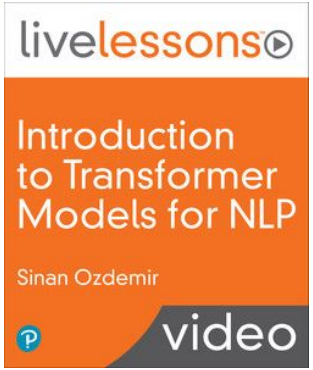
Hugging Face in 4 Hours

Sinan Ozdemir

Data Scientist, Entrepreneur,
Author, Lecturer

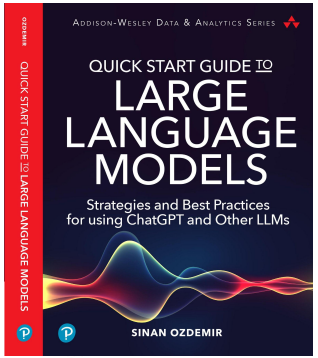
Pearson

Welcome!



My name is **Sinan Ozdemir** (in/sinan-ozdemir + @prof_oz)

- Current **founder** of Loop Genius (using AI to help entrepreneurs get their first 100 customers)
- Current **lecturer** for O'Reilly and Pearson
- Founder of Kylie.ai (Funded by OpenAI Founder + Acquired)
- **Masters** in Theoretical Math from **Johns Hopkins**
- Former lecturer of Data Science at Johns Hopkins



Author of ML textbooks and online series, including

- [Quick Start Guide to LLMs]
- [Introduction to Transformer Models for NLP]

Pearson

Hugging Face in 4 Hours

Sinan Ozdemir

Data Scientist, Entrepreneur,
Author, Lecturer

Pearson

Expectations for Today

We will spend most of our time together with my screen shared.

I will be showing off components of HuggingFace and code for using HuggingFace models, data, and APIs

Hugging Face in 4 Hours

Segment 1: Introduction to Hugging Face and Its Ecosystem

Sinan Ozdemir

Data Scientist, Entrepreneur,
Author, Lecturer

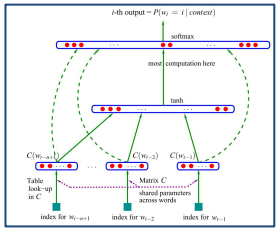
Pearson

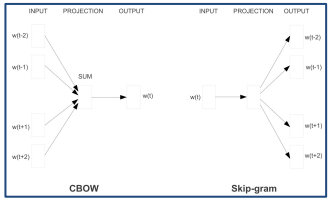
Introduction to transformer models and their significance in NLP

Pearson

Brief History of Modern NLP

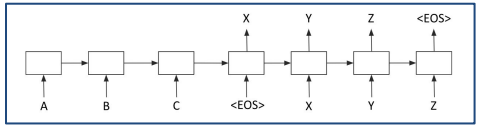
2001
Neural Language Models

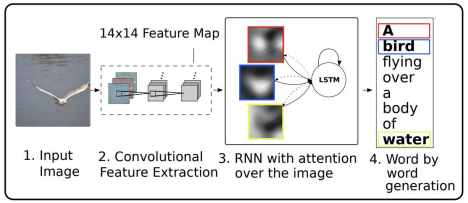




2013
encoding semantic meaning with Word2vec

2014 - 2017
Seq2seq + Attention





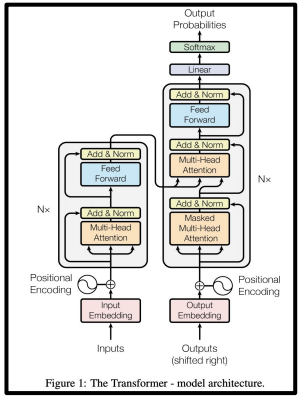


Figure 1: The Transformer - model architecture.

2017 - Present
Transformers + Large Language Models

Pearson

Bengio et al. https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf
Mikolov et al. https://arxiv.org/abs/1301.3781
Xu et al. http://proceedings.mlr.press/v37/xuc15.pdf
https://papers.nips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

2017 – Transformers

“Attention is all you need”

- Introduced the Transformer architecture
- A sequence to sequence model (takes text in and writes text back)
- The parent model of GPT3, BERT, T5, and many more

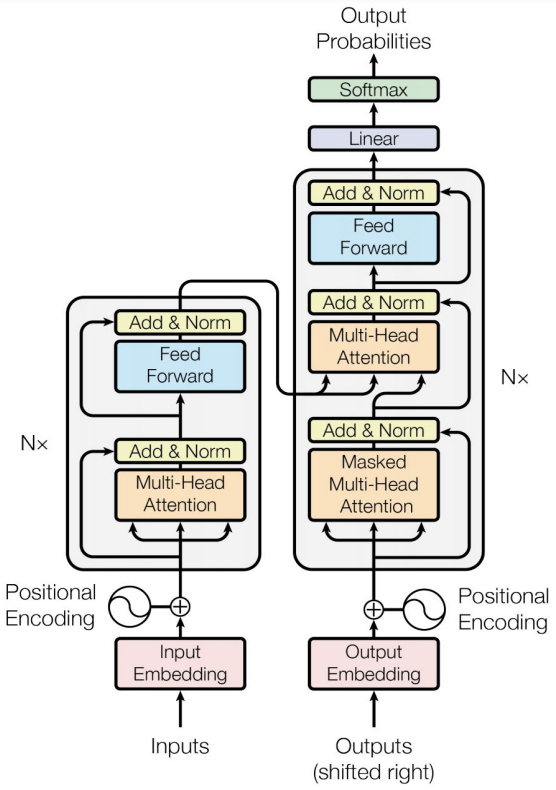


Figure 1: The Transformer - model architecture.

Source: https://papers.nips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Auto-__ Language Models

Auto-regressive Models

Predict a future token (word) given either the past tokens or the future tokens but not both.

If you don't ___ (forward prediction)

Auto-encoding Models

Learn representations of the entire sequence by predicting tokens given both the past and future tokens.

If you don't ___ at the sign, you will get a ticket.

Using LLMs

We can use LLMs in (generally) three ways:

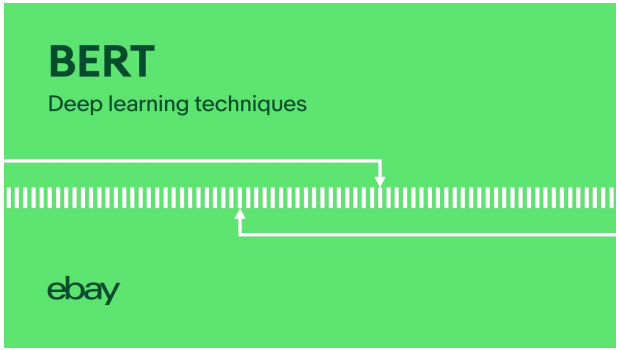
1. **Encode** text into semantic vectors with little/no fine-tuning
   a. Eg. Creating an information retrieval system using BERT vectors

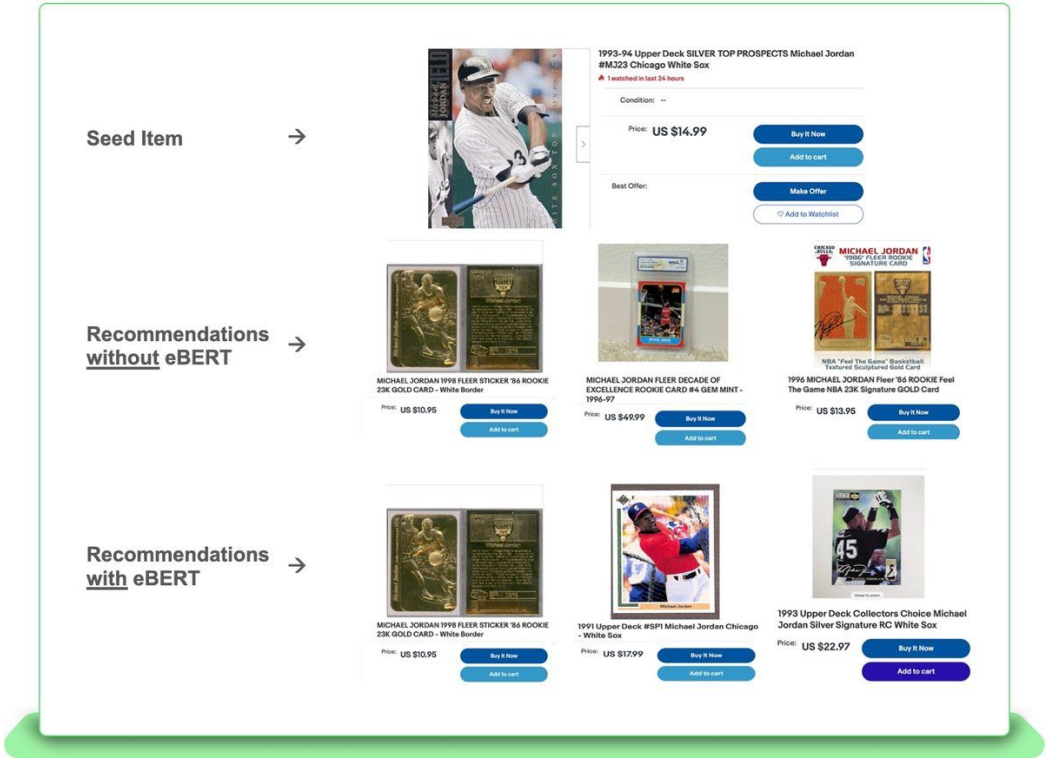
2. Fine-tune a pre-trained LLM to perform a very specific task using **Transfer Learning**
   a. Eg. Fine-tuning BERT to classify sequences with labels

3. Ask an LLM to solve a task it was pre-trained to solve or could intuit
   a. Eg. **Prompting** GPT3 to write a blog post
   b. Eg. **Prompting** T5 to perform language translation

Encoding Ebay's Recommendations with BERT

Ebay uses BERT to generate more relevant recommendations than traditional search techniques





Pearson

Source: https://tech.ebayinc.com/engineering/how-ebay-created-a-language-model-with-three-billion-item-titles

Overview of Hugging Face capabilities and community

Pearson

Huggingface.co / **models**

That's a lot

Choose the type of model you need

Hugging Face | Search models, datasets, u | Models | asets | Spaces | Posts | Docs | Pricing

Tasks | Libraries | Datasets | Languages | Licenses | Other

Filter Tasks by name

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

Models 487,674 | Filter by name | new Full-text search | Sort: Trending

mistralai/Mixtral-8x7B-Instruct-v0.1
Text Generation • Updated Dec 15, 2023 • 1.21M • 2.49k

vikhyatk/moondream1
Updated 31 minutes ago • 192

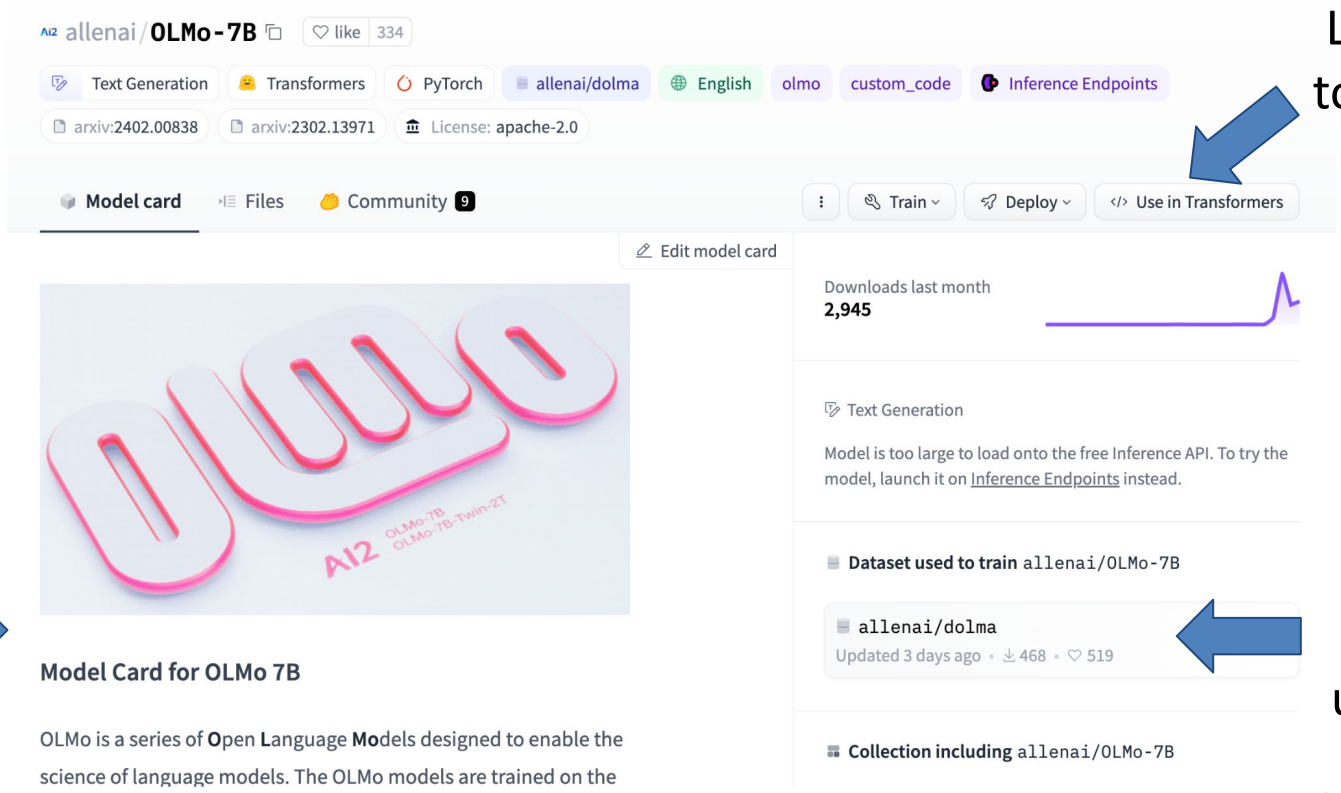
InstantX/InstantID
Text-to-Image • Updated 8 days ago • 36.7k • 213

miqudev/miqu-1-70b
Updated 2 days ago • 134

stabilityai/stable-code-3b
Text Generation • Updated about 23 hours ago • 7.46k • 439

Huggingface.co / **model_page**

Model tags including license



List ways to use the model

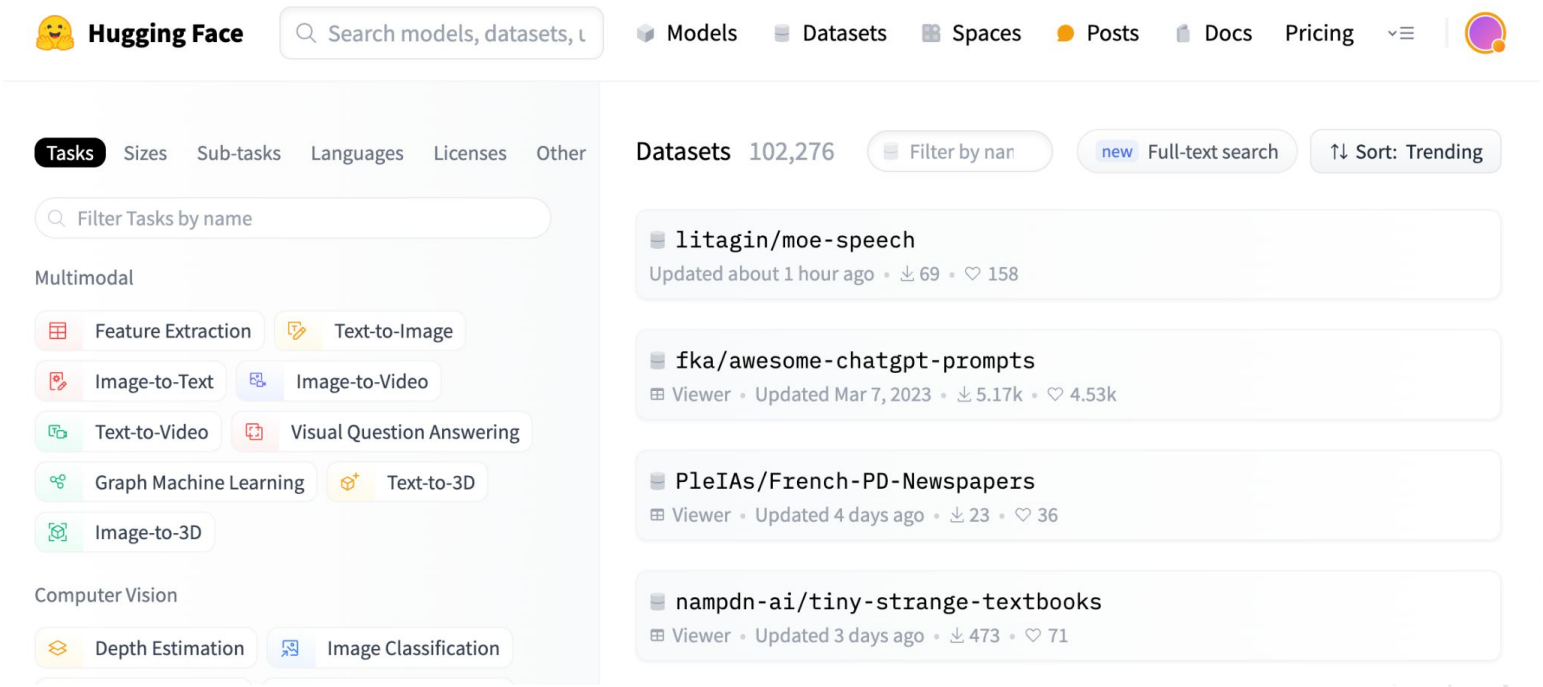
General Information

List the datasets used and spaces used in

Pearson

Huggingface.co / **datasets**

Choose the type of data you need



Pearson

Huggingface.co / data_page

Data tags including license

List ways to use data

Info on columns and showing rows

Show the dataset size

Datasets: math-ai / AutoMathText

like 53

Tasks: Text Generation | Question Answering

Languages: English

Size Categories: 10B<n<100B

Tags: mathematical-reasoning | reasoning | finetuning | +2

License: cc-by-sa-4.0

Dataset card | Files | Community 1

Dataset Viewer (First 5GB)

Auto-converted to Parquet | API | Go to dataset viewer

Subset: arxiv-0.50-to-1.00 (77.5k rows)

Split: train (77.5k rows)

Search this dataset

url string · lengths 31 38	title string · lengths 7 229	abstract string · lengths 39 2.87k	text string · leng 1
https://arxiv.org/abs/2105.10615	Convergence directions of the…	The randomized Gauss--Seidel…	\section{Int: problem is a
https://arxiv.org/abs/1912.01763	A note on semi-infinite program…	Semi-infinite programs are a…	\section{Int: methods for

Downloads last month 115

Use in Datasets library

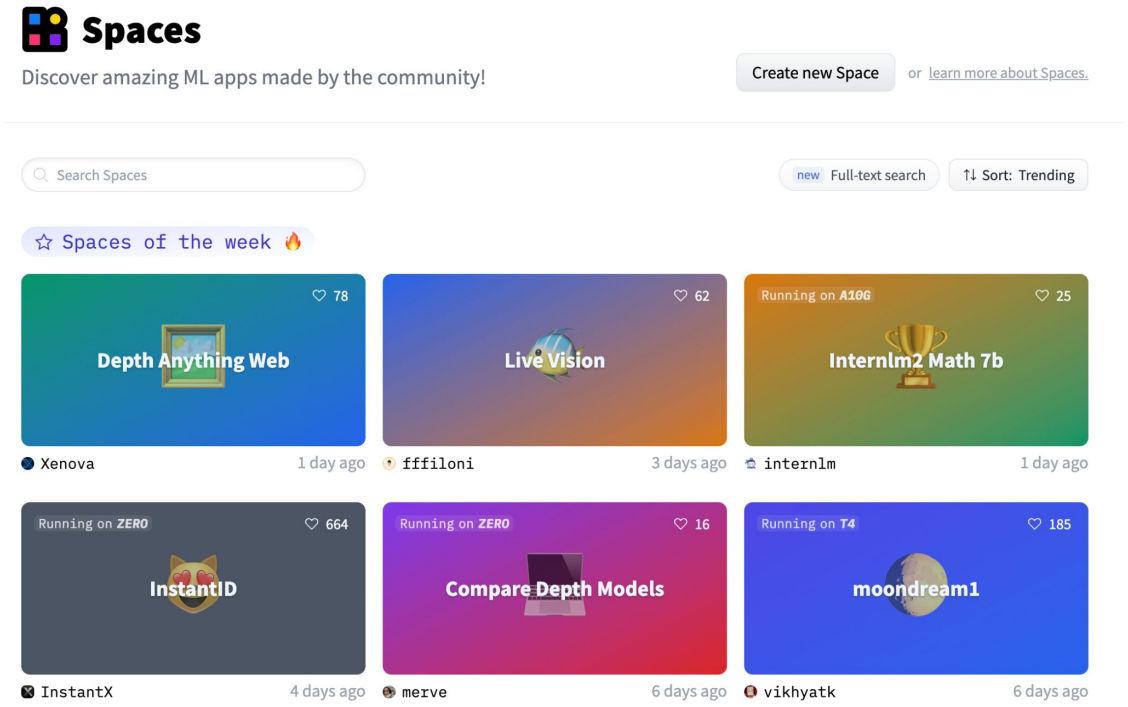
Edit dataset card

Size of the auto-converted Parquet files (First 5GB per…
19.7 GB

Number of rows (First 5GB per split):
3,915,090

Pearson

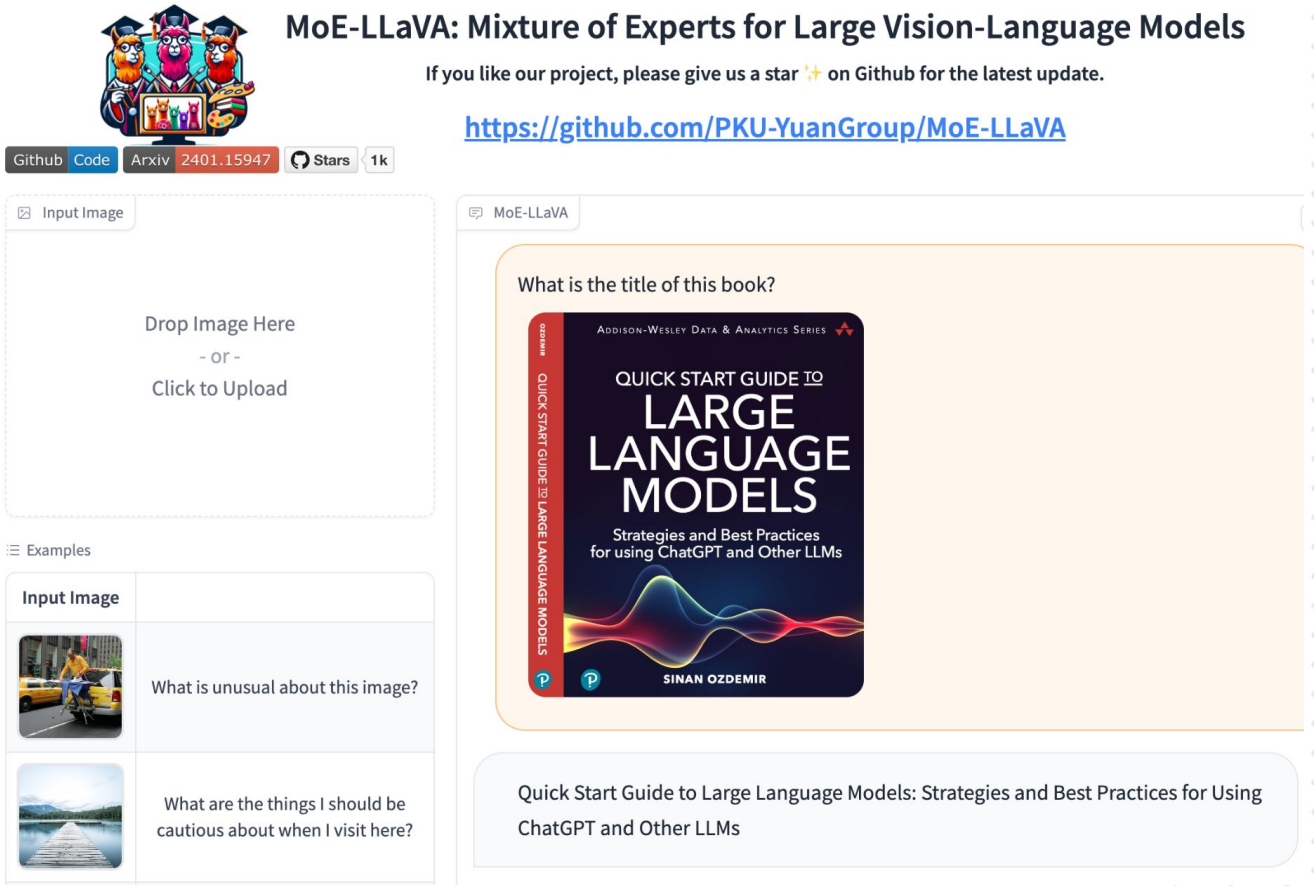
Huggingface.co / **spaces**



Pearson

Huggingface.co / **spaces**

Spaces are a great way to see what's the latest and greatest in open-source



Pearson

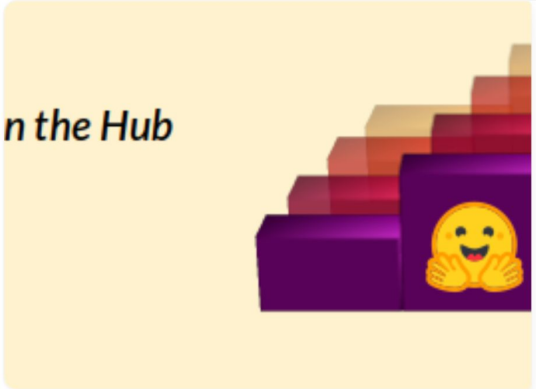
Huggingface.co / **blog**

Posts, articles, and discussions

Everything | Community | Guide | Open Source Collab | Partnerships

Research | NLP | Audio | CV | RL | Ethics | Diffusion | Game Development

Time Series | RLHF | Case Studies



The Hallucinations Leaderboard, an Open Effort to Measure Hallucinations in Large Language Models

By pminervini
January 29, 2024 guest

Community blog posts view all

Building autograd engine tinytorch 03
By joey00072 • about 9 hou...

Building autograd engine tinytorch 02
By joey00072 • about 9 hou...

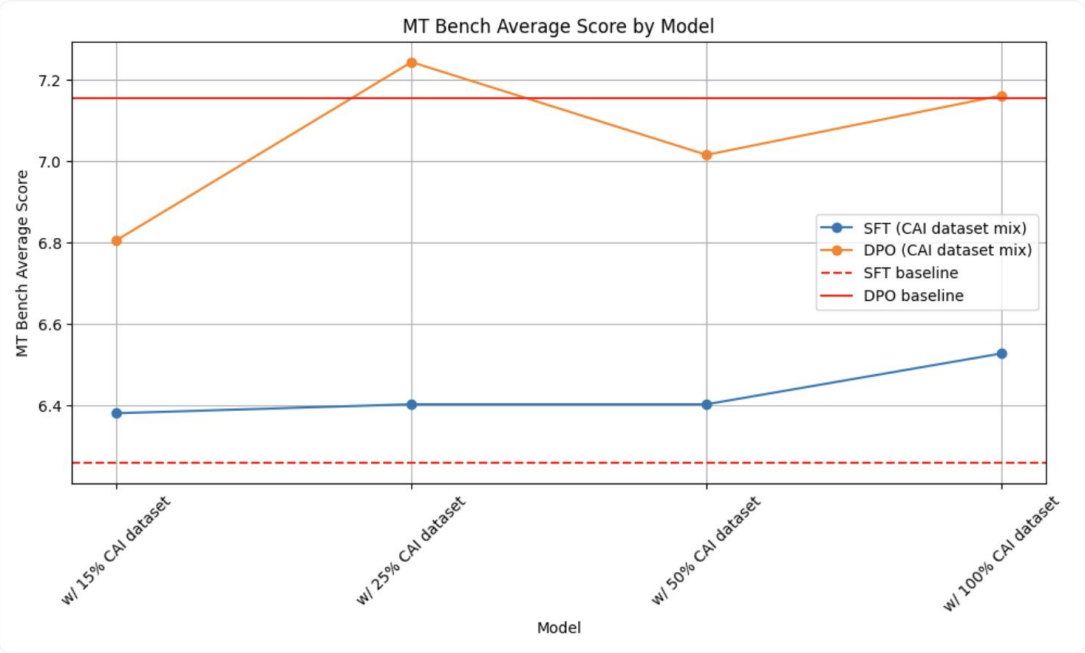
Fine Tuning a LLM Using Kubernetes with Intel® Xeon® Scalable Processors
By dmsuehir • 7 days ago

Create a Web Interface for your LLM in Python
By Alex1337 • 7 days ago

makeMoE: Implement a Sparse Mixture of Experts Language Model from Scratch

Huggingface.co / **blog**

HF often has deep technical dives on things like Constitutional AI



We found training on the CAI dataset does not necessarily reduce helpfulness (i.e., paying the alignment tax). The SFT models obtained higher MT bench scores by training on

Source: https://huggingface.co/blog/constitutional_ai

Huggingface.co / **blog**

Both Giada and Nathan have been on my show!

← Back to blog

Ethics and Society Newsletter #4: Bias in Text-to-Image Models

Published June 26, 2023

Update on GitHub

sasha
Sasha Luccioni

giadap
Giada Pistilli

nazneen
Nazneen Rajani

allendorf
Elizabeth Allendorf

irenesolaiman
Irene Solaiman

natolambert
Nathan Lambert

meg
Margaret Mitchell

TL;DR: We need better ways of evaluating bias in text-to-image models

Source: https://huggingface.co/blog/ethics-soc-4

Code Time!

Pearson

Hugging Face in 4 Hours

Segment 2: Fine-tuning and Utilizing Pre-trained models

Sinan Ozdemir

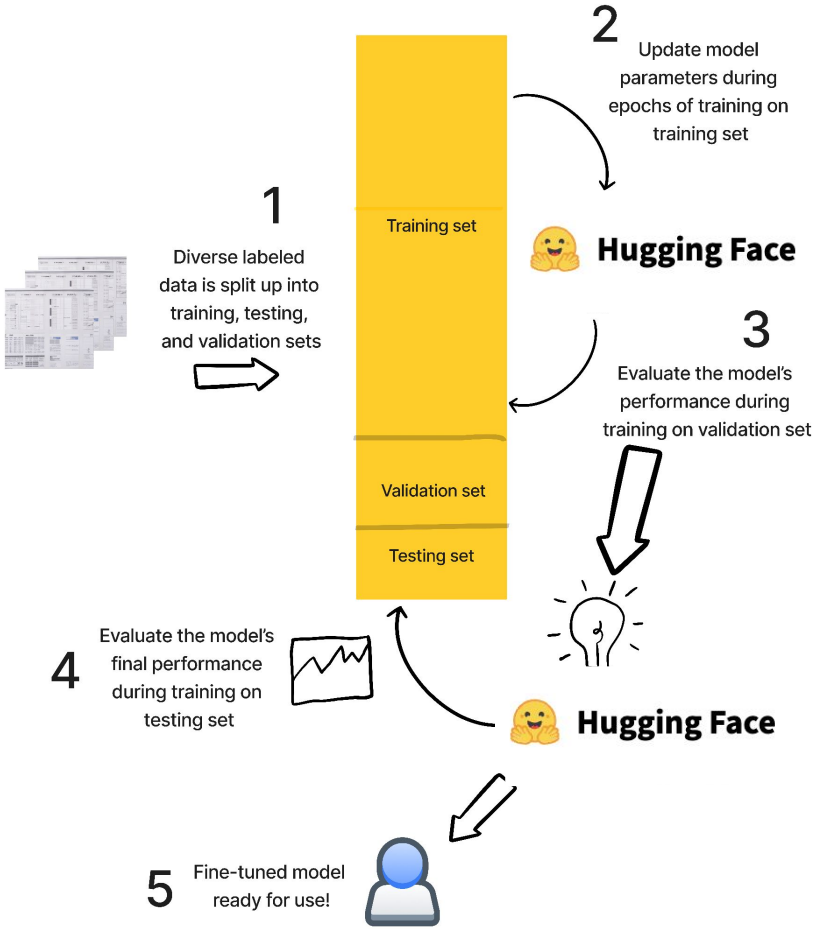
Data Scientist, Entrepreneur,
Author, Lecturer

Pearson

Walkthrough of model fine-tuning process

Pearson

Basic Fine-Tuning Process



Source: Quick Start Guide to LLMs by Sinan Ozdemir

Transfer Learning

Transfer Learning - A model trained for one task is reused as the starting point for a model for a second task.

1. Select a source model from a repository of models (like Huggingface)

Hugging Face

2. Reuse and train the model for a second task using task-specific data

Transfer Learning with BERT

Hugging Face | Search models, datasets, use | Models | Datasets | Spaces | Docs | Soluti

Tasks

- Fill-Mask
- Question Answering
- Summarization
- Table Question Answering
- Text Classification
- Text Generation
- Text2Text Generation
- Token Classification
- Translation
- Zero-Shot Classification
- Sentence Similarity
- + 16

Libraries

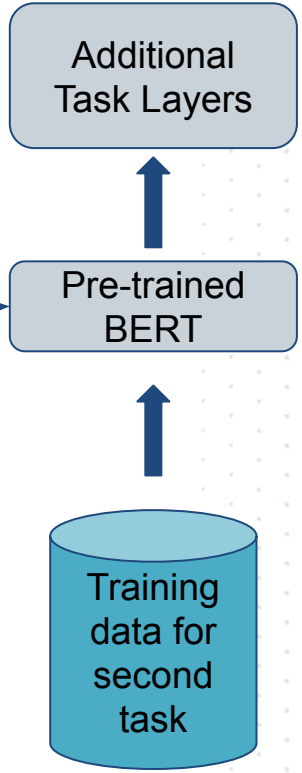
Models 35,367 | Search Models

distilgpt2
Text Generation • Updated May 21, 2021 • ↓ 33.2M • ♡ 39

bert-base-uncased
Fill-Mask • Updated May 18, 2021 • ↓ 16.2M • ♡ 125

cross-encoder/ms-marco-MiniLM-L-12-v2
Text Classification • Updated Aug 5, 2021 • ↓ 11M • ♡ 5

Selecting a source model



Reusing and training model

BERT vs ChatGPT

Hugging Face

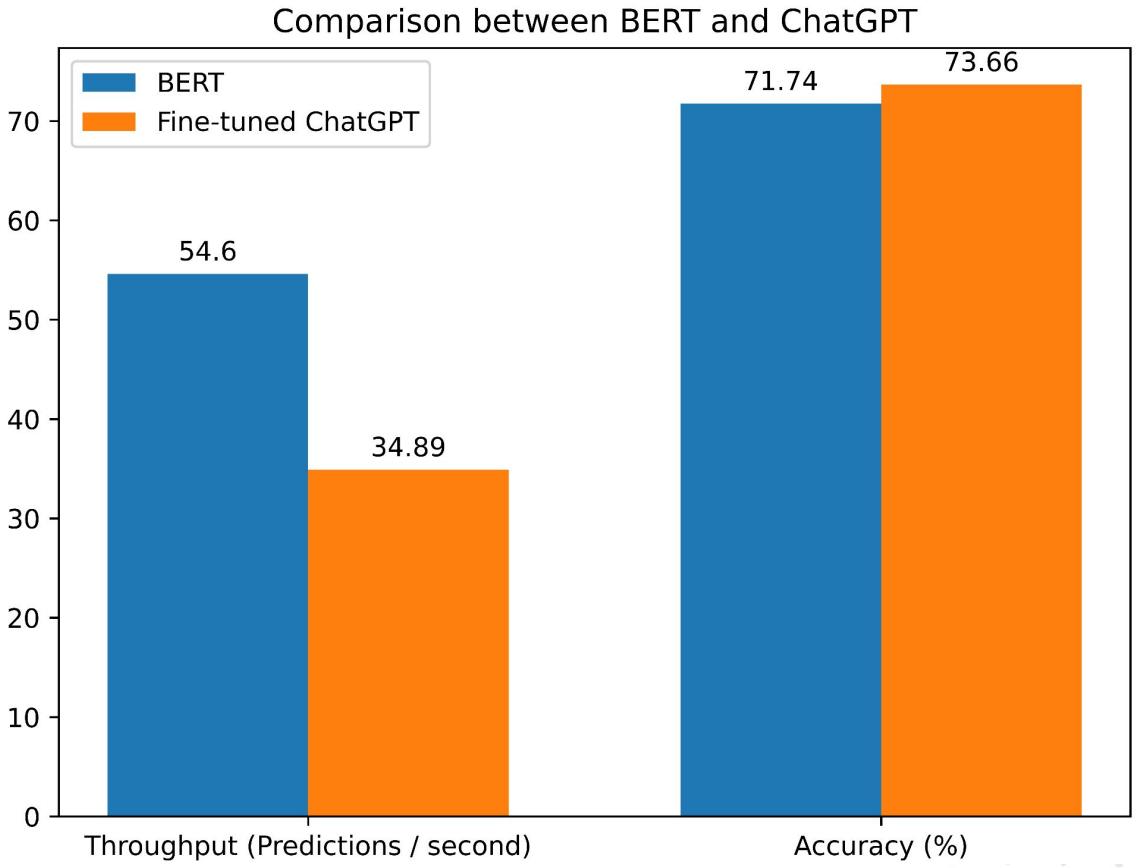
Search moc

Datasets: **app_reviews**

Given a review, predict # stars

The BERT model has roughly 70M params and ChatGPT has ~175B

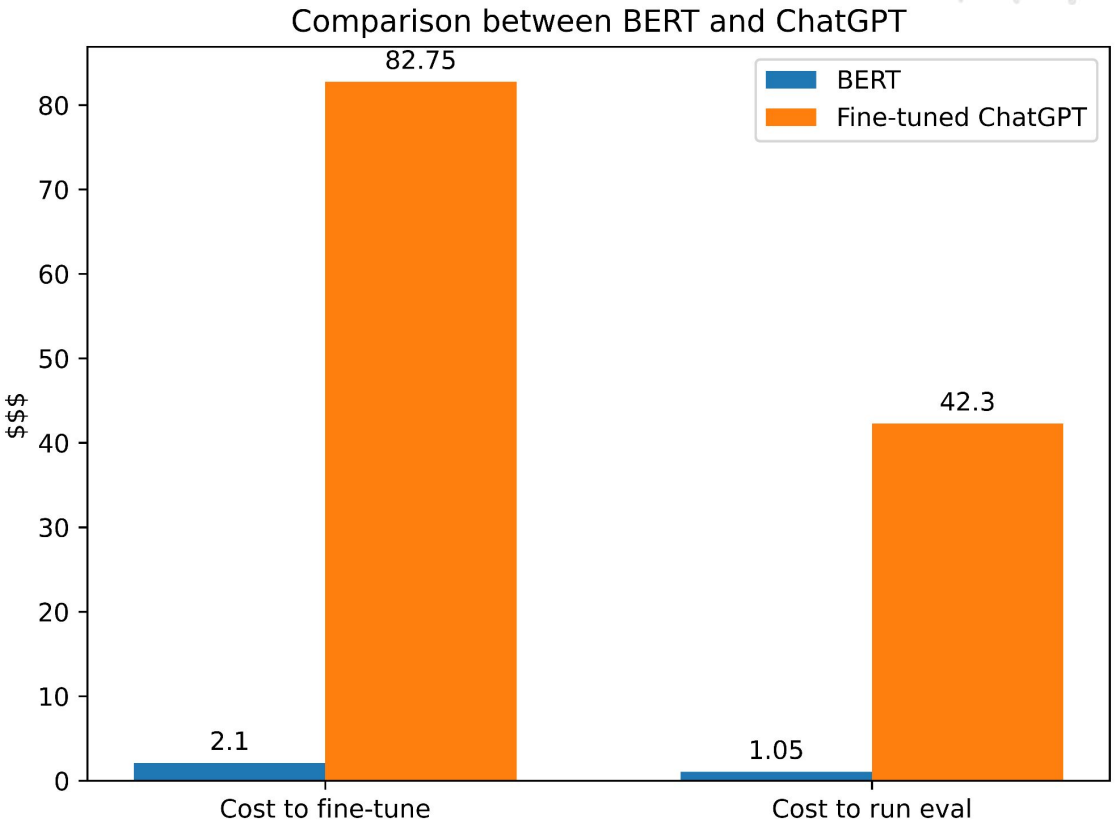
So BERT is ~2,500x smaller than ChatGPT but performances on par



Pearson

BERT vs ChatGPT

BERT is also much cheaper / faster to train



Pearson

Considering Open-source

Auto-encoding LLMs

Learns entire sequences by predicting tokens (words) given past and future context

If you don't __ at the sign, you will get a ticket.

cannot generate text but great for **classification**, **embedding** + **retrieval** tasks

Examples: **BERT**, XLNET, RoBERTa, sBERT

Auto-regressive LLMs

Predict a future token (word) given either past context or future context but not both.

If you don't __ ..... mind? want? have?

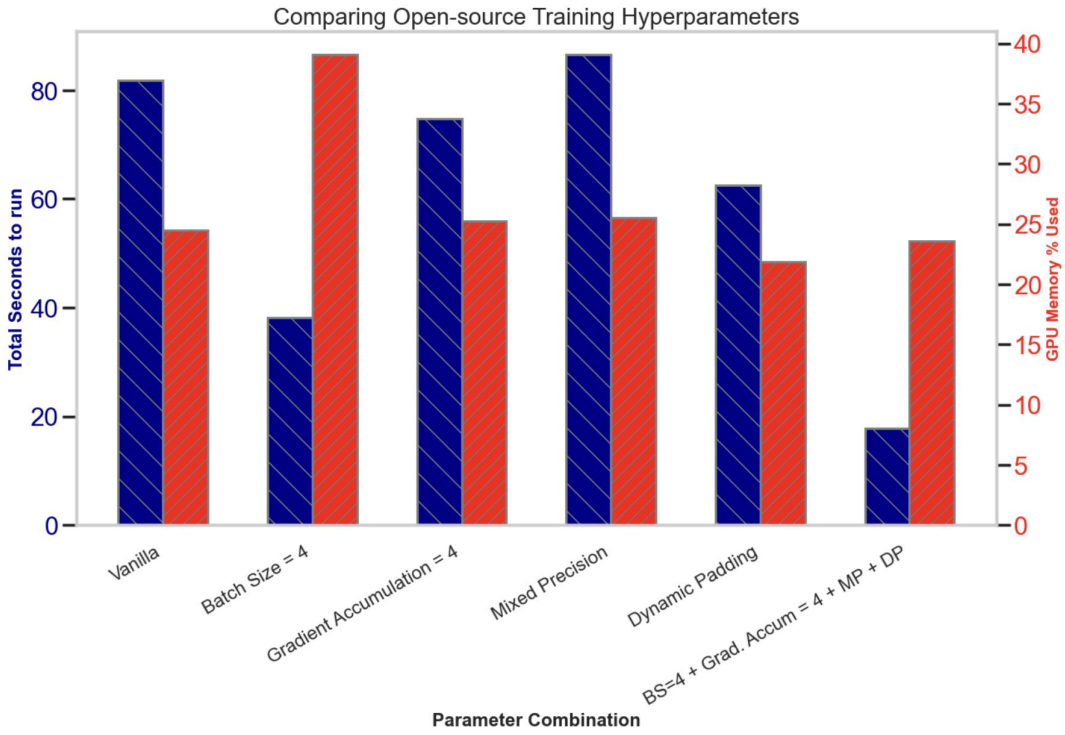
Capable of **generating text**, hence the term Generative LLMs but must be larger to read nearly as well as auto-encoding systems

Examples: **GPT** family, Llama family, Anthropic's Claude family, honestly most of the LLMs you see out there today

Optimizing Fine-tuning

Some smaller techniques (see more in my book or on my other lectures) can also be used to speed up training without consuming more memory.

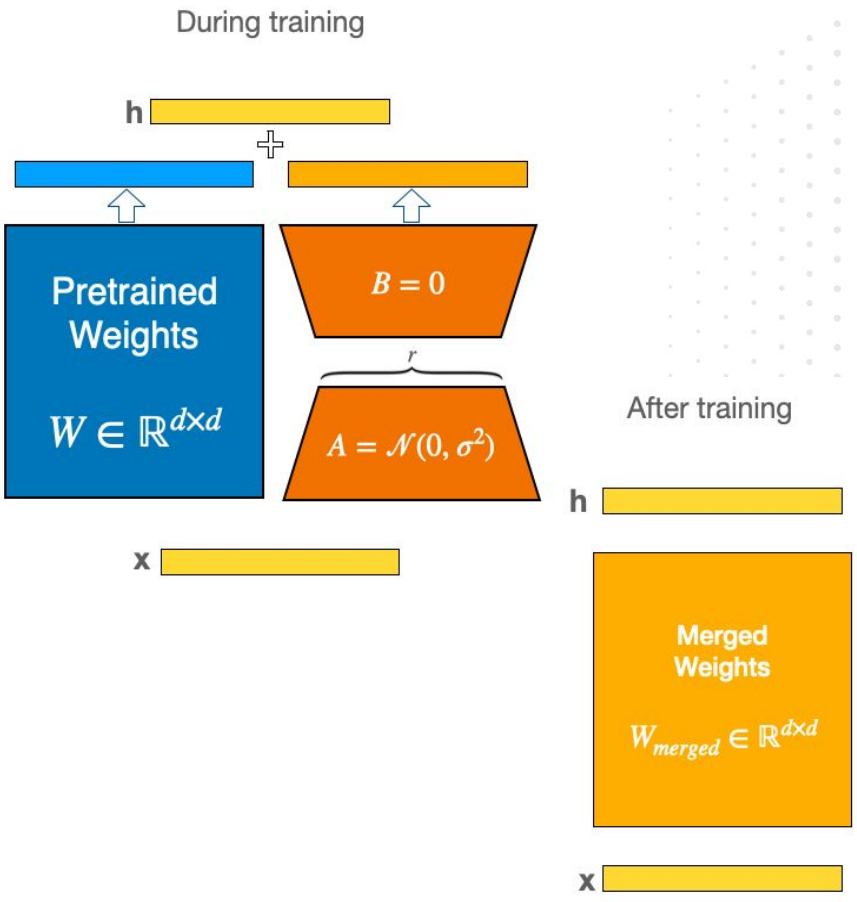
This graph shows a 4x speed up in training a classifier utilizing the same memory footprint



Advanced: PEFT to optimize memory

Parameter **E**fficient **F**ine-**T**uning techniques like

LoRA (**Lo**w-**R**ank **A**daptation) allow for training of larger models on smaller/single GPUs



github.com/huggingface

Hugging Face

The AI community building the future.

27.3k followers | NYC + Paris | https://huggingface.co/ | @huggingface | Verified

Pinned

transformers Public

🤗 Transformers: State-of-the-art Machine Learning for Pytorch, TensorFlow, and JAX.

Python | 120k | 23.9k

diffusers Public

🤗 Diffusers: State-of-the-art diffusion models for image and audio generation in PyTorch

Python | 20.7k | 4.3k

Exercise: Fine-tuning a sample model on a dataset

Code Time!

Pearson

Code Time!

Pearson

Hugging Face in 4 Hours

Segment 3: Deployment Strategies with Hugging Face

Sinan Ozdemir

Data Scientist, Entrepreneur,
Author, Lecturer

Pearson

ui.endpoints . huggingface.co

Deploy virtually any model on HF, even ones you fine-tuned!

Deployed Endpoints + New Browse Catalog

Filter by name, provider, task... Status: All Need help ?

aws-llama-2-7b-hf-3756 Scaled to Zero
text-generation • protected • aws • 1x Nvidia A10G
https://jtjuckgwnk2jqym8.us-east-1.aws.endpoints.huggingface.cloud

aws-llama-2-7b-chat-hf-6139 Scaled to Zero
text-generation • protected • aws • 1x Nvidia A10G
https://sa0b44ky03zvbtds.us-east-1.aws.endpoints.huggingface.cloud

distilbert-toxic-classifier Scaled to Zero
text-classification • protected • aws • Intel Ice Lake
https://d2q5h5r3a1pkorfp.us-east-1.aws.endpoints.huggingface.cloud

aws-t5-aligned-summaries-6363 Paused
text2text-generation • protected • aws • Intel Ice Lake
Endpoint has been paused

aws-flan-t5-small-1837 Paused
text2text-generation • protected • aws • Intel Ice Lake
Endpoint has been paused

aws-all-mpnet-base-v2-3472 Paused
sentence-embeddings • protected • aws • Intel Ice Lake
Endpoint has been paused

ui.endpoints . huggingface.co

distilbert-toxic-classifier Running Pause

Overview | Analytics | Usage & Cost | Logs | Settings

Endpoint URL | Need help?

https://d2q5h5r3a1pkorfp.us-east-1.aws.endpoints.huggingface.cloud

Configuration | Protected

Task	Container Type
text-classification	Default

Created Jan 25 by profoz • Last Edited Feb 5 at 9:16 AM by proxy

Model | Up-to-date

profoz/distilbert-toxic-classifier

Revision db64ff8

Instance | $ 0.06 /h while running

AWS us-east-1 CPU · Intel Ice Lake · 1 vCPU · 2 GB

Scale-to-zero after 15 minutes without activity

Task + endpoint info

Compute info

Pearson

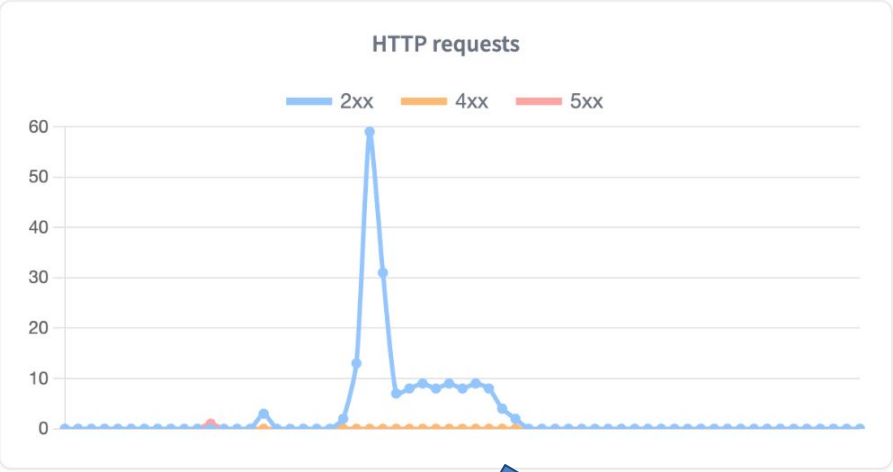
ui.endpoints . huggingface.co

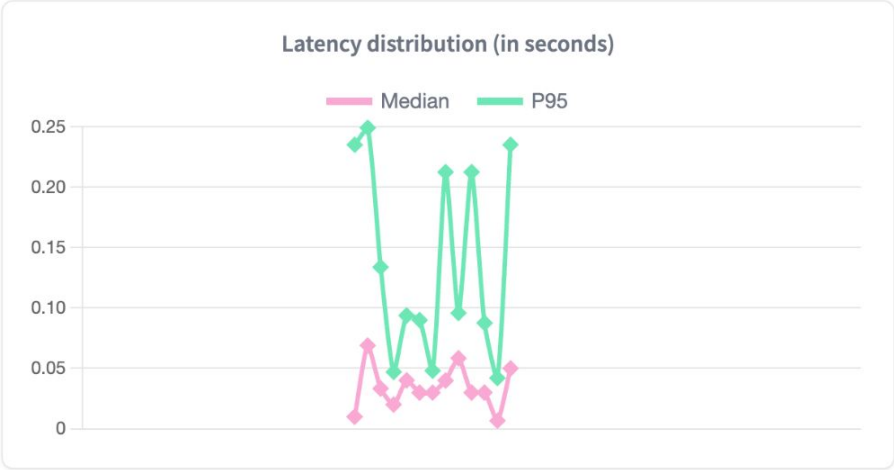
Replicas
1 / 1

Total requests
180 req

Median Latency
34.41 ms

P95 Latency
137.08 ms





Basic Metrics

Pearson

ui.endpoints . huggingface.co

API code to use the model

Playground TEST API

Token • personal

Parameters Doc

Top K

number

Function to Apply

Default

Test your endpoint

Text Classification

You're such a loser

Compute

Toxic 0.665

JSON Output Maximize


```
[
  {
    "label": "Toxic",
    "score": 0.6649302244186401
  }
]
```

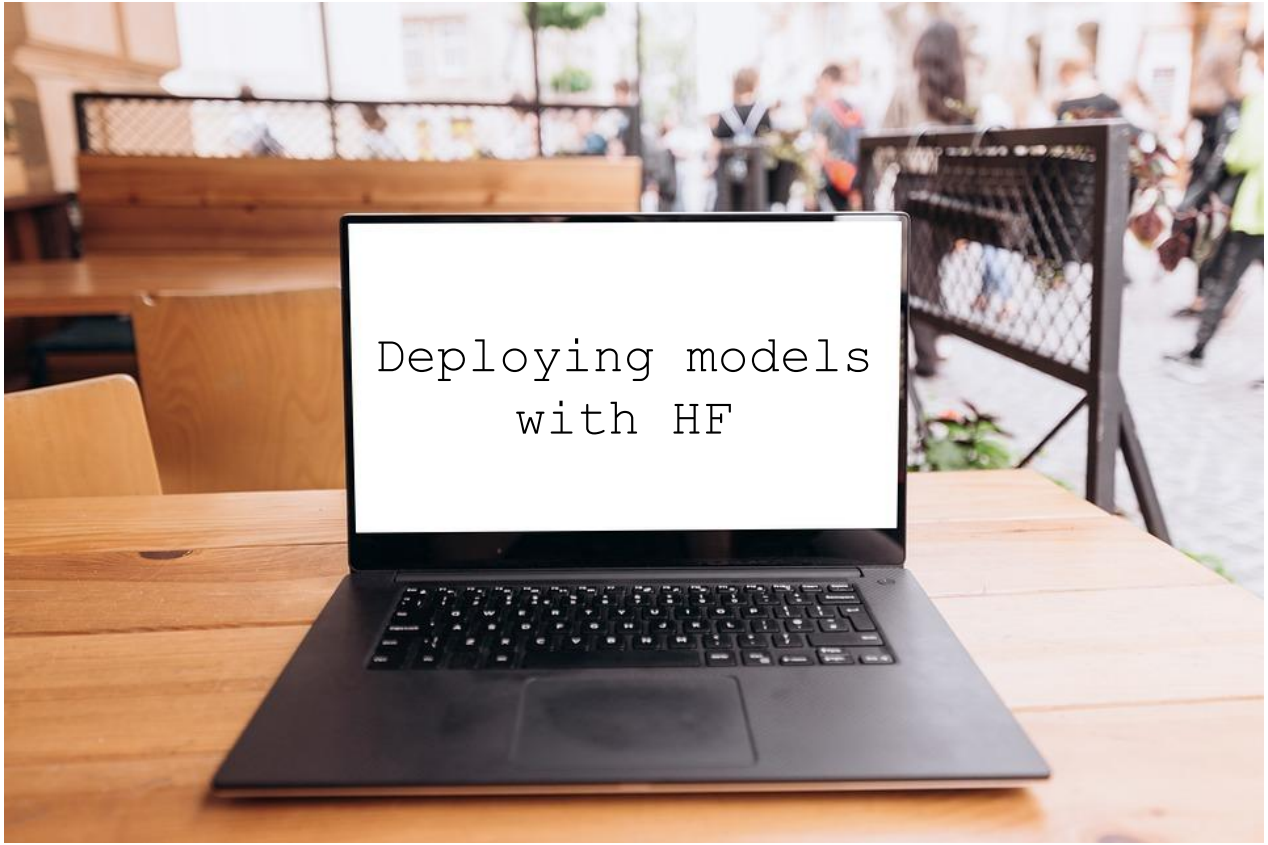

Simple playground for models

Pearson

Exercise: Deploying a fine-tuned model on the Inference API

Pearson

Code Time!



Pearson

github.com/huggingface/**chat-ui**

Chat UI

A chat interface using open source models, eg OpenAssistant or Llama. It is a SvelteKit app and it powers the HuggingChat app on hf.co/chat.

0. No Setup Deploy
1. Setup
2. Launch
3. Web Search
4. Text Embedding Models
5. Extra parameters
6. Deploying to a HF Space
7. Building

Hugging Face in 4 Hours

Segment 4: Multimodal AI and Community Insights

Sinan Ozdemir

Data Scientist, Entrepreneur,
Author, Lecturer

Pearson

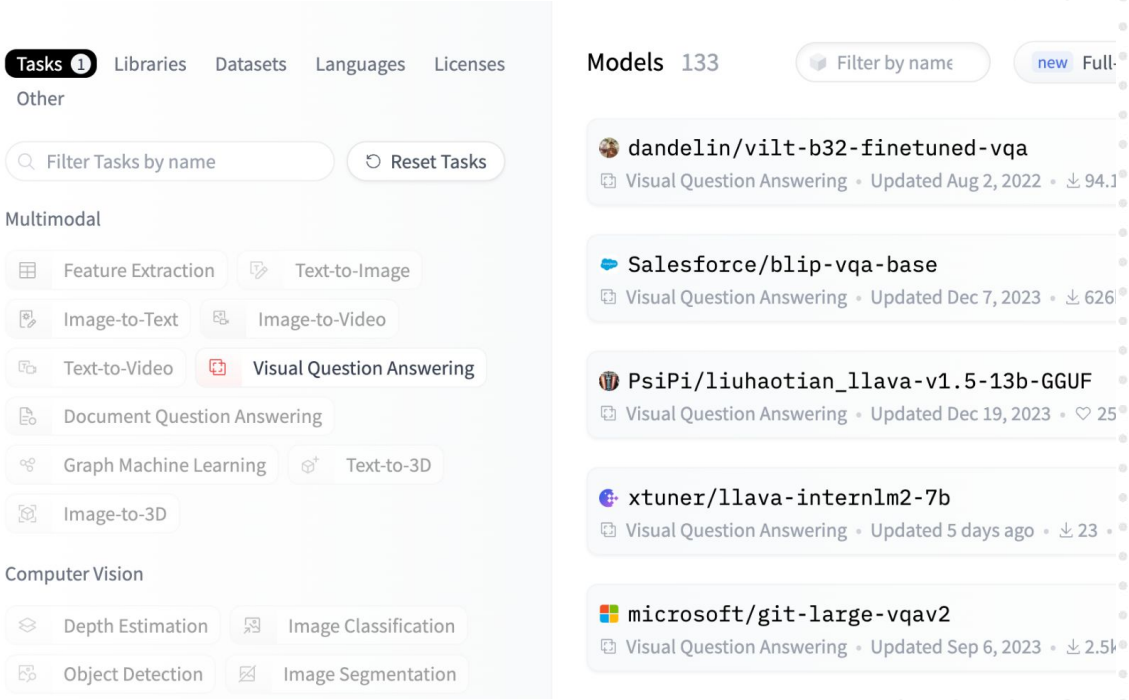
Engaging with multimodality in AI: text, image, and audio processing

Pearson

Multimodal Models

Multimodal models employ a *mix* of data modalities

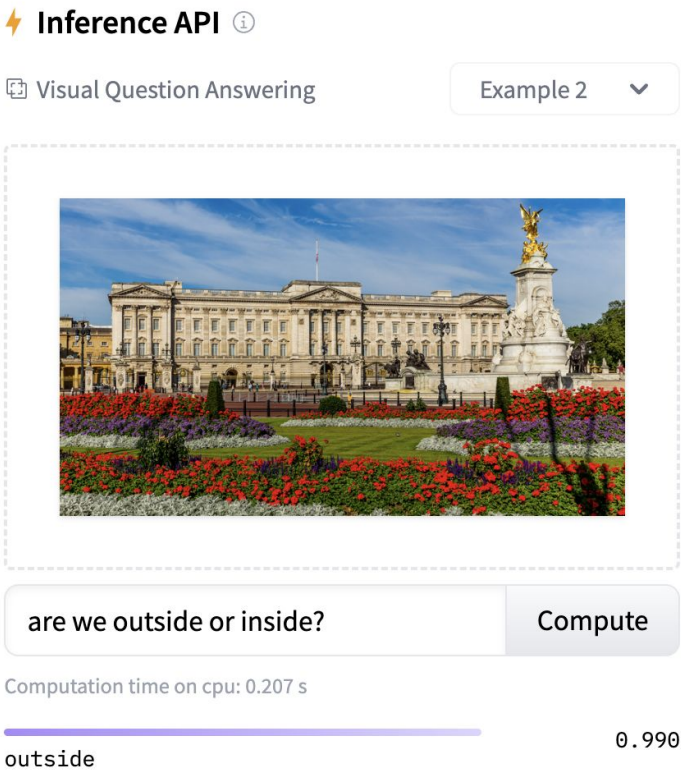
Computer Vision models are strictly working with images



Multimodal Models

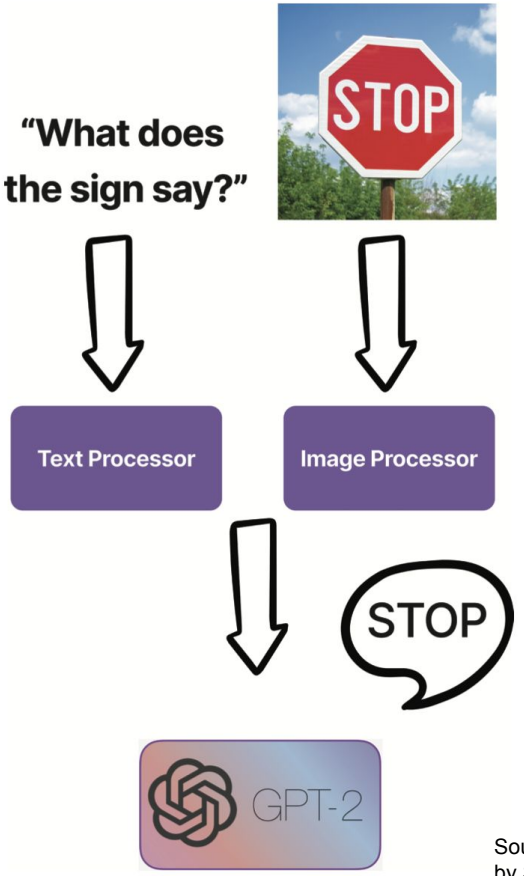
A common multimodal task is **Visual Question/Answer**

Given an image and a question, answer the question (usually only with a word or two)



Multimodal Models

You can built multimodal architectures using open-source components from HuggingFace



Source: Quick Start Guide to LLMs by Sinan Ozdemir

Code Time!

Pearson

Leveraging the community for project collaboration and advancement

Pearson

Considering Open-source

Collaboration

Within org:

Setting up channels of communication between Data teams <> Product teams <> Marketing e.g. for **faster iteration** on meaningful features

Outside of org:

Sharing open-source models/tools is a chance to build **community** - a low-cost marketing strategy

Privacy / Security

Nothing new here, no need to send data to a 3rd party provider like **OpenAI** who have already shown a record of **data leaks** in their (relatively) short time in the limelight.

Ownership

Ownership of models and data provides an opportunity for organizations to get **more hands-on** with their ML use-cases by labeling data and **collecting feedback** from users.

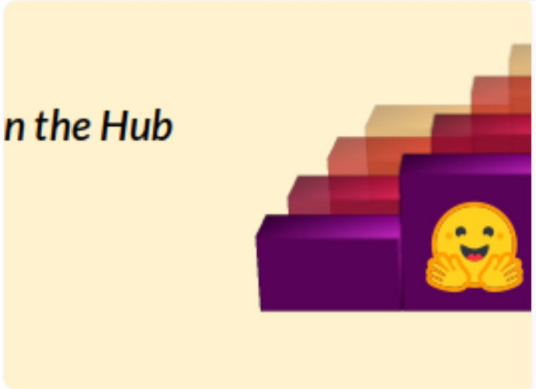
Huggingface.co / **blog**

Posts, articles, and discussions

Everything | Community | Guide | Open Source Collab | Partnerships

Research | NLP | Audio | CV | RL | Ethics | Diffusion | Game Development

Time Series | RLHF | Case Studies



The Hallucinations Leaderboard, an Open Effort to Measure Hallucinations in Large Language Models

By pminervini
January 29, 2024 guest

Community blog posts view all

Building autograd engine tinytorch 03
By joey00072 • about 9 hou...

Building autograd engine tinytorch 02
By joey00072 • about 9 hou...

Fine Tuning a LLM Using Kubernetes with Intel® Xeon® Scalable Processors
By dmsuehir • 7 days ago

Create a Web Interface for your LLM in Python
By Alex1337 • 7 days ago

makeMoE: Implement a Sparse Mixture of Experts Language Model from Scratch

Sinan Ozdemir’s Framework for prototyping with LLMs with a mind for production

Pearson

Sinan’s LLM Framework

1. Define Inputs and Outputs

- Identify and document the specific inputs and outputs for your LLM application.
- Example: Given a user's taste and a list of book descriptions, the model should output a ranked list of book recommendations with reasons.
- Remember, requirements might change during testing or in different contexts.

2. Define Success/Failure States

- Clearly define what constitutes a success or a failure for your model.
- Example of success: The model should return at least 3 recommendations that match the given book list with a rationale for each.
- Example of failure: The model doesn't provide 3 recommendations, or the suggestions aren't from the given list.
- Failures are binary and don't reflect the quality of output, instead indicating whether the model meets the basic requirements.

Sinan's LLM Framework

3. Consider Potential Bias

- Examine if the model's outputs can be influenced by subjective bias or unnecessary information.
- Example: The model might utilize past knowledge or context about the books, leading to bias. Ensure it's "staying on script" and relying on the input given.

4. Create Comprehensive Examples (to be used as few-shot later)

- Develop at least two detailed examples for training (few-shot) or testing.
- Example: real list of wines from a dataset, etc
- This step helps to classify the model's knowledge requirement (Class A, B, or C).

Sinan’s LLM Framework

5. Determine the Model's Knowledge Requirement

- Assess if the model has the necessary information to perform the task.
  - Class A: The model has all the required information encoded.
  - Class B: The model mostly has the necessary information but lacks specific details or updated data.
  - Class C: The model lacks the majority of required knowledge and needs extensive training.

6. Write an MVP (Minimum Viable Product) Prompt

- Create various versions of a prompt and experiment with them in the model's playground. This helps to refine the prompts and assess the model's knowledge requirement.

7. Iterate on Prompt Techniques and Parameters

- Adjust the parameters like temperature and top-p to refine the model's responses.

Sinan's LLM Framework

8. Evaluate and Plan for Scale/Production/Cost/Testing

- Assess the performance of the model, including its computational demands, and plan for potential scaling and production deployment.
- Also, consider the cost of deployment, which includes financial costs (like cloud resources and potential fine-tuning) and resource costs (like time and personnel for testing and maintenance).

9. Prototyping and Iteration

- Create a basic version of the model using tools like Streamlit for quick testing and user feedback.
- Iterate on the model by refining the prompts, adjusting parameters, and fine-tuning the model based on feedback.

Sinan's LLM Framework

10. Labeling Data and Fine-tuning

- Plan for potential data labeling and fine-tuning. This includes considering the cost and time required for these steps.
- Remember, fine-tuning not only requires labeled data but also extensive computational resources, which can add to the overall cost.

11. Evaluation

- Consistently evaluate the model's performance using relevant metrics like semantic similarity, precision, recall, etc. These evaluations will guide the iterations and improvements.

The above framework is not exhaustive but provides a good starting point for designing applications with LLMs like ChatGPT. Each application will have unique needs and constraints, so this framework should be adapted accordingly.

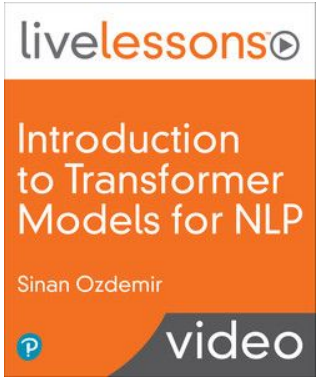
Summary + Next Steps

- The invention of the Transformer in 2017 revitalized of the field of NLP and an explosion of Large Language Models
- There are many types of LLMs with pros/cons and knowing which to use and how to use it makes all the difference
- LLMs are not perfect and **will** eventually produce untrue and harmful statements if left unchecked
- Reinforcement Learning can further align LLMs
- Attention seems to be (mostly) all we need.. for now

Summary + Next Steps

- Libraries like Streamlit help fastrack prototypes and give you the ability to share them for free on Hugging Face
- Knowing which metrics are best for evaluation can make all the difference
- Building prototypes off of a framework and using future-proof techniques like few-shot prompting and chain-of-thought reasoning help us build faster and with more confidence

Summary + Next Steps

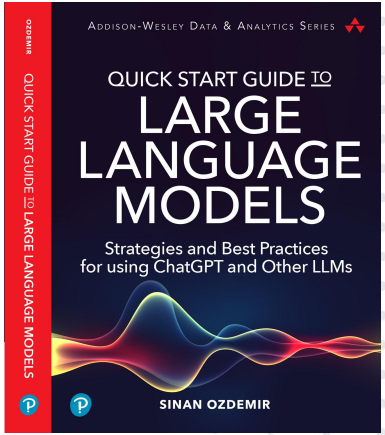


A comprehensive introduction to LLMs + Transformers

https://learning.oreilly.com/videos/introduction-to-transformer/9780137923717

Check out my live trainings for more in depth content!

https://learning.oreilly.com/search/?q=Sinan%20Ozdemir&type=live-event-series

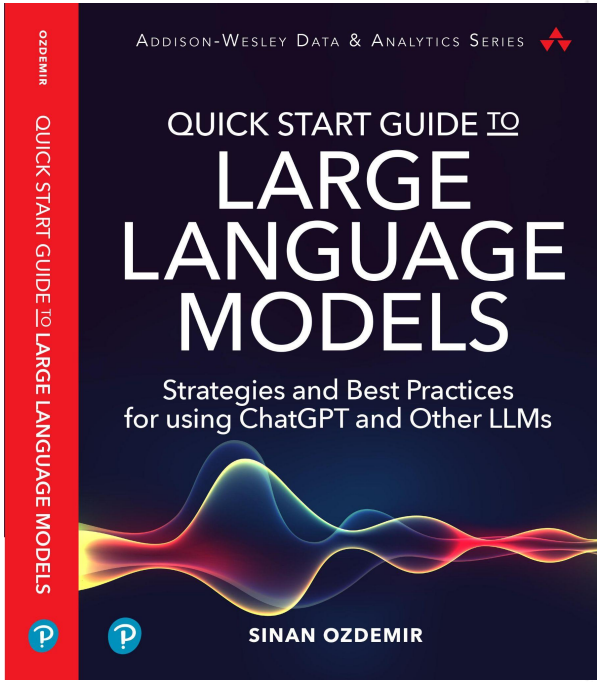


Pearson

Thank you! / Final Q/A

Most of these examples were based off of my new book on LLMs, usually top 10 in many categories on Amazon including NLP

https://a.co/d/fZsOWxd



Pearson

Hugging Face in 4 Hours

Thank you!

Sinan Ozdemir

Data Scientist, Entrepreneur,
Author, Lecturer

Pearson