GEN_AI_LAB

NAME-Amitesh Sinha SECTION-A
SRN-PES2UG23CS054 DATE-23JAN

HUGGING FACE

It is like a big hub where we can download ready made AI models instead of making it from the scratch and companies put up their model which are distilled version can gain popularity

Distilled version vs original model

Original model- Maximum intelligence, slower,expensive to always.

Distilled model-High efficiency ,lightning fast,cheap to run, slightly less intelligent

Feature	Original Model (Teacher)	Distilled Model (Student)
Size (Parameters)	Massive (e.g., 175 Billion). Requires huge memory.	Compressed (e.g., 7 Billion). Fits on smaller devices.
Speed (Latency)	Slower. Generates text at a leisurely pace.	Very Fast. Can generate text in near real-time.
Cost (Inference)	Expensive. Requires clusters of powerful GPUs (A100s/H100s).	Cheap. Can often run on consumer hardware or cheaper cloud instances.
Accuracy	**State-of-the-Art.** Captures deep nuance, rare facts, and complex reasoning.	**Good Enough.** usually retains ~90-97% of the teacher's performance on general tasks.
Capabilities	better at "reasoning," math, coding, and obscure edge cases.	Better at specific, repetitive tasks (classification, summarization) where deep reasoning isn't needed.

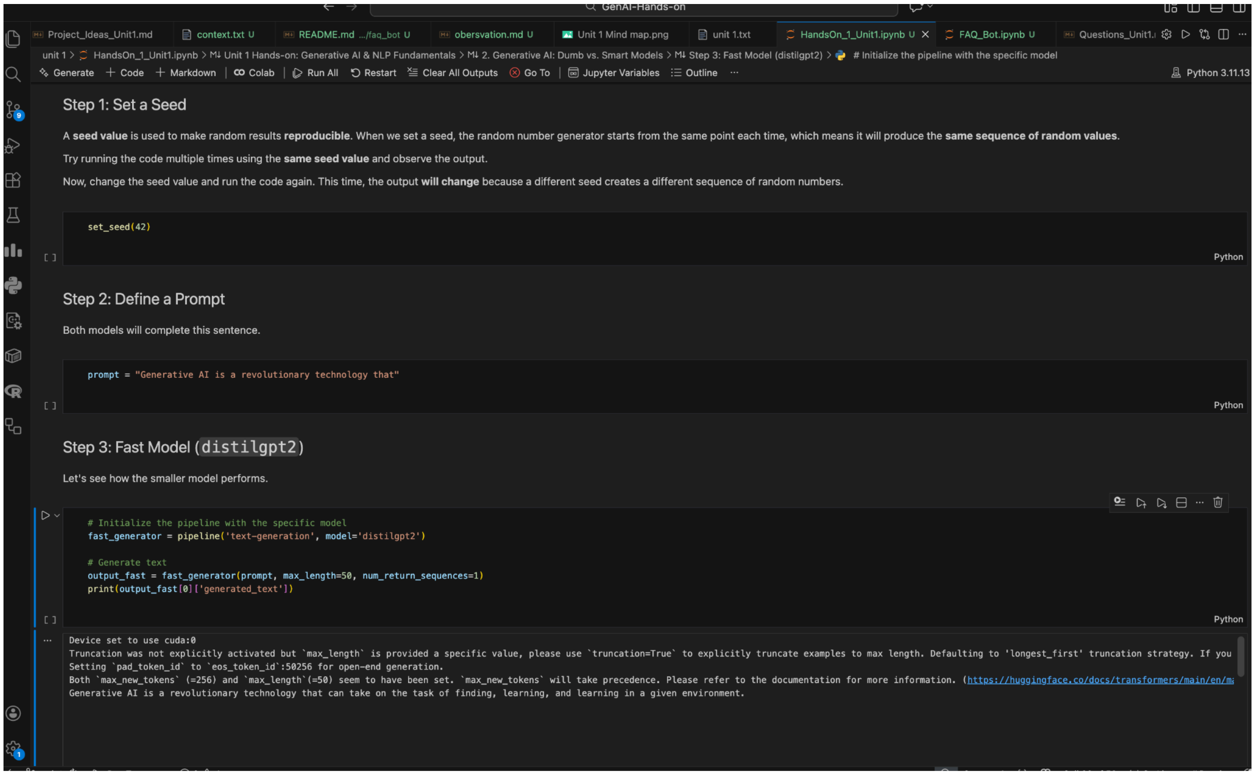
```
- The `transformers` library is the bridge that lets us load and use those models
easily.

- `pipeline()` hides all the complicated steps (tokenization → model → readable
output), so we can focus on results.
```

Seed Value

A seed value defines the starting point of a pseudo-random number generator.

Because computer are logical machines , they cannot generate true randomness instead they use complex algorithms to generate a sequence of numbers that looks random.The seed is the input variable that kicks off that formula.

Think of the random number generator as a massive book having millions of chaotic lists of the numbers.
>The seed is the page number
>if we tell the pc to start at page 42 it will start exactly same list of number every single time
Seed value(42)



Step 1: Set a Seed

A **seed value** is used to make random results **reproducible**. When we set a seed, the random number generator starts from the same point each time, which means it will produce the **same sequence of random values**.

Try running the code multiple times using the **same seed value** and observe the output.

Now, change the seed value and run the code again. This time, the output **will change** because a different seed creates a different sequence of random numbers.

```
set_seed(42)
```

Python

Step 2: Define a Prompt

Both models will complete this sentence.

```
prompt = "Generative AI is a revolutionary technology that"
```

Python

Step 3: Fast Model (`distilgpt2`)

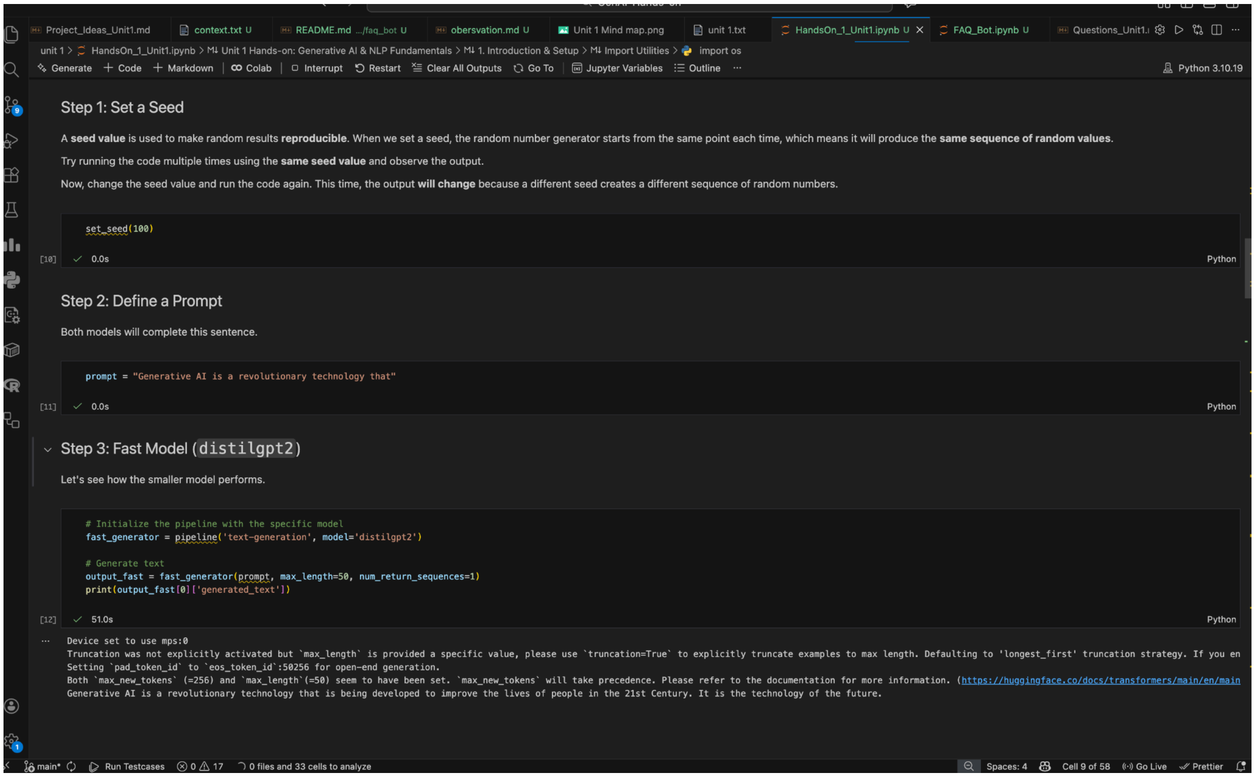
Let's see how the smaller model performs.

```
# Initialize the pipeline with the specific model
fast_generator = pipeline('text-generation', model='distilgpt2')

# Generate text
output_fast = fast_generator(prompt, max_length=50, num_return_sequences=1)
print(output_fast[0]['generated_text'])
```

Python

```
Device set to use cuda:0
Truncation was not explicitly activated but `max_length` is provided a specific value, please use `truncation=True` to explicitly truncate examples to max length. Defaulting to 'longest_first' truncation strategy. If you
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Both `max_new_tokens` (=256) and `max_length`(=50) seem to have been set. `max_new_tokens` will take precedence. Please refer to the documentation for more information. (https://huggingface.co/docs/transformers/main/en/ma
Generative AI is a revolutionary technology that can take on the task of finding, learning, and learning in a given environment.
```

Seed Value(100)



Step 1: Set a Seed

A **seed value** is used to make random results **reproducible**. When we set a seed, the random number generator starts from the same point each time, which means it will produce the **same sequence of random values**.

Try running the code multiple times using the **same seed value** and observe the output.

Now, change the seed value and run the code again. This time, the output **will change** because a different seed creates a different sequence of random numbers.

```
set_seed(100)
```

[10] ✓ 0.0s Python

Step 2: Define a Prompt

Both models will complete this sentence.

```
prompt = "Generative AI is a revolutionary technology that"
```

[11] ✓ 0.0s Python

Step 3: Fast Model (`distilgpt2`)

Let's see how the smaller model performs.

```
# Initialize the pipeline with the specific model
fast_generator = pipeline('text-generation', model='distilgpt2')

# Generate text
output_fast = fast_generator(prompt, max_length=50, num_return_sequences=1)
print(output_fast[0]['generated_text'])
```

[12] ✓ 51.0s Python

```
Device set to use mps:0
Truncation was not explicitly activated but `max_length` is provided a specific value, please use `truncation=True` to explicitly truncate examples to max length. Defaulting to 'longest_first' truncation strategy. If you en
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Both `max_new_tokens` (=256) and `max_length`(=50) seem to have been set. `max_new_tokens` will take precedence. Please refer to the documentation for more information. (https://huggingface.co/docs/transformers/main/en/main
Generative AI is a revolutionary technology that is being developed to improve the lives of people in the 21st Century. It is the technology of the future.
```

Scenario	Same Seed	Different Seed
Video Games	The layout of a specific Minecraft world (if you share the seed, your friend gets the exact same mountains).	A completely new, unique world layout.
Data Science	Splitting your data into "Training" and "Testing" sets exactly the same way every time.	Splitting the data differently, which is useful for checking if your model is robust (Cross-validation).
Generative Art	Recreating an exact image you made yesterday.	Creating 10 different versions of a logo to choose the best one.

Temperature

In the context of LLMs temperature is a setting that controls the creativity vs precision of the model output.
It can be visualized as a Risk dial. It determines how wild of safe it is the AI allows itself to be when choosing the next word in a sentence.

Low Temparture(<0.5):The model becomes highly confident and conservative .It essentially says , will pick the word is most likely to be correct. It amplifies the difference between the best answer and the okay answer.

High temperature(>0.7):The model becomes more adventurous .It flattens the probability curve , giving less common words a fighting chance to be picked .It says,"i might pick a word just to see where it goes."

The spectrum of Temperature

Temperature Value	**Behavior**	**Best Use Cases**
0.0 - 0.3	**Deterministic & Focused.** The AI is repetitive and sticks to the most likely facts. It is very literal.	Coding, Math, Data Extraction, Factual Q&A.
0.4 - 0.7	**Balanced.** A mix of coherence and creativity. The standard setting for most chatbots (like ChatGPT or Gemini default).	General conversation, email writing, summarizing.
0.8 - 1.0+	**Creative & Random.** The AI takes risks. It uses diverse vocabulary but is more prone to making things up (hallucinations).	Creative writing, poetry, brainstorming, generating unique names.

Tokenization

Model doesn't understand words in a particular human language it understands numbers that is tokens which we have multiple ways of generations like ngrams, byte pair tokenization, etc.

POS Tagging

Parts of Speech basically noun, verb , adjective , adverb, preposition , conjunction because model uses these to understand what word to place where, so these parts of speech needs to be tagged in order for our model to place the word at correct position .

The main reason we need POS tagging is that english is ambiguous. The same words can mean completely different things depending on where it sits in a sentence.

IT is important ?
> Text-To-Speech(TTS)
>Sentiment Analysis
>Search Engines

10 standard Part-of-Speech (POS) tags from the **Penn Treebank Tagset**.

#	Tag	Meaning	Example Word
1	NN	Noun, singular or mass	*computer, time, logic*
2	NNP	Proper Noun, singular	*Google, London, John*
3	DT	Determiner	*the, a, these*
4	JJ	Adjective	*fast, blue, numeric*
5	VB	Verb, base form	*eat, run, calculate*
6	VBD	Verb, past tense	*ate, ran, calculated*
7	RB	Adverb	*very, silently, quickly*
8	IN	Preposition / Subordinating Conjunction	*in, of, like, after*
9	CC	Coordinating Conjunction	*and, but, or*
10	MD	Modal	*can, could, might, will*

NER(Named Entity Recognition)

Sub-task of Natural Language Processing (NLP) that automatically identifies specific words in a text and a classifies them into predefined categories.

It answers 2 questions about a text:

>Detection: Where are important proper noun?

>Classification:What type of thing are they?

Task	**Best Architecture**	**Why?**
NER (Classification)	**Encoder Only** (BERT, RoBERTa)	Needs to understand the full context of a word from both directions to label it correctly.
Translation	**Encoder-Decoder** (T5, BART)	Needs to understand the input (Encoder) and then generate a new output in a different language (Decoder).
Summarization	**Encoder-Decoder** (T5, BART) or **Decoder** (GPT)	Needs to generate text. BERT (Encoder) struggles here unless you just want to highlight sentences.
Text Generation	**Decoder Only** (GPT, Llama)	The best at "predicting the next word" to flow naturally.

Documentation — FAQ Bot

Overview

FAQ Bot answers user questions by extracting exact answer from a fixed policy/FAQ document. Ituses a pre-trained extractive questions answering model so no custom training is required.

Objective

Provide accurate, grounded answers to common policy questions (password reset, refunds, privacy, support hours).

Methodology

Task: Extractive Question Answering
Model: DistilBERT fine-tuned on SQuAD
Approach: Load a policy text file as context, then run a QA pipeline to locate the best answer span.

Workflow

Load the policy/FAQ document as context.
Initialize the QA pipeline with a pre-trained model.
Ask questions and return the best matching span with confidence.
(Optional) Use an interactive loop for live Q&A.

Key Features

Fast, lightweight model suitable for CPU.
Grounded answers (no hallucinations) because responses come from the given context.
Easy to customize by replacing the policy document.