GenAI HandsOn-1

Name: Avani Ajith	SRN: PES2UG23CS110	Section: B

1. Model Comparisons: Standard vs. Distilled

We explored the differences between standard Large Language Models (LLMs) and their distilled versions.

- **Main Model (e.g., GPT-2):** These are larger and typically more coherent. They are trained to predict the next word in a sequence based on probability (using a Softmax function).
- **Distilled Model (e.g., DistilGPT-2):** A smaller, faster version optimized for speed and lower memory usage. However, this distillation often leads to a decrease in accuracy and coherence compared to the main model.

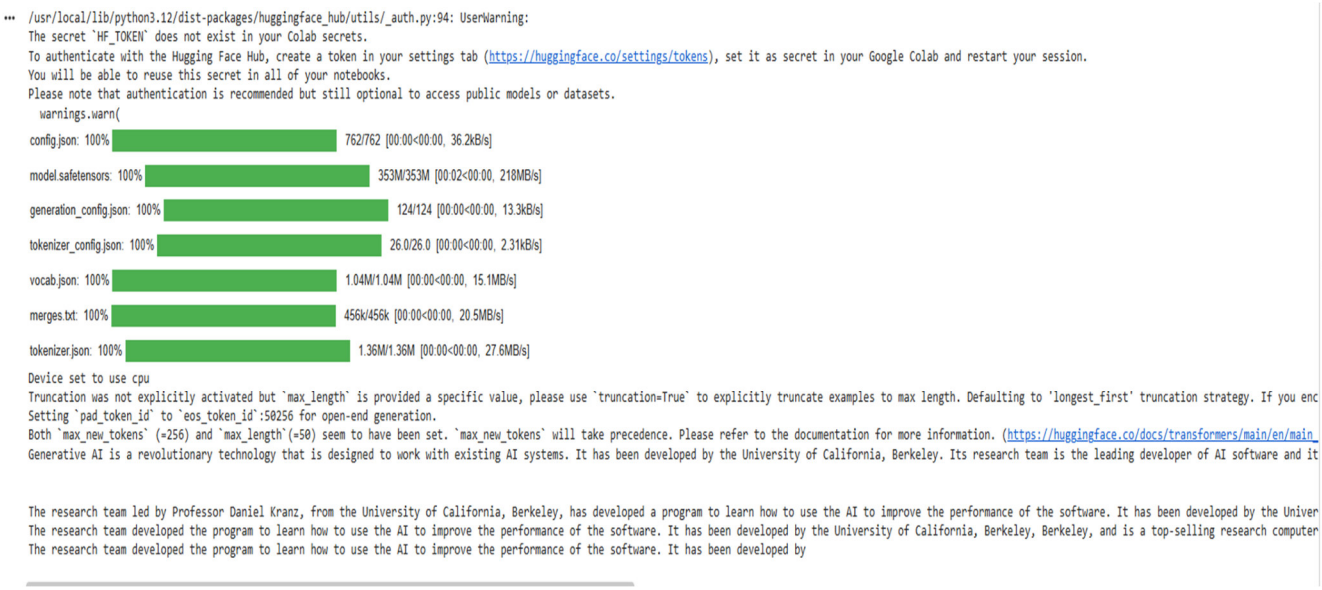
2. The Role of Seed Values

A seed value is used to ensure that random results are reproducible. By setting a seed, the random number generator starts from the same point, producing the same sequence of values every time the code is run.

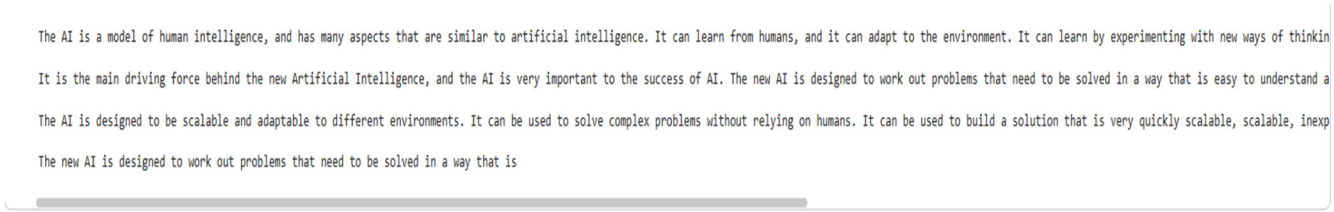
Experimenting with Seed Values:

- **Seed 42:**

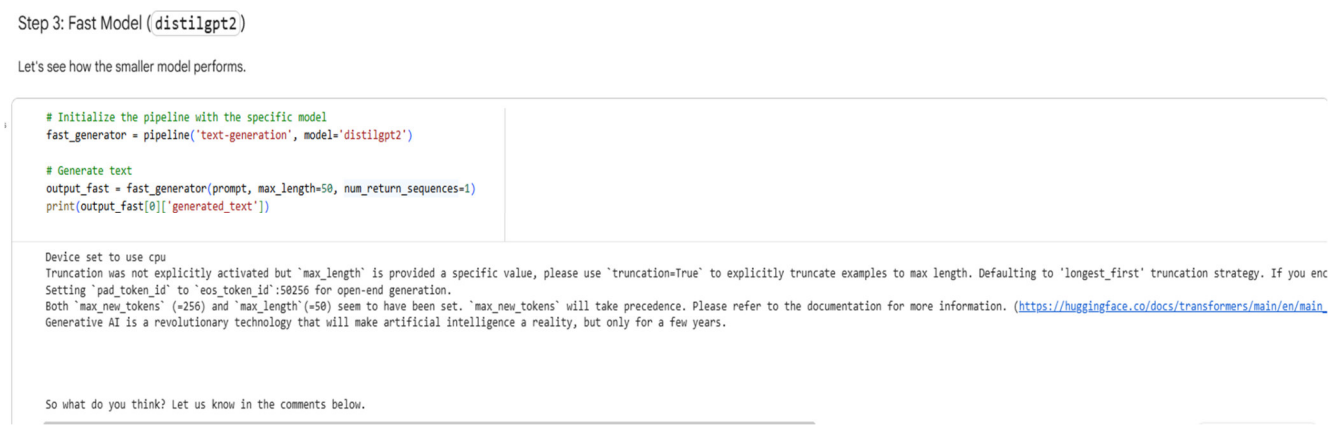
For distilgpt:



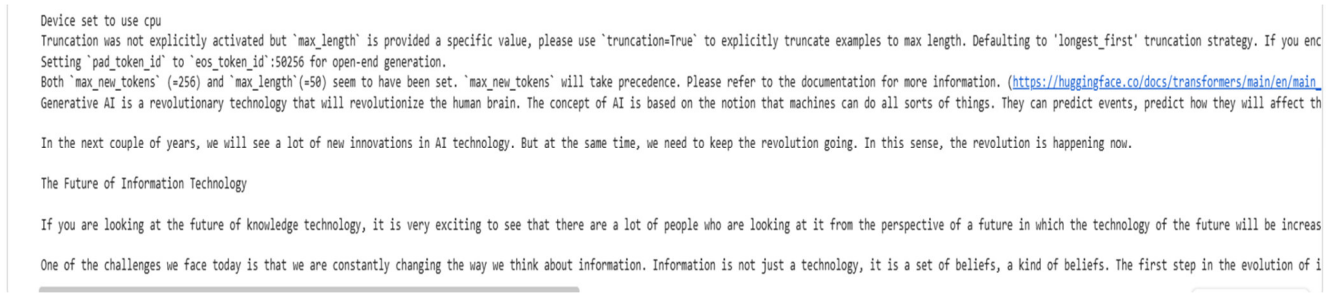
For GPT2:



- **Seed 123:**

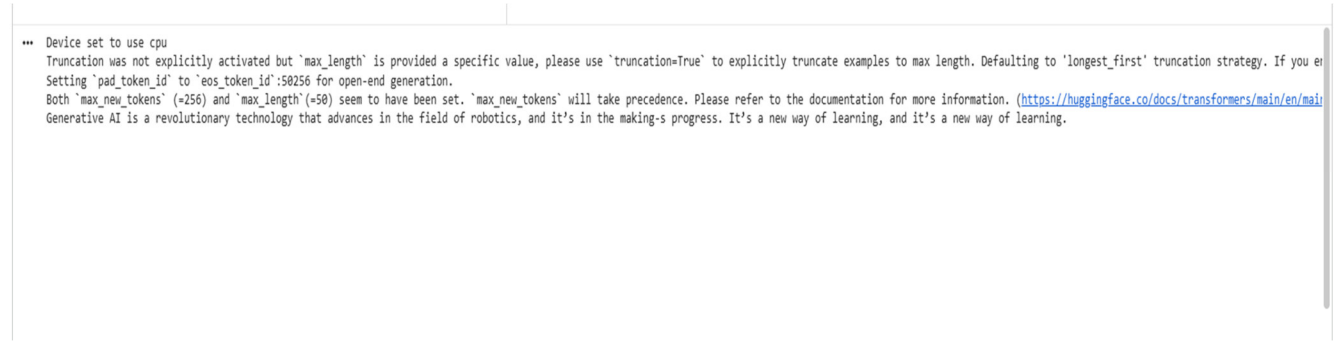


For GPT2:

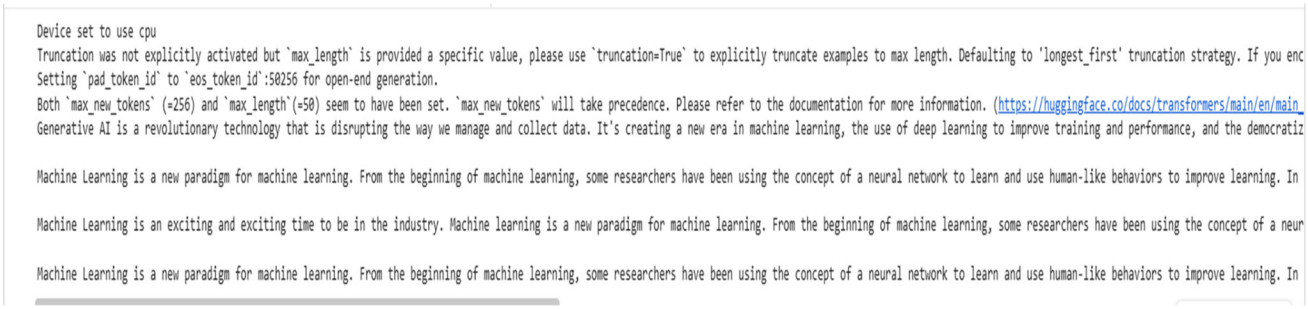


- **Seed 98**

For DistilGPT:



For GPT2:



3. NLP Fundamentals: Under the Hood

Before a model can generate text, several processing steps occur:

- **Tokenization:** Models cannot read text directly. Tokenization breaks text into "Tokens" and assigns each a unique numerical ID.
- **POS Tagging:** This identifies the grammatical parts of speech (e.g., NNS for plural nouns, VBD for past tense verbs) to help the model understand sentence structure.
- **Named Entity Recognition (NER):** Used to extract structured data like names, organizations (ORG), or miscellaneous (MISC) entities from raw text.

4. Key Model Architectures

We differentiated between the "reading" and "writing" components of Transformers:

- **Encoders:** Used to "read" and understand the context of the input. Models like BERT are bidirectional, meaning they read text from both left-to-right and right-to-left to fully grasp context (e.g., distinguishing if "Apple" is a fruit or a company).
- **Decoders:** Used to "write" or generate text. GPT-2 is a **decoder-only** model, trained specifically to predict the next word based on preceding ones.
- **Combined Architectures:** Tasks like Question Answering or filling in the blanks (Masked Language Modeling) often utilize both encoders and decoders.

5. Advanced Decoding Concepts

To control the variety and quality of generated text, we use specific parameters:

- **Temperature:** Controls randomness. A higher temperature makes the output more diverse/random, while a lower temperature makes it more predictable.

- **Top-K:** The model selects the top *k* most likely next words.
- **Top-P:** The model selects the smallest set of words whose cumulative probability is greater than or equal to p.

A major focus for the upcoming year in LLM development includes:

- **Latency:** Reducing the time it takes for a model to generate a response.
- **Context Engineering:** Improving how models handle and remember long-range dependencies within a large text context.

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)