Accionlabs

Accion INNOVATION SUMMIT 2023

02-05 March 2023,
Sofitel Dubai
The Palm Jumeirah
Dubai

Accion
INNOVATION
SUMMIT 2023

Accionlabs

Generative AI, LLMs (like Chat GPT), and ML in Production

Accion
INNOVATION
SUMMIT 2023

Accionlabs

Chetan Khatri

Senior Product Architect

As a seasoned Solutions Architect and Enterprise Product Architect in the field of Data Analytics, AI/ML/MLOps, he has worked with organizations across a wide range of industries, including product development, enterprise, and services. With a hands-on approach, he has built and delivered highly scalable digital enterprise solutions that leverage cutting-edge technologies and methodologies to provide exceptional value to clients.

Table of Contents

What is Generative AI? The landscape of Generative AI.

The landscape of Generative AI

Scopes across landscape

What is LLMs? Evolution of Language Models, How does it differentiate with NLP? Landscape of the LLMs

Business use cases of LLMs and NLP

GPT-3 and ChatGPT: Introduction and an Architecture

Product Development and Integration life cycle with GPT-3

Business use cases of GPT-3/ChatGPT/LLMs

ChatGPT

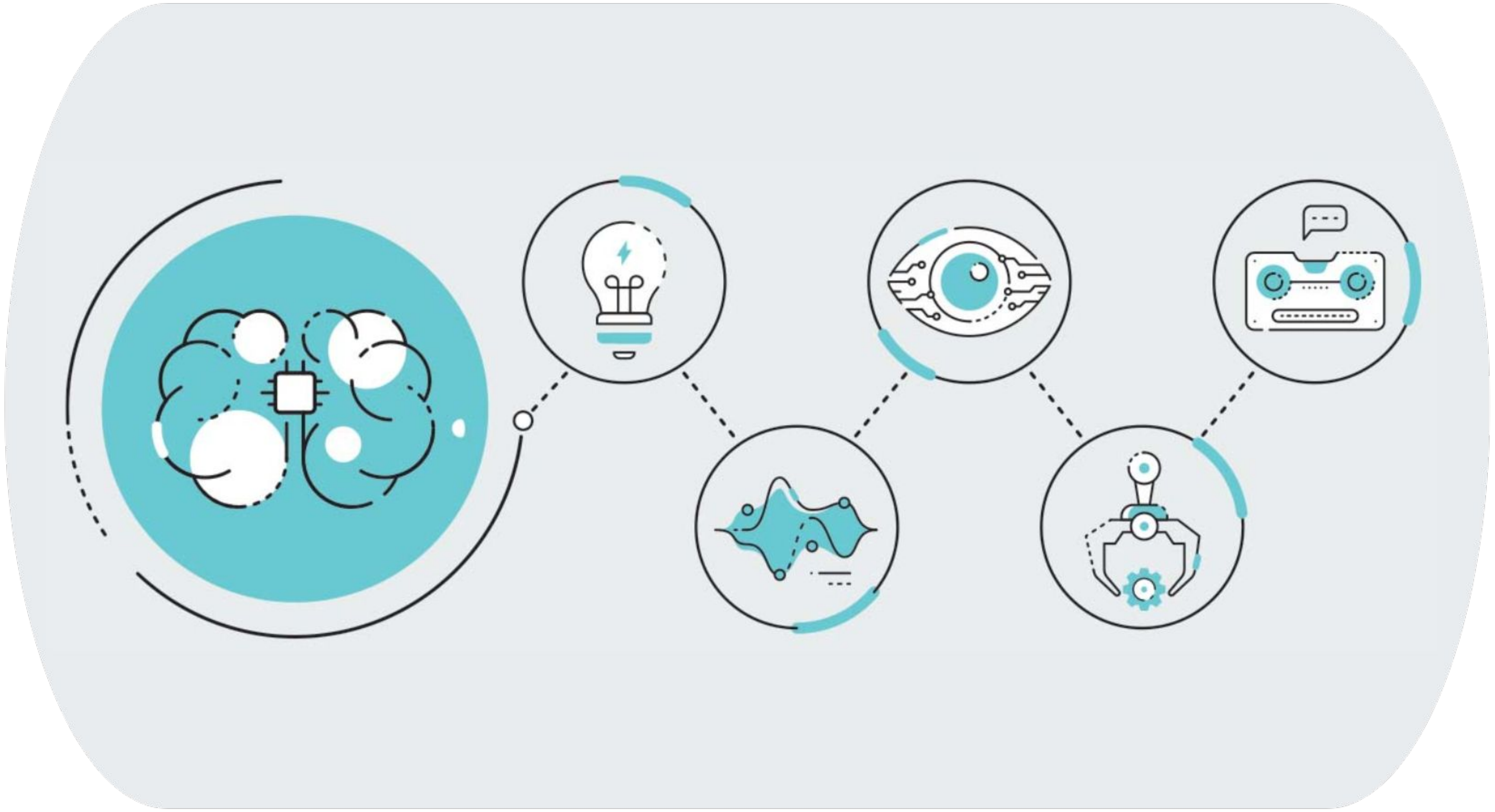
ChatGPT, Can you help me?

Edited by Accion Labs with copyrights belonging to original content owner

Video Editing Credit: Vinay Mune Gowda, Accion Labs

What is Generative AI?

- Generative AI is a powerful technology that enables machines to generate new content that is similar in style or pattern to existing data.
- As an example, a Large Language Model (LLM) can be trained on a vast amount of text data from various sources, such as books, articles, and websites, which allows it to understand the structure and patterns of language.
- Using this understanding, the LLM can generate new text that is both original and coherent.



Landscape of Generative AI

- **Transformers**
  - Transformers are a type of neural network architecture that is particularly suited for generating text.
  - Examples: GPT-3, (Chat)GPT, BERT
- **GANs**
  - A set of two neural networks - a generator and a discriminator - that work together to generate new data that is similar to the original dataset. They are commonly used to generate images and videos.
  - Examples: StyleGAN, CycleGAN
- **VAEs**
  - VAEs are a type of neural network that can learn the underlying structure of a dataset and generate new data that resembles the original data. They are commonly used to generate images, music, and speech.
  - VAEs create new data similar to the original but not identical, allowing for diversity.
  - VAEs generate diverse data that follows the original dataset's patterns.
  - Examples: Beta VAEs, Convolutional VAEs, Adversarial Variational Bayes (AVB), Vector Quantized Variational Autoencoder (VQ-VAE), World Models

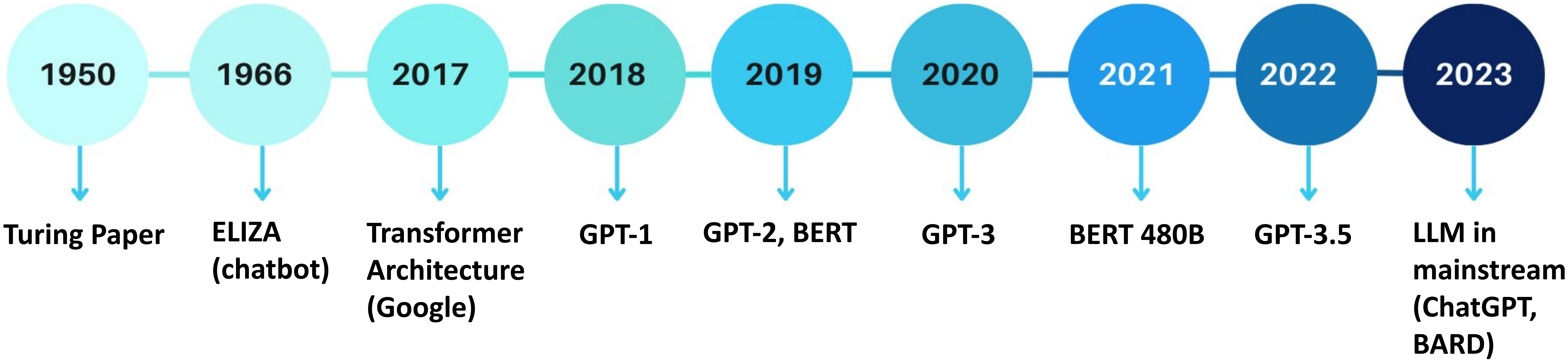
Generative AI: Scopes across landscape

Category	Description	Example(s)
Text	Summarising or automating content	GPT-3
Images	Generating images	DeepDream
Audio	Summarising, generating or converting text in audio	Lyrebird
Video	Generating or editing videos	Flexclip, Magisto
Code	Generating code	Github Copilot, Codota
Chatbots	Automating customer service and more	Bank of America, H&M
ML platforms	Applications / ML platforms.	AWS SageMaker, GoogleVertex.AI
Search	AI-powered insights.	Algolia
Gaming	Gen-AI gaming studios or applications.	AI Dungeon, DeepMind's AlphaGo
Data	Designing, collecting, or summarising data.	Google's AutoML, Appen

What is a Large Language Model (LLMs)?

1. **LLMs can generate human-like language**
   - ○ Analyze large amounts of text data
2. **GPT-3 as an advanced LLM**
   - ○ Can generate high-quality text content
   - ○ Supports various styles and genres
3. **Benefits of LLMs**
   - ○ Automate certain aspects of content creation and communication
   - ○ Improve consistency
   - ○ Reduce costs
4. **Responsible use of LLMs**
   - ○ Acknowledge imperfections
   - ○ Avoid generating errors or biased content
   - ○ Careful consideration of potential impacts.

Timeline of the LLMs

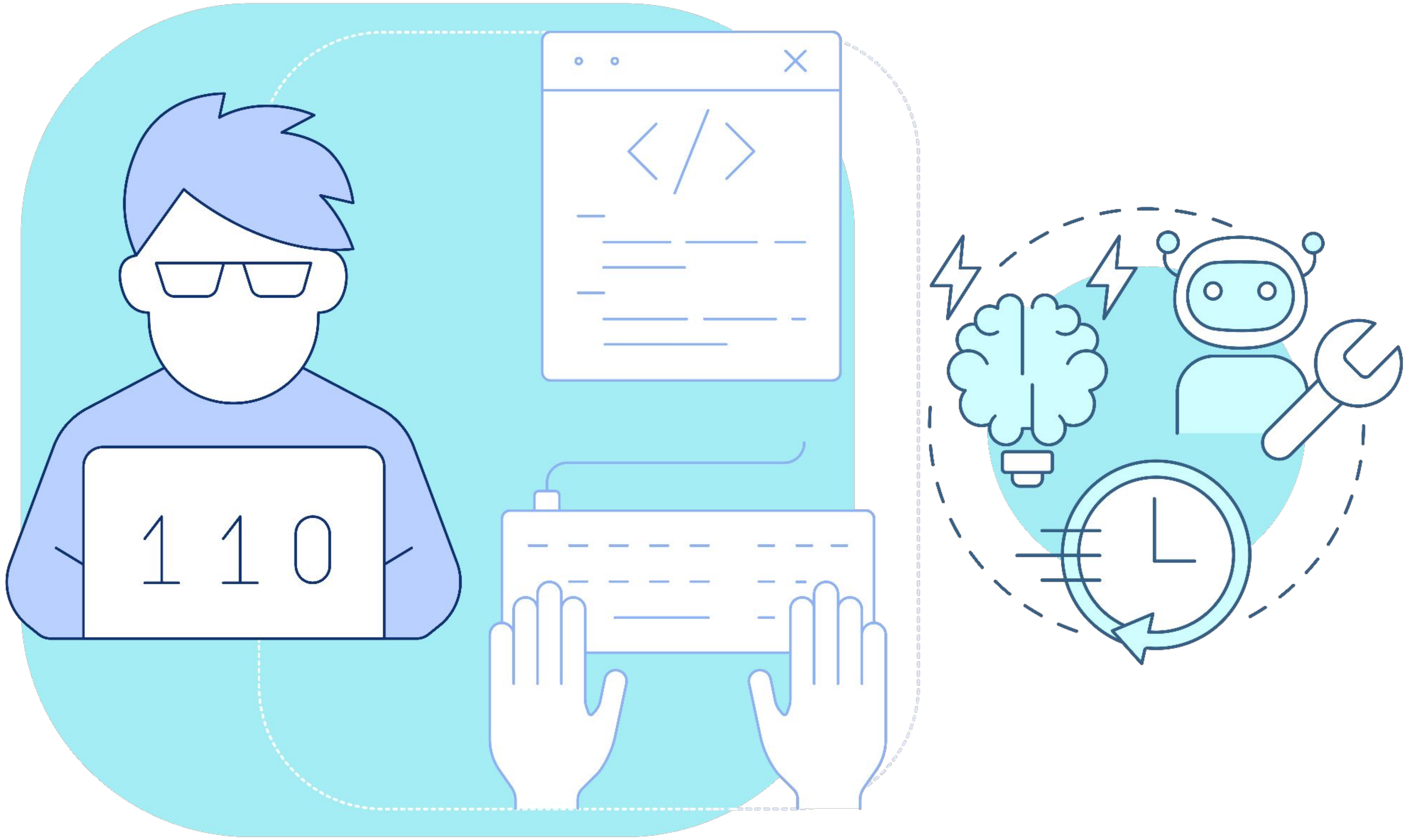


- 1950 — Turing Paper
- 1966 — ELIZA (chatbot)
- 2017 — Transformer Architecture (Google)
- 2018 — GPT-1
- 2019 — GPT-2, BERT
- 2020 — GPT-3
- 2021 — BERT 480B
- 2022 — GPT-3.5
- 2023 — LLM in mainstream (ChatGPT, BARD)

How LLMs differentiate with NLP and business use cases

LLMs and NLP can be used to create intelligent chatbots and virtual assistants that can help customers with their inquiries, provide support, and offer personalized recommendations.

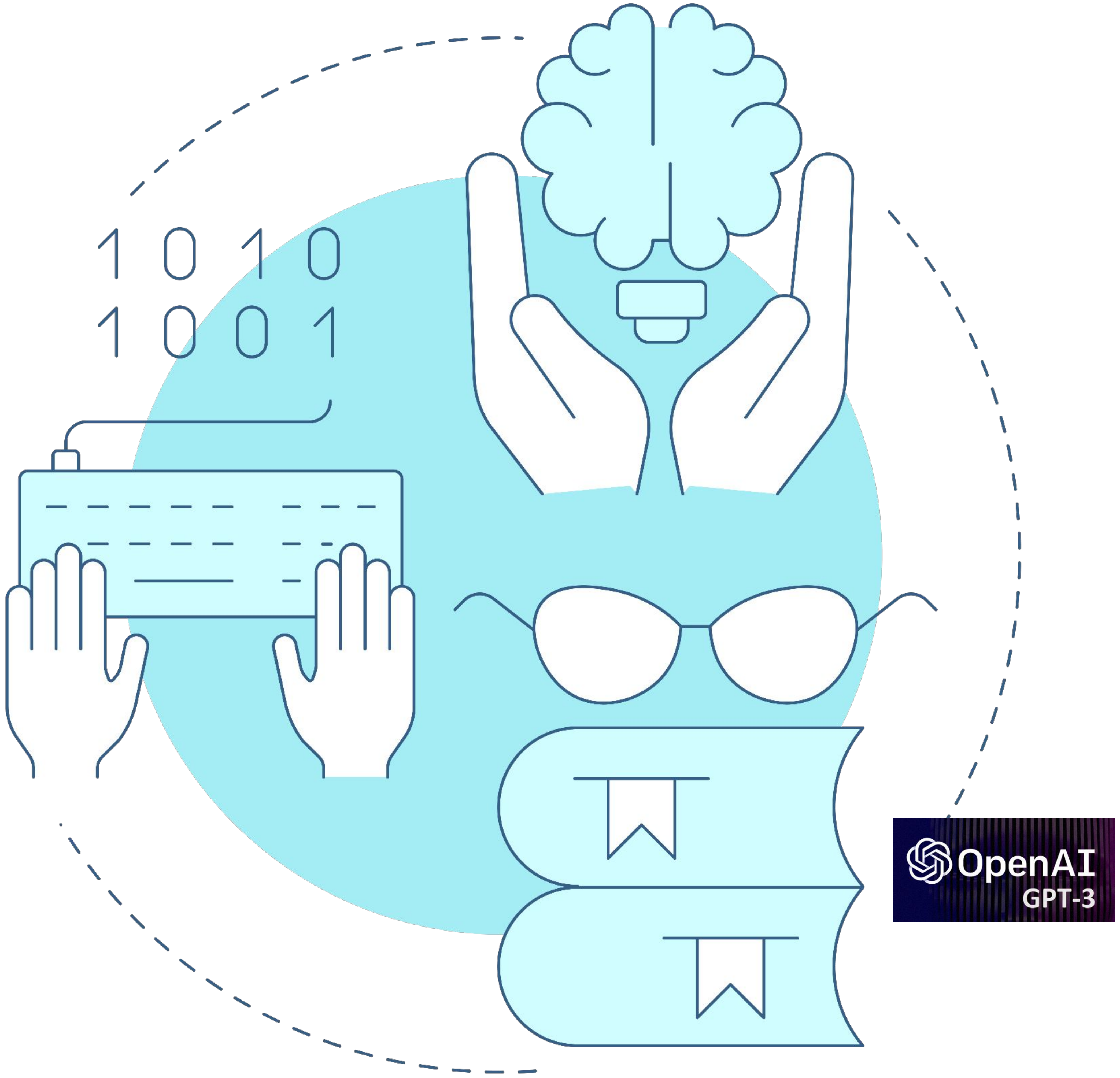
- Sentiment analysis
- Content creation and optimization
- Customer service automation
- Fraud detection
- Language translation
- Resume screening
- Personalization
- Market research
- Speech recognition



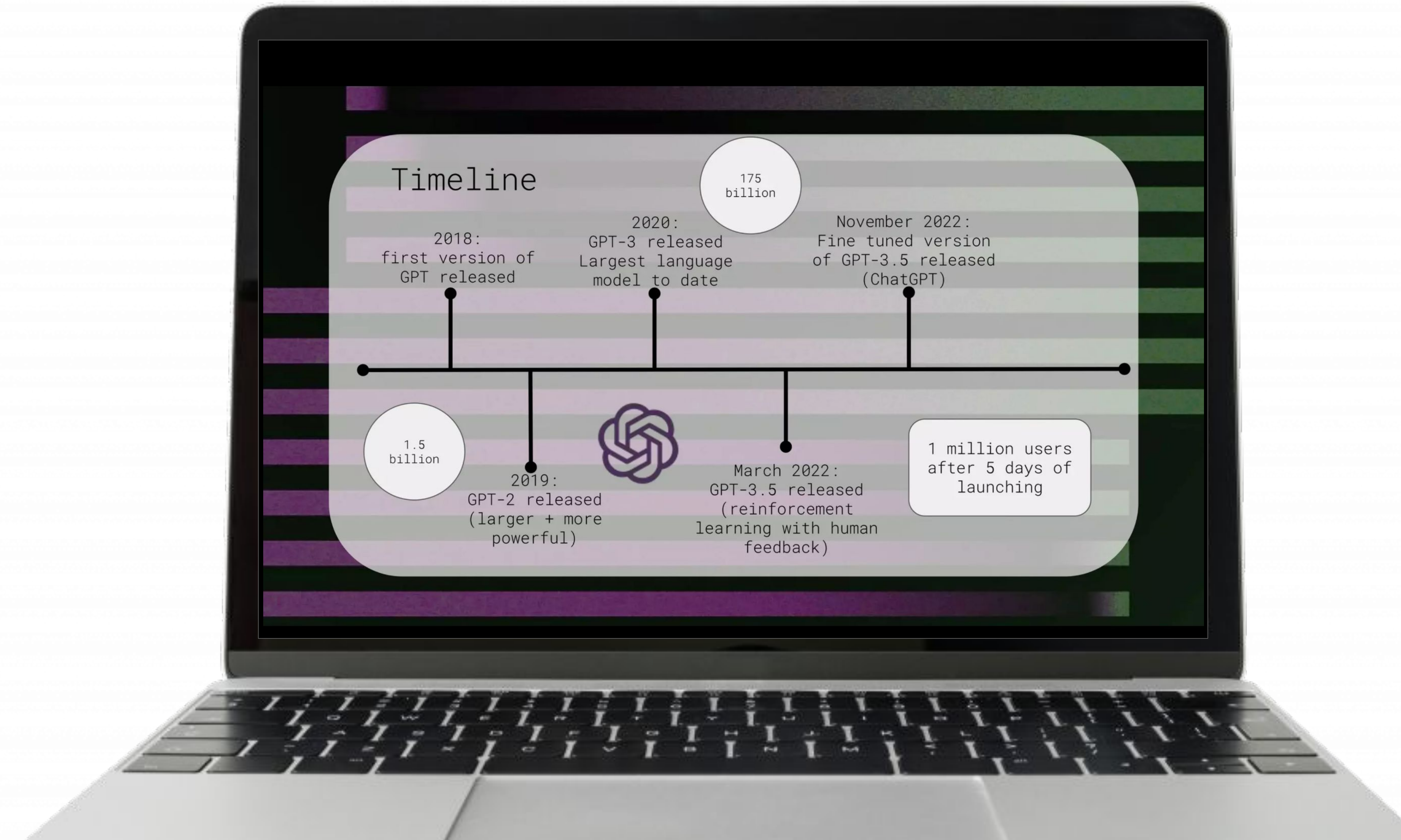
Introduction to GPT-3

- **Generative Pre-trained Transformer** 3 (GPT-3) is an **autoregressive language model** released in 2020 that uses deep learning to produce human-like text. Given an initial text as prompt, it will produce text that continues the prompt.

- Authored by OpenAI



ChatGPT: Introduction and Timeline



Timeline

2018:
first version of
GPT released

1.5
billion

2019:
GPT-2 released
(larger + more
powerful)

175
billion

2020:
GPT-3 released
Largest language
model to date

March 2022:
GPT-3.5 released
(reinforcement
learning with human
feedback)

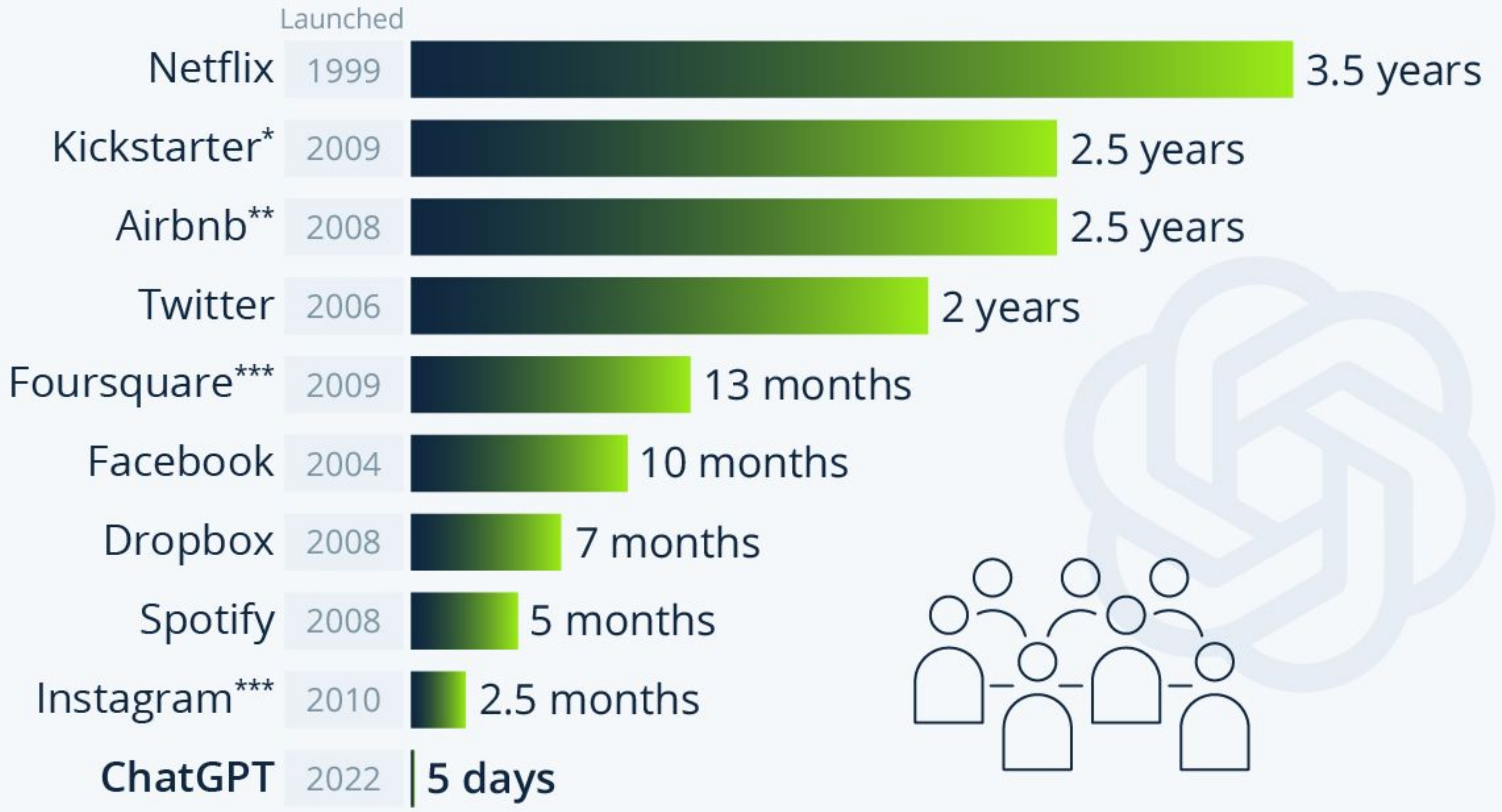
November 2022:
Fine tuned version
of GPT-3.5 released
(ChatGPT)

1 million users
after 5 days of
launching

ChatGPT: Users Timeline

ChatGPT Sprints to One Million Users

Time it took for selected online services to reach one million users

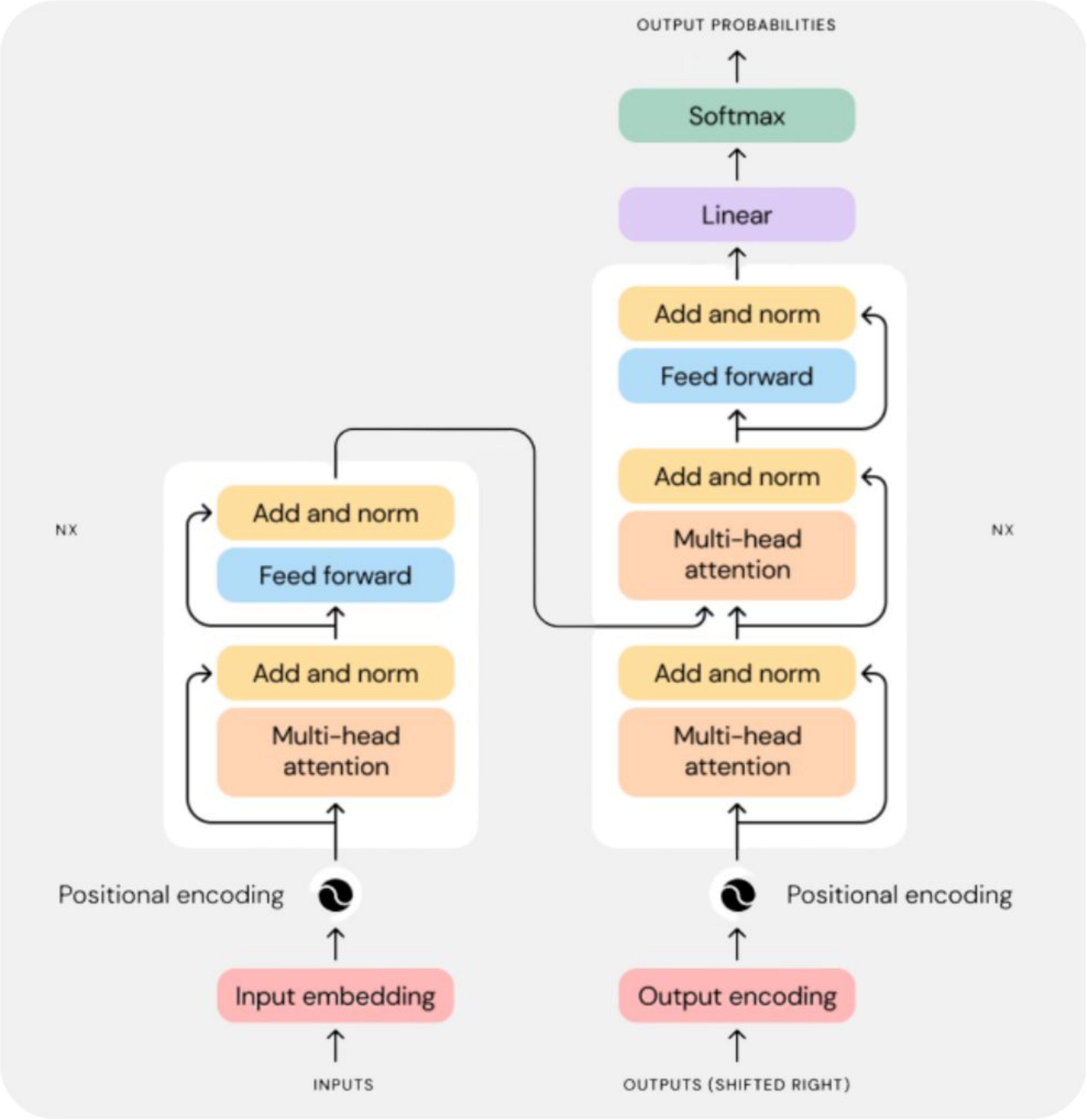


* one million backers ** one million nights booked *** one million downloads

Source: Company announcements via Business Insider/Linkedin

Transformer Architecture (GPT-3 is based on this)

Accionlabs



Architecture of GPT-3

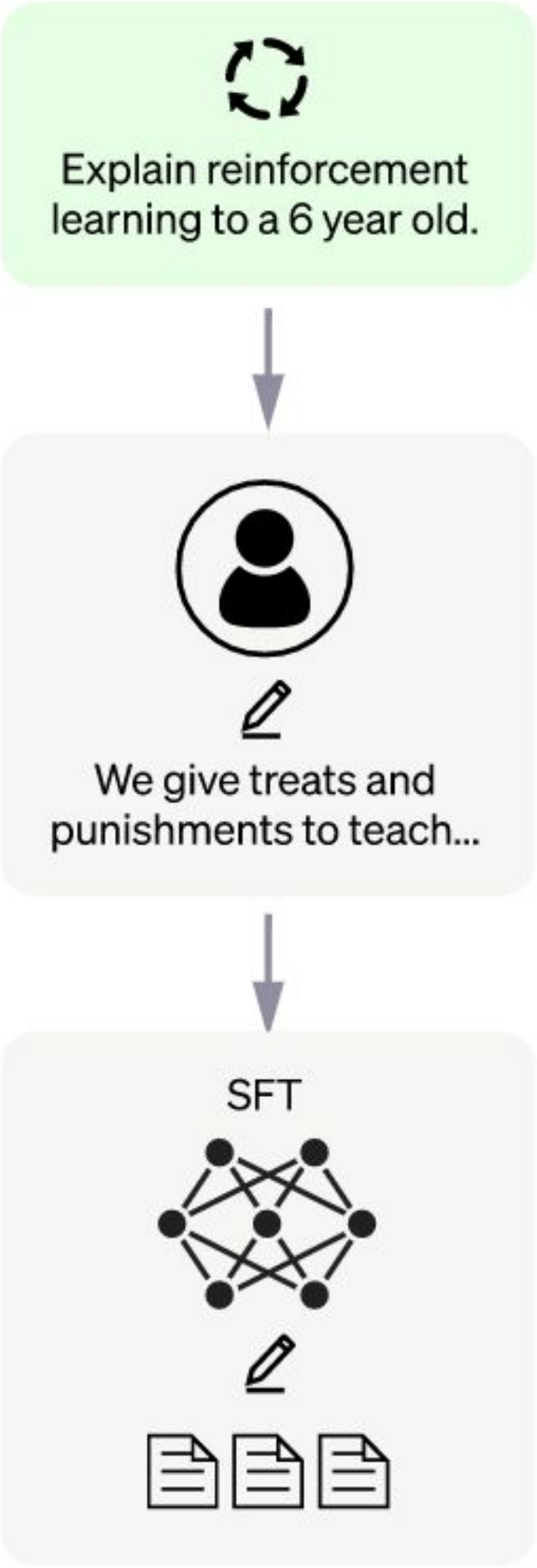
Step 1

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.


Explain reinforcement learning to a 6 year old.
We give treats and punishments to teach...
SFT

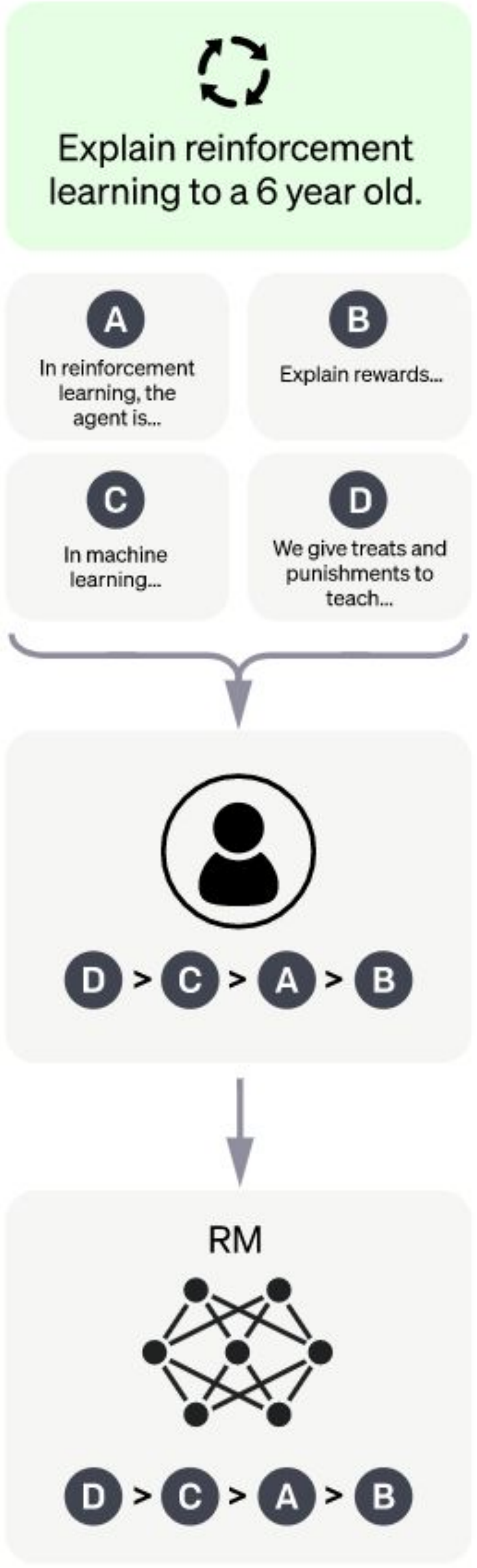
Step 2

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.


Explain reinforcement learning to a 6 year old.
A
In reinforcement learning, the agent is...
B
Explain rewards...
C
In machine learning...
D
We give treats and punishments to teach...
D > C > A > B
RM
D > C > A > B

Step 3

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

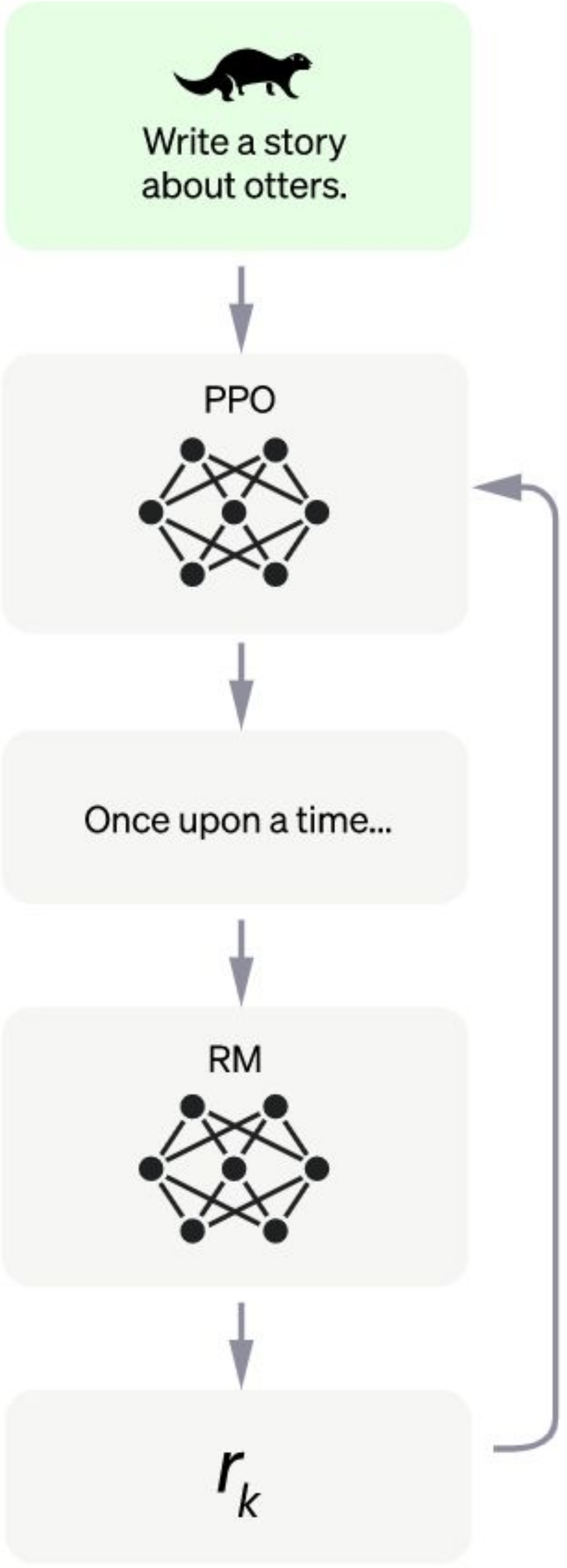
A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

The policy generates an output.

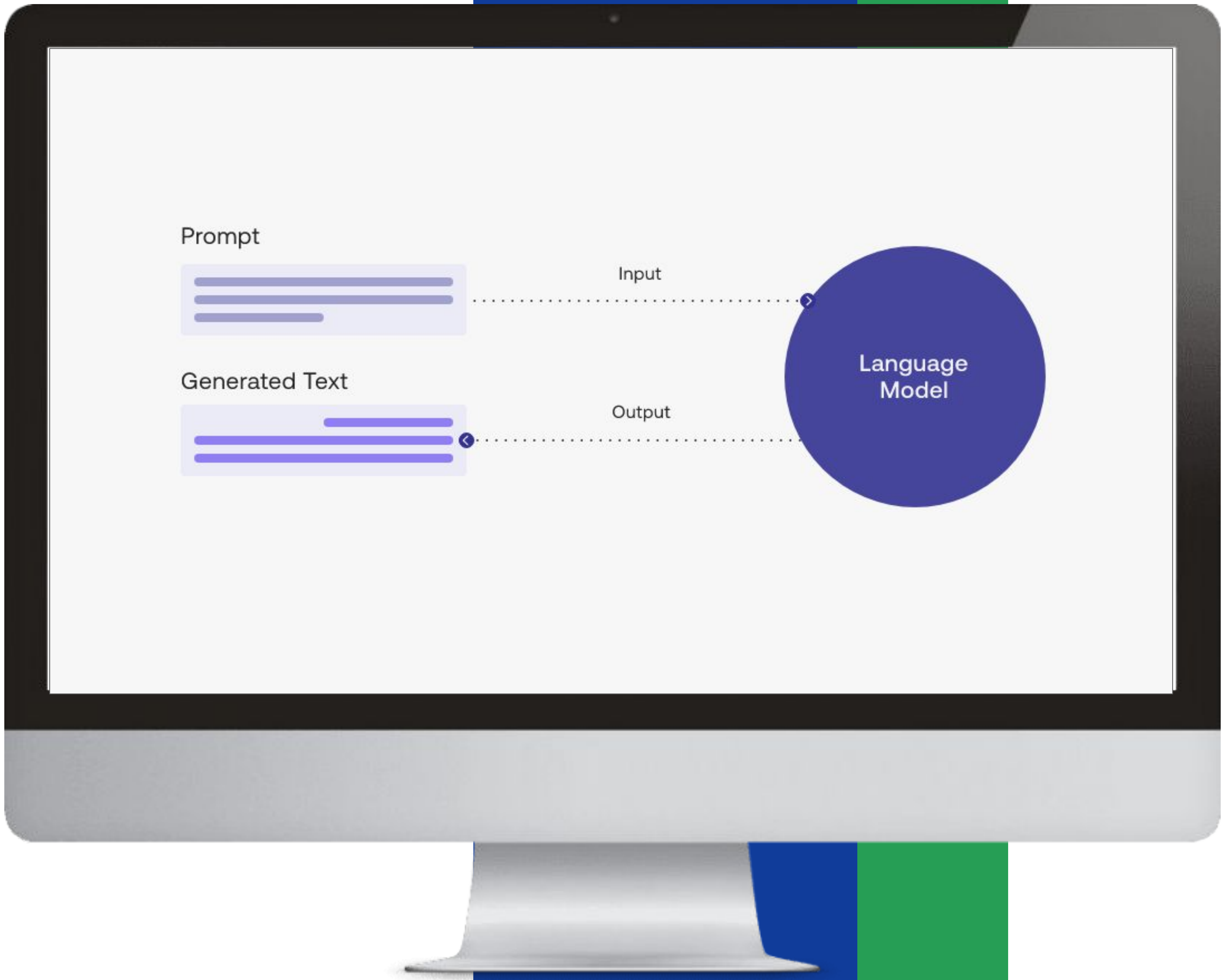
The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

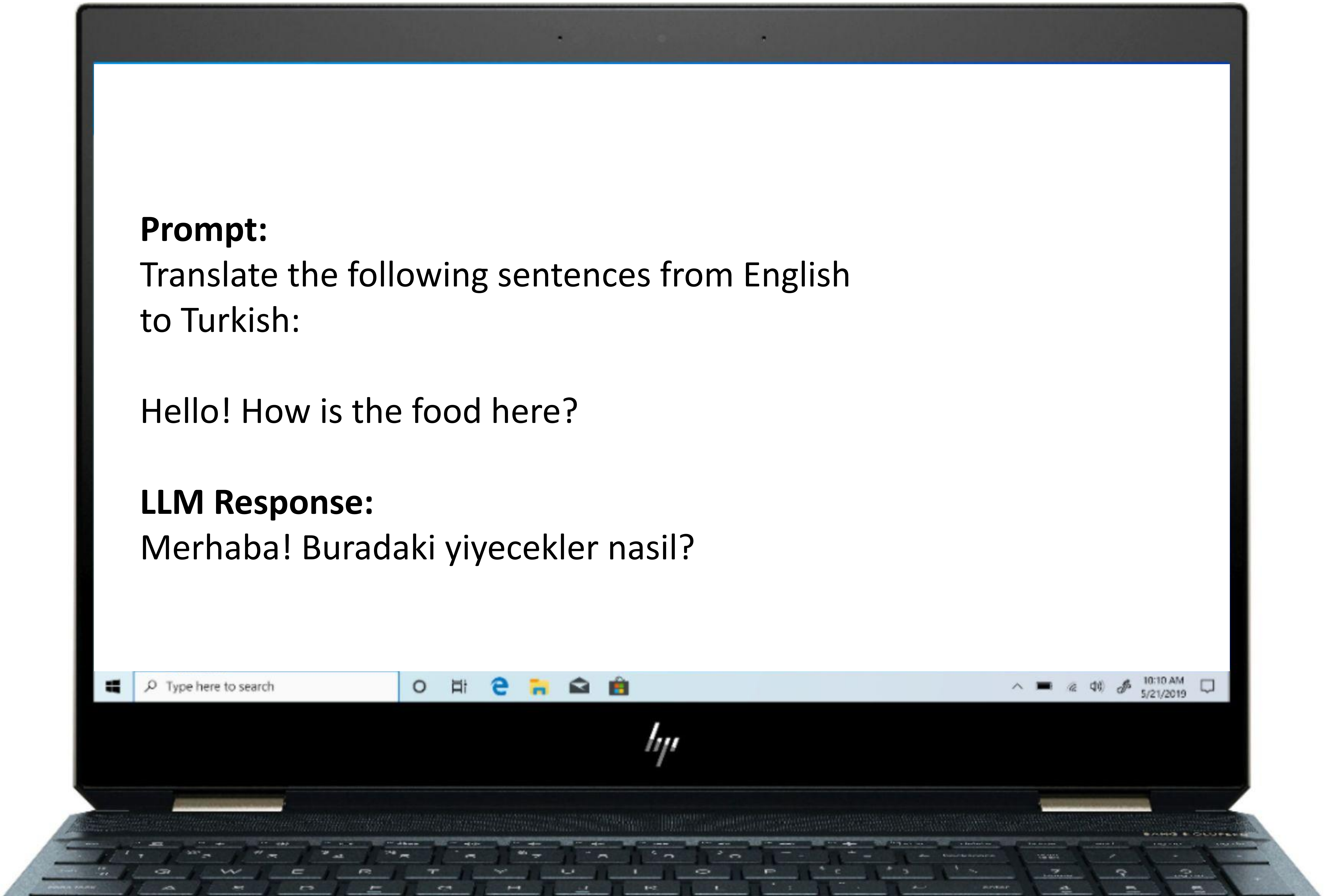


Prompt Engineering

- Prompt engineering is a NLP concept that involves discovering inputs that yield desirable or useful results (or) How to talk to AI to get it to do what you want



Prompt Engineering Examples

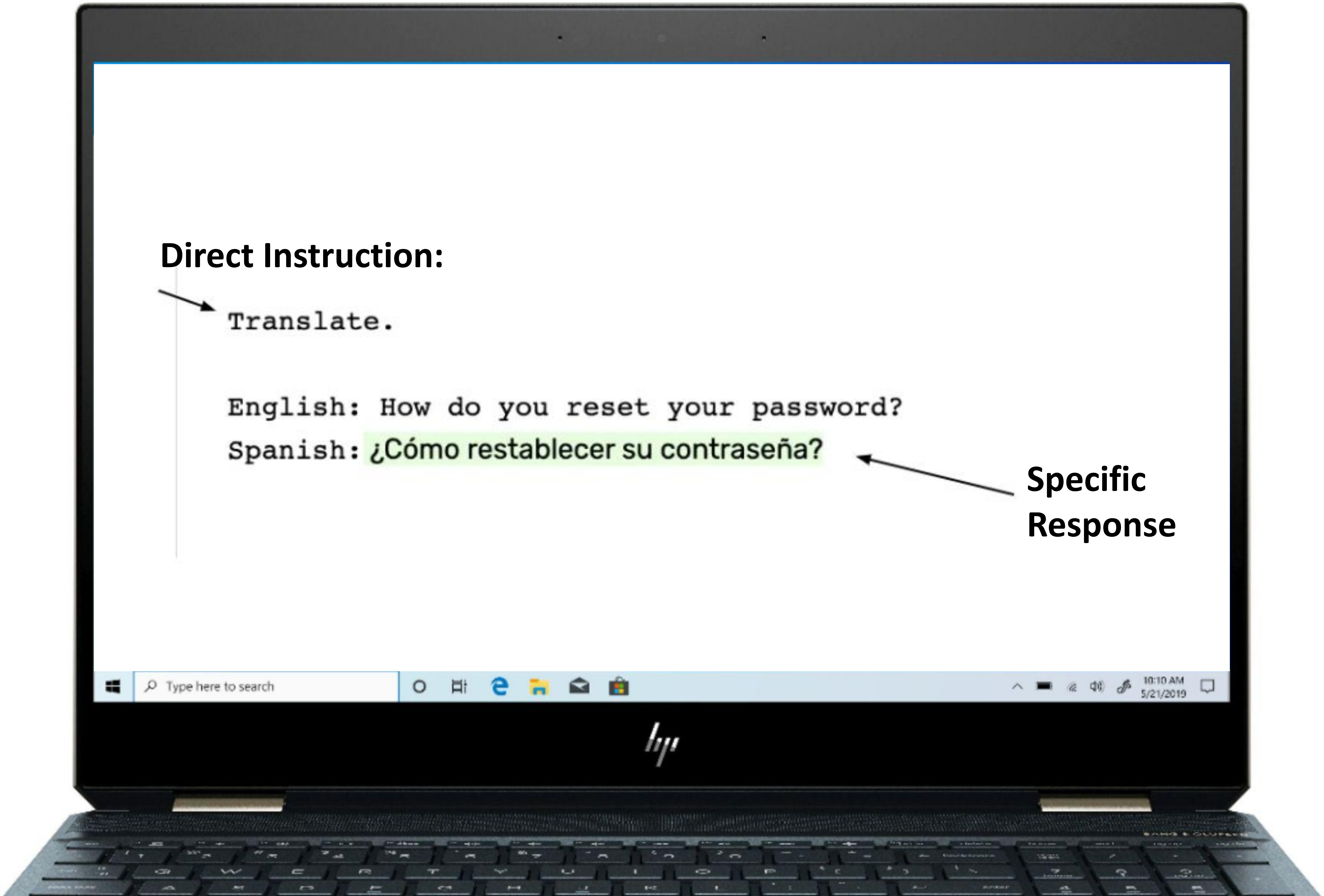


Prompt:
Translate the following sentences from English to Turkish:

Hello! How is the food here?

LLM Response:
Merhaba! Buradaki yiyecekler nasil?

Prompt Engineering Examples



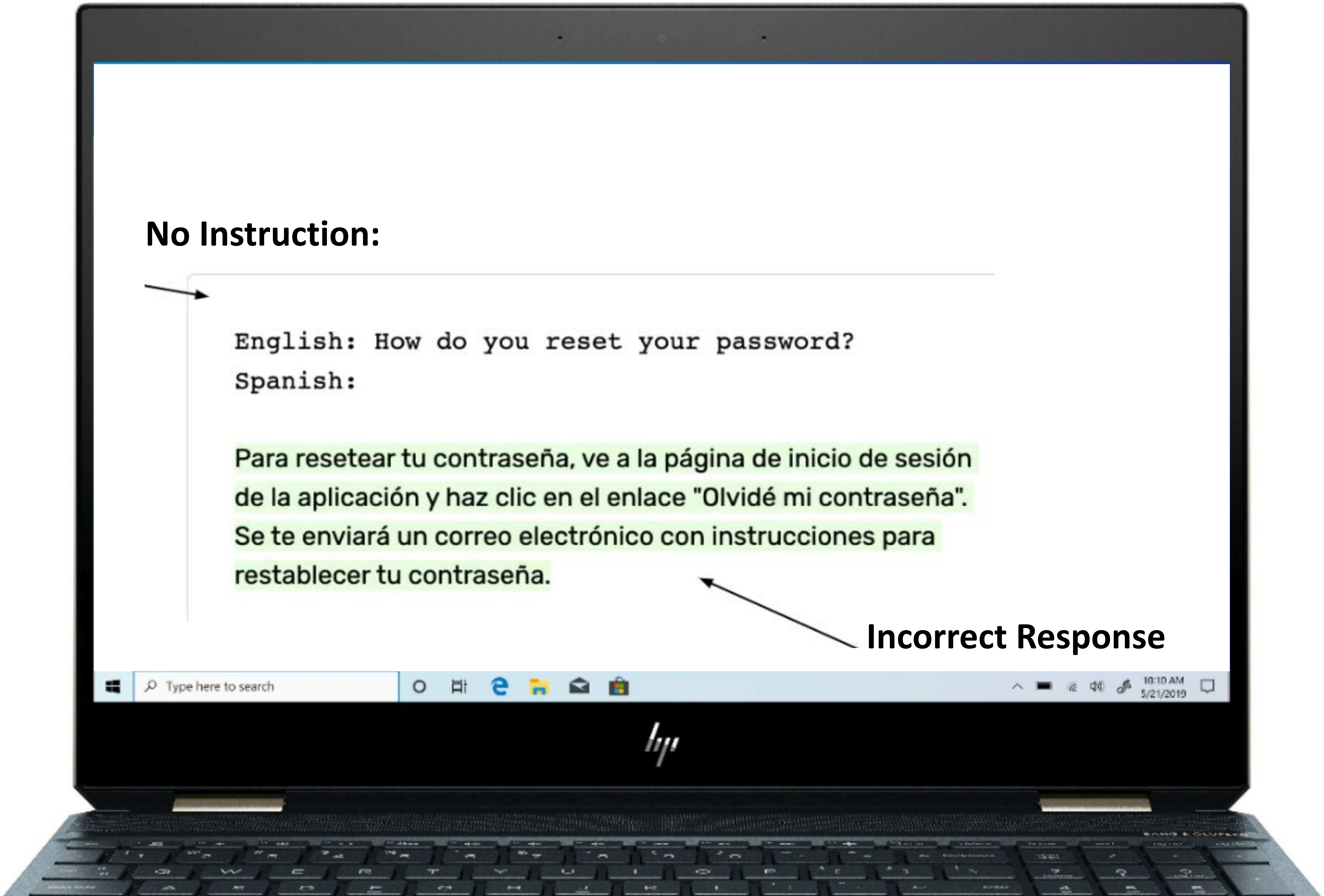
Direct Instruction:

```
Translate.

English: How do you reset your password?
Spanish: ¿Cómo restablecer su contraseña?
```

Specific Response

Prompt Engineering Examples



No Instruction:

```
English: How do you reset your password?
Spanish:
```

Para resetear tu contraseña, ve a la página de inicio de sesión de la aplicación y haz clic en el enlace "Olvidé mi contraseña". Se te enviará un correo electrónico con instrucciones para restablecer tu contraseña.

Incorrect Response

Prompt Engineering: Salient Features

01

Prompt is at the heart of working with LLMs

02

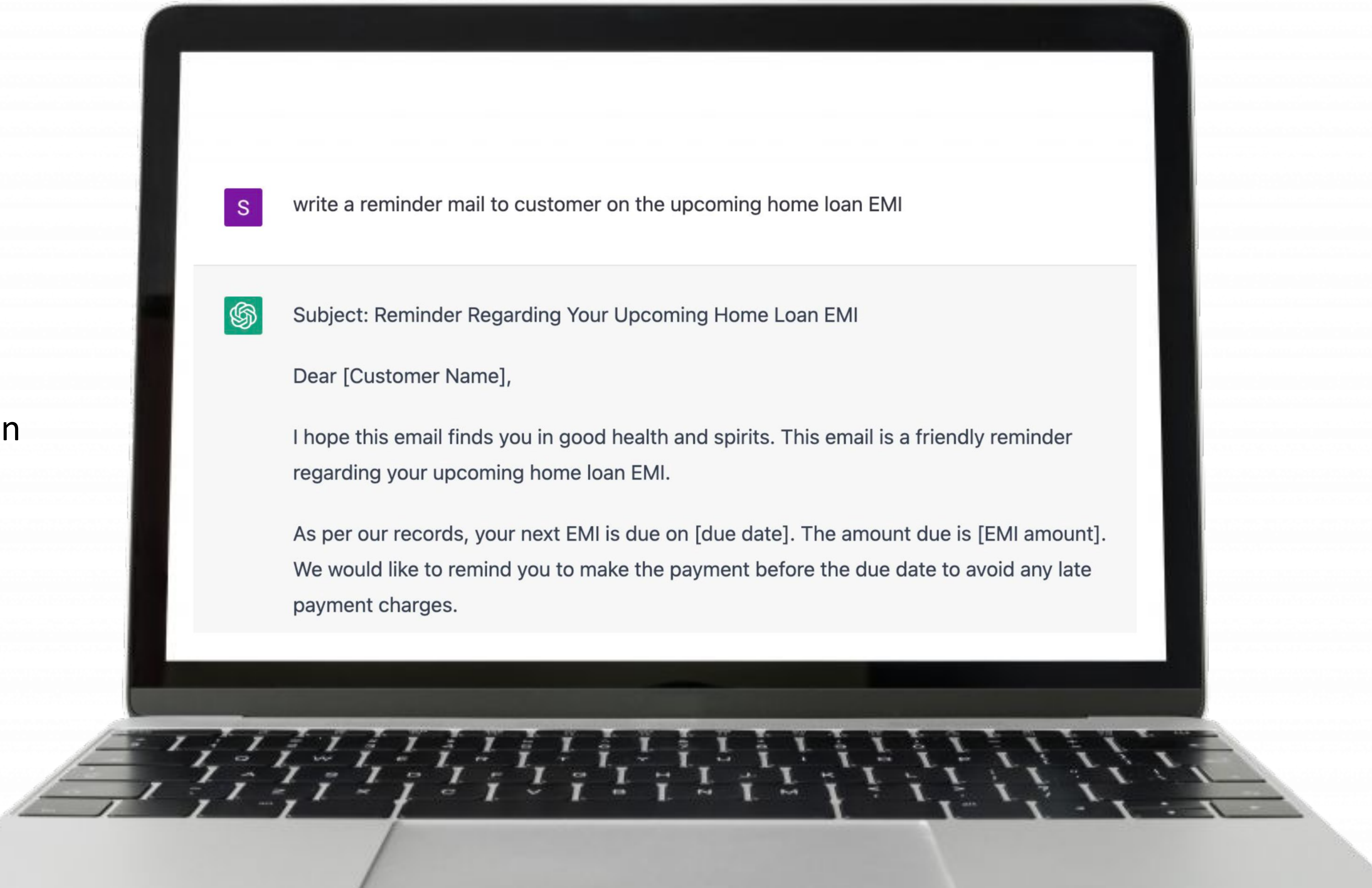
Prompts can be simple text to very complex instructions

03

Coming up with good prompt is a combination of art and science

Types of Prompting

1. Prompting by Instruction



S write a reminder mail to customer on the upcoming home loan EMI

Subject: Reminder Regarding Your Upcoming Home Loan EMI

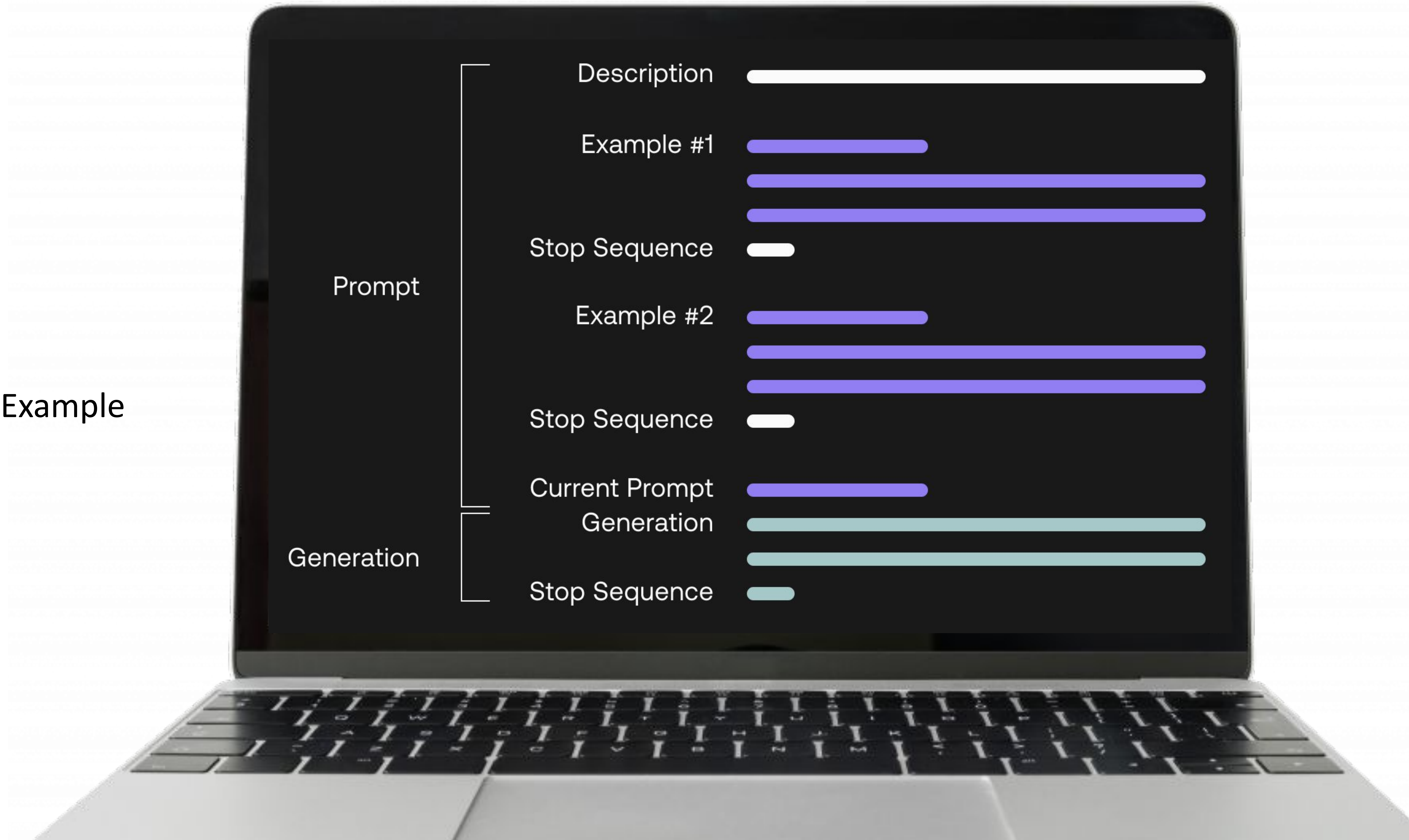
Dear [Customer Name],

I hope this email finds you in good health and spirits. This email is a friendly reminder regarding your upcoming home loan EMI.

As per our records, your next EMI is due on [due date]. The amount due is [EMI amount]. We would like to remind you to make the payment before the due date to avoid any late payment charges.

Types of Prompting

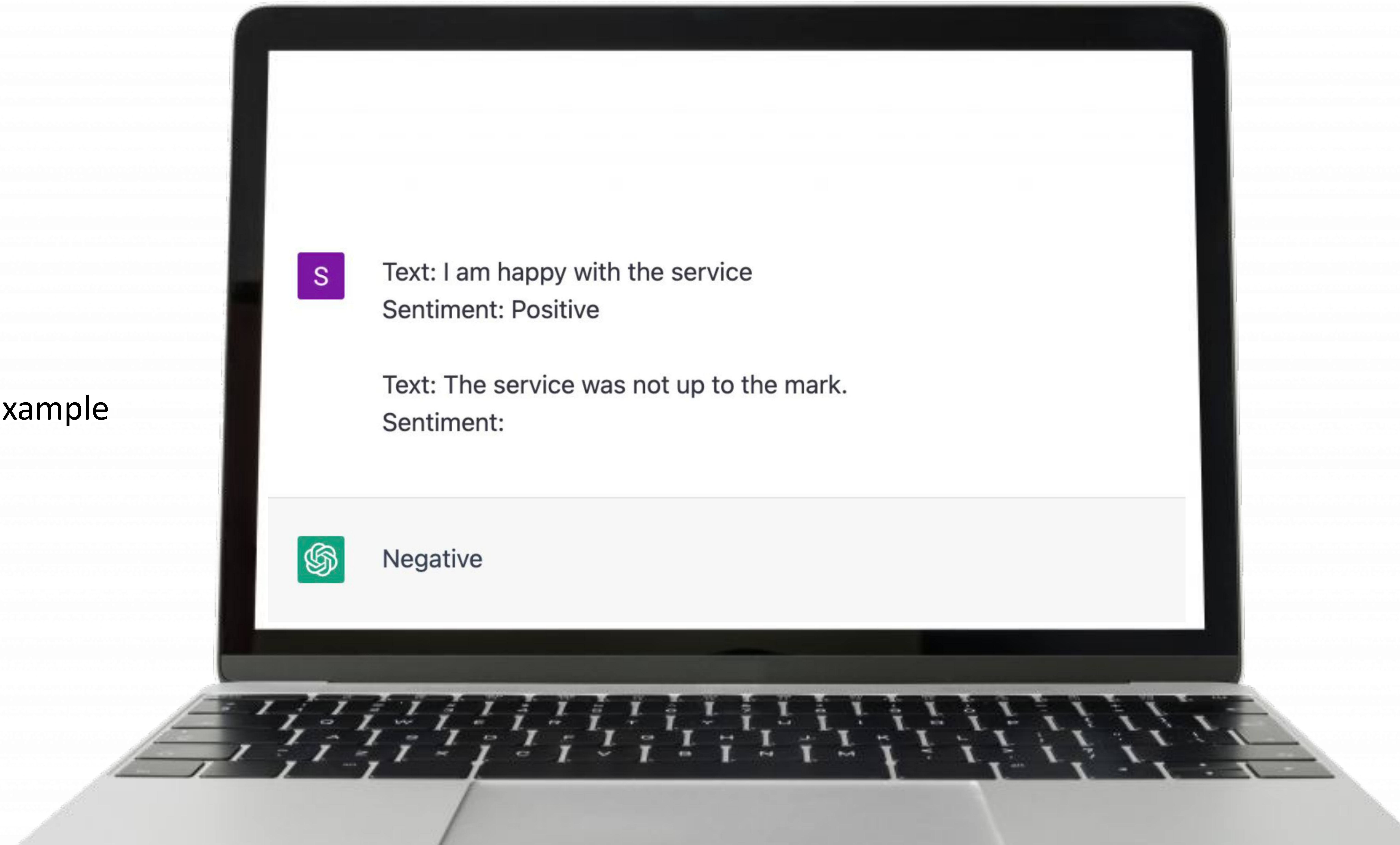
2. Prompting by Example



Types of Prompting

Accionlabs

2. Prompting by Example



S

Text: I am happy with the service
Sentiment: Positive

Text: The service was not up to the mark.
Sentiment:

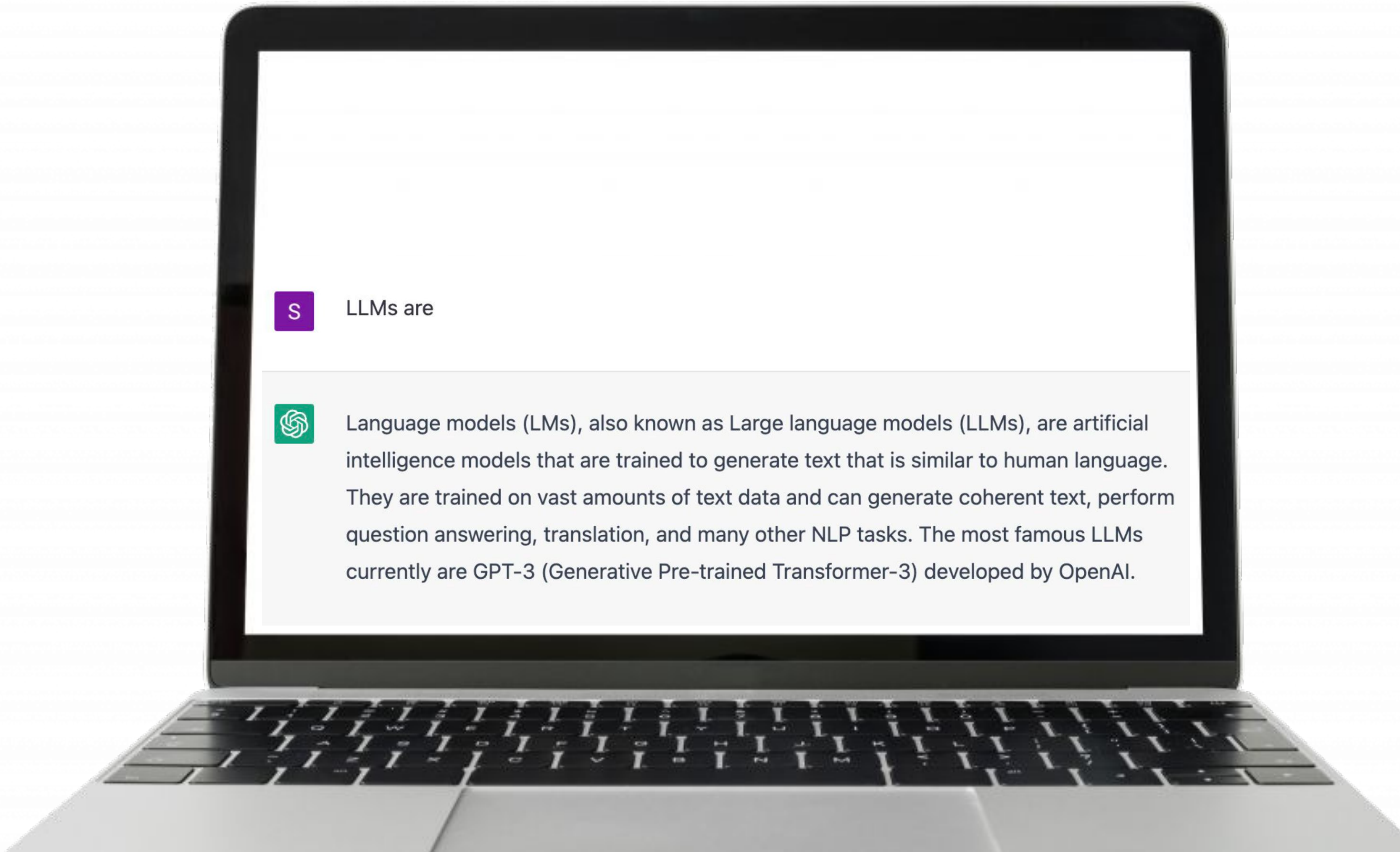
Negative

Accionlabs

ChatGPT: Use cases

Use Case - Text Generation

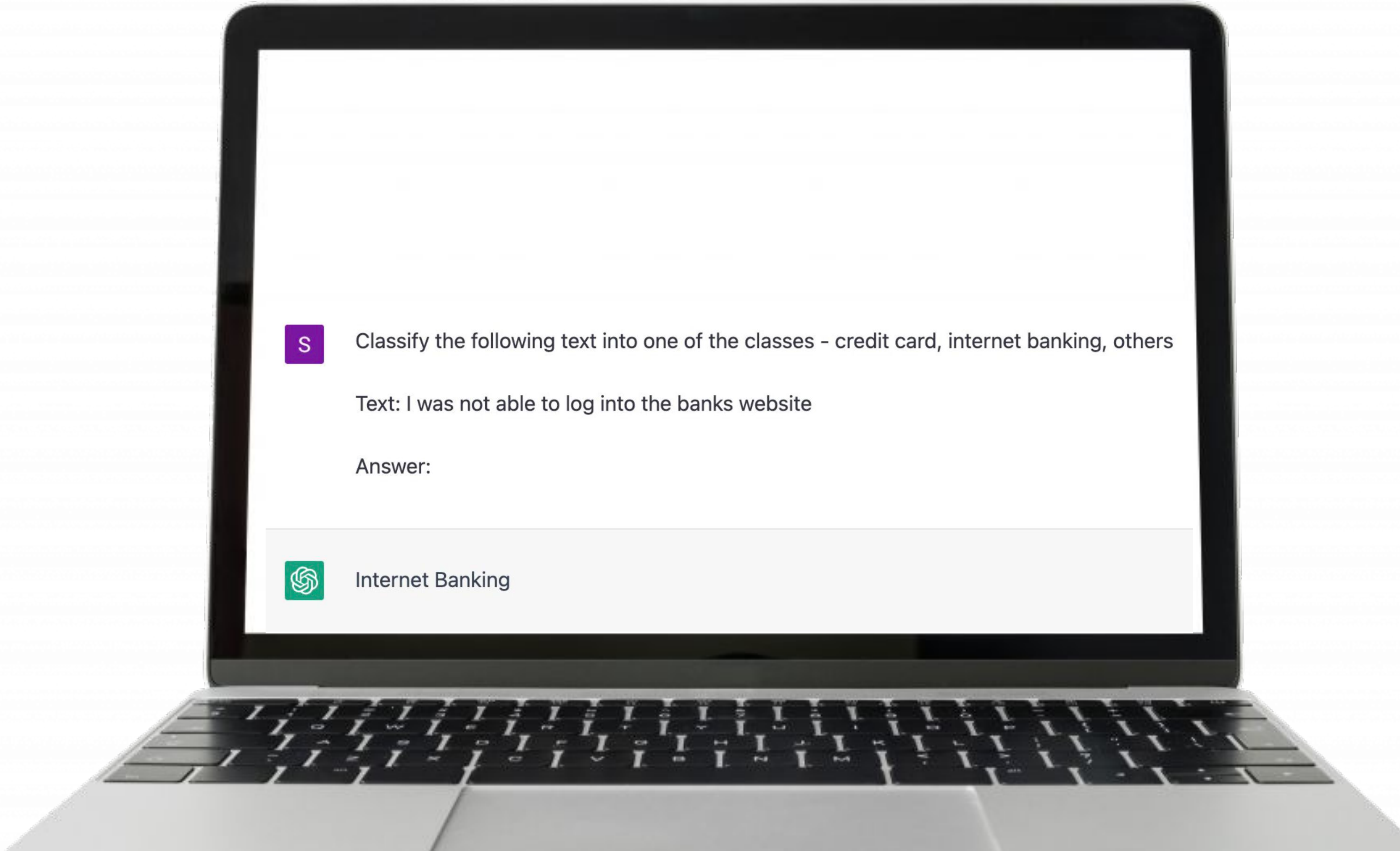
Accionlabs



S LLMs are

Language models (LMs), also known as Large language models (LLMs), are artificial intelligence models that are trained to generate text that is similar to human language. They are trained on vast amounts of text data and can generate coherent text, perform question answering, translation, and many other NLP tasks. The most famous LLMs currently are GPT-3 (Generative Pre-trained Transformer-3) developed by OpenAI.

Use Case - Text Classification



S

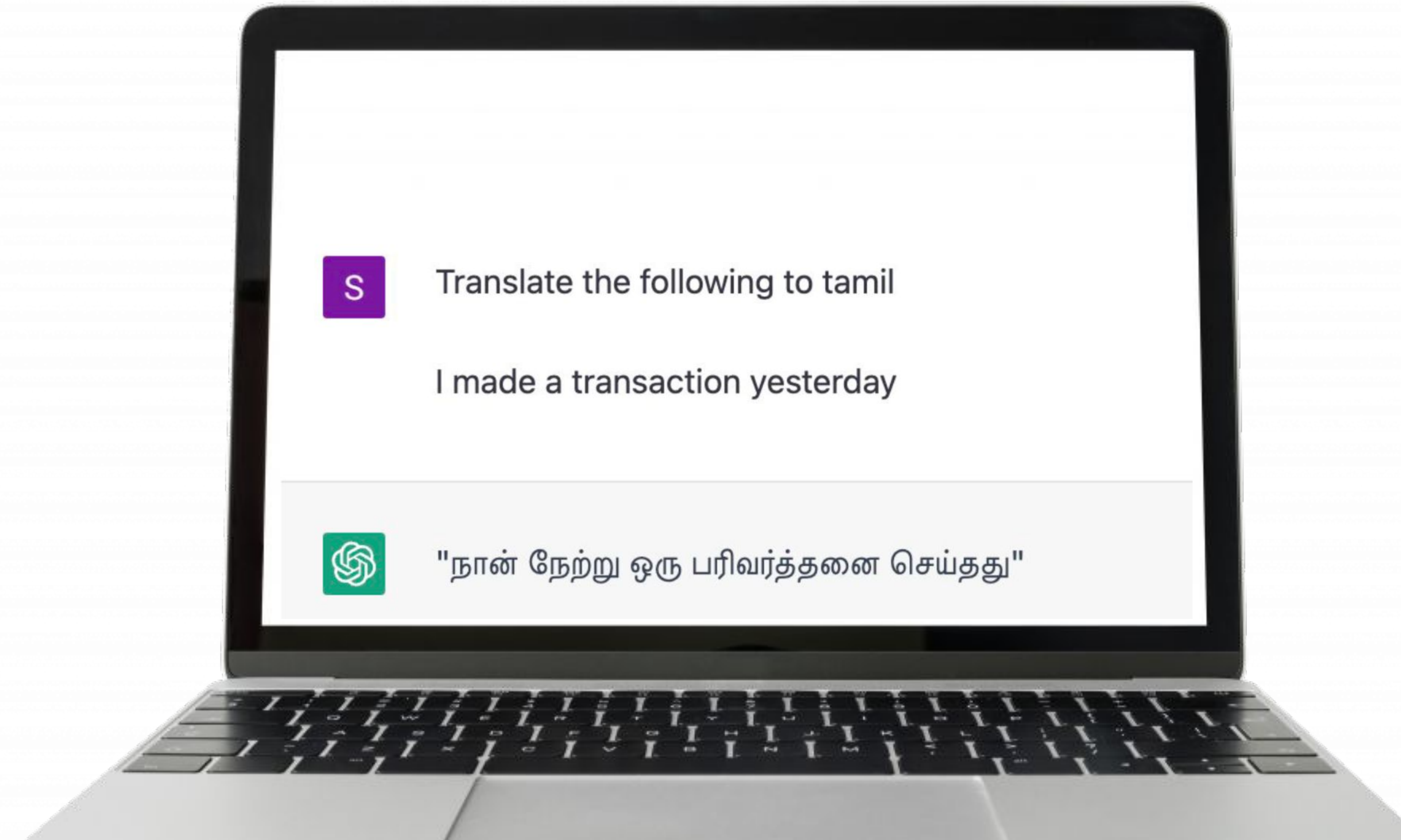
Classify the following text into one of the classes - credit card, internet banking, others

Text: I was not able to log into the banks website

Answer:

Internet Banking

Use Case - Language Translation

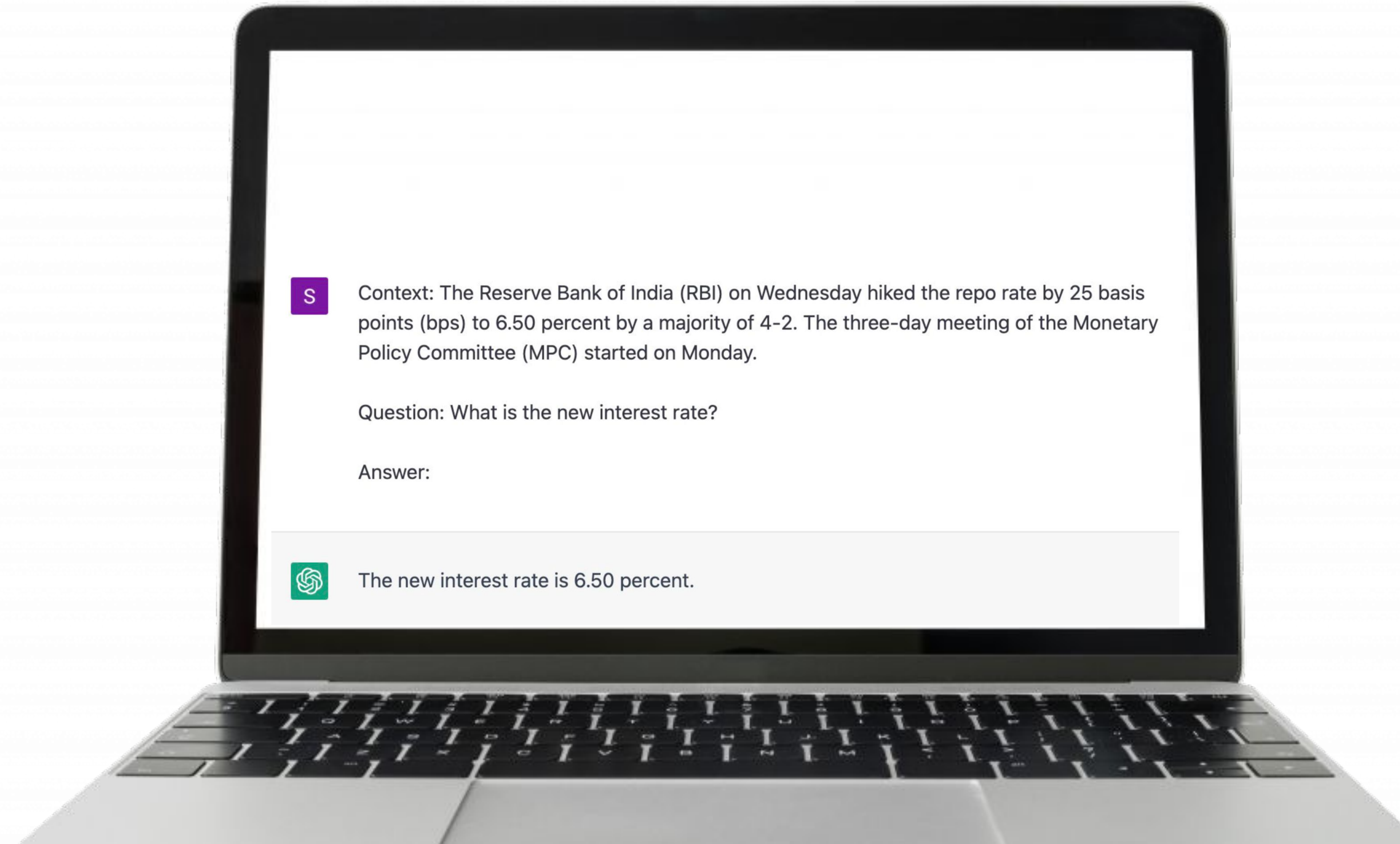


Translate the following to tamil

I made a transaction yesterday

"நான் நேற்று ஒரு பரிவர்த்தனை செய்தது"

Use Case - Question Answering



S

Context: The Reserve Bank of India (RBI) on Wednesday hiked the repo rate by 25 basis points (bps) to 6.50 percent by a majority of 4-2. The three-day meeting of the Monetary Policy Committee (MPC) started on Monday.

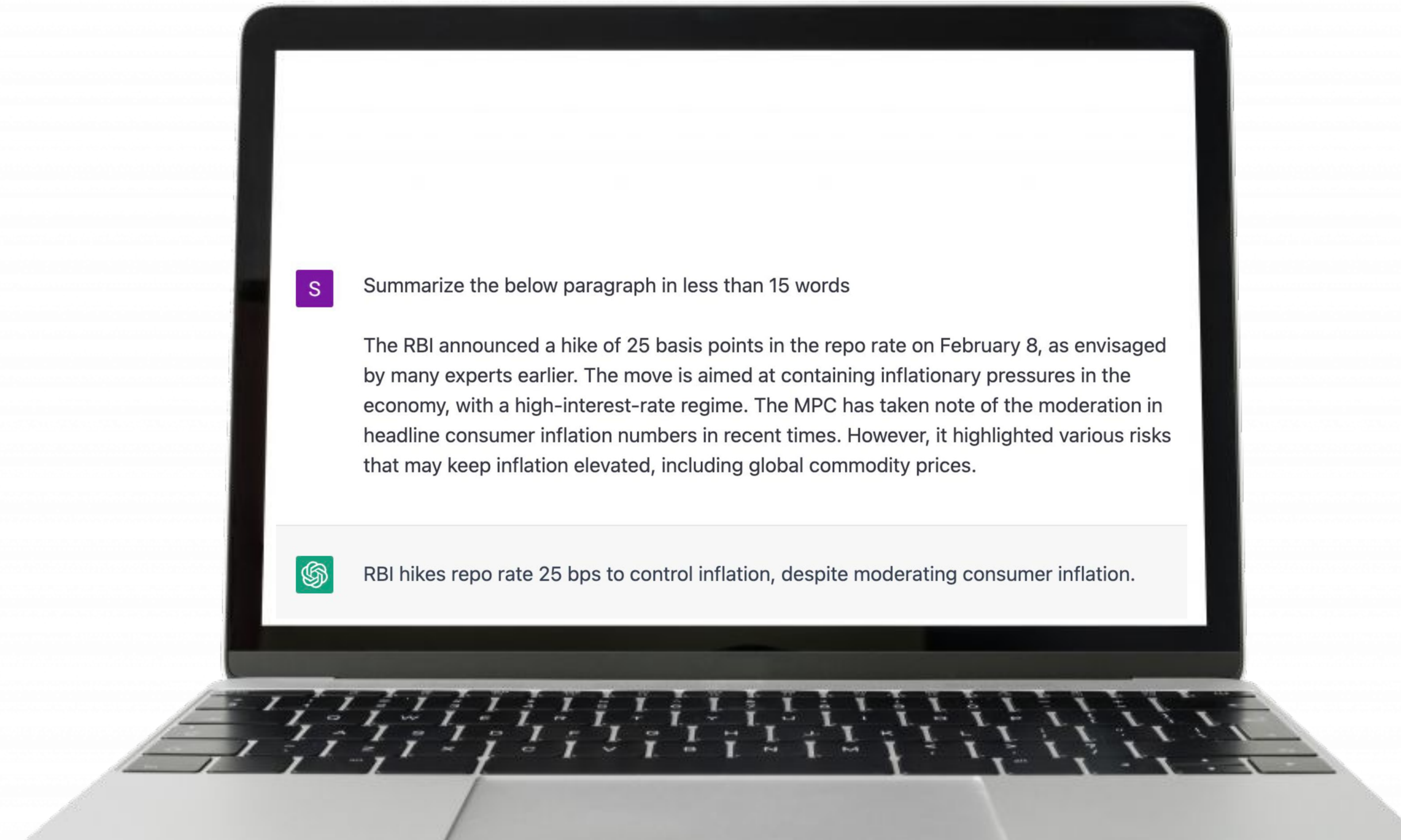
Question: What is the new interest rate?

Answer:

The new interest rate is 6.50 percent.

Use case: Text Summarization

Accionlabs



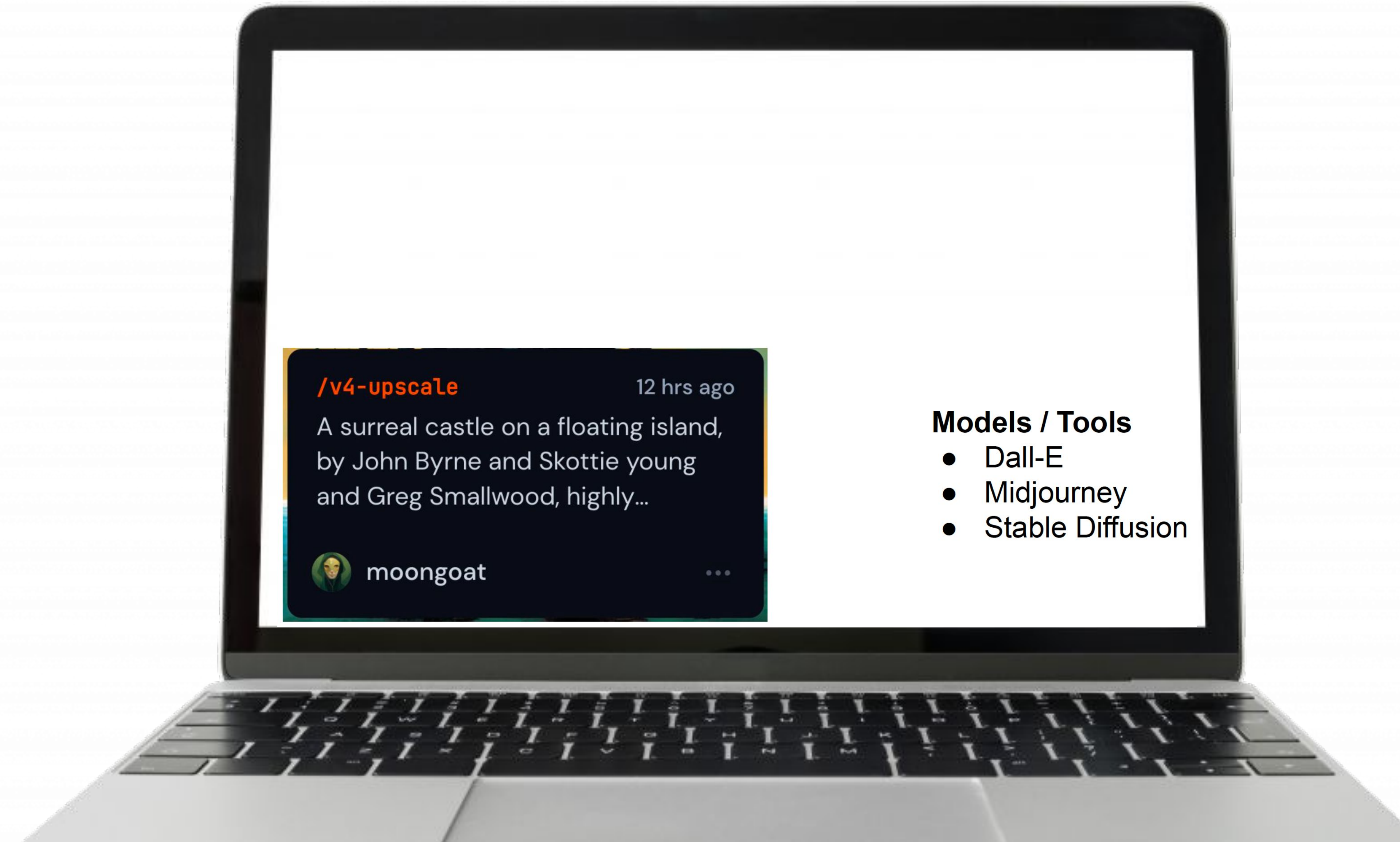
S

Summarize the below paragraph in less than 15 words

The RBI announced a hike of 25 basis points in the repo rate on February 8, as envisaged by many experts earlier. The move is aimed at containing inflationary pressures in the economy, with a high-interest-rate regime. The MPC has taken note of the moderation in headline consumer inflation numbers in recent times. However, it highlighted various risks that may keep inflation elevated, including global commodity prices.

RBI hikes repo rate 25 bps to control inflation, despite moderating consumer inflation.

Use case: Image Generation



/v4-upscale 12 hrs ago

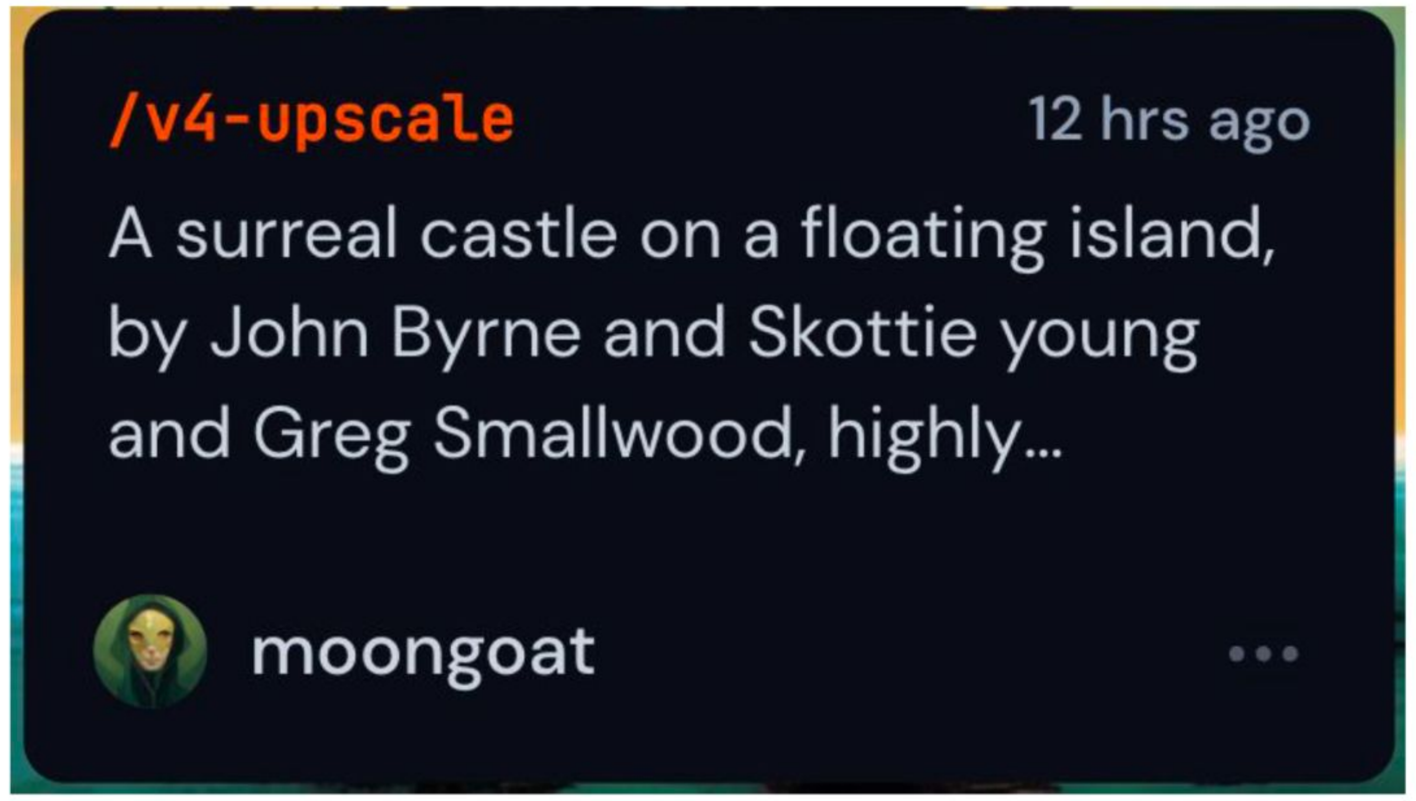
A surreal castle on a floating island, by John Byrne and Skottie young and Greg Smallwood, highly...

moongoat •••

Models / Tools

- Dall-E
- Midjourney
- Stable Diffusion

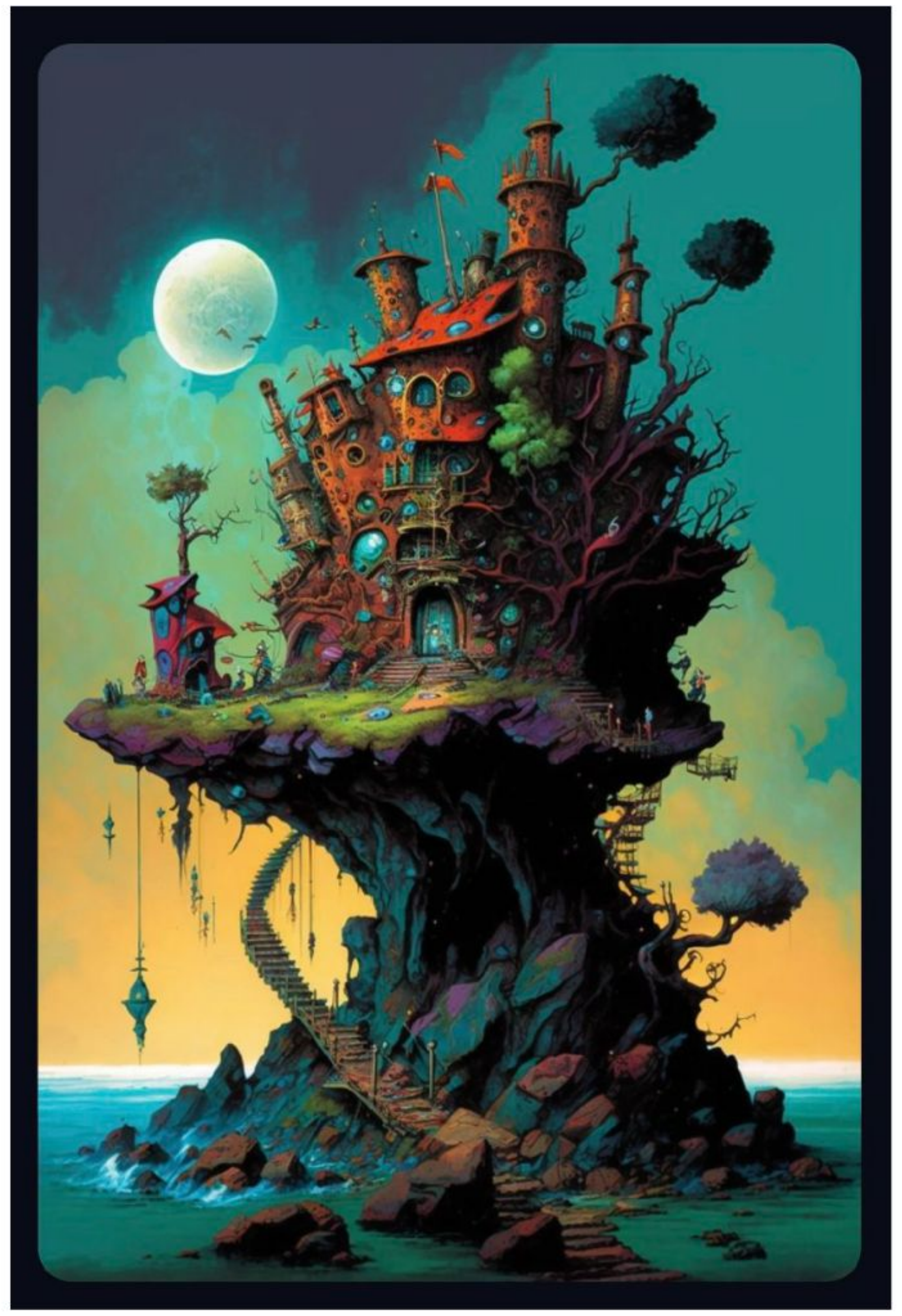
Use case: Image Generation



/v4-upscale 12 hrs ago

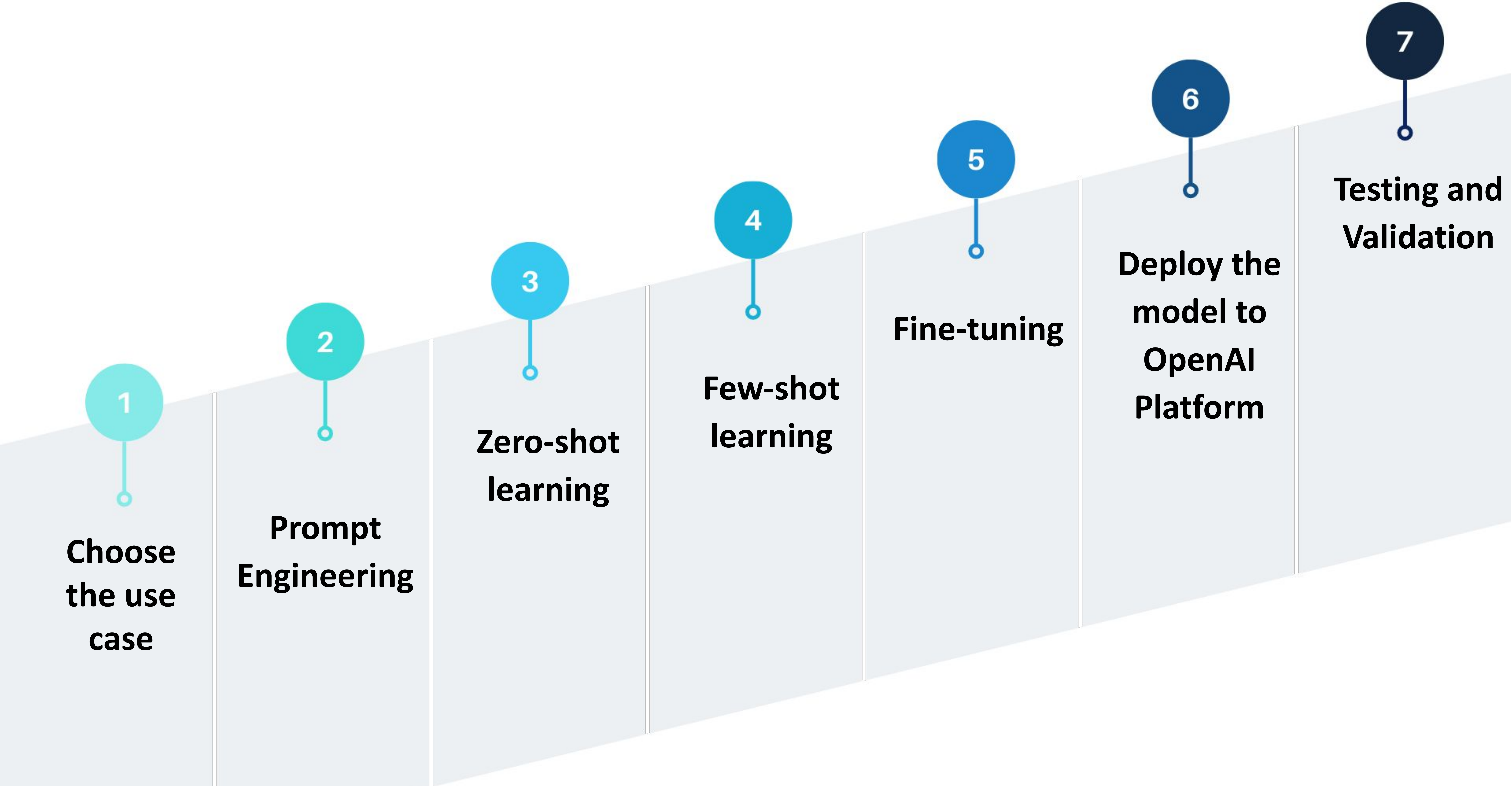
A surreal castle on a floating island, by John Byrne and Skottie young and Greg Smallwood, highly...

moongoat •••



Product Development and Integration life cycle with GPT-3

Accionlabs



1 Choose the use case

2 Prompt Engineering

3 Zero-shot learning

4 Few-shot learning

5 Fine-tuning

6 Deploy the model to OpenAI Platform

7 Testing and Validation

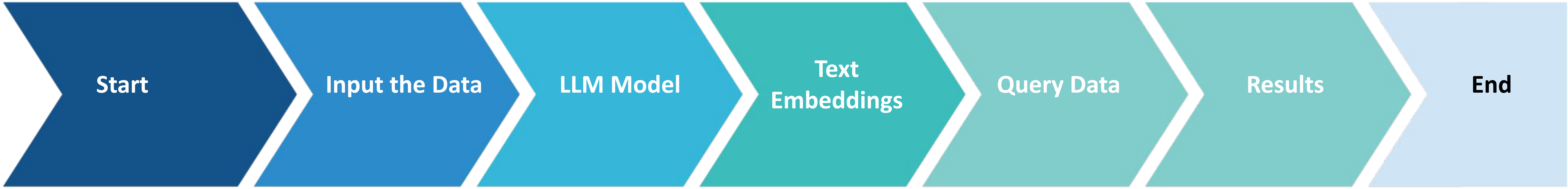
Business use cases of LLMs

Accionlabs

- Low-code development
- Predictive maintenance
- Knowledge management
- Medical diagnosis and treatment
- Fraud prevention
- Personalized education
- Natural language programming
- Automated document processing
- Cybersecurity
- Virtual assistants for employees

What type of requirement can qualify for LLMs/ ChatGPT?

Accionlabs



1. Documents
2. Communication Chat
3. Webpages
4. Transcript

1. Question-Answering
2. Natural Language Query
3. Text Completion
4. Sentiment Analysis
5. Text Summarization
6. Named Entity Recognition
7. Text Classification

New Tech Debt Item(s)

Pricing of Language Models(LLMs) from OpenAI.

Accionlabs

Model	Type	Training Cost	Usage Cost
GPT-3.5-turbo	ChatGPT	N/A	$0.002 / 1K tokens
ADA	Fine-tuning with Custom Data	$0.0004 / 1K tokens	$0.0016 / 1K tokens
Babbage	Fine-tuning with Custom Data	$0.0006 / 1K tokens	$0.0024 / 1K tokens
Curie	Fine-tuning with Custom Data	$0.0030 / 1K tokens	$0.0120 / 1K tokens
Davinci	Fine-tuning with Custom Data	$0.0300 / 1K tokens	$0.1200 / 1K tokens
ADA	**Embedding model:** Advanced search, clustering, topic modeling, and classification functionality	N/A	$0.0008 / 1K tokens

Prices are per 1,000 tokens. You can think of tokens as pieces of words, where 1,000 tokens is about 750 words. This paragraph is 35 tokens.

Questions?

Accionlabs

Accion INNOVATION SUMMIT 2023

Thank you.

Email: chetan.khatri@accionlabs.com