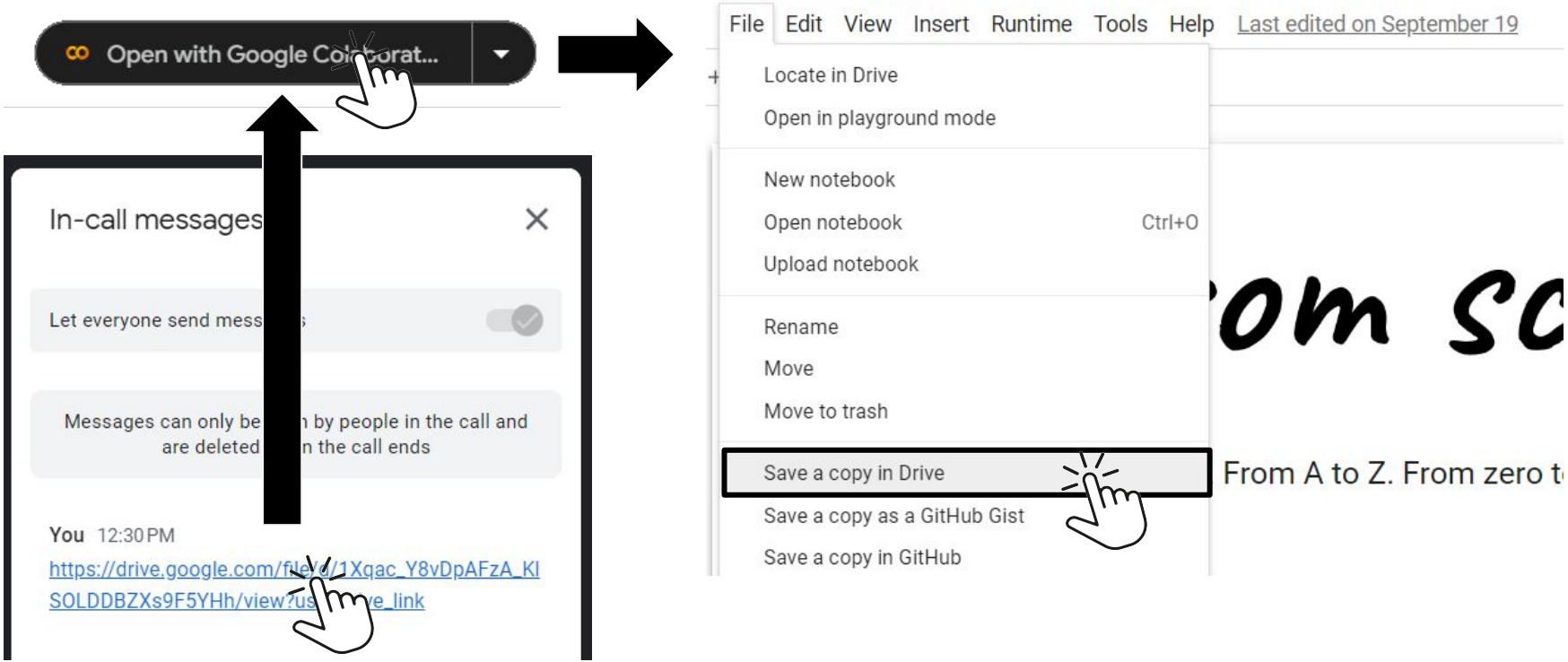
Before we get started:



{"cells":[{"cell_type":"markdown","metadata":{"id":"fwukZZnNTYWE","tags":[]},"source":[" <img src=\"https://drive.google.com/uc?export=view&id=1-1t6Uft8lgBG9jPD0dO6w3dAcv_EUQRP\"> \n","\n","## Learn natural language processing. From A to Z. From zero to hero. Fast.\n","\n","Copyright, NLP from scratch, 2023.\n","\n","[nlpfromscratch.com](https://www.nlpfromscratch.com)\n","\n","----------"]},
{"cell_type":"markdown","metadata":{"id":"gmAd3hBe94HF"},"source":["# Webinar #2 - What the Heck is an LLM?"]},
{"cell_type":"markdown","metadata":{"id":"uuznFjxWVV0n"},"source":["## Introduction \n","In this notebook, we will see several different applications of Large Language Models (LLMs), and show how they can be leveraged the open source libraries from [Hugging Face](https://huggingface.co/).\n","\n","This notebook is best run in [Google Colab](https://colab.research.google.com/), where the majority of dependencies are already installed. However, if you wish to run the notebook locally, please follow the [directions for setting up a local environment](https://drive.google.com/file/d/1EV1seK-dUHRCzj2EDuu3ETAhUyjzOGRd/view?usp=drive_link) and you may then download the notebook as a `.ipynb` and run in either Jupyter or Jupyterlab.\n","\n","Though Google Colab comes with many useful data science libraries included by default (including Pytorch), the Hugging Face libraries are not, so we will first install those here using `pip`, as they will be used in the remainder of the notebook.\n","\n","- The `transformers` library, for general usage of transformer models\n","- The `datasets` library, for working with datasets hosted on Hugging Face\n","- The `diffusers` library, for working with diffusion models for image generation\n","- The `accelerate` library, for using GPU for inference"]},{"cell_type":"code","execution_count":1,"metadata":{"colab":{"base_uri":"https://localhost:8080/"},"executionInfo":{"elapsed":9449,"status":"ok","timestamp":1702487727795,"user":{"displayName":"NLP from scratch","userId":"136364605067828837377"},"user_tz":300},"id":"fjcuFD9_-ajD","outputId":"1480bf7f-0a0e-4d55-bf16-70b955a1aeed"},"outputs":[{"output_type":"stream","name":"stdout","text":["Requirement already satisfied: transformers in /usr/local/lib/python3.10/dist-packages (4.35.2)\n","Collecting datasets\n"," Downloading datasets-2.15.0-py3-none-any.wml (521 kB)\n","\u001b[2K   \u001b[90m━━\u001b[0m \u001b[32m521.2/521.2 kB\u001b[0m \u001b[31m1.9 MB/s\u001b[0m eta \u001b[36m0:00:00\u001b[0m\n","\u001b[?25hCollecting diffusers\n"," Downloading diffusers-0.24.0-py3-none-any.wml (2.8 MB)\n","\u001b[2K   \u001b[90m━━\u001b[0m \u001b[32m2.8/2.8 MB\u001b[0m \u001b[31m4.6 MB/s\u001b[0m eta \u001b[36m0:00:00\u001b[0m\n","Collecting accelerate\n"," Downloading accelerate-0.25.0-py3-none-any.wml (265 kB)\n","\u001b[2K   \u001b[90m━━\u001b[0m \u001b[32m265.7/265.7 kB\u001b[0m \u001b[31m[illegible] MB/s\u001b[0m eta \u001b[36m0:00:00\u001b[0m\n","Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-packages (from transformers) (3.13.1)\n","Requirement already satisfied: huggingface-hub<1.0,>=0.19.3 in /usr/local/lib/python3.10/dist-packages (from transformers) (0.19.4)\n","Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.10/dist-packages (from transformers) (1.23.5)\n","Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-packages (from transformers) (23.2)\n","Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-packages (from transformers) (6.0.1)\n","Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.10/dist-packages (from transformers) (2023.12.25)\n","Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages (from transformers) (2.31.0)\n","Requirement already satisfied: tokenizers<0.19,>=0.14 in /usr/local/lib/python3.10/dist-packages (from transformers) (0.15.0)\n","Requirement already satisfied: safetensors>=0.3.1 in /usr/local/lib/python3.10/dist-packages (from transformers) (0.4.1)\n","Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.10/dist-packages (from transformers) (4.66.1)\n","Requirement already satisfied: pyarrow>=8.0.0 in /usr/local/lib/python3.10/dist-packages (from datasets) (9.0.0)\n","Collecting pyarrow-hotfix (from datasets)\n"," Downloading pyarrow-hotfix-0.6-py3-none-any.wml (7.[illegible] kB)\n","Collecting dill<0.3.8,>=0.3.0 (from datasets)\n"," Downloading dill-0.3.[illegible] (115 kB)\n"]}

NLP from scratch

Learn natural language processing. From A to Z. From zero to hero. Fast.

What the Heck is an LLM?

Monthly Webinar

www.nlpfromscratch.com

Housekeeping

- Camera on if comfortable doing so
- This meeting will not be recorded
- Stay muted unless speaking
- Be professional

Who am I?

- Data Scientist
- Career consultant (SapientNitro, PwC, Accenture)
- Trainer
- Human

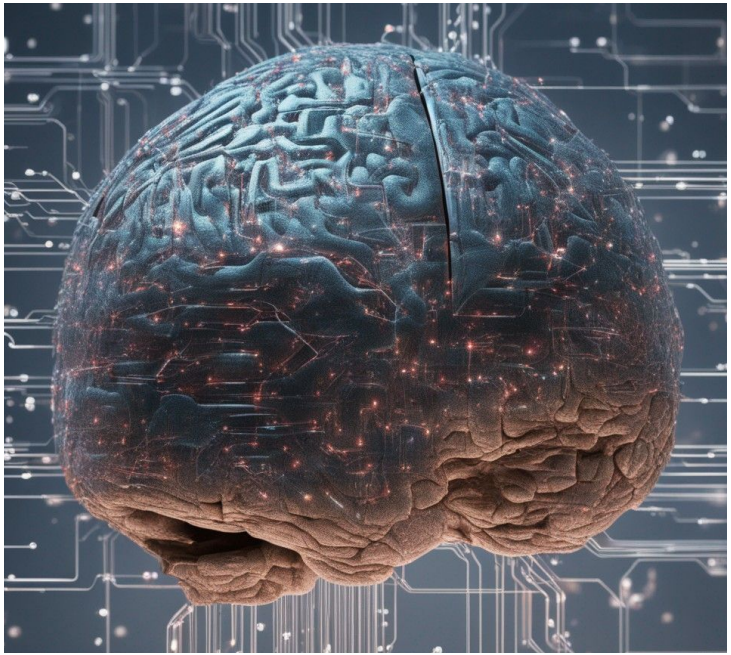

ChatGPT
Capabilities
Remembers what user said earlier in the conversation
Allows user to provide follow-up corrections
Trained to decline inappropriate requests

What is a Large Language Model?

ChatGPT is an example of a large language model (LLM), a type of *deep learning model* trained with hundreds of millions or billions of parameters on very large bodies of text. Large language models currently represent the state of the art in natural language processing (NLP) applications.

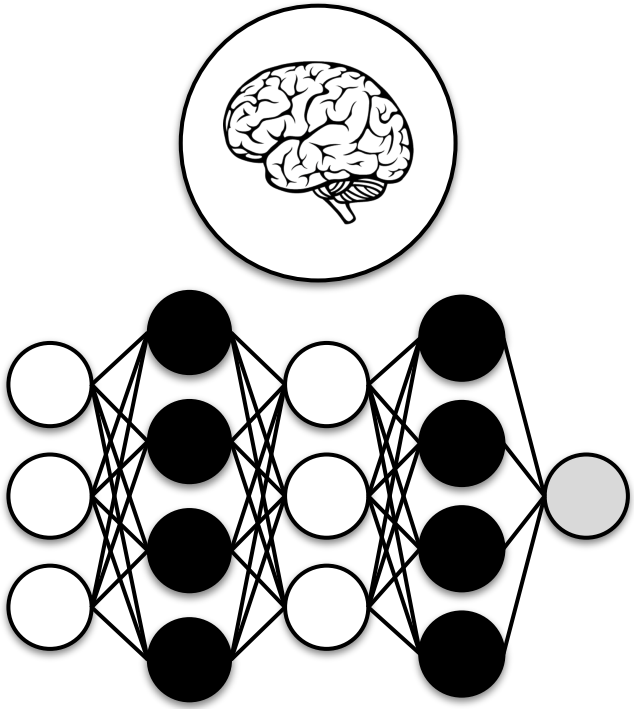
While we're here, ChatGPT is not sentient, nor is it an example of an Artificial General Intelligence (AGI).

Let's take a step back…



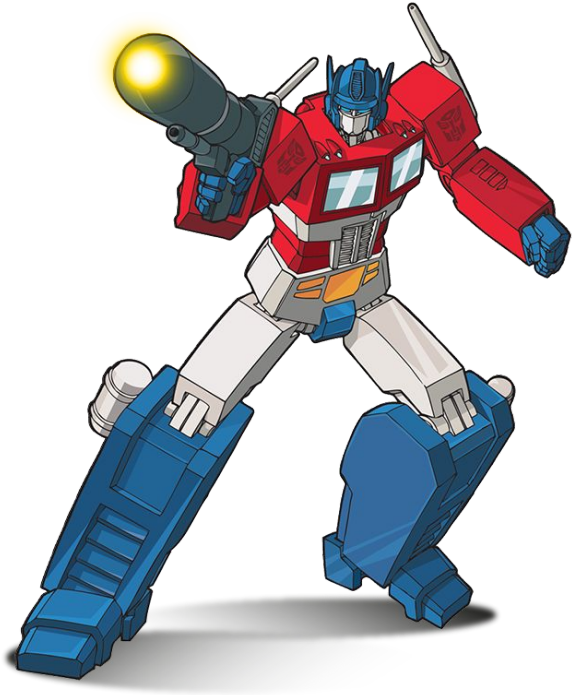
What is Deep Learning?

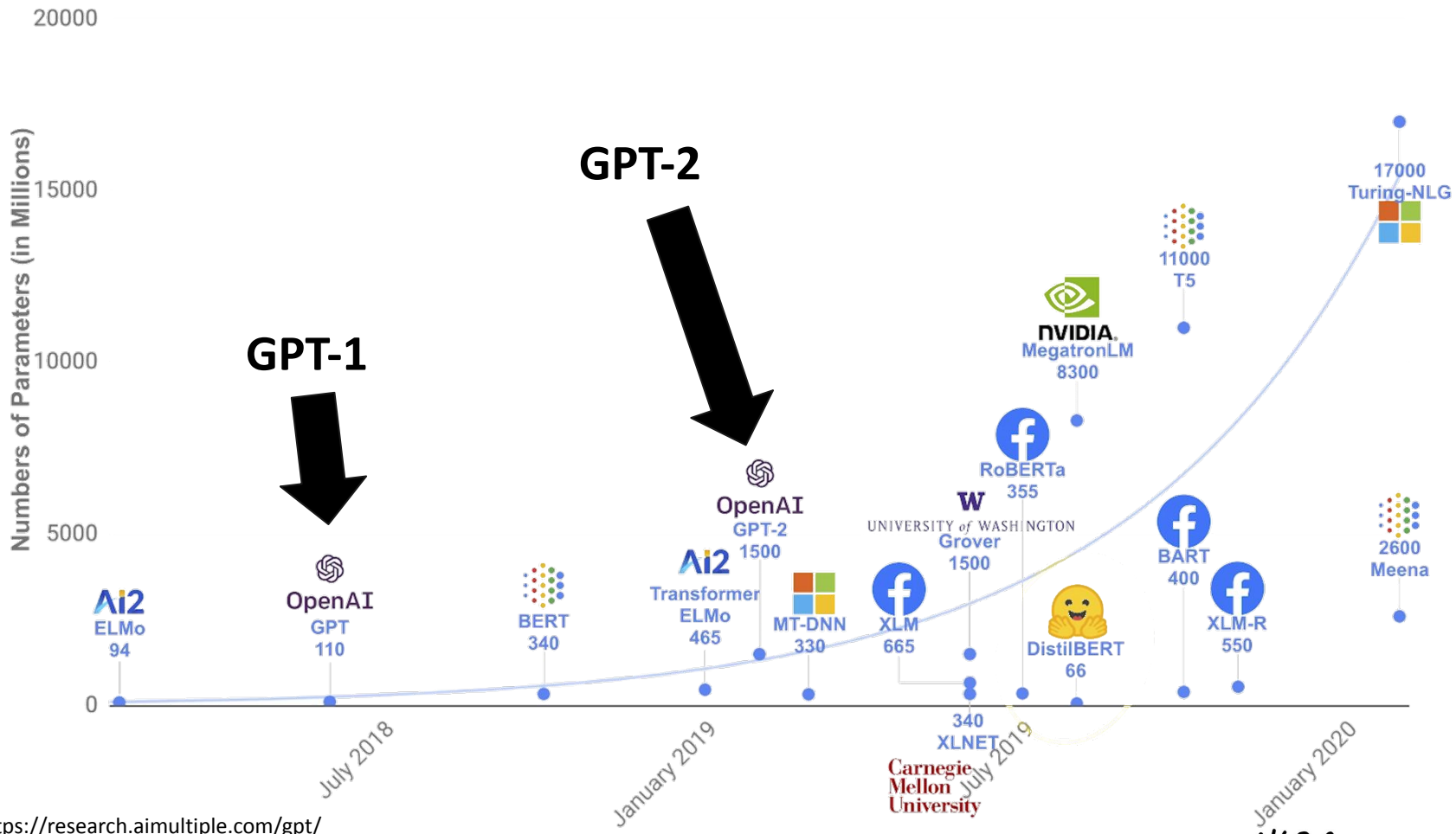
- Deep learning is the branch of machine learning dealing with neural networks - models which mimic the structure of the human brain by having individual “nodes” communicate with each other and pass information along
- Neural networks "learn" by fitting the model parameters or *weights*, by optimizing them against training data and a target objective. This could be text data where the task is to predict the most likely next letter, or classifying types of images.
- LLMs are deep learning models which are massive in size; it is understood that most have billions, hundreds of billions, or even trillions of parameters. It is also understood that these types of models are trained on very large datasets.



The Transformer Architecture

- Groundbreaking paper "Attention is All You Need" from Google researchers (Vaswani et al, 2017) introduced Transformer architecture
- Original application in machine translation but now general purpose and applied to a myriad of other tasks
- Represents the state of the art for LLMs and also applied in domains outside of language (image generation) - virtually all new models based on this architecture
- Popularized by OpenAI and the Generative Pretrained Transformer (GPT) series of models





https://research.aimultiple.com/gpt/

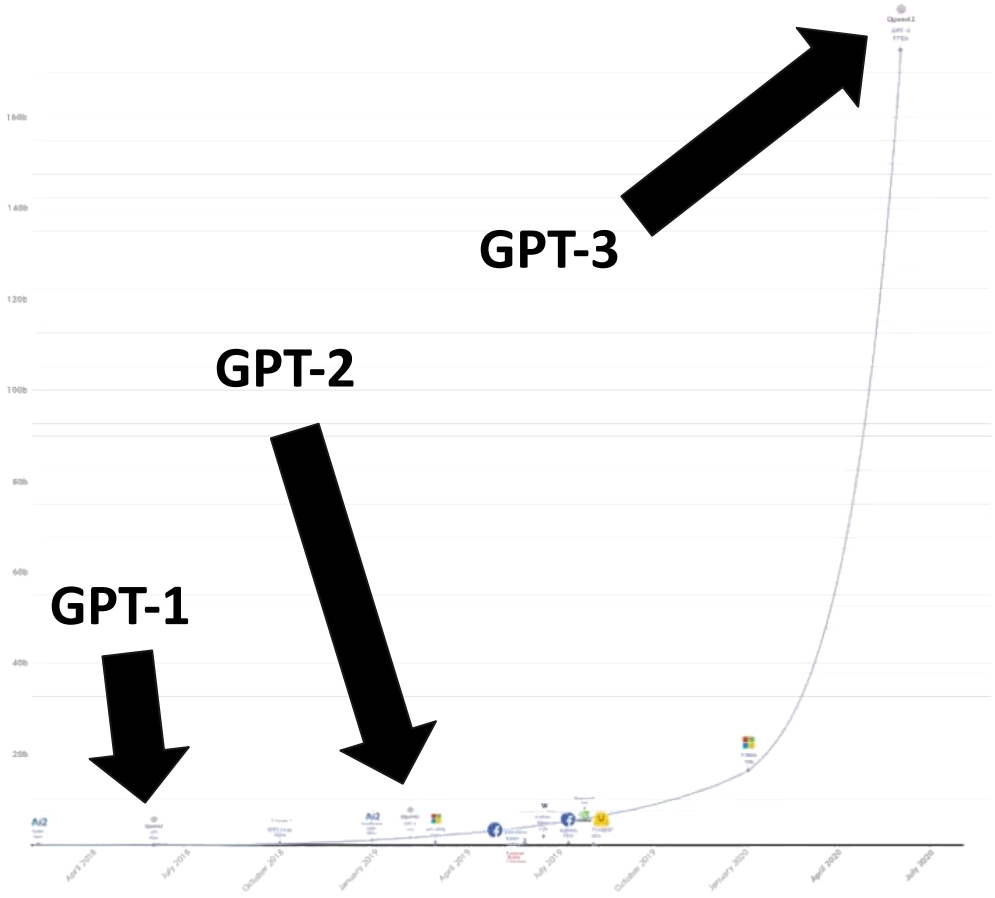
NLP from scratch

To the Moon?

GPT-2 (2019):
1.5B parameters

GPT-3 (2020):
175B parameters

GPT-4 (March 14, 2023):
~1.8T (?) parameters



https://research.aimultiple.com/gpt/

Tools of the Trade

Python 3

Google Colab
/ Jupyter

Hugging Face

Deep Learning and LLM

Deep Learning Frameworks

TensorFlow

- Google product
- Graph-based computation, GPU training
- Other deployment options (Tensorflow Lite, TF.js)
- Easy with integration of Keras into TF 2.x

PyTorch

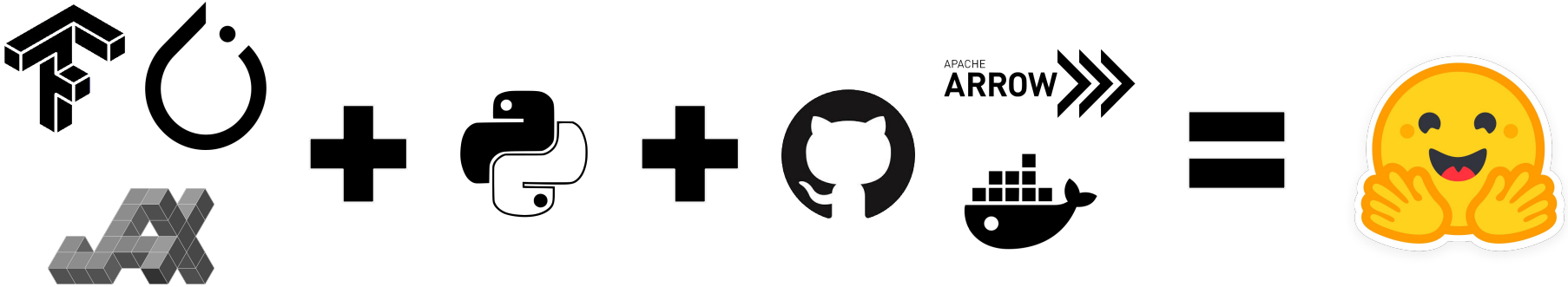
- Meta product
- Graph-based computation, GPU training
- Pytorch Mobile for embedded, no web (ONNX?)
- OOP dev focus (ML eng), Lightning equivalent to Keras

Hugging Face

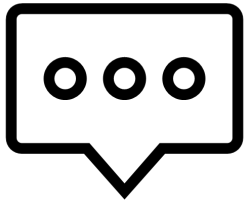
Hugging Face is a software company founded in 2013 and based in New York city. As of August 2023, the company is in Series 'D' funding with a valuation of $4.5B and backing from companies such as Salesforce, Google, Amazon, IBM, Nvidia, AMD, and Intel.

While this name refers to the company, it also refers to the software and platform they develop for working with large language models and data in the natural language processing and other domains.

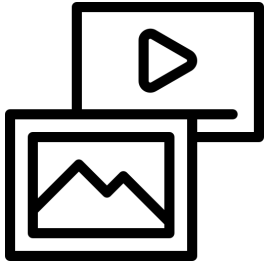
The `datasets` library allows working with data hosted on the platform, and the `transformers` library for working with models of this type. There are also other libraries for working with specialized types of models (*e.g.* `diffusers` for diffusion models) and data processing and model optimization.


APACHE
ARROW

LLM Use Cases



Generative Text



Synthetic Media Generation

Automatic Speech Recognition

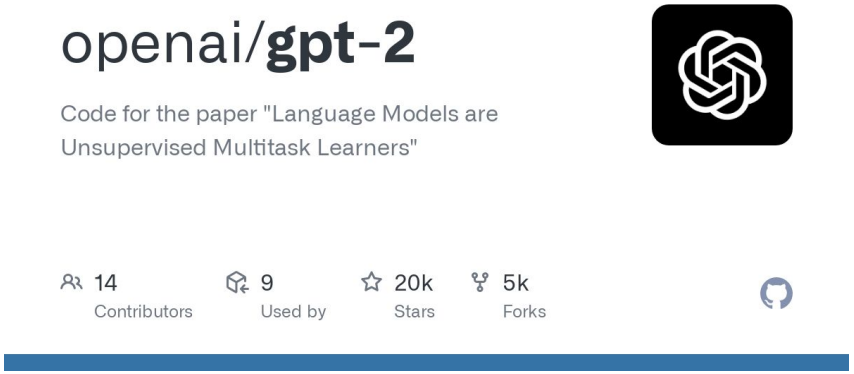
Generative Text: GPT

Undoubtedly, the most popularly known generative text model is that of the Generative Pretrained Transformer (GPT) by OpenAI.

As we've seen, there have been a series of GPT models of increasing size and and trained on increasingly large and more complex datasets.

While GPT-3 remains proprietary and only available to use through the OpenAI API, the weights of GPT-2 are publicly available and can also be accessed through Hugging Face.

Let's take a look at generating text with GPT-2.

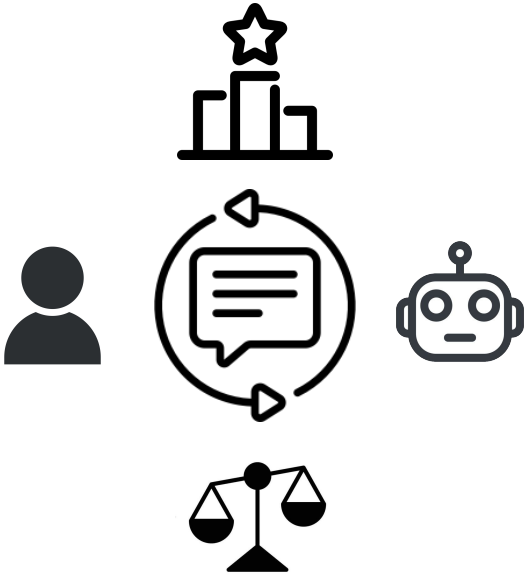


What makes ChatGPT so convincing?

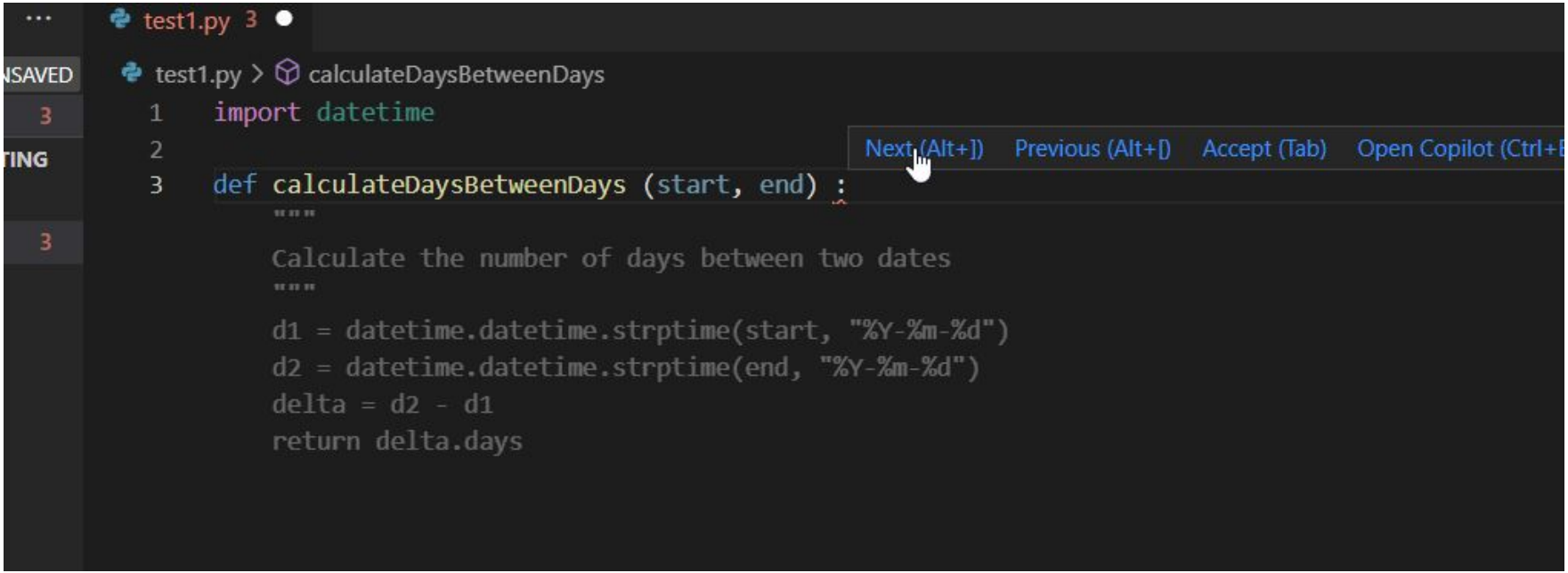
In addition to being “pre-trained” on a very large corpus of publicly available text data and having a very large model size, ChatGPT and other models are also fine-tuned and further refined using *reinforcement learning from human feedback*.

This method of keeping “humans in the loop” by providing feedback to the model of preferred responses, allows the models to display learn the abilities they display for things like detailed question answering and summarization.

The complexity added is that an additional reward model must be trained and incorporated into the overall model development process.



Use Case: Code Completion (Github Copilot)



Adoption

"When we first launched GitHub Copilot for Individuals in June 2022, ***more than 27%*** *of developers' code files on average were generated by GitHub Copilot. Today, GitHub Copilot is behind an* ***average of 46%*** *of a developers' code across all programming languages..."*

- *Github Product Blog, February 2023*

https://github.blog/2023-02-14-github-copilot-now-has-a-better-ai-model-and-new-capabilities/

Generating Images: Stable Diffusion

One of the most popularly known models for media generation is that of Stable Diffusion, created as part of research by the CompVis group at the University of Munich and funded by Stability AI.

Stable Diffusion is a type of latent diffusion model, a neural network that maps images and text to a latent space and does image generation through a denoising (diffusion) process similar to that of generative adversarial networks.

This family of models gained considerable notoriety given Stability's decision to publicly release the code and weights under a license, making the model freely available to use and also making generative image capabilities widely available.

Stable Diffusion XL (SDXL)

- Released July 2023 by researchers at Stability AI, the successor to Stable Diffusion 2.1
- 3x in size to (core of) original model
- Additional refiner model (image-to-image) for denoising used in a supplementary fashion after base model for high fidelity outputs
- Available through Clipdrop (paid) and on Hugging Face spaces (free, various)
- Now near real-time image generation "as you type" with SDXL Turbo

Use Case: AI Photo Editing (Adobe Firefly)

Use Case: Model Texture Generation (Meshy)

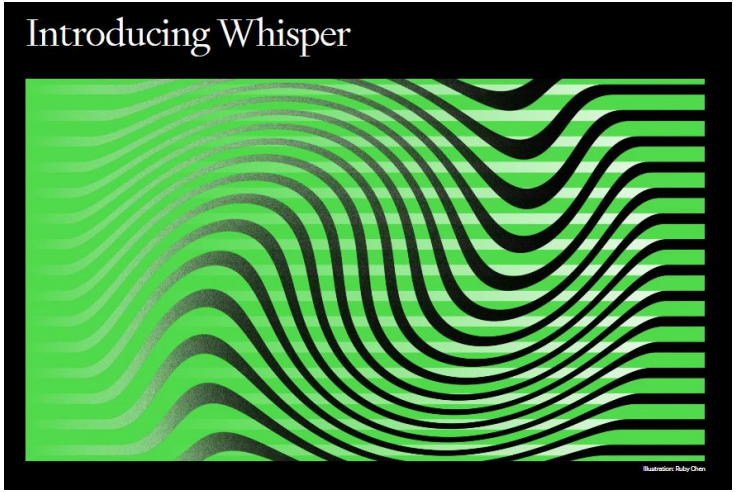
Meshy

Speech Transcription: Whisper

OpenAI's Whisper is a series of multilingual models for high performance transcription of human speech (*i.e.* automatic speech recognition or ASR).

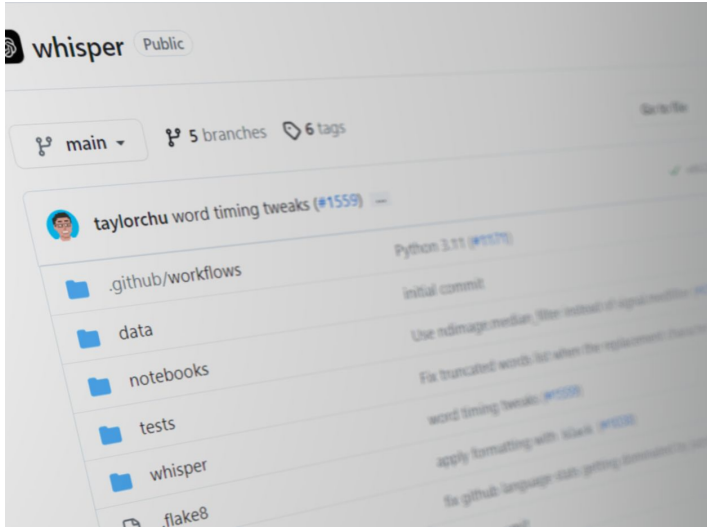
The models are a type of sequence-to-sequence transformer trained on multiple tasks (multilingual speech recognition, translation, spoken language identification, voice activity detection) but fundamentally work by first converting the audio to a spectrogram and then generating sequences of output tokens based on this input.

The company has open-sourced these models and made the weights publicly available. There is also a series of different model sizes to use, ranging from `whisper-tiny` (39M parameters, ~151 MB) to `whisper-large` (1.5B parameters, ~6.2 GB).



Building with Whisper

- The Whisper model is also available to be used directly through the python package
- Model weights and code are also available directly from the github page
- Available directly from OpenAI as a service ($0.006 / minute (rounded to the nearest second)
- Whisper V3 released as part of OpenAI Dev Day (November 6th, 2023) with significant improvements across multiple languages over large V2 model
- Distil-Whisper from Hugging Face is a distilled version 6x faster, 49% smaller, and performs within 1% word error rate (WER)

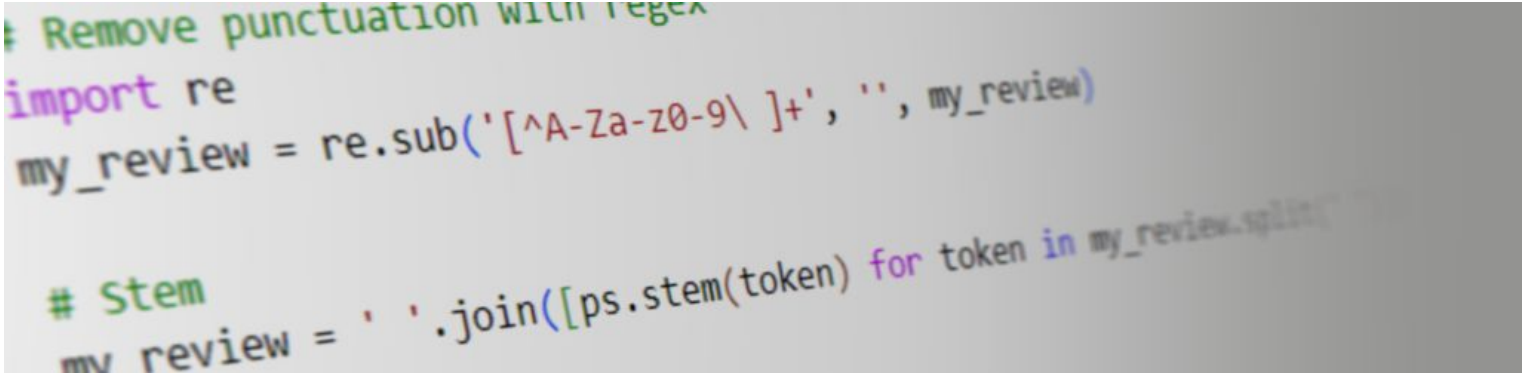


Use Case: Subtitle Generation (*e.g.* Google Meet)

NLP4Free 🔠 ⚡ 🤖 🧠 😀

https://github.com/nlpfromscratch/nlp4free

A Free Natural Language Processing (NLP) microcourse, from basics to deep learning



NLP from scratch

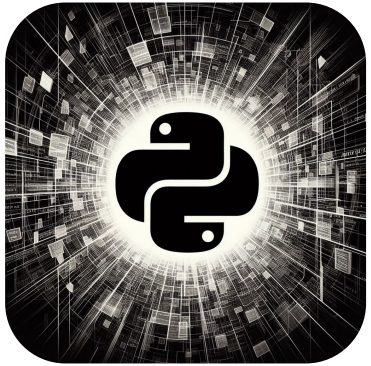
LLM and Generative AI Workshops

Generative Text Models & Fine-tuning LLMs

- Intro to Hugging Face
- LLMs for generative text
- Fine-tuning models
- PEFT (LoRA) and quantization

Building GenAI Apps with OpenAI and GPT

- Intro to APIs and OpenAI
- Working with the OpenAI API
- Setting up a dev environment
- Build a streaming chat app
- Open source alternatives to GPT



Intro to Python & Natural Language Processing

- Intro to NLP
- Intro to Python
- Working with text in Python
- Manipulating text in Pandas

GenAI for Work (1 hr)

- Introduction to GenAI
- Generative AI landscape
- Everyday use cases for GenAI tools
- Prompting and prompt engineering
- Productivity with GenAI

NLP from scratch

TRAINING CONSULTING FAQ A

Today December 2023 Filters

Sun	Mon	Tue	Wed	Thu	Fri	Sat
						2
3	4 What the Heck is a...	5 Deconstruct Care...	6	7 Zero to NLP in 60 ...	8	9
10	11 Large Language M...	12	13 What the Heck is a...	14	15 Large Language M...	16
17	18	19	20	21	22	23

Sign up at **nlpfromscratch.com/training**!

Upcoming eve
Dec 01 - Feb 2

DEC 04 What the LLM? - F

DEC 05 Deconstr Progress

DEC 07 Zero to N FREE Wo

DEC 11 Large La Models a

DEC 13 What the LLM? - F

Manifesto

Knowledge is only valuable if it is useful.

The best way to learn is by doing.

Learning is a non-linear process.

Learning is exploration, not a journey.

Teaching and learning are complementary.

I would value your feedback.

NLP from scratch

www.nlpfromscratch.com