Generative AI in Organisations: Challenges and Opportunities

ODSC Europe 2023 – Heiko Hotz

About myself

Heiko Hotz

- Senior Solutions Architect for Generative AI at AWS
- Founder of not-for-profit Meetup group NLP London
- Data Ambassador at DataKind UK
- Independent consultant at AI/ML Consulting
- More than 20 years of experience in technology
- YouTuber, Writer, Mentor, …

GH Repo

Slides and code samples:

https://github.com/marshmellow77/odsc-europe-2023

Agenda

- What is Generative AI & how does it work?
- Challenges with Large Language Models (LLMs)
- Demo: Document Chatbot
- Bias in Generative AI
- Q&A

Agenda

- What is Generative AI & how does it work?
- Challenges with Large Language Models (LLMs)
- Demo: Document Chatbot
- Bias in Generative AI
- Q&A

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?

Transformer Architecture

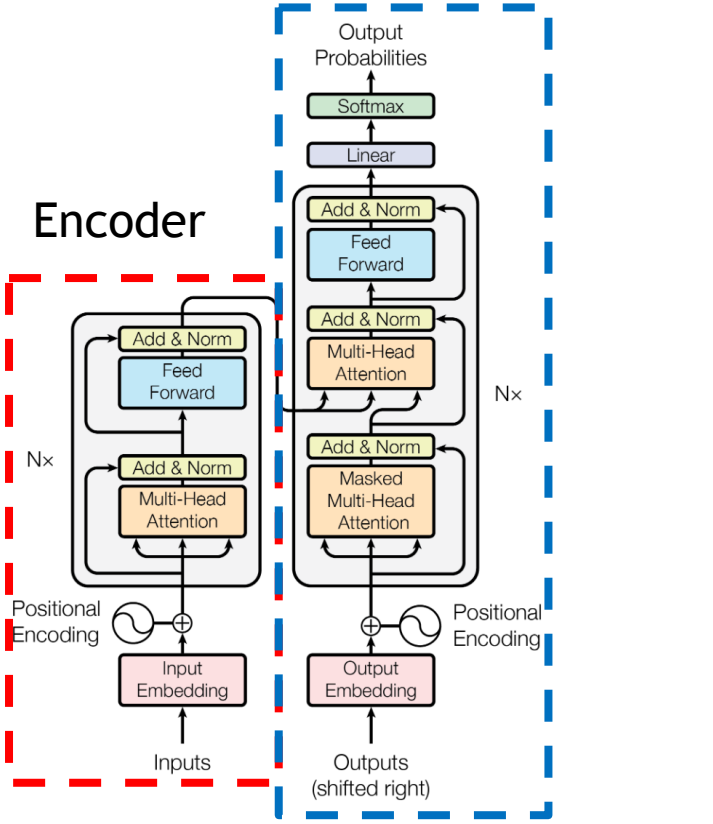


Figure 1: The Transformer - model architecture.

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

Encoder-only models

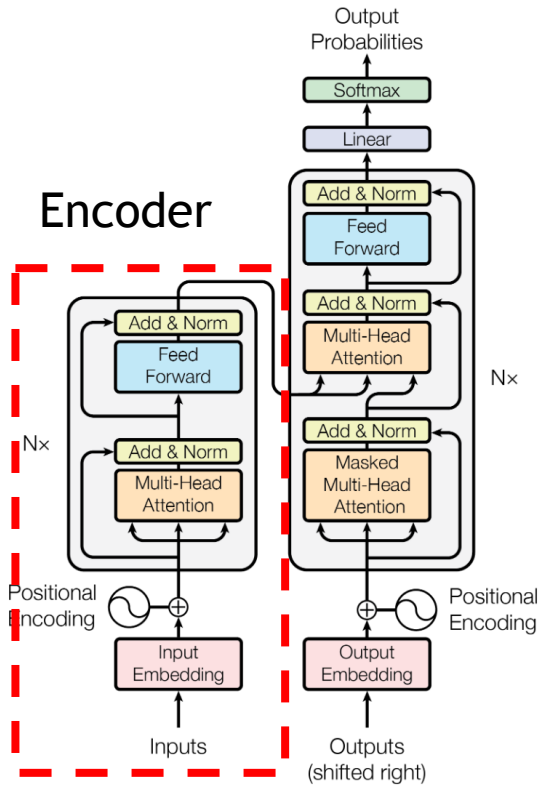


Figure 1: The Transformer - model architecture.

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 (**B**idirectional **E**ncoder **R**epresentation from **T**ransformers)

Decoder-only models

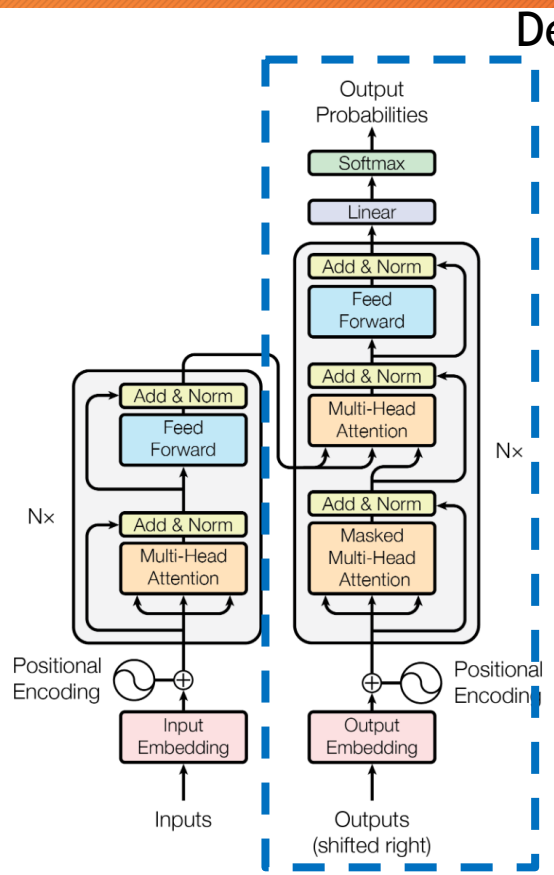


Figure 1: The Transformer - model architecture.

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)

Encoder-decoder models

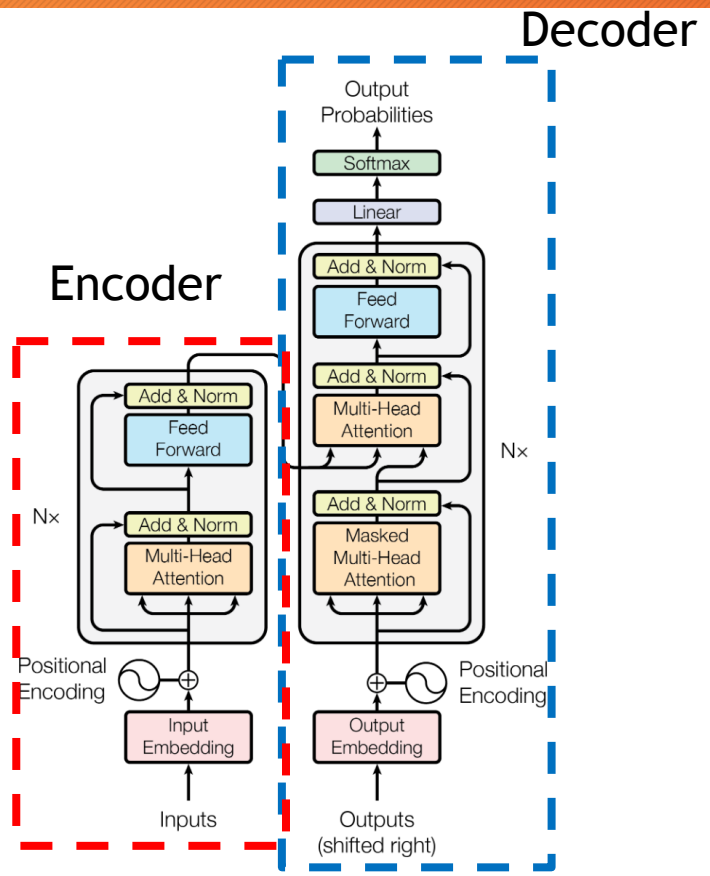


Figure 1: The Transformer - model architecture.

Encoder-Decoder:

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

Examples

Encoder-only models

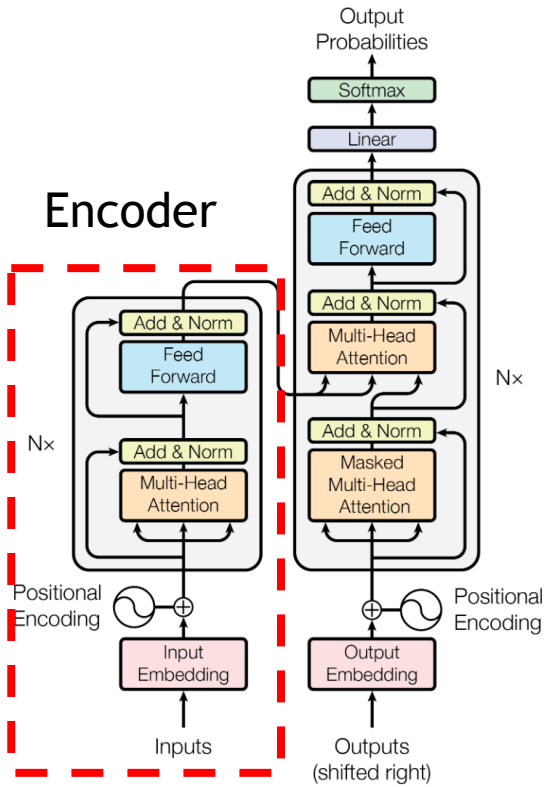


Figure 1: The Transformer - model architecture.

Example:

"She was preparing dinner for her friends who never arrived."

Negative sentiment

Encoder-decoder models

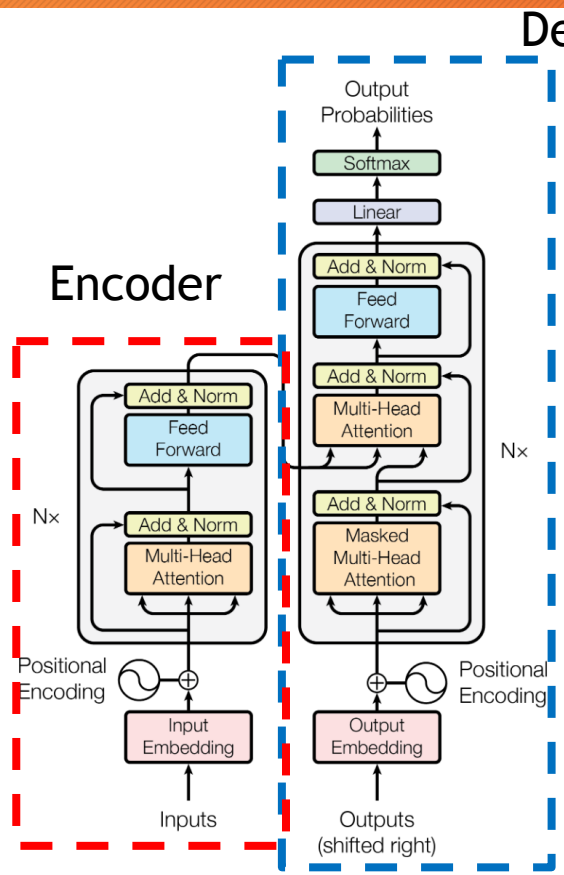


Figure 1: The Transformer - model architecture.

Example:

English: "I have given the book to John."

German: "Ich habe John das Buch gegeben."

Decoder-only models

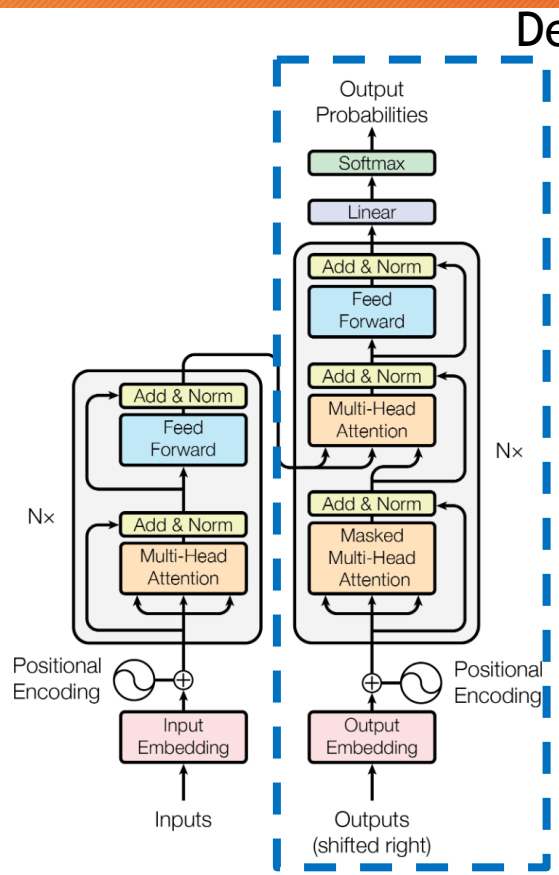


Figure 1: The Transformer - model architecture.

Example:

English: ”I went to the bakery and bought …”

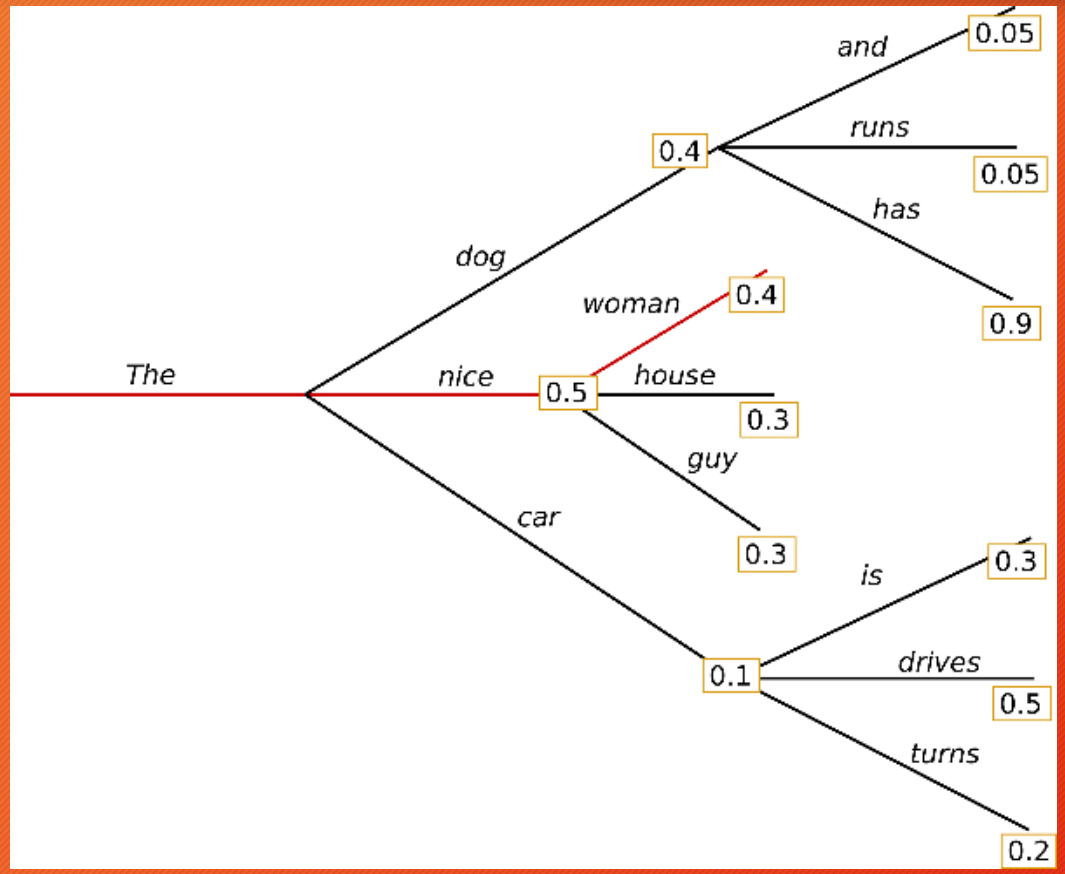
How do decoder-only models work?

How the decoder produces text

$$P(w_{1:T}|W_0) = \prod_{t=1}^{T} P(w_t|w_{1:t-1}, W_0) \text{ ,with } w_{1:0} = \emptyset,$$

Source: https://huggingface.co/blog/how-to-generate

How the decoder produces text



Source: https://huggingface.co/blog/how-to-generate

Demo in GPT Playground

Agenda

- What is Generative AI & how does it work?
- Challenges with Large Language Models (LLMs)
- Demo: Document Chatbot
- Bias in Generative AI
- Q&A

Agenda

- What is Generative AI & how does it work?
- Challenges with Large Language Models (LLMs)
  - GPU Requirements
  - Knowledge cut-off/hallucinations
- Demo: Document Chatbot for private documents
- Bias in Generative AI
- Q&A

How to use LLMs?

There are different ways to use LLMs

Public facing API

co:here

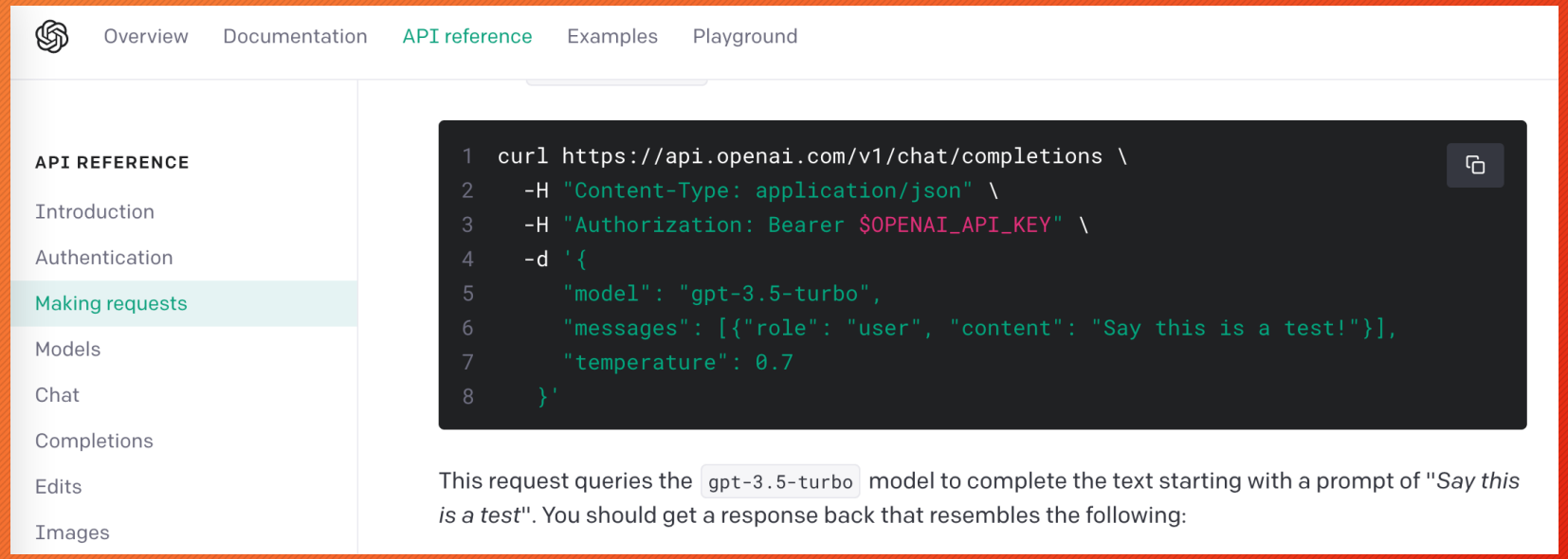
AI21 labs

ANTHROP\C

Self-hosted (locally or cloud)

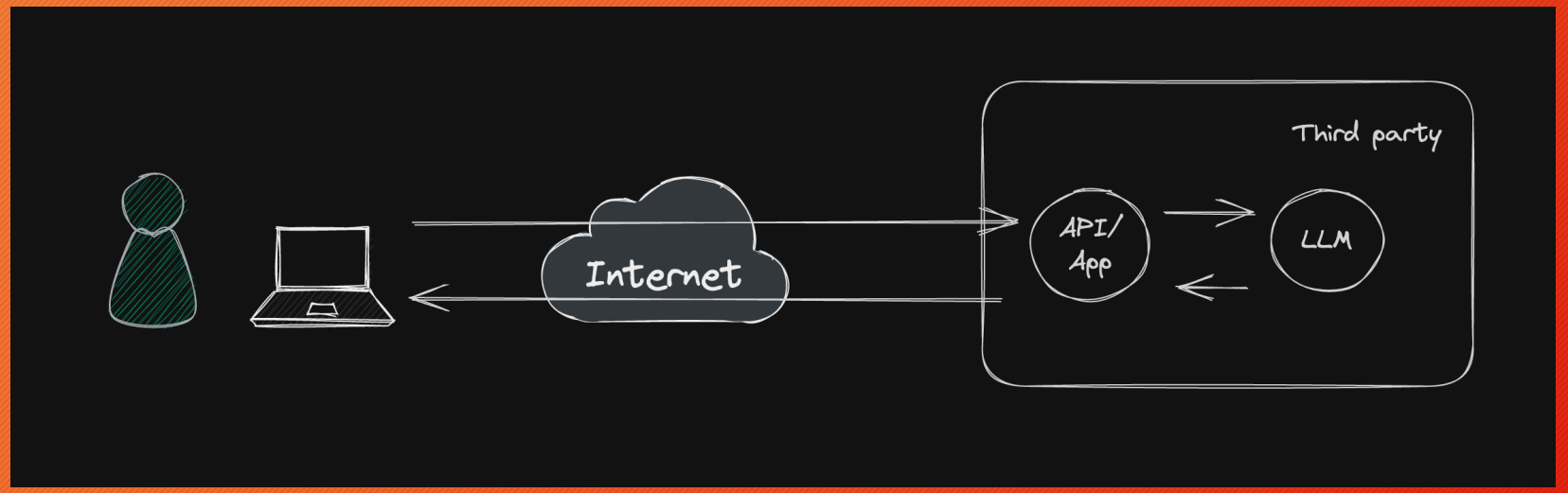
Google Cloud Platform

Public facing APIs

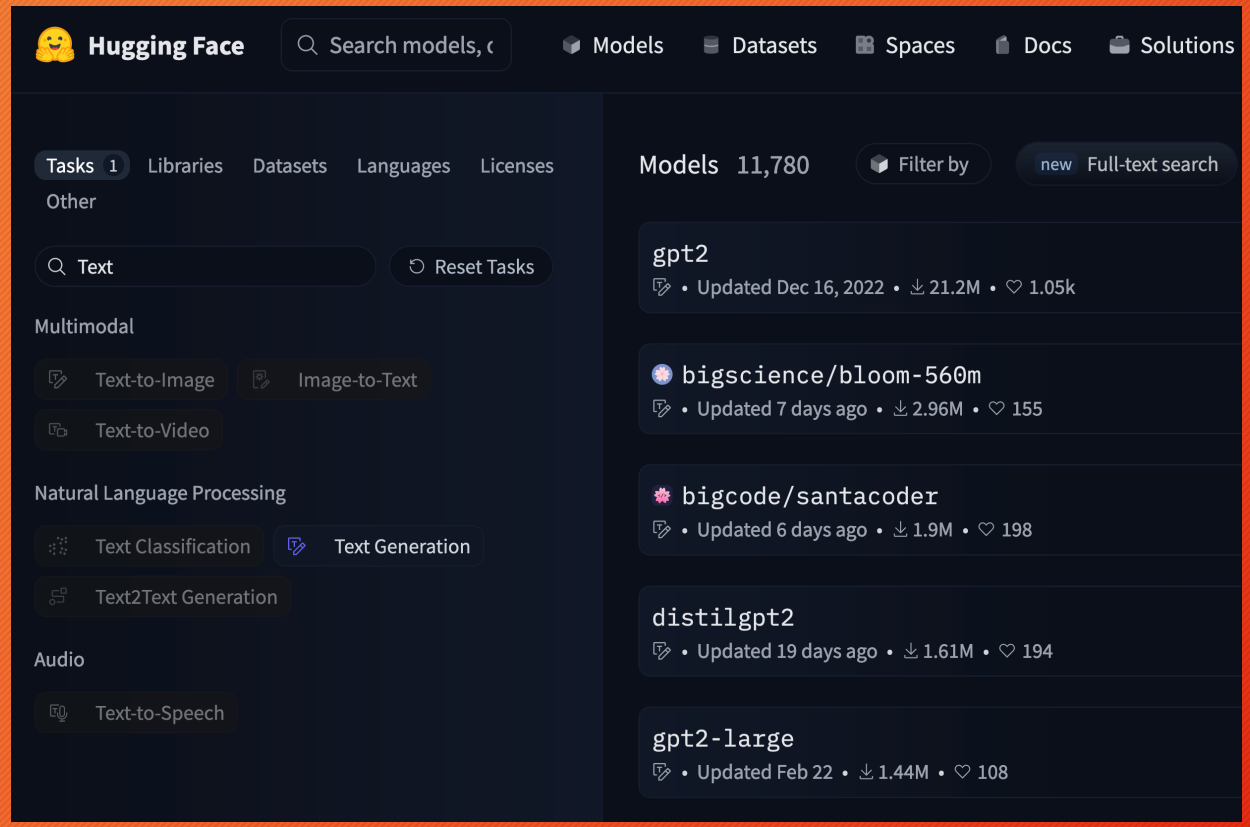


Public facing APIs

Easy to use, but …

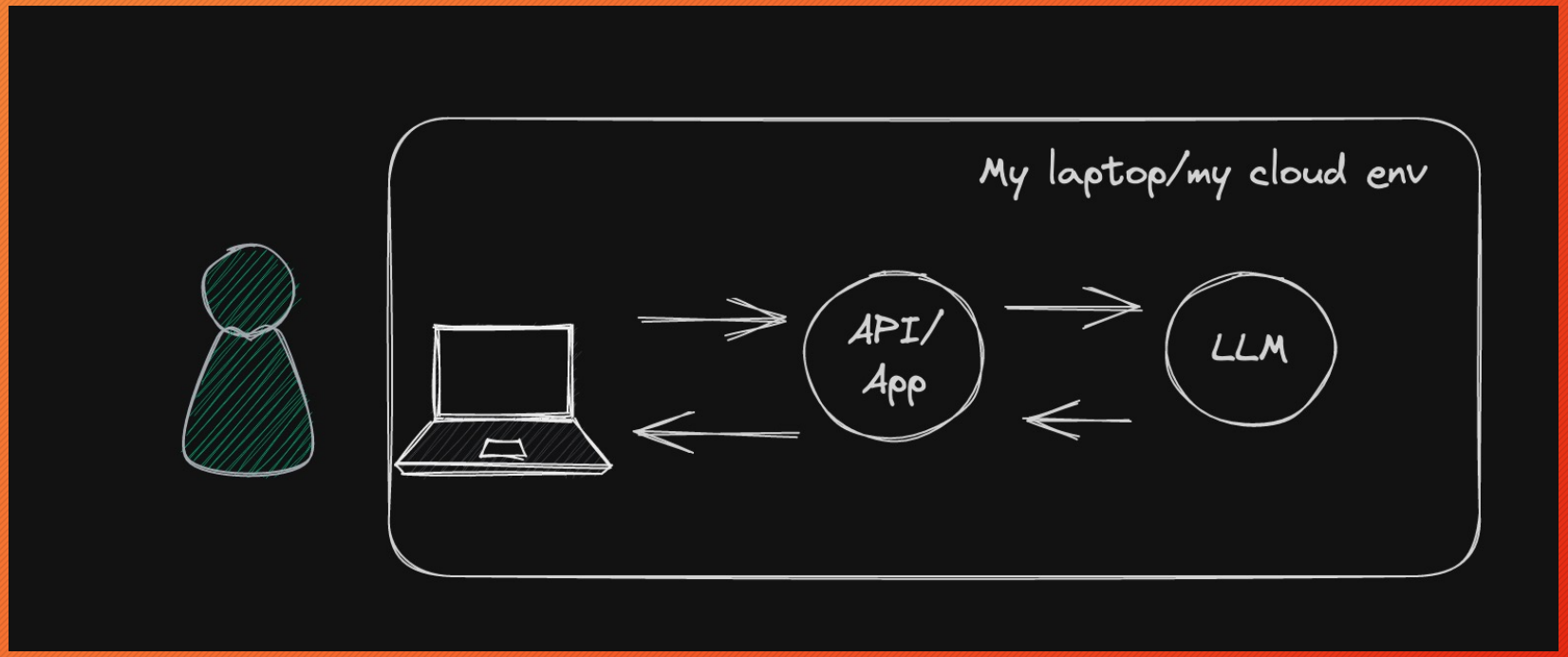


Self-hosting



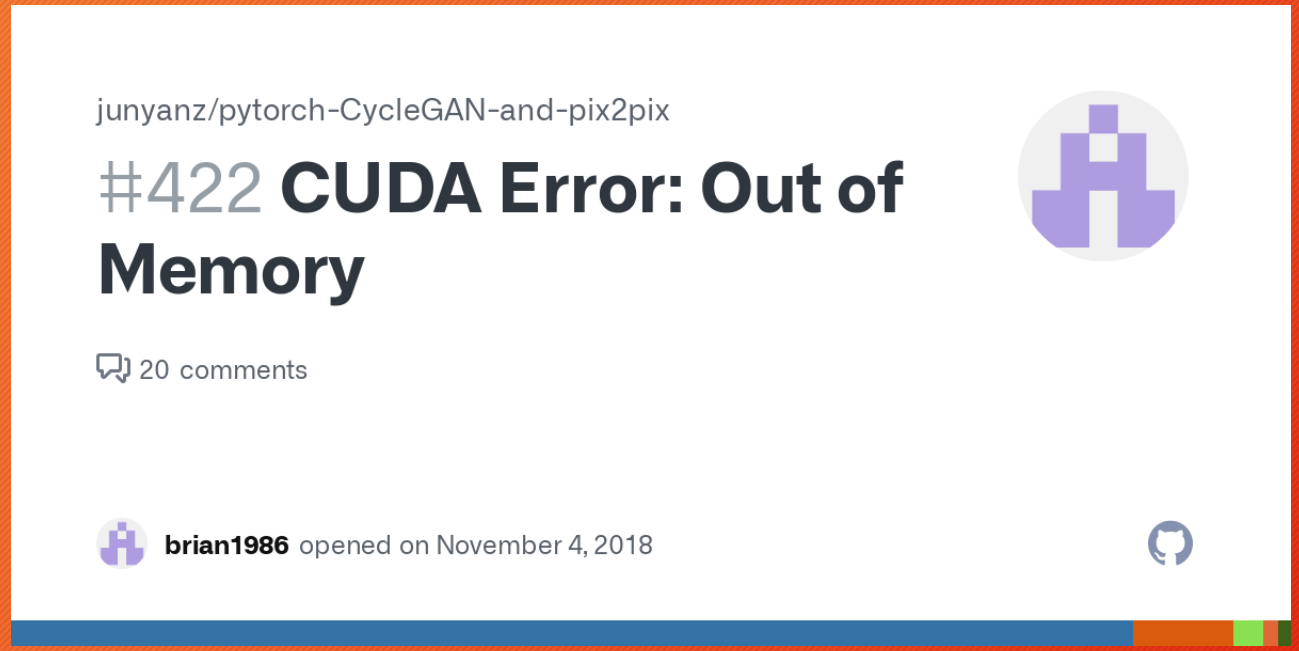
https://huggingface.co/models

Self-hosting



Private, but …

Challenge with self-hosting



Easy-to-use solutions

How 🤗 Accelerate runs very large models thanks to PyTorch

https://huggingface.co/blog/accelerate-large-models

https://github.com/TimDettmers/bitsandbytes

bitsandbytes

The bitsandbytes is a lightweight wrapper around CUDA custom functions, in particular 8-bit optimizers, matrix multiplication (LLM.int8()), and quantization functions.

Resources:

- 8-bit Optimizer Paper -- Video -- Docs
- LLM.int8() Paper -- LLM.int8() Software Blog Post -- LLM.int8() Emergent Features Blog Post

Demo in Notebook

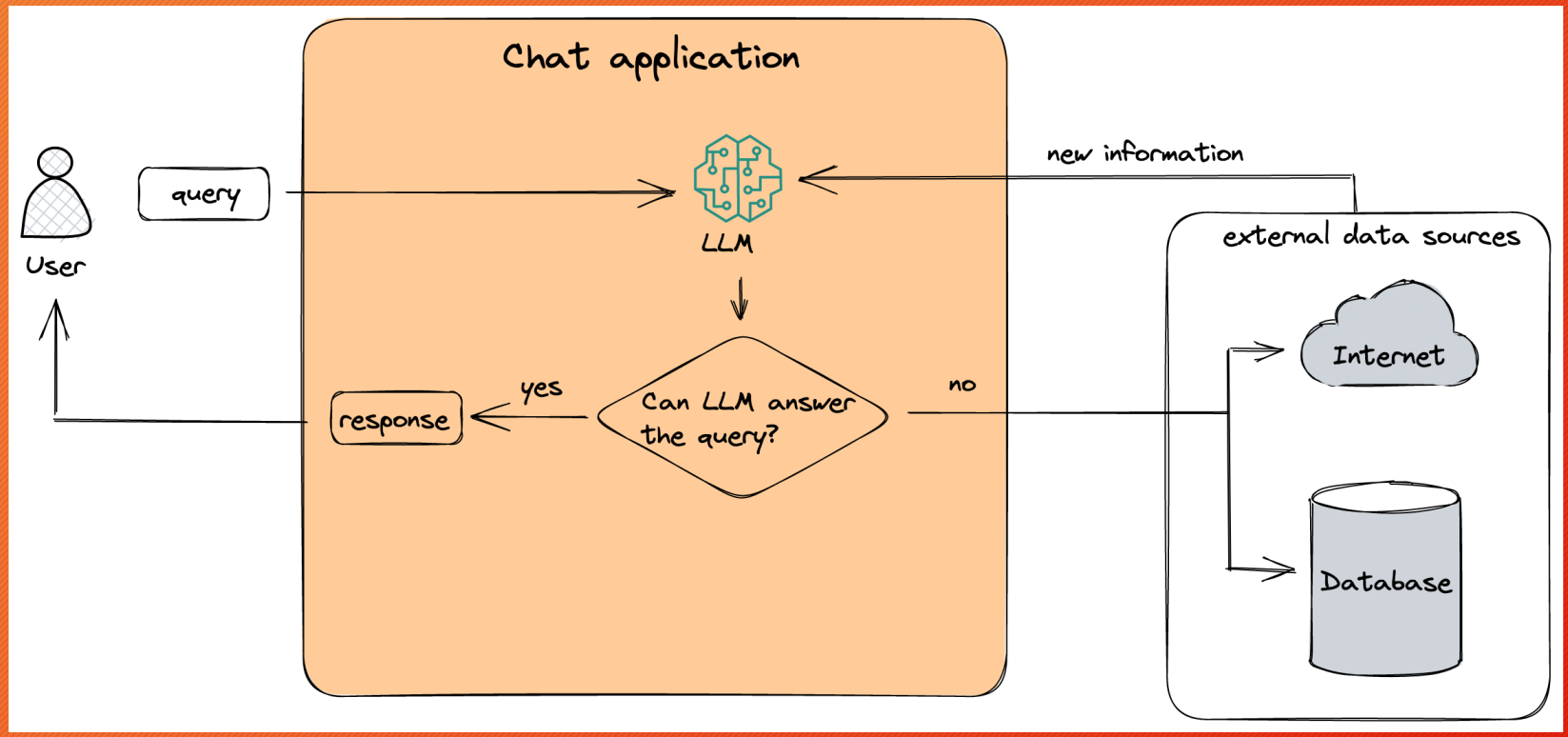
Agenda

- What is Generative AI & how does it work?
- Challenges with Large Language Models (LLMs)
  - GPU Requirements
  - Knowledge cut-off/hallucinations
- Demo: Document Chatbot for private documents
- Bias in Generative AI
- Q&A

Knowledge Cut-off

Demo in ChatGPT

If only LLMs could tell us what they need …



LangChain Demo

If only LLMs could tell us what they needed …

Supercharging Large Language Models With 🦜🔗 Langchain

Building a Modular Reasoning, Knowledge and Language (MRKL) system using prompt chaining

Heiko Hotz
Published in MLearning.ai · 8 min read · Feb 21

178 5

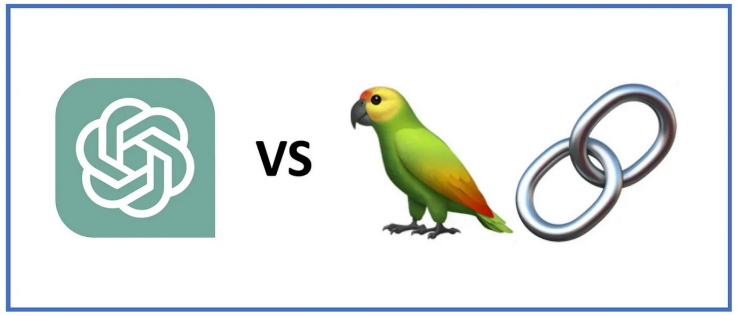
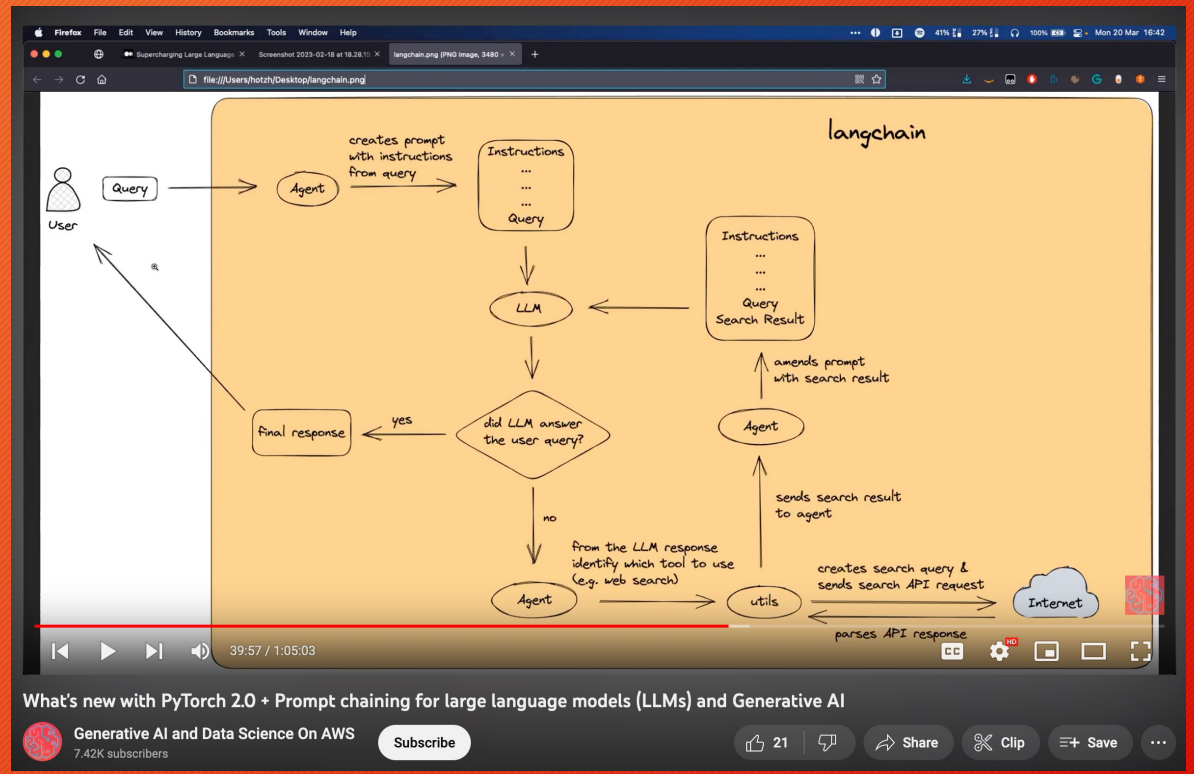


Image by author

https://medium.com/mlearning-ai/supercharging-large-language-models-with-langchain-1cac3c103b52

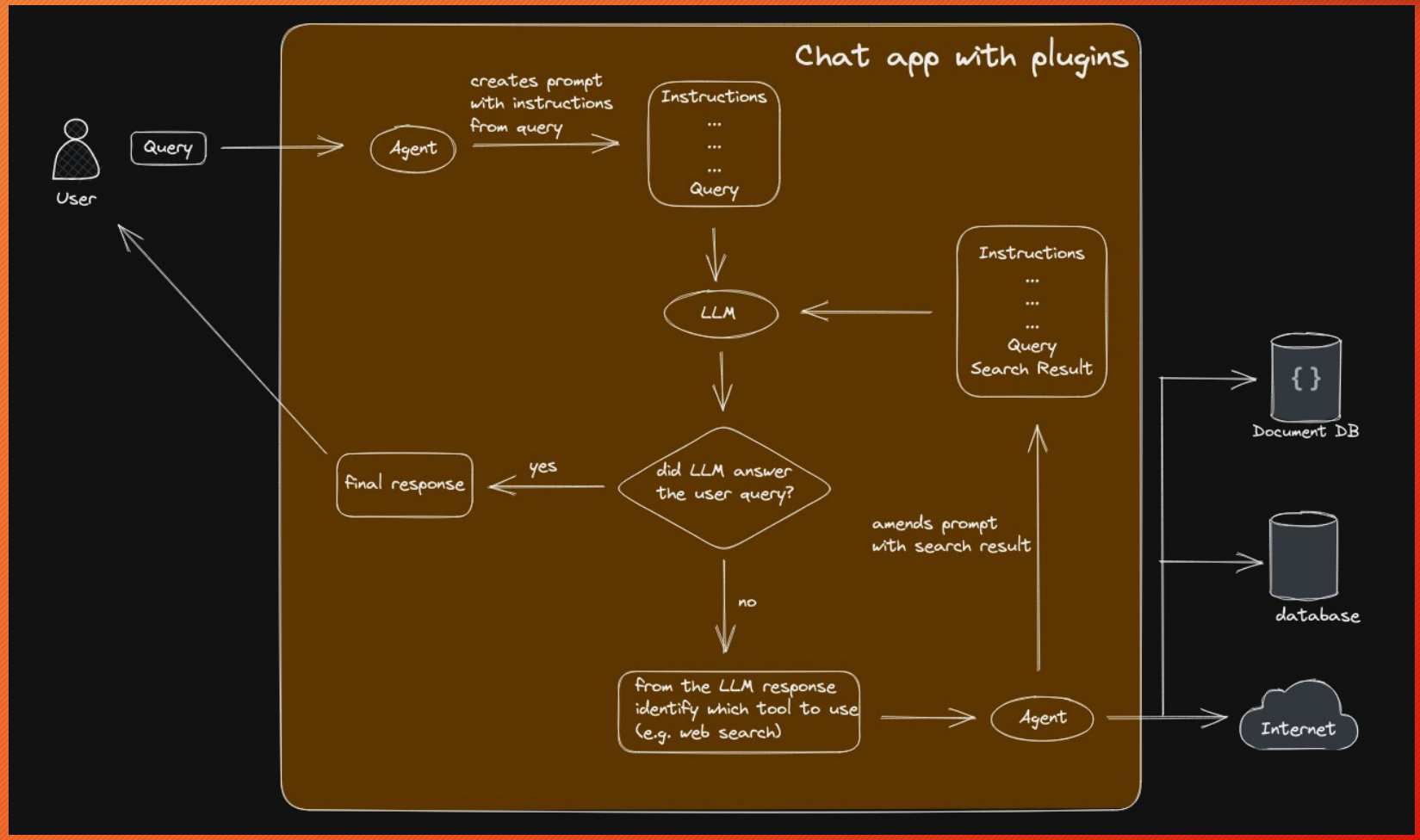


What's new with PyTorch 2.0 + Prompt chaining for large language models (LLMs) and Generative AI

Generative AI and Data Science On AWS
7.42K subscribers
Subscribe
21 Share Clip Save

https://www.youtube.com/watch?v=25dcCFvb4o4&list=PL7pBcJ870QHeNRBXdKirc4fdtbtbB5Xy-&index=4&ab_channel=GenerativeAIandDataScienceOnAWS&t=2055s

If only LLMs could tell us what they needed …



Agenda

- What is Generative AI & how does it work?
- Challenges with Large Language Models (LLMs)
- Demo: Document Chatbot for private documents
- Bias in Generative AI
- Q&A

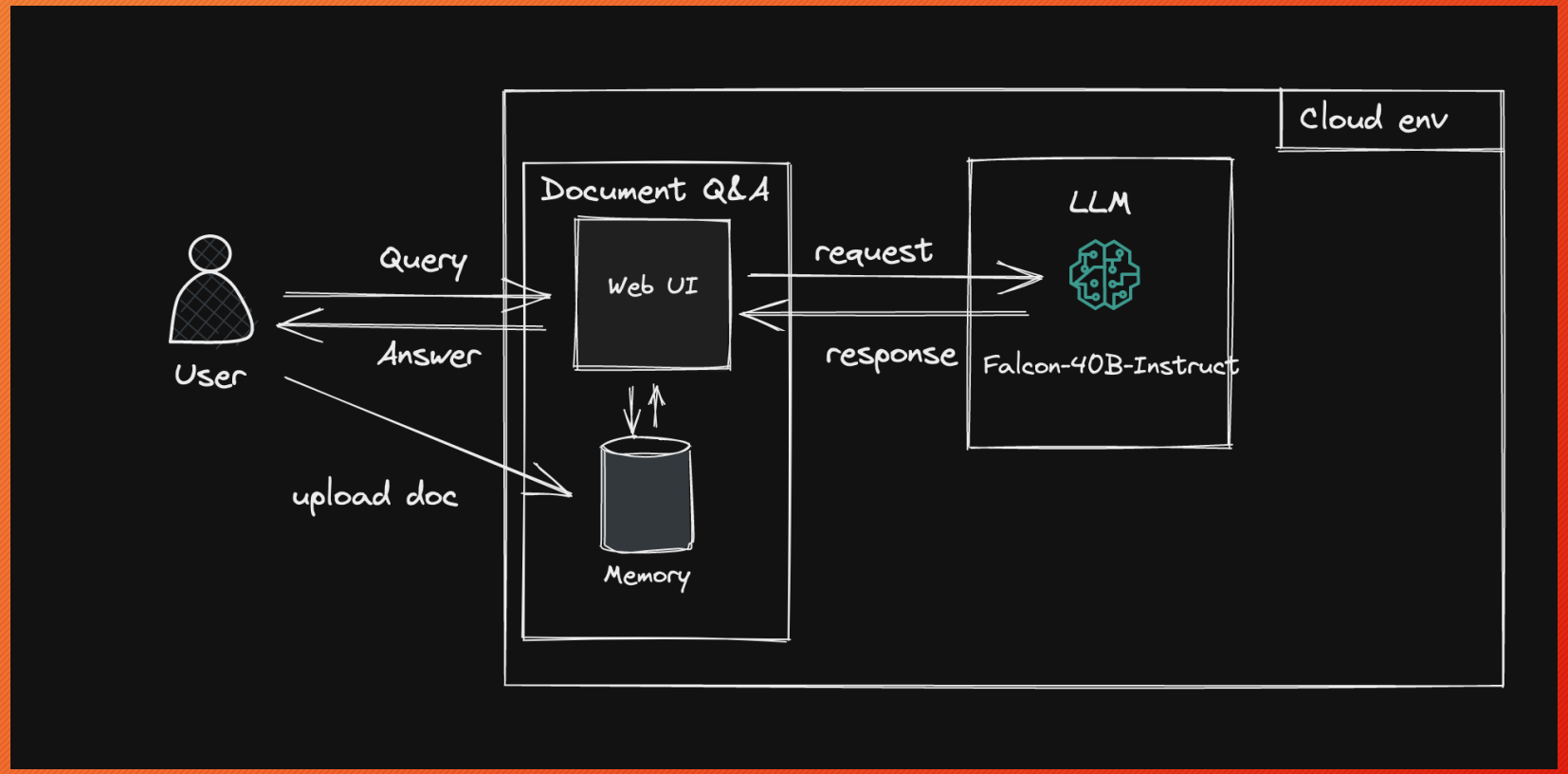
How to build applications with LLMs?

Open source!

Streamlit

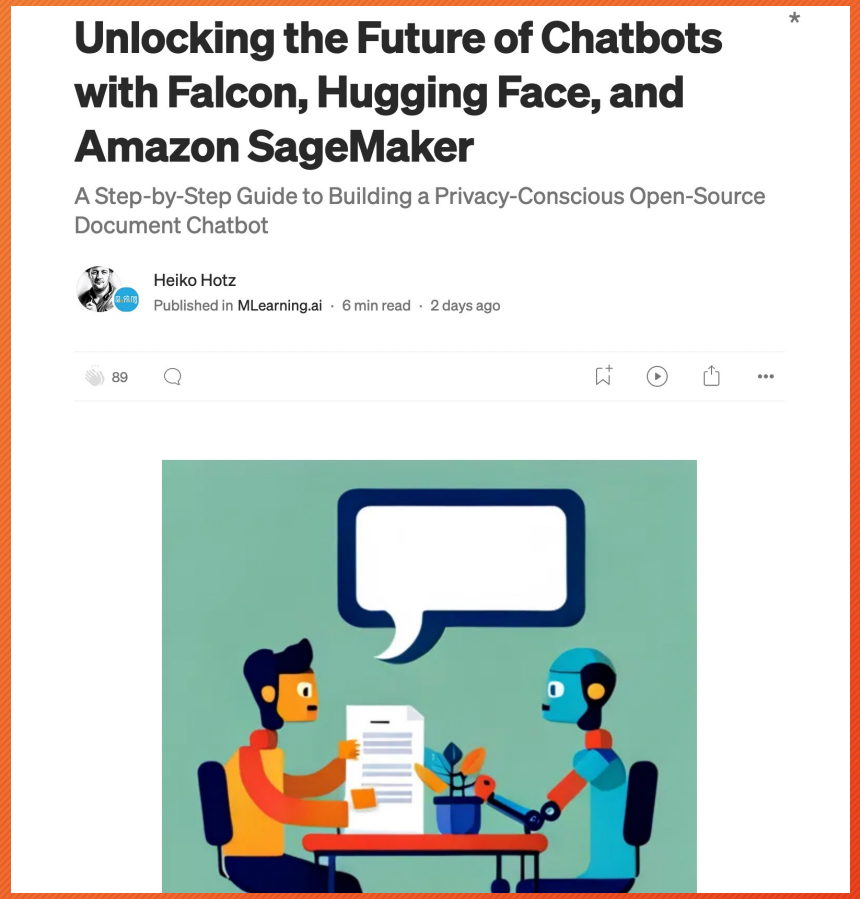
+ 🦜🔗 +

Architecture



Demo

Blog post



Unlocking the Future of Chatbots with Falcon, Hugging Face, and Amazon SageMaker

A Step-by-Step Guide to Building a Privacy-Conscious Open-Source Document Chatbot

Heiko Hotz
Published in MLearning.ai · 6 min read · 2 days ago

89

https://medium.com/mlearning-ai/unlocking-the-future-of-chatbots-with-falcon-hugging-face-and-amazon-sagemaker-cf6bd8aeba54

Agenda

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

Bias: Demo in Playground

Questions?

Connect with me on LinkedIn:

https://www.linkedin.com/in/heikohotz/