

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

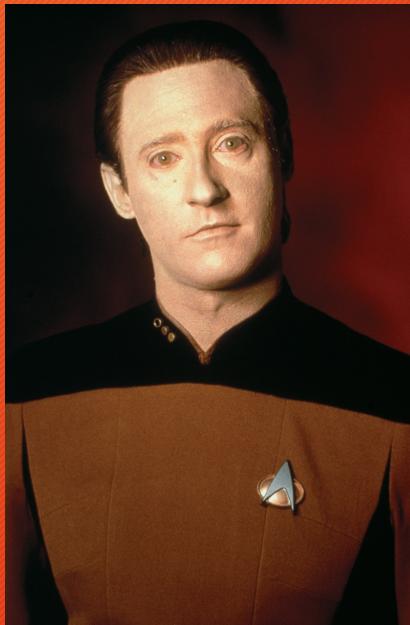
- What is Generative AI & how does it work?
- Challenges with Large Language Models (LLMs)
- Demo: Document Chatbot
- Bias in Generative AI
- Q&A

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>

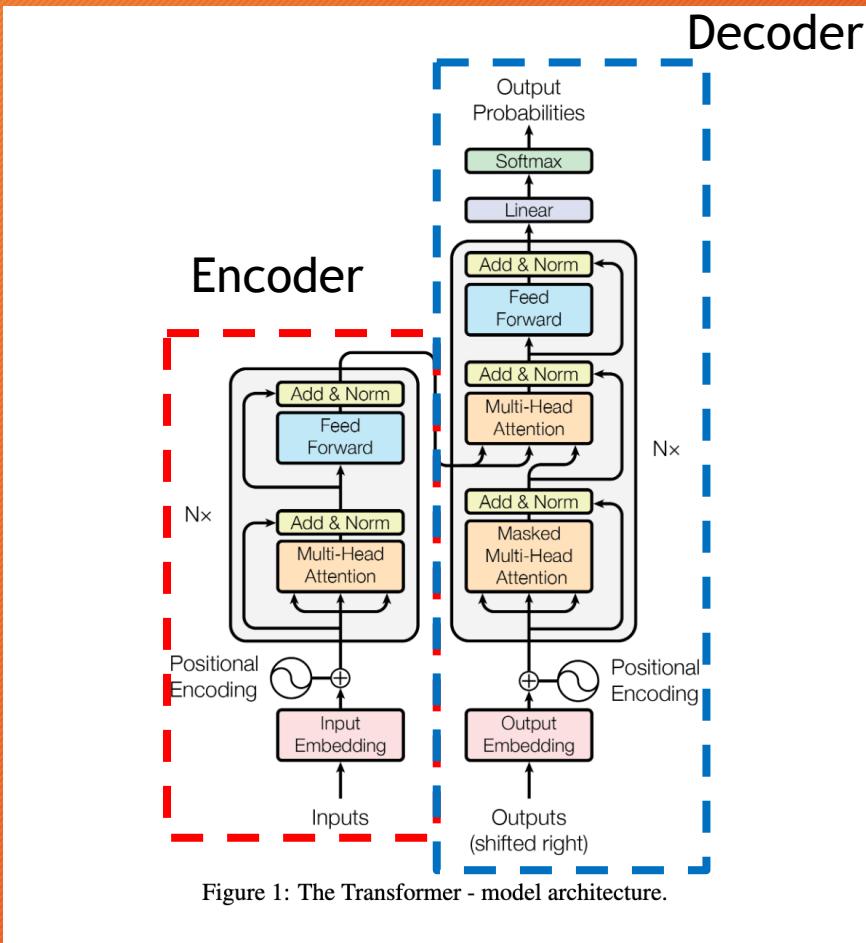
Generative AI - What is it?



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

How does it work?

Transformer Architecture



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

Encoder-only models

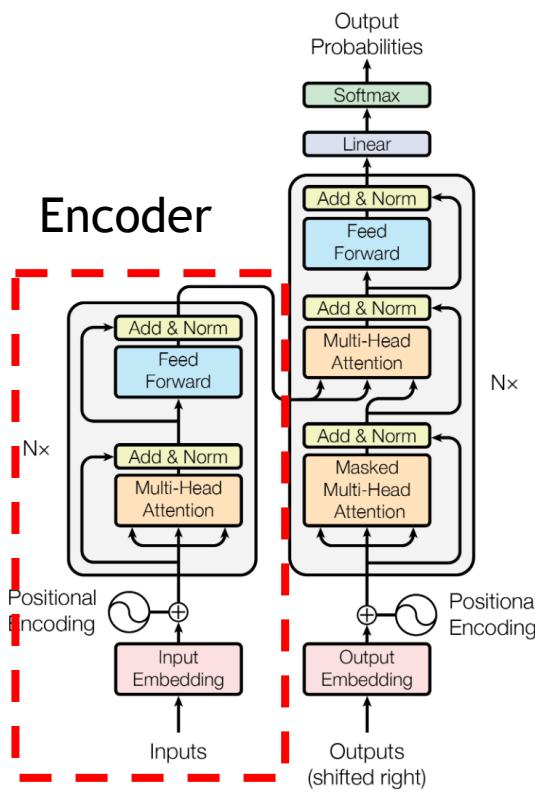


Figure 1: The Transformer - model architecture.

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)

Decoder-only models

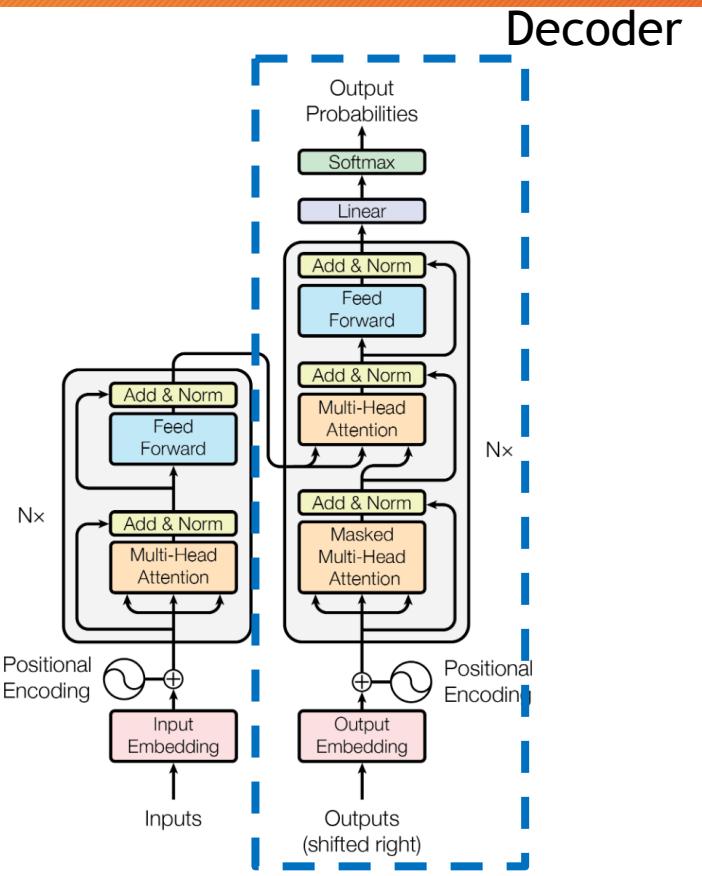
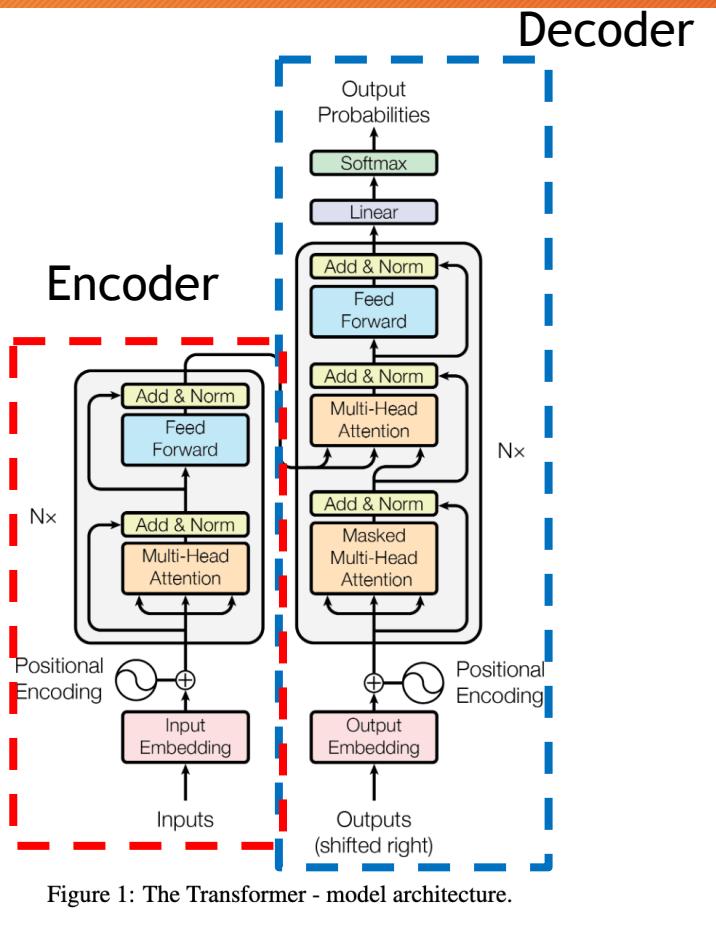


Figure 1: The Transformer - model architecture.

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



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

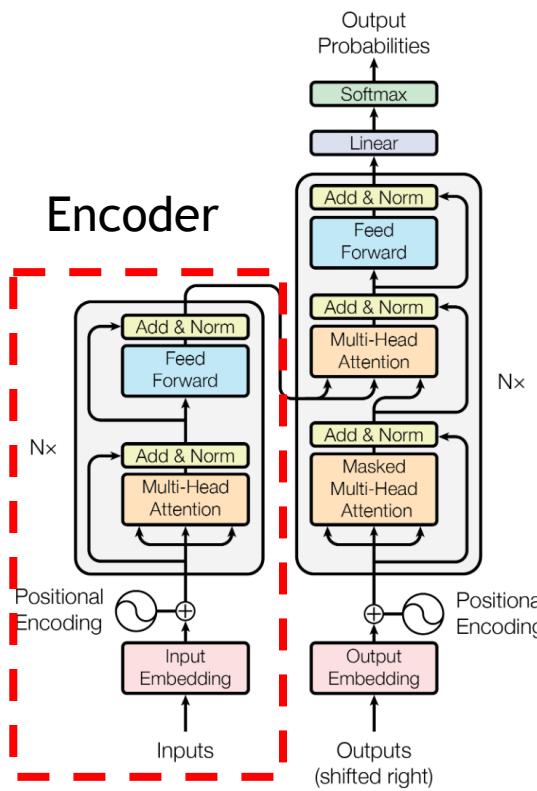


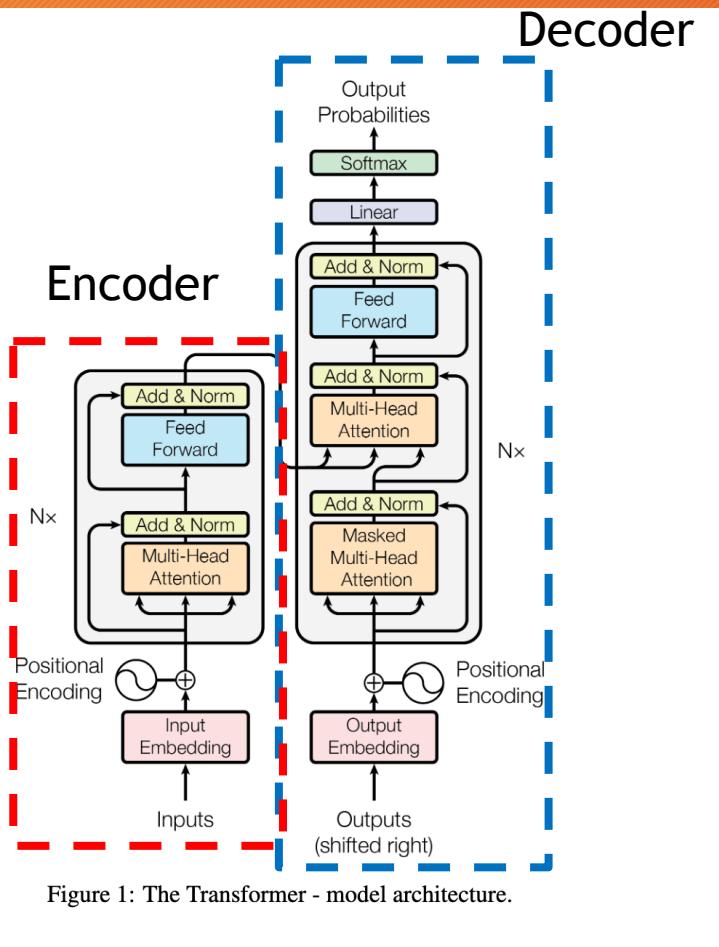
Figure 1: The Transformer - model architecture.

Example:

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



Encoder-decoder models

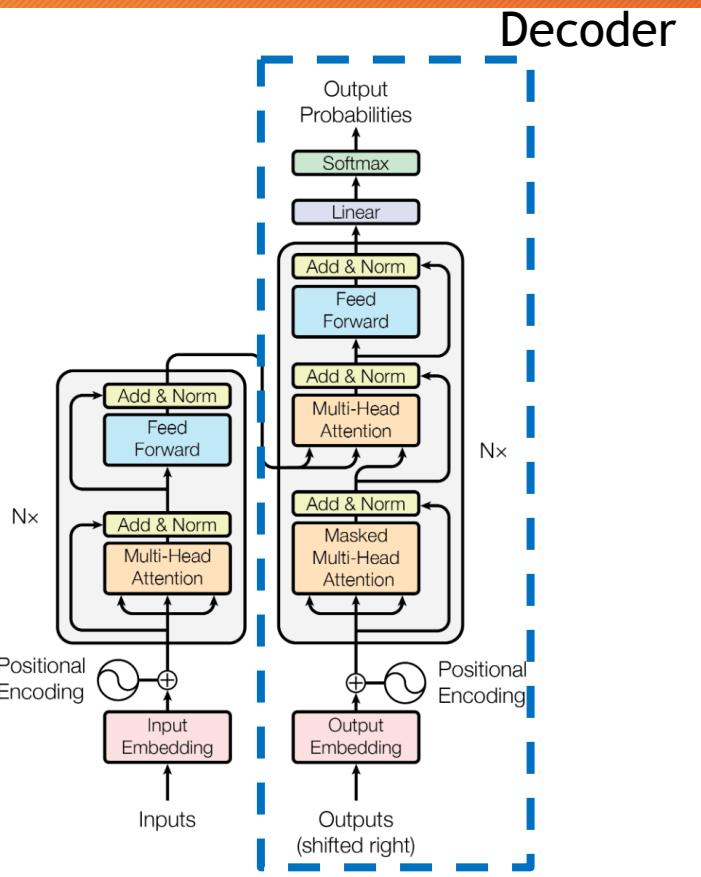


Example:

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

German: "Ich habe John das Buch gegeben."

Decoder-only models



Example:

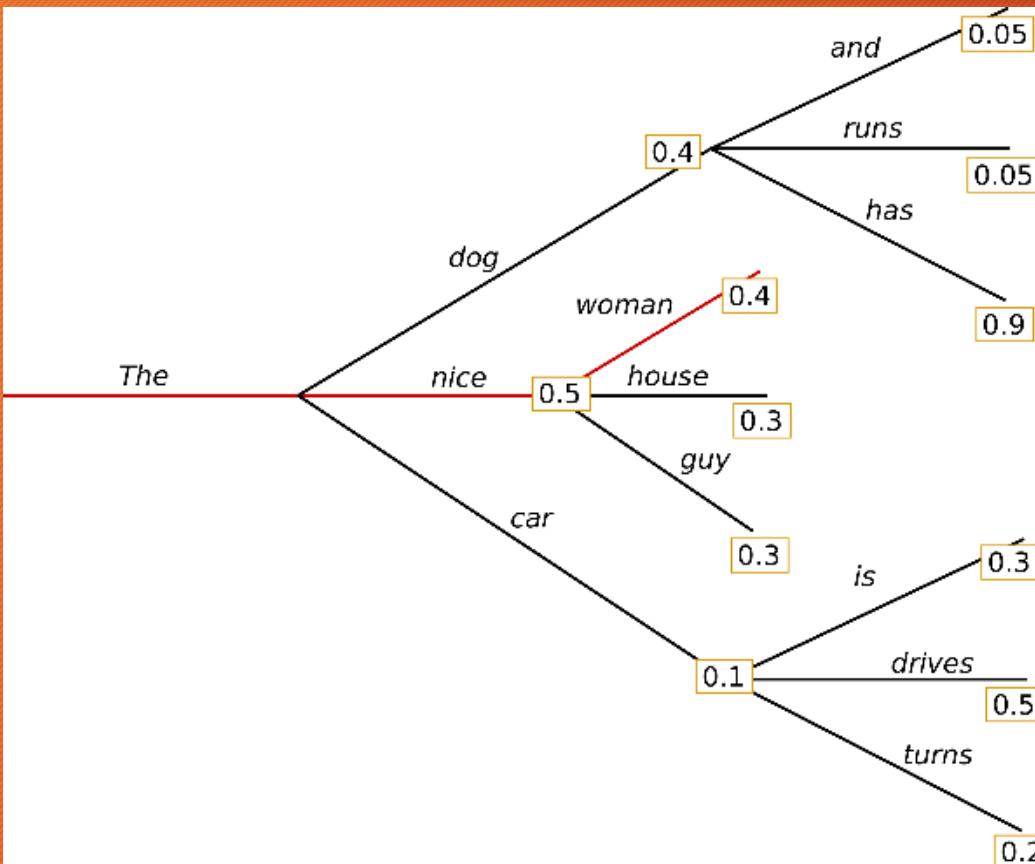
English: "I went to the bakery and bought ..."

How do decoder-only models work?

How the decoder produces text

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

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

AI21labs

ANTHROPIC

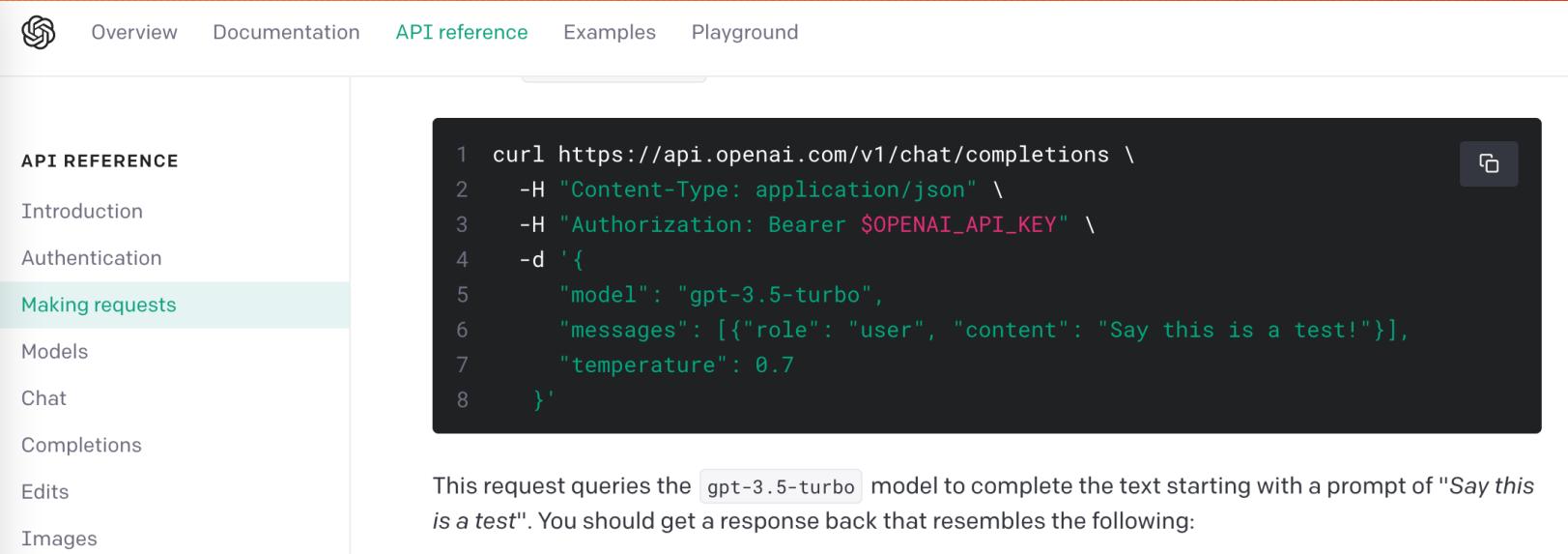
Self-hosted (locally or cloud)



Google Cloud Platform



Public facing APIs



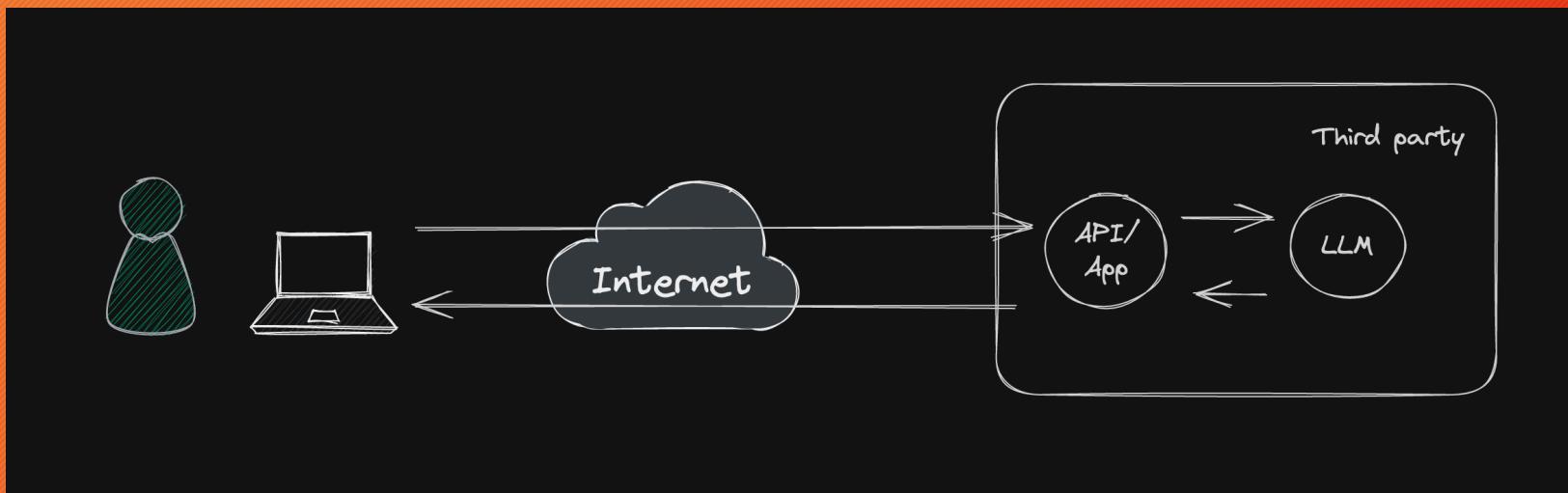
The screenshot shows a web browser displaying the OpenAI API reference documentation. The top navigation bar includes links for Overview, Documentation, API reference (which is the active page), Examples, and Playground. A sidebar on the left is titled 'API REFERENCE' and lists several categories: Introduction, Authentication, Making requests (which is highlighted in green), Models, Chat, Completions, Edits, and Images. The main content area contains a code block showing a curl command to make a request to the /v1/chat/completions endpoint. The command includes headers for Content-Type and Authorization, a JSON payload with a model, messages, and temperature, and a closing brace. Below the code block, a text description explains that the request queries the gpt-3.5-turbo model with a specific prompt and expects a response resembling the following.

```
1 curl https://api.openai.com/v1/chat/completions \
2   -H "Content-Type: application/json" \
3   -H "Authorization: Bearer $OPENAI_API_KEY" \
4   -d '{
5     "model": "gpt-3.5-turbo",
6     "messages": [{"role": "user", "content": "Say this is a test!"}],
7     "temperature": 0.7
8   }'
```

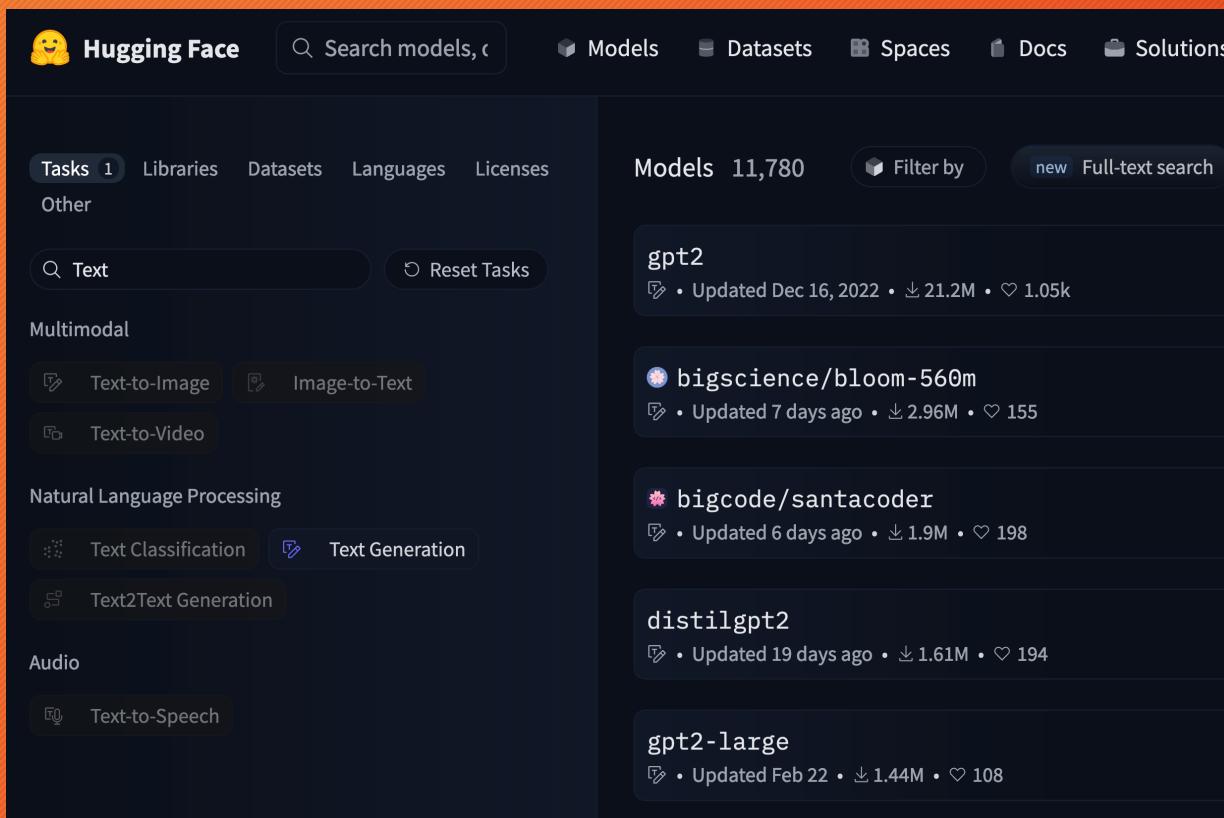
This request queries the `gpt-3.5-turbo` model to complete the text starting with a prompt of "Say this is a test!". You should get a response back that resembles the following:

Public facing APIs

Easy to use, but ...



Self-hosting

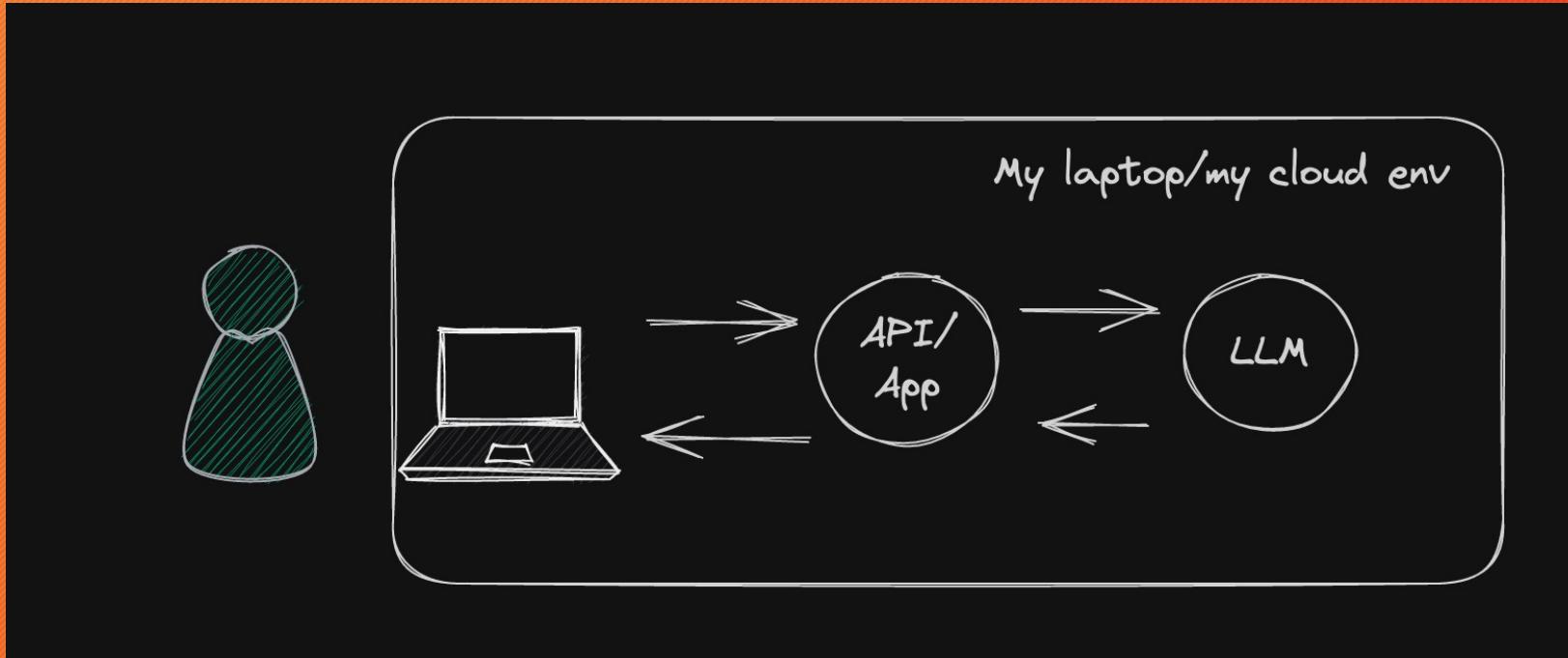


The screenshot shows the Hugging Face website interface. The top navigation bar includes a logo, a search bar, and links for Models, Datasets, Spaces, Docs, and Solutions. The left sidebar features a 'Tasks' section with 1 item, and categories for Libraries, Datasets, Languages, Licenses, and Other. Below this are sections for Text, Multimodal (Text-to-Image, Image-to-Text, Text-to-Video), Natural Language Processing (Text Classification, Text Generation, Text2Text Generation), and Audio (Text-to-Speech). The main content area is titled 'Models 11,780' and includes a 'Filter by' button and a 'new Full-text search' button. It lists several models: '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), and 'gpt2-large' (updated Feb 22, 1.44M, 108).

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

<https://huggingface.co/models>

Self-hosting



Private, but ...

Challenge with self-hosting

junyanz/pytorch-CycleGAN-and-pix2pix

#422 **CUDA Error: Out of Memory**

20 comments

 **brian1986** opened on November 4, 2018



Easy-to-use solutions

How 😊 Accelerate runs very large models thanks to PyTorch

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

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

<https://github.com/TimDettmers/bitsandbytes>

Demo in Notebook

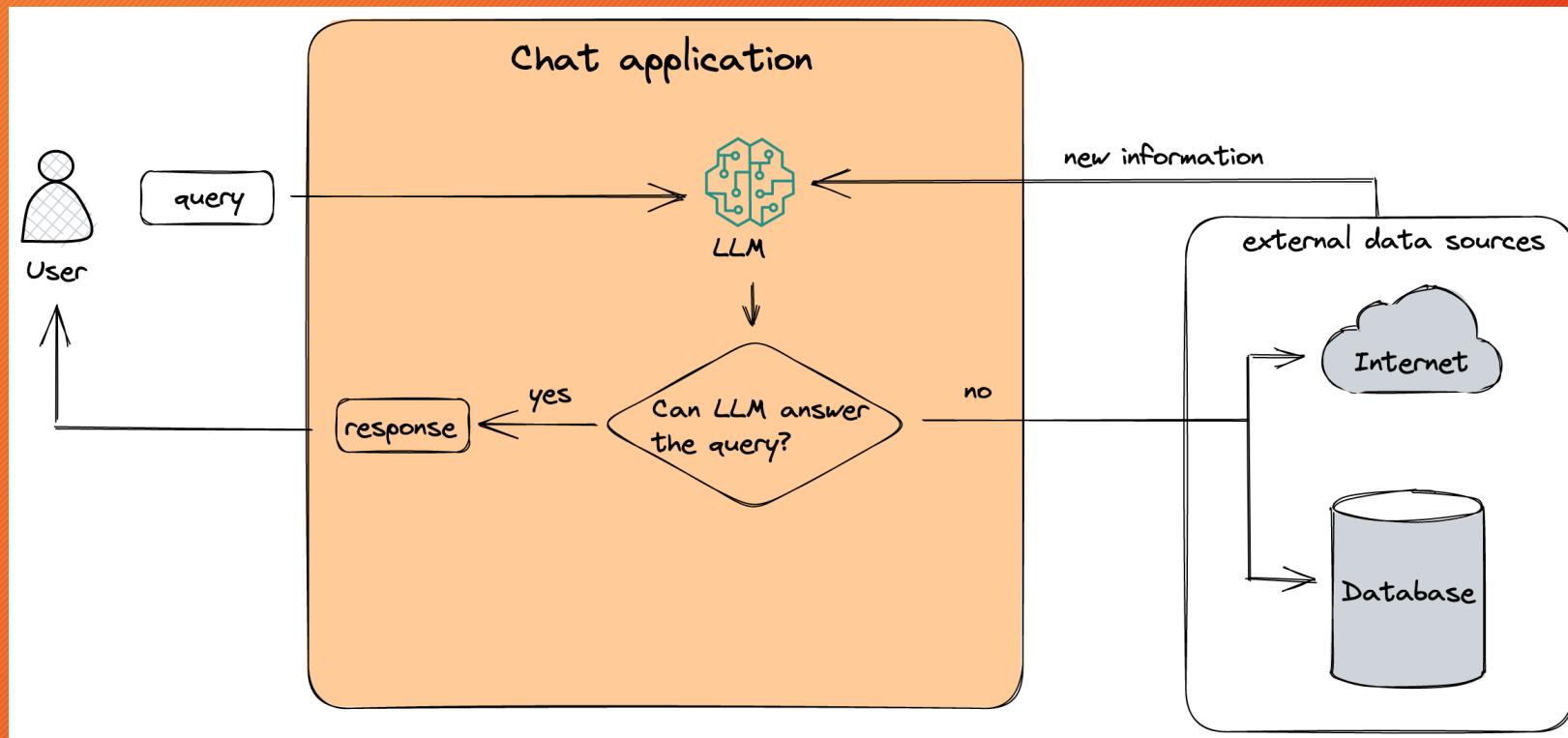
Agenda

- 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

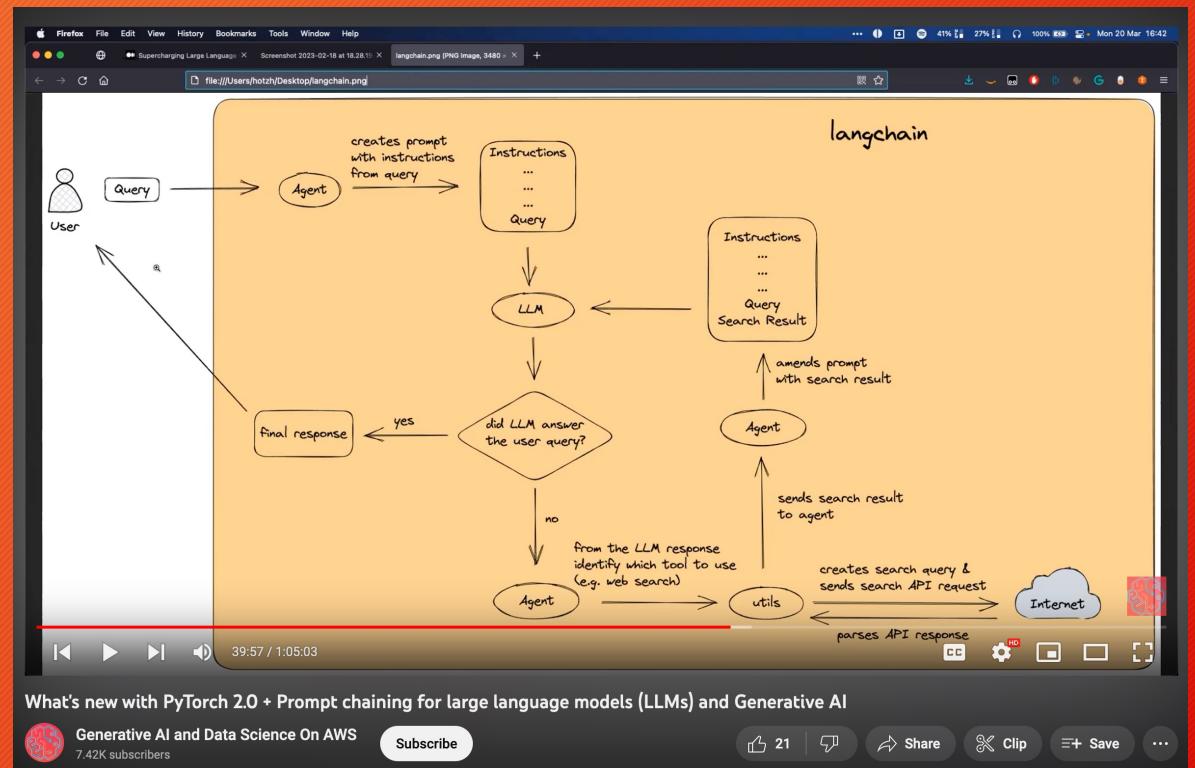
Heiko Hotz
Published in MLearning.ai · 8 min read · Feb 21

178 Q 5



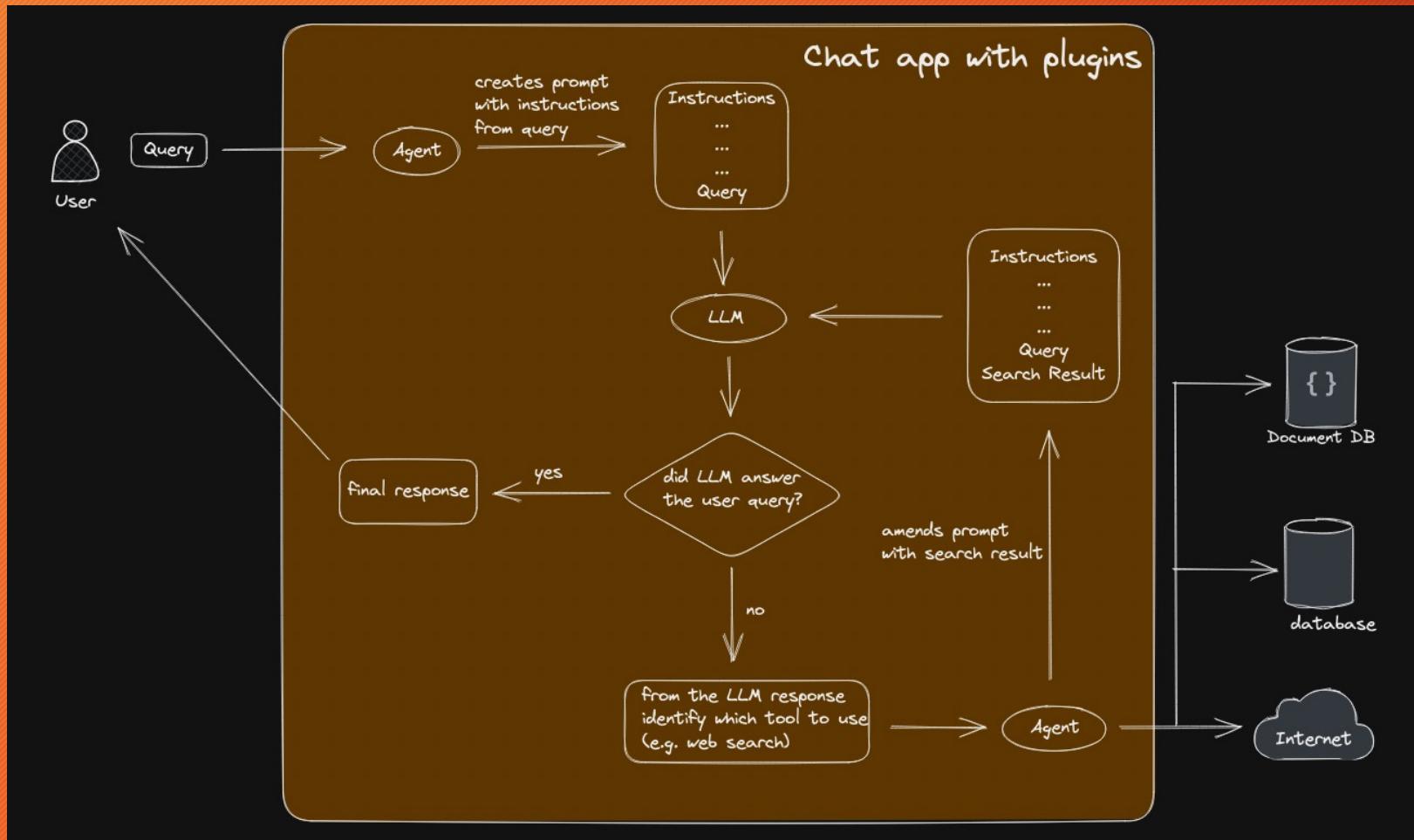
Image by author

[https://medium.com/mlarning-ai/supercharging-large-language-models-with-langchain-1cac3c103b52](https://medium.com/mlearning-ai/supercharging-large-language-models-with-langchain-1cac3c103b52)



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

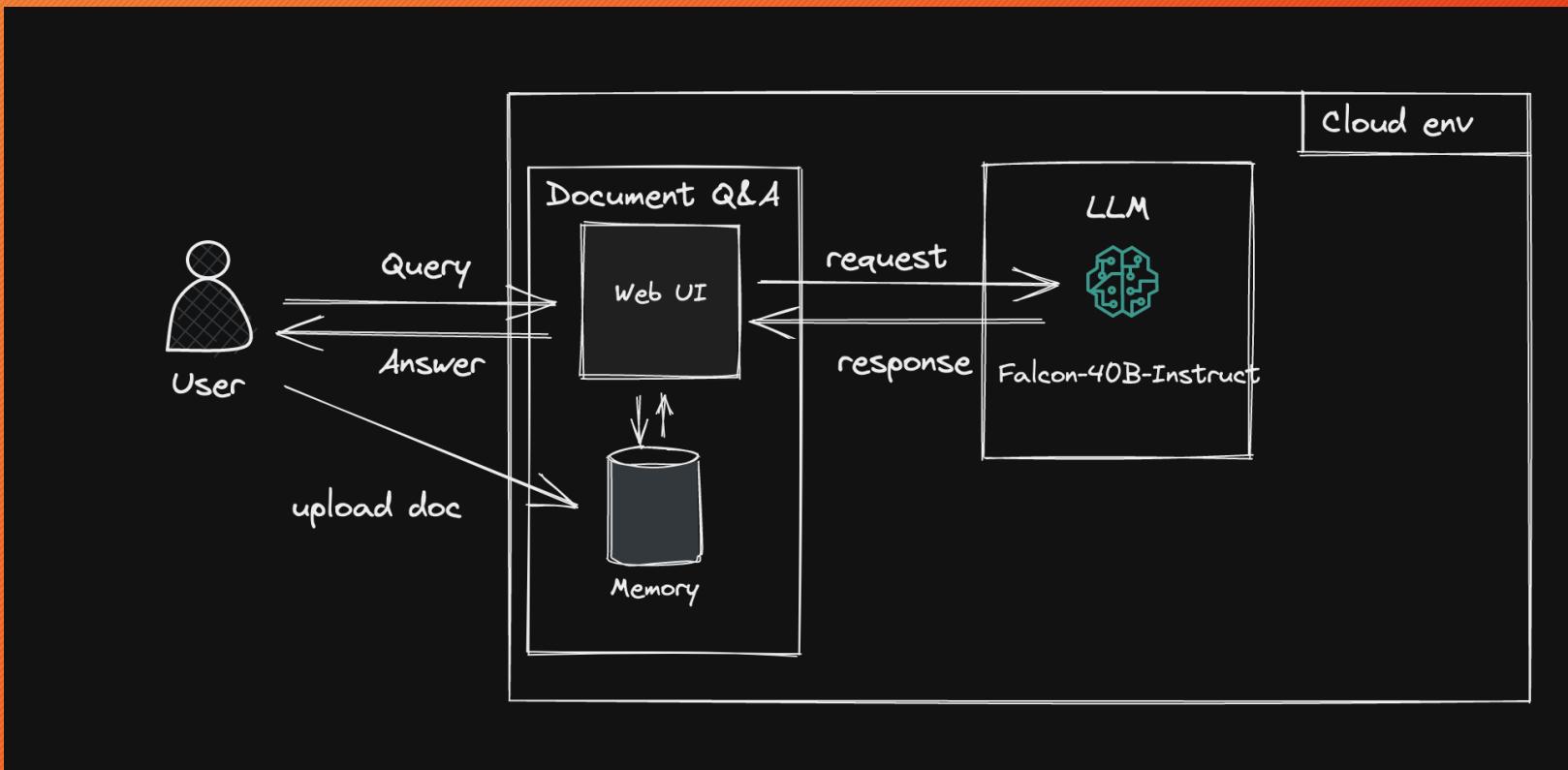
How to build applications with LLMs?

Open source!



Introducing Falcon LLM

Architecture



Demo

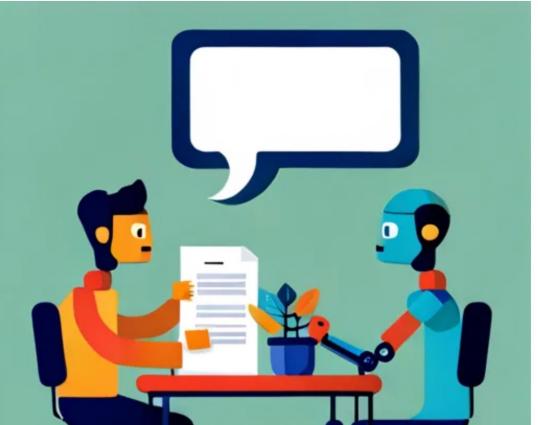
Blog post

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/>

