

✓ Hey human learner: Welcome to the Level 1 of Modern AI Pro!

The best way to predict the future is to invent it.

It is going to be an amazing journey. From Team Mitra, we cannot be more excited for you guys to join us in exploring this amazing field. Here is all the key resources for this class.

All the notebooks are linked from here. Additional code to run in your own environment is found here: <https://github.com/balajivis/modernaipro>

LLM Arena: <https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard>



✓ Recordings of the previous class

Day 1:

1. <https://tldv.io/app/meetings/6662db2b3c3c1900130f59d5/>
2. <https://tldv.io/app/meetings/6662fab5132e2000131c3761/>

Day 2:

1. <https://tldv.io/app/meetings/6663d830777b030013dce9e4/>
2. <https://tldv.io/app/meetings/6663f3216397210013a7546e/>
3. <https://tldv.io/app/meetings/6664162f777b030013dcf8b5/>
4. <https://tldv.io/app/meetings/666435565255810013ef7c7a/>

Day 3:

1. <https://tldv.io/app/meetings/666e64fd22201b0013cd7ea6/>
2. <https://tldv.io/app/meetings/666e8f94412bc2001309fb57/>
3. <https://tldv.io/app/meetings/666ea3129148da001269278e/>
4. <https://tldv.io/app/meetings/66652e69096c340013153bdb/>
5. <https://tldv.io/app/meetings/666ec2a47c2ee200131b4e0e/>

Relevant links from our earlier classes:

1. Finetuning: <https://tldv.io/app/meetings/66471dff1038f30012129f4e/>
2. GANs and research topics: <https://tldv.io/app/meetings/664720644a57b300137b21bd/>

```
from google.colab import files
from IPython.display import Image
uploaded = files.upload()
```



Choose Files

No file chosen

Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

1. Langchain -- the dominant framework for running LLM based applications
2. Ollama -- interface for running open source LLMs locally
3. Together.ai -- training and finetuning models on the cloud.
4. Llama 3 -- best in class open source model
5. Llava -- best in class open source multimodal model
6. Multimodal model -- use a combination of image, text and other types of input, while normal LLMs can take only text
7. Qwen -- a lightweight model you can run locally fast.
8. Stream -- Enables you to run LLM output with the output coming token by token giving a much better responsive output to users

✓ Day 0: Introduction and Setup

Resources: [Slides](#) and [Notebook](#)

✓ Setup

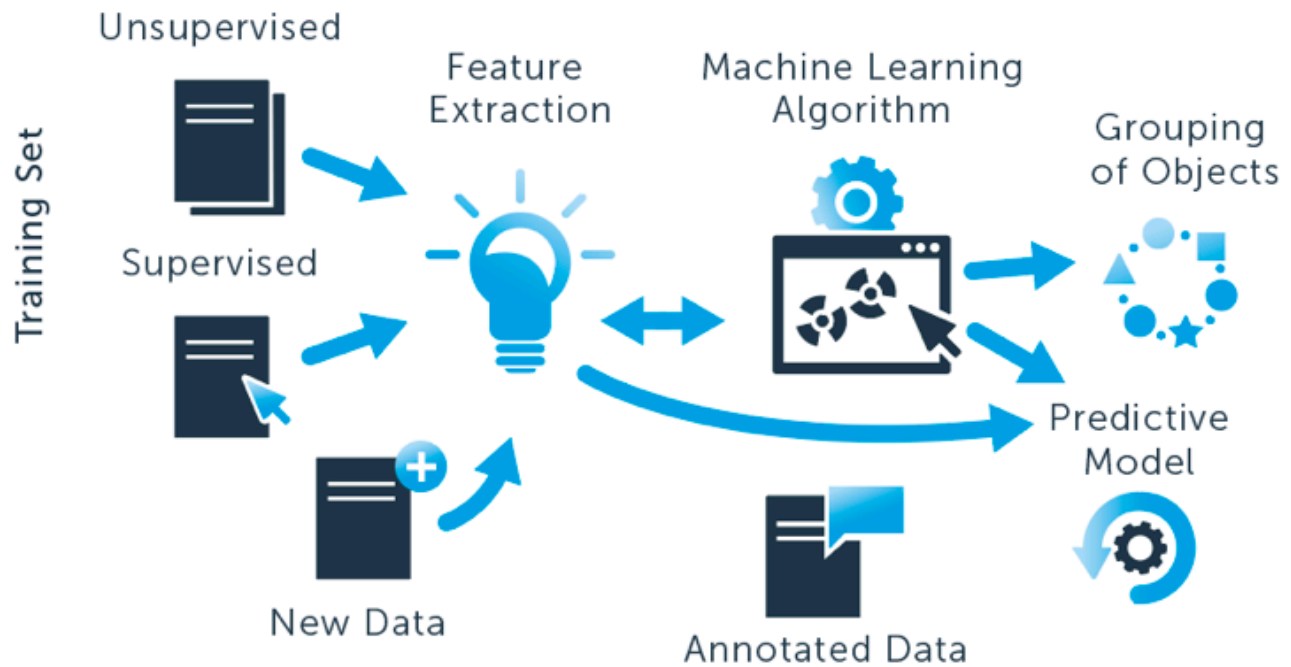
1. Get familiar with [Google Colab](#). This will be the primary work environment for running all the code exercises.
2. Install **VSCode, Docker, Ollama**
3. Sign up & get API keys/token at **Kaggle, Anthropic, HuggingFace, Langfuse, TogetherAI**.
Kaggle gives you datasets, HuggingFace the open source models, Anthropic the top notch LLM (with free credits), Together to deploy & finetune and Langfuse to observe.
4. Learn to keep all these as Colab secrets and as local .env file
5. Sign up for a free account on discord.com and join our Discord Server (link will be provided to you separately). This is where the community collaborates and we have AI bots to assist with questions and quizzes.

Practice

1. Run a simple Python code with Pandas
 2. Learn to keep windows side by side to see what we are doing
 3. Get very familiar with installing Python packages "pip install"
 4. Highly recommended: [The Missing Semester of Your CS Education](#)
-

✓ Day 1: Foundations of Machine Learning

Machine Learning



Session 1: 0 to Hero in LLM

[Slides](#) and [Code](#)

Duration: 2 hours We will start with Ollama and 3 state of the art LLMs to run language models locally.

1. Introduction to Ollama -- a key wrapper to run LLMs on-prem
2. Discovering and downloading models locally
3. Running the best in class Qwen2 a smaller model Gemma that provides best in class output for its size.
4. Practice similarity search --
<https://colab.research.google.com/drive/1cQr0ssoXcYxhJ5MvhtoYwmG67g3l0P5W?usp=sharing>
5. Calling local LLMs from Python and streaming its output to show the tokens generated one by one.
6. Introduction to Langchain -- a top framework for LLM and Cody -- code generation.
7. Introduction to Multimodal LLM with Llava3 and using its output locally.
8. Deploy the model to public with Gradio with a Multimodal model.

You need to install Python, Pip for this to work.

10-min Break 🕒

Session 2: Theory - Intro to ML

Duration: 2 hours

Resources: [Slides](#) and [Notebook -- simple Linear Regression with a few columns and rows.](#)

Outcome: Lay the foundation for the bootcamp with a quick intro to Machine learning & supervised learning, and implement linear regression using a simple example.

Objective: Explain how ML is different from traditional programming, understand types of ML, and implement a regression model.

Key Activities:

Theory for the following topics:

1. Linear Regression
 2. Parametric Models
 3. Cost Function
 4. Iterative learning
 5. Inference
-

✓ More reading through external References

0. [Intro to Gradio](#)
1. [Deploying Gradio in Nginx on the server](#)
2. [Beginner's guide to Ollama](#)
3. [Intro to Langserve](#)
4. [Deploying Langserve on Google Cloud](#)
5. [When to use One Hot Encoding](#)
6. [Top 6 Classification algorithms](#)
7. [Biases in Machine Learning](#)

Additional Notebooks

These notebooks extend the learning you had from this day.

1. [Model math basic spreadsheet](#)
2. [Charting in Colab](#)
3. [Visualization deepdive with Matplotlib and Seaborn](#)

4. [Hands on Linear Regression in 2 variables](#)
 5. [Linear Regression from scratch-optional NB](#)
 6. [Optimization and Learning rate.](#)
 7. [Decision trees and Random forests theory.](#)
 8. [ML Metrics and Performance](#)
 9. [Pandas Deepdive](#)
 10. [Numpy Deepdive](#)
-

✓ Day 2: Foundations of NLP and Deep Learning



Session 3: Train and Deploy Machine Learning algorithms.

Duration: 3.5 hour

Resources: [Slides](#), [Income Class Notebook](#) and [Home Prices notebook](#)

[Advanced Home prices prediction with 100% accuracy](#)

[Income class prediction with Neural Networks](#)

[Fault Prediction in Air Pressure System](#)

Objective To provide hands-on experience with fundamental machine learning (ML) concepts and techniques, focusing on data preprocessing, visualization, model training, and ethics in AI.

Outcomes Participants will learn how to:

- Encode categorical data for ML models.
- Perform train-test splits to evaluate model performance.
- Utilize Seaborn and Matplotlib for data visualization.
- Apply normalization techniques to prepare data for modeling.
- Draw histograms to understand data distribution.
- Understand and discuss the ethical considerations in AI and ML projects.
- Predict income class and home prices using ML models.

Key Activities

1. **Predicting an Income Class:** Participants will use a Colab notebook to predict the income class of individuals based on demographic data, employing various preprocessing and modeling techniques.
2. **Encoding Data:** Learn how to convert categorical data into a format that can be provided to ML models for better predictions.
3. **Train Test Split:** Understand the importance of splitting data into training and testing sets for model evaluation.
4. **Data Visualization:** Gain proficiency in using Seaborn and Matplotlib for creating a variety of visualizations to explore and present data effectively.
5. **Normalization:** Implement normalization techniques to scale numerical data, improving model training and convergence.
6. **Drawing Histograms:** Create histograms to analyze the distribution of data and identify patterns or outliers.
7. **Ethics in AI:** Engage in discussions about the ethical implications of AI and ML, including bias, fairness, and accountability.
8. **Predicting Home Prices with Linear Regression:** Use a Colab notebook to predict home prices based on features like size and location, applying linear regression to real estate data.
9. **Deployment in Gradio**
10. **Mlflow** - Experimentation and tracking results

10-min Break 

Session 4: Practise LLM locally

Duration: 0.5 hours

Objectives: We will build on top of Day 1 to get to superior performance in enterprise settings.

Key activities:

1. Running with the state of the art low latency hardware optimization -- Groq
2. Routing having multiple LLMs
3. Tracking LLMs with observability platform -- Langfuse
4. Evaluation of LLMs
5. Building LLM flows with Flowise

Lunch Break 🍱

12 - 12:30 pm

Session 5: Logistic regression and Model metrics

Duration: 30 min

Resources: [Slides](#)

Outcome: Continue theory from linear regression and expand into classification, decision boundaries, non-linearities.

Objective: Explain how classification is different from regression, understand the role of sigmoid function, and implement the logistic regression model.

Key Activities:

Theory for the following topics:

1. Gradient descent
2. Logistic regression: [Logistic Regression using sklearn](#)
3. Classification
4. Model metrics: [Data Metrics hands-on](#)
5. Sklearn, Numpy, Pandas, Matplotlib

Session 6: Foundations of Neural Networks

Duration: 45 min

Resources: [Slides](#)

Outcome: Classify MNIST digits using a fully connected neural network and understand how the model can be trained using the keras library.

Objective: Intuitively explain the theory of neural networks and how complex functions can be learnt by a bunch of neurons, implement a fully connected network using Keras.

Key Activities Explain theory for the following topics:

1. Neural Network history
 2. Intuitive explanation for a Neuron
 3. Model Architectures
-

Session 7: Computer Vision

Duration: 1 hour

Resources: [Slides](#)

Outcome: Continuation of session 8

Objective: Solve the MNIST task using CNNs and explain how computer vision approaches these problems using an efficient design.

Key Activities: Explain the following:

1. The MNIST dataset - data structure and dataset format
2. MNIST digit classification task: [Introduction to Neural Nets](#)
3. TF Keras library syntax and features
4. Underfit and Overfit problems: [Regularization: Underfitting vs Overfitting.](#)

10-min Break 🕒

✓ Session 8: Intermediate LLM Practice

Duration: 2 hours

1. [Practising Prompt Engineering with Anthropic](#)

2. [Evaluating LLMs Notebook](#)
3. [Foundation of NLP](#)
4. [Similiarity search](#)

Image('/content/Screenshot 2024-05-19 at 9.00.12 PM.png', height=500)



12 Elements of Prompt engineering

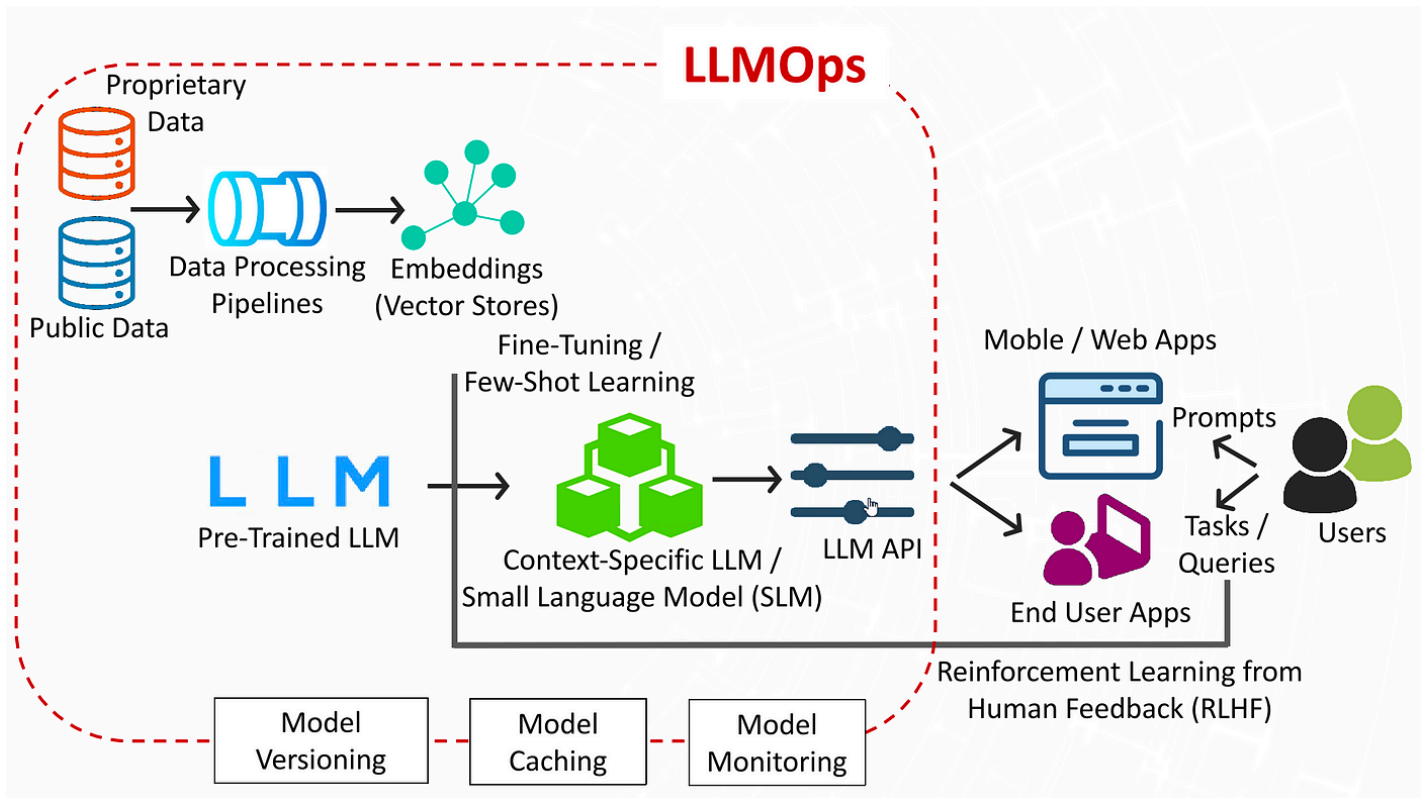
Elements of Prompt	Examples
Character	You are a top product management consultant.
Instruction	Generate a short summary of a product description.
Context	Our startup makes robots and AI tools. Here is ...
Constraints	At most 50 words.
Input Data Delimiter	<code><tag> Mitra Robot </tag></code>
Output Format	JSON
Tone and Style	Use a formal tone suitable for a business proposal.
Examples	Example: "The Mitra Robot is an advanced AI tool designed to..."
Variables and Placeholders	Use <code><product_name></code> as a placeholder for the product name.
Temperature and Sampling	Set temperature to 0.7 for balanced creativity.
Role-specific Instructions	As a financial advisor, provide an analysis of the quarterly...
Output Structure	Output should be in bullet points: Key feature 1, Key feature 2.

Additional notebooks:

1. [Intro to Pytorch](#)
2. [Document analysis and summarizing with traditional NLP](#)
3. [Deeper document analysis and summarizing with older NLP](#) tools
4. [Reference:Transformer Architecture](#)
5. [Qwen deployment](#)
6. [Video Classification with CLIP](#)
7. [CNN for large image classification - dog breeds](#)
8. [Identifying land use from satellite imagery](#)
9. [Face detection with OpenCV](#)

10. [Image Classification and Captioning with CLIP](#)
11. Hyper parameter turning: [Parameters tuning in Neural Nets](#)
12. Convolutions: [Intro to CNN with Keras](#)
13. Convolution and Pooling operations
14. Transfer learning - Introduction to Resnet
15. Data augmentation
16. Hardware dependencies - GPU

✓ Day 3: Advanced Applications and Deployment



✓ Section 9: Understanding Business side of AI

Duration: 45 min

Outcome: Learn Prompt Engineering and how business would use AI.

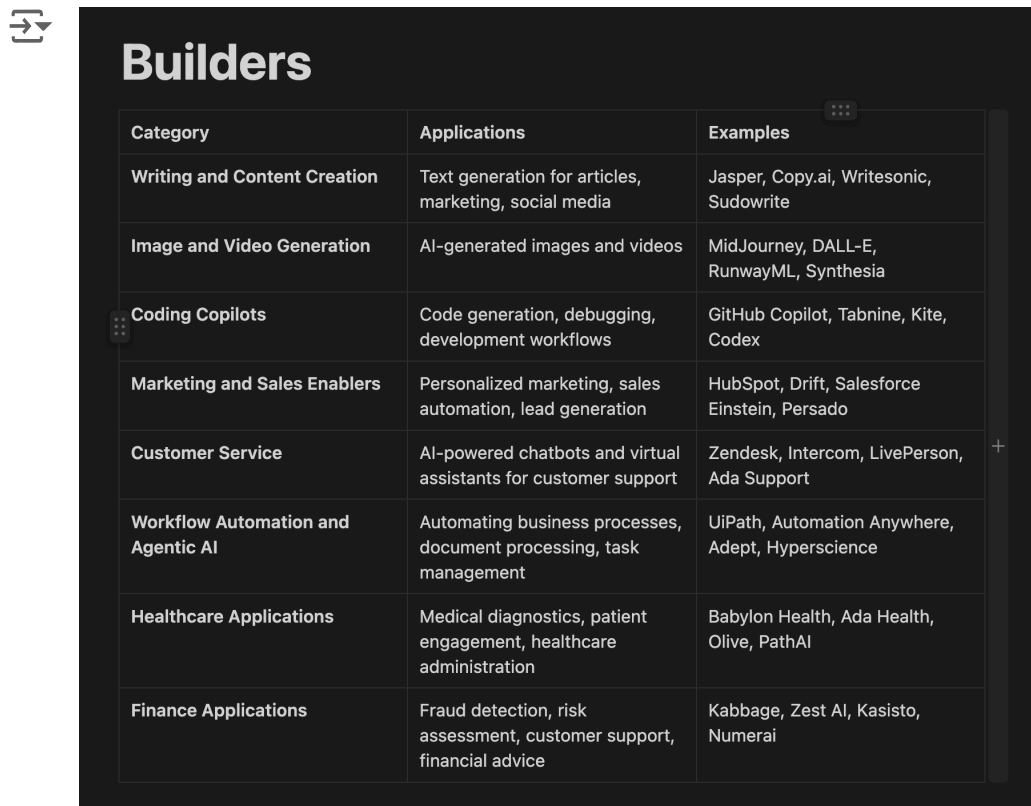
First we will look at the business impact of LLMs. Who are the key players in this?

Gen AI Impact Framework

The concentration of power goes down the levels.

Level	Description	Driver	Unique Aspect	Examples
Innovators	Companies like OpenAI and Meta that develop the foundational LLM technologies.	R&D driven	Pioneering Algorithms	OpenAI, Meta, Google, Anthropic, Mistral
Builders	Companies building specific applications on top of LLMs, such as Salesforce and other software developers.	UX and workflow driven	User Experience Design	Salesforce, Grammarly, Jasper
Enablers	Providers of cloud services, orchestration tools, training data, vector databases, and other essential infrastructure.	Integration and performance driven	Scalability Solutions	AWS, Azure, Databricks, Snowflake, Hugging Face
Customizers	Consultants and firms offering tailored LLM solutions and integrations, like Accenture and Deloitte.	Business value driven	Domain Expertise	Accenture, Deloitte, Capgemini
Utilizers	Enterprises that implement LLM applications internally to drive efficiency and enhance operations.	Optimization and productivity driven	Operational Efficiency	JPMorgan Chase, Walmart, Procter & Gamble
Guardians	Entities focused on education, ethical guidelines, security measures, and regulatory policies.	Safety driven	Ethical Oversight	Parity, Fiddler, Arthur, EthicsGrade, Reliance, Credo AI

Image('/content/Screenshot 2024-05-19 at 8.59.29 PM.png', height=400)



Category	Applications	Examples
Writing and Content Creation	Text generation for articles, marketing, social media	Jasper, Copy.ai, Writesonic, Sudowrite
Image and Video Generation	AI-generated images and videos	MidJourney, DALL-E, RunwayML, Synthesia
Coding Copilots	Code generation, debugging, development workflows	GitHub Copilot, Tabnine, Kite, Codex
Marketing and Sales Enablers	Personalized marketing, sales automation, lead generation	HubSpot, Drift, Salesforce Einstein, Persado
Customer Service	AI-powered chatbots and virtual assistants for customer support	Zendesk, Intercom, LivePerson, Ada Support
Workflow Automation and Agentic AI	Automating business processes, document processing, task management	UiPath, Automation Anywhere, Adept, Hyperscience
Healthcare Applications	Medical diagnostics, patient engagement, healthcare administration	Babylon Health, Ada Health, Olive, PathAI
Finance Applications	Fraud detection, risk assessment, customer support, financial advice	Kabbage, Zest AI, Kasisto, Numerai

Image('/content/Screenshot 2024-05-19 at 8.59.20 PM.png', height=400)



Enablers

Category	Subcategories
Cloud and Compute Infrastructure	Cloud Providers, Hyperscalers, Serverless Computing, Edge AI Infrastructure
Data Management and Pipelines	Data Extraction and Pipeline, Data Storage and Management, Feature Stores, Data and Synthetic Data Creators, Annotation and Labeling Tools
Model Development and Optimization	Building Frameworks, MLOps and Orchestration Tools, Model Optimization and Monitoring, Observability
Security and Compliance	AI Security and Compliance, Cost Management and Optimization
Integration and API Services	API and Integration Services, Networking and Connectivity
Analytics and Visualization	Visualization and BI Tools, AI Development Platforms, AIOps Platforms

Next, we will look at Prompt Engineering

Session 10: Solving Multiple Categories of NLP Problems with Best-in-Class Models from Hugging Face Pipeline

Duration: 1:30

Resources:

- **Notebook 1:** [Introducing Hugging Face Pipeline and NLP Problems](#)
- **Notebook 2:** [Summarizing a YouTube Video with Whisper API and LLM](#)
- **Slides:** [Session Slides](#)

Objectives: This session aims to showcase the versatility and power of Hugging Face's pipelines for solving a wide array of NLP problems. Participants will be introduced to ten different categories of NLP problems and learn how to apply state-of-the-art models for these tasks. Additionally, the session will cover practical applications such as summarizing audio content from YouTube videos using the Whisper API and language models.

Outcomes: Participants will gain insights into:

- The breadth of NLP problem categories that can be addressed with Hugging Face's pipeline.
- How to utilize pre-trained models for tasks such as text classification, sentiment analysis, question answering, and more.

- Practical skills in applying the Whisper API for audio processing and summarization tasks.

Key Activities:

1. **Exploring Hugging Face Pipeline:** Notebook 1 provides a comprehensive introduction to utilizing Hugging Face's pipeline for solving different NLP problems, showcasing the ease with which one can apply these powerful models across a variety of tasks.
2. **Introduction to NLP Problem Categories:** Discover ten categories of NLP problems, understanding the specific challenges and solutions associated with each.
3. **Summarizing YouTube Content:** Notebook 2 demonstrates a practical application of NLP and audio processing technologies by summarizing the content of a YouTube video, combining the Whisper API for speech-to-text conversion with a language model for summarization.

10-min Break 🕒

Section 11: Introduction to RAG

Duration: 2 hours

Outcome: Generate insightful responses to queries integrating LLMs, Dataframes, Embedding models, and Vector DB.

Key Resources From Github repo we will be using Python files 12-15.

Session 12: Advanced RAG and Fine Tuning

Duration: 30 min

Resources:

- **Advanced RAG for Finance Domain:** [Access Here](#)
- **Finetuning with Together:** [Access Here](#)

Objectives: Focusing on Advanced RAG Techniques

Objectives: This session explores advanced applications of Retriever Augmented Generation (RAG) integrated with vector databases. Participants will learn to build and enhance chatbots using enterprise content, apply advanced RAG techniques in the finance domain, and utilize query expansion and visualization techniques for improved information retrieval and user interaction.

Key Activities:

1. Build a simple chatbot capable of scraping web content and PDFs to answer user queries more effectively, showcasing the practical use of RAG in organizing and accessing enterprise

knowledge.

2. Implement an advanced RAG model that serves as a robo advisor for the finance domain, demonstrating the model's ability to handle complex queries and provide insightful responses.
3. Explore augmented RAG techniques with query expansion to enhance the model's understanding and coverage of user queries, leading to more accurate and relevant responses.
4. Visualize the vector space to better understand the distribution and relationship of data points, aiding in the fine-tuning and improvement of RAG models for more effective query handling.

Lunch Break 🍽️

Session 13: Theory of NLP

Duration: 1 hour

Resource: [Slides](#)

Outcome: Understand advances in Natural Language processing from heuristic approaches to neural methods, vectors as embeddings, RNNs and token based modeling.

Objective: Explain how modern deep learning techniques facilitate solving complex problem in the language space and how these models can be trained efficiently.

Key Activities:

1. Explain the foundations of neural approaches to language models
2. Vectors and embeddings
3. How RNNs work and why they are inadequate?

Session 14: Theory of LLMs

Duration: 2 hours


Resources: [Slides](#)

Outcome: Understand the underlying modules of LLMs are built and what they can and cannot achieve.

Objective: Explain how LLMs are built and can be used to solve various problems in NLP and cover the transformer architecture.

Key Activities:

1. LLM settings
2. Perplexity metric
3. Temperature and top-k predictions
4. Memory and continuation
5. The transformer model and how it works
6. Attention and various uses

10-min Break 

✓ Session 15: Wrapping Together with Advanced topics.

-- <https://colab.research.google.com/drive/1GKnxasdBqh81YooV8oCqekTwB3YYTXuY>

- Real time data with LLMs: [Access Here](#)
- Types of Memory in LLMs Notebook: [Access Here](#)
- Agents: [Access Here](#)
- Combining SQL data with LLMs: [Access Here](#)
- RAG with SQL: [Access Here](#)
- RAG with authorization levels: [Access Here](#)
- Analyzing and Querying a Document with LLM Notebook: [Access Here](#)
- Table Analysis with Unstructured IO Notebook: [Access Here](#)
- Multimodal advanced table and image analysis: [Access Here](#)
- Additional Resources: [Slides](#)

Key Activities

0. Importing and setting up dataframes from a world history dataset.
1. Visualizing data using Geopandas and Wordcloud.
2. Transitioning from a relational DB to a Vector DB.
3. Encoding text with an embeddings model.
4. Implementing ChromaDB for storing embeddings.
5. Deploying and quantizing an open-source LLM, Mistral, for Retriever Augmented Generation (RAG).
6. Benchmarking against GPT-4 and generating SQL queries with LLM.

✓ Additional Notebooks:

1. Langchain Intro [Access here](#)
 2. Simple Chatbot on Your Enterprise Content: [Access Here](#)
 3. [Generate natural sounding audio](#)
 4. [Summarize a large book](#)
 5. [Generative Models: GANs](#)
 6. [Deep dive into Stable Diffusion](#)
 7. [K-means Clustering](#)
 8. [PCA and Dimensionality reduction](#) -- External Reference.
 9. [Audio - Spoken keyword detection with CNN and spectrogram](#)
 10. [Dataset creation: Web Scraping](#) (use it with caution)
 11. [Chaining in Langchain](#)
 12. [Tokenizing and Vectorising](#)
 13. Building a Simple Fake News Classifier Notebook: [Access Here](#)
-

✓ Projects

Project 1: Binary Classification on Titanic Disaster data

Project Overview The objective of this project is to predict whether a passenger survived the Titanic disaster, based on a set of features such as age, sex, passenger class, etc. This is a binary classification problem, which is a key concept in machine learning.

Dataset The dataset for this project is provided by [Kaggle](#) and is split into two parts: a training set and a test set. The training set includes the passenger's outcome (survived or not) along with the features, while the test set only includes the features.

Skills You Will Learn and Practice

This project will give you hands-on experience with several important data science and machine learning skills:

1. **Data Exploration and Visualization:** You'll use libraries like Matplotlib and Seaborn to explore the data and visualize the relationships between features. Data Preprocessing: The dataset includes missing values and categorical features, so you'll learn how to handle these common issues.

2. **Binary Classification:** You'll learn about and apply algorithms like logistic regression, decision trees, and random forests.
3. **Model Evaluation:** You'll learn how to evaluate your models using techniques like cross-validation and metrics like accuracy.

Submission

Once you've built your model, you can submit your predictions to Kaggle and see how your model performs on the unseen test data. This will give you a sense of how well your model generalizes to new data, which is a key aspect of machine learning.

This project provides a good balance of challenge and accessibility for beginners, and it covers many of the fundamental concepts and techniques in machine learning. I believe it will be a valuable learning experience for all of you.

Project 2: Quora Insincere Questions Classification

Project Overview Quora is a platform that empowers people to learn from each other. It's a place to gain and share knowledge, about anything. It's a platform to ask questions and connect with people who contribute unique insights and quality answers. Unfortunately, not all questions asked are sincere. Some are posted with malicious intent, and these insincere questions need to be identified.

Your task for this project is to build a model that identifies and flags insincere questions. To achieve this, we will be using an LLM such as Mistral, one of the most advanced language models currently available.

Dataset

The dataset for this project is provided by Quora and is available on [Kaggle](#). It consists of a large number of questions, some of which are labeled as insincere.

Skills You Will Learn and Practice

This project will give you hands-on experience with several important concepts in Natural Language Processing and AI:

1. **Working with Large Language Models:** You'll get hands-on experience with LLM such as Mistral, learning how to interact with the model via the OpenAI API.
2. **Text Classification:** You'll learn how to apply AI tools to classify text into different categories, a common task in NLP.
3. **Interpreting AI Outputs:** You'll gain experience in interpreting and making use of the outputs from an AI model. Submission

Once you've built your model and used it to classify the Quora questions, you can submit a report on LinkedIn with a link on Discord detailing your approach and the results. We'll be looking for creative and effective uses of LLM such as Mistral, as well as clear presentation of your methods and findings.

This project will provide a challenging but rewarding experience as you apply state-of-the-art AI to a real-world problem. As always, please don't hesitate to reach out if you have any questions or need any help.

Project 3: Classify images from CIFAR-10 dataset

I trust that you are all doing well and enjoyed working on the Titanic survival prediction project and the Quora project. We will be diving into the fascinating world of Computer Vision, a field of Artificial Intelligence that trains computers to interpret and understand the visual world.

Project Overview

For this week, your project will be to build a Convolutional Neural Network (CNN) to classify images from the [CIFAR-10 dataset](#). The CIFAR-10 dataset consists of 60,000 32x32 color images in 10 classes, with 6,000 images per class. The classes include objects like cars, birds, cats, ships, etc. Your task is to build a CNN that can accurately classify these images.

Skills You Will Learn and Practice

This project will give you hands-on experience with several important concepts in computer vision and deep learning:

- 1. Image Preprocessing:** You'll learn how to preprocess images for machine learning, including resizing images, normalizing pixel values, and converting labels to one-hot vectors.
- 2. Convolutional Neural Networks (CNNs):** You'll learn about CNNs, which are a type of deep learning model that are especially good at processing images.
- 3. Training Deep Learning Models:** You'll learn how to train a CNN using techniques like stochastic gradient descent and backpropagation.
- 4. Model Evaluation:** You'll learn how to evaluate your model's performance, understand confusion matrices, and compute metrics like accuracy.

Submission

Once you've built and trained your model, you will test it on the CIFAR-10 test set to see how well it can classify new images. This will give you a sense of how well your model generalizes, which is a key aspect of machine learning.

This project will be a bit more challenging than the last one, but I'm confident that you're all up to the task. As always, please don't hesitate to reach out if you have any questions or need any help.

Project 4: Sentiment Analysis on Amazon Reviews

Project Overview

Amazon, as one of the world's largest online retailers, has an immense repository of customer reviews. These reviews are invaluable for understanding customer satisfaction and guiding potential buyers. However, with millions of reviews, it's challenging to distinguish between positive and negative sentiments manually. This project aims to automate the process by building a model that can accurately perform sentiment analysis on Amazon reviews.

Your task is to develop a model that analyzes the sentiment of reviews—classifying them into positive or negative categories. We will leverage a sophisticated Large Language Model (LLM), such as GPT (a successor to Mistral and among the most advanced models available), to accomplish this task.

Dataset

The dataset for this project is a popular collection of Amazon reviews available on [Kaggle](#). It includes reviews from various product categories, each tagged with helpfulness votes, scores, and text. This rich dataset will serve as the foundation for your sentiment analysis model.

Skills You Will Learn and Practice

This project will equip you with practical experience in several key areas of AI and Natural Language Processing (NLP):

- 1. Working with Large Language Models:** You will gain hands-on experience with an advanced LLM, such as GPT, learning how to harness its capabilities via the OpenAI API for sentiment analysis.
- 2. Sentiment Analysis:** This project will deepen your understanding of how AI can be used to interpret complex human emotions expressed in text, teaching you to classify sentiments effectively.
- 3. Data Preprocessing and Analysis:** You'll learn how to prepare a large dataset for processing, including cleaning, tokenization, and feature extraction techniques essential for NLP tasks.
- 4. Model Evaluation:** You will develop skills in assessing the performance of your model, understanding metrics such as accuracy, precision, recall, and F1 score, which are crucial for evaluating AI models.

Submission

After developing your sentiment analysis model, you can submit a comprehensive report to us on Discord and LinkedIn detailing your methodology, tools used, analysis of the dataset, and your findings. We are interested in innovative approaches to using LLMs like GPT4, effective data preprocessing strategies, and insightful interpretations of your model's performance.

This project offers a challenging opportunity to apply cutting-edge AI technology to solve a real-world problem, enhancing your skills in machine learning and NLP. We encourage creativity and critical thinking throughout this project and are here to support you with any questions or guidance you may need.

✓ Project 5: Identifying Fake Jobs

[Real / Fake Job Posting Prediction](#)

Project Overview

The internet has transformed the job market, making it easier than ever for job seekers to find potential opportunities. However, this ease of access has also led to an increase in fraudulent job postings designed to deceive or scam job seekers. The challenge of identifying these fake job postings is significant, not only for job seekers but also for legitimate employers and job boards. This project aims to address this challenge by developing a model capable of accurately predicting whether a job posting is real or fake.

You will use a sophisticated Large Language Model (LLM), such as GPT (a successor to models like Mistral and among the most advanced available), to analyze job postings and classify them as real or fake. This involves not only understanding the textual content but also identifying subtle cues and patterns that distinguish legitimate postings from fraudulent ones.

Dataset

The dataset for this project consists of a mixture of genuine and fake job advertisements and is available on [Kaggle](#). It includes various features such as company descriptions, job descriptions,