```
!nvidia-smi
```

```
Sat Jan 24 23:38:47 2026
+-----------------------------------------------------------------------------------------+
| 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   65C    P8             10W /   70W |       0MiB /  15360MiB |      0%      Default |
|                                         |                        |                  N/A |
+-----------------------------------------+------------------------+----------------------+

+-----------------------------------------------------------------------------------------+
| Processes:                                                                              |
|  GPU   GI   CI        PID   Type   Process name                              GPU Memory |
|        ID   ID                                                               Usage      |
|=========================================================================================|
|  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: accelerate in /usr/local/lib/python3.12/dist-packages (1.12.0)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.12/dist-packages (from accelerate) (2.0.2)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.12/dist-packages (from accelerate) (25.0)
Requirement already satisfied: psutil in /usr/local/lib/python3.12/dist-packages (from accelerate) (5.9.5)
Requirement already satisfied: pyyaml in /usr/local/lib/python3.12/dist-packages (from accelerate) (6.0.3)
Requirement already satisfied: torch>=2.0.0 in /usr/local/lib/python3.12/dist-packages (from accelerate) (2.9.0+cu126)
Requirement already satisfied: huggingface_hub>=0.21.0 in /usr/local/lib/python3.12/dist-packages (from accelerate) (0.36.0)
Requirement already satisfied: safetensors>=0.4.3 in /usr/local/lib/python3.12/dist-packages (from accelerate) (0.7.0)
Requirement already satisfied: filelock in /usr/local/lib/python3.12/dist-packages (from huggingface_hub>=0.21.0->accelerate)
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.12/dist-packages (from huggingface_hub>=0.21.0->acc
Requirement already satisfied: requests in /usr/local/lib/python3.12/dist-packages (from huggingface_hub>=0.21.0->accelerate)
Requirement already satisfied: tqdm>=4.42.1 in /usr/local/lib/python3.12/dist-packages (from huggingface_hub>=0.21.0->acceler
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.12/dist-packages (from huggingface_hub>=0
Requirement already satisfied: hf-xet<2.0.0,>=1.1.3 in /usr/local/lib/python3.12/dist-packages (from huggingface_hub>=0.21.0-
Requirement already satisfied: setuptools in /usr/local/lib/python3.12/dist-packages (from torch>=2.0.0->accelerate) (75.2.0)
Requirement already satisfied: sympy>=1.13.3 in /usr/local/lib/python3.12/dist-packages (from torch>=2.0.0->accelerate) (1.14
Requirement already satisfied: networkx>=2.5.1 in /usr/local/lib/python3.12/dist-packages (from torch>=2.0.0->accelerate) (3.
Requirement already satisfied: jinja2 in /usr/local/lib/python3.12/dist-packages (from torch>=2.0.0->accelerate) (3.1.6)
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.6.77 in /usr/local/lib/python3.12/dist-packages (from torch>=2.0.0-
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.6.77 in /usr/local/lib/python3.12/dist-packages (from torch>=2.0.
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.6.80 in /usr/local/lib/python3.12/dist-packages (from torch>=2.0.0-
Requirement already satisfied: nvidia-cudnn-cu12==9.10.2.21 in /usr/local/lib/python3.12/dist-packages (from torch>=2.0.0->ac
Requirement already satisfied: nvidia-cublas-cu12==12.6.4.1 in /usr/local/lib/python3.12/dist-packages (from torch>=2.0.0->ac
Requirement already satisfied: nvidia-cufft-cu12==11.3.0.4 in /usr/local/lib/python3.12/dist-packages (from torch>=2.0.0->acc
Requirement already satisfied: nvidia-curand-cu12==10.3.7.77 in /usr/local/lib/python3.12/dist-packages (from torch>=2.0.0->a
Requirement already satisfied: nvidia-cusolver-cu12==11.7.1.2 in /usr/local/lib/python3.12/dist-packages (from torch>=2.0.0->
Requirement already satisfied: nvidia-cusparse-cu12==12.5.4.2 in /usr/local/lib/python3.12/dist-packages (from torch>=2.0.0->
Requirement already satisfied: nvidia-cusparselt-cu12==0.7.1 in /usr/local/lib/python3.12/dist-packages (from torch>=2.0.0->a
Requirement already satisfied: nvidia-nccl-cu12==2.27.5 in /usr/local/lib/python3.12/dist-packages (from torch>=2.0.0->accele
Requirement already satisfied: nvidia-nvshmem-cu12==3.3.20 in /usr/local/lib/python3.12/dist-packages (from torch>=2.0.0->acc
Requirement already satisfied: nvidia-nvtx-cu12==12.6.77 in /usr/local/lib/python3.12/dist-packages (from torch>=2.0.0->accel
Requirement already satisfied: nvidia-nvjitlink-cu12==12.6.85 in /usr/local/lib/python3.12/dist-packages (from torch>=2.0.0->
Requirement already satisfied: nvidia-cufile-cu12==1.11.1.6 in /usr/local/lib/python3.12/dist-packages (from torch>=2.0.0->ac
Requirement already satisfied: triton==3.5.0 in /usr/local/lib/python3.12/dist-packages (from torch>=2.0.0->accelerate) (3.5.
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.12/dist-packages (from sympy>=1.13.3->torch>=2.0.
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.12/dist-packages (from jinja2->torch>=2.0.0->acceler
Requirement already satisfied: charset_normalizer<4,>=2 in /usr/local/lib/python3.12/dist-packages (from requests->huggingfac
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.12/dist-packages (from requests->huggingface_hub>=0.21.
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.12/dist-packages (from requests->huggingface_hub>
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.12/dist-packages (from requests->huggingface_hub>
Found existing installation: transformers 4.57.6
Uninstalling transformers-4.57.6:
  Successfully uninstalled transformers-4.57.6
Found existing installation: accelerate 1.12.0
Uninstalling accelerate-1.12.0:
  Successfully uninstalled accelerate-1.12.0
Collecting transformers
  Using cached transformers-4.57.6-py3-none-any.whl.metadata (43 kB)
Collecting accelerate
  Using cached accelerate-1.12.0-py3-none-any.whl.metadata (19 kB)
Requirement already satisfied: filelock in /usr/local/lib/python3.12/dist-packages (from transformers) (3.20.3)
Requirement already satisfied: huggingface-hub<1.0,>=0.34.0 in /usr/local/lib/python3.12/dist-packages (from transformers) (0
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.12/dist-packages (from transformers) (2.0.2)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.12/dist-packages (from transformers) (25.0)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.12/dist-packages (from transformers) (6.0.3)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.12/dist-packages (from transformers) (2025.11.3)
Requirement already satisfied: requests in /usr/local/lib/python3.12/dist-packages (from transformers) (2.32.4)
```

```
import torch
import pandas as pd
from tqdm import tqdm
import matplotlib.pyplot as plt
```

```
import matplotlib.pyplot as plt
from transformers import pipeline, set_seed
from datasets import load_dataset, load_from_disk
from transformers import AutoModelForSeq2SeqLM, AutoTokenizer

import nltk
from nltk.tokenize import sent_tokenize
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"
print("Device : ", device)
```

```
Device :  cuda
```

```
# from transformers import PegasusForConditionalGeneration
# # Check for GPU
# device = "cuda" if torch.cuda.is_available() else "cpu"

# model_name = "google/pegasus-xsum"
# model = PegasusForConditionalGeneration.from_pretrained(model_name).to(device)
# tokenizer = AutoTokenizer.from_pretrained(model_name)

# ARTICLE_TO_SUMMARIZE = [
#     "The quick brown fox jumps over the lazy dog.",
#     "Machine learning is a subset of artificial intelligence that focuses on algorithms and statistical models.",
#     "Natural language processing enables computers to understand and interpret human language.",
# ]

# # Note: We added padding=True for batch processing
# inputs = tokenizer(
#     ARTICLE_TO_SUMMARIZE,
#     max_length=1024,
#     return_tensors="pt",
#     truncation=True,
#     padding=True,
# ).to(device)

# # Generate Summary
# summary_ids = model.generate(inputs["input_ids"])
# summaries = tokenizer.batch_decode(
#     summary_ids, skip_special_tokens=True, clean_up_tokenization_spaces=False
# )

# print(summaries[0])
```

```
model = "google/pegasus-cnn_dailymail"
tokenizer = AutoTokenizer.from_pretrained(model)
model_pegasus = AutoModelForSeq2SeqLM.from_pretrained(model).to(device)

print(f"Model loaded successfully on {device}")
```

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

```
# Download the actual raw binary file
!wget https://github.com/SiddhuShkya/Text-Summarizer-With-HF/raw/main/data/summarizer-data.zip

# Now unzip will work
!unzip summarizer-data.zip
```

```
--2026-01-24 23:41:22--  https://github.com/SiddhuShkya/Text-Summarizer-With-HF/raw/main/data/summarizer-data.zip
Resolving github.com (github.com)... 20.205.243.166
Connecting to github.com (github.com)|20.205.243.166|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://raw.githubusercontent.com/SiddhuShkya/Text-Summarizer-With-HF/main/data/summarizer-data.zip [following]
--2026-01-24 23:41:22--  https://raw.githubusercontent.com/SiddhuShkya/Text-Summarizer-With-HF/main/data/summarizer-data.zip
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.108.133, 185.199.109.133, 185.199.110.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.108.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 4004830 (3.8M) [application/zip]
Saving to: ‘summarizer-data.zip.1’

summarizer-data.zip 100%[===================>]   3.82M  --.-KB/s    in 0.02s

2026-01-24 23:41:23 (224 MB/s) - ‘summarizer-data.zip.1’ saved [4004830/4004830]
```

```
Archive:  summarizer-data.zip
replace summarizer-data/validation.csv? [y]es, [n]o, [A]ll, [N]one, [r]ename: A
  inflating: summarizer-data/validation.csv
  inflating: summarizer-data/train.csv
  inflating: summarizer-data/test.csv
```

```
train_df = pd.read_csv('./summarizer-data/train.csv')
test_df = pd.read_csv('./summarizer-data/test.csv')
val_df = pd.read_csv('./summarizer-data/validation.csv')

print("Features in the dataset: ", train_df.columns.tolist())
print("=" * 70)
print("Number of samples in each dataset:")
print("Train data samples: ", len(train_df))A
print("Test data samples: ", len(test_df))
print("Validation data samples: ", len(val_df))
```

```
Features in the dataset:  ['id', 'dialogue', 'summary']
======================================================================
Number of samples in each dataset:
Train data samples:  14731
Test data samples:  819
Validation data samples:  818
```

```
print(train_df["dialogue"][0])
print("\nSummary: ", train_df["summary"][0])
```

```
Amanda: I baked  cookies. Do you want some?
Jerry: Sure!
Amanda: I'll bring you tomorrow :-)

Summary:  Amanda baked cookies and will bring Jerry some tomorrow.
```

```
def convert_examples_to_features(example_batch):
    input_encodings = tokenizer(
        example_batch["dialogue"],
        max_length=512,
        truncation=True,
    )
    with tokenizer.as_target_tokenizer():
        target_encodings = tokenizer(
            example_batch["summary"],
            max_length=128,
            truncation=True,
        )
    return {
        'input_ids': input_encodings['input_ids'],
        'attention_mask': input_encodings['attention_mask'],
        'labels': target_encodings['input_ids']
    }
```

```
from datasets import Dataset

# Convert dataframes → Dataset
train_dataset = Dataset.from_pandas(train_df)
test_dataset = Dataset.from_pandas(test_df)
val_dataset = Dataset.from_pandas(val_df)
# Apply the function with .map()
train_dataset = train_dataset.map(convert_examples_to_features, batched=True)
test_dataset = test_dataset.map(convert_examples_to_features, batched=True)
val_dataset = val_dataset.map(convert_examples_to_features, batched=True)
```

```
Map: 100%  14731/14731 [00:19<00:00, 770.80 examples/s]
/usr/local/lib/python3.12/dist-packages/transformers/tokenization_utils_base.py:4174: UserWarning: `as_target_tokenizer` is de
  warnings.warn(
Map: 100%  819/819 [00:00<00:00, 1138.14 examples/s]
Map: 100%  818/818 [00:00<00:00, 1651.21 examples/s]
```

```
train_dataset
```

```
Dataset({
    features: ['id', 'dialogue', 'summary', 'input_ids', 'attention_mask', 'labels'],
    num_rows: 14731
})
```

```
test_dataset
```

```
Dataset({
    features: ['id', 'dialogue', 'summary', 'input_ids', 'attention_mask', 'labels'],
    num_rows: 819
})
```

```
val_dataset
```

```
Dataset({
    features: ['id', 'dialogue', 'summary', 'input_ids', 'attention_mask', 'labels'],
    num_rows: 818
})
```

```
print("Train Dataset:\n", train_dataset)
print("Test Dataset:\n", test_dataset)
print("Val Dataset:\n", val_dataset)
```

```
Train Dataset:
 Dataset({
    features: ['id', 'dialogue', 'summary', 'input_ids', 'attention_mask', 'labels'],
    num_rows: 14731
})
Test Dataset:
 Dataset({
    features: ['id', 'dialogue', 'summary', 'input_ids', 'attention_mask', 'labels'],
    num_rows: 819
})
Val Dataset:
 Dataset({
    features: ['id', 'dialogue', 'summary', 'input_ids', 'attention_mask', 'labels'],
    num_rows: 818
})
```

```
from transformers import DataCollatorForSeq2Seq

seq2seq_data_collator = DataCollatorForSeq2Seq(tokenizer, model=model_pegasus)
```

```
from transformers import TrainingArguments, Trainer

training_args = TrainingArguments(
    output_dir="pegasus-finetuned",
    num_train_epochs=1,
    warmup_steps=500,
    per_device_train_batch_size=1,
    per_device_eval_batch_size=1,
    weight_decay=0.01,
    logging_steps=10,
    eval_strategy="steps",
    eval_steps=500,
    save_steps=100000,
    gradient_accumulation_steps=16,
)
```

```
trainer = Trainer(
    model=model_pegasus,
    args=training_args,
    processing_class=tokenizer,
    data_collator=seq2seq_data_collator,
    train_dataset=test_dataset,
    eval_dataset=val_dataset,
)
```

```
trainer.train()
```

```
The tokenizer has new PAD/BOS/EOS tokens that differ from the model config and generation config. The model config and generat
wandb: (1) Create a W&B account
wandb: (2) Use an existing W&B account
wandb: (3) Don't visualize my results
wandb: Enter your choice: 3
wandb: You chose "Don't visualize my results"
wandb: Using W&B in offline mode.
wandb: W&B API key is configured. Use `wandb login --relogin` to force relogin
Tracking run with wandb version 0.24.0
W&B syncing is set to `offline` in this directory. Run `wandb online` or set WANDB_MODE=online to enable cloud syncing.
Run data is saved locally in /content/wandb/offline-run-20260124_234739-z2r69h29
```

```
[52/52 08:46, Epoch 1/1]
```

Step	Training Loss	Validation Loss

```
/usr/local/lib/python3.12/dist-packages/transformers/modeling_utils.py:3918: UserWarning: Moving the following attributes in t
  warnings.warn(
TrainOutput(global_step=52, training_loss=2.9682337779265184, metrics={'train_runtime': 549.9829, 'train_samples_per_second':
1.489, 'train_steps_per_second': 0.095, 'total_flos': 313176745058304.0, 'train_loss': 2.9682337779265184, 'epoch': 1.0})
```

```
def generate_batch_sized_chunks(list_of_elements, batch_size):
    """split the dataset into smaller batches"""
    for i in range(0, len(list_of_elements), batch_size):
        yield list_of_elements[i : i + batch_size]
```

```
def calculate_metric_on_test_ds(
    dataset,
    metric,
    model,
    tokenizer,
    batch_size=1,
    device=device,
    column_text="article",
    column_summary="highlights",
):
    model.eval()  # 🔴 add

    article_batches = list(generate_batch_sized_chunks(dataset[column_text], batch_size))
    target_batches = list(generate_batch_sized_chunks(dataset[column_summary], batch_size))

    with torch.no_grad():  # 🔴 add
        for article_batch, target_batch in zip(article_batches, target_batches):

            inputs = tokenizer(
                article_batch,
                max_length=256,          # 🔴 reduce
                truncation=True,
                padding="max_length",
                return_tensors="pt",
            )

            summaries = model.generate(
                input_ids=inputs["input_ids"].to(device),
                attention_mask=inputs["attention_mask"].to(device),
                max_new_tokens=128,
                num_beams=1,             # 🔴 critical
                do_sample=False,
                use_cache=True,
            )

            decoded_summaries = tokenizer.batch_decode(
                summaries, skip_special_tokens=True
            )

            metric.add_batch(
                predictions=decoded_summaries,
                references=target_batch,
            )

    return metric.compute()
```

```
!pip install evaluate
```

```
Requirement already satisfied: evaluate in /usr/local/lib/python3.12/dist-packages (0.4.6)
Requirement already satisfied: datasets>=2.0.0 in /usr/local/lib/python3.12/dist-packages (from evaluate) (4.0.0)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.12/dist-packages (from evaluate) (2.0.2)
Requirement already satisfied: dill in /usr/local/lib/python3.12/dist-packages (from evaluate) (0.3.8)
Requirement already satisfied: pandas in /usr/local/lib/python3.12/dist-packages (from evaluate) (2.2.2)
Requirement already satisfied: requests>=2.19.0 in /usr/local/lib/python3.12/dist-packages (from evaluate) (2.32.4)
Requirement already satisfied: tqdm>=4.62.1 in /usr/local/lib/python3.12/dist-packages (from evaluate) (4.67.1)
Requirement already satisfied: xxhash in /usr/local/lib/python3.12/dist-packages (from evaluate) (3.6.0)
Requirement already satisfied: multiprocess in /usr/local/lib/python3.12/dist-packages (from evaluate) (0.70.16)
Requirement already satisfied: fsspec>=2021.05.0 in /usr/local/lib/python3.12/dist-packages (from fsspec[http]>=2021.05.0->eva
Requirement already satisfied: huggingface-hub>=0.7.0 in /usr/local/lib/python3.12/dist-packages (from evaluate) (0.36.0)
Requirement already satisfied: packaging in /usr/local/lib/python3.12/dist-packages (from evaluate) (25.0)
Requirement already satisfied: filelock in /usr/local/lib/python3.12/dist-packages (from datasets>=2.0.0->evaluate) (3.20.3)
Requirement already satisfied: pyarrow>=15.0.0 in /usr/local/lib/python3.12/dist-packages (from datasets>=2.0.0->evaluate) (18
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.12/dist-packages (from datasets>=2.0.0->evaluate) (6.0.3)
Requirement already satisfied: aiohttp!=4.0.0a0,!=4.0.0a1 in /usr/local/lib/python3.12/dist-packages (from fsspec[http]>=2021.
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.12/dist-packages (from huggingface-hub>=0.
Requirement already satisfied: hf-xet<2.0.0,>=1.1.3 in /usr/local/lib/python3.12/dist-packages (from huggingface-hub>=0.7.0->e
Requirement already satisfied: charset_normalizer<4,>=2 in /usr/local/lib/python3.12/dist-packages (from requests>=2.19.0->eva
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.12/dist-packages (from requests>=2.19.0->evaluate) (3.11
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.12/dist-packages (from requests>=2.19.0->evaluate)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.12/dist-packages (from requests>=2.19.0->evaluate)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.12/dist-packages (from pandas->evaluate) (2.9.
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.12/dist-packages (from pandas->evaluate) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.12/dist-packages (from pandas->evaluate) (2025.3)
Requirement already satisfied: aiohappyeyeballs>=2.5.0 in /usr/local/lib/python3.12/dist-packages (from aiohttp!=4.0.0a0,!=4.0
Requirement already satisfied: aiosignal>=1.4.0 in /usr/local/lib/python3.12/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1->f
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.12/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1->fssp
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.12/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1->
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.12/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1
Requirement already satisfied: propcache>=0.2.0 in /usr/local/lib/python3.12/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1->f
Requirement already satisfied: yarl<2.0,>=1.17.0 in /usr/local/lib/python3.12/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1->
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.12/dist-packages (from python-dateutil>=2.8.2->pandas->evalu
```

```
import evaluate

rouge_metric = evaluate.load("rouge")
rouge_names = ["rouge1", "rouge2", "rougeL", "rougeLsum"]
```

```
score = calculate_metric_on_test_ds(
    dataset=val_dataset[0:10],
    metric=rouge_metric,
    model=trainer.model,
    tokenizer=tokenizer,
    column_text="dialogue",
    column_summary="summary"
)

rouge_dict = {name: score[name] for name in rouge_names}
pd.DataFrame(rouge_dict, index=[f'pegasus-finetuned'])
```

```
The following generation flags are not valid and may be ignored: ['length_penalty']. Set `TRANSFORMERS_VERBOSITY=info` for mor
```

	rouge1	rouge2	rougeL	rougeLsum
pegasus-finetuned	0.305614	0.1115	0.247998	0.248458

```
model_pegasus.save_pretrained("pegasus-model")
tokenizer.save_pretrained("pegasus-tokenizer")
```

```
('pegasus-tokenizer/tokenizer_config.json',
 'pegasus-tokenizer/special_tokens_map.json',
 'pegasus-tokenizer/spiece.model',
 'pegasus-tokenizer/added_tokens.json',
 'pegasus-tokenizer/tokenizer.json')
```

```
model_path = "./pegasus-model"
tokenizer_path = "./pegasus-tokenizer"
```

```
from transformers import PegasusForConditionalGeneration, PegasusTokenizer

# Load with local_files_only=True to prevent it from checking the internet
tokenizer = PegasusTokenizer.from_pretrained(tokenizer_path)
model = PegasusForConditionalGeneration.from_pretrained(model_path)
```

```
gen_kwargs = {
    "length_penalty": 0.8,
    "num_beams": 8,
    "max_length": 128,
}
```

```
sample_text = train_dataset[0]['dialogue']
reference = train_dataset[0]['summary']
pipe = pipeline("summarization", model="pegasus-model", tokenizer=tokenizer)
```

```
Device set to use cuda:0
```

```
print("Dialogue : \n", sample_text)
print("\nReference Summary : \n", reference)
print("\nModel Summary : \n", pipe(sample_text, **gen_kwargs)[0]["summary_text"])
```

```
Your max_length is set to 128, but your input_length is only 25. Since this is a summarization task, where outputs shorter tha
Dialogue :
 Amanda: I baked  cookies. Do you want some?
Jerry: Sure!
Amanda: I'll bring you tomorrow :-)

Reference Summary :
 Amanda baked cookies and will bring Jerry some tomorrow.

Model Summary :
 Amanda: I'll bring you tomorrow :-) <n> Amanda: I baked cookies. Do you want some? <n> Amanda: I'll bring you tomorrow :-)
```

```
# 1. Grab your text from the dataset
sample_text = train_dataset[0]['dialogue']
reference = train_dataset[0]['summary']

# 2. Tokenize the input dialogue
# We use truncation=True to ensure it fits within the model's 1024 token limit
inputs = tokenizer(sample_text, truncation=True, padding="longest", return_tensors="pt")

# 3. Generate the summary
# The model produces token IDs
summary_ids = model.generate(
    inputs["input_ids"],
    max_length=128,
    num_beams=4,
```

```
    length_penalty=2.0,
    early_stopping=True
)

# 4. Decode the IDs back into a string
decoded_summary = tokenizer.decode(summary_ids[0], skip_special_tokens=True)

# 5. Compare the results
print("--- DIALOGUE ---")
print(sample_text)
print("\n--- REFERENCE SUMMARY (Ground Truth) ---")
print(reference)
print("\n--- MODEL GENERATED SUMMARY ---")
print(decoded_summary)
```

```
--- DIALOGUE ---
Amanda: I baked  cookies. Do you want some?
Jerry: Sure!
Amanda: I'll bring you tomorrow :-)

--- REFERENCE SUMMARY (Ground Truth) ---
Amanda baked cookies and will bring Jerry some tomorrow.

--- MODEL GENERATED SUMMARY ---
Amanda: I baked cookies. Do you want some? Jerry: Sure! <n> Amanda: . <n> I'll bring you tomorrow :-), Jerry
```

```
# Load the pipeline
pipe = pipeline("summarization", model=model, tokenizer=tokenizer)

# Use it
result = pipe(sample_text, max_length=25, min_length=10, do_sample=False)
print(result[0]['summary_text'])
```

```
Device set to use cuda:0
Amanda: I baked cookies. <n> Amanda: I'll bring you tomorrow :-)
```

Start coding or generate with AI.