

!nvidia-smi

```
Thu Aug 14 08:54:49 2025
```

NVIDIA-SMI 550.54.15		Driver Version: 550.54.15		CUDA Version: 12.4	
GPU	Name	Persistence-M	Bus-Id	Disp.A	Volatile Uncorr. ECC
Fan	Temp	Perf	Pwr:Usage/Cap	Memory-Usage	GPU-Util Compute M. MIG M.
0	Tesla T4	Off	00000000:00:04.0	Off	0
N/A	74C	P0	43W / 70W	0MiB / 15360MiB	0% Default N/A

Processes:						
GPU	GI	CI	PID	Type	Process name	GPU Memory Usage
	ID	ID				
No running processes found						

!pip install transformers[sentencepiece] datasets sacrebleu rouge_score py7zr -q

!pip install --upgrade accelerate
!pip uninstall -y transformers accelerate
!pip install transformers accelerate

```
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from jinja2->torch>=2.0.0->accelerate) (3.0)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests->huggingface_hub>=0.
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests->huggingface_hub>=0.21.0->accele
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests->huggingface_hub>=0.21.0->
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests->huggingface_hub>=0.21.0->
Found existing installation: transformers 4.55.2
Uninstalling transformers-4.55.2:
  Successfully uninstalled transformers-4.55.2
Found existing installation: accelerate 1.10.0
Uninstalling accelerate-1.10.0:
  Successfully uninstalled accelerate-1.10.0
Collecting transformers
  Using cached transformers-4.55.2-py3-none-any.whl.metadata (41 kB)
Collecting accelerate
  Using cached accelerate-1.10.0-py3-none-any.whl.metadata (19 kB)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from transformers) (3.18.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.34.0 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.34.4)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-packages (from transformers) (2.0.2)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from transformers) (25.0)
```

```
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (2025.8.3)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from jinja2->torch>=2.0.0->accelerate) (3.0)
Using cached transformers-4.55.2-py3-none-any.whl (11.3 MB)
Using cached accelerate-1.10.0-py3-none-any.whl (374 kB)
Installing collected packages: transformers, accelerate
Successfully installed accelerate-1.10.0 transformers-4.55.2
```

```
!pip install -U transformers --quiet
```

```
!pip install evaluate
```

```
Requirement already satisfied: evaluate in /usr/local/lib/python3.11/dist-packages (0.4.5)
Requirement already satisfied: datasets>=2.0.0 in /usr/local/lib/python3.11/dist-packages (from evaluate) (4.0.0)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-packages (from evaluate) (2.0.2)
Requirement already satisfied: dill in /usr/local/lib/python3.11/dist-packages (from evaluate) (0.3.8)
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages (from evaluate) (2.2.2)
Requirement already satisfied: requests>=2.19.0 in /usr/local/lib/python3.11/dist-packages (from evaluate) (2.32.3)
Requirement already satisfied: tqdm>=4.62.1 in /usr/local/lib/python3.11/dist-packages (from evaluate) (4.67.1)
Requirement already satisfied: xxhash in /usr/local/lib/python3.11/dist-packages (from evaluate) (3.5.0)
Requirement already satisfied: multiprocessing in /usr/local/lib/python3.11/dist-packages (from evaluate) (0.70.16)
Requirement already satisfied: fsspec>=2021.05.0 in /usr/local/lib/python3.11/dist-packages (from fsspec[http]>=2021.05.0->evaluate) (2025.10.0)
Requirement already satisfied: huggingface-hub>=0.7.0 in /usr/local/lib/python3.11/dist-packages (from evaluate) (0.34.4)
Requirement already satisfied: packaging in /usr/local/lib/python3.11/dist-packages (from evaluate) (25.0)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from datasets>=2.0.0->evaluate) (3.18.0)
Requirement already satisfied: pyarrow>=15.0.0 in /usr/local/lib/python3.11/dist-packages (from datasets>=2.0.0->evaluate) (18.1.0)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-packages (from datasets>=2.0.0->evaluate) (6.0.2)
Requirement already satisfied: aiohttp!=4.0.0a0,!<4.0.0a1 in /usr/local/lib/python3.11/dist-packages (from fsspec[http]>=2021.05.0->evaluate) (4.0.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.7.0->evaluate) (4.12.0)
Requirement already satisfied: hf-xet<2.0.0,>=1.1.3 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.7.0->evaluate) (1.2.0)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests>=2.19.0->evaluate) (3.4.0)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests>=2.19.0->evaluate) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests>=2.19.0->evaluate) (2.5.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests>=2.19.0->evaluate) (2025.8.3)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas->evaluate) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from pandas->evaluate) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas->evaluate) (2025.2)
Requirement already satisfied: aiohappyeyeballs>=2.5.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!<4.0.0a1->fsspec[http]>=2021.05.0->evaluate) (2.5.0)
Requirement already satisfied: aiosignal>=1.4.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!<4.0.0a1->fsspec[http]>=2021.05.0->evaluate) (1.4.0)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!<4.0.0a1->fsspec[http]>=2021.05.0->evaluate) (25.0.1)
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!<4.0.0a1->fsspec[http]>=2021.05.0->evaluate) (1.5.0)
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!<4.0.0a1->fsspec[http]>=2021.05.0->evaluate) (6.1.0)
Requirement already satisfied: propcache>=0.2.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!<4.0.0a1->fsspec[http]>=2021.05.0->evaluate) (0.2.0)
Requirement already satisfied: yarl<2.0,>=1.17.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!<4.0.0a1->fsspec[http]>=2021.05.0->evaluate) (1.17.0)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->pandas->evaluate) (1.17.0)
```

```
from transformers import pipeline, set_seed
from datasets import load_dataset, load_from_disk
import matplotlib.pyplot as plt
from datasets import load_dataset
import pandas as pd
from datasets import load_dataset
from evaluate import load
```

```
from transformers import AutoModelForSeq2SeqLM, AutoTokenizer
```

```
import nltk
from nltk.tokenize import sent_tokenize
```

```
from tqdm import tqdm
import torch
```

```
nltk.download("punkt")
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data] Package punkt is already up-to-date!
True
```

```
device = "cuda" if torch.cuda.is_available() else "cpu"
device
```

```
'cuda'
```

```
from transformers import AutoModelForSeq2SeqLM, AutoTokenizer
```

```
model_ckpt = "google/pegasus-cnn_dailymail"
tokenizer = AutoTokenizer.from_pretrained(model_ckpt)
```

```

/usr/local/lib/python3.11/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
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(

```

```
model_pegasus = AutoModelForSeq2SeqLM.from_pretrained(model_ckpt).to(device)
```

```

Some weights of PegasusForConditionalGeneration were not initialized from the model checkpoint at google/pegasus-cnn_dailymail and are n
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.

```

```
dataset_samsum = load_dataset("knkarthick/samsum")
```

```
dataset_samsum
```

```

DatasetDict({
  train: Dataset({
    features: ['id', 'dialogue', 'summary'],
    num_rows: 14732
  })
  validation: Dataset({
    features: ['id', 'dialogue', 'summary'],
    num_rows: 818
  })
  test: Dataset({
    features: ['id', 'dialogue', 'summary'],
    num_rows: 819
  })
})

```

```
dataset_samsum["train"][1]["dialogue"][1]
```

```
'Olivia: Who are you voting for in this election? \nOliver: Liberals as always.\nOlivia: Me too!!\nOliver: Great '
```

```
dataset_samsum["train"][1]["summary"]
```

```
'Olivia and Olivier are voting for liberals in this election. '
```

```
split_lengths = [len(dataset_samsum[split])for split in dataset_samsum]
```

```

print(f"Split lengths: {split_lengths}")
print(f"Features: {dataset_samsum['train'].column_names}")
print("\nDialogue:")

```

```
print(dataset_samsum["test"][1]["dialogue"])
```

```
print("\nSummary:")
```

```
print(dataset_samsum["test"][1]["summary"])
```

```

Split lengths: [14732, 818, 819]
Features: ['id', 'dialogue', 'summary']

```

```

Dialogue:
Eric: MACHINE!
Rob: That's so gr8!
Eric: I know! And shows how Americans see Russian ;)
Rob: And it's really funny!
Eric: I know! I especially like the train part!
Rob: Hahaha! No one talks to the machine like that!
Eric: Is this his only stand-up?
Rob: Idk. I'll check.
Eric: Sure.
Rob: Turns out no! There are some of his stand-ups on youtube.
Eric: Gr8! I'll watch them now!
Rob: Me too!
Eric: MACHINE!
Rob: MACHINE!
Eric: TTYL?
Rob: Sure :)

```

```

Summary:
Eric and Rob are going to watch a stand-up on youtube.

```

```
def preprocess_function(batch):
    dialogues = [str(x) for x in batch["dialogue"]]
    summaries = [str(x) for x in batch["summary"]]

    # Tokenize the dialogues
    model_inputs = tokenizer(
        dialogues,
        max_length=1024,
        truncation=True,
        padding="max_length"
    )

    # Tokenize the summaries (labels)
    with tokenizer.as_target_tokenizer():
        labels = tokenizer(
            summaries,
            max_length=128,
            truncation=True,
            padding="max_length"
        )

    model_inputs["labels"] = labels["input_ids"]
    return model_inputs
```

```
dataset_samsum_pt = dataset_samsum.map(
    preprocess_function,
    batched=True,
    remove_columns=dataset_samsum["train"].column_names
)
```

↻ Map: 100% 818/818 [00:01<00:00, 560.39 examples/s]
 /usr/local/lib/python3.11/dist-packages/transformers/tokenization_utils_base.py:4006: UserWarning: `as_target_tokenizer` is deprecated a
 warnings.warn()

```
dataset_samsum_pt["train"]
```

↻ Dataset({
 features: ['input_ids', 'attention_mask', 'labels'],
 num_rows: 14732
 })

```
from transformers import DataCollatorForSeq2Seq, Seq2SeqTrainingArguments, Seq2SeqTrainer
```

```
data_collator = DataCollatorForSeq2Seq(tokenizer=tokenizer, model=model_pegasus)
```

```
import evaluate
```

```
rouge = evaluate.load("rouge")
```

```
def compute_metrics(eval_pred):
    predictions, labels = eval_pred

    # Decode predictions
    decoded_preds = tokenizer.batch_decode(predictions, skip_special_tokens=True)

    # Replace -100 in the labels (ignored tokens) before decoding
    labels = np.where(labels != -100, labels, tokenizer.pad_token_id)
    decoded_labels = tokenizer.batch_decode(labels, skip_special_tokens=True)

    # Rouge expects newline separated sentences for each summary
    decoded_preds = ["\n".join(pred.strip().split(" ")) for pred in decoded_preds]
    decoded_labels = ["\n".join(label.strip().split(" ")) for label in decoded_labels]


    result = rouge.compute(
        predictions=decoded_preds,
        references=decoded_labels,
        use_stemmer=True
    )

    # Return results as percentages
```


```
return {k: round(v * 100, 2) for k, v in result.items()}
```

```
training_args = Seq2SeqTrainingArguments(
    output_dir="./pegasus-samsum",
    learning_rate=2e-5,
    per_device_train_batch_size=2,
    per_device_eval_batch_size=2,
    weight_decay=0.01,
    save_total_limit=2,
    num_train_epochs=1, # Increase later
    predict_with_generate=True,
    fp16=torch.cuda.is_available(),
    logging_dir='./logs',
    logging_strategy="steps",
    logging_steps=100
)
```

```
trainer = Seq2SeqTrainer(
    model=model_pegasus,
    args=training_args,
    train_dataset=dataset_samsum_pt["test"],
    eval_dataset=dataset_samsum_pt["validation"],
    tokenizer=tokenizer,
    data_collator=data_collator,
    compute_metrics=compute_metrics
)
```

 /tmp/ipython-input-1488543101.py:1: FutureWarning: `tokenizer` is deprecated and will be removed in version 5.0.0 for `Seq2SeqTrainer.__`
 trainer = Seq2SeqTrainer(

```
trainer.train()
```

 **wandb**: Currently logged in as: **gogogoyes785** (**gogogoyes785-tata-consultancy-services**) to <https://api.wandb.ai>. Use `wandb login --relogir`
 Tracking run with wandb version 0.21.1
 Run data is saved locally in /content/wandb/run-20250814_085638-ec1obttv
 Syncing run [decent-sun-2](#) to [Weights & Biases \(docs\)](#)
 View project at <https://wandb.ai/gogogoyes785-tata-consultancy-services/huggingface>
 View run at <https://wandb.ai/gogogoyes785-tata-consultancy-services/huggingface/runs/ec1obttv>
 _____ [410/410 06:12, Epoch 1/1]

Step	Training Loss
------	---------------

100	9.534100
200	8.812000
300	8.595800
400	8.376900

```
/usr/local/lib/python3.11/dist-packages/transformers/modeling_utils.py:3917: UserWarning: Moving the following attributes in the config
warnings.warn(
TrainOutput(global_step=410, training_loss=8.815876137338034, metrics={'train_runtime': 377.1069, 'train_samples_per_second': 2.172,
'train_steps_per_second': 1.087, 'total_flos': 2366471355236352.0, 'train_loss': 8.815876137338034, 'epoch': 1.0})
```

```
trainer.save_model("./pegasus-samsum")
```

```
test_text = dataset_samsum["test"][0]["dialogue"]
inputs = tokenizer(test_text, return_tensors="pt", truncation=True, padding="max_length", max_length=1024)
```

```
if torch.cuda.is_available():
    inputs = {k: v.to("cuda") for k, v in inputs.items()}
    model_pegasus.to("cuda")
```

```
summary_ids = model_pegasus.generate(  
    inputs["input_ids"],  
    attention_mask=inputs["attention_mask"],  
    max_length=128,  
    num_beams=4,  
    length_penalty=2.0,  
    early_stopping=True  
)
```

```
print("\nOriginal Dialogue:\n", test_text)  
print("\nGenerated Summary:\n", tokenizer.decode(summary_ids[0], skip_special_tokens=True))  
print("\nReference Summary:\n", dataset_samsum["test"][0]["summary"])
```



Original Dialogue:

Hannah: Hey, do you have Betty's number?

Amanda: Lemme check

Hannah: <file_gif>

Amanda: Sorry, can't find it.

Amanda: Ask Larry

Amanda: He called her last time we were at the park together

Hannah: I don't know him well

Hannah: <file_gif>

Amanda: Don't be shy, he's very nice

Hannah: If you say so..

Hannah: I'd rather you texted him

Amanda: Just text him 😊

Hannah: Urgh.. Alright

Hannah: Bye

Amanda: Bye bye

Generated Summary:

Amanda is looking for Betty's number, but Hannah can't find it.<n>Hannah would rather she text Betty, because she'd rather she text him

Reference Summary:

Hannah needs Betty's number but Amanda doesn't have it. She needs to contact Larry.
