

# Generative AI in Organisations: Challenges and Opportunities

ODSC Europe 2023 - Heiko Hotz

# Agenda

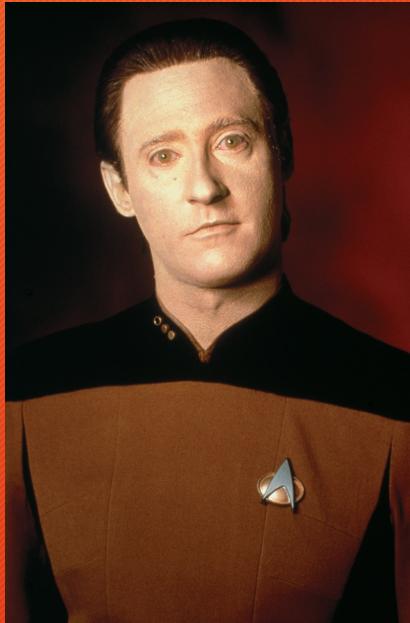
- What is Generative AI & how does it work?
- How to use Large Language Models (LLMs)?
- Demo: Chat with your document
- Bias and ethics in Generative AI

# What is Generative AI?

# Generative AI - I grew up with it 😊



Source: <https://theconversation.com/star-treks-holodeck-from-science-fiction-to-a-new-reality-74839>



Source: <https://www.fanpop.com/clubs/star-trek-the-next-generation/images/9406565/title/lt-commander-data-photo>

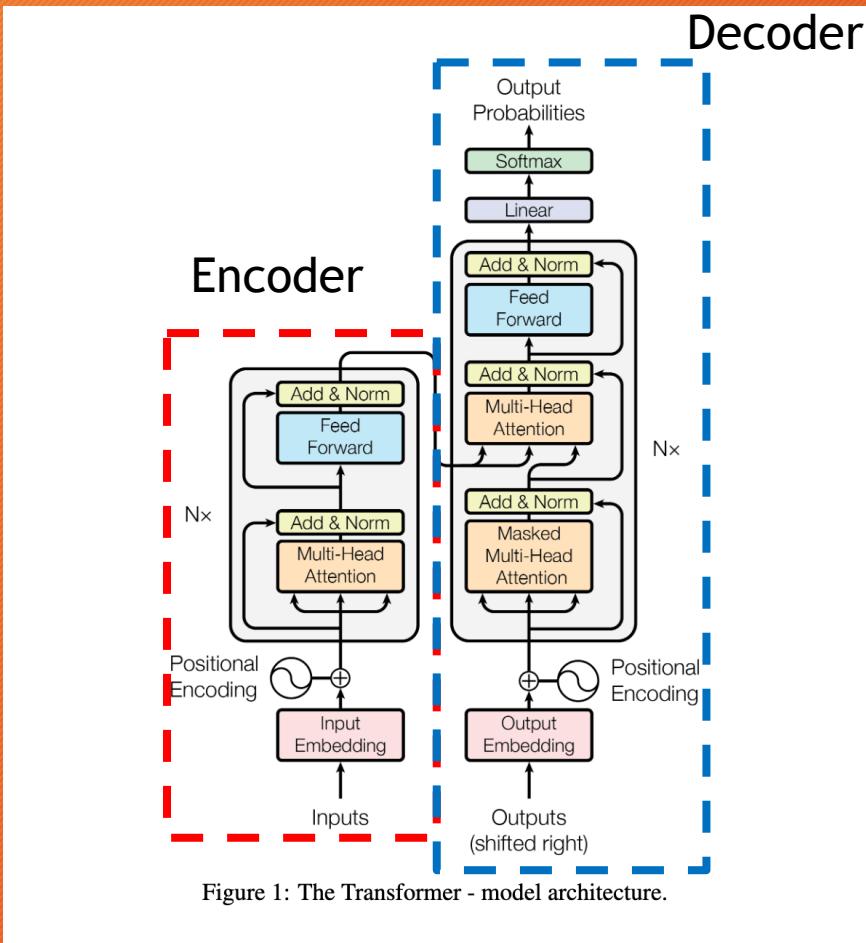
# Generative AI - What is it?



AI that can produce original content close enough to human generated content for real-world tasks

# How does it work (Text Generation Edition)?

# Transformer Architecture



Source:  
<https://arxiv.org/pdf/1706.03762.pdf>  
(Attention is all you need)

# Encoder-only models

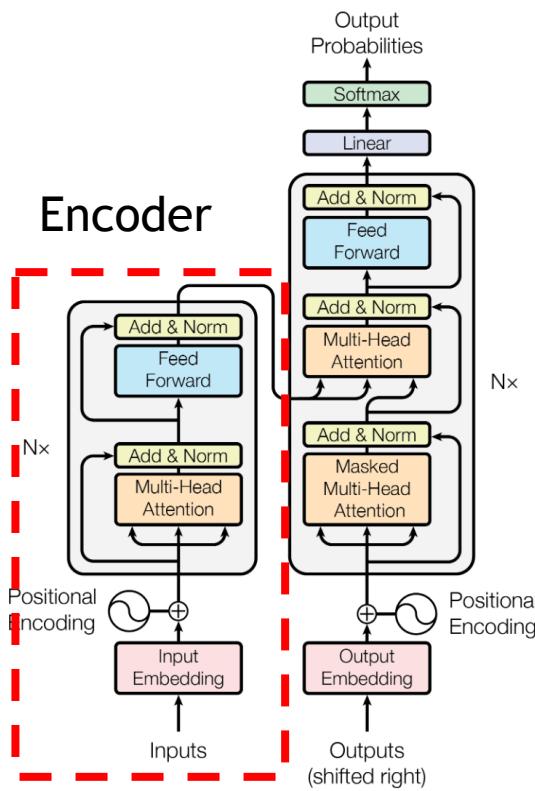
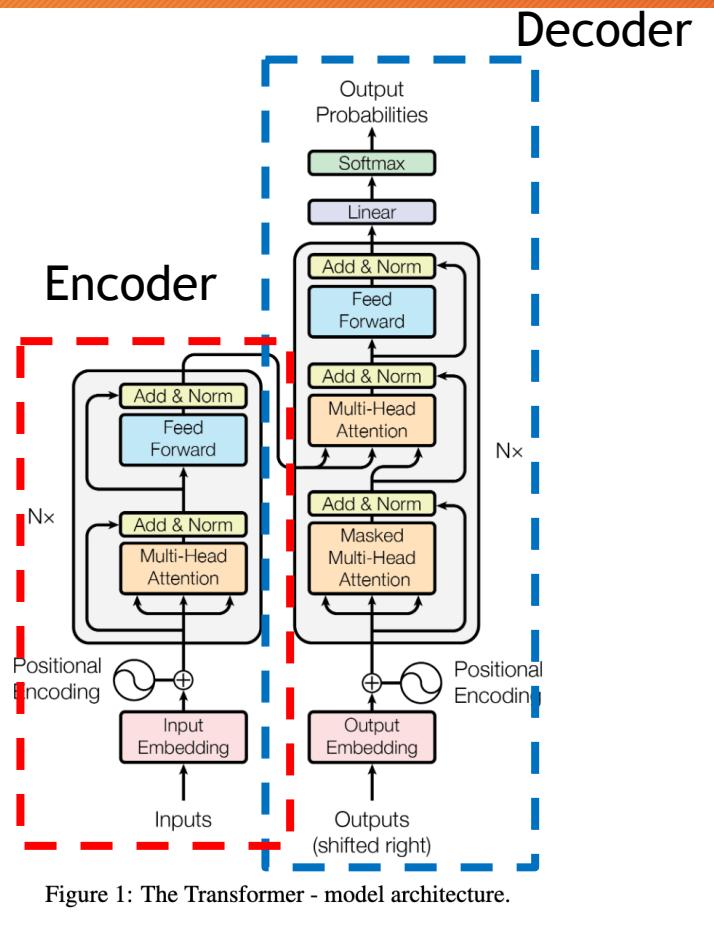


Figure 1: The Transformer - model architecture.

## Encoder:

- Reads the entire input sequence & encodes the information
- Bidirectional: It has access to all the words in the input sequence
- Useful for tasks where the entire context needs to be taken into account
- Example: Text classification with BERT (Bidirectional Encoder Representation from Transformers)

# Encoder-decoder models



## Encoder-Decoder:

- Combines both abilities: Reading and understanding an input sequence and generating an output sequence
- Sequence-to-sequence models (T5, BART)
- Example: Translation

# Decoder-only models

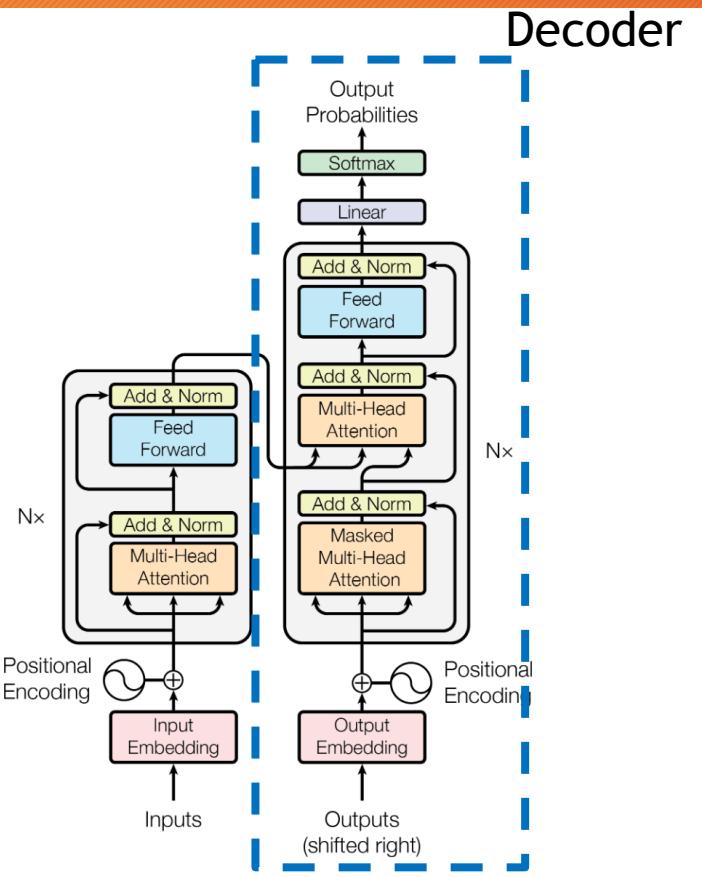


Figure 1: The Transformer - model architecture.

## Decoder:

- Generates output sequence one token at a time
- Unidirectional: It has access only to the words to the left
- Autoregressive: Each next token is predicted based on the tokens that came before
- Example: Text generation (All GPT models are decoder-only)

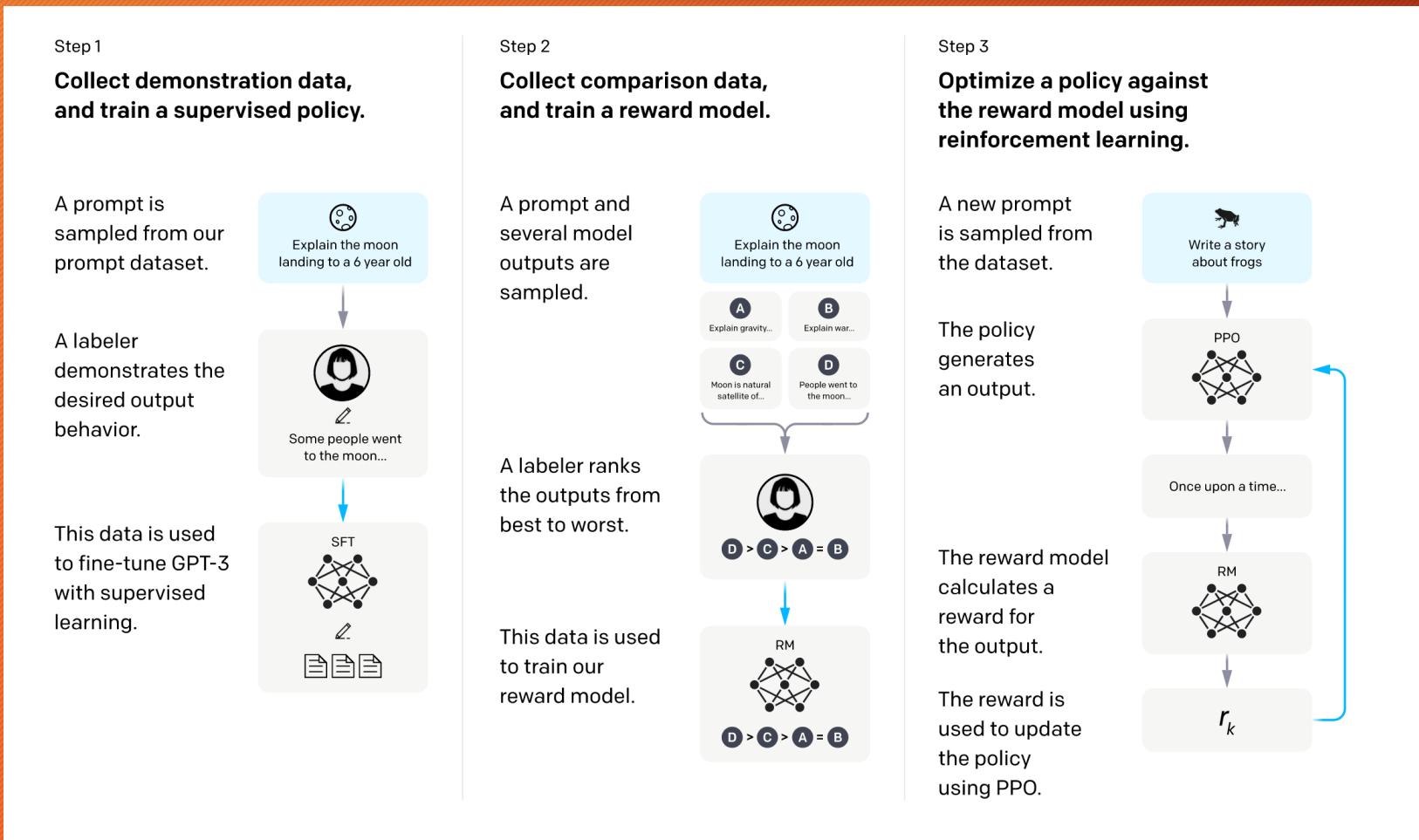
# How to train decoder-only models

## Step1: Self-supervised *pretraining*

- Feed the model with text from left to right and it learns to predict the next ...
- Self-supervised because it uses unlabelled data and creates the label on its own (by masking the next word)

## Step 2: Train (finetune) the model on human feedback

# Training using human feedback



# Agenda

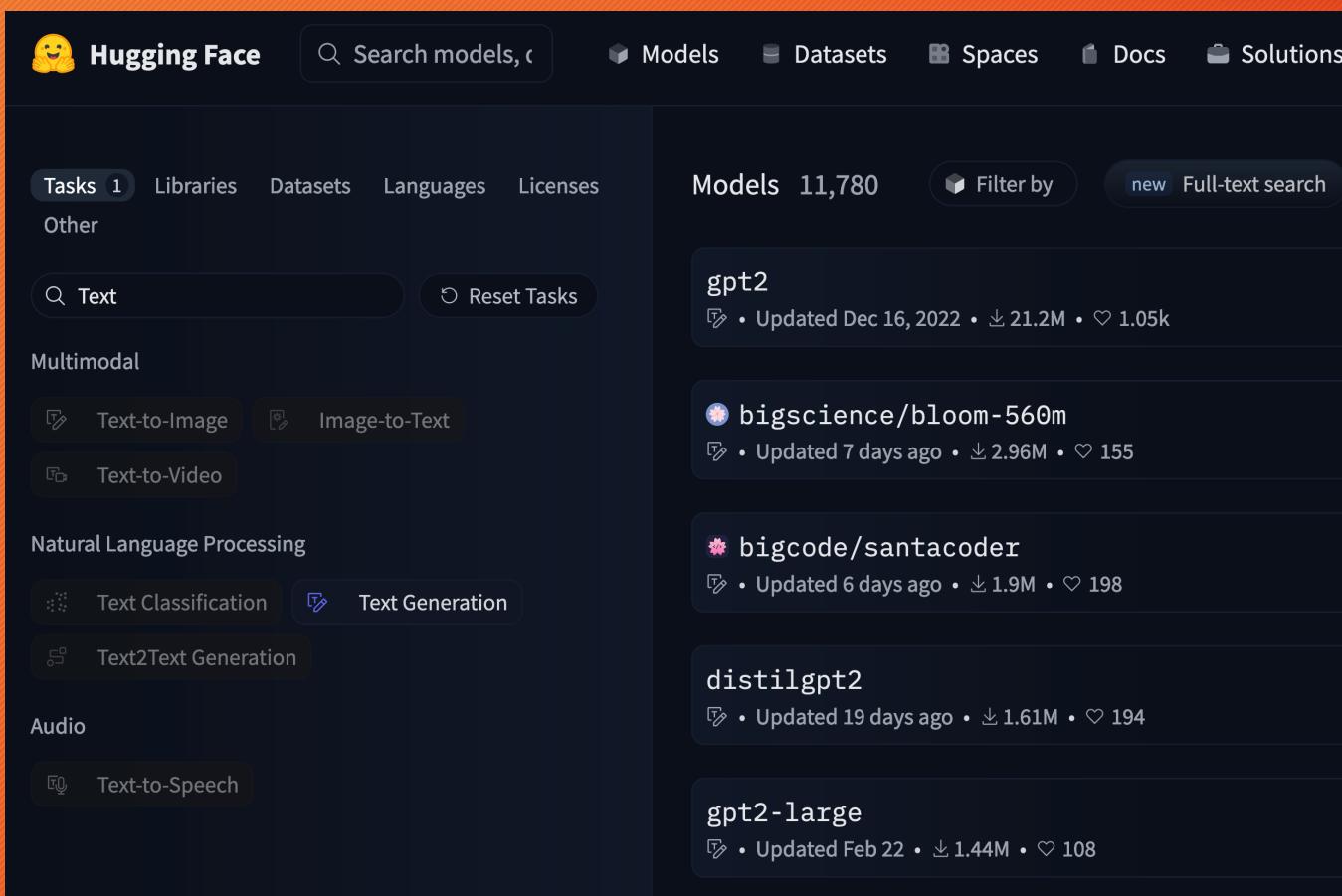
- What is Generative AI & how does it work?
- How to use Large Language Models (LLMs)?
- Demo: Chat with your document
- Bias and ethics in Generative AI

# How to deploy Large Language Models?

# There are different ways to use LLMs

- API-based (OpenAI)
- Self-hosted on local machine
- Hosted in Cloud

# Self-hosting



The screenshot shows the Hugging Face website interface. At the top, there is a navigation bar with the Hugging Face logo, a search bar, and links for Models, Datasets, Spaces, Docs, and Solutions. Below the navigation bar, there is a sidebar with categories: Tasks (1), Libraries, Datasets, Languages, Licenses, Other, and a search bar for Text. The main content area shows a list of 11,780 models. The first few models listed are:

- gpt2**  
• Updated Dec 16, 2022 • 21.2M • 1.05k
- bigscience/bloom-560m**  
• Updated 7 days ago • 2.96M • 155
- bigcode/santacoder**  
• Updated 6 days ago • 1.9M • 198
- distilgpt2**  
• Updated 19 days ago • 1.61M • 194
- gpt2-large**  
• Updated Feb 22 • 1.44M • 108

# Challenge with self-hosting

- GPU Memory!

# Cloud providers can help

	Instance Size	GPU	GPU Memory (GiB)	vCPUs	Memory (GiB)	Storage (GB)	Network Bandwidth (Gbps)	EBS Bandwidth (Gbps)	On Demand Price/hr*
Single GPU VMs	g5.xlarge	1	24	4	16	1x250	Up to 10	Up to 3.5	\$1.006
	g5.2xlarge	1	24	8	32	1x450	Up to 10	Up to 3.5	\$1.212
	g5.4xlarge	1	24	16	64	1x600	Up to 25	8	\$1.624
	g5.8xlarge	1	24	32	128	1x900	25	16	\$2.448
	g5.16xlarge	1	24	64	256	1x1900	25	16	\$4.096
Multi GPU VMs	g5.12xlarge	4	96	48	192	1x3800	40	16	\$5.672
	g5.24xlarge	4	96	96	384	1x3800	50	19	\$8.144
	g5.48xlarge	8	192	192	768	2x3800	100	19	\$16.288

Source: <https://aws.amazon.com/ec2/instance-types/g5/>

# How to use Large Language Models?

# API use

<SCREENSHOT>

# Better: playground

## Create Your Own Large Language Model Playground in SageMaker Studio

Now you can deploy LLMs and experiment with them all in one place

Heiko Hotz  
Published in Towards Data Science · 4 min read · Mar 20

31 1



Image by author — created with Midjourney

### Flan-T5 Parameters

Stop word

Min/Max length

Temperature

Repetition Penalty

Flan-T5-XXL Playground

Enter your prompt here:

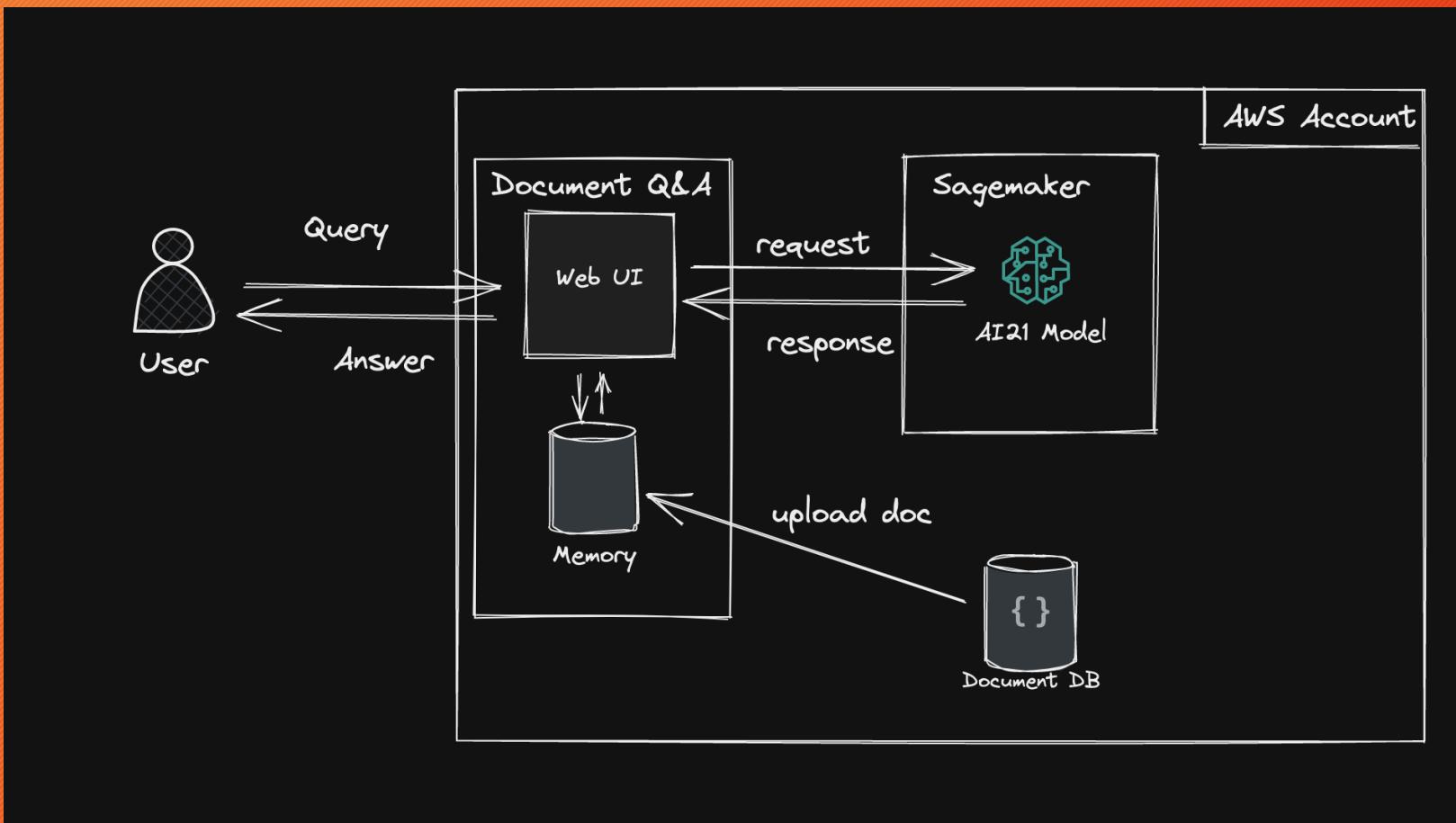
Q: Could Albert Einstein have had a conversation with George Washington? Give the rationale before answering.

Run

George Washington died in 1799. Albert Einstein was born in 1879. The final answer: no.

# How to build applications with LLMs?

# Architecture



# Demo

# Agenda

- What is Generative AI & how does it work?
- How to use Large Language Models (LLMs)?
- Demo: Chat with your document
- Bias and ethics in Generative AI

# Bias in LLMs

SCREENSHOT

# Ethics with (Generative) AI

**TV and film writers are fighting to save their jobs from AI. They won't be the last**



By [Samantha Murphy Kelly](#), CNN Business

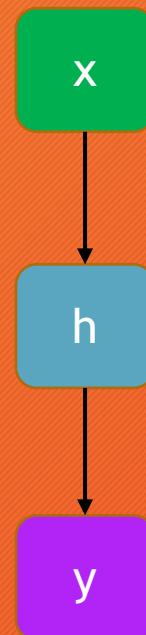
Updated 5:28 PM EDT, Thu May 4, 2023

Source: <https://edition.cnn.com/2023/05/04/tech/writers-strike-ai/index.html>

# Appendix

# Neural Networks Recap

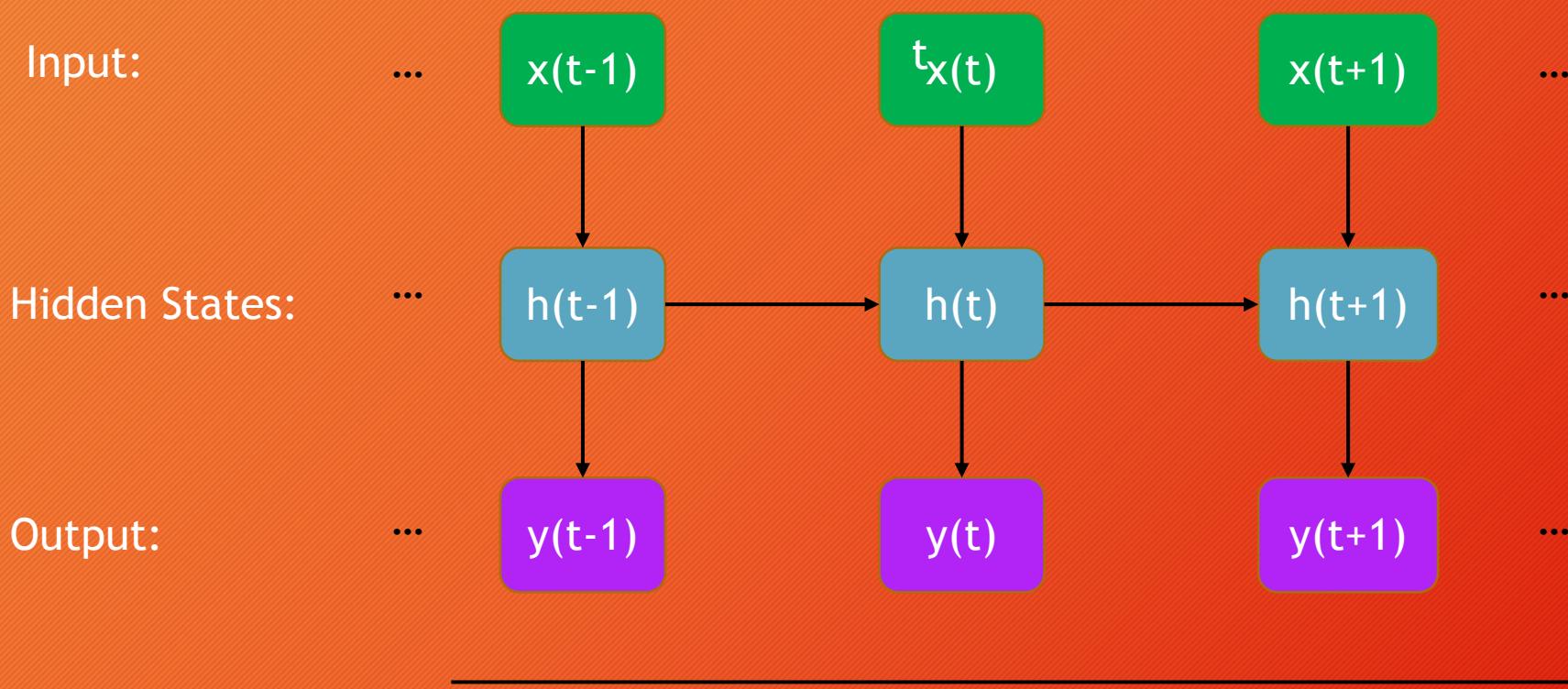
Input:



Hidden States:

Output:

# Recurrent Neural Networks Recap



# RNNs are useful for text generation

