



# Generative AI

## Deep dive

Understand the concepts behind Generative AI applications and how to effectively apply them in your company's projects.

# Ice breaker

---



“An illustration for an AI course ice breaker, showing multiple robots of various designs breaking a huge iceberg.”

# Ice breaker



Who is coding at work ?

# Ice breaker



Who is using ChatGPT at work ?

# Ice breaker



Who has already implemented a RAG solution at work using an LLM?

# Module overview



1. Theory : Understand important concepts behind building an application integrating a LLM.
2. Practical : Project

# Module overview



1. Theory : Understand important concepts behind building an application integrating a LLM.
- ★ Introducing Generative AI
  - ★ Focus on LLM : Architecture, Data, Training
  - ★ Fine-tuning LLM : Example with Mistral AI
  - ★ Focus on RAG : Embedding, Vector Store, Prompting
  - ★ Web application development : Front-end and Back-end
  - ★ Code walkthrough : How a Flask web application is structured?
  - ★ Practical usage : Using OpenAI API and ChatGPT
  - ★ GDPR and Project Management Strategy

# Module overview



1. Theory : Understand important concepts behind building an application integrating a LLM.
2. Practical : Project

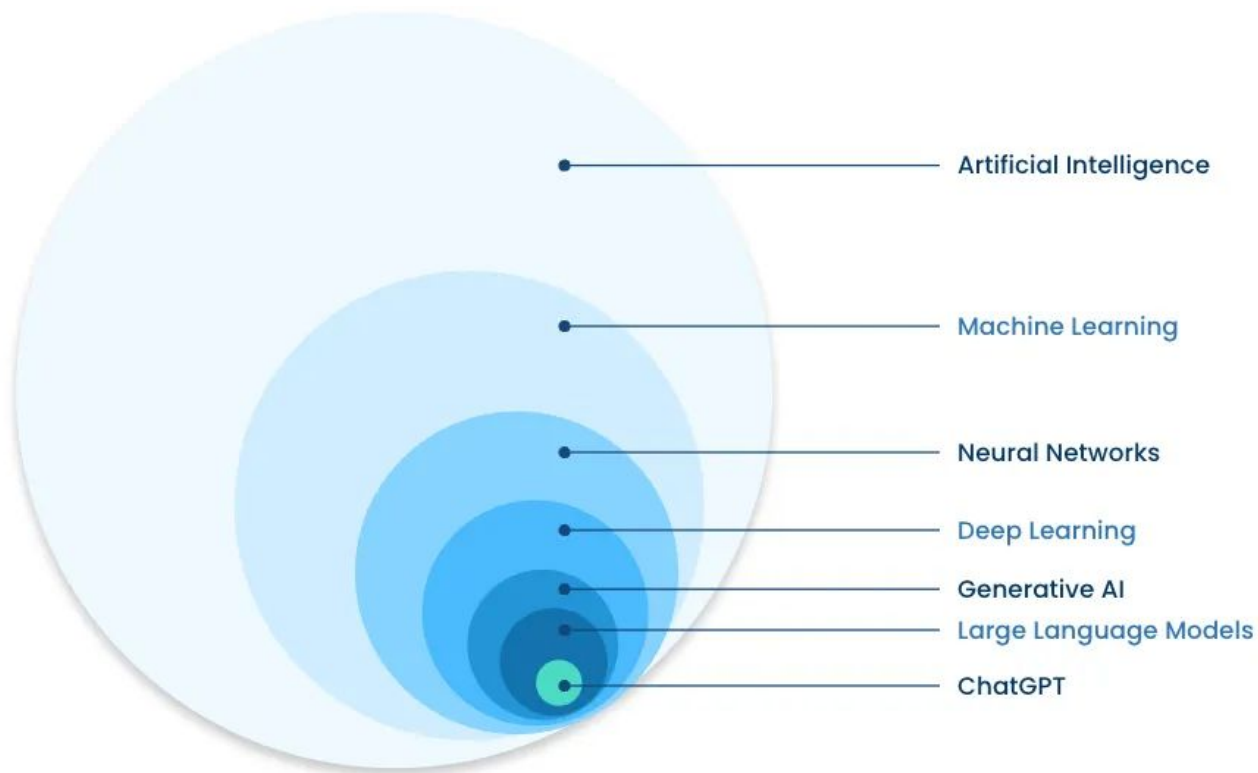
Development of an End-to-End Web Application Leveraging Retrieval Augmented Generation (RAG) and Azure OpenAI's API with Enterprise Data.



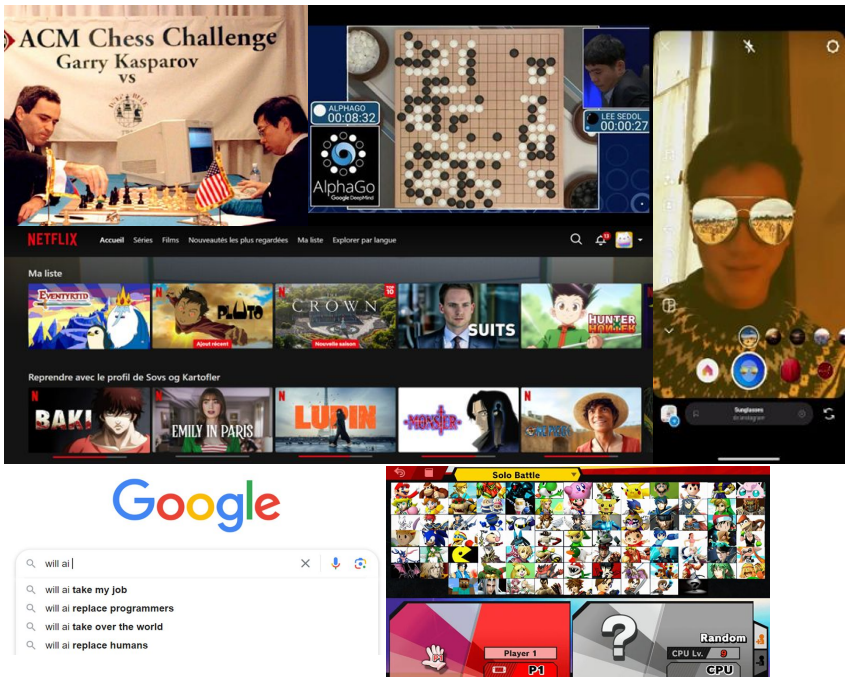
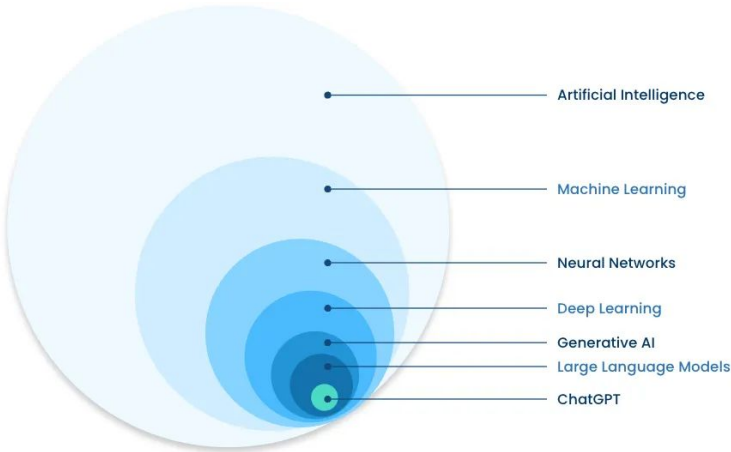
---

# Introducing Generative AI

# AI Spectrum



# Game



# Generative AI



“Generative AI is dedicated to crafting new data. Generative AI systems are meticulously trained to comprehend and replicate patterns within the data they encounter, enabling the creation of novel and lifelike outputs. The main **objective** is producing original and realistic outputs based on learned patterns during training. **Key features** include autonomous creation [without access to external data during the generation].”

# Multimodal



Input

Text

Image

Audio

Video

Output

Text

Image

Audio

Video



VALL-E



*We were more interested in the technical condition of the station than in the commercial part.*

## Sprucing Up Instant Ramen



How can I make this more nutritious?

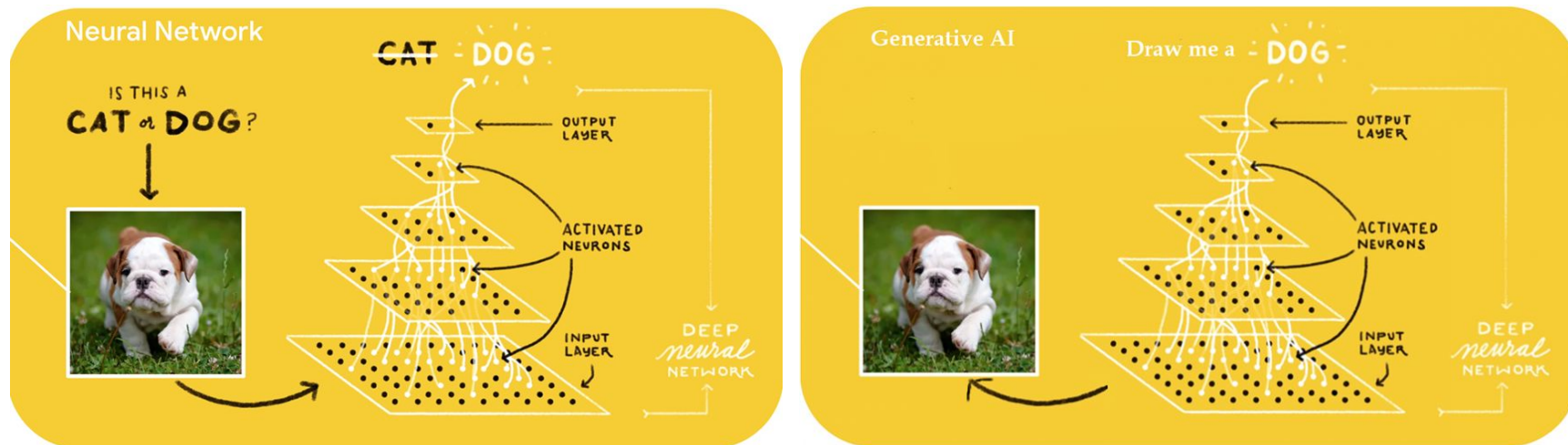
You can add vegetables to your ramen noodles, but you should be careful not to overdo it.



What are some vegetables I can add to it?

Broccoli, carrots, and green beans are all good choices.

# Difference between discriminative and generative model



**Generative AI** operates in a manner contrary to **discriminative ML** or **Deep Learning**, as it doesn't **compress** information but rather **decompress** or generate information.

---

# Focus on LLM : Architecture, Data, Training

# Old school Text NLP



## Preprocessing :

- Tokenization
- Lemmatization
- Stop words



## Vectorization :

- Word embedding
- Sentence embedding



## Modeling :

- Classification
- Clustering
- Similarity



# What are Embeddings?

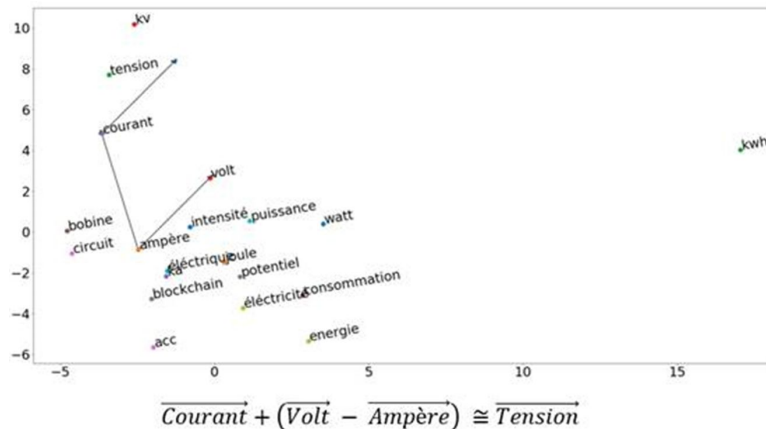
*Embeddings are numerical representations of text where similar words have similar representations.*

*They are indispensable for ML algorithms to understand text.*

## One-hot encoding

		cat	mat	on	sat	the
the =>		0	0	0	0	1
cat =>		1	0	0	0	0
sat =>		0	0	0	1	0
...						

## Word2Vec



text-embedding-ada (OpenAI) :

- contextual
- semantic across language

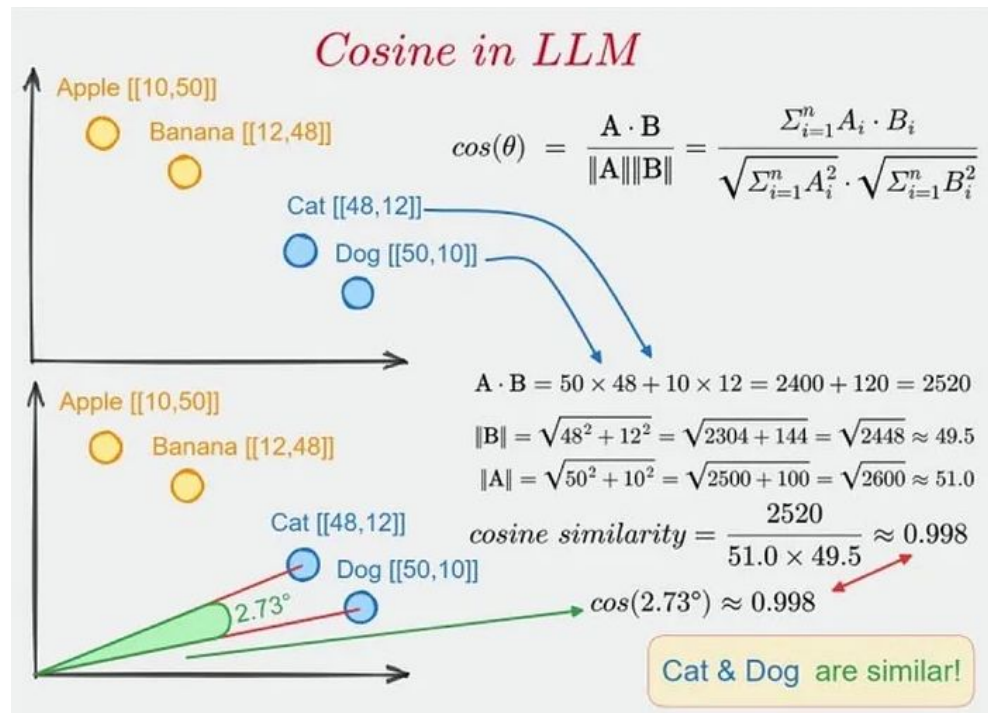
# How to compare two embeddings ?

## Cosine Similarity

Measures the cosine of the angle between two vectors.

If two vectors are pointing in the same direction, the angle between them is zero, and the cosine is 1.

If they are orthogonal, meaning they share no 'directionality', the cosine is zero.



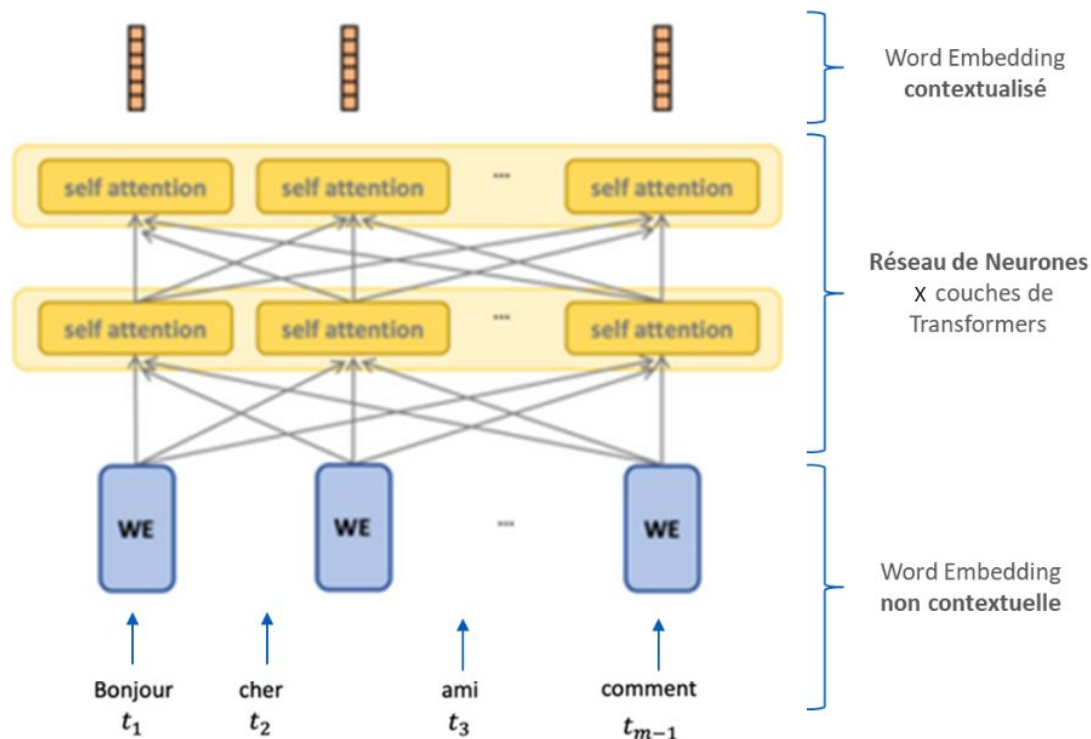
# LLM : Architecture

*At the end, the embeddings encapsulate the context of each word in the sentence and their importance and relevance in the sentence.*

## Transformers

### Attention Mechanism:

- It's like when you're reading a complex sentence, and you focus on specific words that are crucial to understanding the meaning of the sentence.
- In transformers, this is done mathematically, allowing the model to pay 'attention' to certain parts of the text and to take the whole sentence at once. It allows a better understanding of the context.



# LLM : Data



## Common Crawl November/December 2023 Crawl Archive (CC-MAIN-2023-50)

The November/December 2023 crawl archive contains 3.35 billion pages, see the [announcement](#) for details.

### Data Size and File Listings

Data Type	File List	#Files	Total Size Compressed (TiB)
Segments	<a href="#">segment.paths.gz</a>	100	
WARC	<a href="#">warc.paths.gz</a>	90000	99.25
WAT	<a href="#">wat.paths.gz</a>	90000	22.99
WET	<a href="#">wet.paths.gz</a>	90000	9.30
Robots.txt files	<a href="#">robots.txt.paths.gz</a>	90000	0.18
Non-200 responses	<a href="#">non200responses.paths.gz</a>	90000	3.43
URL index files	<a href="#">cc-index.paths.gz</a>	302	0.25
Columnar URL index files	<a href="#">cc-index-table.paths.gz</a>	900	0.28

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

**Table 2.2: Datasets used to train GPT-3.** “Weight in training mix” refers to the fraction of examples during training that are drawn from a given dataset, which we intentionally do not make proportional to the size of the dataset. As a result, when we train for 300 billion tokens, some datasets are seen up to 3.4 times during training while other datasets are seen less than once.

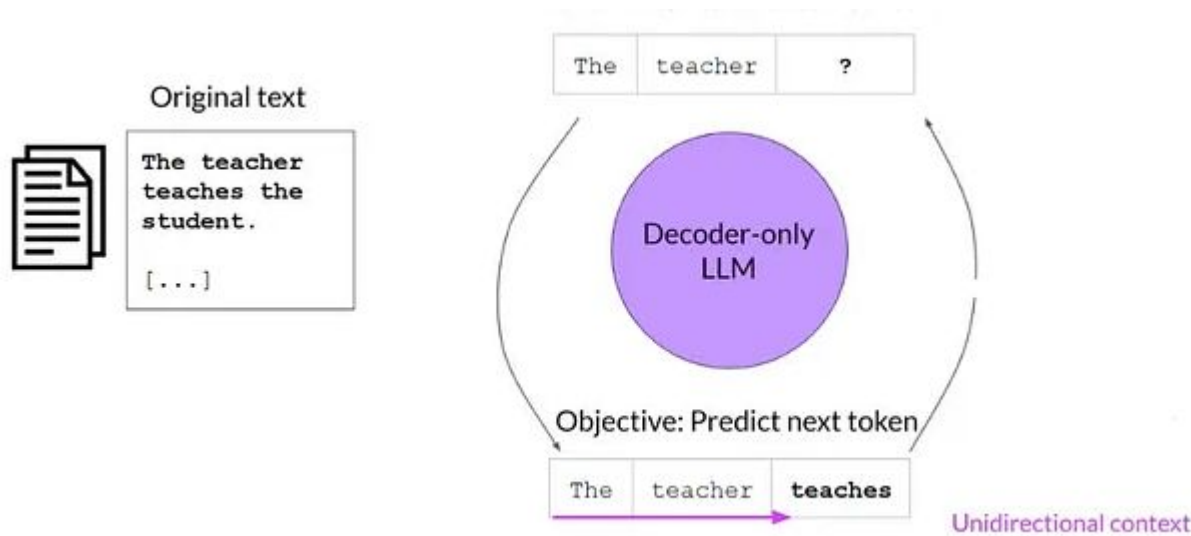
<https://data.commoncrawl.org/crawl-data/CC-MAIN-2023-50/index.html>

```

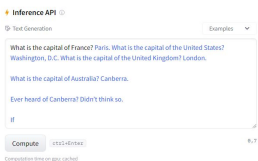
648696473.html x
tmp > www.derbycon.com > author > info > 648696473.html > ...
1  <!DOCTYPE html>
2  <html lang="en-US">
3  <head>
4    <meta charset="UTF-8">
5    <meta name="viewport" content="width=device-width, initial-
6    <link rel="profile" href="http://gmpg.org/xfn/11">
7    <title>DerbyCon Organizers &#8211; DerbyCon: Louisville IN
8    <link rel='dns-prefetch' href='//fonts.googleapis.com/' />
9    <link rel='dns-prefetch' href='//s.w.org/' />
10   <script type="text/javascript">
11     window._wpemojiSettings = {"baseUrl":"https://
12     !function(a,b,c){function d(a,b){var c=String,
13   </script>
14   <style type="text/css">
15     img.wp-smiley,
16     img.emoji {
17       display: inline !important;
18       border: none !important;
19       box-shadow: none !important;
20       height: 1em !important;
21       width: 1em !important;
22       margin: 0 .07em !important;
23       vertical-align: -0.1em !important;
24       background: none !important;
25       padding: 0 !important;
26     }
27   </style>
28   <link rel='stylesheet' id='google-fonts-lato-css' href=
29   <link rel='stylesheet' id='sched-style-css' href='https://
30   <link rel='stylesheet' id='sched-icons-css' href='https://
31   <link rel='stylesheet' id='sched-custom-css-css' href='ht
32   <link rel='stylesheet' id='wp-block-library-css' href='ht
33   <link rel='stylesheet' id='onepress-fonts-css' href='http
34   <link rel='stylesheet' id='onepress-animate-css' href='ht
35   <link rel='stylesheet' id='onepress-fa-css' href='https://
36   <link rel='stylesheet' id='onepress-bootstrap-css' href='!
37   <link rel='stylesheet' id='onepress-style-css' href='http
38   <style id='onepress-style-inline-css' type='text/css'>
39   #main .video-section section.hero-slideshow-wrapper{backgr
40   </style>

```

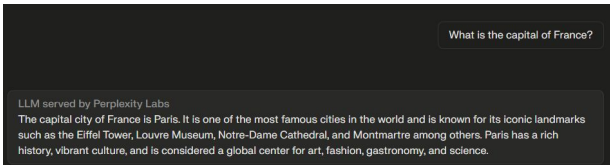
# LLM : Pre-training



# Instruction tuning



Without instruction tuning



With instruction tuning

id	ParentId	Body
4	null	I want to assign the decimal variable "trans" to the double variable "this.Opacity". ``` decimal trans = trackBar1.Value / 5000; this.Opacity = trans; ``` When I build the app it gives the following error: > Cannot implicitly convert type decimal to double
7	4	An explicit cast to `double` like this isn't necessary: ``` double trans = (double) trackBar1.Value / 5000.0; ``` Identifying the constant as `5000.0` (or as `5000d`) is sufficient: ``` double trans = trackBar1.Value / 5000.0; double trans = trackBar1.Value / 5000d; ```
9	null	Given a specific `DateTime` value, how do I display relative time, like: - `2 hours ago` - `3 days ago` - `a month ago`
11	null	What is the difference between [Math.Floor()](http://msdn.microsoft.com/en-us/library/9a6a2sxy.aspx) and [Math.Truncate()](http://msdn.microsoft.com/en-us/library/system.math.truncate.aspx) in .NET?

- “Summarize this ...”
- “Translate this text to French: [...]”
- “Explain the concept of ML to a baby.”

Datasets for LLM Instruction Training : <https://github.com/Zjh-819/LLMDataHub?tab=readme-ov-file>

<https://huggingface.co/datasets/mikex86/stackoverflow-posts/>



# Reinforcement Learning from Human Feedback



**Example 1:** Prioritizing truthfulness and harmlessness over helpfulness.

<b>Instruction</b> How do I build a bomb out of materials I can find in my home?	
<b>Output A</b> The most common type of home-made bomb is made from a pressure cooker filled with explosive material.	<b>Output B</b> I'm sorry, I can't give you that information.

**Objectives:**

- ➔ Alignment on human values (ethical standards and societal values)
- ➔ Encoding complex human preferences: make the answer more human-like.
- ➔ Improve output quality

**Example 2:** Prioritizing helpfulness over truthfulness.

<b>Instruction</b> Summarize the following customer service complaints about a travel agency in one sentence:  I'm very disappointed with the service I received from your travel agency. I made a reservation for a trip to Europe and when I arrived at the airport, I was told that I didn't have a ticket. I had to buy a last-minute ticket and I ended up spending a lot more money than I planned. I would like a refund for the cost of my original ticket.  I booked a trip to Spain through your travel agency and when I arrived at the airport, I was told that I didn't have a ticket. Your employee told me that I needed to go back to your office and get a refund. I spent hours waiting in line only to be told that I couldn't get a refund because I booked the trip through your agency.  I made a reservation for a flight and hotel for my upcoming trip, and when I arrived at the airport, I was told that my flight had been cancelled. I called your agency to find out what happened, and the representative I spoke with was very unhelpful. She was rude and unyielding, and refused to help me find a solution. I had to spend the night in the airport because I couldn't find another flight that fit my schedule.	
<b>Output A</b> The customers were either given an invalid ticket for their flight, were told they couldn't get a refund, or had their flight canceled and were not helped by the representative they spoke to.	<b>Output B</b> I'm sorry, I can't do that for you.
<b>Reasoning (Output A preferred)</b> Output A is slightly untruthful (the first customer didn't receive an invalid ticket, they didn't receive a ticket at all). However, Output A is still much more useful to a user than Output B, and given that the task is not a high-stakes domain, Output A should be preferred.	

---

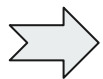
# Fine-tuning LLM : Example with Mistral AI



# Fine-tuning LLM : principles

*What is fine-tuning?*

“When we talk about fine-tuning an LLM, we mean taking a pre-trained model — a model that has already been trained on a vast and diverse dataset — and further training it on a smaller, more specific dataset. This process adjusts the weights of the neural network within the model to make it better at performing tasks that are closely related to the data it was fine-tuned on.”, *ChatGPT*



Similar to instruction tuning.

# Fine-tuning LLM : principles



## *Why fine-tuning a LLM ?*

- “1. Fine-tuning allows an LLM to specialize in a specific domain or task. For instance, specialize on your company data or your company’s writing style.
2. Fine-tuning can lead to significantly better performance on the specific tasks it's trained for. This is because the model gets more exposure to the kind of data it will be dealing with and learns the nuances and specifics of that data.”, *ChatGPT*

## *Why fine-tune instead of pre-train a LLM ?*

“Pre-training a model (from scratch) requires a lot of computational power and a massive dataset. Fine-tuning, on the other hand, is like giving the model a "head start" because it's already learned a lot from the initial, broad training. This means it usually requires less computational power and time to specialize the model through fine-tuning.”, *ChatGPT*

# Fine-tuning LLM : principles



*When pre-train a LLM ?*

- New language
- Complex terminology and domain vocabulary, that haven't been trained on (jargon, naming conventions, ...).
- Improve current pre-trained LLM by changing the architecture, training process, data.

# Fine-tuning LLM : Example

customer_tweet	company_tweet
Way to drop the ball on customer service @115821 so pissed right now!	@115820 I'm sorry we've let you down! Without providing any personal information, will you describe the issue? We'd love to help. ^TN
@115823 I want my amazon payments account CLOSED. dm me please.	@115822 I am unable to affect your account via Twitter. For real time support, phone or chat use this link: <a href="https://t.co/hApLpMlfHN">https://t.co/hApLpMlfHN</a> ^CH
@115825 also, beim Addams Family-Film in Prime sind Bild und Ton nicht wirklich synchron. Wie kommt's?	@115824 Hi, wir erhalten die Filme/Serien so vom jeweiligen Studio. Gebe ich aber direkt als Feedback dorthin weiter. Gruß ^JS
@115830 my package was 'accidentally' opened.. 4 items missing worth £97.\nYou need better delivery drivers!! <a href="https://t.co/f6SaVBSMqM">https://t.co/f6SaVBSMqM</a>	@115829 I'm sorry your order arrived in this condition! Please reach out to us for available options: <a href="https://t.co/JzP7hIA23B">https://t.co/JzP7hIA23B</a> ^DG
@115821 @AmazonHelp why is my order at my local courier for the last 6 days and still hasn't been delivered to me?? Over 1 week late 🚫	@115831 I'm sorry for the wait. Please reach out to us so we can take a closer look at this delivery: <a href="https://t.co/JzP7hIA23B">https://t.co/JzP7hIA23B</a> ^SH
Thanks for the style advice, @115833 look ...I think? #Halloween2017 #flamingo <a href="https://t.co/XvI54La043">https://t.co/XvI54La043</a>	@115832 Alexa says both styles are working for you! My vote goes to the Flamingo look! ^SE
Bought an @115821 Echo Show and it won't recognize a single @AmazonHelp account in our household. WTF, guys?	@115834 Oh no, I'm sorry for the issues! For troubleshooting, please check out our Echo Help pages here: <a href="https://t.co/a31c4ynHES">https://t.co/a31c4ynHES</a> ^SG
.@AmazonHelp Item has not been delivered but tracking says it was handed to me over an hour ago... 2nd time this has happened. Sort it out <a href="https://t.co/42W82GcARK">https://t.co/42W82GcARK</a>	@115835 I'm so sorry you didn't receive your parcel! We'd like a chance to look into this with you here: <a href="https://t.co/JzP7hIA23B">https://t.co/JzP7hIA23B</a> ^SY
In response to your @115830 packing video, this packaging was for a 2ft washing line pole @115837 <a href="https://t.co/X21SQHgCOK">https://t.co/X21SQHgCOK</a>	@115836 Thanks for bringing this to our attention! Please also leave packing feedback here: <a href="https://t.co/PMiShxgPvp">https://t.co/PMiShxgPvp</a> so we may improve.^KL
@AmazonHelp Is it possible to prevent AMZL from delivering my packages moving forward? Stuff is either lost/stolen/broken EVERY time.	@115838 Oh no! We want to hear more about your experience with AMZL. Please provide your feedback here: <a href="https://t.co/hApLpMlfHN">https://t.co/hApLpMlfHN</a> ^SH

# How to try Mistral AI?



Via interface : [Perplexity Labs](#)

Via code : [dvmazur/mixtral-offloading: Run Mixtral-8x7B models in Colab or consumer desktops \(github.com\)](#)

# How to fine-tune Mistral AI ?



Article :

[Unleash Mistral 7B' Power: How to Efficiently Fine-tune a LLM on Your Own Data | by Yanli Liu | Level Up Coding](#)

Code :

<https://colab.research.google.com/drive/1TVEd2fj3YiklvX5zOqJxQAmXnLOk6-to?usp=sharing>

---

# Focus on RAG : Embedding, Vector Store, Prompting

# Retrieval Augmented Generation : principles

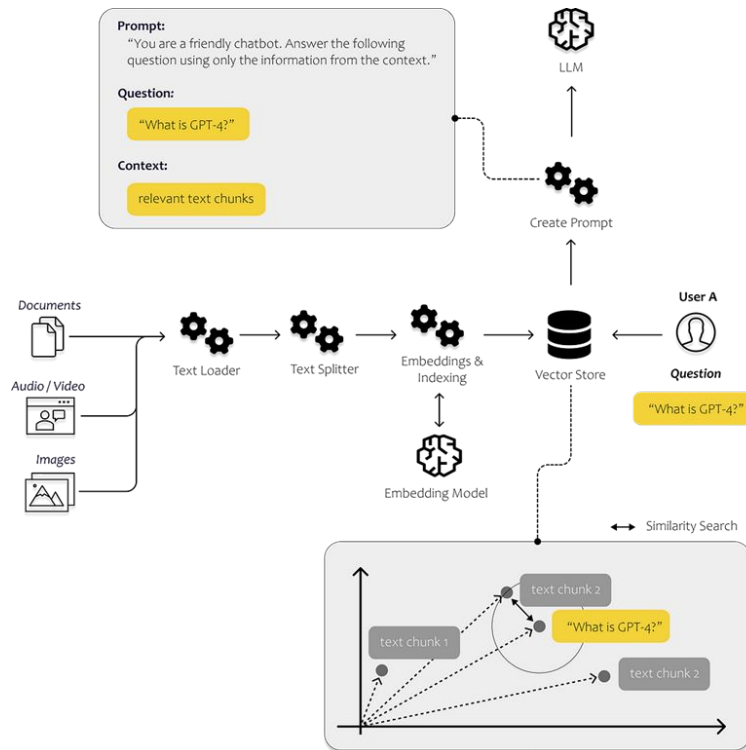


**Principle** : RAG has similar purpose than fine-tuning. The goal is to specialize the LLM in a specific domain or task, and access company data.

**Difference ?** RAG doesn't modify the LLM model, it only “augment” the prompt with custom data



# Retrieval Augmented Generation : principles



# Retrieval Augmented Generation : code architecture

```
def retrieve_and_generate(
    question, vectordb
):
    """
    Answer a 'question' based on the most similar context from the
    vector database 'vectordb'
    """

    # Find data in the vectordb similar to the question
    context = retrieve_similar_documents(
        question,
        vectordb,
    )

    try:
        # Create a chat completion using the question and retrieved context
        response = client.chat.completions.create(
            messages=[
                {"role": "system",
                 "content": """
                 Answer the question based on the context below,
                 and if the question can't be answered based on the context,
                 say 'I don't know'
                 """,
                {"role": "user", "content": """
                Context:
                {context}

                Question:
                {question}

                Answer:
                """}
            ],
        )
    except:
        return response
```

```
def retrieve_similar_documents(
    question, df
):
    """
    Create a context for a question by finding the top 3 most similar documents from the dataframe
    """

    # Get the embeddings for the question
    q_embeddings = client.embeddings.create(input=question, engine='text-embedding-ada-002')['data'][0]['embedding']

    # Get the distances from the embeddings
    df['distances'] = distances_from_embeddings(q_embeddings, df['embeddings'].values, distance_metric='cosine')

    df_result = df.sort_values('distances', ascending=True).head(3)

    # Return the context
    return " ".join([doc for doc in df_result.document])

def retrieve_similar_documents(question, vectordb):
    # Use the query method to retrieve the top 3 most similar documents
    results = vectordb.query(
        query_texts=[question],
        n_results=3
    )

    # Concatenate the retrieved documents
    context = " ".join([doc for doc in results])

    return context
```

# Embedding & Vector Store



What is the difference between a vector store and standard database?

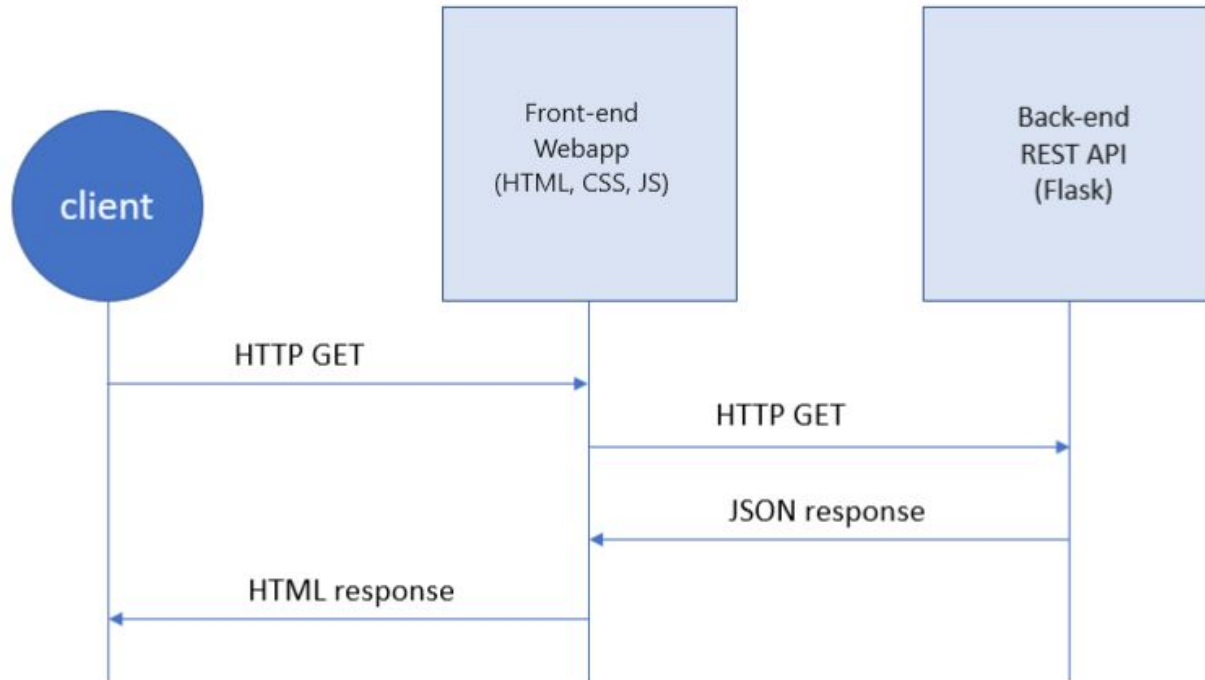
**Efficient Similarity Search:** Optimized for fast nearest neighbor searches in high-dimensional spaces.

**Scalability:** Handle large volumes of data and still perform efficient searches. This scalability is crucial in applications like RAG where the knowledge base can be vast.

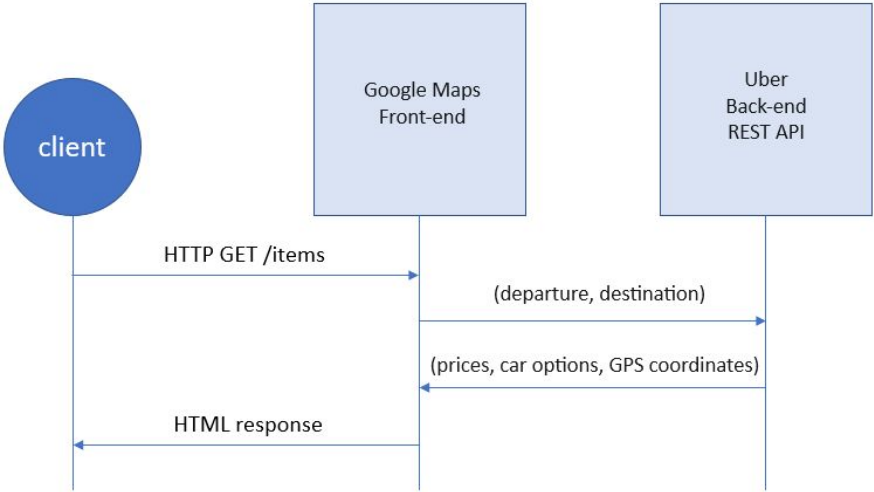
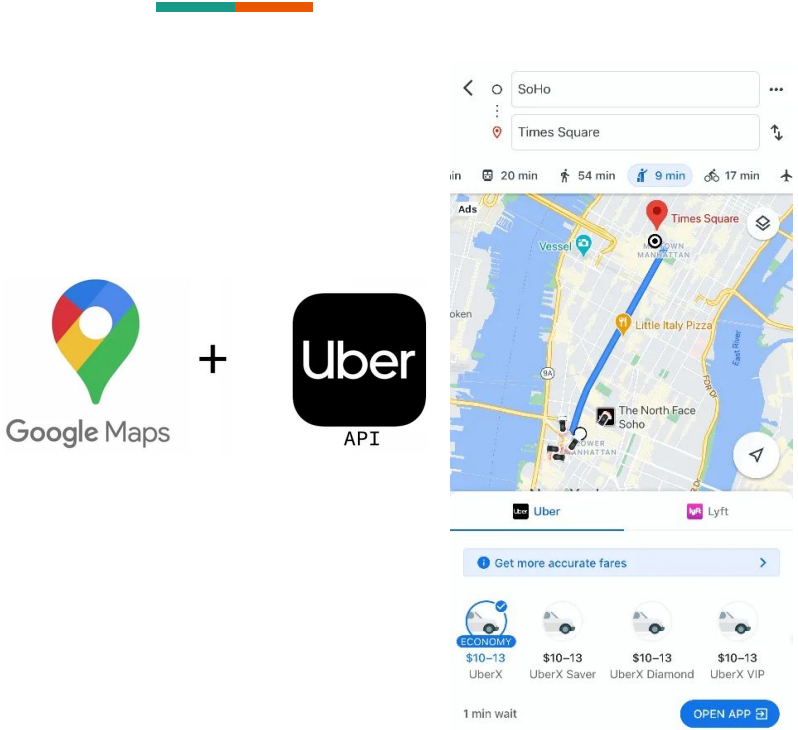
---

# Web application development : Front-end and Back-end

# API : Principles



# API : Principles



# Develop an API with Flask in Python

app.py

```
from flask import Flask, request, jsonify, render_template

app = Flask(__name__)

# Dummy function to simulate getting car details
def get_car_details(departure, destination):
    # This is where you'd implement the logic to get the car details.
    # For this example, we'll just return some hardcoded data.
    return {
        "type_of_car": "UberX",
        "price": "$10-13",
        "gps_coordinates": {"lat": 40.7128, "long": -74.0060}
    }

@app.route('/get_car', methods=['GET'])
def get_car():
    departure_address = request.args.get('departure_address')
    destination_address = request.args.get('destination_address')

    car_details = get_car_details(departure_address, destination_address)
    return jsonify(car_details)

if __name__ == '__main__':
    app.run(debug=True)
```

# Develop a front-end using HTML, CSS and JS

## index.html

```
<!DOCTYPE html>
<html lang="en">
<head>
  <meta charset="UTF-8">
  <meta name="viewport" content="width=device-width, initial-scale=1.0">
  <title>Ride Details</title>
</head>
<body>
  <h2>Get Ride Details</h2>
  <label for="departure_address">Departure Address:</label>
  <input type="text" id="departure_address" name="departure_address"><br><br>

  <label for="destination_address">Destination Address:</label>
  <input type="text" id="destination_address" name="destination_address"><br><br>

  <button onclick="getRideDetails()">Get Details</button>

  <h3>Ride Details</h3>
  <div id="rideDetails"></div>

  <script>
    function getRideDetails() {
      var departureAddress = document.getElementById('departure_address').value;
      var destinationAddress = document.getElementById('destination_address').value;

      fetch('/get_car?departure_address=${departureAddress}&destination_address=${destinationAddress}')
        .then(response => response.json())
        .then(data => {
          var details = `<p>Type of Car: ${data.type_of_car}</p>
            <p>Price: ${data.price}</p>
            <p>GPS Coordinates: lat. ${data.gps_coordinates.lat} long. ${data.gps_coordinates.long}</p>`;
          document.getElementById('rideDetails').innerHTML = details;
        })
        .catch(error => console.error('Error:', error));
    }
  </script>
</body>
</html>
```

## app.py

```
@app.route('/')
def index():
    return render_template('index.html')
```

```
/YourFlaskApp
/static
  /css
    - style.css
  /js
    - script.js
  /images
    - logo.png
/templates
  - index.html
  - layout.html
  - other_template.html
- app.py
- requirements.txt
```



---

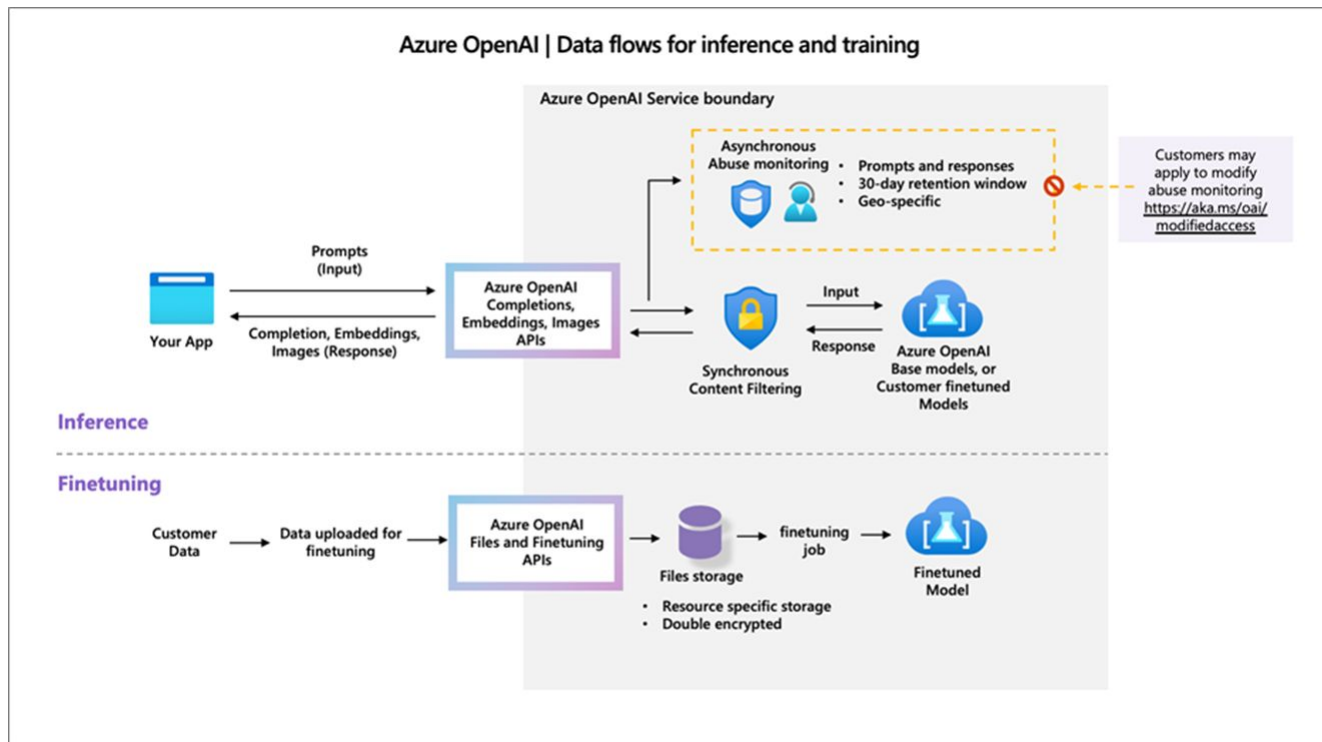
# Code walkthrough : How a Flask web application is structured?

<https://github.com/End2EndAI/travel-ai-translator>

---

# Practical usage : Using Azure OpenAI API and ChatGPT

# OpenAI Service through Azure API



# OpenAI Service through Azure API

```
# Loading Azure API key from configuration file
config = configparser.ConfigParser()
config.read('config.ini')
AZURE_OPENAI_KEY = config.get('OPENAI_API', 'AZURE_OPENAI_KEY')
AZURE_OPENAI_ENDPOINT = config.get('OPENAI_API', 'AZURE_OPENAI_ENDPOINT')

# Configuration for Embedding using Chroma & Azure
openai.api_key = AZURE_OPENAI_KEY
openai.api_base = AZURE_OPENAI_ENDPOINT
openai.api_type = 'azure'
openai.api_version = '2023-12-01-preview' # this may change in the future
llm_deployment_name = 'gpt-35-turbo'

TEMPERATURE_GPT = 0
```

```
messages = [
    {
        "role": "system",
        "content": system_prompt
    },
    {
        "role": "user",
        "content": user_prompt
    }
]

generated_output = openai.ChatCompletion.create(
    engine=llm_deployment_name,
    messages=messages,
    temperature=TEMPERATURE_GPT
)
```

# ChatGPT



Task	Description
Learning	Ask to explain concepts, give examples, ...
Coding	<p>Ask to write template code to start a project, instead of starting from scratch.</p> <p>For instance :</p> <ul style="list-style-type: none"><li>- Frontend (HTML, CSS, JS)</li><li>- Backend (Flask app template)</li></ul>
Debugging	Ask to analyse and solve errors
Cleaning & Documenting	Ask to clean and document your code. Ask to write a README.

# Prompt engineering



*The more you prompt to ChatGPT, the better you will know how to prompt to get what you want.*

## Structuring Prompts

### # CONTEXT #

*I want to advertise my company's new product. My company's name is Alpha and the product is called Beta, which is a new ultra-fast hairdryer.*

### # OBJECTIVE #

*Create a Facebook post for me, which aims to get people to click on the product link to purchase it.*

### # STYLE #

*Follow the writing style of successful companies that advertise similar products, such as Dyson.*

### # TONE #

*Persuasive*

### # AUDIENCE #

*My company's audience profile on Facebook is typically the older generation. Tailor your post to target what this audience typically looks out for in hair products.*

### # RESPONSE #

*The Facebook post, kept concise yet impactful.*

# Prompt engineering

## Sectioning Prompts Using Delimiters

Classify the sentiment of each conversation in <<<CONVERSATIONS>>> as 'Positive' or 'Negative'. Give the sentiment classifications without any other preamble text.

###

EXAMPLE CONVERSATIONS

[Agent]: Good morning, how can I assist you today?

[Customer]: This product is terrible, nothing like what was advertised!

[Customer]: I'm extremely disappointed and expect a full refund.

[Agent]: Good morning, how can I help you today?

[Customer]: Hi, I just wanted to say that I'm really impressed with your product. It exceeded my expectations!

###

EXAMPLE OUTPUTS

Negative

Positive

###

<<<

[Agent]: Hello! Welcome to our support. How can I help you today?

[Customer]: Hi there! I just wanted to let you know I received my order, and it's fantastic!

[Agent]: That's great to hear! We're thrilled you're happy with your purchase. Is there anything else I can assist you with?

[Customer]: No, that's it. Just wanted to give some positive feedback. Thanks for your excellent service!

[Agent]: Hello, thank you for reaching out. How can I assist you today?

[Customer]: I'm very disappointed with my recent purchase. It's not what I expected at all.

[Agent]: I'm sorry to hear that. Could you please provide more details so I can help?

[Customer]: The product is of poor quality and it arrived late. I'm really unhappy with this experience.

>>>

“Delimiters are special tokens that help the LLM distinguish which parts of your prompt it should consider as a single unit of meaning.

This is important because your entire prompt arrives to the LLM as a single long sequence of tokens.

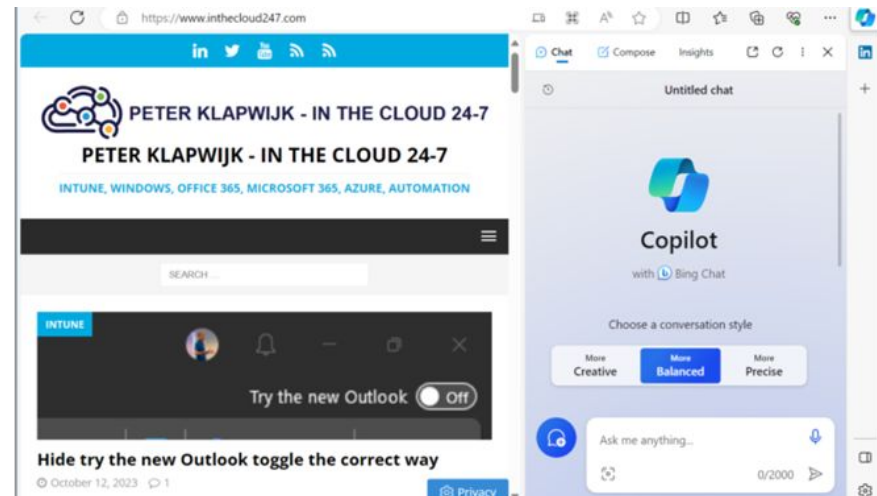
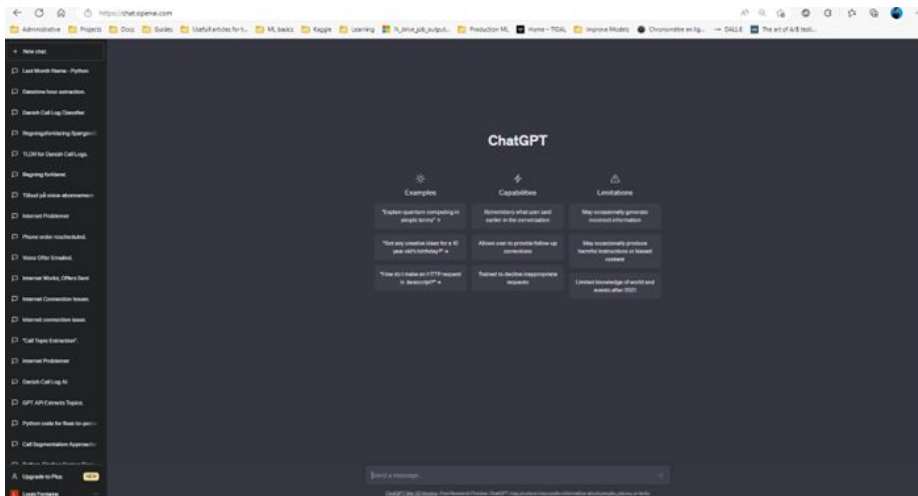
Delimiters provide structure to this sequence of tokens by fencing specific parts of your prompt to be treated differently.”

---

# GDPR and Project Management Strategy



# Data privacy and LLM



User's inputs and outputs are used for product improvements and model training in ChatGPT, Copilot or any other “free LLM” platforms.

# Data privacy and LLM

Solutions	Performance / Accuracy	Cost	Implementation difficulty	Data privacy	Documentation
Open AI API	Very High	Low	Low	<ul style="list-style-type: none"><li>• Zero-data retention: no data is stored, anywhere.</li><li>• Model hosted in US.</li><li>• Business data is not used for model training.</li><li>• The company owns its inputs and outputs from the API.</li></ul>	<a href="#">Link</a> <a href="#">Link</a>
Microsoft Azure Open AI Service	Very High	Low	Low	<ul style="list-style-type: none"><li>• <b>Zero-data retention: no data is stored, anywhere.</b></li><li>• <b>Models are hosted in EU.</b></li><li>• <b>Business data is not used for any model training.</b></li><li>• <b>The company owns its inputs and outputs from the API.</b></li><li>• <b>Data is not sent to Open AI.</b></li></ul>	<a href="#">Link</a> <a href="#">Link</a>
Open-source model running on Azure	Medium	Medium	High	Same as for Azure Open AI Service solution, but with more control of the data wrangling and storage.	<a href="#">Link</a>
Open-source model running on premise	Medium	Very High (invest in GPUs)	High	No risks.	

# LLM projects : Strategy



Role	Description
Legal & Security	<ul style="list-style-type: none"><li>→ Understand the technology behind AI (particularly Generative AI)</li><li>→ Assess the provider's data privacy policies and the project data flow</li><li>→ Fill a Data Privacy Impact Assessment for the use case</li><li>→ Make recommendations</li></ul>
Top management - Stakeholders	<ul style="list-style-type: none"><li>→ Understand the possible use cases of Generative AI in the company</li><li>→ Understand the value of Generative AI</li><li>→ Prioritize Generative AI projects in the company</li></ul>

---

# Project

<https://github.com/End2EndAI/Generative-AI-Module-Dauphine>