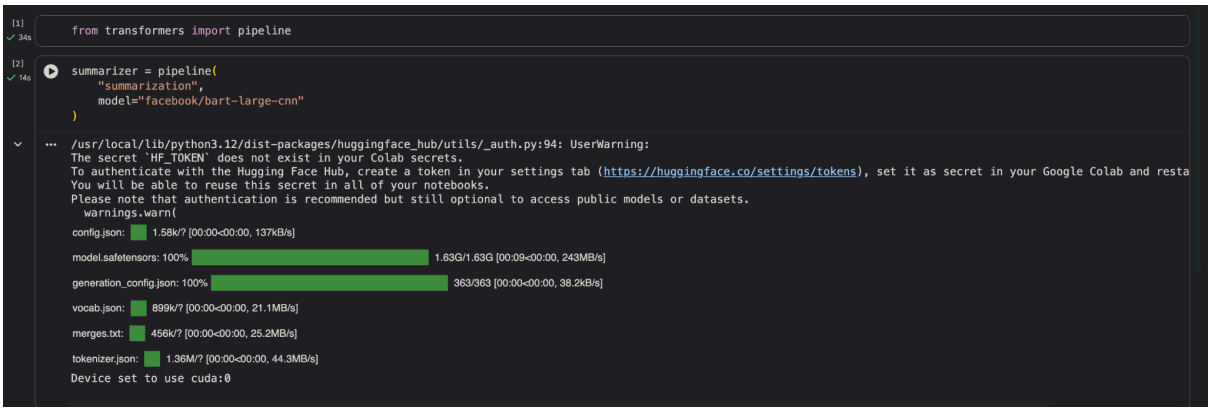
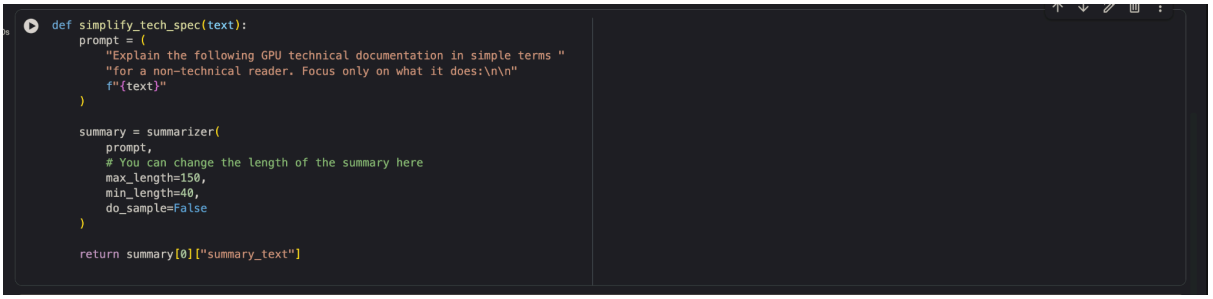
GenAI Project Unit-1

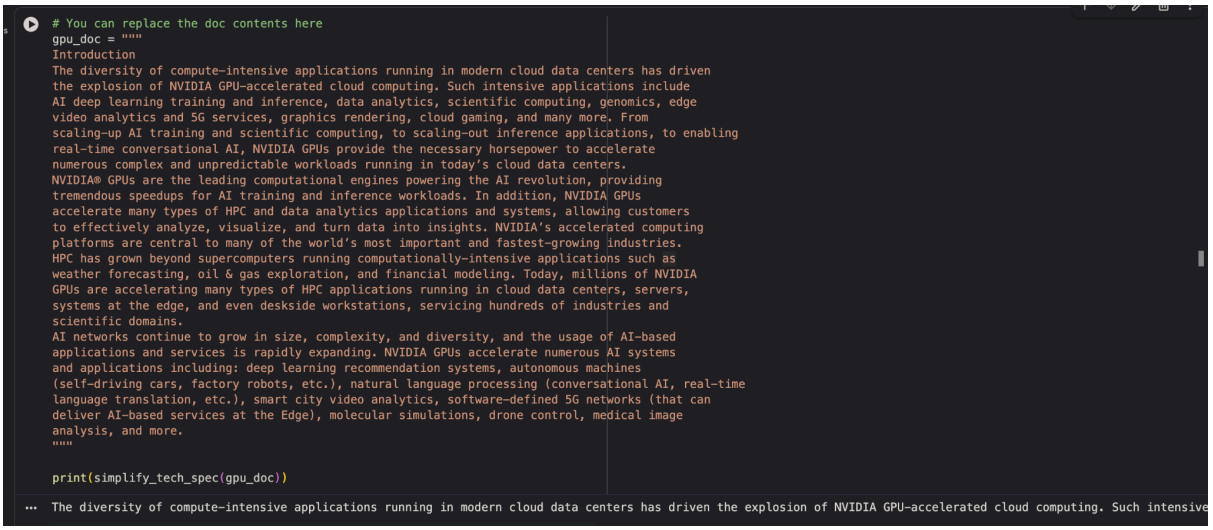
PES2UG23CS072
ANIRUDH SRIPADA KOUNDINYA M



```
[1] 34s
from transformers import pipeline

[2] 14s
summarizer = pipeline(
    "summarization",
    model="facebook/bart-large-cnn"
)

/usr/local/lib/python3.12/dist-packages/huggingface_hub/utils/_auth.py:94: UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab (https://huggingface.co/settings/tokens), set it as secret in your Google Colab and resta
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access public models or datasets.
  warnings.warn(
config.json: 1.58k/? [00:00<00:00, 137kB/s]
model.safetensors: 100% 1.63G/1.63G [00:09<00:00, 243MB/s]
generation_config.json: 100% 363/363 [00:00<00:00, 38.2kB/s]
vocab.json: 899k/? [00:00<00:00, 21.1MB/s]
merges.txt: 456k/? [00:00<00:00, 25.2MB/s]
tokenizer.json: 1.36M/? [00:00<00:00, 44.3MB/s]
Device set to use cuda:0
```



```
def simplify_tech_spec(text):
    prompt = (
        "Explain the following GPU technical documentation in simple terms "
        "for a non-technical reader. Focus only on what it does:\n\n"
        f"{text}"
    )

    summary = summarizer(
        prompt,
        # You can change the length of the summary here
        max_length=150,
        min_length=40,
        do_sample=False
    )

    return summary[0]["summary_text"]
```



```
# You can replace the doc contents here
gpu_doc = """
Introduction
The diversity of compute-intensive applications running in modern cloud data centers has driven
the explosion of NVIDIA GPU-accelerated cloud computing. Such intensive applications include
AI deep learning training and inference, data analytics, scientific computing, genomics, edge
video analytics and 5G services, graphics rendering, cloud gaming, and many more. From
scaling-up AI training and scientific computing, to scaling-out inference applications, to enabling
real-time conversational AI, NVIDIA GPUs provide the necessary horsepower to accelerate
numerous complex and unpredictable workloads running in today's cloud data centers.
NVIDIA® GPUs are the leading computational engines powering the AI revolution, providing
tremendous speedups for AI training and inference workloads. In addition, NVIDIA GPUs
accelerate many types of HPC and data analytics applications and systems, allowing customers
to effectively analyze, visualize, and turn data into insights. NVIDIA's accelerated computing
platforms are central to many of the world's most important and fastest-growing industries.
HPC has grown beyond supercomputers running computationally-intensive applications such as
weather forecasting, oil & gas exploration, and financial modeling. Today, millions of NVIDIA
GPUs are accelerating many types of HPC applications running in cloud data centers, servers,
systems at the edge, and even deskside workstations, servicing hundreds of industries and
scientific domains.
AI networks continue to grow in size, complexity, and diversity, and the usage of AI-based
applications and services is rapidly expanding. NVIDIA GPUs accelerate numerous AI systems
and applications including: deep learning recommendation systems, autonomous machines
(self-driving cars, factory robots, etc.), natural language processing (conversational AI, real-time
language translation, etc.), smart city video analytics, software-defined 5G networks (that can
deliver AI-based services at the Edge), molecular simulations, drone control, medical image
analysis, and more.
"""

print(simplify_tech_spec(gpu_doc))

... The diversity of compute-intensive applications running in modern cloud data centers has driven the explosion of NVIDIA GPU-accelerated cloud computing. Such intensive
```

1. Problem Statement

GPU technical documentation is often complex and difficult for non-technical users to understand. This makes it harder for developers, product managers, and stakeholders to quickly grasp what a GPU actually does.

This project aims to automatically convert detailed GPU specifications into a short, plain-English “What it does” description.

2. Technical Architecture

The system is built using a text summarization pipeline based on a transformer model.

2.1 Summarization Logic

1. Input: Raw GPU documentation text.
2. Prompting: The input is prefixed with *“What it does:”* to guide the model toward functional summarization.
3. Model Processing: A pre-trained transformer model analyzes the text and condenses it while preserving core functionality.
4. Output: A short, readable explanation describing the GPU’s purpose and capabilities.

3. Implementation Details

- Model Used: `facebook/bart-large-cnn`
- Task Type: Abstractive summarization

- Library: Hugging Face `transformers`
- Language: Python
- Inference Method: Hugging Face `pipeline` API

4. Output Interpretation

The output is a concise natural-language summary that explains what the GPU does, rather than listing technical specifications.
This text can be used in:

- Product documentation
- Dashboards
- AI-generated reports
- Non-technical summaries