

LLMs: A Crash Course in Powering up Applications

Instructor: Andrew Suter-Morris

Introduction – Andrew Suter-Morris

CTO, Forming AI & PreNav - GenAI and Diffusion Based Modeling. Synthetic Imagery

Senior Tech Lead, Ex-Microsoft - ~10 years on HoloLens and Synthetic Data

BS/MS in CompSci + Mathematics - Colorado School of Mines



Today's Agenda

	10:00 - 10:20 MT
h:00 - h:05	Brief introduction to LLMs and Use Cases
h:05 - h:10	<ul style="list-style-type: none">• Tools & Setup• Introduction to Gradio• HuggingFace - Datasets and Models
h:10-h:15	<ul style="list-style-type: none">• Text Analysis• Text Analysis w/ Hugging Face• Chatbot w/ ChatGPT
h:15 - h:19	LangChain & LangSmith
h:19-h:20	Summary and Questions

LLMs

Home > AI & Machine Learning

Nvidia's new coding LLM will make you a better programmer and can run on a CPU

Nvidia, in collaboration with HuggingFace and ServiceNow, has released StarCoder2, and it will help you generate code.

BY ADAM CONWAY PUBLISHED MAR 1, 2024

MANY THINGS FREQUENTLY —

Words are flowing out like endless rain: Recapping a busy week of LLM news

Gemini 1.5 Pro launch, new version of GPT-4 Turbo, new Mistral model, and more.

BENJ EDWARDS - 4/12/2024, 2:31 PM

ChatGPT has entered the classroom: how LLMs could transform education

Researchers, educators and companies are experimenting with ways to turn flawed but famous large language models into trustworthy, accurate 'thought partners' learning.

By [Andy Exance](#)

Multi-AI collaboration helps reasoning and factual accuracy in large language models

Researchers use multiple AI models to collaborate, debate and improve their reasoning abilities to advance the performance of LLMs while increasing accountability and factual accuracy.

Rachel Gordon | MIT CSAIL

September 18, 2023

Use Cases?

- What use cases can you think of for using LLMs?

Use Cases

- Text Summarization
- Chatbot and Agents
- Code Analysis and CodeGen
- Synthetic and/or Structured Data
- Translation
- Content Creation

Goal

- View different use case implementations and frameworks
- You'll get a view of the possibilities, and how that may apply in your own line of work or interviews

Tools & Setup

- I prefer Debian based systems, but Win, *Nix and MacOS should all work
- Things you'll need
 - [VSCode](#) (or editor of your choice)
 - [Anaconda](#) (or venv of your choice)
 - I use Python 3.10 for my environment
 - [LangChain](#), [OpenAPI](#) and [Apify](#) API keys

- Setup

```
git clone https://github.com/asutermo/llm-lesson
```

```
cd <src>
```

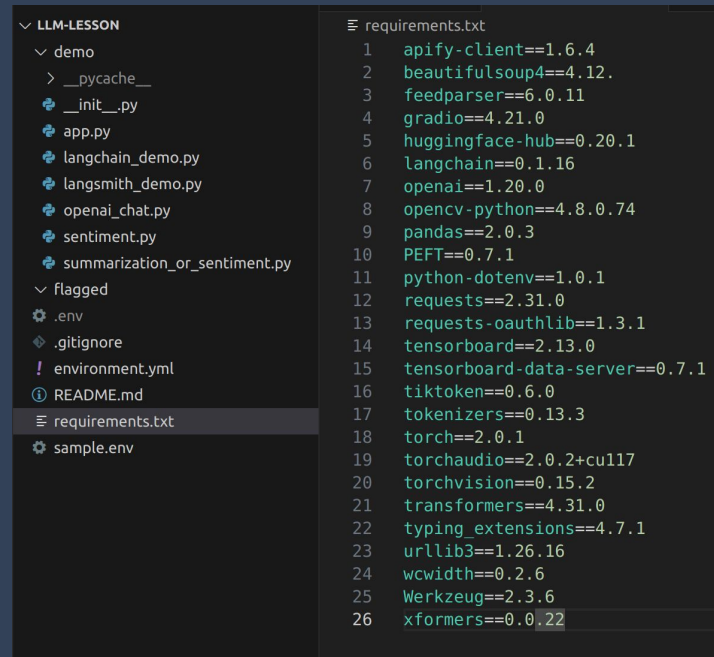
```
touch .env
```

```
conda install -n base conda-libmamba-solver
```

```
conda config --set solver libmamba
```

```
conda env create -f environment.yml
```

```
conda activate llm-lesson
```



Questions

- Do you have any familiarity with Gradio or Streamlit? Which do you prefer and why?

Gradio

- An easy way to get started with low barrier to entry

```
import gradio as gr
```

```
def greet(name):
```

```
    return "Hello " + name + "!"
```

```
demo = gr.Interface(fn=greet, inputs="text", outputs="text")
```

```
demo.launch()
```

name

Andrew

Clear

Submit

output

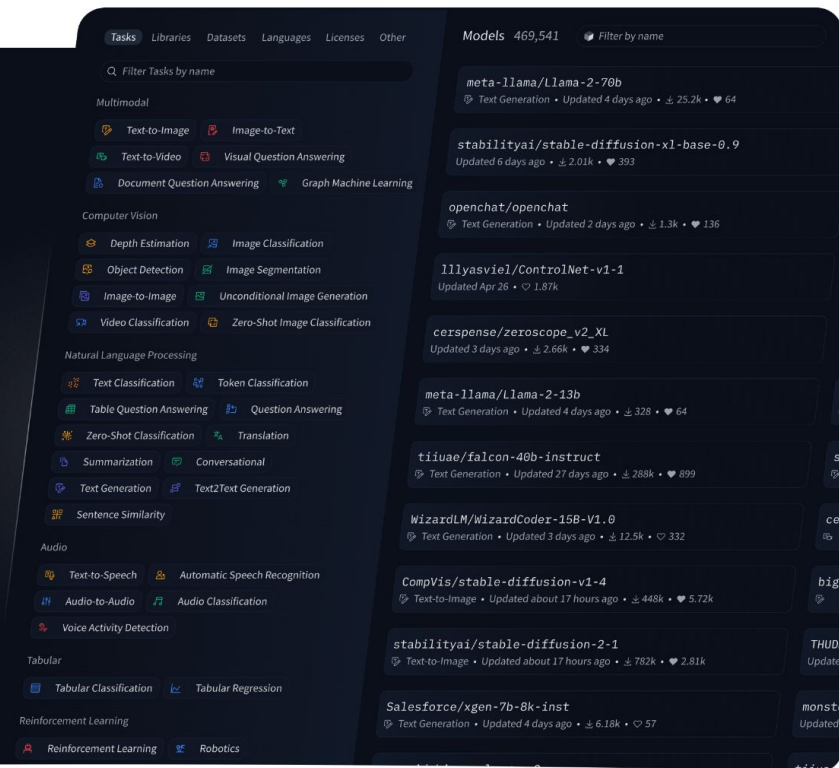
Hello Andrew!

HuggingFace - Datasets and Models



The AI community building the future.

The platform where the machine learning community collaborates on models, datasets, and applications.



Text Sentiment using HuggingFace

Demo

Simple Sentiment

Simple Summarization

Hugging Face Pipelines

OpenAPI Demo

LangChain Demo

LangSmith Demo

RobertaForSequenceClassification

Input

Hi, I'm feeling great

Clear

Submit

Classification

POSITIVE

POSITIVE

100%

Flag

Text Summarization And Sentiment using HuggingFace

article_title

List of articles to summarize or get sentiment of!

Louisiana's flagship university lets oil firms influence research – for a price

summarize_or_sentiment

Summarize or sentiment analysis of article

☒ summarize ☐ sentiment

Clear

Submit

output

Louisiana State University allowed Shell to influence studies after a \$25m donation and sought funds from other fossil fuel firms. Director of LSU's Institute for Energy Innovation said being able to work with oil and gas companies is 'really a key to advancing energy innovation'

Flag

Examples

article_title	summarize_or_sentiment
\$18k in stolen antlers: poaching on the rise in Wyoming as collectors 'cheat the system'	summarize
'I want to go home': passengers stranded by Dubai extreme floods – video	sentiment

Text Analysis

- What are some real world use-cases for this? Why might you consider using this?

Chat with OpenAI

Gradio and ChatGPT 3.5 Turbo

ChatGPT Personality

You are an instructional, informative, and kind AI assistant.

ChatGPT 3.5 Turbo

Can you write me a hello world app in python?

Clear

Submit

Reply

Sure! Here's a simple "Hello World" program in Python:

```
```\n# Hello world program in Python\nprint("Hello, World!")\n```
```

You can copy and paste this code into a Python IDE or terminal to run the program and see the output "Hello, World!" displayed.

Chatbot

I am an artificial intelligence designed to assist you with any questions or information you may need. I am here to provide assistance, guidance, and support on a wide range of topics. Feel free to ask me anything, and I will do my best to help you.

What do you think of tools like GitHub CoPilot?

Tools like GitHub CoPilot can be very helpful for developers by providing code suggestions, improv

Textbox

|

Clear



# GPT and The Rise of AI Assistants

- ChatGPT
- HuggingChat
- Microsoft's series of CoPilots
- Google Gemini (Bard)

# LangChain

Careers

## Build

LangChain gives developers a framework to construct LLM-powered apps easily.

# LangChain and Apify

## Gradio and LangChain Knowledge Base Query Answering

LangChain KB Query

How do I do synthetic data creation?

Clear

Submit

LangChain KB Reply

To create synthetic data, you can use a synthetic data generator tool or library. This tool will allow you to specify the structure or schema of your data, and then generate artificial data that follows that structure. You can also provide real-world examples as a "seed" to guide the generator in creating data that looks similar to your desired data. It's important to use synthetic data carefully, as it may not always capture real-world complexities.

Flag

# LangChain

- What do you think that Apify might be useful for?
- Brief Aside: Why use Synthetic Data?

# LangSmith

## Observe

LangSmith gives visibility into what's happening with your LLM-powered app, whether it's built with LangChain or not, so you know how to take action and improve quality.

# LangSmith

## Gradio and ChatGPT 3.5 Turbo with Langsmith Tracing

ChatGPT Trace

Write an obfuscated hello world app in python.

Clear

Submit

Reply

```
'''
exec('print(''.join([chr(ord(c)+1) for c in "gdkkn$twfnf"])[::-1])')
'''
```

Flag

# LangSmith

Why might tracing be useful?

# LangSmith Analysis

Personal > Projects > default

default

Project IDAdd RuleEdit

RunsThreadsMonitorSetup

1 filterLast 7 daysRoot RunsLLM CallsAll Runs

Columns

		✓	Name	Input	Output	Start Time	Latency	Dataset	Annotation Queue	Tokens	Cost
<input type="checkbox"/>	>	✓	RetrievalQA	How do I do synthetic ...	To create syntheti...	4/20/2024, 11:14:58 ...	🕒 2.33s	<input type="checkbox"/> 📄	<input type="checkbox"/> ✎	968	\$0.0
<input type="checkbox"/>	>	✓	trace_oai_pipeline	Write an obfuscated h...	"" exec("").join([chr(...	4/20/2024, 10:11:08 ...	🕒 1.20s	<input type="checkbox"/> 📄	<input type="checkbox"/> ✎	49	\$0.0
<input type="checkbox"/>	>	✓	trace_oai_pipeline	Write an obfuscated h...	""python exec("").joi...	4/18/2024, 8:30:53 ...	🕒 1.37s	<input type="checkbox"/> 📄	<input type="checkbox"/> ✎	47	\$0.0
<input type="checkbox"/>	>	✓	trace_oai_pipeline	wooo, what's my trace ...	I'm sorry, I'm not s...	4/17/2024, 10:52:38 ...	🕒 1.99s	<input type="checkbox"/> 📄	<input type="checkbox"/> ✎	42	\$0.0
<input type="checkbox"/>	>	✓	trace_oai_pipeline	Hello, how are you?	Hello! I'm just a vir...	4/17/2024, 10:46:32 ...	🕒 2.21s	<input type="checkbox"/> 📄	<input type="checkbox"/> ✎	44	\$0.0
<input type="checkbox"/>	>	✓	Sample Agent Trace	What is a document lo...	BEEP BOOP! A do...	4/17/2024, 4:29:20 ...	🕒 7.31s	<input type="checkbox"/> 📄	<input type="checkbox"/> ✎	540	

**Stats**  
Last 7 days

RUN COUNT  
6

TOTAL TOKENS  
1,690 / \$0.00 ⓘ

MEDIAN TOKENS  
48

ERROR RATE  
0%

% STREAMING  
0%

LATENCY  
P50: 2.10s P99: 7.06s

**Filter Shortcuts**

**Metadata**  
☐ custom-metadata-key



# Summary

Interfaces: Gradio

Model Hubs: HuggingFace and OpenAI

LLM Dev & Analysis Frameworks: LangChain and LangSmith