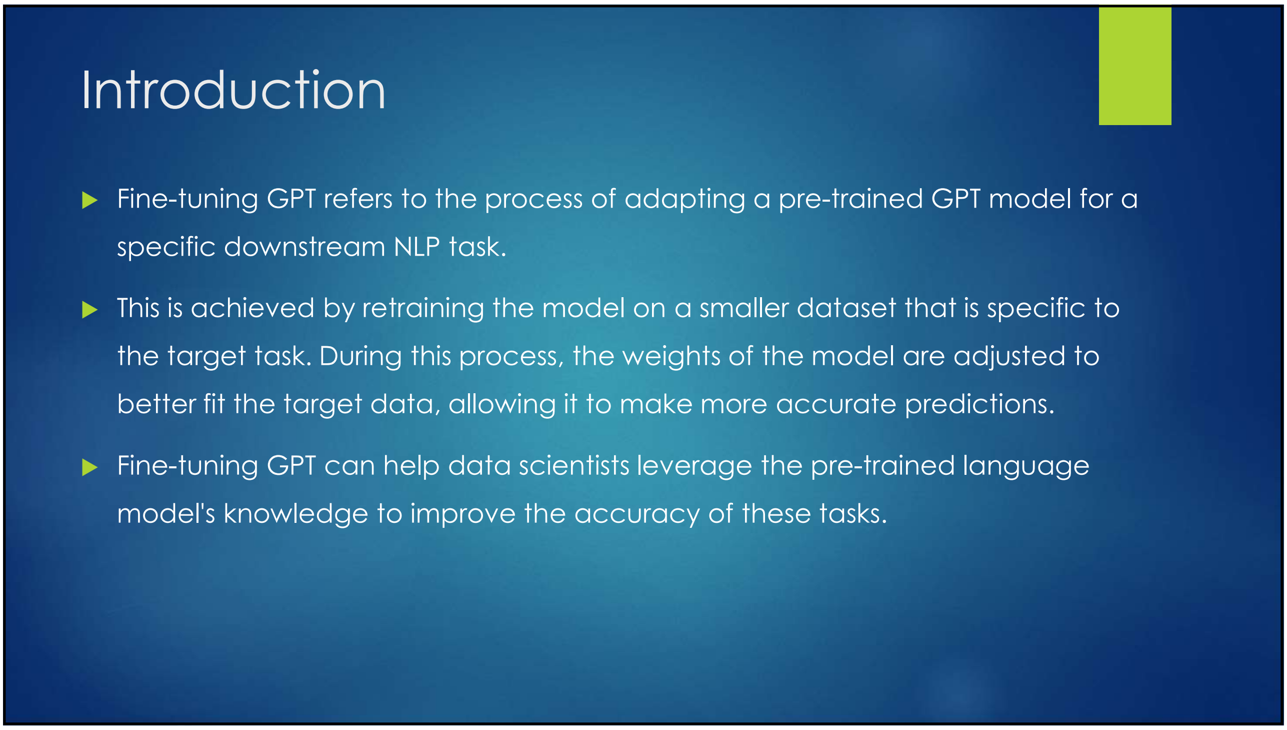
Fine-tuning GPT-3

DR FAKHRELDIN SAEED

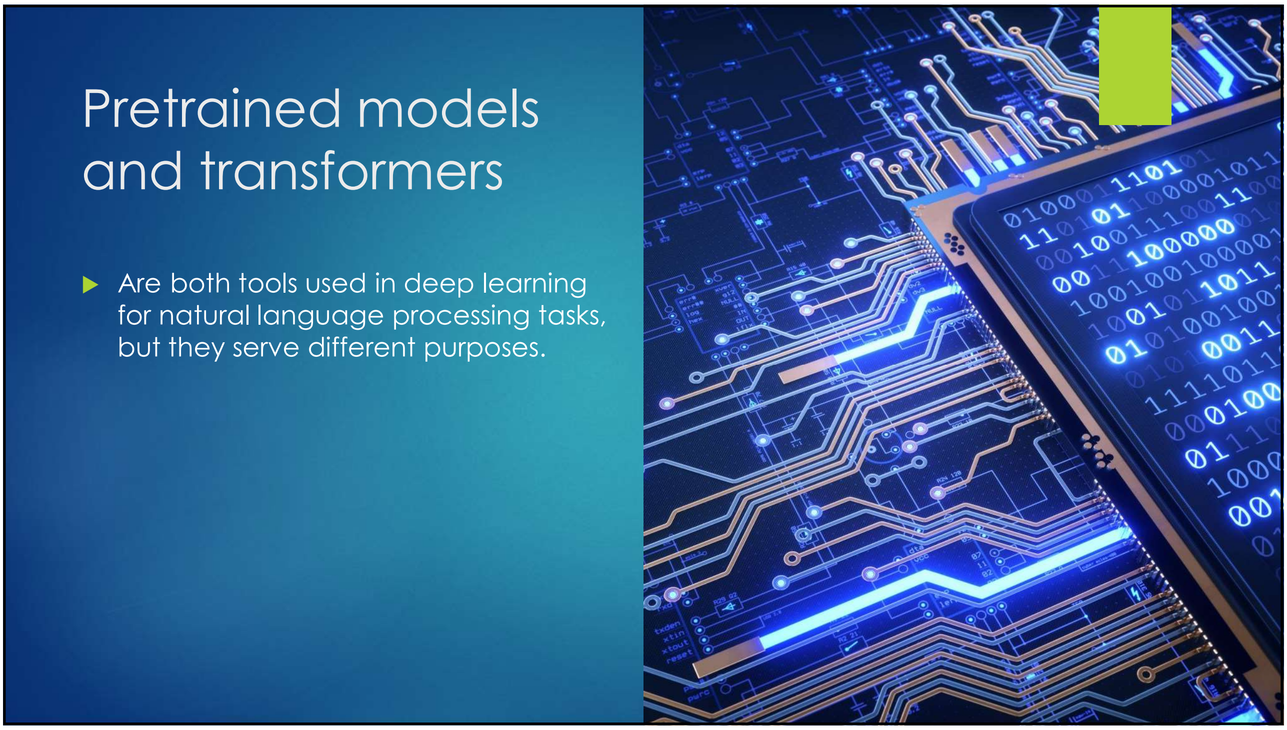
1



Introduction

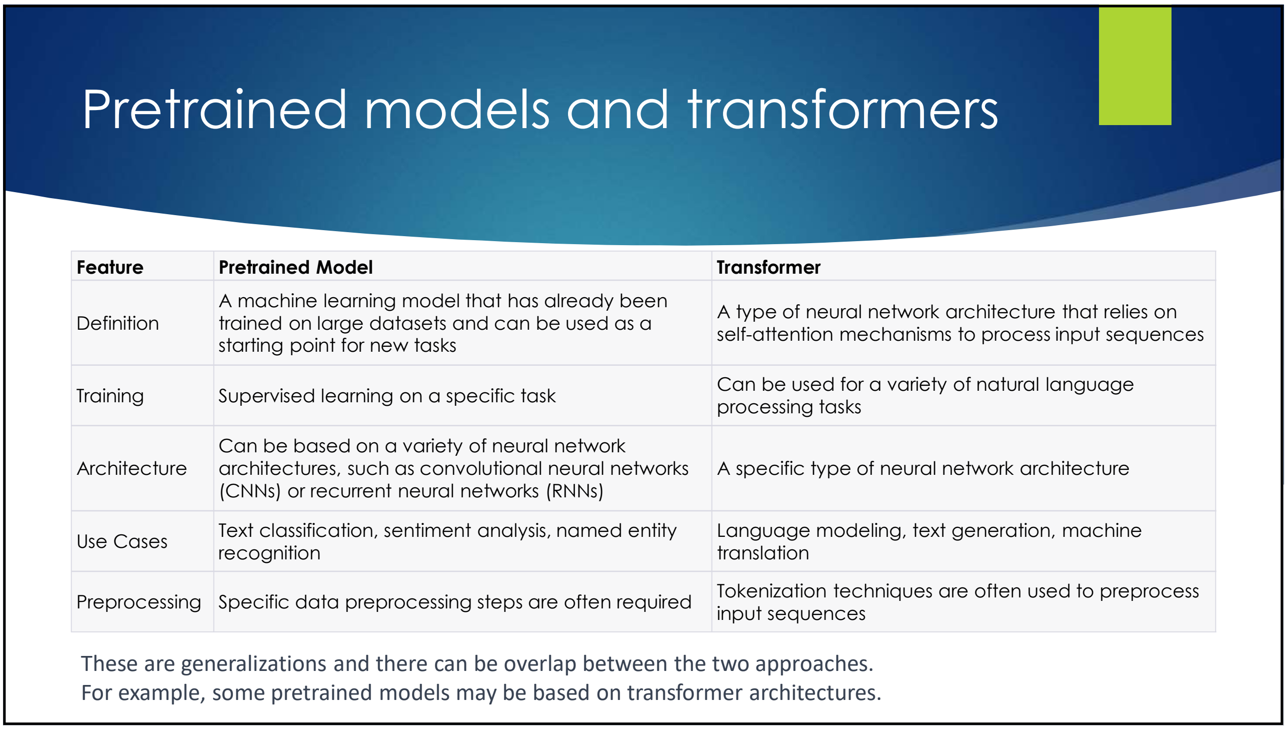
- Fine-tuning GPT refers to the process of adapting a pre-trained GPT model for a specific downstream NLP task.
- This is achieved by retraining the model on a smaller dataset that is specific to the target task. During this process, the weights of the model are adjusted to better fit the target data, allowing it to make more accurate predictions.
- Fine-tuning GPT can help data scientists leverage the pre-trained language model's knowledge to improve the accuracy of these tasks.

2



Pretrained models and transformers

- Are both tools used in deep learning for natural language processing tasks, but they serve different purposes.



Pretrained models and transformers

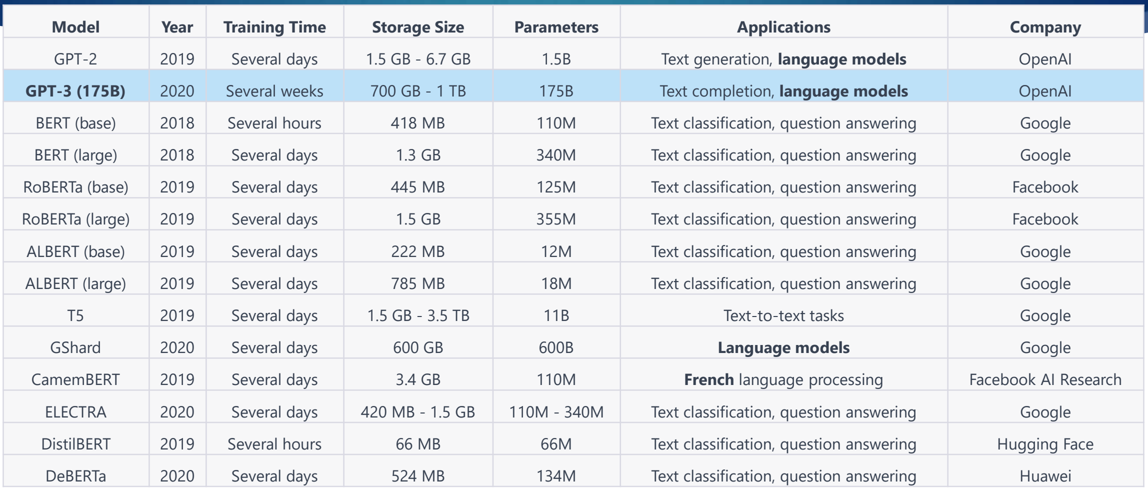
Feature	Pretrained Model	Transformer
Definition	A machine learning model that has already been trained on large datasets and can be used as a starting point for new tasks	A type of neural network architecture that relies on self-attention mechanisms to process input sequences
Training	Supervised learning on a specific task	Can be used for a variety of natural language processing tasks
Architecture	Can be based on a variety of neural network architectures, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs)	A specific type of neural network architecture
Use Cases	Text classification, sentiment analysis, named entity recognition	Language modeling, text generation, machine translation
Preprocessing	Specific data preprocessing steps are often required	Tokenization techniques are often used to preprocess input sequences

These are generalizations and there can be overlap between the two approaches.
For example, some pretrained models may be based on transformer architectures.

Self-attention mechanisms

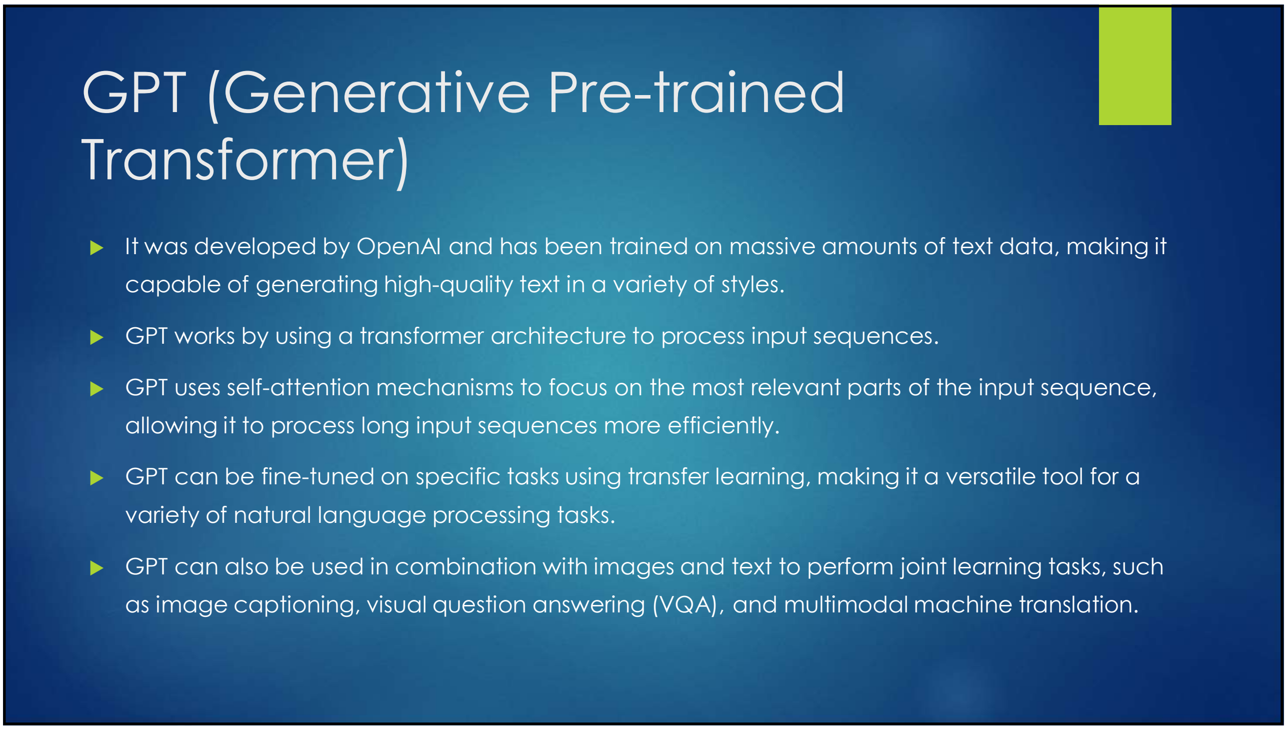
- Are a key component of transformer-based models used in natural language processing.
- Allows the model to attend to different parts of the input sequence when processing each individual element of the sequence.
- It works by computing attention weights between every pair of positions in the input sequence.
- The model learns to weigh each position in the input sequence based on how relevant it is to each other position in the sequence.
- Its advantages over traditional RNN.
    - capture long-range dependencies between different parts of the input sequence.
    - it allows the model to process input sequences in parallel (more efficient)
    - it allows us to visualize which parts of the input sequence are most important for each output element(more interpretable).

Popular transformer models used for NLP



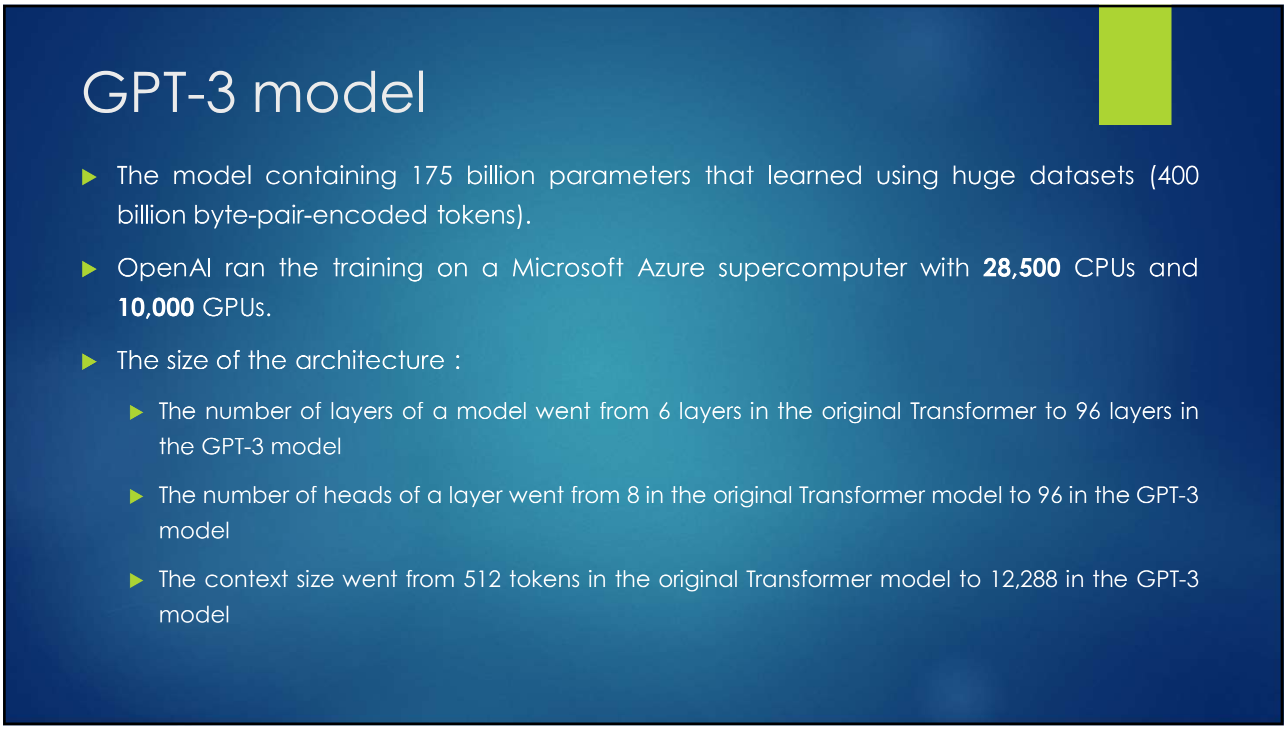
Model	Year	Training Time	Storage Size	Parameters	Applications	Company
GPT-2	2019	Several days	1.5 GB - 6.7 GB	1.5B	Text generation, **language models**	OpenAI
GPT-3 (175B)	2020	Several weeks	700 GB - 1 TB	175B	Text completion, **language models**	OpenAI
BERT (base)	2018	Several hours	418 MB	110M	Text classification, question answering	Google
BERT (large)	2018	Several days	1.3 GB	340M	Text classification, question answering	Google
RoBERTa (base)	2019	Several days	445 MB	125M	Text classification, question answering	Facebook
RoBERTa (large)	2019	Several days	1.5 GB	355M	Text classification, question answering	Facebook
ALBERT (base)	2019	Several days	222 MB	12M	Text classification, question answering	Google
ALBERT (large)	2019	Several days	785 MB	18M	Text classification, question answering	Google
T5	2019	Several days	1.5 GB - 3.5 TB	11B	Text-to-text tasks	Google
GShard	2020	Several days	600 GB	600B	**Language models**	Google
CamemBERT	2019	Several days	3.4 GB	110M	**French** language processing	Facebook AI Research
ELECTRA	2020	Several days	420 MB - 1.5 GB	110M - 340M	Text classification, question answering	Google
DistilBERT	2019	Several hours	66 MB	66M	Text classification, question answering	Hugging Face
DeBERTa	2020	Several days	524 MB	134M	Text classification, question answering	Huawei

Language Models like Text generation, Language translation, Sentiment analysis, Text summarization, Question answering, Named entity recognition, and Speech recognition



GPT (Generative Pre-trained Transformer)

- It was developed by OpenAI and has been trained on massive amounts of text data, making it capable of generating high-quality text in a variety of styles.
- GPT works by using a transformer architecture to process input sequences.
- GPT uses self-attention mechanisms to focus on the most relevant parts of the input sequence, allowing it to process long input sequences more efficiently.
- GPT can be fine-tuned on specific tasks using transfer learning, making it a versatile tool for a variety of natural language processing tasks.
- GPT can also be used in combination with images and text to perform joint learning tasks, such as image captioning, visual question answering (VQA), and multimodal machine translation.



GPT-3 model

- The model containing 175 billion parameters that learned using huge datasets (400 billion byte-pair-encoded tokens).
- OpenAI ran the training on a Microsoft Azure supercomputer with **28,500** CPUs and **10,000** GPUs.
- The size of the architecture :
  - The number of layers of a model went from 6 layers in the original Transformer to 96 layers in the GPT-3 model
  - The number of heads of a layer went from 8 in the original Transformer model to 96 in the GPT-3 model
  - The context size went from 512 tokens in the original Transformer model to 12,288 in the GPT-3 model

Steps involved in Fine-tuning GPT

Dataset Preparation: The dataset should be specific to the task and should contain enough data to train the model effectively.

Pre-processing the Data: This includes removing stop words, stemming, and converting the text to lowercase.

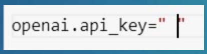
Fine-tuning the GPT Model: During this process, the weights of the model are adjusted to fit the target data. The fine-tuning process typically involves training the model on the target task for several epochs.

Evaluating the Fine-tuned Model: Once the fine-tuning is complete, the model's performance is evaluated on a validation set.

9

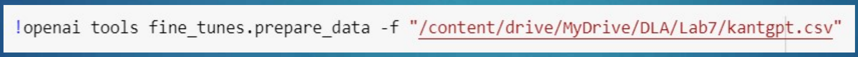
Preparing the dataset

- Step 1: Installing OpenAI & Wandb
- Step 2: Your API Key



```
openai.api_key=" "
```

- Step 3: Preparing the data

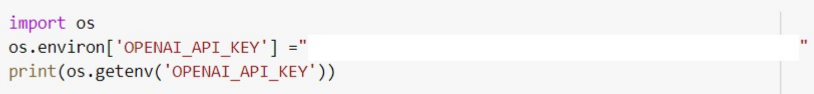


```
!openai tools fine_tunes.prepare_data -f "/content/drive/MyDrive/DLA/Lab7/kantgpt.csv"
```

10

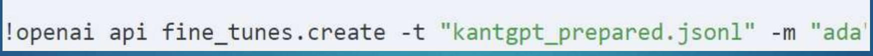
Fine-tuning ADA

- Step 4: Creating an OS environment for the API key

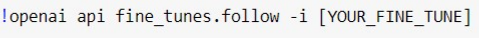


```
import os
os.environ['OPENAI_API_KEY'] ="                                                  "
print(os.getenv('OPENAI_API_KEY'))
```

- Step 5: Fine-tuning GPT-3 with the ADA engine

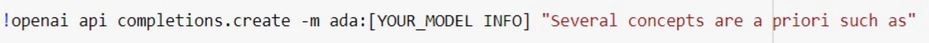


```
!openai api fine_tunes.create -t "kantgpt_prepared.jsonl" -m "ada'
```



```
!openai api fine_tunes.follow -i [YOUR_FINE_TUNE]
```

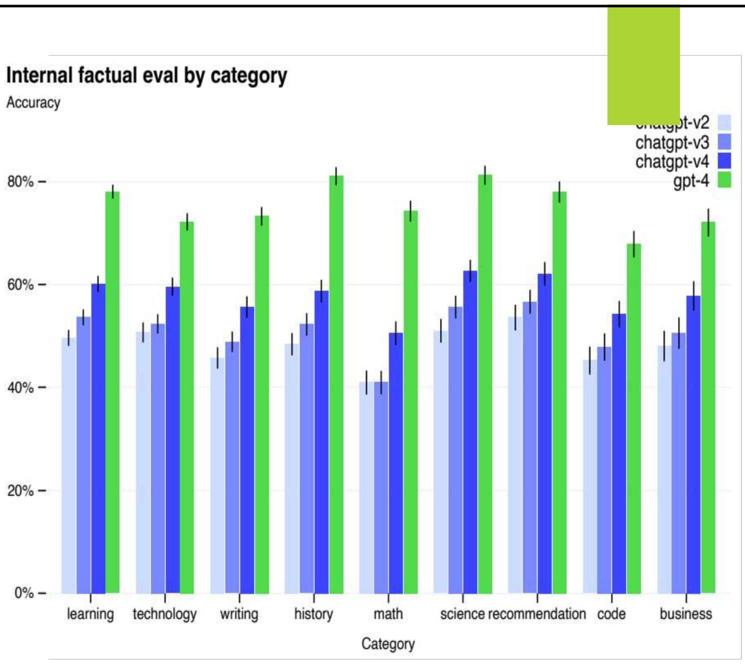
- Step 6: Using the fine-tuned GPT-3 for a completion task

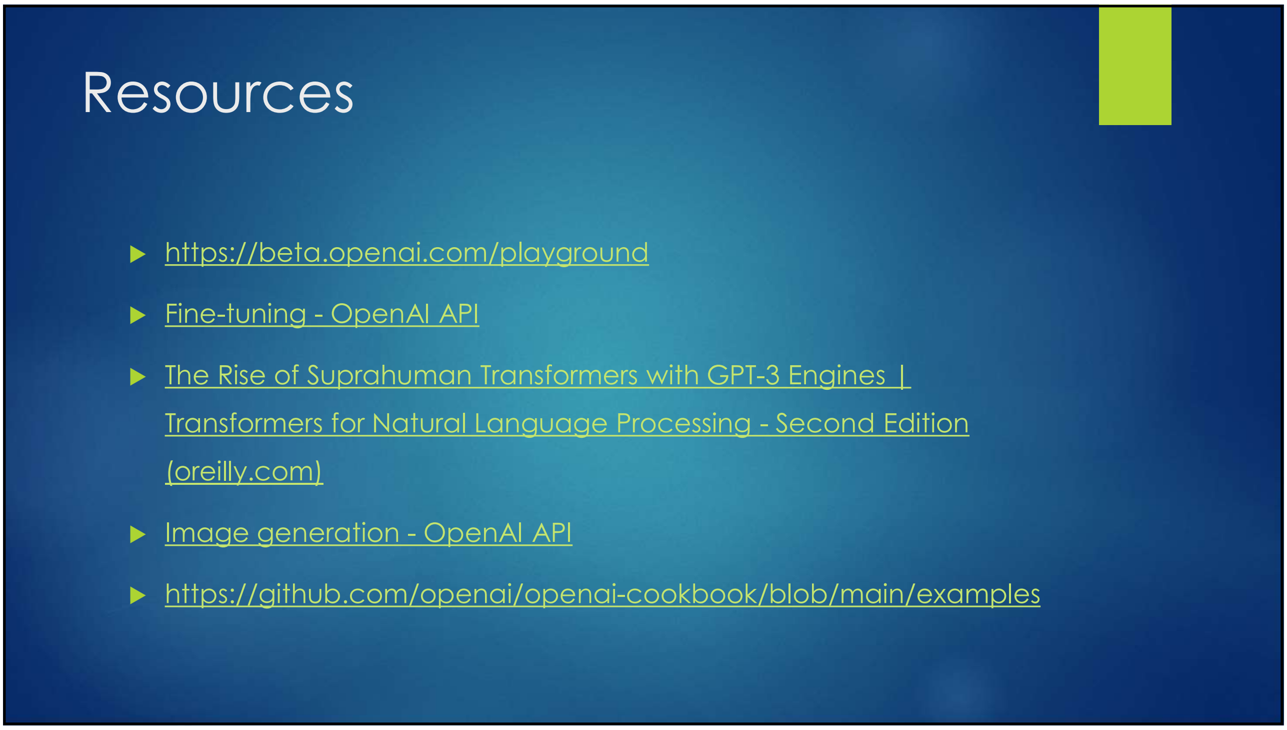


```
!openai api completions.create -m ada:[YOUR_MODEL INFO] "Several concepts are a priori such as"
```

GPT-4–100X More Powerful than GPT-3

- GPT-4 is the latest and most advanced version of the GPT (16th March 2023).
- GPT-4 is significantly larger and more powerful than GPT-3, with **170 trillion** parameters compared to GPT-3's **175 billion** parameters.
- GPT-4 can use image inputs.





Resources

- https://beta.openai.com/playground
- Fine-tuning - OpenAI API
- The Rise of Suprahuman Transformers with GPT-3 Engines | Transformers for Natural Language Processing - Second Edition (oreilly.com)
- Image generation - OpenAI API
- https://github.com/openai/openai-cookbook/blob/main/examples