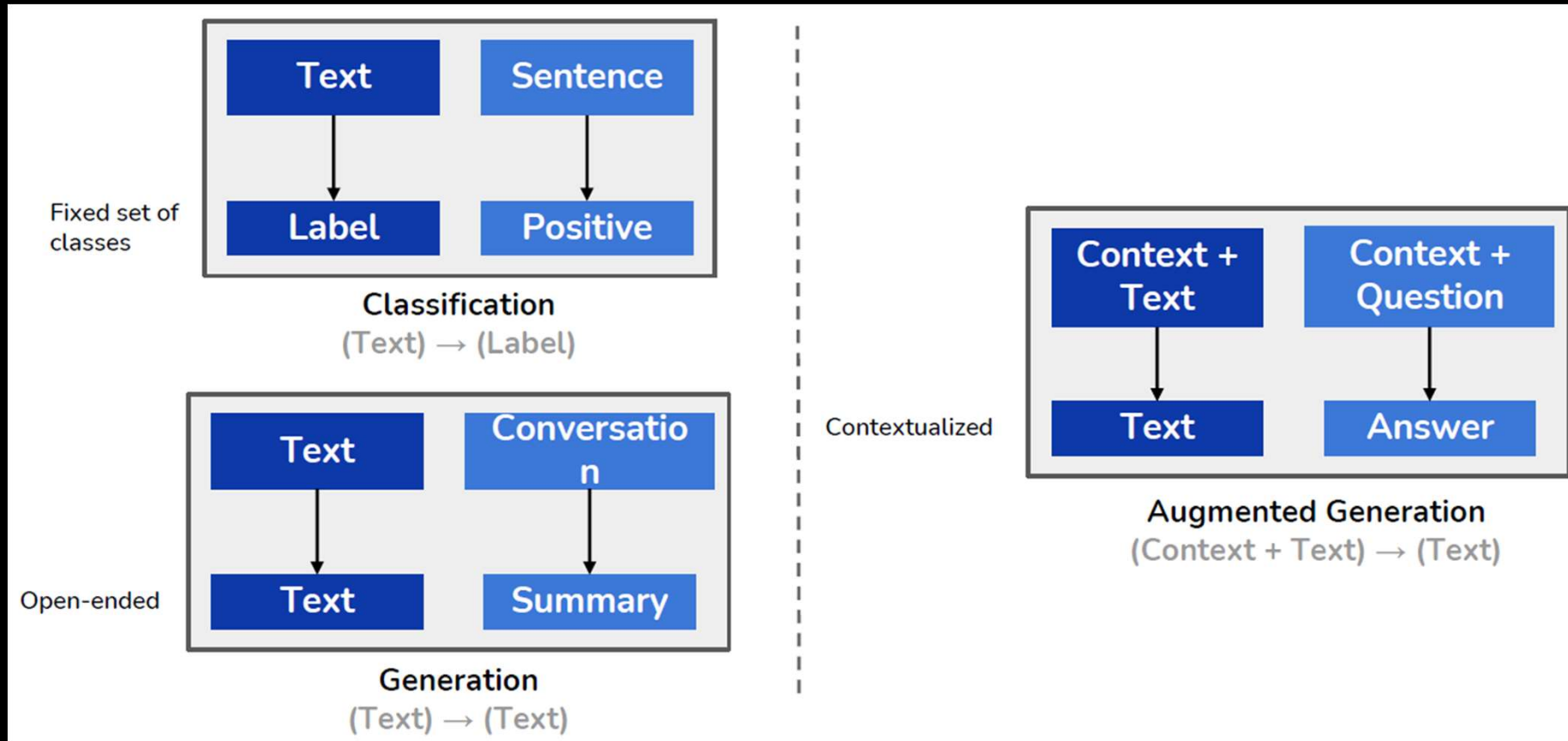




Generative AI Academy

Business problems solved by LLMs – a taxonomy



Accessing LLMs through Azure Open AI

Azure provides controlled, fast access to a host of Open AI LLMs.

Base models

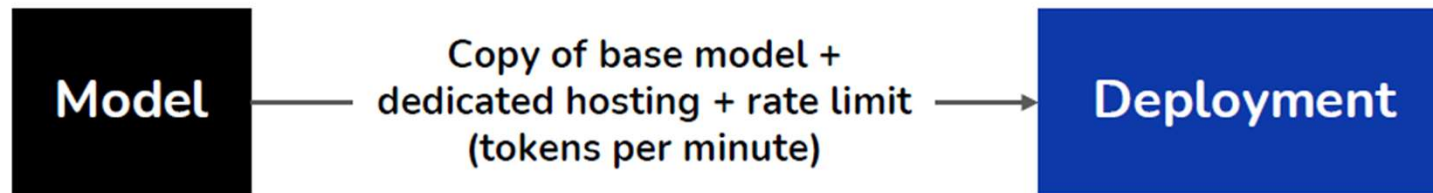
 Deploy  Create a custom model  Column options  Refresh

 Search

Model name ▾	Model version ▾	Created at ▾	Status ▾	Deployable ▾
gpt-35-turbo	0613	6/19/2023 5:30 AM	✓ Succeeded	✓ Yes
gpt-35-turbo	0301	3/9/2023 5:30 AM	✓ Succeeded	✓ Yes
gpt-35-turbo-16k	0613	6/19/2023 5:30 AM	✓ Succeeded	✓ Yes
text-embedding-ada-002	2	4/3/2023 5:30 AM	✓ Succeeded	✓ Yes
text-embedding-ada-002	1	2/2/2023 5:30 AM	✓ Succeeded	✓ Yes

Accessing LLMs through Azure Open AI

Azure provides controlled, fast access to a host of Open AI LLMs.



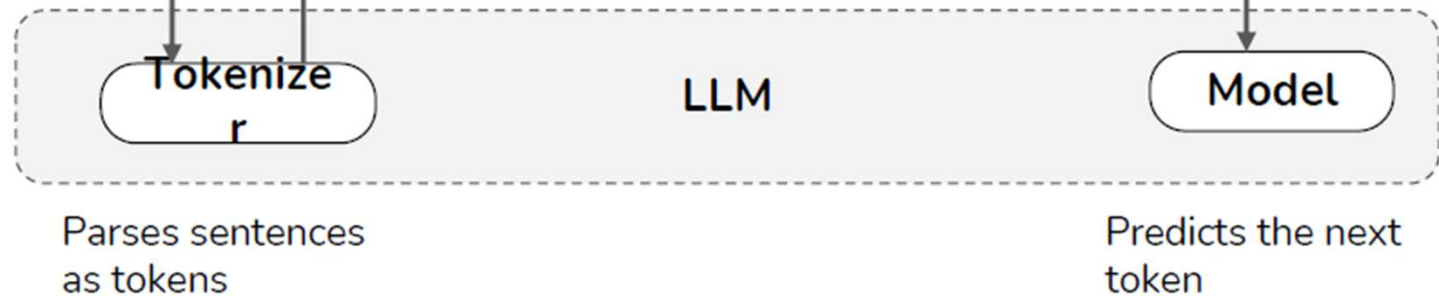
Deployment name ▾	Model version ▾	Capacity	Model deprecati... ▾	Content Filter ▾	Rate limit (Tokens per minute) ▾
<input checked="" type="radio"/> gpt-35-turbo	0613	134K TPM	7/5/2024	Default	134000
<input type="radio"/> gpt-35-turbo-16k	0613	135K TPM	1/15/2024	Default	135000
<input type="radio"/> text-embedding-ada-002	2	133K TPM	2/2/2025	Default	133000

Understanding tokens

A token refers to a segment or piece of text, such as a word, punctuation, or other meaningful element, into which input text is divided for processing by the model.

The movie was awesome.
Overall, the experience
was positive.

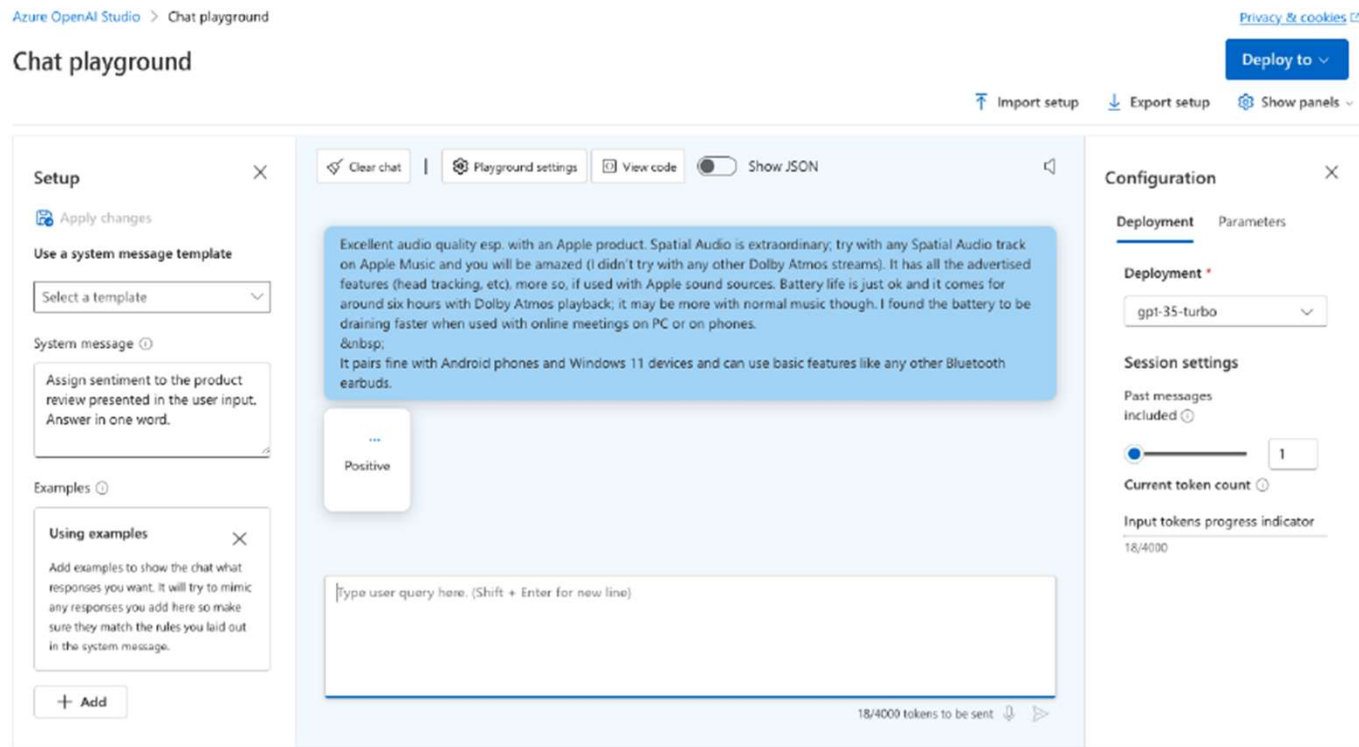
[The,movie,was,awesome,.,Overall,,the
,experience,was,positive,.]



Demo: tokenizer: <https://platform.openai.com/tokenizer>

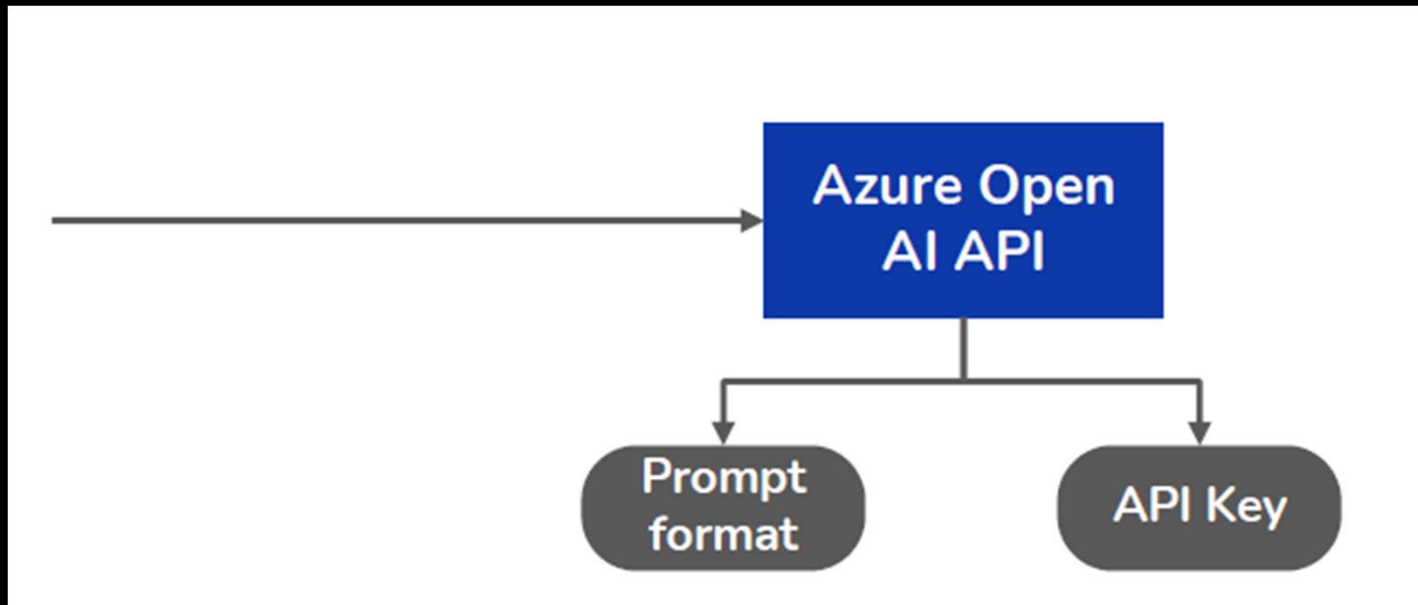
Azure Open AI Playground

The Azure Open AI Playground enables iterative development for prompt engineering, that is, designing specific instructions for LLMs to accomplish a task.



Accessing Azure Open AI via APIs

- Colab notebook
- Python or any other application



Accessing Azure Open AI via APIs

Demo and hands-on

- Accessing Azure OpenAI from Colab notebook
- Accessing Azure Open AI from your IDE
(VS Code or PyCharm or Spyder or any other)

Prompt Engineering

Prompt Engineering

Prompt engineering is the process of crafting a high-quality prompt for a GenAI model to generate response

Why are Prompts Important ?



Provide clear direction: Prompts provide clear direction to AI, ensuring it understands your needs.



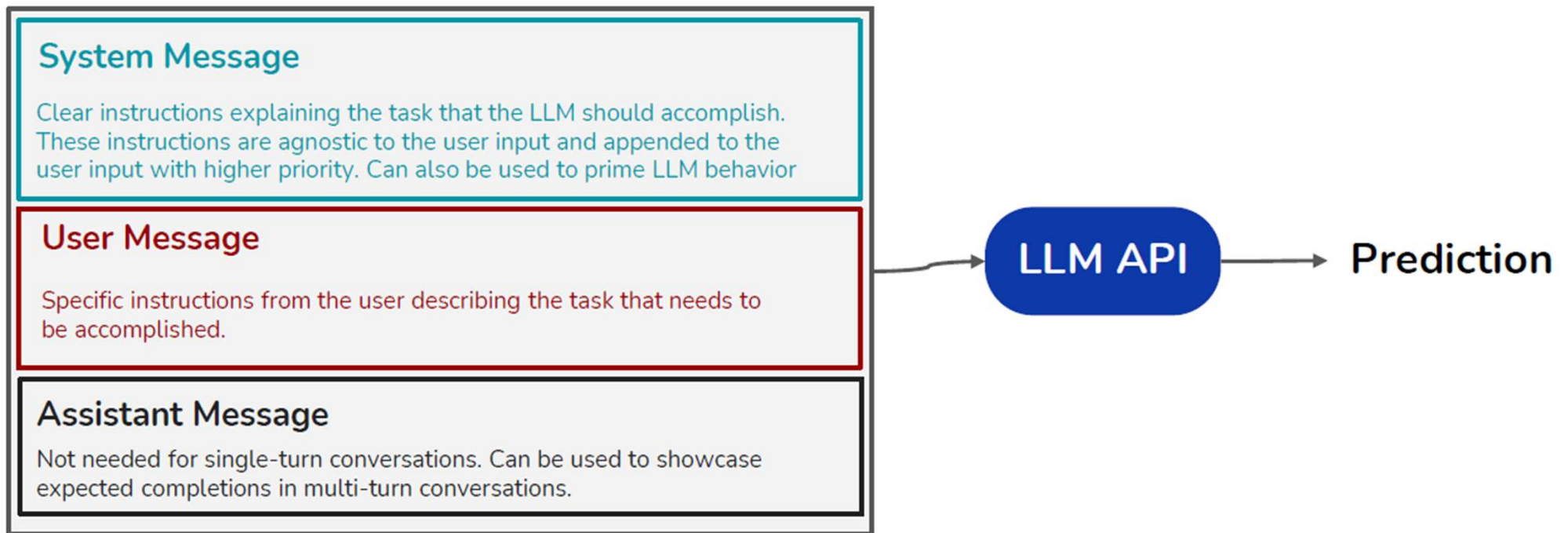
Improve accuracy: Well-crafted prompts enhance the accuracy of AI responses.



Save time: Specific prompts save time by reducing back-and-forth clarifications.

Prompt Format

Azure Open AI APIs are compatible with the Open AI APIs and have the following three components.



Prompt Format - example

System Message

You are an AI assistant designed to help users with financial planning and investment strategies. Provide accurate and helpful information while maintaining a friendly and professional tone.

User Message

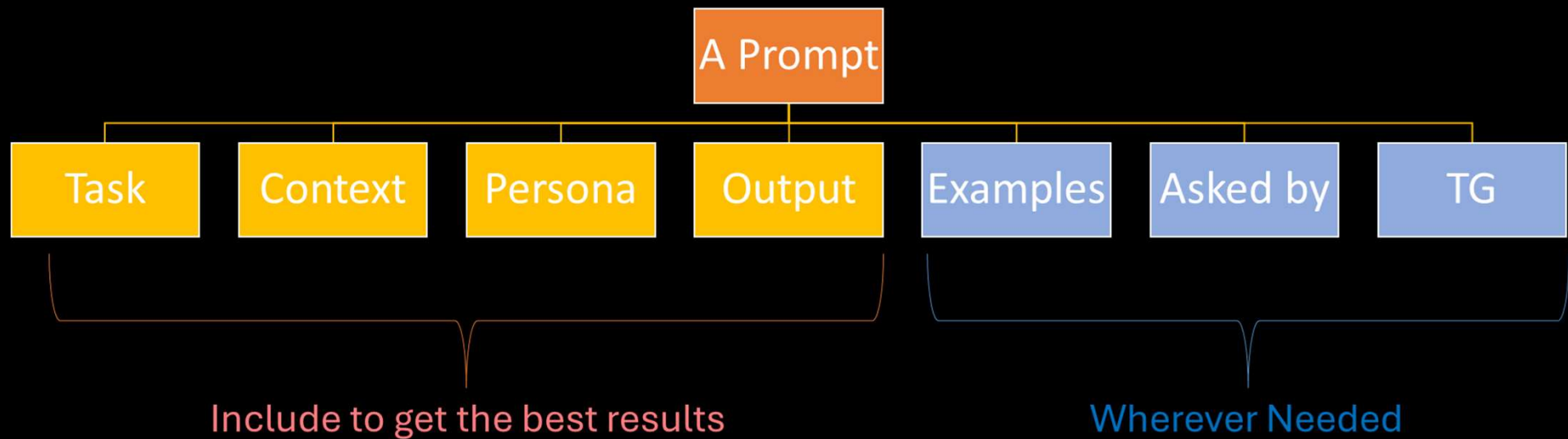
I'm looking for advice on how to diversify my investment portfolio. Can you suggest some strategies?

Assistant Message

Diversifying your investment portfolio is a smart way to manage risk and enhance potential returns. Here are some strategies you might consider:

1. ****Asset Allocation****: Divide your investments among different...
2.

Components of an effective prompt



Relevance and specificity

- Clear and specific prompt
- Clearly define the task or information that the user wants from the model
- Avoids ambiguity or confusion

 **You**
Explain banking services

 **ChatGPT**
Banking Services: A Comprehensive Overview

Banking services encompass a broad range of financial products and offerings provided by banks to individuals, businesses, and governments. These services play a crucial role in facilitating economic activities, managing finances, and promoting financial well-being. Here is a comprehensive overview of key banking services:


****1. Checking and Savings Accounts:**

- **Checking Accounts:** Designed for everyday transactions, offering features like debit cards and check-writing.
- **Savings Accounts:** Provide a safe place to save money while earning interest.

****2. Credit Services:**

- **Credit Cards:** Allow users to make purchases on credit, with the option to pay over time.
- **Personal Loans:** Unsecured loans for various purposes, often with fixed terms and



 **You**
Compose an overview highlighting the key differences between savings accounts and certificates of deposit (CDs) in banking, including their respective interest rates, withdrawal restrictions, and suitability for different financial goals

 **ChatGPT**
Savings Accounts vs. Certificates of Deposit (CDs): A Comparative Overview

Savings Accounts:

1. Interest Rates:

- Typically offer lower interest rates compared to CDs.
- Interest rates may vary based on the bank and market conditions.

2. Withdrawal Flexibility:

- Allow more flexibility for withdrawals, often permitting frequent transactions.
- Accessible for daily transactions and emergencies without penalties.

3. Suitability:

- Ideal for short-term financial goals or emergency funds.



Format

- Structure and style of the prompt
- Influence the model's response. E.g., Comparison prompts, command prompts, Q&A, etc.



You

A diversified portfolio helps spread risk by investing in a variety of assets, such as stocks, bonds, and _____.



ChatGPT

A diversified portfolio helps spread risk by investing in a variety of assets, such as stocks, bonds, and **other financial instruments or asset classes.**

Fill-in-the-blanks

Role



You

what is swap



ChatGPT

A "swap" refers to the exchange or substitution of one thing for another.



You

As a financial analyst, explain what swap is.



ChatGPT

The term "swap" in finance refers to a derivative contract where two parties agree to exchange financial instruments, such as interest rates or currencies. As a financial analyst, understanding swaps is crucial. Interest rate swaps involve exchanging fixed-rate and floating-rate interest payments, while currency swaps involve exchanging cash flows in different currencies. These agreements are used to manage risks, hedge against fluctuations, and optimize financial positions within the market. Analyzing the terms, risks, and potential benefits of swaps is integral to making informed financial decisions and managing a portfolio effectively.

Finance related – company financials analysis

Reliance Industries Limited reported earnings results for the first quarter ended June 30, 2023.

For the first quarter, the company reported sales was INR 2,278,600 million compared to INR 2,387,200 million a year ago.

Revenue was INR 2,146,440 million compared to INR 2,249,010 million a year ago.

Net income was INR 160,110 million compared to INR 179,550 million a year ago.

Basic earnings per share from continuing operations was INR 23.66 compared to INR 26.48 a year ago. Diluted earnings per share from continuing operations was INR 23.66 compared to INR 26.48 a year ago.

Basic earnings per share was INR 23.66 compared to INR 26.54 a year ago.

Diluted earnings per share was INR 23.66 compared to INR 26.54 a year ago.

Extract the data into a table.

Perform a simple descriptive analysis on it

Understanding tokens



Token:

- Smallest unit into which text data can be broken down for an AI model to process (Think 1,000 tokens to be about 750 words)
- Exact tokenization process varies between models.
- Request Tokens are generated from the input prompt. Response Tokens are generated by the model from the output.

Token Limit:

- = Total tokens (input + output) a model can process in one request

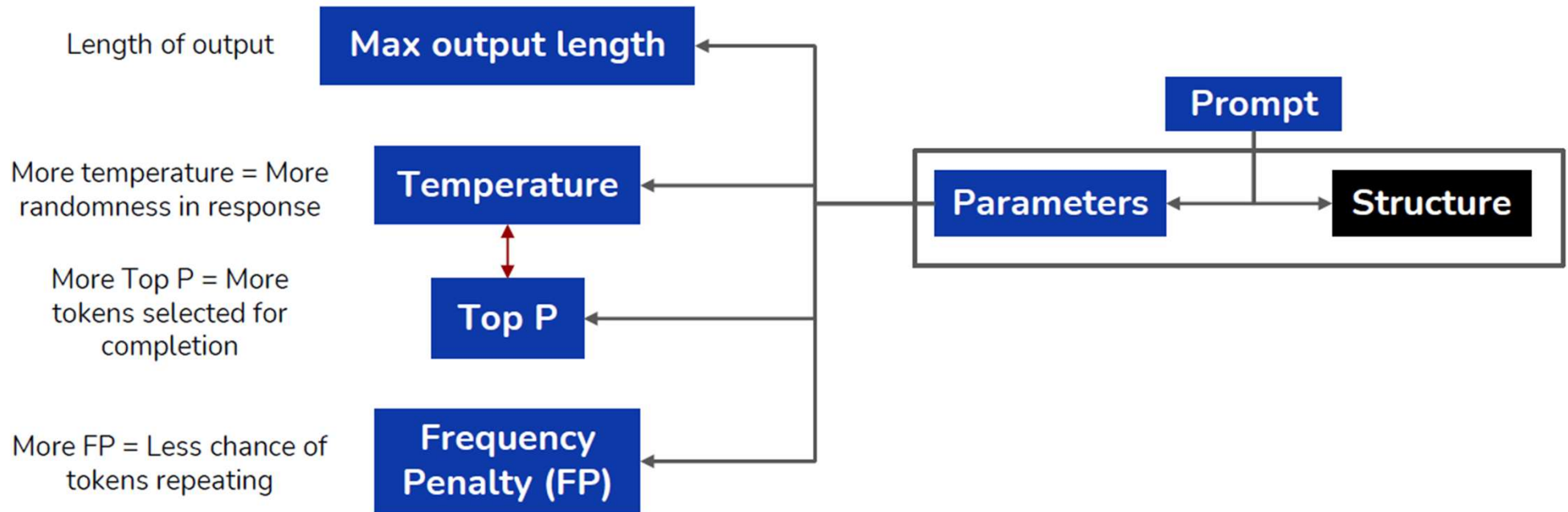
Context Window:

- The model's capacity to retain and consider tokens for generating responses.

Prompt – tunable settings

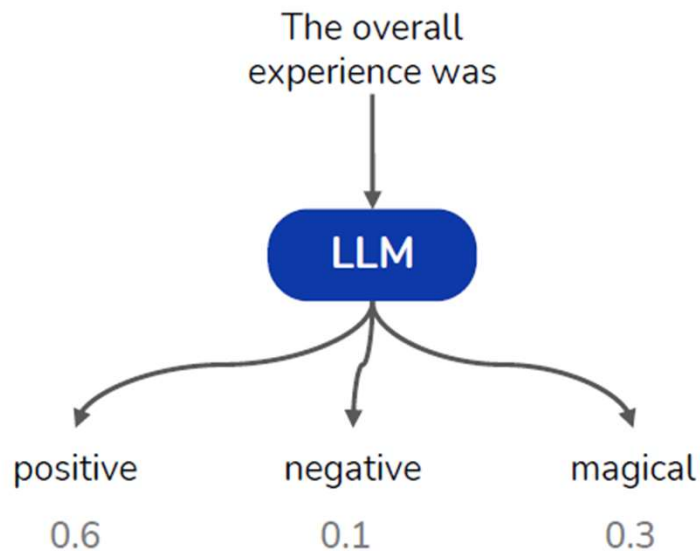
Prompt = Specific set of instructions sent to a LLM to accomplish a task

Engineering = Iteratively deriving a specific prompt for the task



Prompt – understanding temperature

Understanding temperature



Temperature = 0

The overall experience was positive
The overall experience was positive
The overall experience was positive

Repeated execution produces the same results

Temperature = 1

The overall experience was positive
The overall experience was magical
The overall experience was negative

Repeated execution can produce different results

Prompt – understanding top-p (nucleus sampling)

Range: 0.0 to 1.0

Default: 1.0

Effect: Limits the response to the most probable tokens.

Low (0.1 to 0.3): Focuses only on the topmost likely responses.

Medium (0.4 to 0.7): Balances diversity and relevance.

High (0.8 to 1.0): Allows more variation in responses.

👉 Tip: Lowering top-p makes responses more focused. It works well with a moderate temperature.

Prompt – understanding max tokens

- Range: Up to 4096 tokens (for gpt-4).
- Effect: Limits response length.

One token \approx 4 characters or 0.75 words.

👉 Example: "Summarize this article in 100 words." Max tokens: 150

Prompt – understanding max tokens

- Range: Up to 4096 tokens (for gpt-4).
- Effect: Limits response length.

One token \approx 4 characters or 0.75 words.

👉 Example: "Summarize this article in 100 words." Max tokens: 150

Prompt – understanding frequency penalty

- Range: -2.0 to 2.0
- Default: 0.0
- Effect: Reduces repetition.
 - Positive values (0.5 to 2.0): Discourages frequent words or phrases.
 - Negative values (-2.0 to -0.5): Encourages repetition.

👉 Example: "Write a report on climate change." Frequency penalty: 1.0

→ Less repetitive phrasing.

Prompt – understanding presence penalty

- Range: -2.0 to 2.0
 - Default: 0.0
 - Effect: Encourages new topics or words.
 - Positive values (0.5 to 2.0): Promotes novelty.
 - Negative values (-2.0 to -0.5): Promotes sticking to existing topics.
- 👉 Example: "Write a story about space exploration." Presence penalty: 1.5
- More unique ideas.

Common settings combinations

Use Case	Temperature	Top-p	Max Tokens	Frequency Penalty	Presence Penalty
Factual Answers	0.2 to 0.4	0.3 to 0.7	500–1000	0.0	0.0
Creative Writing	1.0 to 1.5	0.8 to 1.0	1000–2000	0.0	0.5 to 1.0
Brainstorming Ideas	1.2 to 1.8	0.9 to 1.0	500–1500	0.0	1.0 to 2.0
Code Generation	0.2 to 0.5	0.3 to 0.7	500–1500	0.0	0.0

Prompt Engineering

- Zero-shot, Few-shot
- Chain-of-Thought
- Rephrase & Respond
- Self-Consistency
- LLM-as-a-judge
- Tree-of-thought

Walkthrough:

<https://www.promptingguide.ai/techniques/zeroshot>

[Prompt enhancement using PromptPerfect.](#)

<https://www.aiforwork.co/>

Prompt Engineering – demo – generate hospital admission letter (1/2)

You are an AI assistant that generates a hospital admission letter.

****Your task**:**

- Collect the following details from the user:

1. Patient's full name
2. Date of birth
3. Diagnosis
4. Doctor's name
5. Admission date
6. Reason for admission

Generate hospital admission letter (2/2)

****Instructions**:**

1. If any detail is missing or unclear, ask the user follow-up questions to obtain it.
2. Once you have all the required details, confirm them with the user.
3. Finally, produce a ****professional hospital admission letter**** that includes all the information in a clear and formal format.

****Prompting Strategy**:**

- First, greet the user and ask them for the required information one by one.
- If any piece of information is missing or conflicting, request clarification or additional details.
- Once all data is collected, generate the final letter.
- The final output should look like an official admission letter, including a header, body, and signature line.

Begin.

Prompt Engineering pitfalls



Ambiguity: Prompts that are vague or unclear can lead to confusing or irrelevant responses from the AI.

Lack of Context: Providing insufficient background information can result in responses that are out of context or irrelevant.

Overly Complex Prompts: Including too many variables or complicated structures can overwhelm the AI and reduce the quality of the response.

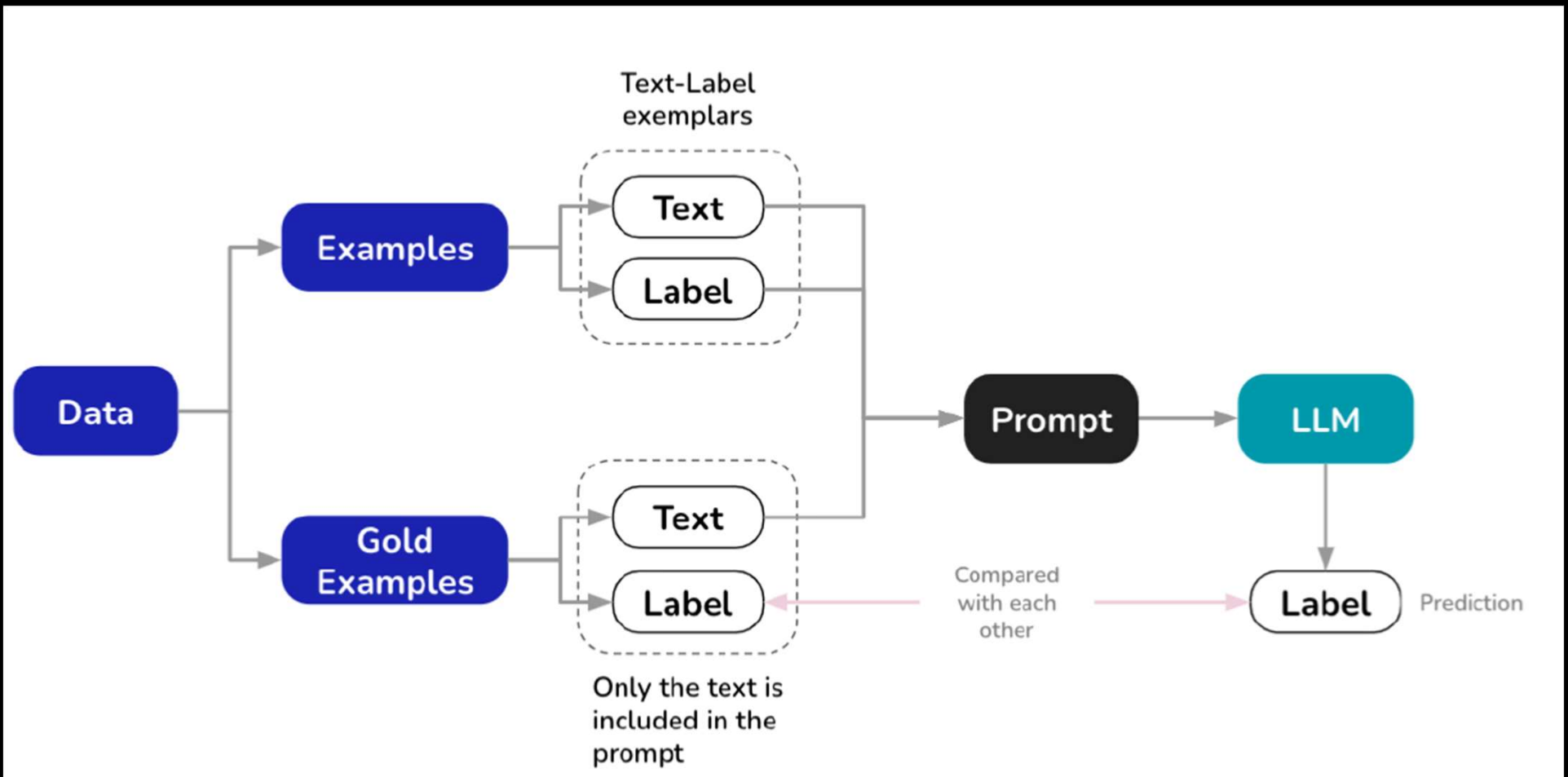
Unclear Objectives: Prompts without clear goals or expected outcomes can result in generic or unhelpful responses.

Misunderstanding Model Capabilities: Models have a knowledge cut-off date and are not infallible. It can make mistakes or misunderstand requests.

Prompt Engineering demo and hands-on

LLMs for classification and generation tasks

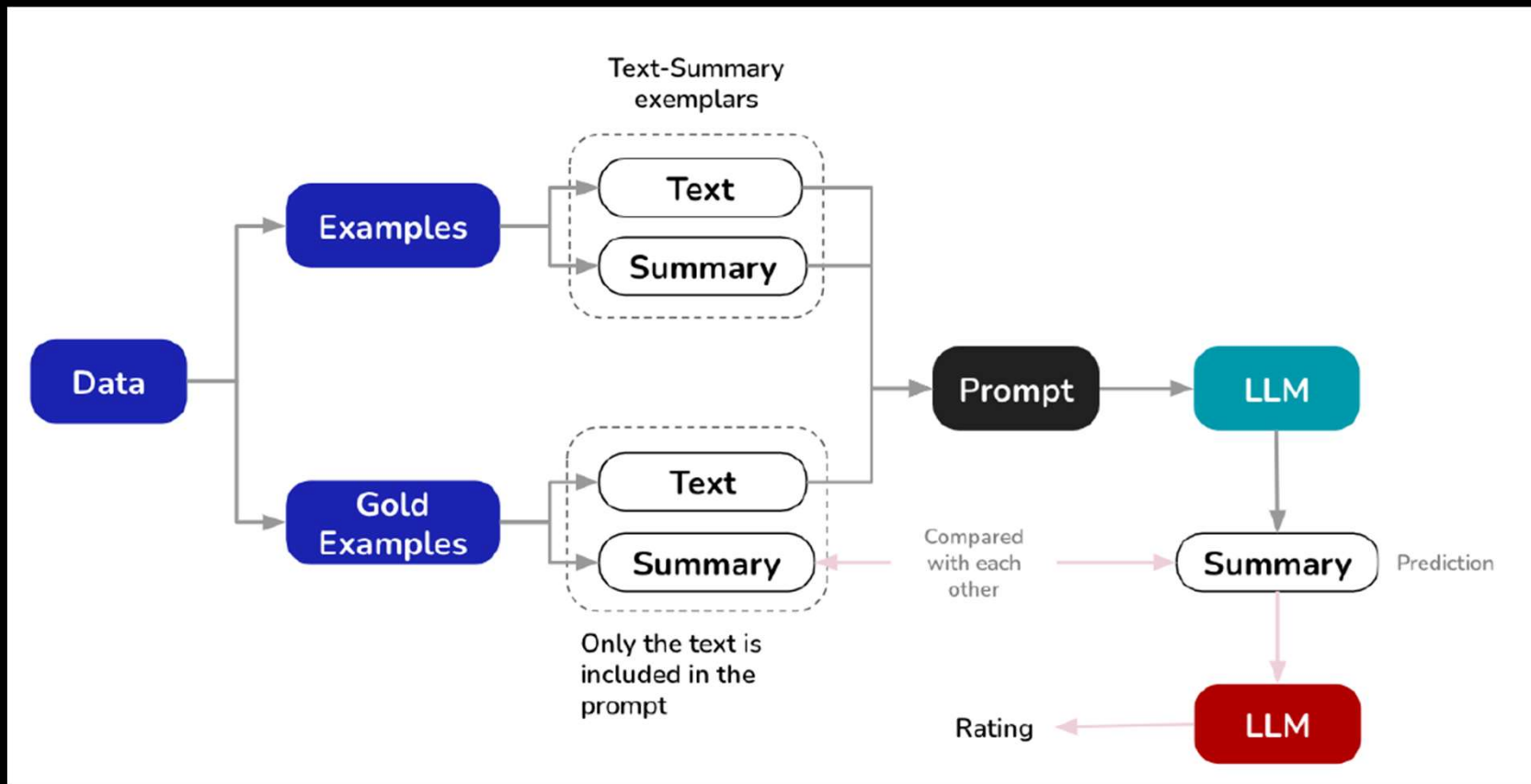
LLMs for classification tasks - overall flow



Classification Hands-on

- Sentiment Analysis
- Aspect-based Sentiment Analysis

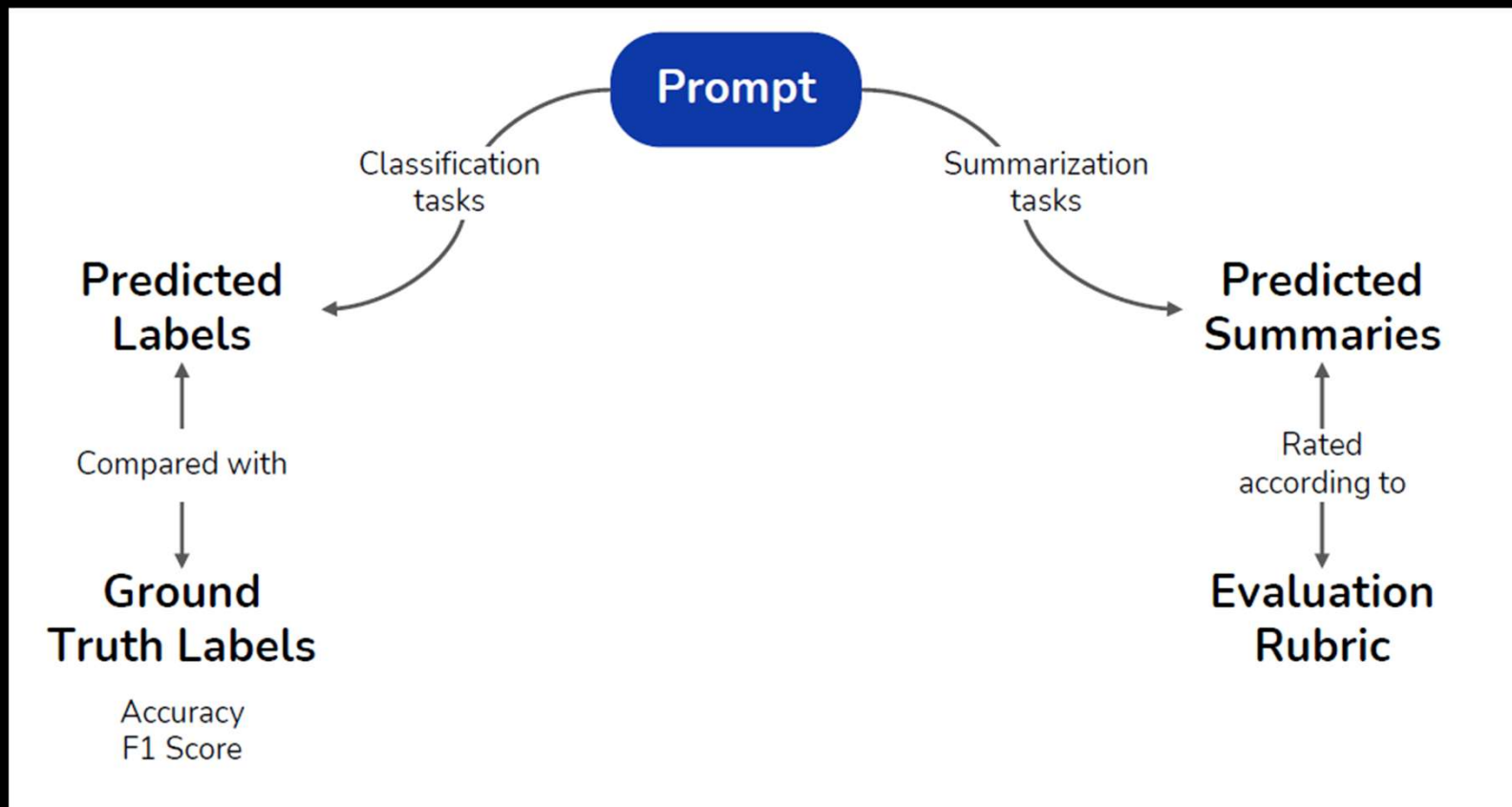
LLMs for generation tasks - approach



Generation Hands-on

- Dialogue Summarization

Evaluation for classification and generation



Summary of full flow

