GENAI LAB 1

NAME : ANJALI GUNTI	SRN : PES2UG23CS075	SECTION : B

Project Summary: Generative AI & NLP Fundamentals

1. Introduction: Setting up the Workshop

We started by learning about Hugging Face, which acts like the "GitHub for AI." Instead of building models from scratch (which is expensive and hard), we use the transformers library to download pre-trained models.

- The Magic Tool: We used the pipeline() function. This function is a lifesaver because it automates three hard steps:
    1. Preprocessing (turning text into numbers).
    2. Inference (the model making a prediction).
    3. Post-processing (turning numbers back into text we can read).

2. Model Efficiency: Main vs. Distilled Models

We compared a "Smart" model (GPT-2) with a "Fast" model (DistilGPT-2). This taught us about the trade-offs in AI:

- Storage (Size): The distill model is much smaller (fewer parameters), saving disk space and RAM.
- Accuracy: The accuracy of the distill model decreases slightly compared to the main model, but it is usually efficient enough for most apps.
- Latency (Speed): This is the time it takes for the model to reply. Latency is a good factor to build a model for real-time apps (like chatbots) because distilled models are much faster.

3. The Architecture: Encoders vs. Decoders

We learned that the "Transformer" (the brain of these models) has different parts for different jobs:

- Encoder (The Reader): This part knows how to read. It looks at the whole sentence at once (bi-directional).
    - BERT is an Encoder model. It needs to iterate from both sides (read left-to-right and right-to-left) to understand context. This makes it perfect for Named Entity Recognition (NER) (finding names, places, orgs).
- Decoder (The Writer): This part knows how to write. It reads only from left-to-right.
    - GPT-2 is a decoder model. It is built to generate text by guessing the next word.
- Encoder-Decoder (The Hybrid): This combines both. BART is the example we used.

- Summarization Question: We use BART (not BERT) for summarization. *Why?* Because summarization requires *reading* a long text (Encoder) and *writing* a new, short summary (Decoder). BERT is only a reader, so it can't write the summary.

4. How GPT-2 Generates Text (The Mechanics)

GPT-2 doesn't just "know" the answer; it predicts the next value based on probability (math).

- Softmax: This is the math function that converts the model's raw scores into probabilities (percentages).
- Sampling: This is how GPT predicts next words. Instead of just picking the #1 most likely word (which is boring), it samples from a list of good options.
- Temperature: This controls the "creativity" or risk level.
  - More temperature = more unpredictable (creative/random).
  - Less temperature = more predictable (safe/repetitive).
- Top-K: The model looks at the Top-K number of elements (e.g., the top 50 most likely words) and ignores the unlikely ones.
- Top-P (Nucleus Sampling): This is the aggregation of Top-P. The model sums up the probabilities of the best words until it reaches the P value (e.g., 0.90) and only picks from that group.

5. Reproducibility: Controlling Randomness

- Seed Values: Since the model uses random sampling, different types of seed values will produce different stories.
- If you want the exact same output every time (for debugging), you must set the seed to a fixed number (like 42).

EXAMPLE:

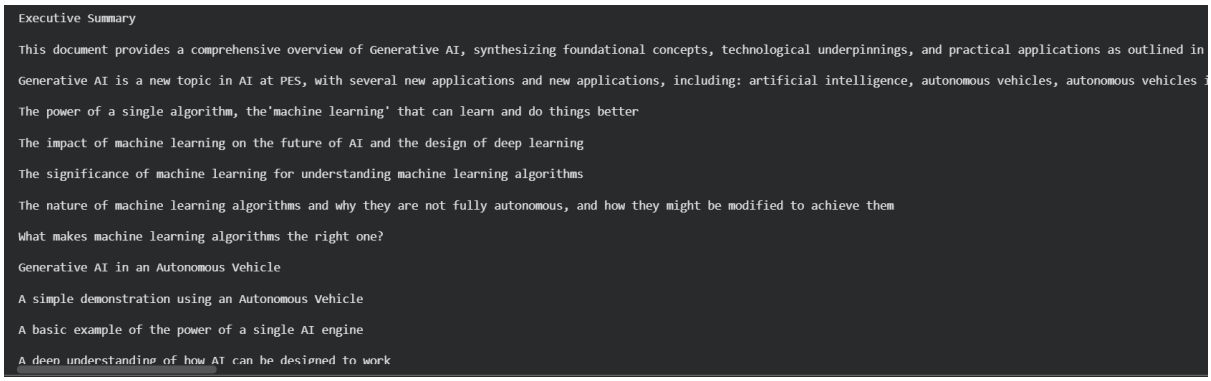
Seed value 42:

```
Executive Summary
This document provides a comprehensive overview of Generative AI, synthesizing foundational concepts, technological underpinnings, and practical applications as outlined in
In this book the goal is to summarize and describe the major issues, as well as the research areas covered, for a broad range of approaches to AI research, including:
The challenges of machine learning to address the potential problems of machine learning, as well as the computational and cognitive complexities associated with machine lea
The advantages of new methods for extracting insights from data generated from the context of machine learning, and by using data and models generated from them to create sy
The opportunities to integrate natural language systems with AI, and the
```

Seed value 98

```
This document provides a comprehensive overview of Generative AI, synthesizing foundational concepts, technological underpinnings, and practical applications as outlined
1 Introduction
Generative AI is an emerging field of AI research, where the main focus is on the computational capabilities of individual users. In this document, we outline three core
```

Seed value 123



6. Linguistics: Parts of Speech (POS)

We used NLTK to tag the parts of speech in the sentence generated by GPT-2 (*"Generative AI is a revolutionary technology..."*) to prove the model understands grammar.

- Generative (JJ - Adjective)
- AI (NNP - Proper Noun, Singular)
- is (VBZ - Verb, 3rd person singular)
- a (DT - Determiner)
- revolutionary (JJ - Adjective)
- technology (NN - Noun, Singular)
- that (WDT - WH-determiner)
- has (VBZ - Verb)
- the (DT - Determiner)
- potential (NN - Noun)

7. Question Answering (QA)

- Why is question answering used? It allows us to extract specific facts from a massive text file (Knowledge Base) without reading the whole thing.
- We used DistilBERT for this because it is an Encoder (good at reading/finding) and it is distilled (fast).

8. Beyond Accuracy & Latency: Other Critical Factors

You asked if we only care about accuracy and latency. The answer is NO. Based on the notebook, here are the other factors you must write about:

1. Storage/Memory Footprint:
    - Even if a model is fast (low latency), if it is too big (storage), it won't fit on a phone or a small server. We chose distilgpt2 partly because it is smaller.
2. Coherence (Quality):
    - The notebook mentioned that while distilgpt2 is fast, it might be less coherent (make less sense) over long paragraphs compared to the larger gpt2. Accuracy measures "right or wrong," but coherence measures "does this sound like a human?"
3. Reproducibility:
    - In a real project, being able to get the same result twice is a critical factor. This is why we emphasized Seed Values.
4. Context Window (Attention):
    - For tasks like NER or Summarization, the model's ability to handle ambiguity and long-range dependencies (remembering the start of the sentence at the end) is a key factor. This is why the Transformer's attention mechanism was highlighted.