

# Generative AI — An Introduction for Beginners

## IIJ Research Laboratory セミナー



2024年11月19日

株式会社インターネットイニシアティブ  
技術研究所  
Dimitrios GIAKATOS

# What is Generative AI?

## Image generation

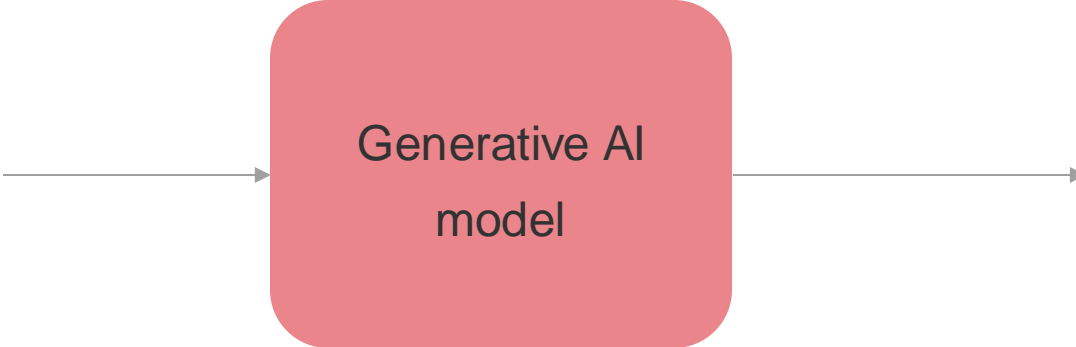
A drawing of a cat  
playing violin in front  
of Tokyo Tower

Generative AI  
model



## Text generation

I am ill. Write a letter to my teacher about me missing a class today.



Generative AI model

I hope this message finds you well. I am writing to inform you that I am unable to attend class today due to illness. I have been feeling unwell and believe it is best for my health and the well-being of my classmates to stay home and recover.

- 1. History**
- 2. Generative AI in action**
- 3. How it works**
- 4. Hosting your own LLM**
- 5. How can you use an LLM for your use case?**

- **Artificial Intelligence (1956)**

Introduced for first time in Dartmouth Conference. Researchers aimed to develop machines capable of simulating human cognitive functions.



*Credit: This week in The History of AI at AIWS.net – the Dartmouth Conference began on 18 June 1956*

Artificial Intelligence

- **Machine Learning (1997)**

Focusing on algorithms that enable computers to learn from data. A notable achievement was IBM's Deep Blue defeating chess champion Garry Kasparov.



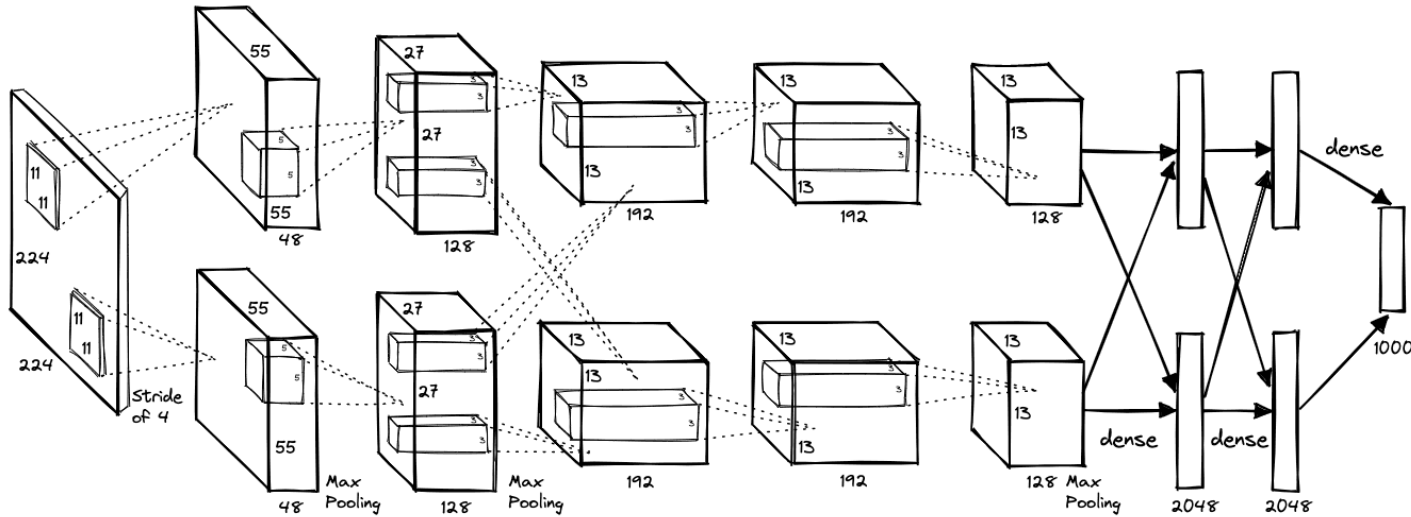
*Credit: World chess champion Garry Kasparov (left) playing against IBM's supercomputer Deep Blue in 1996 during the ACM Chess Challenge in Philadelphia. Photo: Tom Mihalek/AFP/Getty Images*

Artificial Intelligence

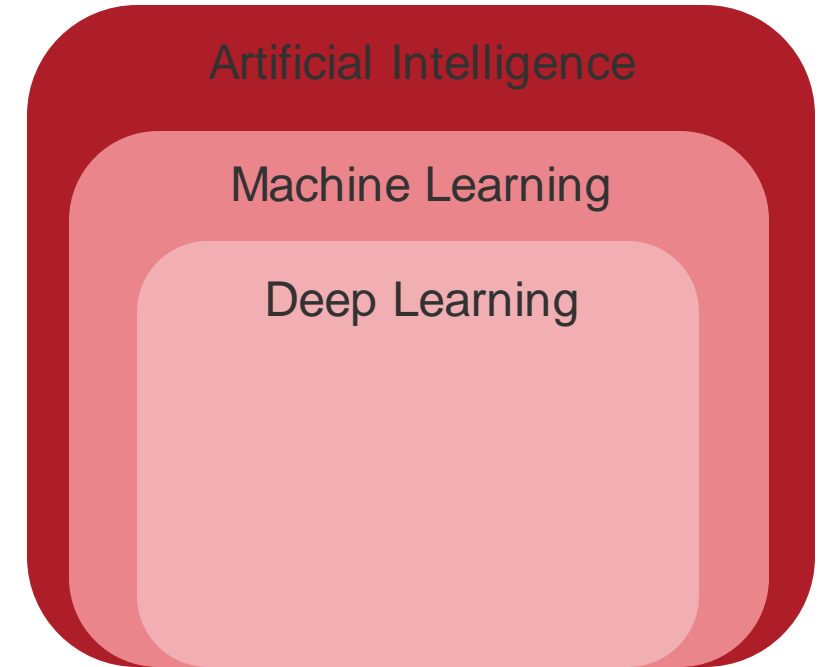
Machine Learning

- **Deep Learning (2012)**

Multi-layered neural networks to analyze large datasets. AlexNet's success in the ImageNet competition highlighted its transformative impact on image recognition.



Credit: AlexNet and ImageNet: The Birth of Deep Learning

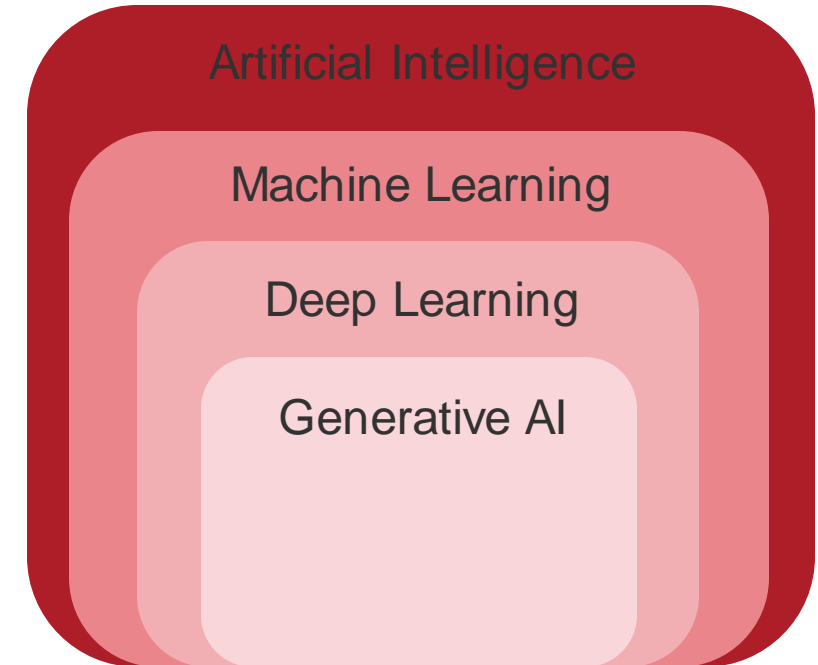




- **Generative AI (2021)**  
Capable of producing human-like text.

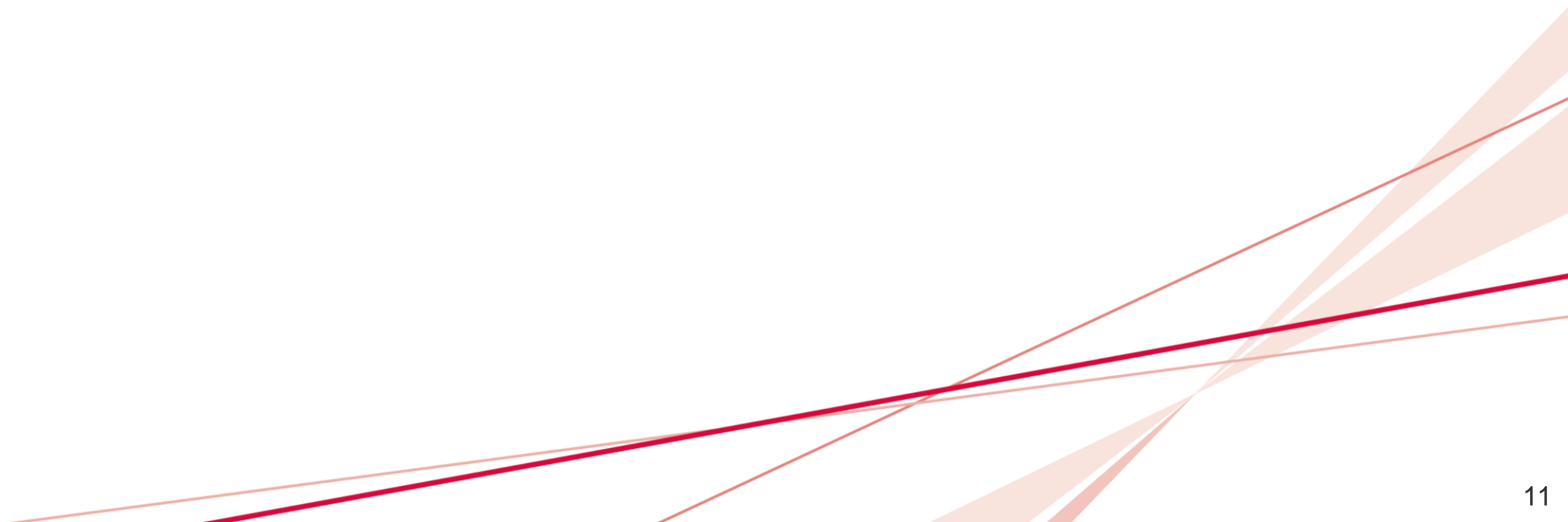


Credit: Shutterstock



**Generative AI in action**  
<https://duck.ai>

# How Generative AI works



**Tokenization is the process of breaking down text into smaller units, called tokens, which can be words, phrases, or subwords.**

Internet Initiative Japan, Inc. was founded in 1992

Input

Internet Initiative Japan, Inc. was founded in 1992

Tokenizer

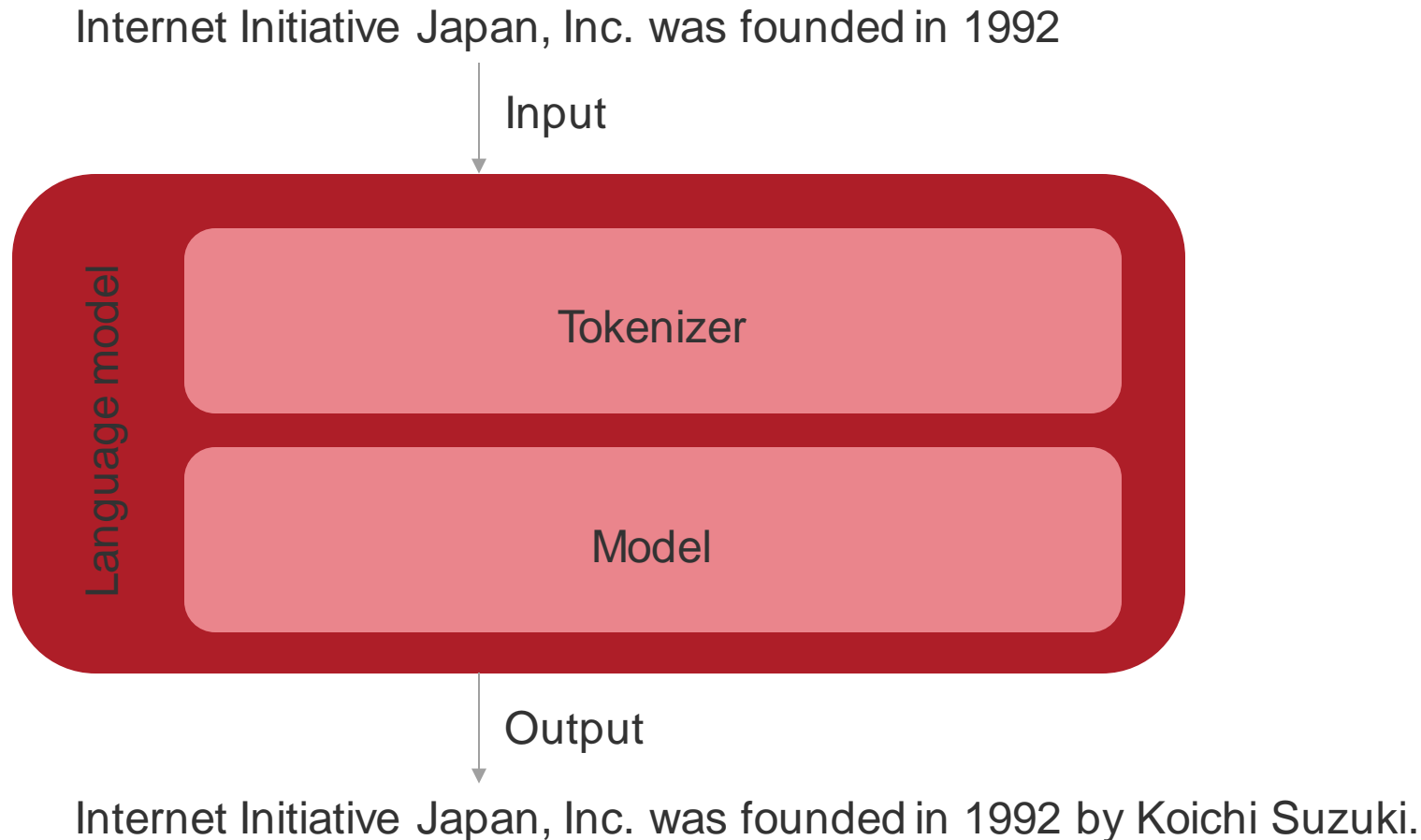
[34831, 55152, 10198, 11, 7079, 13, 673, 24303, 306, 220, 3204, 17]

Output

[34831, 55152, 10198, 11, 7079, 13, 673, 24303, 306, 220, 3204, 17]

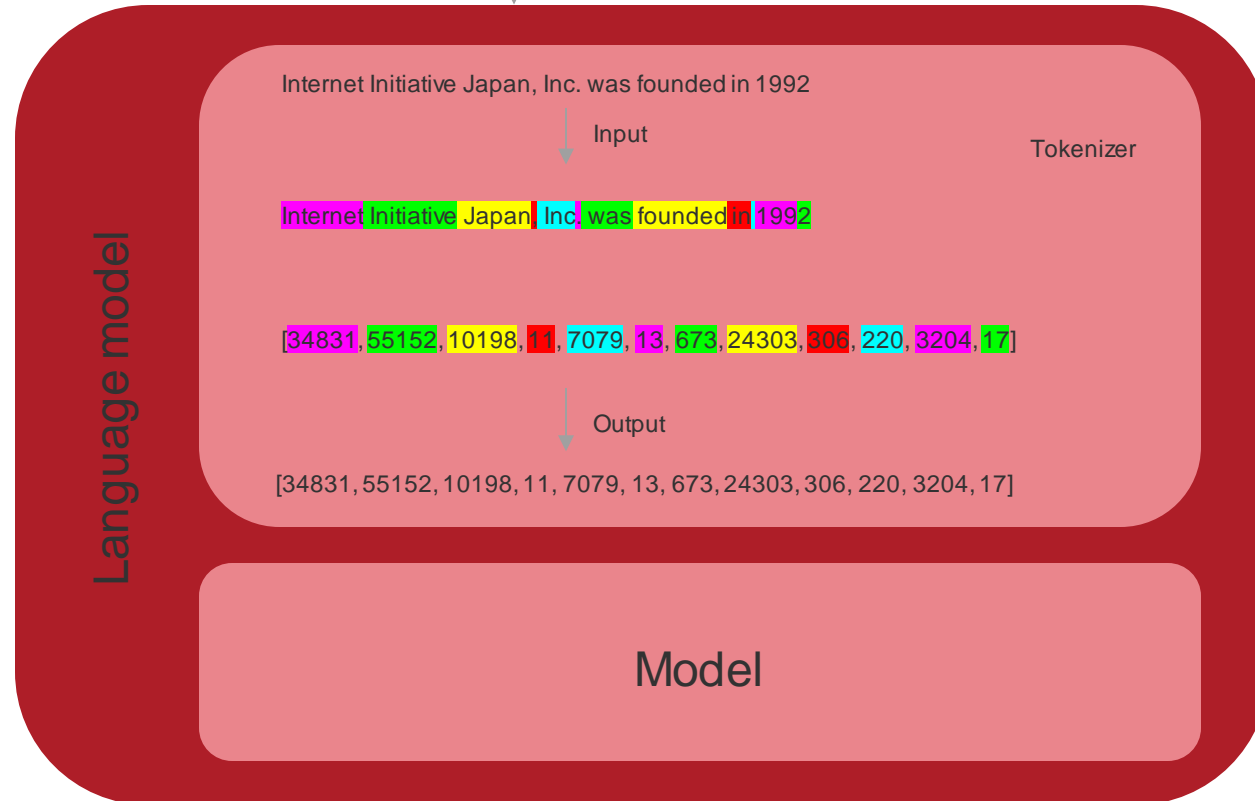
Live example: <https://platform.openai.com/tokenizer>

- **Prompt** is the input provided to a language model.
- **Completion** is the output generated a language model (predicting the next token).



Internet Initiative Japan, Inc. was founded in 1992

Input



Output

Internet Initiative Japan, Inc. was founded in 1992 by Koichi Suzuki.

## How language models generates text

Internet Initiative Japan, Inc. was founded in 1992

[34831, 55152, 10198, 11, 7079, 13, 673, 24303, 306, 220, 3204, 17]

Input

[0.01, 0.005, 0.003, ..., 0.95, 0.13, 0.85, ..., 0.006] p Model

Select max probability

[ 64, 378, 26682, ..., 656, 65, 2051, ..., 1778]

Output

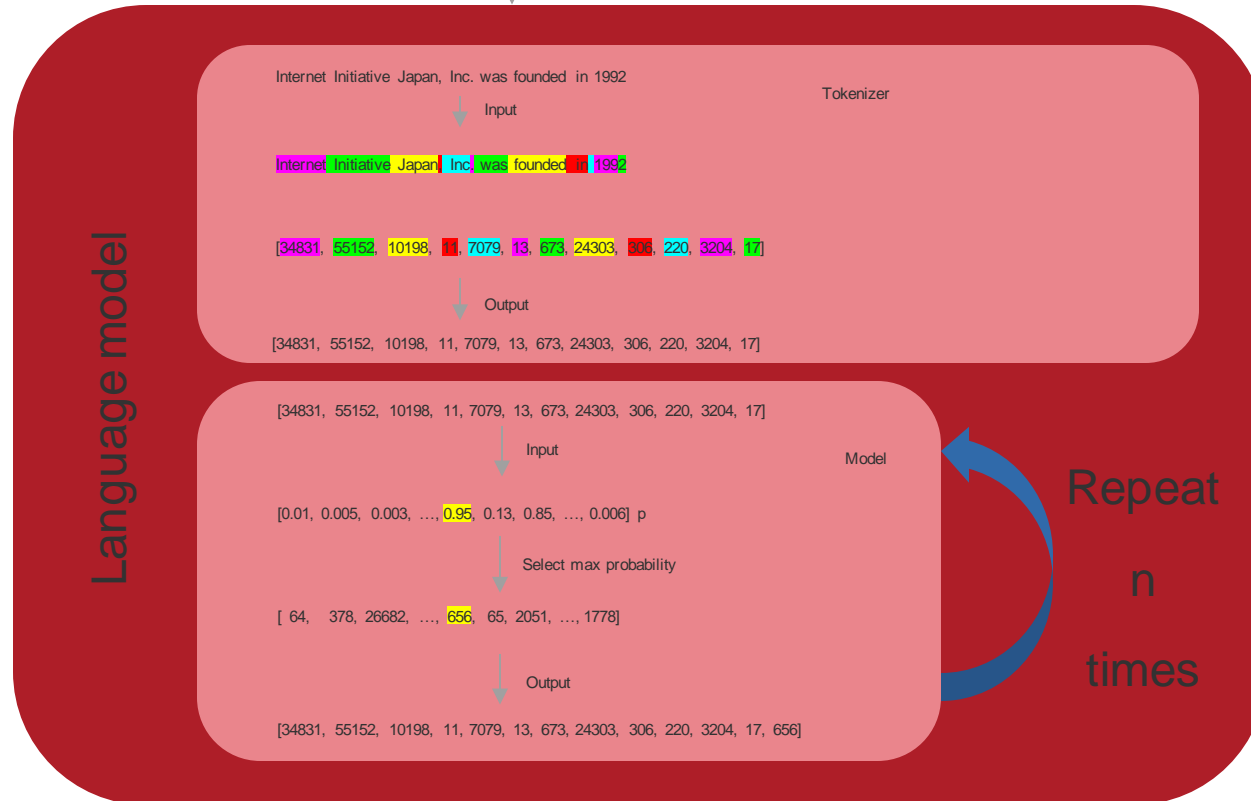
[34831, 55152, 10198, 11, 7079, 13, 673, 24303, 306, 220, 3204, 17, 656]

Internet Initiative Japan, Inc. was founded in 1992 by

# How language models generates text

Internet Initiative Japan, Inc. was founded in 1992

Input



Output

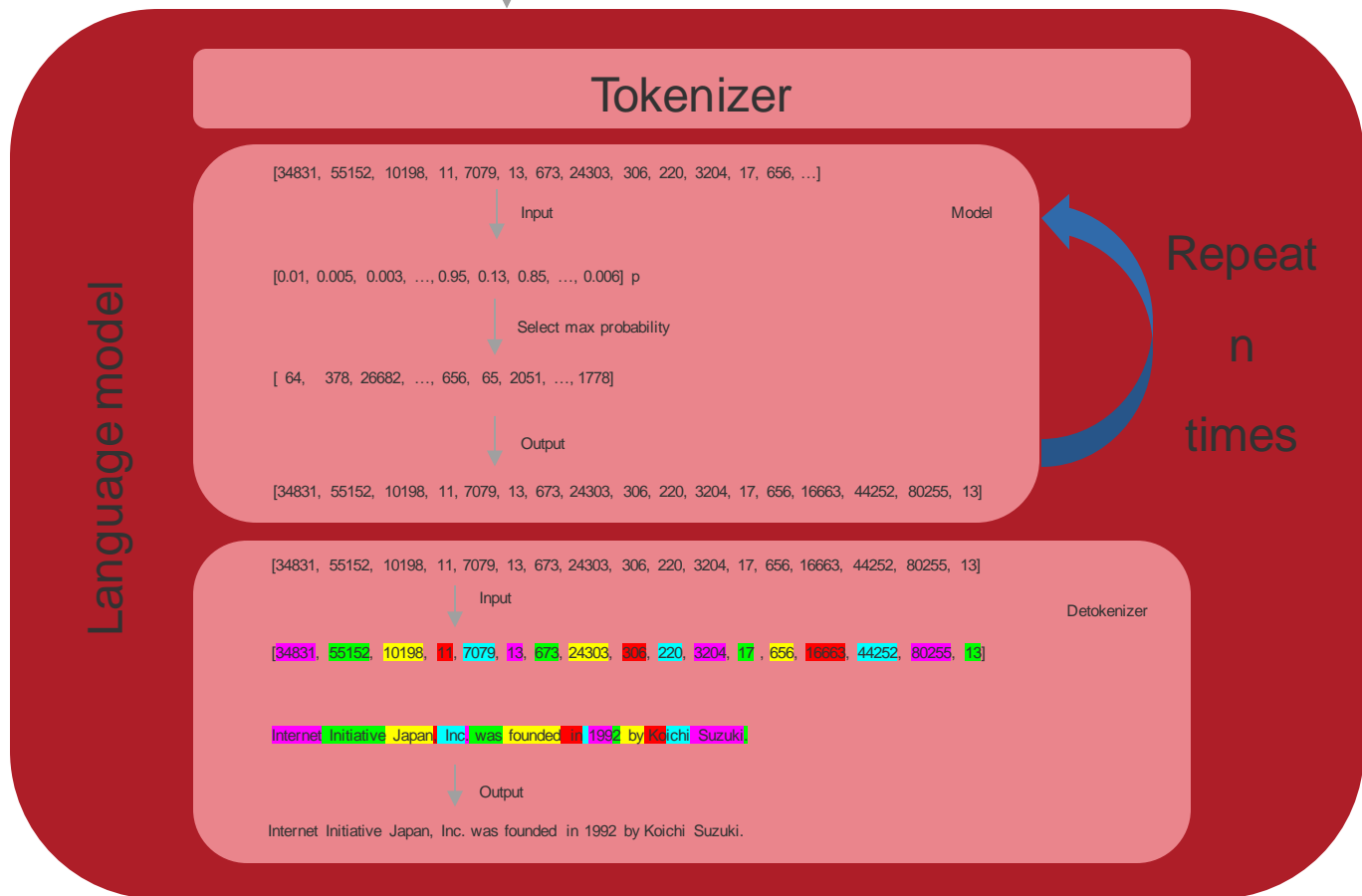
Internet Initiative Japan, Inc. was founded in 1992 by Koichi Suzuki.



# How language models generates text

Internet Initiative Japan, Inc. was founded in 1992

Input



Output

Internet Initiative Japan, Inc. was founded in 1992 by Koichi Suzuki.

**Detokenization involves decoding the numerical representations (vectors) back into human-readable text.**

[34831, 55152, 10198, 11, 7079, 13, 673, 24303, 306, 220, 3204, 17, 656, 16663, 44252, 80255, 13]

Input

[34831, 55152, 10198, 11, 7079, 13, 673, 24303, 306, 220, 3204, 17, 656, 16663, 44252, 80255, 13]

Internet Initiative Japan, Inc. was founded in 1992 by Koichi Suzuki.

Detokenizer

Output

Internet Initiative Japan, Inc. was founded in 1992 by Koichi Suzuki.

# Hosting your own LLM

- **Control and Customization:** Tune to specific use case, or industry.
- **Security and Compliance:** Ensure sensitive data and conversations remain private and secure.
- **Cost-Effective:** Avoid recurring cloud costs and optimize resource allocation for workloads.
- **Flexibility and Scalability:** Easily scale to meet changing demands, without relying on third-party infrastructure.
- **IP Protection:** Keep data, and models in-house, reducing the risk of exposure or theft.

- **Choose a Model:** Select a pre-trained LLM that fits your use case, such as a language translation or text generation model.
- **Select a Framework:** Decide on a framework to host your LLM, such as Ollama, or llama.cpp.
- **Prepare Infrastructure:** Ensure you have the necessary computational resources, storage, and memory.

- **Use Open WebUI:**

<https://github.com/open-webui/open-webui>

Using only **one** command:

```
docker run -d -p 3000:8080 -v ollama:/root/.ollama -v open-webui:/app/backend/data --name open-webui --restart always ghcr.io/open-webui/open-webui:ollama
```

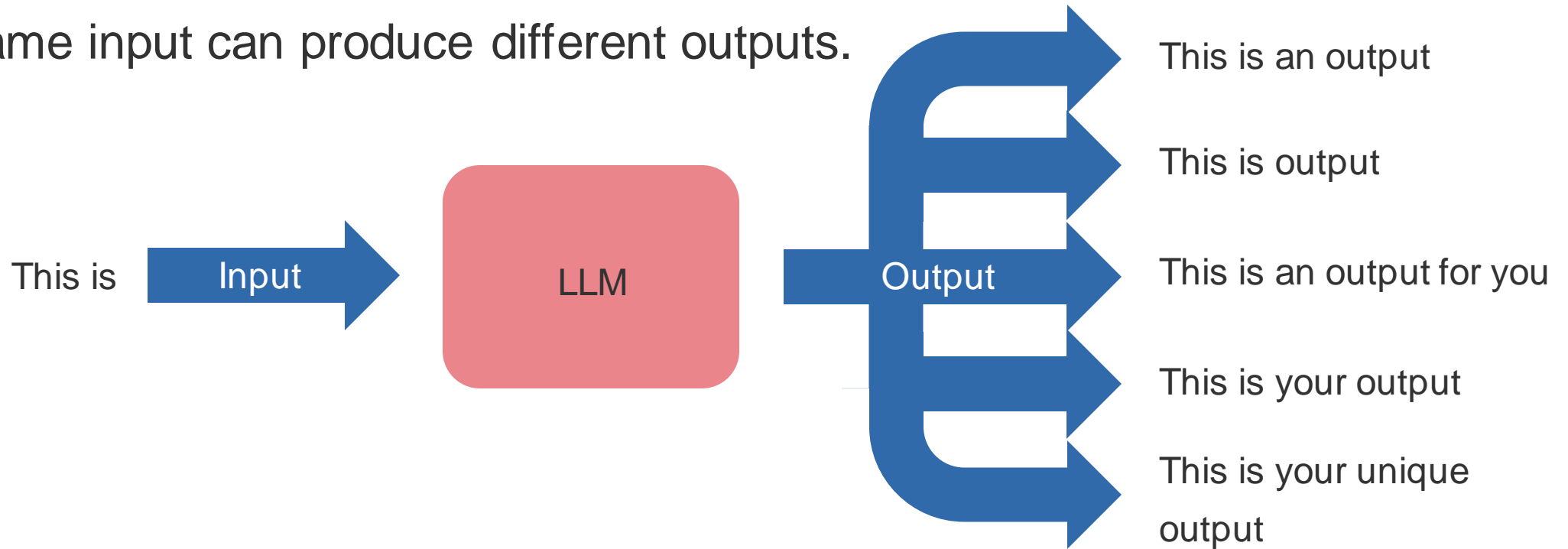
# Open WebUI in action

**How can you use an LLM for your use case?**



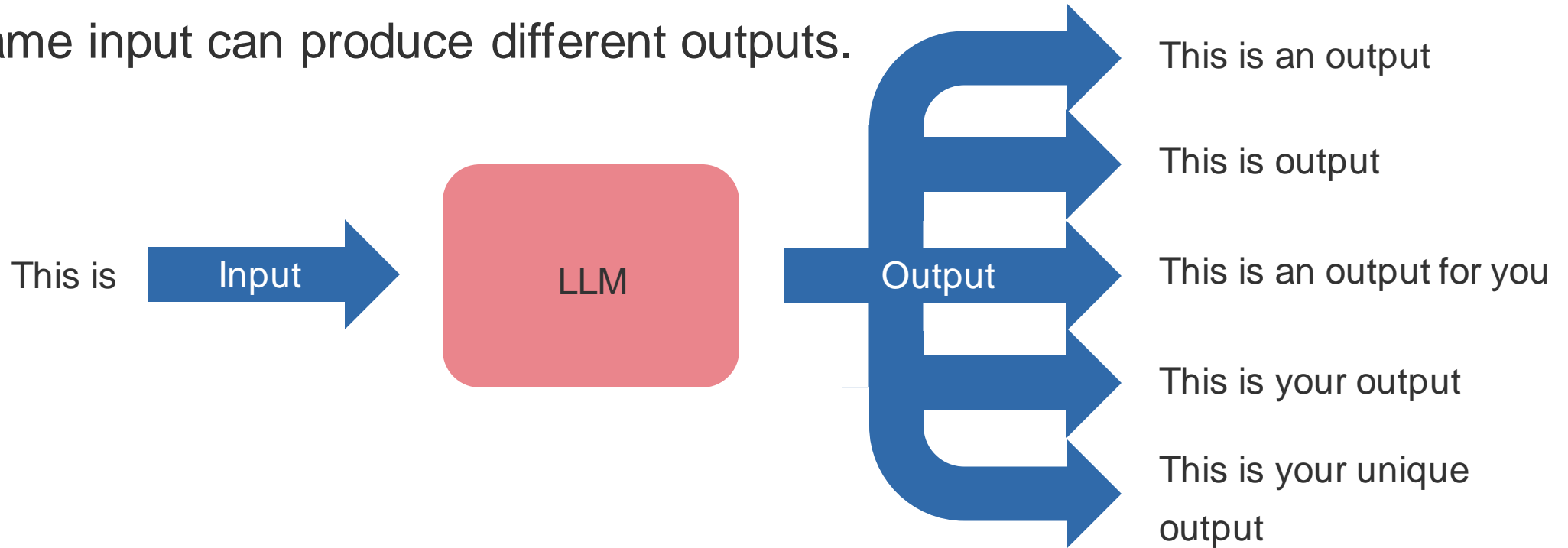
- **Language models are stochastic.**

- The same input can produce different outputs.

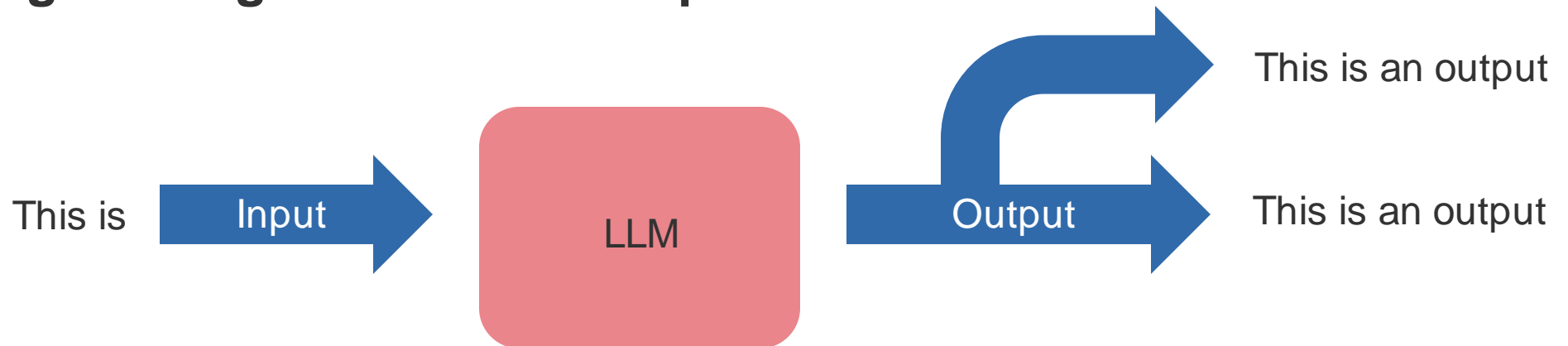


- **Language models are stochastic.**

- The same input can produce different outputs.



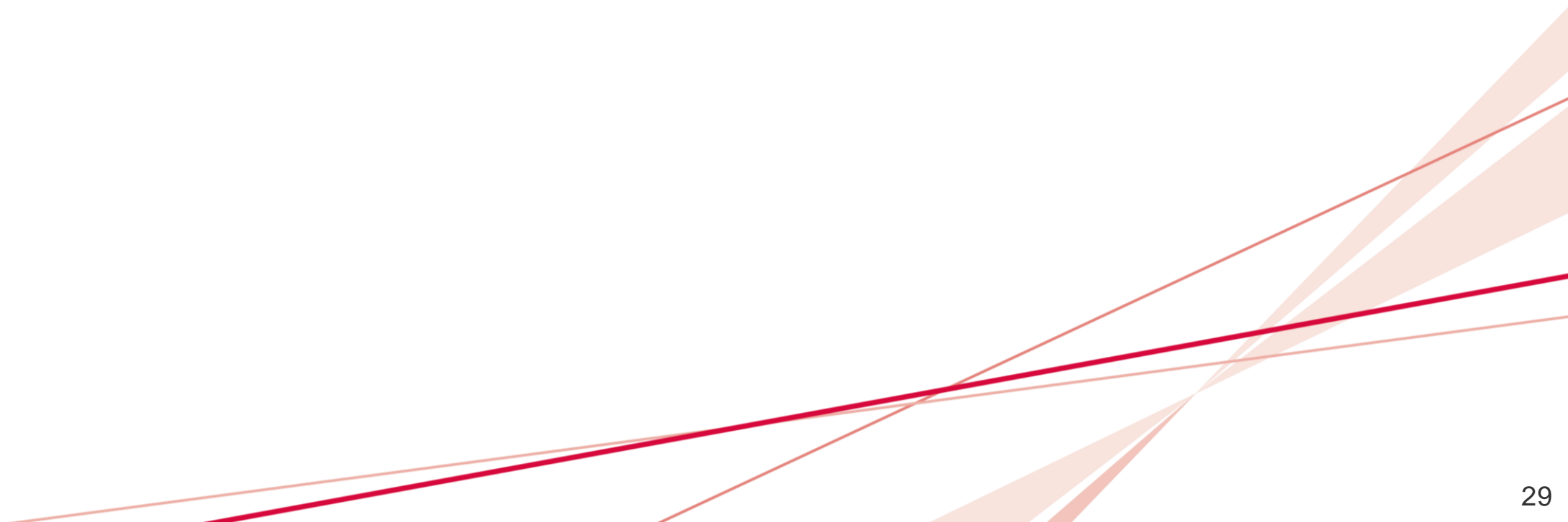
⇒ **Prompt engineering controls the responses.**



# Prompt engineering in action

- **Language models cannot think.**  
They simply predict the next token in a sequence based on learned patterns and probabilities from the data they were trained on.
- **Fabrication occurs when they generate responses that appear realistic but are not factually accurate.**

# Fabrication in action

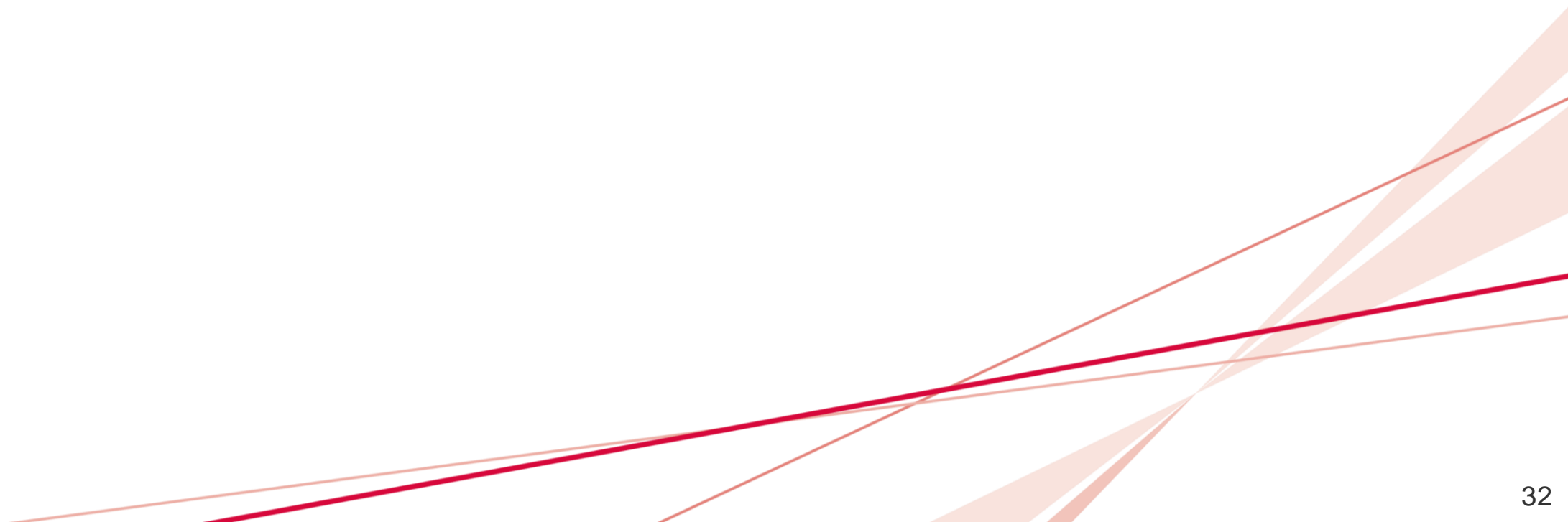


## **Write clear instructions**

- **Basic prompt for Completion**
- **Iterate prompts for Conversation**
- **Instruct Task**
- **Refine Context for Clarity and/or Length Format**
- **Primary Context**
- **Provide Guides**
- **Provide Examples**

# Prompt engineering in action

# Conclusion





If you want to learn more about Generative AI and LLMs, you can watch Microsoft's 'Generative AI for Beginners' lessons.

<https://github.com/microsoft/generative-ai-for-beginners>

The screenshot shows the GitHub repository page for 'generative-ai-for-beginners'. At the top, it indicates the repository is public and shows 557 watches, 33.2k forks, and 65.1k stars. The main content area displays a list of files and folders with their commit messages and dates. The 'About' section on the right provides a brief description of the repository, including a link to the lessons and a list of related topics like 'ai', 'azure', 'transformers', and 'openai'.

File/Folder	Commit Message	Time Ago
.devcontainer	Add pip dep, azure-ai-inference	2 months ago
.github	remove chores/fixes branch	8 months ago
.vscode	update UX section, add more references	last year
00-course-setup	Fix#540: Renamed AZURE_OPENAI_KEY to AZURE_OPE...	2 months ago
01-introduction-to-genai	docs(chapter-1): Fix broken image URL and list indentation	3 weeks ago
02-exploring-and-comparing-different-llms	fix broken urls	3 months ago
03-using-generative-ai-responsibly	Merge pull request #529 from Inder24/Adding-OpenAIFil...	4 months ago
04-prompt-engineering-fundamentals	Merge pull request #590 from bmerkle/fix#589	2 months ago
05-advanced-prompts	Add zh-tw translations	5 months ago
06-text-generation-apps	added the chat application lesson	2 months ago
07-building-chat-applications	added the chat application lesson	2 months ago
08-building-search-applications	adding search applications app	2 months ago

**About**  
21 Lessons, Get Started Building with Generative AI <https://microsoft.github.io/generative-ai-for-beginners/>  
[microsoft.github.io/generative-ai-for-...](https://microsoft.github.io/generative-ai-for-...)

ai azure transformers openai gpt language-model semantic-search dall-e prompt-engineering llms generative-ai generativeai chatgpt

Readme MIT license Code of conduct Security policy Activity Custom properties 65.1k stars 557 watching 33.2k forks Report repository

# This presentation covered lessons 1, 2, and 4.

## 📖 Lessons

#	Lesson Link	Description	Video	Extra Learning
00	<a href="#">Course Setup</a>	<b>Learn:</b> How to Setup Your Development Environment	Coming Soon	<a href="#">Learn More</a>
01	<a href="#">Introduction to Generative AI and LLMs</a>	<b>Learn:</b> Understanding what Generative AI is and how Large Language Models (LLMs) work.	<a href="#">Video</a>	<a href="#">Learn More</a>
02	<a href="#">Exploring and comparing different LLMs</a>	<b>Learn:</b> How to select the right model for your use case	<a href="#">Video</a>	<a href="#">Learn More</a>
03	<a href="#">Using Generative AI Responsibly</a>	<b>Learn:</b> How to build Generative AI Applications responsibly	<a href="#">Video</a>	<a href="#">Learn More</a>
04	<a href="#">Understanding Prompt Engineering Fundamentals</a>	<b>Learn:</b> Hands-on Prompt Engineering Best Practices	<a href="#">Video</a>	<a href="#">Learn More</a>
05	<a href="#">Creating Advanced Prompts</a>	<b>Learn:</b> How to apply prompt engineering techniques that improve the outcome of your prompts.	<a href="#">Video</a>	<a href="#">Learn More</a>
06	<a href="#">Building Text Generation Applications</a>	<b>Build:</b> A text generation app using Azure OpenAI / OpenAI API	<a href="#">Video</a>	<a href="#">Learn More</a>
07	<a href="#">Building Chat Applications</a>	<b>Build:</b> Techniques for efficiently building and integrating chat applications.	<a href="#">Video</a>	<a href="#">Learn More</a>
08	<a href="#">Building Search Apps Vector Databases</a>	<b>Build:</b> A search application that uses Embeddings to search for data.	<a href="#">Video</a>	<a href="#">Learn More</a>
09	<a href="#">Building Image Generation</a>	<b>Build:</b> A image generation application	<a href="#">Video</a>	<a href="#">Learn</a>

- <https://spectrum.ieee.org/how-ibms-deep-blue-beat-world-champion-chess-player-garry-kasparov>
- <https://aiws.net/the-history-of-ai/this-week-in-the-history-of-ai-at-aiws-net-the-dartmouth-conference-began-on-18-june-1956-2/>
- <https://www.pinecone.io/learn/series/image-search/imagenet/>
- <https://www.computerworld.com/article/1627101/what-are-large-language-models-and-how-are-they-used-in-generative-ai.html>