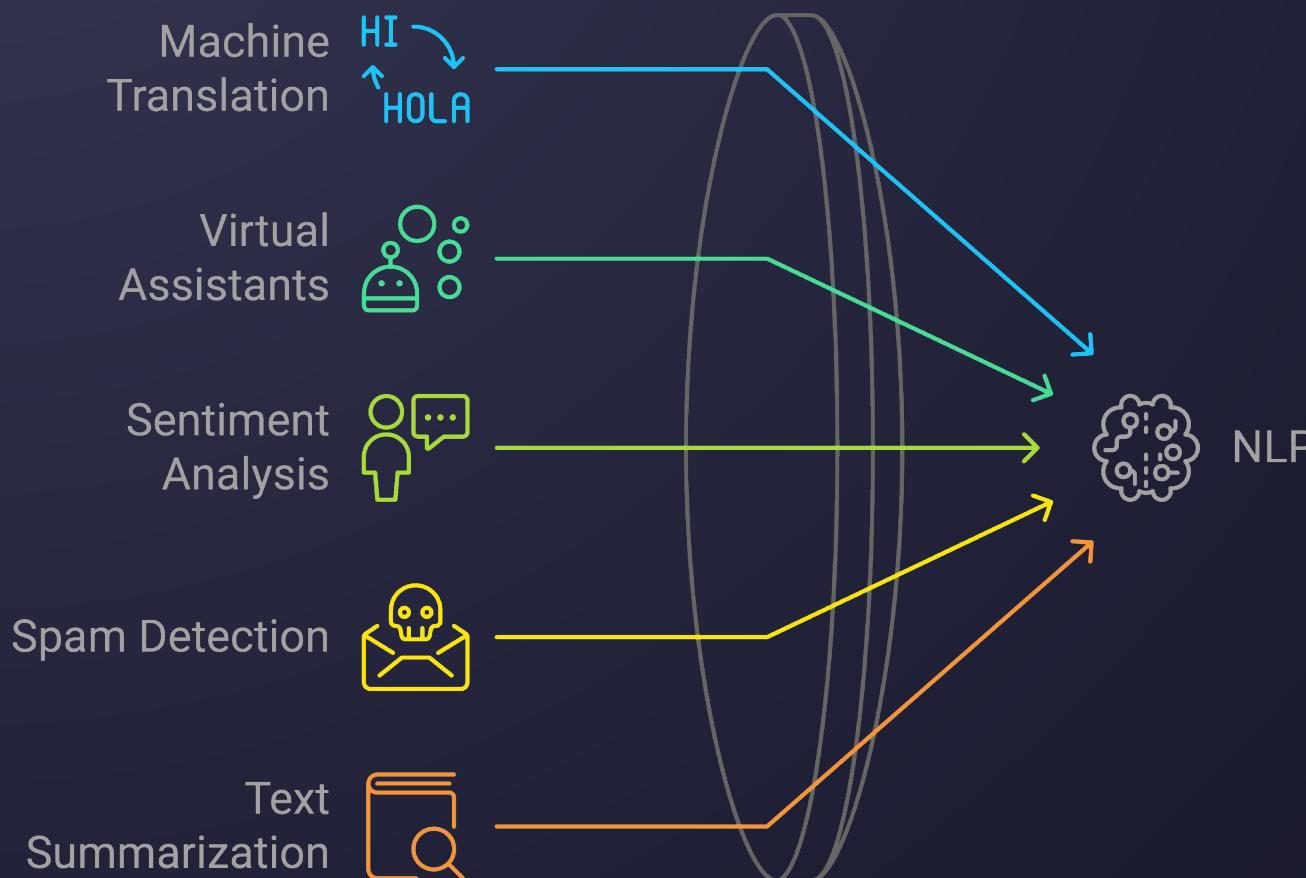
NLP is a branch of AI that enables computers to **comprehend, generate, and manipulate** human language.



# A Brief History of NLP



# Examples of NLP Tasks



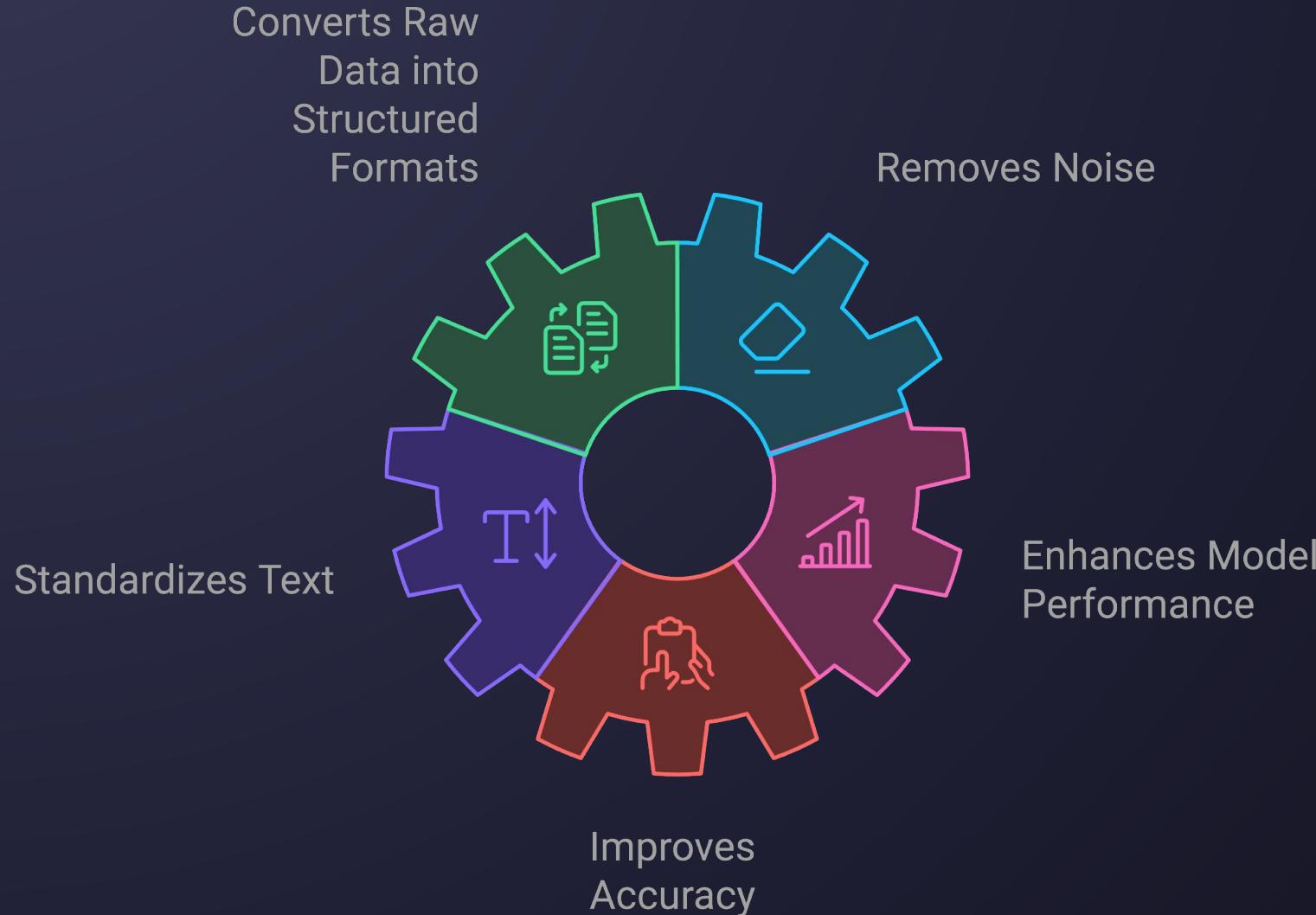
**Discussion question:** How have you used NLP in your work, or how have you seen it used in your daily life?

# Challenges in NLP



However, with the recent innovations in LLMs, some of these challenges are being mitigated...

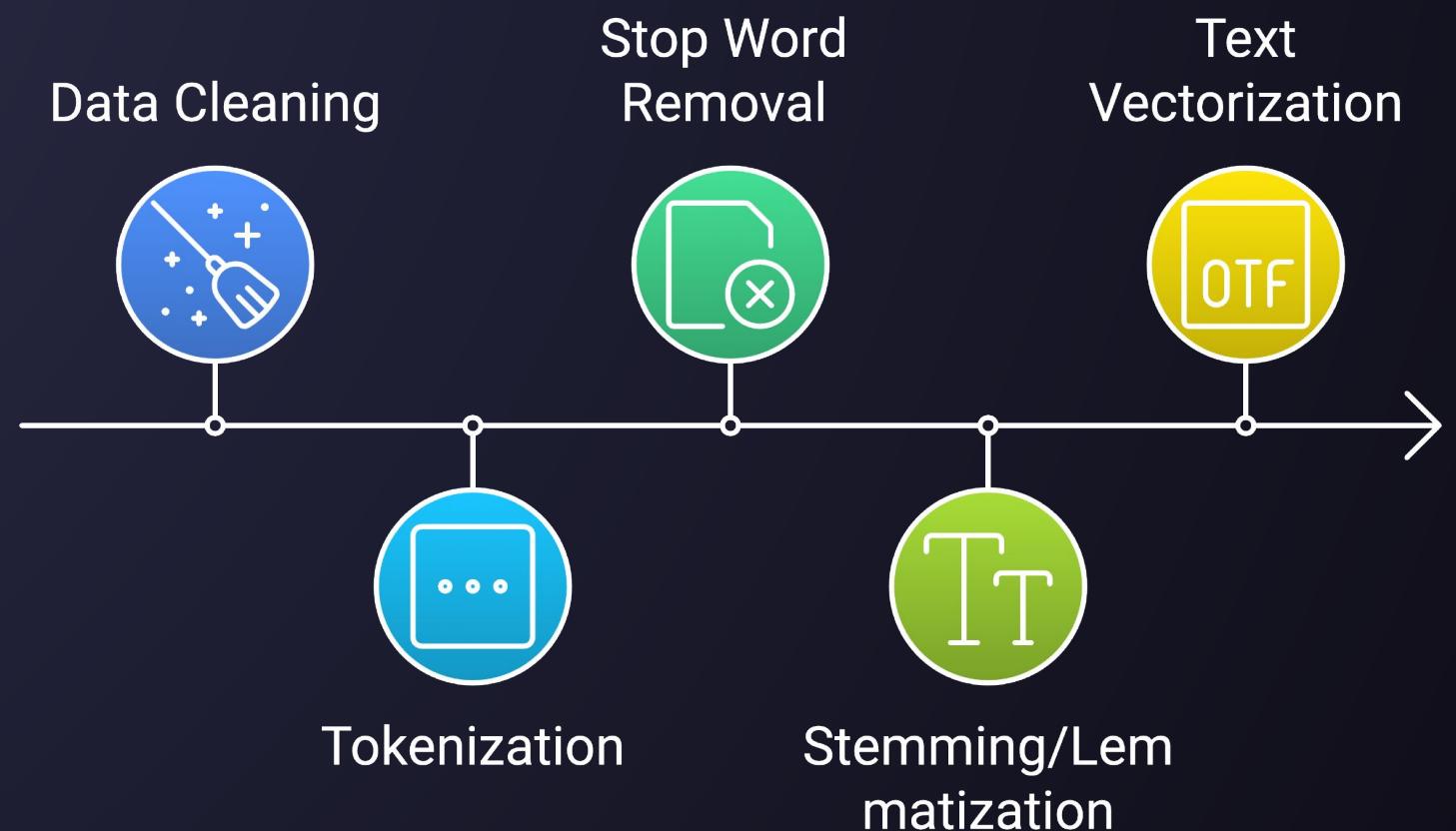
# Understanding Text Processing



**Discussion question:**  
What are some examples  
of how you think textual  
data should be “cleaned”  
before it is analyzed?

# Key Text Processing Techniques

These are generally the steps in an NLP data cleaning / text processing pipeline, but it may vary depending on your use case.



# Key Text Processing Techniques (Code)

[https://github.com/angelaateng/AIR\\_AI\\_Engineering\\_Course\\_2024/blob/  
main/Day1/0\\_NLP\\_TextPreprocessing.ipynb](https://github.com/angelaateng/AIR_AI_Engineering_Course_2024/blob/main/Day1/0_NLP_TextPreprocessing.ipynb)

# Hands on Example: Text Preprocessing using NLTK and SpaCy

# Text Preprocessing Activity

## Instructions:

1. Download the IMDB Dataset from [here](#) or directly from our Github [here](#)
2. Tokenization with NLTK and SpaCy
3. Stopword Removal
4. Stemming & Lemmatization
5. **Compare Results.** Once you're done, look at the differences between NLTK and spaCy. Which approach do you think is more suitable for this dataset? What are some differences that you saw in these 2 packages?



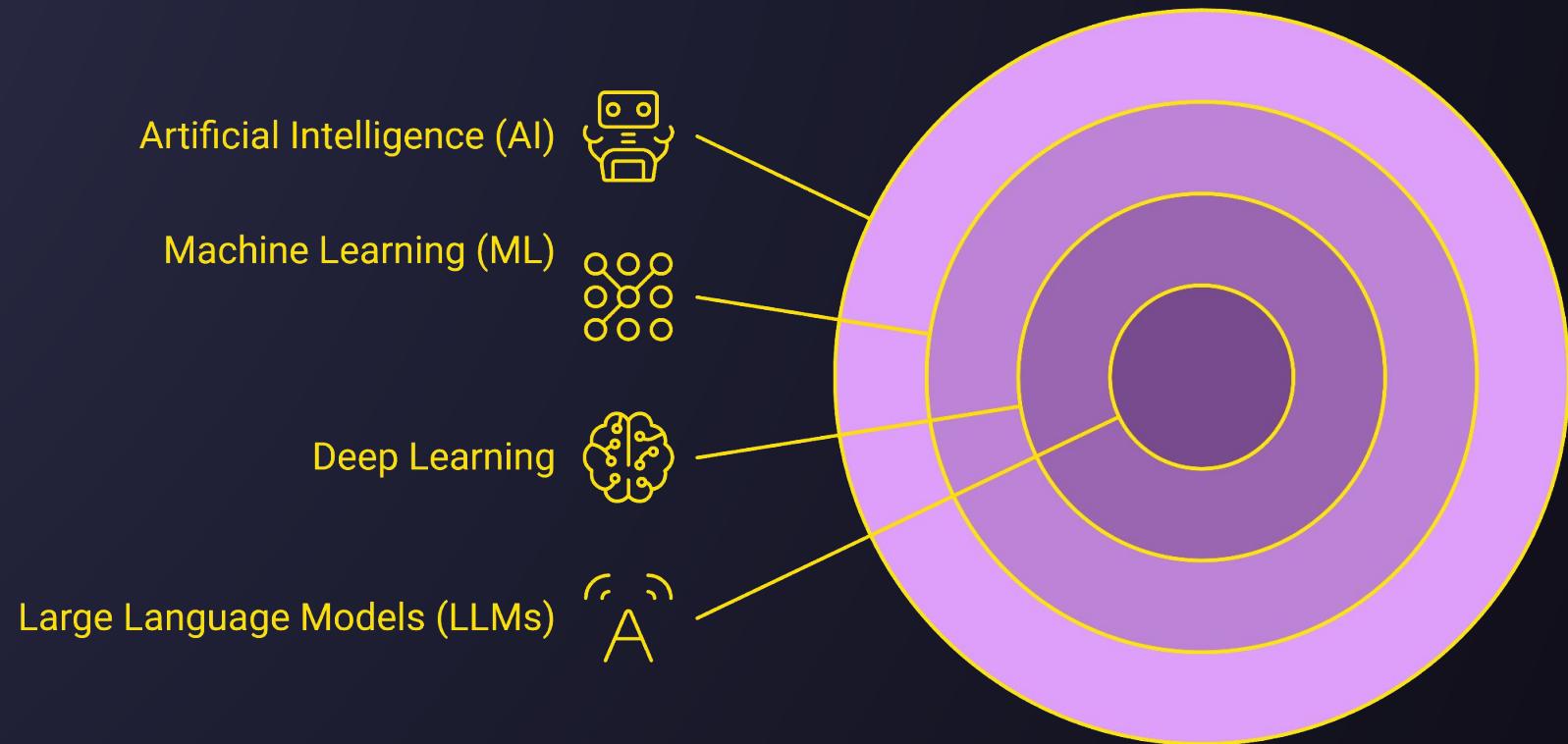
# Introduction to LLMs

# What are Large Language Models?

## Large Language Models

refer to large,  
general-purpose language  
models that can be  
pre-trained and fine-tuned  
for specific purposes

- Large training dataset
- Large number of parameters



# What are Large Language Models?

An LLM represents a ***probability distribution*** over sequences of tokens (words or subwords). Given a vocabulary, the model **assigns a probability to each sequence of tokens** it encounters. For example:

$$p(\text{the, mouse, ate, the, cheese}) = 0.02$$

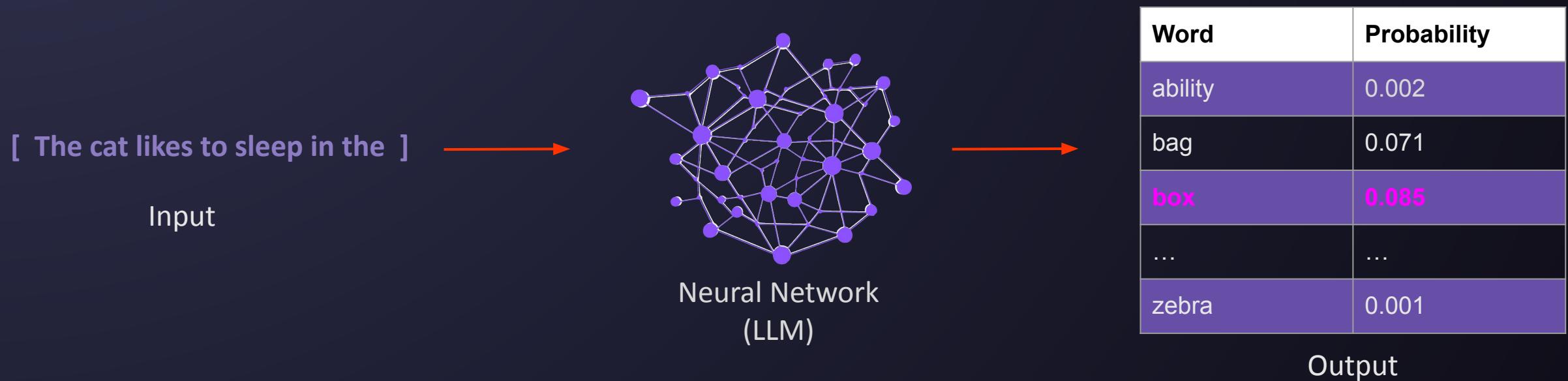
$$p(\text{the, cheese, ate, the, mouse}) = 0.01$$

$$p(\text{mouse, the, the, cheese, ate}) = 0.0001$$

# What are Large Language Models?

A **language model** is a machine learning model that aims to predict and generate plausible language.

What is the next word in a sequence of words?

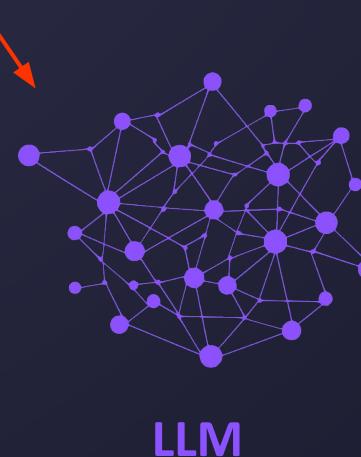


# What are Large Language Models?

**After training:** We can **generate text** by predicting **one word at a time**

[ A trained language model can ]

Input



Word	Probability
speak	0.065
<b>generate</b>	<b>0.072</b>
politics	0.001
...	...
walk	0.003



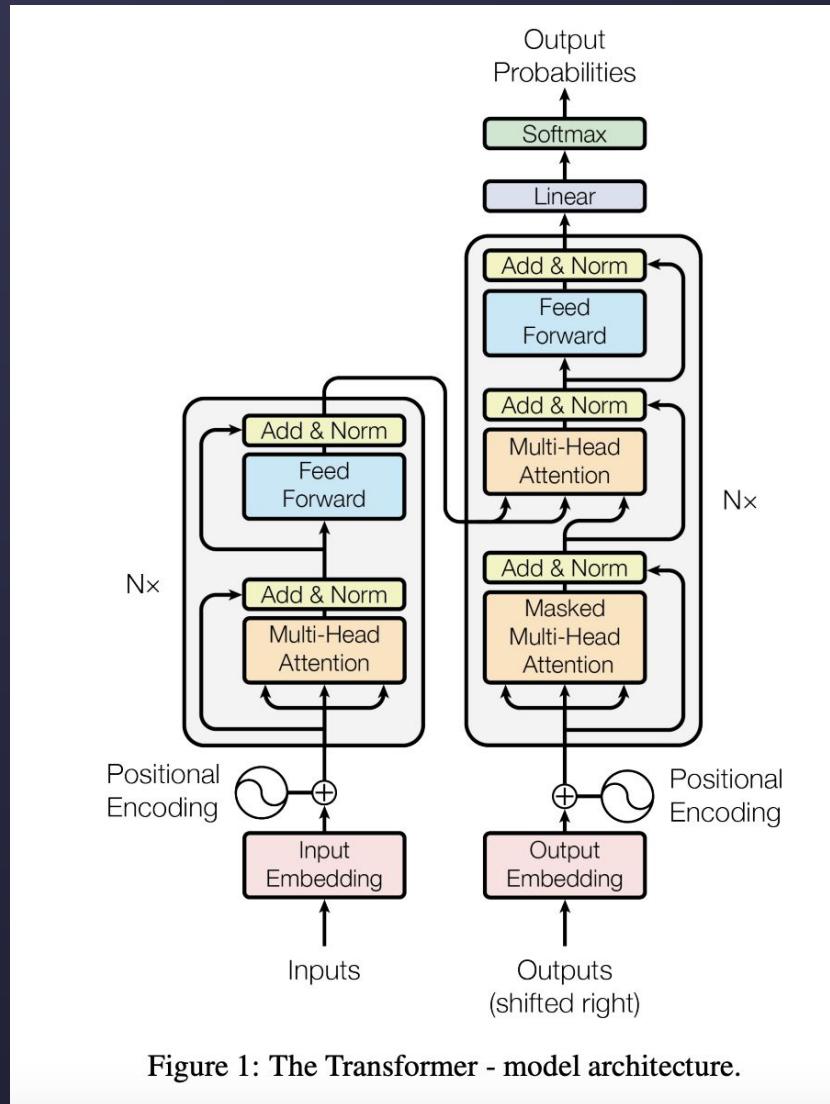
Word	Probability
ability	0.002
text	0.084
<b>coherent</b>	<b>0.085</b>
...	...
ideas	0.041

LLMs are an example of what's called "Generative AI"

Output at Step 1

Output at Step 2

# A Brief Note on Transformers



- **Attention:** Focuses on key words in a sentence.
- **Self-Attention:** Understands word relationships within the entire sentence.
- **Positional Encoding:** Keeps track of word order and structure.

Figure 1: The Transformer - model architecture.

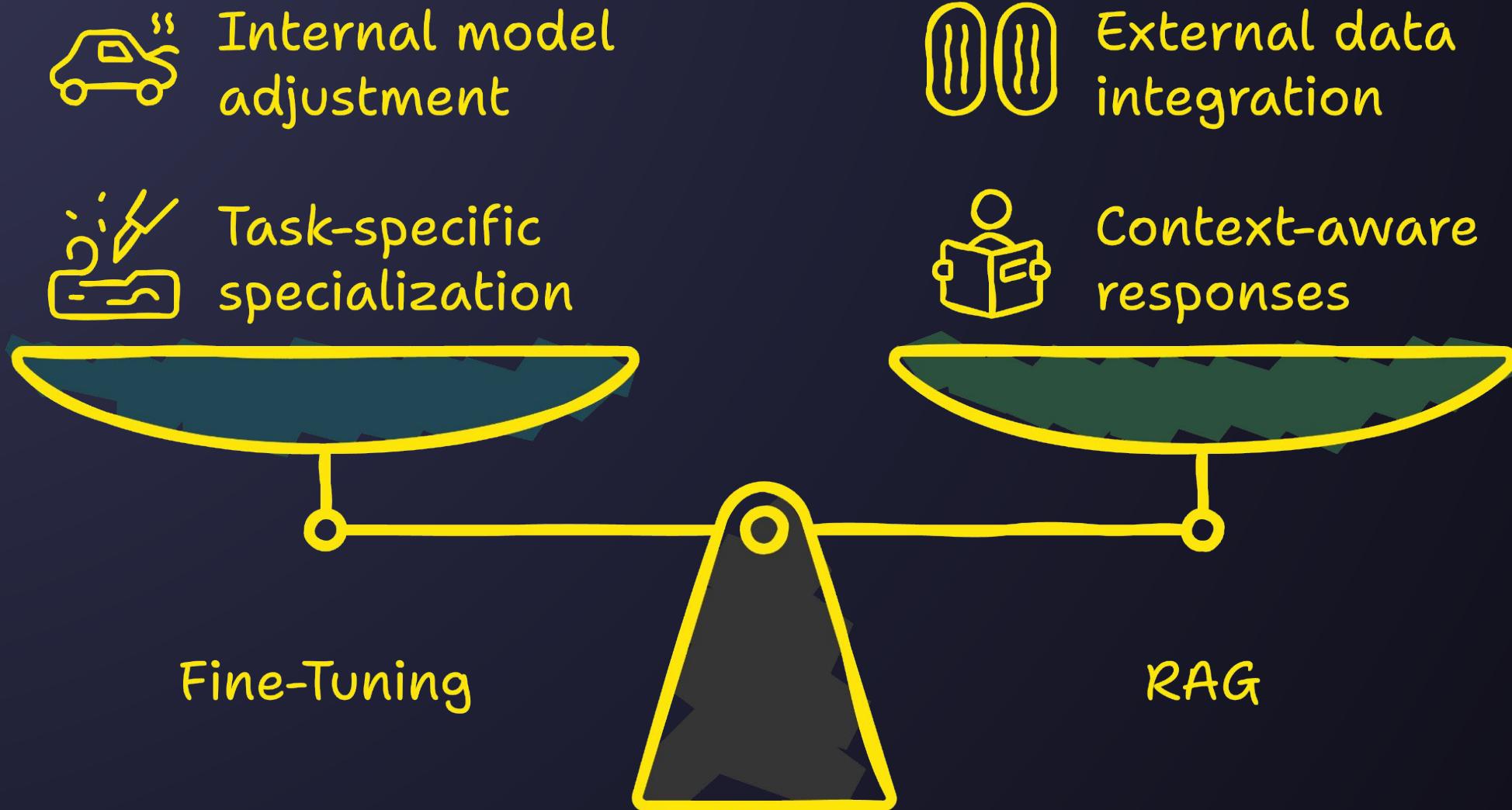
# What are Large Language Models?

```
1 import torch
2 from transformers import AutoTokenizer, GPTNeoForCausallM
3
4 # Load the tokenizer and model
5 tokenizer = AutoTokenizer.from_pretrained("EleutherAI/gpt-neo-1.3B")
6 model = GPTNeoForCausallM.from_pretrained("EleutherAI/gpt-neo-1.3B")
7
8 # Tokenize the input text
9 input_ids = tokenizer("Language models are", return_tensors="pt")
10
```

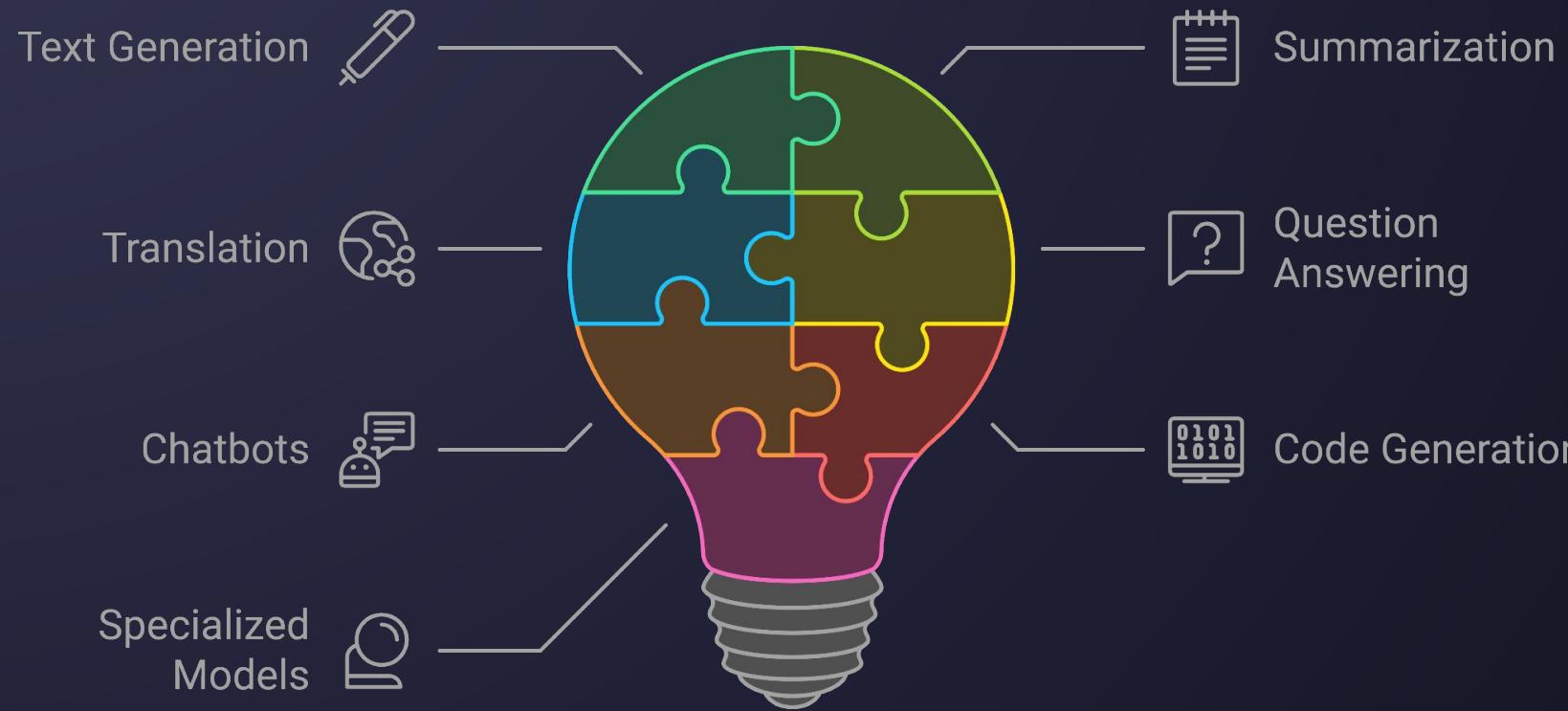
We can also [re-create this in Python](#), and using the LLM of our preference.

Token: a, Score: -1.1022074222564697
Token: used, Score: -1.4315834045410156
Token: the, Score: -1.7675762176513672
Token: often, Score: -1.841552972793579
Token: an, Score: -2.465235471725464
Token: widely, Score: -2.657839775085449
Token: not, Score: -2.6726551055908203
Token: increasingly, Score: -2.756849765777588
Token: ubiquitous, Score: -2.868807792663574
Token: important, Score: -2.902836322784424
Token: one, Score: -2.908351182937622
Token: defined, Score: -3.08156681060791
Token: being, Score: -3.2117607593536377
Token: commonly, Score: -3.311004161834717
Token: very, Score: -3.3173515796661377
Token: typically, Score: -3.4478561878204346
Token: complex, Score: -3.521364212036133
Token: powerful, Score: -3.533857822418213
Token: language, Score: -3.550966739654541
Token: pervasive, Score: -3.563505172729492

# A Brief Note on Finetuning vs RAG



# Applications of LLMs



**Discussion question:**  
How have you used  
LLMs in your own life?

# Strengths and Limitations of LLMs

Language  
Understanding and  
Generation



Multi-lingual  
capabilities without  
too much fine-tuning



Robust performance  
across various fields



Strengths



Sycophancy

Reasoning gaps

Hallucinations

Limitations

# LLM Prompting Techniques



# Zero-Shot Prompting

- Asking an AI model to perform a task without providing any examples or specific training for that task.
- Relies on the model's pre-existing knowledge and generalization abilities.

Claude ❁ Translating "The cat is on the mat" to French ▾

A Translate the following sentence to French: 'The cat is on the mat.'

Le chat est sur le tapis.

 Copy  Retry 

Claude can make mistakes. Please double-check responses.



A Generate a one-sentence marketing slogan for a new coffee shop in Makati.

"Sip the spirit of Makati in every brew."

This slogan concisely captures the essence of a local coffee shop while tying it to the vibrant city of Makati. It suggests that the coffee embodies the energy and character of the area, appealing to both locals and visitors who want to experience something uniquely tied to the location. Would you like me to explain the thinking behind this slogan or generate alternative options?

 Copy  Retry 

Claude can make mistakes. Please double-check responses.



# Few-Shot Prompting

- Providing the AI model with a small number of examples before asking it to perform a similar task.
- Helps guide the model's response by demonstrating the desired output format or style.

Q Artisanal Bakery Slogans ▾

A "Our coffee is brewed with passion!"  
"Experience the taste of quality."  
---  
Generate a similar slogan for an artisanal bakery in Quezon City.

Here are two similar slogans for an artisanal bakery in Quezon City:  
"Our bread is kneaded with love!"  
"Savor the flavor of homemade perfection."

These slogans aim to evoke the craftsmanship, care, and quality associated with artisanal baking, while also appealing to the senses and emotions of potential customers. Would you like me to generate more options or explain the thought process behind these slogans?

Copy Retry ⌂ ⌂



Claude can make mistakes. Please double-check responses.

# Chain-of-Thought Prompting

- Encouraging the AI model to break down complex problems into smaller steps and explain its reasoning process.
- Improves performance on multi-step problems and provides transparency in the model's decision-making.

A screenshot of a conversational interface showing a user prompt and a generated response from an AI model.

**User Prompt:** Solve the equation.  $34 + 44 + 3 * 23 / 3 * 2$ . Think step by step.

**AI Response:**

Certainly! Let's solve this equation step by step, keeping in mind the order of operations (PEMDAS: Parentheses, Exponents, Multiplication/Division from left to right, Addition/Subtraction from left to right).

Given equation:  $34 + 44 + 3 * 23 / 3 * 2$

Step 1: Solve the multiplication and division from left to right

$3 * 23 = 69$

$69 / 3 = 23$

$23 * 2 = 46$

Now our equation looks like this:  $34 + 44 + 46$

Step 2: Perform the additions from left to right

$34 + 44 = 78$

$78 + 46 = 124$

Therefore, the final answer is 124.

Would you like me to explain any part of this solution in more detail?

**Bottom Bar:** Claude can make mistakes. Please double-check responses.

**Buttons:** Copy, Retry, Like, Dislike

**Subscription Information:** Subscribe to Pro for 5x more usage with Claude 3.5 Sonnet. | [Subscribe to Pro](#)

# Chain-of-Thought Prompting

A What are the steps to create a business expansion plan? Start by considering market analysis, followed by financial planning and risk assessment.

Certainly. Here's a step-by-step guide to creating a business expansion plan, starting with the elements you mentioned:

## 1. Market Analysis

- Research target markets
- Analyze competitors
- Identify customer needs and preferences
- Assess market trends and growth potential

## 2. Financial Planning

- Estimate expansion costs
- Project revenue and profitability
- Analyze cash flow requirements
- Determine funding sources (e.g., loans, investors)

## 3. Risk Assessment

- Identify potential risks and challenges
- Develop mitigation strategies
- Conduct SWOT analysis (Strengths, Weaknesses, Opportunities, Threats)

## 4. Define Expansion Goals

- Set clear, measurable objectives
- Establish timelines for achieving goals

## 4. Define Expansion Goals

- Set clear, measurable objectives
- Establish timelines for achieving goals

## 5. Operational Strategy

- Plan for increased production or service capacity
- Determine staffing needs
- Develop logistics and supply chain strategies

## 6. Marketing Strategy

- Create marketing plans for new markets or products
- Budget for marketing and advertising

## 7. Legal and Regulatory Considerations

- Research relevant laws and regulations in new markets
- Plan for necessary licenses or permits

## 8. Technology and Infrastructure

- Assess IT needs for expansion
- Plan for scaling systems and processes

## 9. Human Resources Planning

- Develop hiring and training strategies
- Plan for organizational structure changes

## 10. Implementation Timeline

- Create a detailed schedule for expansion activities
- Set milestones and checkpoints

## 11. Monitoring and Evaluation

- Establish key performance indicators (KPIs)
- Develop systems for tracking progress and adjusting plans

# Prompt Engineering Best Practices

1. Write clear instructions
2. Provide reference text
3. Split complex tasks into simpler subtasks
4. Give the model time to “think”
5. Use external tools
6. Test changes systematically
7. \*Note that there will be variations in prompt engineering across different LLMs

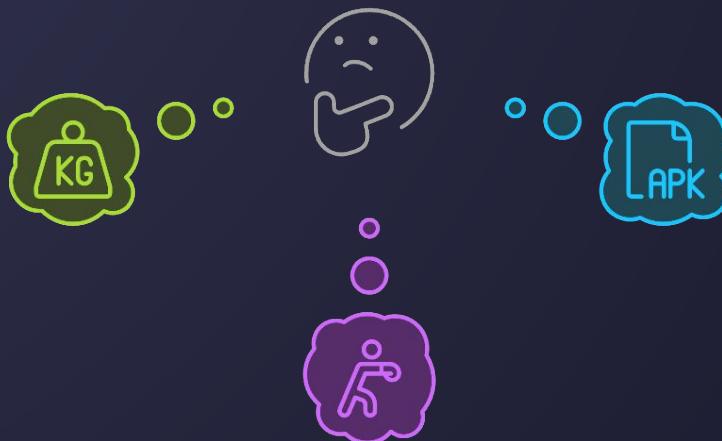


# Open-Source vs Closed-Source LLMs

# Open-Source LLMs

## Model Weights

includes all the parameters of the model and the model configuration



## Training Data

includes training data used for the model, code on how it was sourced\*, data at different stages of transformation, the order in which the data was fed to the model

## Model Code

includes model training code, hyperparameter settings, and code used for pre-processing training data



GPT-Neo



Gemma



BLOOM



LLaMA



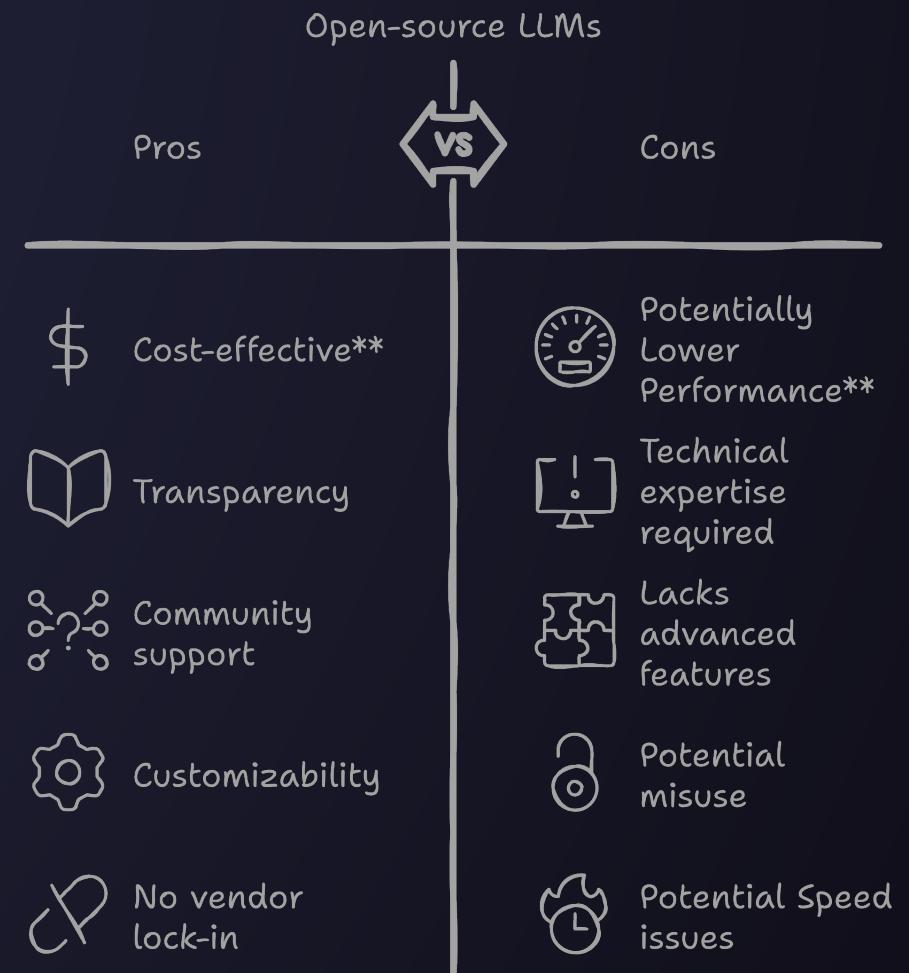
Falcon



Dolphin

# Open-Source LLMs

```
1 # Install the necessary libraries
2 !pip install transformers
3
4 # Import the T5 model and tokenizer
5 from transformers import T5ForConditionalGeneration, T5Tokenizer
6
7 # Load the T5 model and tokenizer (T5-small for faster execution)
8 model_name = "t5-small" # You can also use "t5-base" for a bigger model
9 model = T5ForConditionalGeneration.from_pretrained(model_name)
10 tokenizer = T5Tokenizer.from_pretrained(model_name)
11
12 # Set the translation prompt (T5 expects prompts in the format 'translate English to French: ...')
13 prompt = "translate English to French: Hello, how are you?"
14
15 # Tokenize the input prompt
16 input_ids = tokenizer.encode(prompt, return_tensors="pt")
17
18 # Generate the translation
19 output = model.generate(input_ids, max_length=50)
20
21 # Decode the output to text
22 response = tokenizer.decode(output[0], skip_special_tokens=True)
23
24 # Print the translated text
25 print(response)
26
```



\*\*it depends

# Proprietary LLMs

Closed Source LLMs	
Pros	Cons
 Powerful and refined	 Higher costs
 Regular updates	 Limited customization
 Well-documented APIs**	 Dependency on provider
 Managed infrastructure	 Data privacy concerns
 Built-in safety**	 Less transparency

\*\*it depends

```
10  from openai import OpenAI
11  client = OpenAI(api_key=api_key)
12
13  completion = client.chat.completions.create(
14      model="gpt-4o-mini",
15      messages=[
16          {"role": "system", "content": "You are a helpful assistant."},
17          {
18              "role": "user",
19              "content": "Translate the following from english to french: 'hello, how are you?' "
20          }
21      ]
22  )
23
24
25  print(completion.choices[0].message)
26  response
```

ChatCompletionMessage(content='The translation of "hello, how are you?" in French is "Bonjour, comment ça va ?"', refusal=None, role='assistant', 'Bonjour, comment êtes-vous?')

'Bonjour, comment êtes-vous?'

# Proprietary LLMs



Gemini ▾

Try Gemini Advanced

See the latest updates to the Gemini Apps Privacy Hub

Write a business plan outline for an marketing agency in Tagalog. Highlight all the section titles, and make it less than 300 words.

Show drafts ▾

Outline ng Business Plan para sa isang Marketing Agency sa Tagalog

1. Executive Summary

- Buod ng negosyo at mga layunin
- Panimula sa mga produkto at serbisyo
- Pagtataya ng merkado
- Pangkalahatang estratehiya
- Pinansyal na proyeksiyon

2. Paglalarawan ng Negosyo

Business Planning for Marketing Agency in Tagalog

Write a business plan outline for an marketing agency in Tagalog. Highlight all the section titles, and make it less than 300 words.

Talakayan ng Pagsasakatayuan ng Negosyo para sa Agency ng Marketing

Pangalan ng Agency

[Pangalan ng Agency] ay isang marketing agency na nagbibigay ng mga solusyon sa marketing na nakatuon sa pagtulong sa mga negosyo upang makapag-establish ng malawak na audience at maging kilala sa kanilang industriya.

ChatGPT 4o ▾

Write a business plan outline for an marketing agency in Tagalog. Highlight all the section titles, and make it less than 300 words.

Panimula

Ang aming ahensya sa marketing ay naglalayong magbigay ng epektibong digital marketing services para sa maliliit at katamtamang laki ng mga negosyo sa Pilipinas. Ang layunin namin ay palawakin ang online presence ng aming mga kliyente, pataasin ang kanilang benta, at magtagumpay sa kanilang industriya.

Serbisyo

- Social Media Management: Pamamahala ng mga social media accounts para mapalakas ang

# Open-Source vs Proprietary?



# The Challenge of Evaluating LLMs

## HELM Leaderboard

The HELM leaderboard shows how the various models perform across different scenarios and metrics.

Model		Mean win rate	NarrativeQA - F1	NaturalQuestions (open) - F1	NaturalQuestions (closed) - F1	OpenbookQA - EM
GPT-4o (2024-05-13)	0.963	0.804	0.803	0.501	0.966	
Claude 3.5 Sonnet (20240620)	0.915	0.746	0.749	0.502	0.972	
GPT-4 (0613)	0.915	0.768	0.79	0.457	0.96	
GPT-4 Turbo (2024-04-09)	0.905	0.761	0.795	0.482	0.97	
Llama 3.1 Instruct Turbo (405B)	0.895	0.749	0.756	0.456	0.94	
Palmyra-X-004	0.861	0.773	0.754	0.457	0.926	
Llama 3.1 Instruct Turbo (70B)	0.854	0.772	0.738	0.452	0.938	
Llama 3 (70B)	0.835	0.798	0.743	0.475	0.934	

Select a group:

Core scenarios

Rank* (UB)	Model	Arena Score	95% CI	Votes	Organization	License	Knowledge Cutoff
1	o1-preview	1355	+12/-11	2991	OpenAI	Proprietary	2023/10
2	ChatGPT-4o-latest (2024-09-03)	1335	+5/-6	10213	OpenAI	Proprietary	2023/10
2	o1-mini	1324	+12/-9	3009	OpenAI	Proprietary	2023/10
4	Gemini-1.5-Pro-Exp-0827	1299	+5/-4	28229	Google	Proprietary	2023/11
4	Grok-2-08-13	1294	+4/-4	23999	xAI	Proprietary	2024/3
6	GPT-4o-2024-05-13	1285	+3/-3	90695	OpenAI	Proprietary	2023/10

T	Model	Average	IFEva
...	MaziyarPanahi/calme-2.4-rys-78b	50.26	80.11
◆	dnhkng/RYS-XLarge	44.75	79.96
...	MaziyarPanahi/calme-2.1-rys-78b	44.14	81.36
...	MaziyarPanahi/calme-2.2-rys-78b	43.92	79.86
...	MaziyarPanahi/calme-2.1-qwen2-72b	43.61	81.63
...	arcee-ai/Arcee-Nova	43.5	79.07
...	MaziyarPanahi/calme-2.2-qwen2-72b	43.4	80.08

# LLM Prompt Engineering Activity 1

1. Create a Marketing Slogan for our AI Engineering Bootcamp: Use LLMs we discussed ([GPT3](#), [Cohere](#), [Claude](#), [Gemini](#)) to generate a one-line slogan for the AI Engineering Bootcamp.
2. Develop a Marketing Plan: Include target audience, messaging, channels, and strategy.
3. Compare Outputs: Evaluate slogans and plans across LLMs for creativity, clarity, and business relevance.
4. Save your results on 1 slide and be prepared to discuss with the rest of the class :-)

# LLM Prompt Engineering Activity 2

1. Choose a Model (or multiple): [DALL-E](#) via ChatGPT, [Dall-e Mini](#), [Midjourney \(paid\)](#), [Imagen 3 via Gemini \(trial\)](#), [Flux](#), or [Stable Diffusion](#) or [any image gen model](#) of your choice
2. Generate an Image with the theme of **2024 Dalagang Filipina**
3. Analyze: Compare outputs from different models based on cultural representation and style, as well as the prompts that you used
4. Share: Present your images, discuss prompt modifications, and evaluate model strengths. Also discuss any relevant observations you had when generating image prompts versus text prompts.

# Prompt Engineering Best Practices

## Include details

Worse	Better
How do I add numbers in Excel?	How do I add up a row of dollar amounts in Excel? I want to do this automatically for a whole sheet of rows with all the totals ending up on the right in a column called "Total".
Who's president?	Who was the president of Mexico in 2021, and how frequently are elections held?
Write code to calculate the Fibonacci sequence.	Write a TypeScript function to efficiently calculate the Fibonacci sequence. Comment the code liberally to explain what each piece does and why it's written that way.
Summarize the meeting notes.	Summarize the meeting notes in a single paragraph. Then write a markdown list of the speakers and each of their key points. Finally, list the next steps or action items suggested by the speakers, if any.

# Prompt Engineering Best Practices

## Include details

<b>Worse</b>	<b>Better</b>
How do I add numbers in Excel?	How do I add up a row of dollar amounts in Excel? I want to do this automatically for a whole sheet of rows with all totals ending up on the right in a column called “Total”.
Who's president?	Who was the president of Mexico in 2021, and how frequently are elections held?
Write code to calculate the Fibonacci sequence.	Write a TypeScript function to efficiently calculate the Fibonacci sequence. Comment the code liberally to explain what each piece does and why it's written that way.
Summarize the meeting notes.	Summarize the meeting notes in a single paragraph. Then write a markdown list of the speakers and each of their key points. Finally, list the next steps or action items suggested by the speakers, if any.

# Prompt Engineering Best Practices

## Ask the model to adopt a persona

- |               |   |
|---------------|---|
| <b>SYSTEM</b> | When I ask for help to write something, you will reply with a document that contains at least one joke or playful comment in every paragraph.                       |
| <b>USER</b>   | Write a thank you note to my steel bolt vendor for getting the delivery in on time and in short notice. This made it possible for us to deliver an important order. |

# Prompt Engineering Best Practices

## Use delimiters to indicate the different parts of the input

**USER** Summarize the text delimited by triple quotes with a haiku.

"""insert text here"""

**SYSTEM** You will be provided with a pair of articles (delimited with XML tags) about the same topic. First summarize the arguments of each article. Then indicate which of them makes a better argument and explain why.

**USER** <article> insert first article here </article>

<article> insert second article here </article>

# Prompt Engineering Best Practices

## **Use delimiters to indicate the different parts of the input**

<b>SYSTEM</b>	You will be provided with a thesis abstract and a suggested title for it. The thesis title should give the reader a good idea of the topic of the thesis but should also be eye-catching. If the title does not meet these criteria, suggest 5 alternatives.
<b>USER</b>	Abstract: insert abstract here
	Title: insert title here

# Prompt Engineering Best Practices

## Specify the steps required to complete a task

SYSTEM	<p>Use the following step-by-step instructions to respond to user inputs.</p> <p>Step 1 - The user will provide you with text in triple quotes. Summarize this text in one sentence with a prefix that says "Summary: ".</p> <p>Step 2 - Translate the summary from Step 1 into Spanish, with a prefix that says "Translation: ".</p>
USER	""""insert text here""""

# Prompt Engineering Best Practices

## Provide examples

<b>SYSTEM</b>	Answer in a consistent style.
<b>USER</b>	Teach me about patience.
<b>ASSISTANT</b>	The river that carves the deepest valley flows from a modest spring; the grandest symphony originates from a single note; the most intricate tapestry begins with a solitary thread.
<b>USER</b>	Teach me about the ocean.

# Prompt Engineering Best Practices

## Specify the desired length of the output

**USER** Summarize the text delimited by triple quotes in about 50 words.

"""insert text here"""

**USER** Summarize the text delimited by triple quotes in 2 paragraphs.

"""insert text here"""

# Prompt Engineering Best Practices

## Instruct the model to answer using reference text

**SYSTEM** Use the provided articles delimited by triple quotes to answer questions. If the answer cannot be found in the articles, write "I could not find an answer."

**USER** <insert articles, each delimited by triple quotes>

Question: <insert question here>

# Prompt Engineering Best Practices

## Use intent classification

**SYSTEM** Use the provided articles delimited by triple quotes to answer questions. If the answer cannot be found in the articles, write "I could not find an answer."

**USER** <insert articles, each delimited by triple quotes>

Question: <insert question here>

# Prompt Engineering Best Practices

## Give the model time to think

<b>SYSTEM</b>	First work out your own solution to the problem. Then compare your solution to the student's solution and evaluate if the student's solution is correct or not. Don't decide if the student's solution is correct until you have done the problem yourself.
<b>USER</b>	<p>Problem Statement: I'm building a solar power installation and I need help working out the financials.</p> <ul style="list-style-type: none"><li>- Land costs \$100 / square foot</li><li>- I can buy solar panels for \$250 / square foot</li><li>- I negotiated a contract for maintenance that will cost me a flat \$100k per year, and an additional \$10 / square foot</li></ul> <p>What is the total cost for the first year of operations as a function of the number of square feet.</p>
<b>ASSISTANT</b>	<p>Student's Solution: Let <math>x</math> be the size of the installation in square feet.</p> <ol style="list-style-type: none"><li>1. Land cost: <math>100x</math></li><li>2. Solar panel cost: <math>250x</math></li><li>3. Maintenance cost: <math>100,000 + 100x</math></li></ol> <p>Total cost: <math>100x + 250x + 100,000 + 100x = 450x + 100,000</math></p> <p>Let <math>x</math> be the size of the installation in square feet.</p> <ol style="list-style-type: none"><li>1. Land cost: <math>100x</math></li><li>2. Solar panel cost: <math>250x</math></li><li>3. Maintenance cost: <math>100,000 + 10x</math></li></ol> <p>Total cost: <math>100x + 250x + 100,000 + 10x = 360x + 100,000</math></p> <p>The student's solution is incorrect. They made an error in the maintenance cost calculation, using <math>100x</math> instead of <math>10x</math>. The correct total cost for the first year of operations is <math>360x + 100,000</math>.</p>

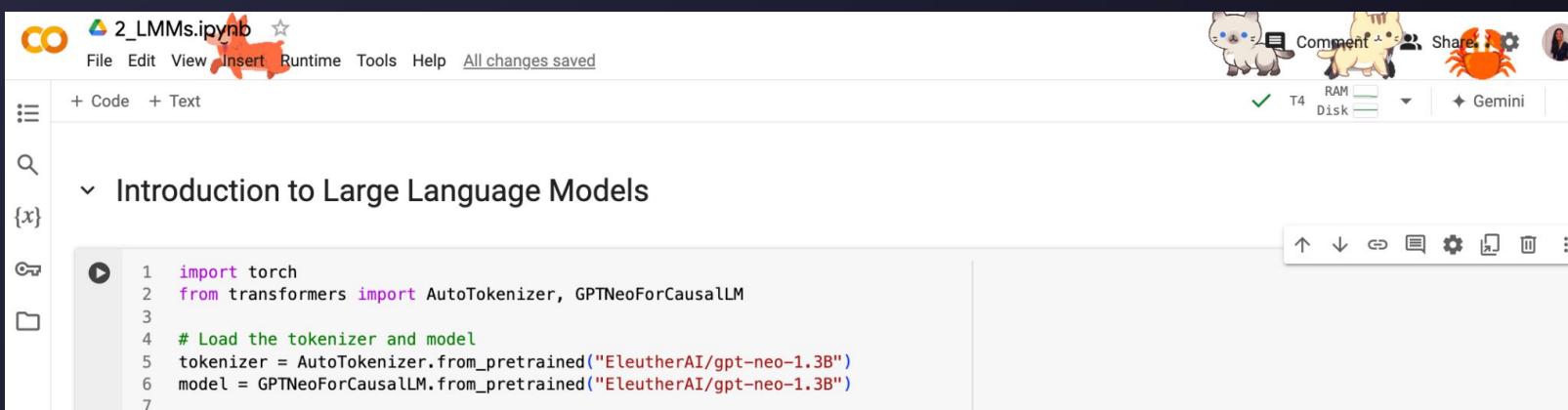


# Environment Setup

# Google Colab

- Free cloud-based development environment.
- Supports Python and machine learning libraries.
- No need for local setup – runs on Google's servers.

1. Open your browser and visit: colab.research.google.com.
2. Sign in with your Google account.
3. Start a new notebook by clicking "File" → "New Notebook".



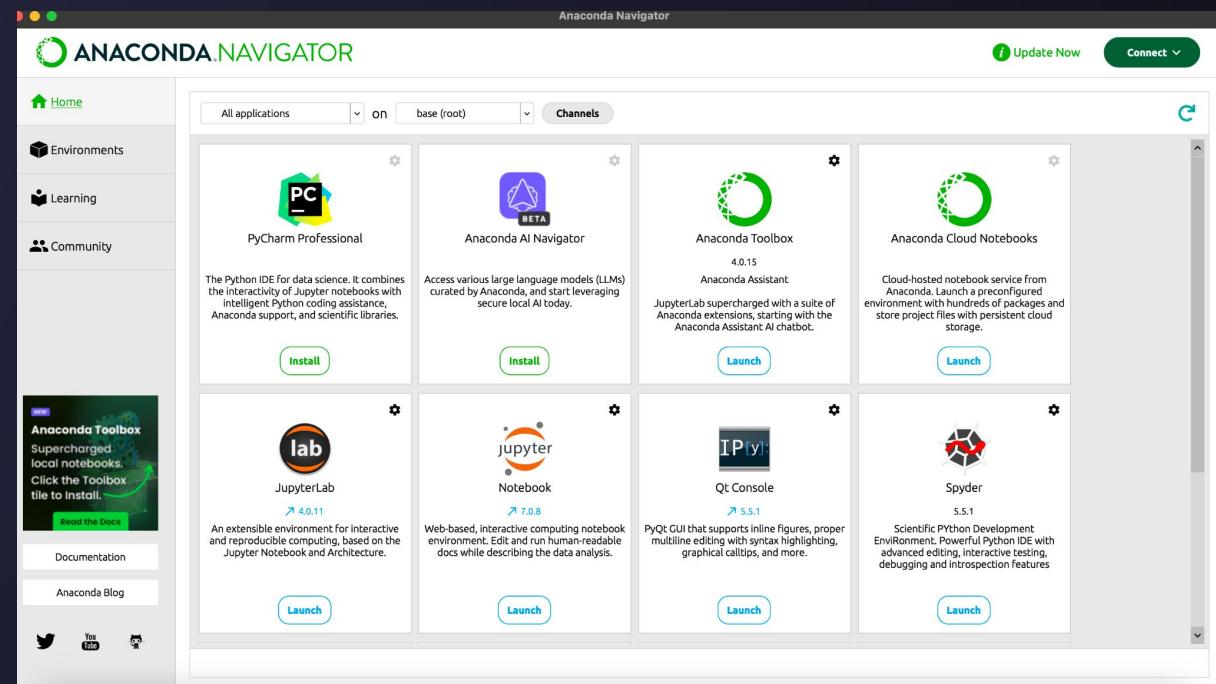
```
1 import torch
2 from transformers import AutoTokenizer, GPTNeoForCausalLM
3
4 # Load the tokenizer and model
5 tokenizer = AutoTokenizer.from_pretrained("EleutherAI/gpt-neo-1.3B")
6 model = GPTNeoForCausalLM.from_pretrained("EleutherAI/gpt-neo-1.3B")
7
```

# Anaconda

1. Open your browser and visit:

<https://www.anaconda.com/download>

2. Skip registration, download according to your computer settings



# Basic Terminal Commands

- ls: Lists all files and directories in the current directory.
- cd <directory\_name>: Changes the directory to the specified folder.
- mkdir <folder\_name>: Creates a new directory/folder.
- pwd: Displays the present working directory path.
- touch <file\_name>: Creates a new, empty file.
- rm <file\_name>: Deletes a specified file.
- pip install <package\_name>: Installs a Python package.
- python <file\_name>.py: Runs a Python script.

# Cursor.ai

## Edit in natural language

Cursor lets you write code using instructions. Update entire classes or functions with a simple prompt.

```
parallelize this because the api is slow
Submit Edit ⌘+Q quick question ⌘+G gpt:40 ⌘+K to toggle

def insert_failed_transactions_from_stripe():
    import sqlite3
    import stripe
    from datetime import datetime, timedelta

    # Set up Stripe API key
    stripe.api_key = 'your_stripe_api_key_here'

    # Connect to SQLite database (or create it if it doesn't exist)
    conn = sqlite3.connect('transactions.db')
    cursor = conn.cursor()

    # Create table if it doesn't exist
    cursor.execute('''
        CREATE TABLE IF NOT EXISTS transactions (
            id TEXT PRIMARY KEY,
            amount INTEGER,
            currency TEXT,
            status TEXT,
            created TIMESTAMP
        )
    ''')

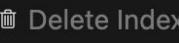
    # Insert failed transactions
    cursor.executemany('''
        INSERT INTO transactions (amount, currency, status, created)
        VALUES (?, ?, ?, ?)
    ''', failed_transactions)

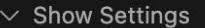
    conn.commit()
    conn.close()
```

## Codebase indexing

Embeddings improve your codebase-wide answers. Embeddings and metadata are stored in the cloud, but all code is stored locally.

Synced  100%

 Resync Index  Delete Index

 Show Settings

# The AI Code Editor

Built to make you extraordinarily productive, Cursor is the best way to code with AI.

## Just hit tab

Cursor lets you breeze through changes by predicting your next edit.

```
todo.tsx
import React, { useState } from 'react';

const TodoComponent: React.FC = () => {
  const [todos, setTodos] = useState<string>([]);
  const [input, setInput] = useState<string>('');

  return (
    <>
      <button onClick={() => {
        setTodos([...todos, input]);
        setInput('');
      }}>
        Add
      </button>
      <input
        value={input}
        onChange={(e) => setInput(e.target.value)}
      />
      <ul>
        {todos.map((todo) => (
          <li>{todo}</li>
        ))}
      </ul>
    </>
  );
}
```



# Introduction to Streamlit

# Introduction to Streamlit

## Python-based Framework

Streamlit simplifies building interactive web apps using Python, making it accessible to data scientists.

## Data-centric Applications

Ideal for showcasing data insights, building dashboards, and creating interactive applications.

## Rapid Prototyping

Quickly create interactive visualizations and dashboards with minimal code.

## User-friendly Syntax

Streamlit's intuitive syntax makes it easy to build engaging user interfaces.

# Embrace scripting

Build an app in a few lines of code with our magically API. Then see it automatically update as you iterate the source file.

MyApp.py

```
import streamlit as st  
import pandas as pd
```

```
t.write("""  
My first app  
Hello *world!*  
""")
```

```
f = pd.read_csv("my_data.csv")  
t.line_chart(df)
```

My

Hello



# Setting up Streamlit Environment

## Install Anaconda

Anaconda is a popular Python distribution that simplifies package management. It includes Streamlit, making installation easy.

## Create a Virtual Environment

Isolate your Streamlit project within a virtual environment to manage dependencies and prevent conflicts.

## Install Streamlit

Use pip within your virtual environment to install the latest version of Streamlit.

## Start a Streamlit App

Run your Streamlit app using the `streamlit run` command in your terminal.

# Hands-on Activity: Sentiment Analysis

# Sentiment Analysis with SpaCy on IMDB Dataset

- Download the IMDB Dataset from Github [here](#)
- Use what we've learned today in terms of text preprocessing to clean the dataset
- apply sentiment analysis using the SpaCy Library. Please refer to the documentation to explore spacy's capabilities.
- Does the spacy predicted sentiment match the actual rating of each movie review? (i.e. was spacy correct?)
- If you feel inclined, please feel free to deploy your code via streamlit