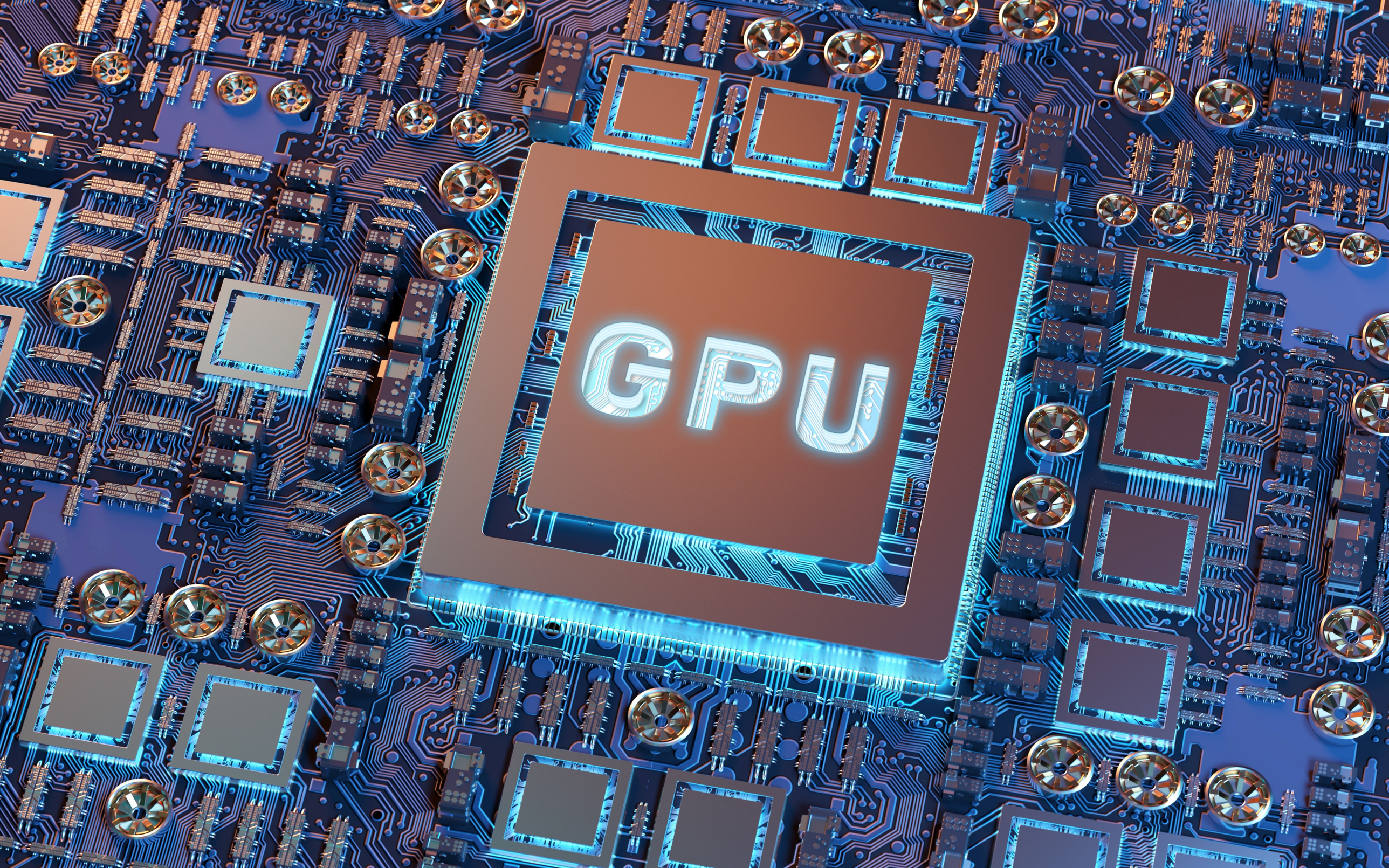
DEEP LEARNING

ARTIFICIAL NEURAL NETWORK

- Scientific Advances - Deep Learning
- Availability of Big Data (You need data to configure these neural networks)
- Lots of compute power


GPU

LARGE LANGUAGE MODELS (LLM)

ATTENTION IS ALL YOU NEED

Bigger is Better

- Very Large Neural Networks
- Vast Amounts of Training Data
- Huge Compute Power to Train Data
- General Purpose AI

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

1 Introduction

Recurrent neural networks, long short-term memory [12] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and transduction problems such as language modeling and machine translation [29, 2, 5]. Numerous efforts have since continued to push the boundaries of recurrent language models and encoder-decoder architectures [31, 21, 13].

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

ATTENTION IS ALL YOU NEED

Transformer Architecture

- Specialized architecture for token prediction
- Key Innovation
  - Attention mechanisms
- Not Just a big neural network

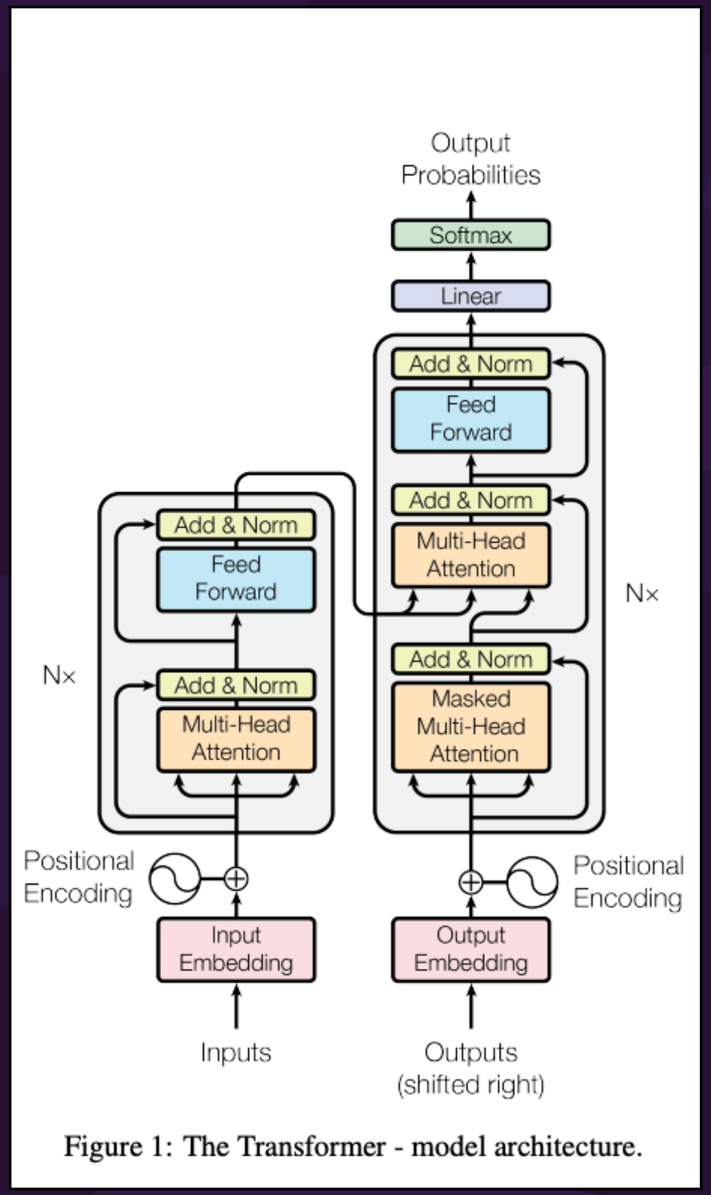


Figure 1: The Transformer - model architecture.

LARGE LANGUAGE MODELS

So what are LLMs

- LLMs are a type of artificial intelligence that can generate text, understand language, answer questions and more.
- They are large because they have a vast number of parameters, which are the parts of the model that are learned from data during training.
- ChatGPT
  - 175 Billion Parameters
  - Training Data - Hundreds of Billions of words

Tasks Libraries Datasets Languages Licenses Other

Filter Tasks by name

Multimodal

- Image-Text-to-Text
- Visual Question Answering
- Document Question Answering

Computer Vision

- Depth Estimation
- Image Classification
- Object Detection
- Image Segmentation
- Text-to-Image
- Image-to-Text
- Image-to-Image
- Image-to-Video
- Unconditional Image Generation
- Video Classification
- Text-to-Video
- Zero-Shot Image Classification
- Mask Generation
- Zero-Shot Object Detection
- Text-to-3D
- Image-to-3D
- Image Feature Extraction

Natural Language Processing

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

Audio

- Text-to-Speech
- Text-to-Audio
- Automatic Speech Recognition
- Audio-to-Audio
- Audio Classification
- Voice Activity Detection

Tabular

- Tabular Classification
- Tabular Regression

Models 570,761

Filter by name

new Full-text search | Sort: Trending

- xai-org/grok-1
  Text Generation • Updated 9 days ago • ♡ 1.74k
- databricks/dbrx-instruct
  Text Generation • Updated about 13 hours ago • ↓ 438 • ♡ 266
- mistralai/Mistral-7B-Instruct-v0.2
  Text Generation • Updated 4 days ago • ↓ 2.02M • ♡ 1.43k
- ByteDance/AnimateDiff-Lightning
  Text-to-Video • Updated 7 days ago • ↓ 57.9k • ♡ 344
- stabilityai/sv3d
  Image-to-Video • Updated 9 days ago • ♡ 395
- databricks/dbrx-base
  Text Generation • Updated about 13 hours ago • ↓ 381 • ♡ 165
- google/gemma-7b
  Text Generation • Updated 29 days ago • ↓ 221k • ♡ 2.63k
- meta-llama/Llama-2-7b-chat-hf
  Text Generation • Updated 9 days ago • ↓ 1.35M • ♡ 3.21k
- Nexusflow/Starling-LM-7B-beta
  Text Generation • Updated about 22 hours ago • ↓ 3.29k • ♡ 129
- alpindale/Mistral-7B-v0.2-hf
  Text Generation • Updated 3 days ago • ↓ 5.99k • ♡ 107
- mistralai/Mixtral-8x7B-Instruct-v0.1
  Text Generation • Updated 28 days ago • ↓ 969k • ♡ 3.48k
- ByteDance/SDXL-Lightning
  Text-to-Image • Updated 14 days ago • ↓ 680k • ♡ 1.46k
- NousResearch/Hermes-2-Pro-Mistral-7B
  Text Generation • Updated 13 days ago • ↓ 37.5k • ♡ 337
- cagliostrolab/animagine-xl-3.1
  Text-to-Image • Updated 10 days ago • ↓ 46.5k • ♡ 242
- stabilityai/stable-diffusion-xl-base-1.0
  Text-to-Image • Updated Oct 30, 2023 • ↓ 3.2M • ♡ 4.82k
- openai/whisper-large-v3
  Automatic Speech Recognition • Updated Feb 8 • ↓ 1.17M • ♡ 2.1k
- stabilityai/stable-code-instruct-3b
  Text Generation • Updated 2 days ago • ↓ 1.2k • ♡ 58
- BAAI/bge-m3
  Sentence Similarity • Updated 2 days ago • ↓ 2.24M • ♡ 607
- hpcai-tech/Open-Sora
  Updated 8 days ago • ♡ 120
- runwayml/stable-diffusion-v1-5
  Text-to-Image • Updated Aug 23, 2023 • ↓ 4M • ♡ 10.5k
- hpcai-tech/grok-1
  Text Generation • Updated 2 days ago • ↓ 1k • ♡ 52
- distil-whisper/distil-large-v3
  Automatic Speech Recognition • Updated about 10 hours ago • ↓ 10.1k • ♡ 51
- CohereForAI/c4ai-command-r-v01
- ostris/ip-composition-adapter

GENERATIVE AI

GENERATIVE PRE-TRAINED TRANSFORMER (GPT)

WHAT IS GENERATIVE AI?

NOT JUST Machine Learning

- Unlike the facial recognition example we saw earlier Generative AI can take the training data and generate something completely new
- NOT Generative Ai
  - Number
  - Classification
  - Probability
- IS Generative AI
  - Natural Language (Text or Speech)
  - Image
  - Audio

JAVA & AI

JAVA & AI

How can we leverage AI?

- Why AI + Java
  - AI is becoming ubiquitous across the IT landscape
  - Java is the language of enterprise, creating Java AI apps is a new requirement
- Spring AI
  - Provide the necessary API access and components for developing AI apps
- Use Cases
  - Q&A over docs
  - Documentation summarization
  - Text, Code, Image, Audio and Video Generation

```
#!/bin/bash
echo "Calling Open AI..."
MY_OPENAI_KEY="YOUR_API_KEY_HERE"
PROMPT="When was the first version of Java released?"

curl https://api.openai.com/v1/chat/completions \
  -H "Content-Type: application/json" \
  -H "Authorization: Bearer $MY_OPENAI_KEY" \
  -d '{"model": "gpt-3.5-turbo", "messages": [{"role":"user", "content": "'"${PROMPT}"'"}] }'
```


vega@Dans-MacBook-Pro-M1-MAX:~/dev/spring-ai/scripts
⌥⌘1
dev/spring-ai/scripts 🚀 ./hello-open-ai.sh
Calling Open AI...
{
"id": "chatcmpl-96qnh95F1pliuO9tTk5rSvpj7pZCR",
"object": "chat.completion",
"created": 1711419961,
"model": "gpt-3.5-turbo-0125",
"choices": [
{
"index": 0,
"message": {
"role": "assistant",
"content": "The first version of Java was released on January 23, 1996."
},
"logprobs": null,
"finish_reason": "stop"
}
],
"usage": {
"prompt_tokens": 16,
"completion_tokens": 16,
"total_tokens": 32
},
"system_fingerprint": "fp_3bc1b5746c"
}
dev/spring-ai/scripts 🚀 _

```java
public static void main(String[] args) throws IOException, InterruptedException {
    var apiKey = "YOUR_API_KEY_HERE";
    var body = """
            {
                "model": "gpt-4",
                "messages": [
                    {
                        "role": "user",
                        "content": "What is Spring Boot?"
                    }
                ]
            }""";

    HttpRequest request = HttpRequest.newBuilder()
            .uri(URI.create("https://api.openai.com/v1/chat/completions"))
            .header("Content-Type", "application/json")
            .header("Authorization", "Bearer " + apiKey)
            .POST(HttpRequest.BodyPublishers.ofString(body))
            .build();

    var client = HttpClient.newHttpClient();
    var response = client.send(request, HttpResponse.BodyHandlers.ofString());
    System.out.println(response.body());
}
```

SPRING AI PROVIDES US SO MUCH MORE THAN A FACILITY FOR MAKING REST API CALLS

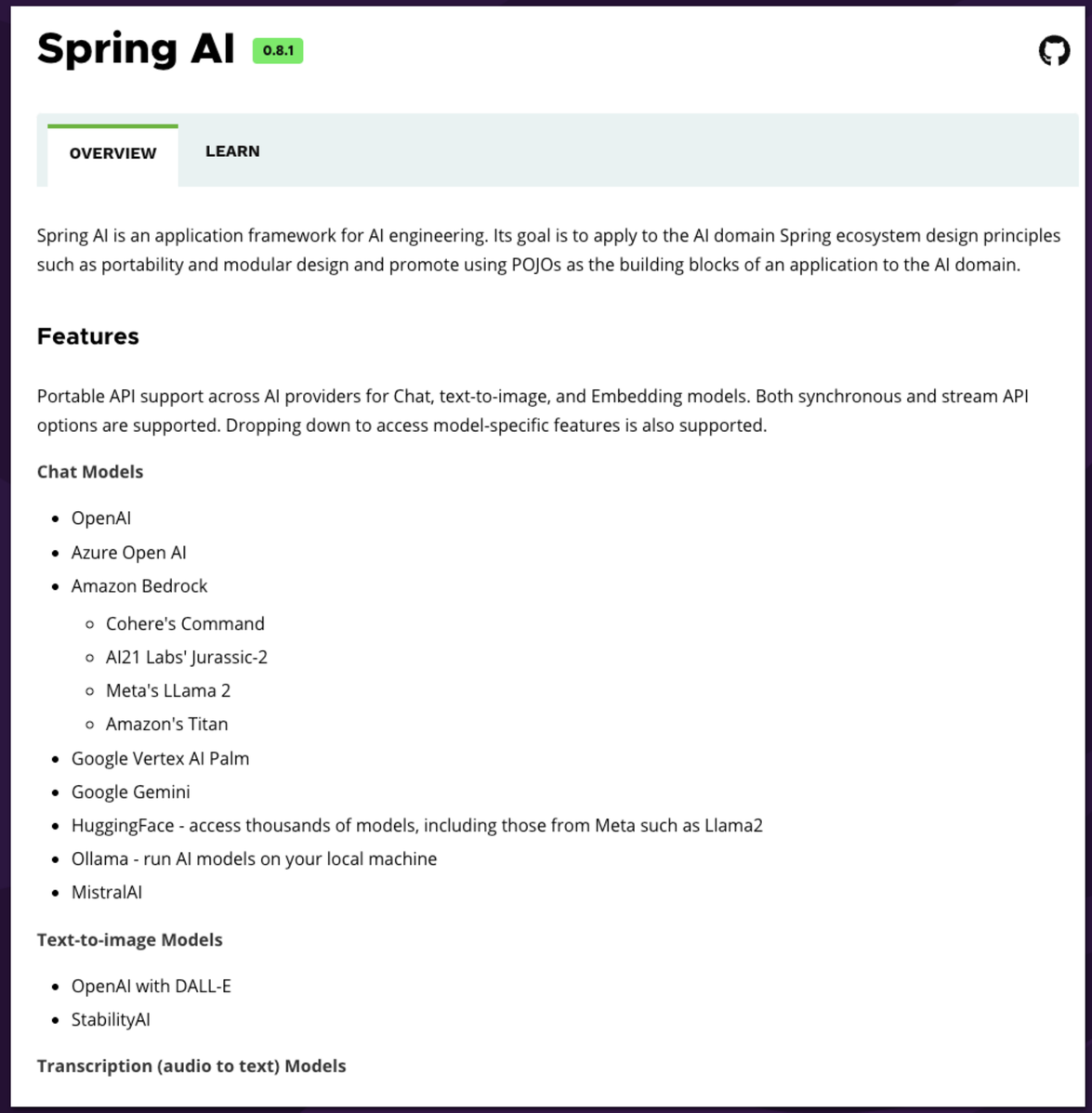
SPRING AI

WHAT IS SPRING AI?

Spring AI

AI for Spring Developers

- A new Spring Project
  - Mark Pollack
  - Current Version 1.0.0-M1
  - https://spring.io/projects/spring-ai
- Inspired by Python projects
  - LangChain
  - LlamaIndex



Spring AI 0.8.1

OVERVIEW LEARN

Spring AI is an application framework for AI engineering. Its goal is to apply to the AI domain Spring ecosystem design principles such as portability and modular design and promote using POJOs as the building blocks of an application to the AI domain.

Features

Portable API support across AI providers for Chat, text-to-image, and Embedding models. Both synchronous and stream API options are supported. Dropping down to access model-specific features is also supported.

Chat Models

- OpenAI
- Azure Open AI
- Amazon Bedrock
  - Cohere's Command
  - AI21 Labs' Jurassic-2
  - Meta's LLama 2
  - Amazon's Titan
- Google Vertex AI Palm
- Google Gemini
- HuggingFace - access thousands of models, including those from Meta such as Llama2
- Ollama - run AI models on your local machine
- MistralAI

Text-to-image Models

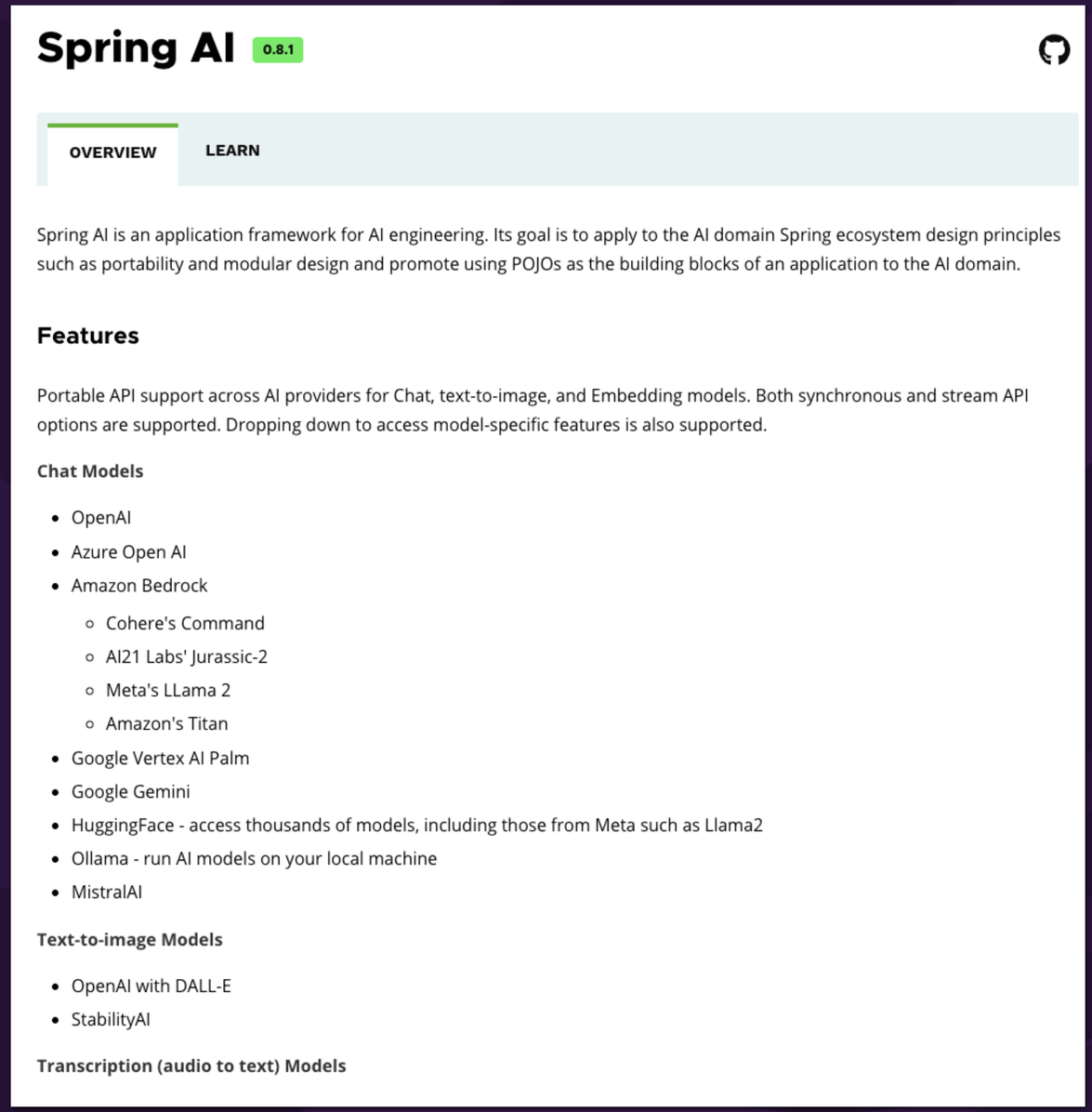
- OpenAI with DALL-E
- StabilityAI

Transcription (audio to text) Models

Spring AI

AI for Spring Developers

- Aligns with Spring project design values
  - Component Abstractions & Default Implementations
  - Portable Chat Completion and EmbeddingClient
  - Multimodality Support
  - Portable Vector Store API & Query Language
  - Function Calling
- Key Components
  - Models
  - Data
  - Chain
  - Evaluation



Spring AI 0.8.1

OVERVIEW LEARN

Spring AI is an application framework for AI engineering. Its goal is to apply to the AI domain Spring ecosystem design principles such as portability and modular design and promote using POJOs as the building blocks of an application to the AI domain.

Features

Portable API support across AI providers for Chat, text-to-image, and Embedding models. Both synchronous and stream API options are supported. Dropping down to access model-specific features is also supported.

Chat Models

- OpenAI
- Azure Open AI
- Amazon Bedrock
  - Cohere's Command
  - AI21 Labs' Jurassic-2
  - Meta's LLama 2
  - Amazon's Titan
- Google Vertex AI Palm
- Google Gemini
- HuggingFace - access thousands of models, including those from Meta such as Llama2
- Ollama - run AI models on your local machine
- MistralAI

Text-to-image Models

- OpenAI with DALL-E
- StabilityAI

Transcription (audio to text) Models