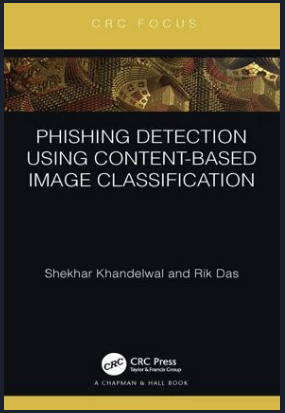
Shekhar Khandelwal

Senior Data Scientist
Hamburg, Germany

15+ years industry experience
Ex Accenture - IBM - EY
Currently MMT GmbH, Germany

Linkedin: https://www.linkedin.com/in/shekhar-khandelwal-05310130/
Medium: https://medium.com/@khandelwal-shekhar

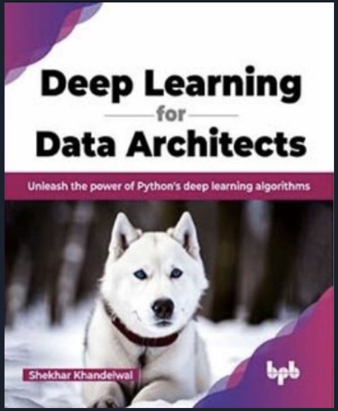


CRC FOCUS

PHISHING DETECTION USING CONTENT-BASED IMAGE CLASSIFICATION

Shekhar Khandelwal and Rik Das

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

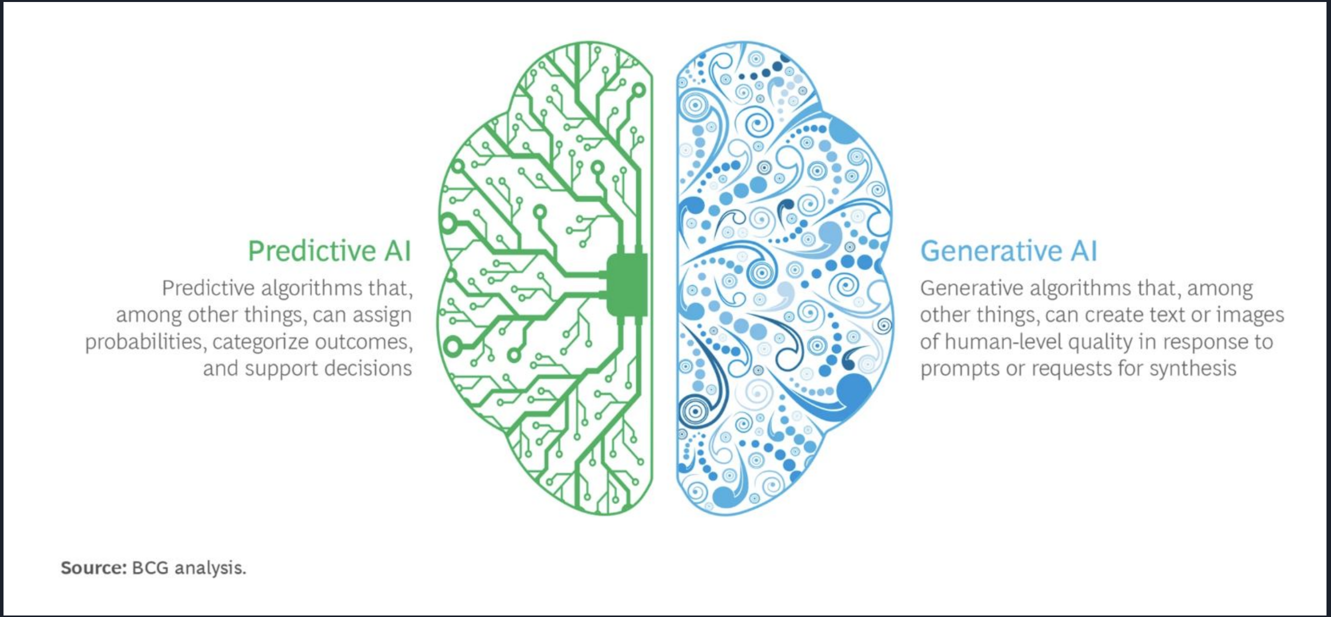


Deep Learning for Data Architects

Unleash the power of Python's deep learning algorithms

Shekhar Khandelwal

bpb



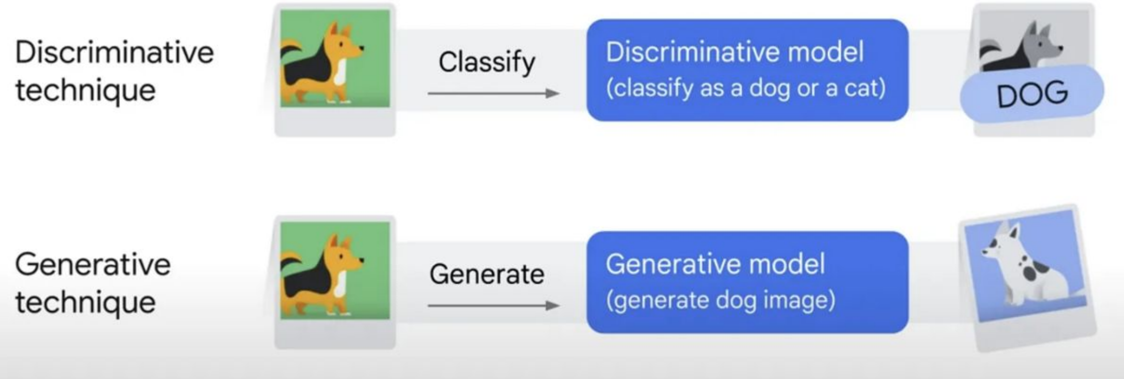
Predictive AI

Predictive algorithms that, among other things, can assign probabilities, categorize outcomes, and support decisions

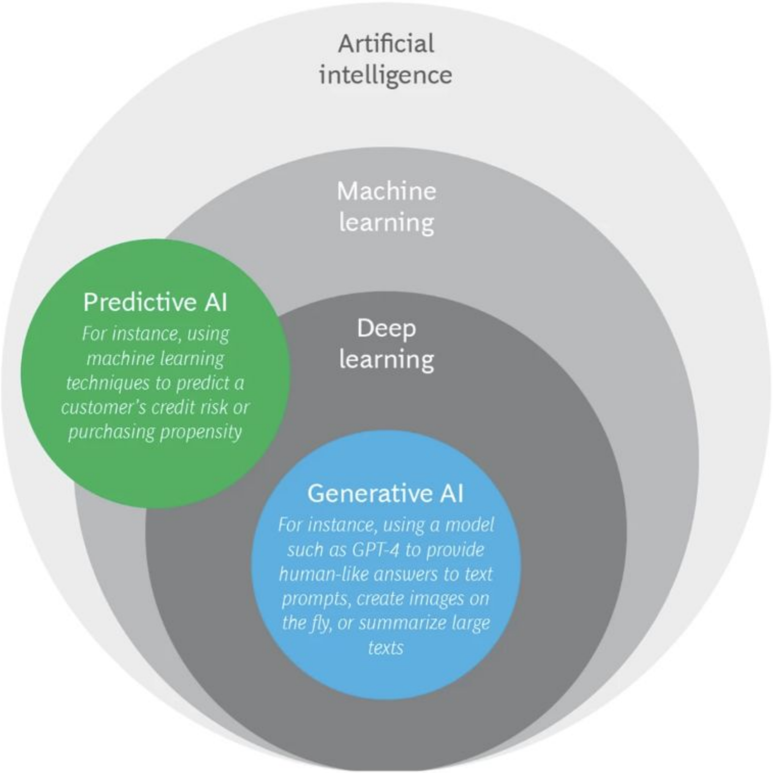
Generative AI

Generative algorithms that, among other things, can create text or images of human-level quality in response to prompts or requests for synthesis

Source: BCG analysis.



Disclaimer: All images are borrowed from web for better conceptual understanding.



Artificial intelligence

Machine learning

Deep learning

Predictive AI

For instance, using machine learning techniques to predict a customer's credit risk or purchasing propensity

Generative AI

For instance, using a model such as GPT-4 to provide human-like answers to text prompts, create images on the fly, or summarize large texts

Artificial intelligence. A broad term for nonhuman "intelligence" or problem-solving ability embedded in machines or software.

Machine learning. A subset of artificial intelligence algorithms in which computers figure out how to tackle problems and discover solutions independently, often by using artificial neural networks.

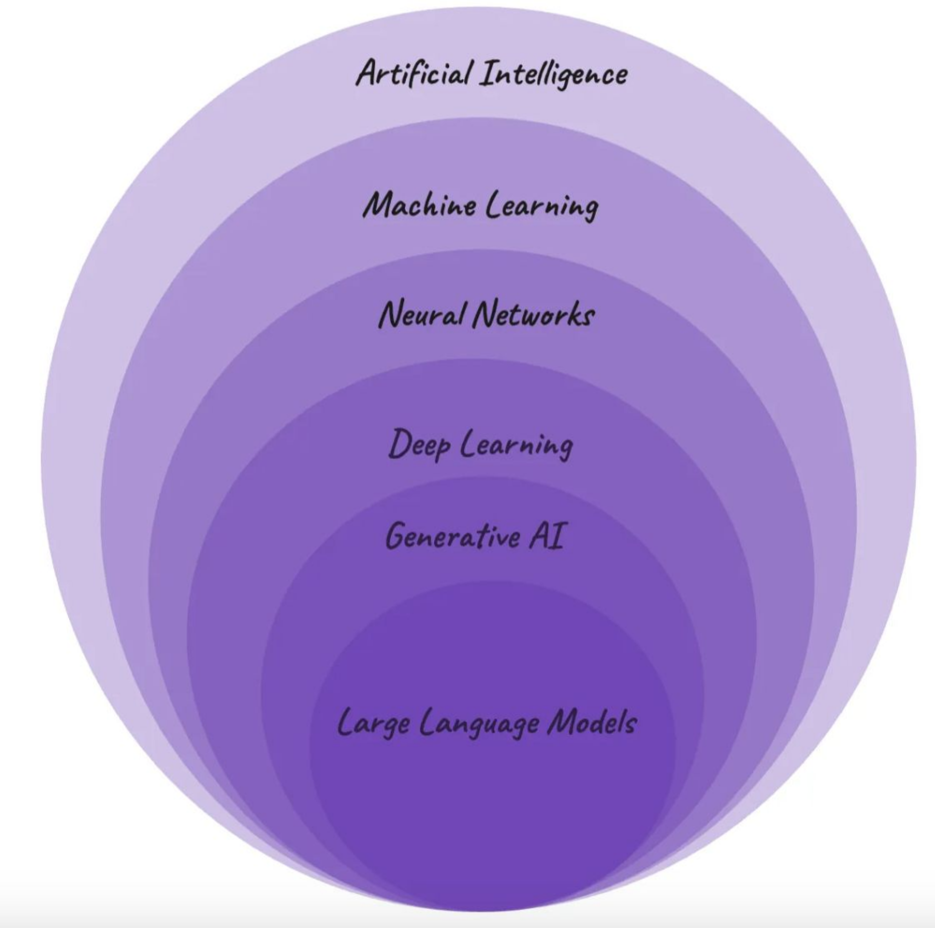
Deep learning. A subset of machine learning algorithms in which computers leverage multilayer ("deep") artificial neural networks to perform complex learning tasks that in many cases involve large amounts of text or images.

Generative AI (GenAI). A subset of deep learning algorithms in which computers focus on generating apparently new, realistic content from unstructured inputs such as text, images, or audio. Widely known examples include ChatGPT (for text) and DALL-E (for images).

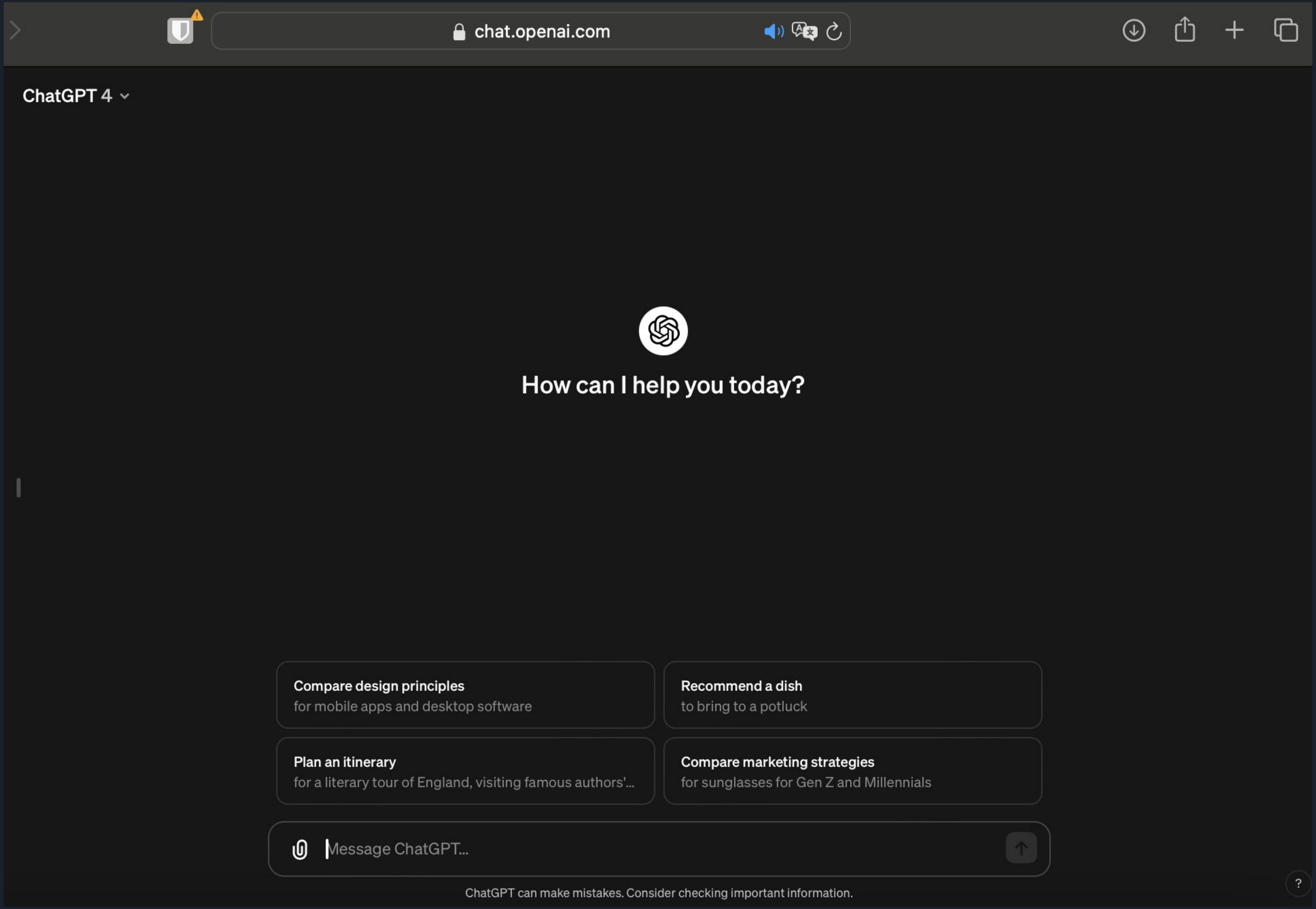
Predictive AI. Predictive modeling techniques that are widespread in industries such as banking and that can leverage a variety of AI techniques, sometimes including machine learning or deep learning.

Source: BCG analysis.

Disclaimer: All images are borrowed from web for better conceptual understanding.



Disclaimer: All images are borrowed from web for better conceptual understanding.



chat.openai.com

ChatGPT 4

How can I help you today?

Compare design principles
for mobile apps and desktop software

Recommend a dish
to bring to a potluck

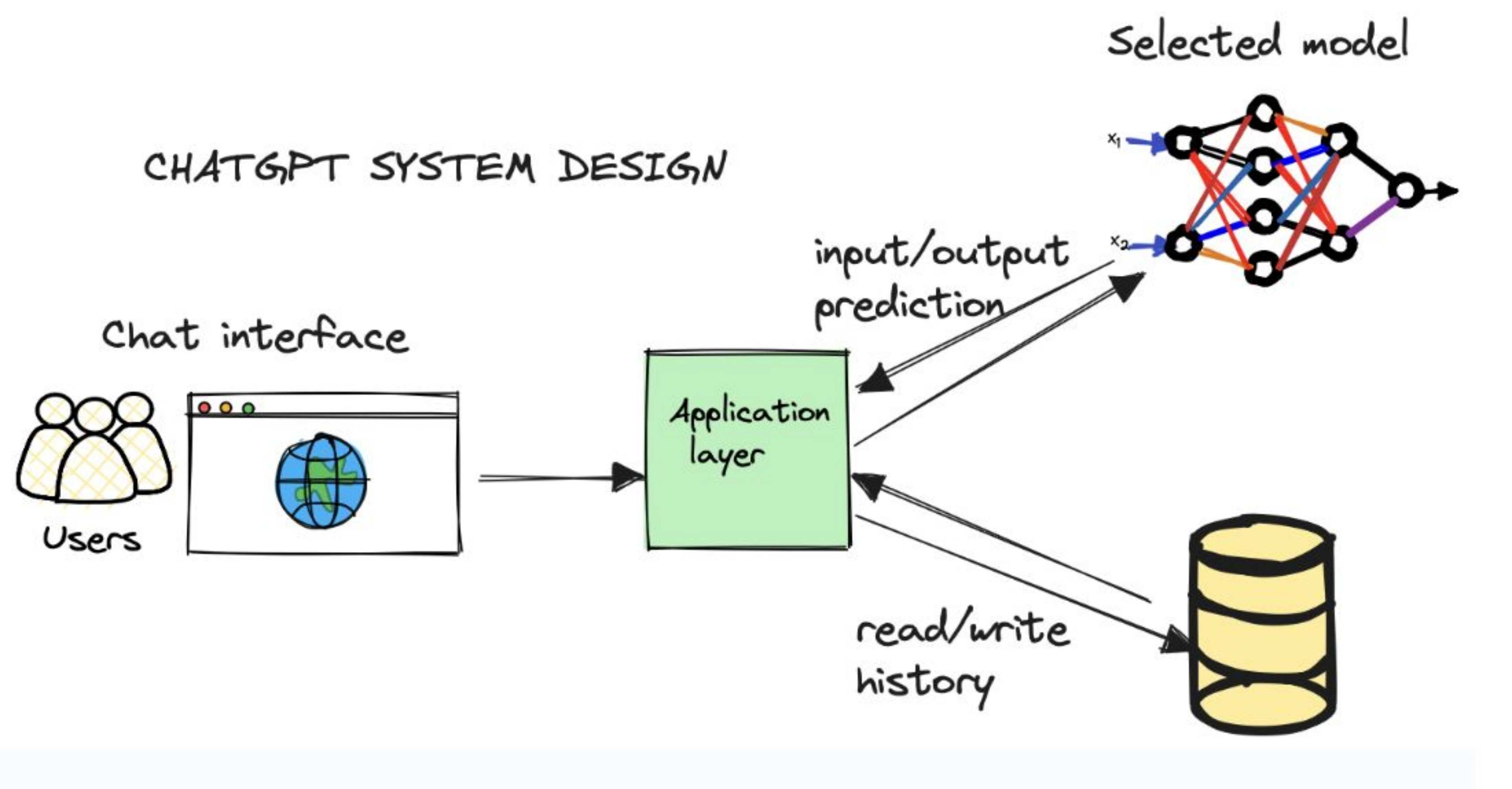
Plan an itinerary
for a literary tour of England, visiting famous authors'...

Compare marketing strategies
for sunglasses for Gen Z and Millennials

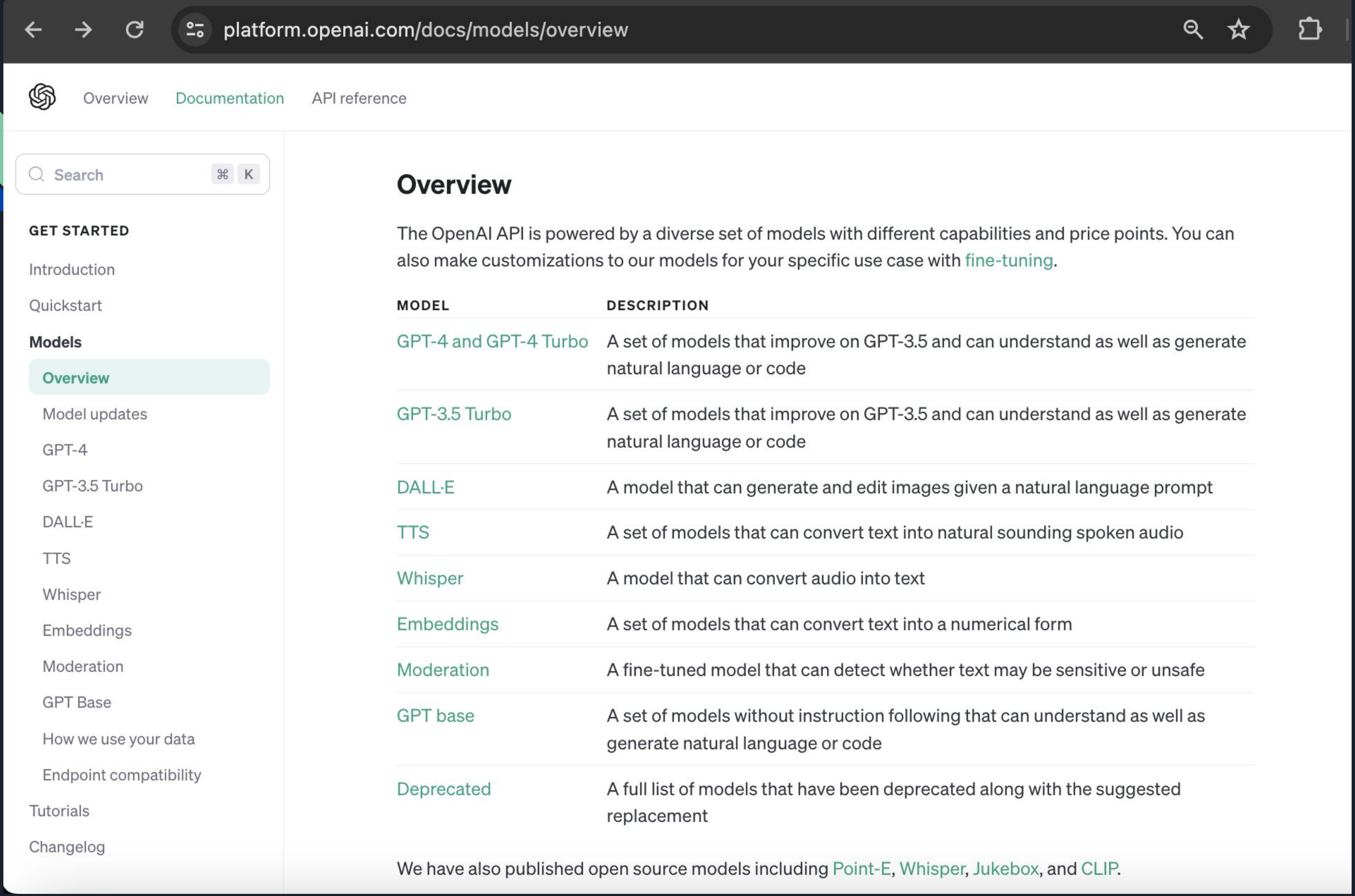
Message ChatGPT...

ChatGPT can make mistakes. Consider checking important information.

Image by author



Disclaimer: All images are borrowed from web for better conceptual understanding.



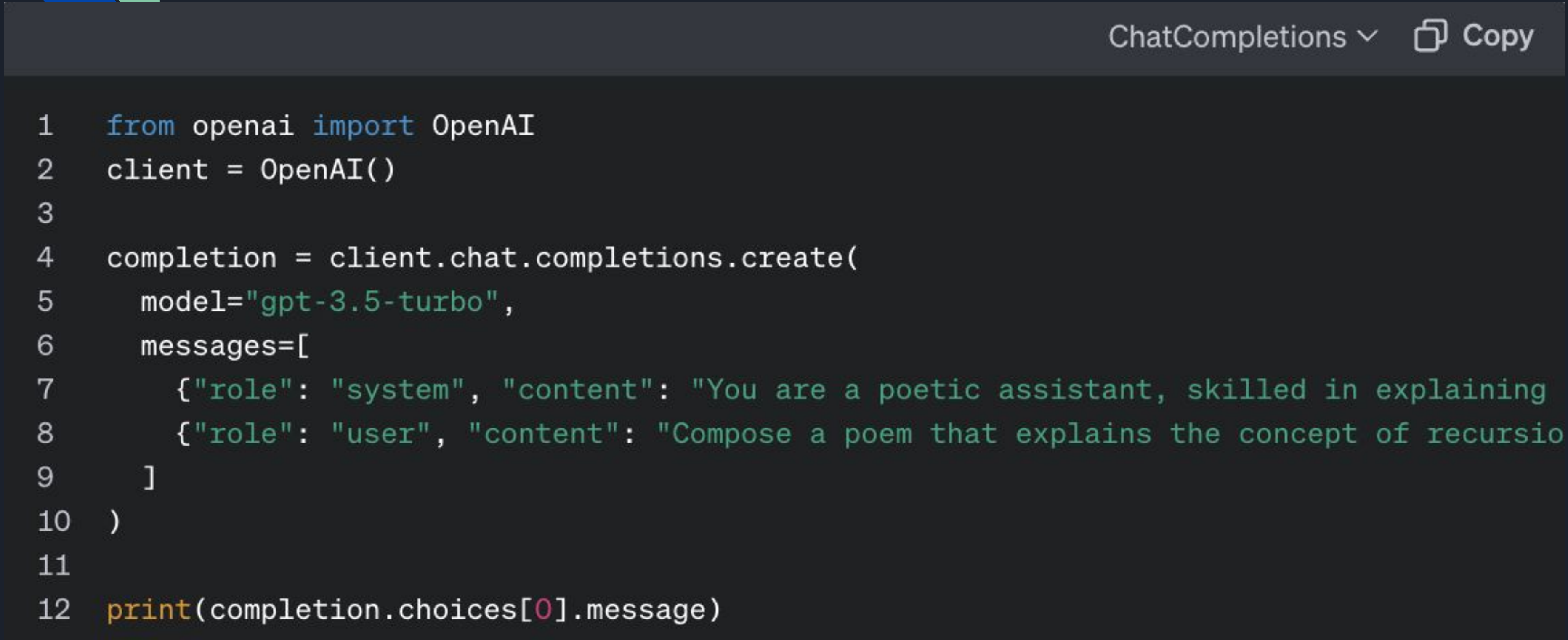
Overview

The OpenAI API is powered by a diverse set of models with different capabilities and price points. You can also make customizations to our models for your specific use case with fine-tuning.

MODEL	DESCRIPTION
GPT-4 and GPT-4 Turbo	A set of models that improve on GPT-3.5 and can understand as well as generate natural language or code
GPT-3.5 Turbo	A set of models that improve on GPT-3.5 and can understand as well as generate natural language or code
DALL·E	A model that can generate and edit images given a natural language prompt
TTS	A set of models that can convert text into natural sounding spoken audio
Whisper	A model that can convert audio into text
Embeddings	A set of models that can convert text into a numerical form
Moderation	A fine-tuned model that can detect whether text may be sensitive or unsafe
GPT base	A set of models without instruction following that can understand as well as generate natural language or code
Deprecated	A full list of models that have been deprecated along with the suggested replacement

We have also published open source models including Point-E, Whisper, Jukebox, and CLIP.

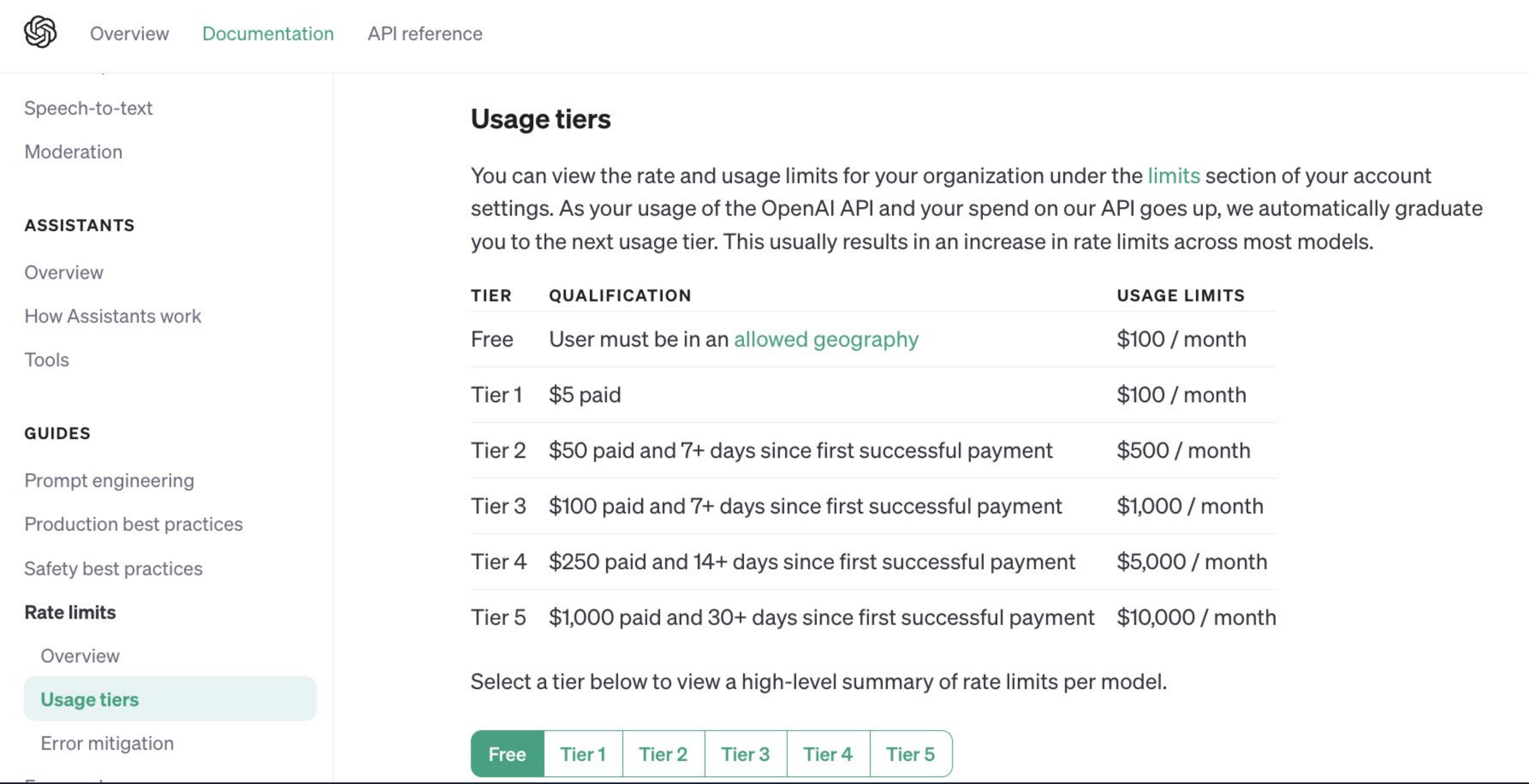
Application development using OpenAI API



```python
from openai import OpenAI
client = OpenAI()

completion = client.chat.completions.create(
  model="gpt-3.5-turbo",
  messages=[
    {"role": "system", "content": "You are a poetic assistant, skilled in explaining
    {"role": "user", "content": "Compose a poem that explains the concept of recursio
  ]
)

print(completion.choices[0].message)
```



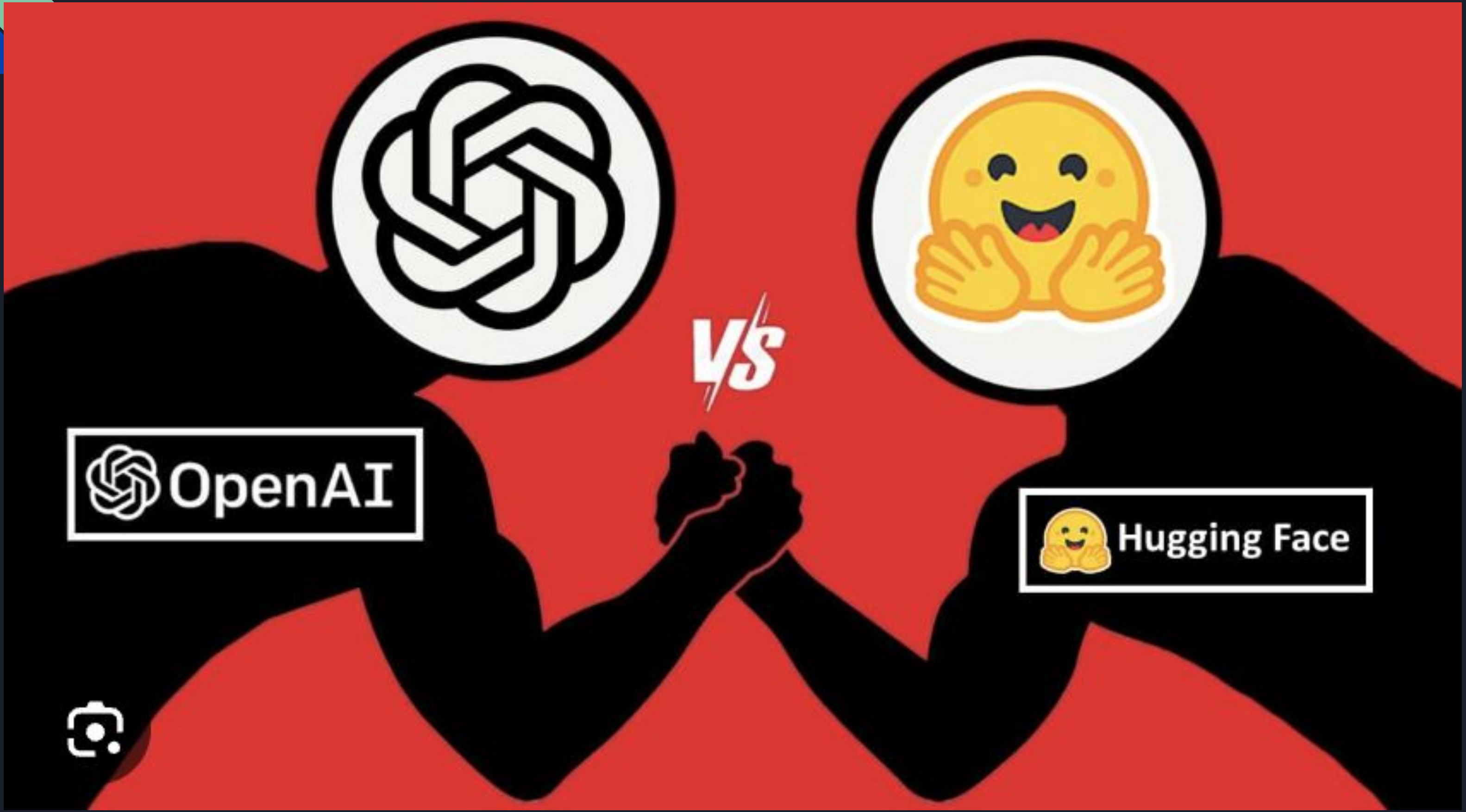
Usage tiers

You can view the rate and usage limits for your organization under the limits section of your account settings. As your usage of the OpenAI API and your spend on our API goes up, we automatically graduate you to the next usage tier. This usually results in an increase in rate limits across most models.

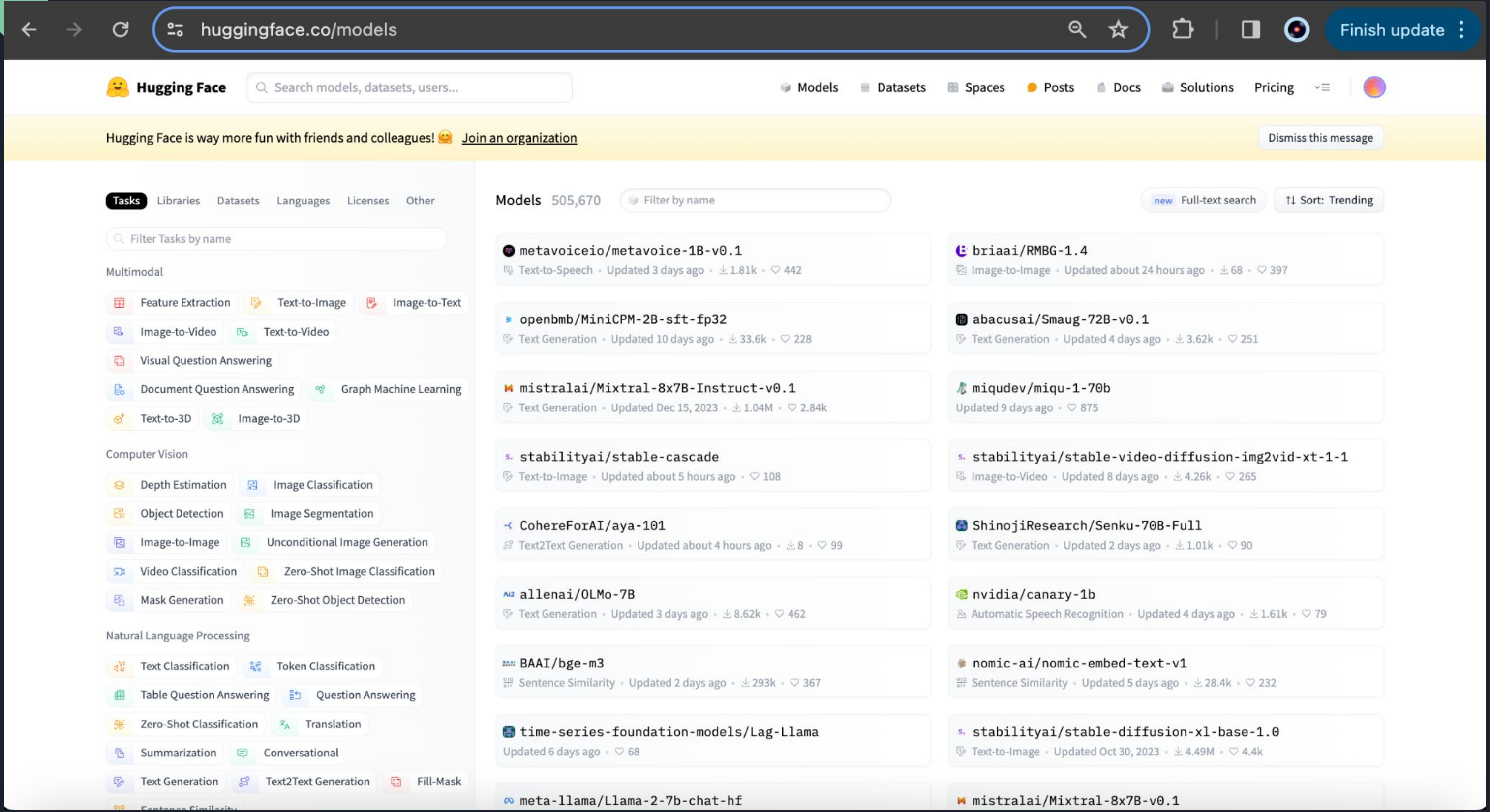
TIER	QUALIFICATION	USAGE LIMITS
Free	User must be in an allowed geography	$100 / month
Tier 1	$5 paid	$100 / month
Tier 2	$50 paid and 7+ days since first successful payment	$500 / month
Tier 3	$100 paid and 7+ days since first successful payment	$1,000 / month
Tier 4	$250 paid and 14+ days since first successful payment	$5,000 / month
Tier 5	$1,000 paid and 30+ days since first successful payment	$10,000 / month

Select a tier below to view a high-level summary of rate limits per model.

Proprietary vs Open Source



Hugging Face - Open Source Model Repository



Hugging Face is way more fun with friends and colleagues! 🤗 Join an organization

Dismiss this message

Multimodal

- Feature Extraction
- Text-to-Image
- Image-to-Text
- Image-to-Video
- Text-to-Video
- Visual Question Answering
- Document Question Answering
- Graph Machine Learning
- Text-to-3D
- Image-to-3D

Computer Vision

- Depth Estimation
- Image Classification
- Object Detection
- Image Segmentation
- Image-to-Image
- Unconditional Image Generation
- Video Classification
- Zero-Shot Image Classification
- Mask Generation
- Zero-Shot Object Detection

Natural Language Processing

- Text Classification
- Token Classification
- Table Question Answering
- Question Answering
- Zero-Shot Classification
- Translation
- Summarization
- Conversational
- Text Generation
- Text2Text Generation
- Fill-Mask
- Sentence Similarity

Models 505,670

metavoiceio/metavoice-1B-v0.1
Text-to-Speech · Updated 3 days ago · 1.81k · ♡ 442

briaai/RMBG-1.4
Image-to-Image · Updated about 24 hours ago · 68 · ♡ 397

openbmb/MiniCPM-2B-sft-fp32
Text Generation · Updated 10 days ago · 33.6k · ♡ 228

abacusai/Smaug-72B-v0.1
Text Generation · Updated 4 days ago · 3.62k · ♡ 251

mistralai/Mixtral-8x7B-Instruct-v0.1
Text Generation · Updated Dec 15, 2023 · 1.04M · ♡ 2.84k

miqudev/miqu-1-70b
Updated 9 days ago · ♡ 875

stabilityai/stable-cascade
Text-to-Image · Updated about 5 hours ago · ♡ 108

stabilityai/stable-video-diffusion-img2vid-xt-1-1
Image-to-Video · Updated 8 days ago · 4.26k · ♡ 265

CohereForAI/aya-101
Text2Text Generation · Updated about 4 hours ago · 8 · ♡ 99

ShinojiResearch/Senku-70B-Full
Text Generation · Updated 2 days ago · 1.01k · ♡ 90

allenai/OLMo-7B
Text Generation · Updated 3 days ago · 8.62k · ♡ 462

nvidia/canary-1b
Automatic Speech Recognition · Updated 4 days ago · 1.61k · ♡ 79

BAAI/bge-m3
Sentence Similarity · Updated 2 days ago · 293k · ♡ 367

nomic-ai/nomic-embed-text-v1
Sentence Similarity · Updated 5 days ago · 28.4k · ♡ 232

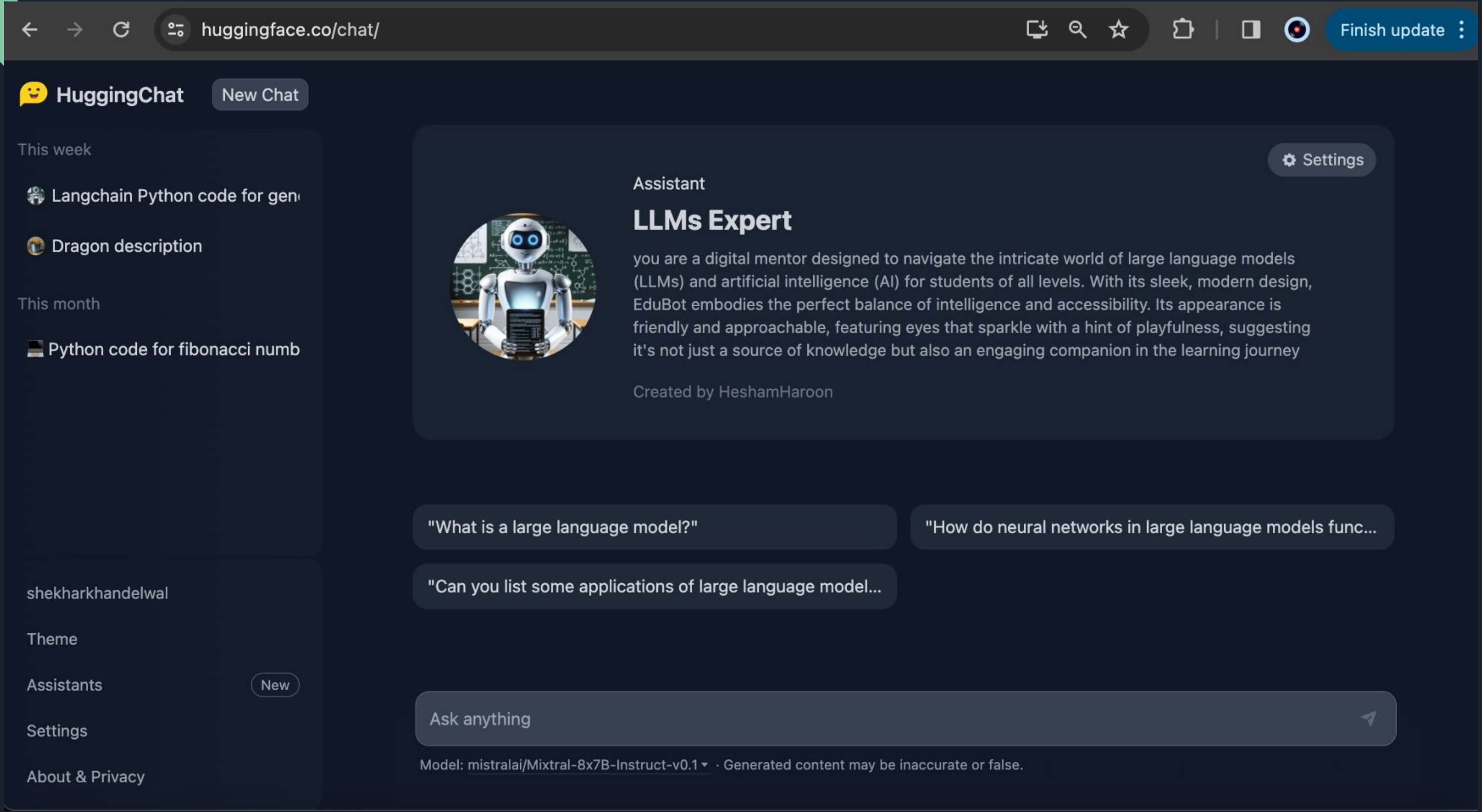
time-series-foundation-models/Lag-Llama
Updated 6 days ago · ♡ 68

stabilityai/stable-diffusion-xl-base-1.0
Text-to-Image · Updated Oct 30, 2023 · 4.49M · ♡ 4.4k

meta-llama/Llama-2-7b-chat-hf

mistralai/Mixtral-8x7B-v0.1

Hugging Face Chat



Assistant

LLMs Expert

you are a digital mentor designed to navigate the intricate world of large language models (LLMs) and artificial intelligence (AI) for students of all levels. With its sleek, modern design, EduBot embodies the perfect balance of intelligence and accessibility. Its appearance is friendly and approachable, featuring eyes that sparkle with a hint of playfulness, suggesting it's not just a source of knowledge but also an engaging companion in the learning journey

Created by HeshamHaroon

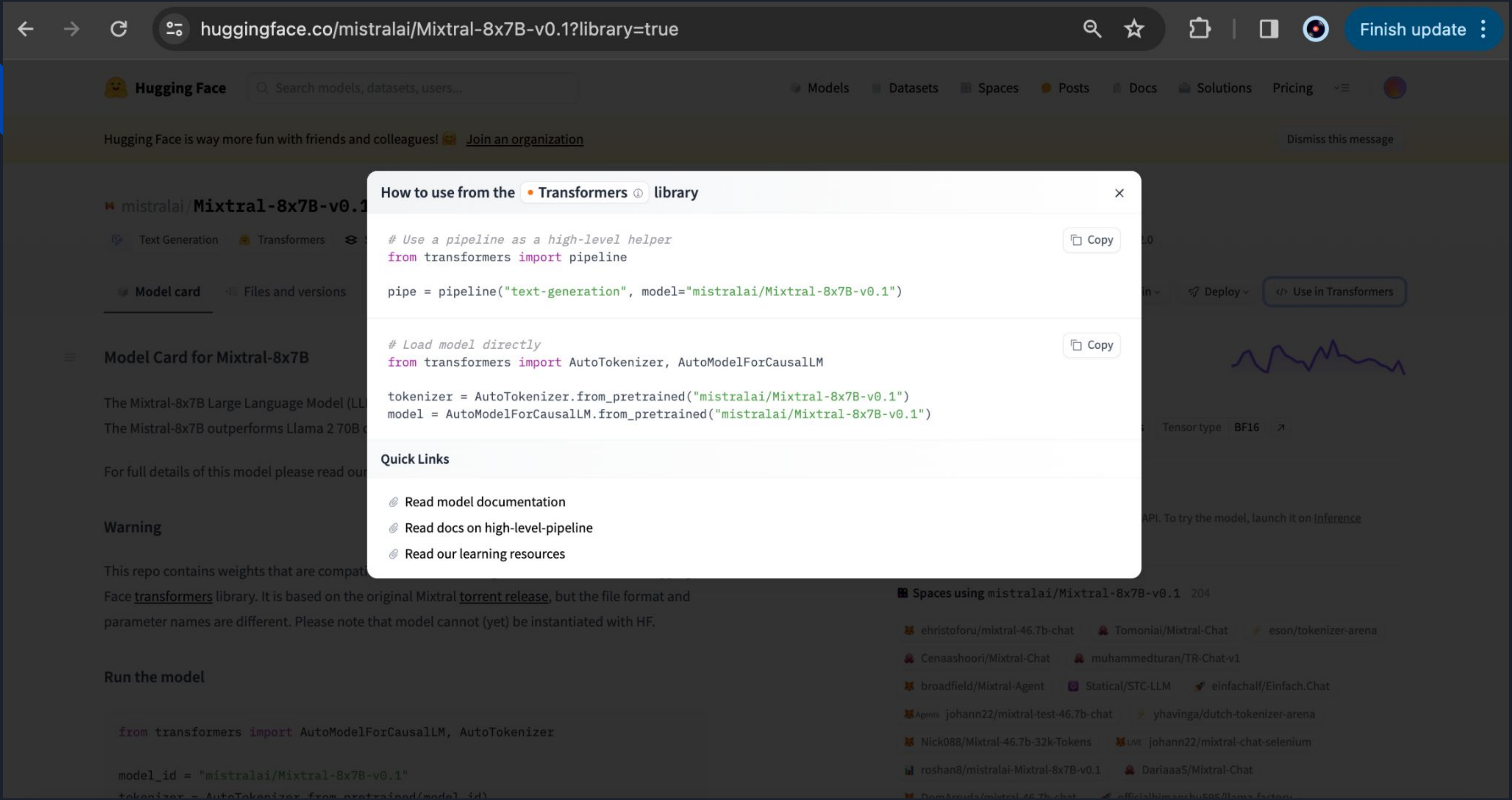
"What is a large language model?"

"How do neural networks in large language models func...

"Can you list some applications of large language model...

Ask anything

Model: mistralai/Mixtral-8x7B-Instruct-v0.1 · Generated content may be inaccurate or false.



Hugging Face is way more fun with friends and colleagues! 🤗 Join an organization

How to use from the Transformers library

```
# Use a pipeline as a high-level helper
from transformers import pipeline

pipe = pipeline("text-generation", model="mistralai/Mixtral-8x7B-v0.1")
```

```
# Load model directly
from transformers import AutoTokenizer, AutoModelForCausalLM

tokenizer = AutoTokenizer.from_pretrained("mistralai/Mixtral-8x7B-v0.1")
model = AutoModelForCausalLM.from_pretrained("mistralai/Mixtral-8x7B-v0.1")
```

Quick Links

- Read model documentation
- Read docs on high-level-pipeline
- Read our learning resources

Model Card for Mixtral-8x7B

The Mixtral-8x7B Large Language Model (LL

The Mistral-8x7B outperforms Llama 2 70B o

For full details of this model please read our

API. To try the model, launch it on Inference

Warning

This repo contains weights that are compati

Face transformers library. It is based on the original Mixtral torrent release, but the file format and parameter names are different. Please note that model cannot (yet) be instantiated with HF.

Run the model

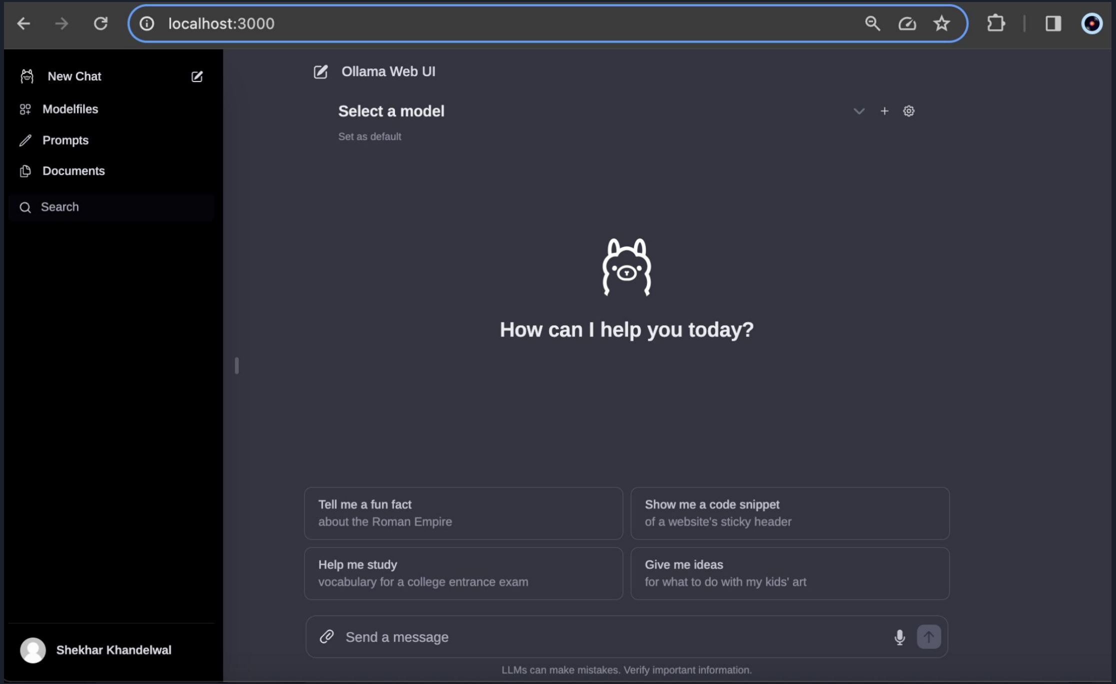
```
from transformers import AutoModelForCausalLM, AutoTokenizer

model_id = "mistralai/Mixtral-8x7B-v0.1"
tokenizer = AutoTokenizer.from_pretrained(model_id)
```

Spaces using mistralai/Mixtral-8x7B-v0.1 204

ehristoforu/mixtral-46.7b-chat
Tomoniai/Mixtral-Chat
eson/tokenizer-arena
Cenaashoori/Mixtral-Chat
muhammedturan/TR-Chat-v1
broadfield/Mixtral-Agent
Statical/STC-LLM
einfachalf/Einfach.Chat
johann22/mixtral-test-46.7b-chat
yhavinga/dutch-tokenizer-arena
Nick088/Mixtral-46.7b-32k-Tokens
johann22/mixtral-chat-selenium
roshan8/mistralai-Mixtral-8x7B-v0.1
DariaaaS/Mixtral-Chat
DomArruda/mixtral-46.7b-chat
officialhimanshu595/llama-factory

The Local Chat GPT - Ollama



localhost:3000

How can I help you today?

Tell me a fun fact
about the Roman Empire

Show me a code snippet
of a website's sticky header

Help me study
vocabulary for a college entrance exam

Give me ideas
for what to do with my kids' art

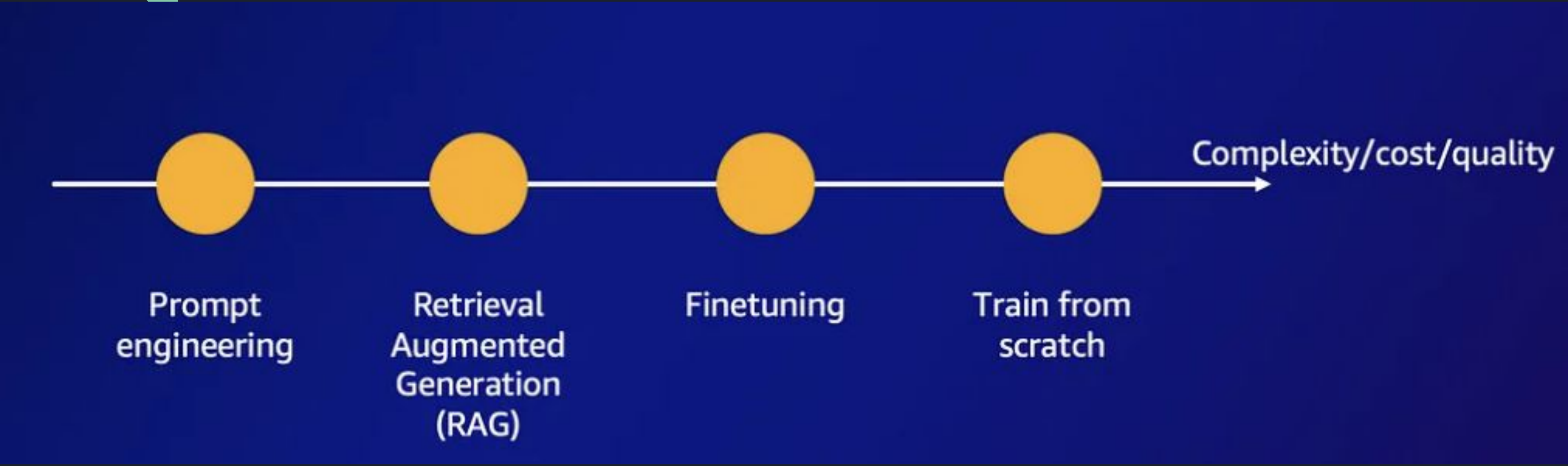
Send a message

LLMs can make mistakes. Verify important information.

https://medium.com/@khandelwal-shekhar/ollama-webui-a-revolutionary-llm-local-deployment-framework-with-chatgpt-like-web-interface-ecea44b80102https://medium.com/@khandelwal-shekhar/ollama-webui-a-revolutionary-llm-local-deployment-framework-with-chatgpt-like-web-interface-ecea44b80102

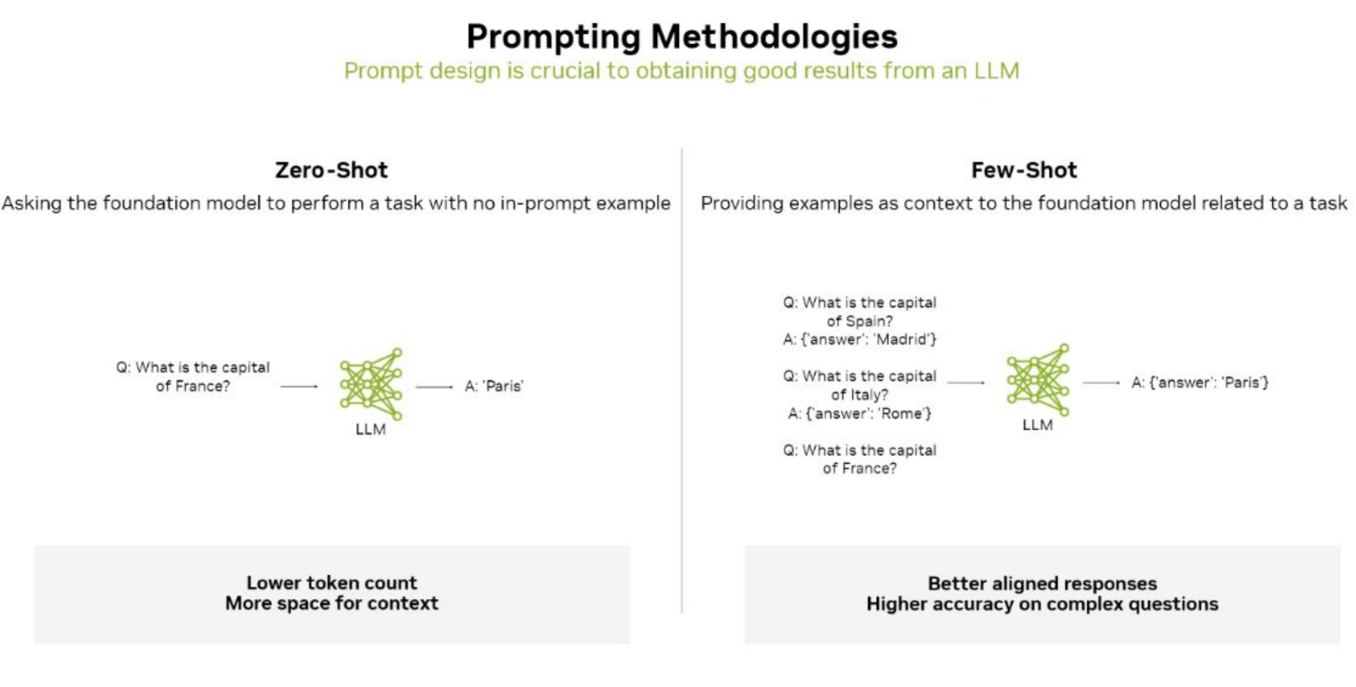
https://medium.com/@khandelwal-shekhar/bring-any-huggingface-model-to-ollama-a457235dd5b8

How to develop industry grade applications with LLMs ?



Disclaimer: All images are borrowed from web for better conceptual understanding.

Prompt Engineering



Prompting Methodologies

Prompt design is crucial to obtaining good results from an LLM

Zero-Shot

Asking the foundation model to perform a task with no in-prompt example

Q: What is the capital of France?

A: 'Paris'

Lower token count
More space for context

Few-Shot

Providing examples as context to the foundation model related to a task

Q: What is the capital of Spain?
A: {'answer': 'Madrid'}

Q: What is the capital of Italy?
A: {'answer': 'Rome'}

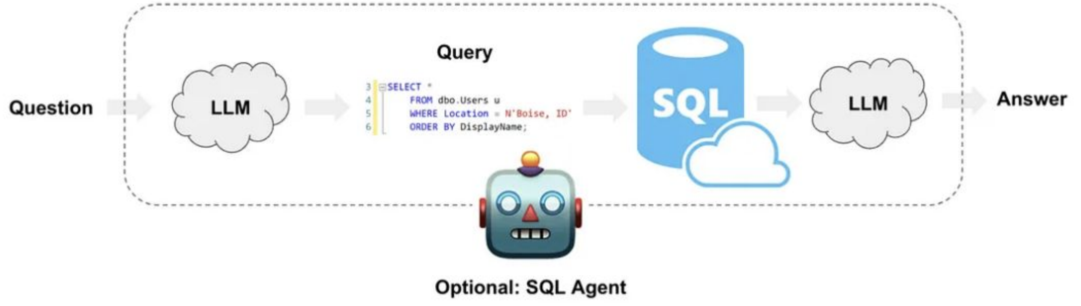
Q: What is the capital of France?

A: {'answer': 'Paris'}

Better aligned responses
Higher accuracy on complex questions

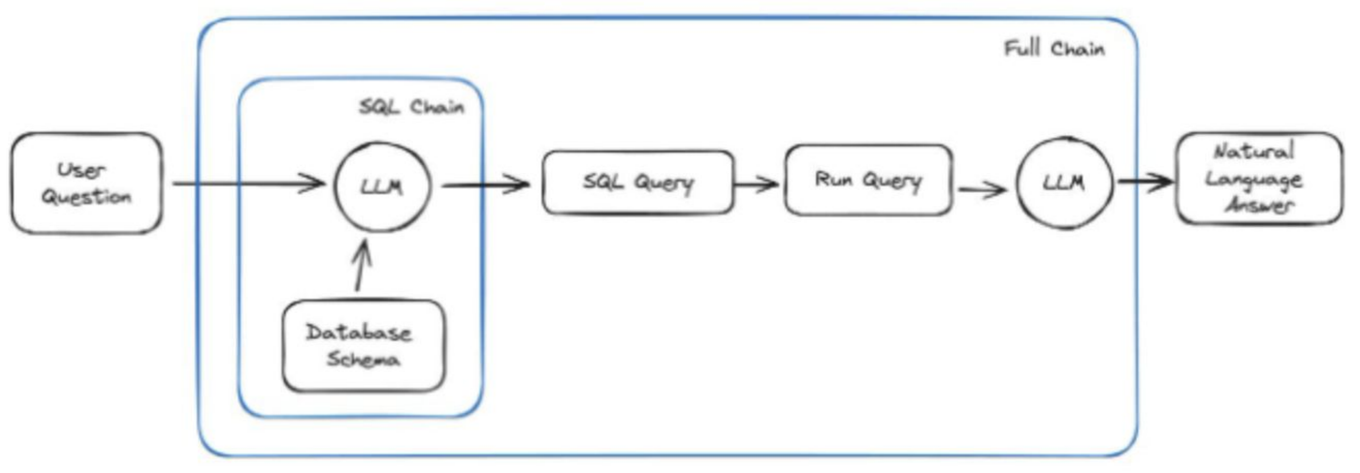
Disclaimer: All images are borrowed from web for better conceptual understanding.

Talk to DB

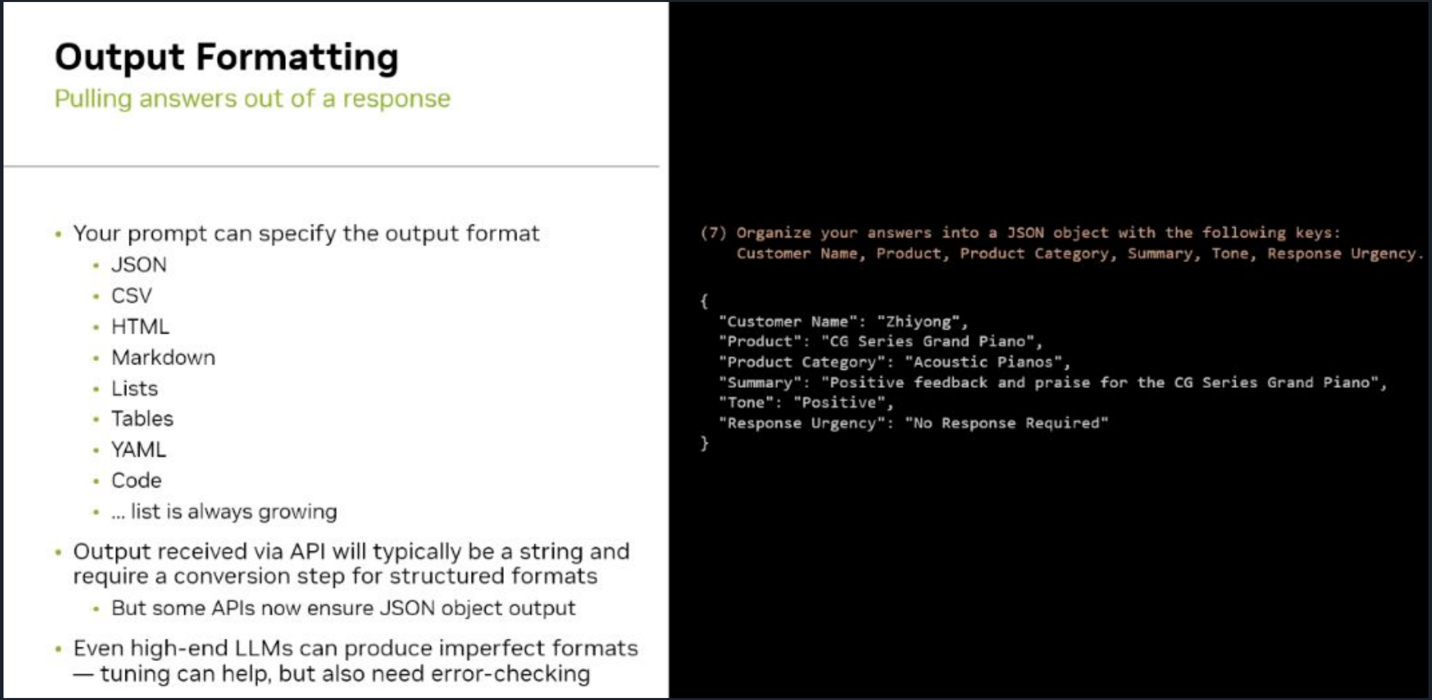


```
SELECT *
    FROM dbo.Users u
    WHERE Location = N'Boise, ID'
    ORDER BY DisplayName;
```

Examples of using LLMs to generate SQL queries from user inputs, and summarize output to provide an answer. Sources: Langchain SQL Agents



Disclaimer: All images are borrowed from web for better conceptual understanding.



Output Formatting

Pulling answers out of a response

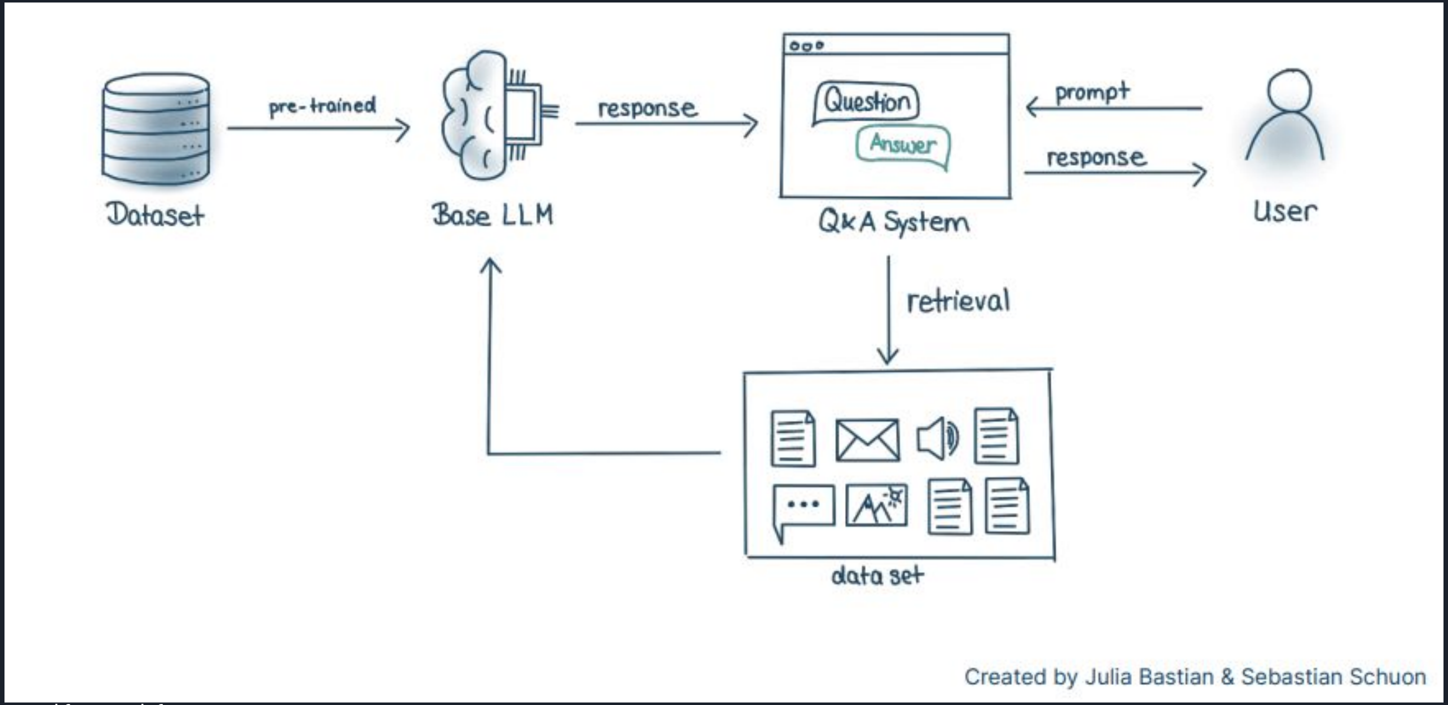
- Your prompt can specify the output format
  - JSON
  - CSV
  - HTML
  - Markdown
  - Lists
  - Tables
  - YAML
  - Code
  - ... list is always growing
- Output received via API will typically be a string and require a conversion step for structured formats
  - But some APIs now ensure JSON object output
- Even high-end LLMs can produce imperfect formats — tuning can help, but also need error-checking

```
(7) Organize your answers into a JSON object with the following keys:
    Customer Name, Product, Product Category, Summary, Tone, Response Urgency.

{
  "Customer Name": "Zhiyong",
  "Product": "CG Series Grand Piano",
  "Product Category": "Acoustic Pianos",
  "Summary": "Positive feedback and praise for the CG Series Grand Piano",
  "Tone": "Positive",
  "Response Urgency": "No Response Required"
}
```

Disclaimer: All images are borrowed from web for better conceptual understanding.

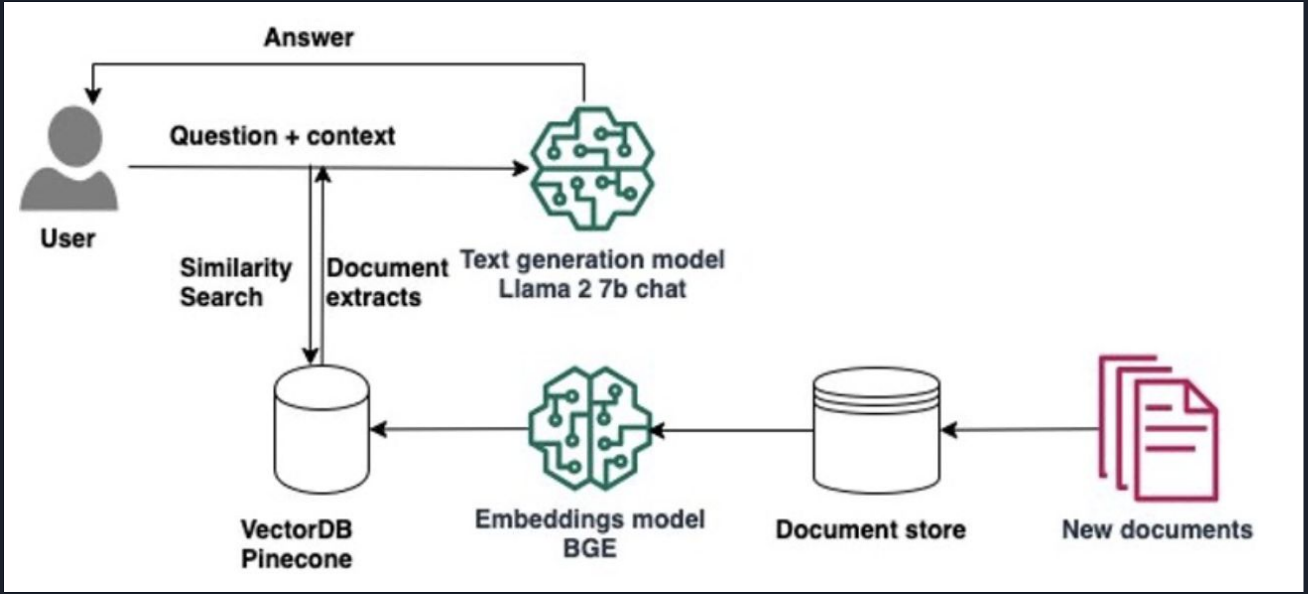
RAG - Retrieval Augmented Generation



Created by Julia Bastian & Sebastian Schuon

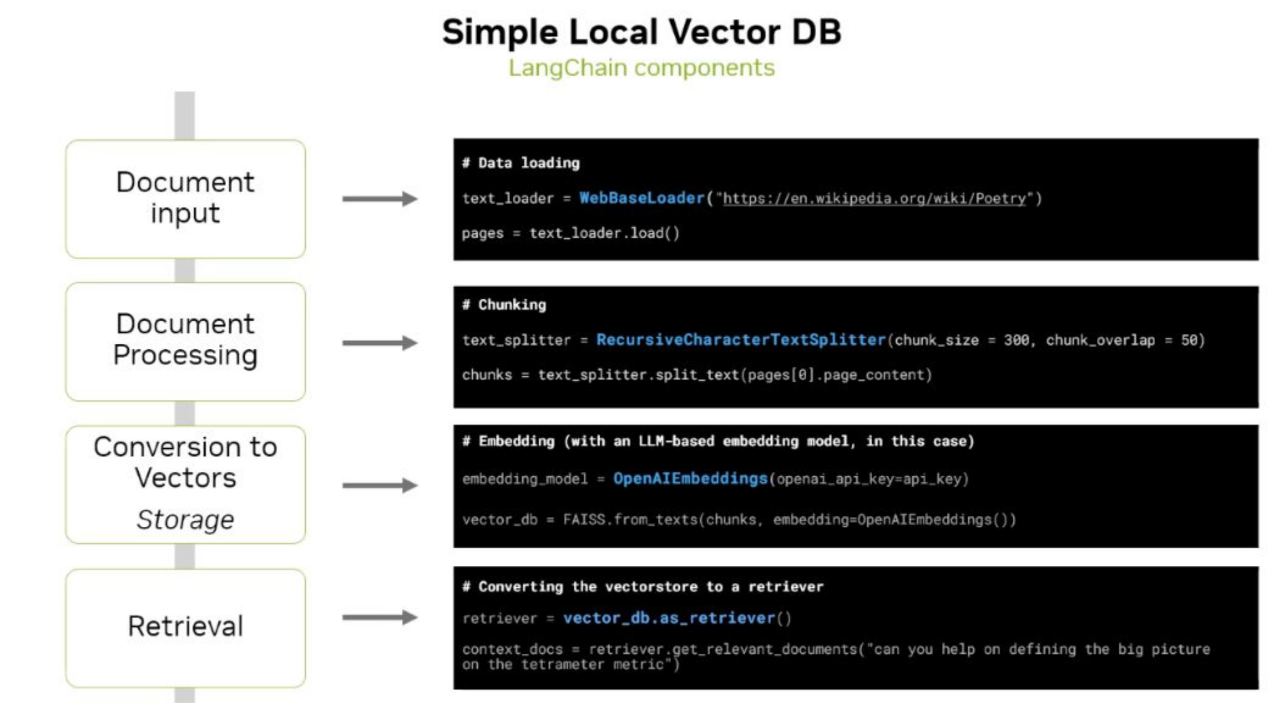
Disclaimer: All images are borrowed from web for better conceptual understanding.

RAG Technical Architecture



Disclaimer: All images are borrowed from web for better conceptual understanding.

RAG code



Simple Local Vector DB

LangChain components

Document input

```
# Data loading
text_loader = WebBaseLoader("https://en.wikipedia.org/wiki/Poetry")
pages = text_loader.load()
```

Document Processing

```
# Chunking
text_splitter = RecursiveCharacterTextSplitter(chunk_size = 300, chunk_overlap = 50)
chunks = text_splitter.split_text(pages[0].page_content)
```

Conversion to Vectors

Storage

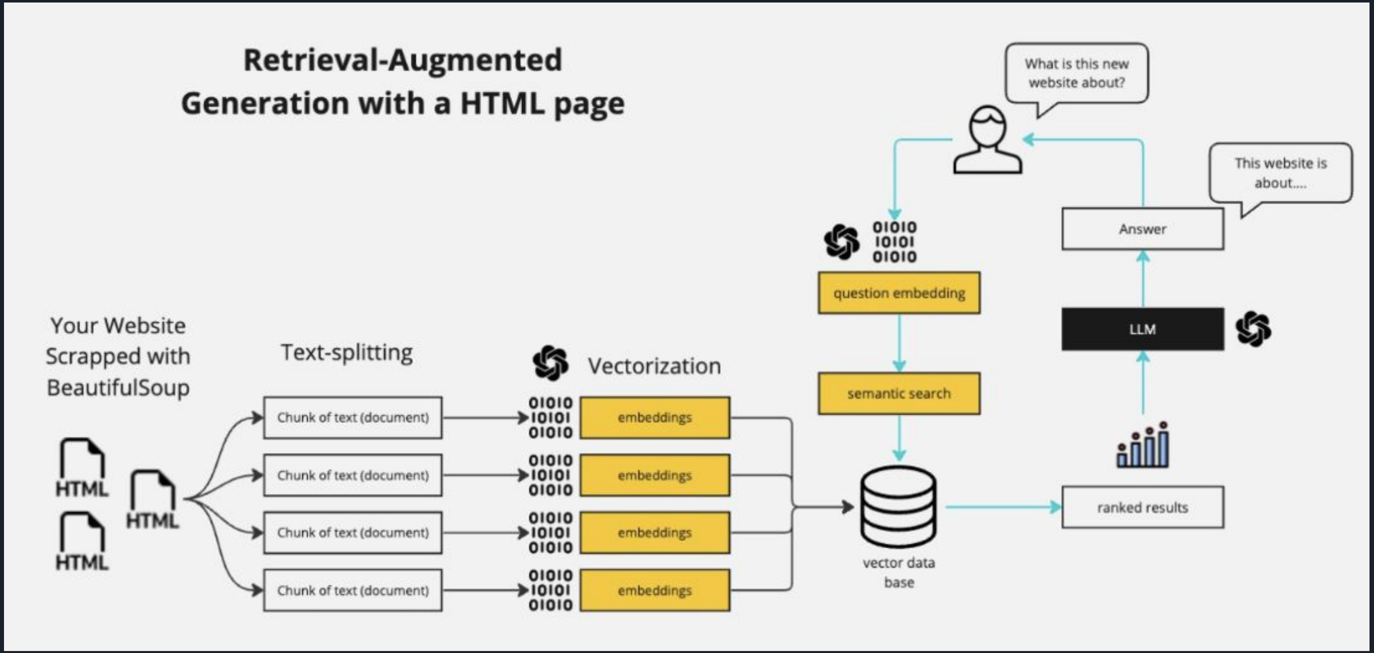
```
# Embedding (with an LLM-based embedding model, in this case)
embedding_model = OpenAIEmbeddings(openai_api_key=api_key)
vector_db = FAISS.from_texts(chunks, embedding=OpenAIEmbeddings())
```

Retrieval

```
# Converting the vectorstore to a retriever
retriever = vector_db.as_retriever()
context_docs = retriever.get_relevant_documents("can you help on defining the big picture
on the tetrameter metric")
```

Disclaimer: All images are borrowed from web for better conceptual understanding.

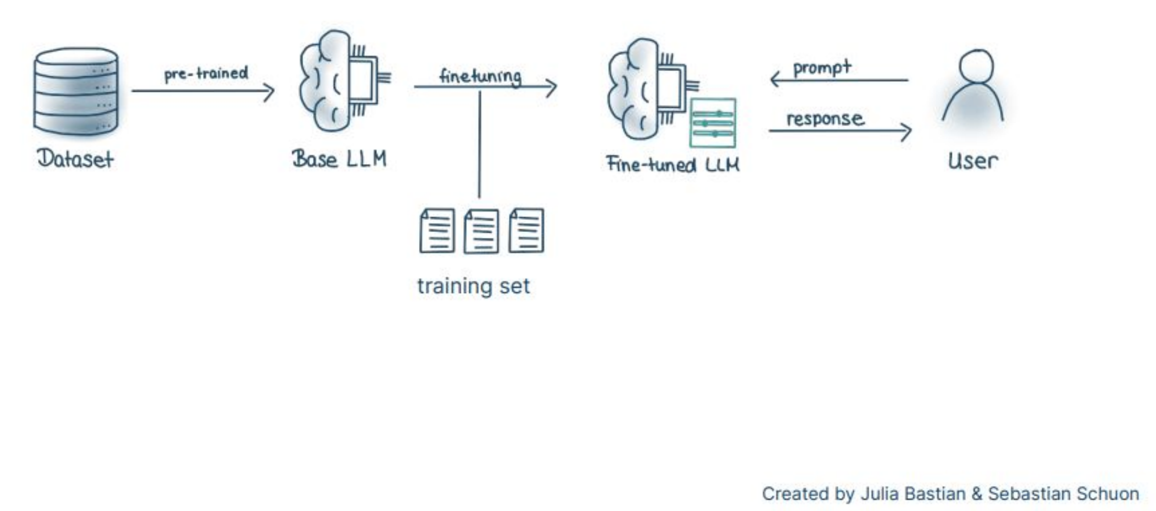
Low Level RAG architecture



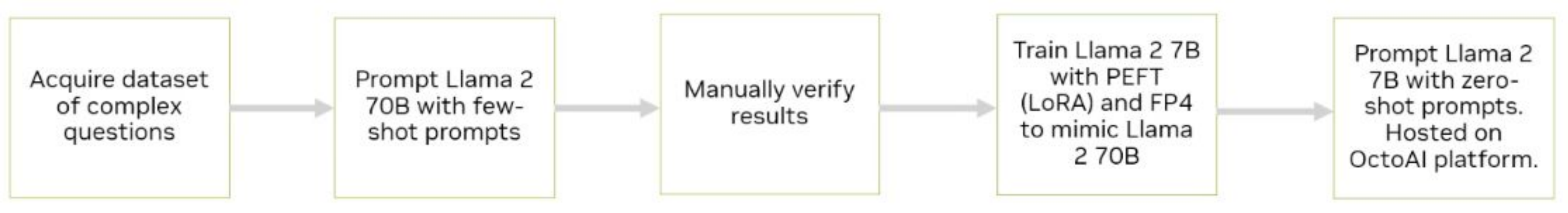
Disclaimer: All images are borrowed from web for better conceptual understanding.

Disclaimer: All images are borrowed from web for better conceptual understanding.

Fine Tuning



Created by Julia Bastian & Sebastian Schuon



Acquire dataset of complex questions

Prompt Llama 2 70B with few-shot prompts

Manually verify results

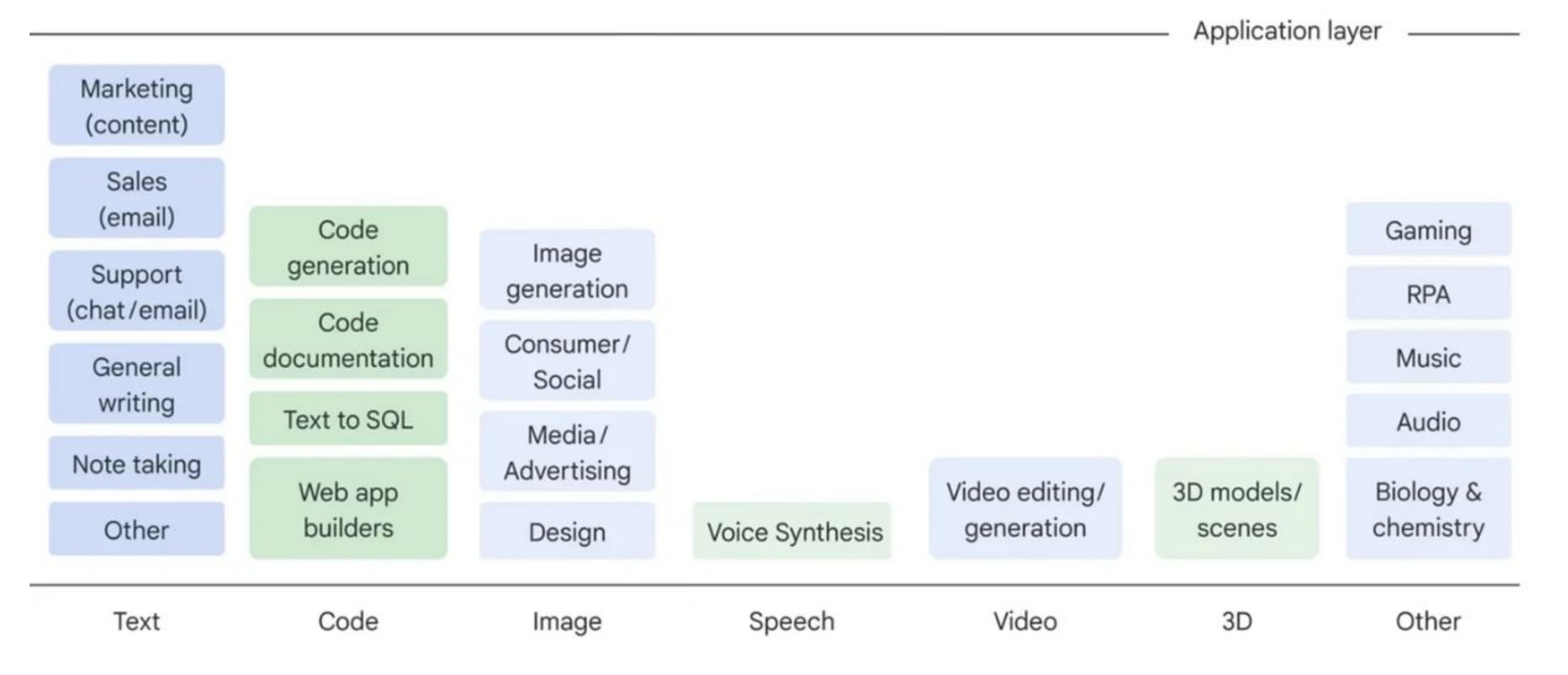
Train Llama 2 7B with PEFT (LoRA) and FP4 to mimic Llama 2 70B

Prompt Llama 2 7B with zero-shot prompts. Hosted on OctoAI platform.

Agentic Applications

LangGraph
Introduction to LangGraph: Enhancing LLM Applications with Cyclic Computations

The Generative AI Application Landscape



Application layer

Text	Code	Image	Speech	Video	3D	Other
Marketing (content) Sales (email) Support (chat/email) General writing Note taking Other	Code generation Code documentation Text to SQL Web app builders	Image generation Consumer/ Social Media/ Advertising Design	Voice Synthesis	Video editing/ generation	3D models/ scenes	Gaming RPA Music Audio Biology & chemistry

Thanks !!