Fine-Tune a Generative AI Model for Dialogue Summarization

In this notebook, you will fine-tune an existing LLM from Hugging Face for enhanced dialogue summarization. You will use the FLAN-T5 model, which provides a high quality instruction tuned model and can summarize text out of the box. To improve the inferences, you will explore a full fine-tuning approach and evaluate the results with ROUGE metrics. Then you will perform Parameter Efficient Fine-Tuning (PEFT), evaluate the resulting model and see that the benefits of PEFT outweigh the slightly-lower performance metrics.

Table of Contents

1 - Load Required Dependencies, Dataset and LLM (5 points)

1.1 - Set up Required Dependencies (1 point)

Now install the required packages for the LLM and datasets.

The next cell may take a few minutes to run. Please be patient.

Ignore the warnings and errors, along with the note about restarting the kernel at the end.

In [27]:
```
# Installing required dependencies
!pip install datasets torch transformers evaluate rouge_score loralib peft wandb
```

```
huggingface/tokenizers: The current process just got forked, after parallelism has already been used. Disabling paralle
lism to avoid deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true | false)
```

```
Requirement already satisfied: datasets in /usr/local/lib/python3.11/dist-packages (4.4.1)
Requirement already satisfied: torch in /usr/local/lib/python3.11/dist-packages (2.6.0+cu124)
Requirement already satisfied: transformers in /usr/local/lib/python3.11/dist-packages (4.53.3)
Requirement already satisfied: evaluate in /usr/local/lib/python3.11/dist-packages (0.4.6)
Requirement already satisfied: rouge_score in /usr/local/lib/python3.11/dist-packages (0.1.2)
Requirement already satisfied: loralib in /usr/local/lib/python3.11/dist-packages (0.1.2)
Requirement already satisfied: peft in /usr/local/lib/python3.11/dist-packages (0.16.0)
Requirement already satisfied: wandb in /usr/local/lib/python3.11/dist-packages (0.21.0)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from datasets) (3.20.0)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-packages (from datasets) (1.26.4)
Requirement already satisfied: pyarrow>=21.0.0 in /usr/local/lib/python3.11/dist-packages (from datasets) (22.0.0)
Requirement already satisfied: dill<0.4.1,>=0.3.0 in /usr/local/lib/python3.11/dist-packages (from datasets) (0.4.0)
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages (from datasets) (2.2.3)
Requirement already satisfied: requests>=2.32.2 in /usr/local/lib/python3.11/dist-packages (from datasets) (2.32.5)
Requirement already satisfied: httpx<1.0.0 in /usr/local/lib/python3.11/dist-packages (from datasets) (0.28.1)
Requirement already satisfied: tqdm>=4.66.3 in /usr/local/lib/python3.11/dist-packages (from datasets) (4.67.1)
Requirement already satisfied: xxhash in /usr/local/lib/python3.11/dist-packages (from datasets) (3.6.0)
Requirement already satisfied: multiprocess<0.70.19 in /usr/local/lib/python3.11/dist-packages (from datasets) (0.70.1
8)
Requirement already satisfied: fsspec<=2025.10.0,>=2023.1.0 in /usr/local/lib/python3.11/dist-packages (from fsspec[htt
p]<=2025.10.0,>=2023.1.0->datasets) (2025.10.0)
Requirement already satisfied: huggingface-hub<2.0,>=0.25.0 in /usr/local/lib/python3.11/dist-packages (from datasets)
(0.36.0)
Requirement already satisfied: packaging in /usr/local/lib/python3.11/dist-packages (from datasets) (25.0)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-packages (from datasets) (6.0.3)
Requirement already satisfied: typing-extensions>=4.10.0 in /usr/local/lib/python3.11/dist-packages (from torch) (4.15.
0)
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-packages (from torch) (3.5)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.11/dist-packages (from torch) (3.1.6)
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch)
(12.4.127)
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torc
h) (12.4.127)
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch)
(12.4.127)
Requirement already satisfied: nvidia-cudnn-cu12==9.1.0.70 in /usr/local/lib/python3.11/dist-packages (from torch) (9.
1.0.70)
Requirement already satisfied: nvidia-cublas-cu12==12.4.5.8 in /usr/local/lib/python3.11/dist-packages (from torch) (1
2.4.5.8)
Requirement already satisfied: nvidia-cufft-cu12==11.2.1.3 in /usr/local/lib/python3.11/dist-packages (from torch) (11.
2.1.3)
Requirement already satisfied: nvidia-curand-cu12==10.3.5.147 in /usr/local/lib/python3.11/dist-packages (from torch)
(10.3.5.147)
Requirement already satisfied: nvidia-cusolver-cu12==11.6.1.9 in /usr/local/lib/python3.11/dist-packages (from torch)
(11.6.1.9)
Requirement already satisfied: nvidia-cusparse-cu12==12.3.1.170 in /usr/local/lib/python3.11/dist-packages (from torch)
(12.3.1.170)
Requirement already satisfied: nvidia-cusparselt-cu12==0.6.2 in /usr/local/lib/python3.11/dist-packages (from torch)
```

```
(0.6.2)
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in /usr/local/lib/python3.11/dist-packages (from torch) (2.21.
5)
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch) (12.
4.127)
Requirement already satisfied: nvidia-nvjitlink-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch)
(12.4.127)
Requirement already satisfied: triton==3.2.0 in /usr/local/lib/python3.11/dist-packages (from torch) (3.2.0)
Requirement already satisfied: sympy==1.13.1 in /usr/local/lib/python3.11/dist-packages (from torch) (1.13.1)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.11/dist-packages (from sympy==1.13.1->torc
h) (1.3.0)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.11/dist-packages (from transformers) (2025.1
1.3)
Requirement already satisfied: tokenizers<0.22,>=0.21 in /usr/local/lib/python3.11/dist-packages (from transformers)
(0.21.2)
Requirement already satisfied: safetensors>=0.4.3 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.5.
3)
Requirement already satisfied: absl-py in /usr/local/lib/python3.11/dist-packages (from rouge_score) (1.4.0)
Requirement already satisfied: nltk in /usr/local/lib/python3.11/dist-packages (from rouge_score) (3.9.2)
Requirement already satisfied: six>=1.14.0 in /usr/local/lib/python3.11/dist-packages (from rouge_score) (1.17.0)
Requirement already satisfied: psutil in /usr/local/lib/python3.11/dist-packages (from peft) (7.1.3)
Requirement already satisfied: accelerate>=0.21.0 in /usr/local/lib/python3.11/dist-packages (from peft) (1.9.0)
Requirement already satisfied: click!=8.0.0,>=7.1 in /usr/local/lib/python3.11/dist-packages (from wandb) (8.3.0)
Requirement already satisfied: gitpython!=3.1.29,>=1.0.0 in /usr/local/lib/python3.11/dist-packages (from wandb) (3.1.4
5)
Requirement already satisfied: platformdirs in /usr/local/lib/python3.11/dist-packages (from wandb) (4.5.0)
Requirement already satisfied: protobuf!=4.21.0,!=5.28.0,<7,>=3.19.0 in /usr/local/lib/python3.11/dist-packages (from w
andb) (6.33.0)
Requirement already satisfied: pydantic<3 in /usr/local/lib/python3.11/dist-packages (from wandb) (2.12.4)
Requirement already satisfied: sentry-sdk>=2.0.0 in /usr/local/lib/python3.11/dist-packages (from wandb) (2.33.2)
Requirement already satisfied: aiohttp!=4.0.0a0,!=4.0.0a1 in /usr/local/lib/python3.11/dist-packages (from fsspec[http]
<=2025.10.0,>=2023.1.0->datasets) (3.13.2)
Requirement already satisfied: gitdb<5,>=4.0.1 in /usr/local/lib/python3.11/dist-packages (from gitpython!=3.1.29,>=1.
0.0->wandb) (4.0.12)
Requirement already satisfied: anyio in /usr/local/lib/python3.11/dist-packages (from httpx<1.0.0->datasets) (4.11.0)
Requirement already satisfied: certifi in /usr/local/lib/python3.11/dist-packages (from httpx<1.0.0->datasets) (2025.1
0.5)
Requirement already satisfied: httpcore==1.* in /usr/local/lib/python3.11/dist-packages (from httpx<1.0.0->datasets)
(1.0.9)
Requirement already satisfied: idna in /usr/local/lib/python3.11/dist-packages (from httpx<1.0.0->datasets) (3.11)
Requirement already satisfied: h11>=0.16 in /usr/local/lib/python3.11/dist-packages (from httpcore==1.*->httpx<1.0.0->d
atasets) (0.16.0)
Requirement already satisfied: hf-xet<2.0.0,>=1.1.3 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<2.
0,>=0.25.0->datasets) (1.2.0)
Requirement already satisfied: mkl_fft in /usr/local/lib/python3.11/dist-packages (from numpy>=1.17->datasets) (1.3.8)
Requirement already satisfied: mkl_random in /usr/local/lib/python3.11/dist-packages (from numpy>=1.17->datasets) (1.2.
4)
Requirement already satisfied: mkl_umath in /usr/local/lib/python3.11/dist-packages (from numpy>=1.17->datasets) (0.1.
```

```
1)
Requirement already satisfied: mkl in /usr/local/lib/python3.11/dist-packages (from numpy>=1.17->datasets) (2025.3.0)
Requirement already satisfied: tbb4py in /usr/local/lib/python3.11/dist-packages (from numpy>=1.17->datasets) (2022.3.
0)
Requirement already satisfied: mkl-service in /usr/local/lib/python3.11/dist-packages (from numpy>=1.17->datasets) (2.
4.1)
Requirement already satisfied: annotated-types>=0.6.0 in /usr/local/lib/python3.11/dist-packages (from pydantic<3->wand
b) (0.7.0)
Requirement already satisfied: pydantic-core==2.41.5 in /usr/local/lib/python3.11/dist-packages (from pydantic<3->wand
b) (2.41.5)
Requirement already satisfied: typing-inspection>=0.4.2 in /usr/local/lib/python3.11/dist-packages (from pydantic<3->wa
ndb) (0.4.2)
Requirement already satisfied: charset_normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests>=2.3
2.2->datasets) (3.4.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->da
tasets) (2.5.0)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from jinja2->torch) (3.0.3)
Requirement already satisfied: joblib in /usr/local/lib/python3.11/dist-packages (from nltk->rouge_score) (1.5.2)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas->dataset
s) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from pandas->datasets) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas->datasets) (2025.
2)
Requirement already satisfied: aiohappyeyeballs>=2.5.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a
0,!=4.0.0a1->fsspec[http]<=2025.10.0,>=2023.1.0->datasets) (2.6.1)
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]<=2025.10.0,>=2023.1.0->datasets) (1.4.0)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a
1->fsspec[http]<=2025.10.0,>=2023.1.0->datasets) (25.4.0)
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]<=2025.10.0,>=2023.1.0->datasets) (1.8.0)
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]<=2025.10.0,>=2023.1.0->datasets) (6.7.0)
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]<=2025.10.0,>=2023.1.0->datasets) (0.4.1)
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]<=2025.10.0,>=2023.1.0->datasets) (1.22.0)
Requirement already satisfied: smmap<6,>=3.0.1 in /usr/local/lib/python3.11/dist-packages (from gitdb<5,>=4.0.1->gitpyt
hon!=3.1.29,>=1.0.0->wandb) (5.0.2)
Requirement already satisfied: sniffio>=1.1 in /usr/local/lib/python3.11/dist-packages (from anyio->httpx<1.0.0->datase
ts) (1.3.1)
Requirement already satisfied: onemkl-license==2025.3.0 in /usr/local/lib/python3.11/dist-packages (from mkl->numpy>=1.
17->datasets) (2025.3.0)
Requirement already satisfied: intel-openmp<2026,>=2024 in /usr/local/lib/python3.11/dist-packages (from mkl->numpy>=1.
17->datasets) (2024.2.0)
Requirement already satisfied: tbb==2022.* in /usr/local/lib/python3.11/dist-packages (from mkl->numpy>=1.17->datasets)
(2022.3.0)
Requirement already satisfied: tcmlib==1.* in /usr/local/lib/python3.11/dist-packages (from tbb==2022.*->mkl->numpy>=1.
```

```
17->datasets) (1.4.0)
Requirement already satisfied: intel-cmplr-lib-rt in /usr/local/lib/python3.11/dist-packages (from mkl_umath->numpy>=1.
17->datasets) (2024.2.0)
Requirement already satisfied: intel-cmplr-lib-ur==2024.2.0 in /usr/local/lib/python3.11/dist-packages (from intel-open
mp<2026,>=2024->mkl->numpy>=1.17->datasets) (2024.2.0)
```

Import the necessary components. Some of them are new for this week, they will be discussed later in the notebook.

In [82]:
```python
# Importing necessary components
from datasets import load_dataset
from transformers import AutoModelForSeq2SeqLM, AutoTokenizer
from transformers import Trainer, TrainingArguments
import evaluate
import torch
import time
import wandb
import pandas as pd
import numpy as np
```

1.2 - Load Dataset and LLM (2 points)

You are going to continue experimenting with the DialogSum Hugging Face dataset. It contains 10,000+ dialogues with the corresponding manually labeled summaries and topics.

In [4]:
```python
# Loading Dataset
huggingface_dataset_name = "knkarthick/dialogsum"
dataset = load_dataset(huggingface_dataset_name)
dataset
```

```
README.md: 0.00B [00:00, ?B/s]
train.csv:   0%|          | 0.00/11.3M [00:00<?, ?B/s]
validation.csv: 0.00B [00:00, ?B/s]
test.csv: 0.00B [00:00, ?B/s]
Generating train split:   0%|          | 0/12460 [00:00<?, ? examples/s]
Generating validation split:   0%|          | 0/500 [00:00<?, ? examples/s]
Generating test split:   0%|          | 0/1500 [00:00<?, ? examples/s]
```

Out[4]:
```
DatasetDict({
    train: Dataset({
        features: ['id', 'dialogue', 'summary', 'topic'],
        num_rows: 12460
    })
    validation: Dataset({
        features: ['id', 'dialogue', 'summary', 'topic'],
        num_rows: 500
    })
    test: Dataset({
        features: ['id', 'dialogue', 'summary', 'topic'],
        num_rows: 1500
    })
})
```

Load the pre-trained FLAN-T5 model and its tokenizer directly from HuggingFace. Notice that you will be using the small version of FLAN-T5. Setting `torch_dtype=torch.bfloat16` specifies the memory type to be used by this model.

In [6]:
```
# Loading pre-trained FLAN-T5 small model and its tokenizer directly from HuggingFace
model_name = "google/flan-t5-small"
original_model = AutoModelForSeq2SeqLM.from_pretrained(model_name, torch_dtype=torch.bfloat16)
tokenizer = AutoTokenizer.from_pretrained(model_name)
```

```
config.json: 0.00B [00:00, ?B/s]
model.safetensors:   0%|          | 0.00/308M [00:00<?, ?B/s]
generation_config.json:   0%|          | 0.00/147 [00:00<?, ?B/s]
tokenizer_config.json: 0.00B [00:00, ?B/s]
spiece.model:   0%|          | 0.00/792k [00:00<?, ?B/s]
tokenizer.json: 0.00B [00:00, ?B/s]
special_tokens_map.json: 0.00B [00:00, ?B/s]
```

It is possible to pull out the number of model parameters and find out how many of them are trainable. The following function can be used to do that, at this stage, you do not need to go into details of it.

In [7]:
```
# Function to print number of parameters in model and number of parameters in model that are trainable
def print_number_of_trainable_model_parameters(model):
    trainable_model_params = sum(p.numel() for p in model.parameters() if p.requires_grad)
    all_model_params = sum(p.numel() for p in model.parameters())
    print("Total Number of Parameters: " + str(all_model_params))
    print("Total Number of Trainable Parameters: " + str(trainable_model_params))

print_number_of_trainable_model_parameters(original_model)
```

```
Total Number of Parameters: 76961152
Total Number of Trainable Parameters: 76961152
```

1.3 - Test the Model with Zero Shot Inferencing (2 Points)

Test the model with the zero shot inferencing. You can see that the model struggles to summarize the dialogue compared to the baseline summary, but it does pull out some important information from the text which indicates the model can be fine-tuned to the task at hand.

In [9]:
```
# Get random dialogue and it's summary from the test dataset
index = 200
dialogue = dataset['test'][index]['dialogue']
summary = dataset['test'][index]['summary']

# Create prompt for zero shot inferencing
prompt = "Summarize the following dialogue.\n\n" + dialogue + "\n\nSummary:"

# Tokenize prompt
inputs = tokenizer(prompt, return_tensors="pt")

# Get model to generate a response to input prompt
response = original_model.generate(**inputs)

# Decode model's response
output = tokenizer.decode(
    response[0],
    skip_special_tokens=True
)

# Compare zero shot inferencing output of our model to baseline human summary
dash_line = '-'.join('' for x in range(100))
print(dash_line)
print(f'INPUT PROMPT:\n{prompt}')
print(dash_line)
print(f'BASELINE HUMAN SUMMARY:\n{summary}\n')
print(dash_line)
print(f'MODEL GENERATION - ZERO SHOT:\n{output}')
```

```
---------------------------------------------------------------------------------------------------
INPUT PROMPT:
Summarize the following dialogue.

#Person1#: Have you considered upgrading your system?
#Person2#: Yes, but I'm not sure what exactly I would need.
#Person1#: You could consider adding a painting program to your software. It would allow you to make up your own flyers
and banners for advertising.
#Person2#: That would be a definite bonus.
#Person1#: You might also want to upgrade your hardware because it is pretty outdated now.
#Person2#: How can we do that?
#Person1#: You'd probably need a faster processor, to begin with. And you also need a more powerful hard disc, more mem
ory and a faster modem. Do you have a CD-ROM drive?
#Person2#: No.
#Person1#: Then you might want to add a CD-ROM drive too, because most new software programs are coming out on Cds.
#Person2#: That sounds great. Thanks.

Summary:
---------------------------------------------------------------------------------------------------
BASELINE HUMAN SUMMARY:
#Person1# teaches #Person2# how to upgrade software and hardware in #Person2#'s system.

---------------------------------------------------------------------------------------------------
MODEL GENERATION - ZERO SHOT:
You'd like to add a CD-ROM drive to your software.
```

1.3

Compare the generated summary with the human baseline using qualitative analysis

As seen in the above output, the zero-shot generated summary does capture the information that appears towards the end of the conversation, but fails to capture the main purpose of the conversation as a whole like the baseline human summary does. Also, the zero-shot generated summary thinks Person 2 is me instead of a third party conversation between two random people.

2 - Perform Full Fine-Tuning (10 points)

2.1 - Preprocess the Dialog-Summary Dataset (2 points)

You need to convert the dialog-summary (prompt-response) pairs into explicit instructions for the LLM. Prepend an instruction to the start of the dialog with `Summarize the following conversation` and to the start of the summary with `Summary` as follows:

Training prompt (dialogue):

```
Summarize the following conversation.

    Chris: This is his part of the conversation.
    Antje: This is her part of the conversation.

Summary:
```

Training response (summary):

```
Both Chris and Antje participated in the conversation.
```

Then preprocess the prompt-response dataset into tokens and pull out their `input_ids` (1 per token).

In [57]:
```
def tokenize_function(examples):
    # Create prompt for every example in batch
    inputs = ["Summarize the following conversation.\n\n" + dialogue + "\n\nSummary:"
              for dialogue in examples['dialogue']]

    # Tokenize the prompts
    model_inputs = tokenizer(inputs, max_length=512, truncation=True, padding="max_length")
    # Tokenize the labels(baseline human summaries)
    labels = tokenizer(examples['summary'], max_length=128, truncation=True, padding="max_length")
    model_inputs["labels"] = labels["input_ids"]
    # Return tokenized model inputs
    return model_inputs

# The dataset actually contains 3 diff splits: train, validation, test.
# The tokenize_function code is handling all data across all splits in batches.
tokenized_datasets = dataset.map(tokenize_function, batched=True, remove_columns=dataset['train'].column_names)
```

```
Map:   0%|          | 0/12460 [00:00<?, ? examples/s]
```

To save some time in the lab, you will subsample the dataset:

In [58]:
```
# Create a subsampled version of the dataset for efficient training
tokenized_datasets = tokenized_datasets.filter(lambda example, index: index % 3 == 0, with_indices=True)
```

```
Filter:   0%|          | 0/12460 [00:00<?, ? examples/s]
Filter:   0%|          | 0/500 [00:00<?, ? examples/s]
Filter:   0%|          | 0/1500 [00:00<?, ? examples/s]
```

Check the shapes of all three parts of the dataset:

In [59]:
```
print(f"Shapes of the datasets:")
print(f"Training: {tokenized_datasets['train'].shape}")
```

```
print(f"Validation: {tokenized_datasets['validation'].shape}")
print(f"Test: {tokenized_datasets['test'].shape}")

print(tokenized_datasets)
```

```
Shapes of the datasets:
Training: (4154, 3)
Validation: (167, 3)
Test: (500, 3)
DatasetDict({
    train: Dataset({
        features: ['input_ids', 'attention_mask', 'labels'],
        num_rows: 4154
    })
    validation: Dataset({
        features: ['input_ids', 'attention_mask', 'labels'],
        num_rows: 167
    })
    test: Dataset({
        features: ['input_ids', 'attention_mask', 'labels'],
        num_rows: 500
    })
})
```

The output dataset is ready for fine-tuning.

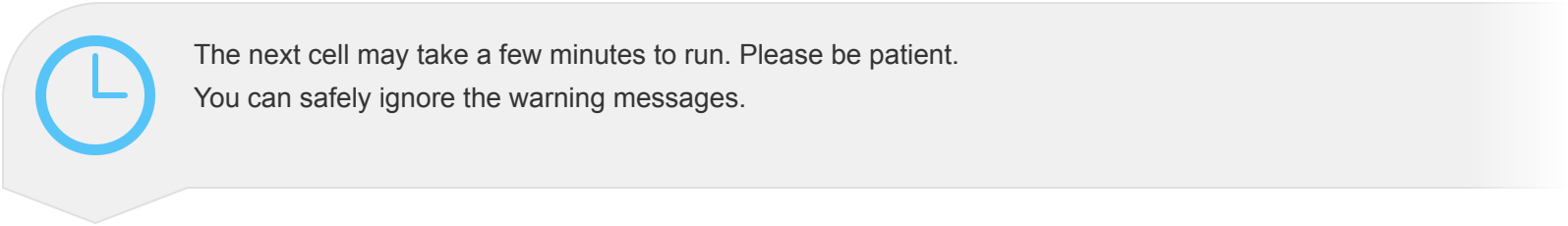
2.2 - Fine-Tune the Model with the Preprocessed Dataset (3 points)

Now utilize the built-in Hugging Face `Trainer` class (see the documentation here). Pass the preprocessed dataset with reference to the original model. Other training parameters are found experimentally and there is no need to go into details about those at the moment.

In [60]:
```
output_dir = f'./dialogue-summary-training-{str(int(time.time()))}'

# Configure TrainingArguments with appropriate learning rate, epochs, and other hyperparameters
training_args = TrainingArguments(
    output_dir=output_dir,
    learning_rate=5e-5,
    num_train_epochs=3,
    auto_find_batch_size=True,
    logging_steps=100
)

# Initialize the Hugging Face Trainer class with the model, training arguments, and datasets
trainer = Trainer(
    model=original_model,
```

```
    args=training_args,
    train_dataset=tokenized_datasets['train'],
    eval_dataset=tokenized_datasets['validation']
)
```

Start training process...



The code trainer.train() utilizes the Weights & Biases (wandb) library to track and visualize the training process. To proceed, you'll need to sign up for a wandb account using your Gmail and then enter your unique API token to authenticate and enable logging of the training progress.

In [62]:
```
# Login to wandb and initialize it
wandb.login(key = "43d56b51a7a89074cdecf6fd4ef49d1d5256762e")
wandb.init(project="hw4_problem1", name="2.2")

# Execute the training process using trainer.train()
trainer.train()

# Save the fine-tuned model checkpoint
trainer.save_model("./flan-t5-finetuned")
tokenizer.save_pretrained("./flan-t5-finetuned")
```

```
wandb: WARNING If you're specifying your api key in code, ensure this code is not shared publicly.
wandb: WARNING Consider setting the WANDB_API_KEY environment variable, or running `wandb login` from the command line.
wandb: Appending key for api.wandb.ai to your netrc file: /root/.netrc
```

Tracking run with wandb version 0.21.0

Run data is saved locally in /kaggle/working/wandb/run-20251115_173629-z4f57jnq

Syncing run **2.2** to Weights & Biases (docs)

View project at https://wandb.ai/pkh2120-columbia-university/hw4_problem1

View run at https://wandb.ai/pkh2120-columbia-university/hw4_problem1/runs/z4f57jnq

```
/usr/local/lib/python3.11/dist-packages/torch/nn/parallel/_functions.py:70: UserWarning: Was asked to gather along dime
nsion 0, but all input tensors were scalars; will instead unsqueeze and return a vector.
  warnings.warn(
```

[780/780 07:06, Epoch 3/3]

Step	Training Loss
100	11.767500
200	7.639700
300	5.739400
400	5.090000
500	4.869700
600	4.734400
700	4.730600

```
/usr/local/lib/python3.11/dist-packages/torch/nn/parallel/_functions.py:70: UserWarning: Was asked to gather along dime
nsion 0, but all input tensors were scalars; will instead unsqueeze and return a vector.
  warnings.warn(
```

Out[62]:
```
('./flan-t5-finetuned/tokenizer_config.json',
 './flan-t5-finetuned/special_tokens_map.json',
 './flan-t5-finetuned/spiece.model',
 './flan-t5-finetuned/added_tokens.json',
 './flan-t5-finetuned/tokenizer.json')
```

Create an instance of the `AutoModelForSeq2SeqLM` class for the instruct model:

In [64]:
```python
# Load tokenizer and models
# Import T5Tokenizer from transformers
from transformers import T5Tokenizer, AutoModelForSeq2SeqLM

# Define the model path using the config.json path
model_path = "./flan-t5-finetuned"

# Load tokenizer and models
# Use the default T5 tokenizer
instruct_tokenizer = T5Tokenizer.from_pretrained("t5-base") # TODO  # or "t5-base", "t5-large", etc.

# Load the model in a way that is compatible with single-GPU environments
instruct_model = AutoModelForSeq2SeqLM.from_pretrained(
    model_path,
    torch_dtype=torch.bfloat16,
    # The following line addresses the multi-GPU loading issue
    device_map="auto",
```

```
)

# Move model to GPU if available (optional, as device_map="auto" should handle it)
device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
instruct_model.to(device)
```

Out[64]:
```
T5ForConditionalGeneration(
  (shared): Embedding(32128, 512)
  (encoder): T5Stack(
    (embed_tokens): Embedding(32128, 512)
    (block): ModuleList(
      (0): T5Block(
        (layer): ModuleList(
          (0): T5LayerSelfAttention(
            (SelfAttention): T5Attention(
              (q): Linear(in_features=512, out_features=384, bias=False)
              (k): Linear(in_features=512, out_features=384, bias=False)
              (v): Linear(in_features=512, out_features=384, bias=False)
              (o): Linear(in_features=384, out_features=512, bias=False)
              (relative_attention_bias): Embedding(32, 6)
            )
            (layer_norm): T5LayerNorm()
            (dropout): Dropout(p=0.1, inplace=False)
          )
          (1): T5LayerFF(
            (DenseReluDense): T5DenseGatedActDense(
              (wi_0): Linear(in_features=512, out_features=1024, bias=False)
              (wi_1): Linear(in_features=512, out_features=1024, bias=False)
              (wo): Linear(in_features=1024, out_features=512, bias=False)
              (dropout): Dropout(p=0.1, inplace=False)
              (act): NewGELUActivation()
            )
            (layer_norm): T5LayerNorm()
            (dropout): Dropout(p=0.1, inplace=False)
          )
        )
      )
      (1-7): 7 x T5Block(
        (layer): ModuleList(
          (0): T5LayerSelfAttention(
            (SelfAttention): T5Attention(
              (q): Linear(in_features=512, out_features=384, bias=False)
              (k): Linear(in_features=512, out_features=384, bias=False)
              (v): Linear(in_features=512, out_features=384, bias=False)
              (o): Linear(in_features=384, out_features=512, bias=False)
            )
            (layer_norm): T5LayerNorm()
            (dropout): Dropout(p=0.1, inplace=False)
          )
          (1): T5LayerFF(
            (DenseReluDense): T5DenseGatedActDense(
              (wi_0): Linear(in_features=512, out_features=1024, bias=False)
              (wi_1): Linear(in_features=512, out_features=1024, bias=False)
```

```
            (wo): Linear(in_features=1024, out_features=512, bias=False)
            (dropout): Dropout(p=0.1, inplace=False)
            (act): NewGELUActivation()
          )
          (layer_norm): T5LayerNorm()
          (dropout): Dropout(p=0.1, inplace=False)
        )
      )
    )
  )
  (final_layer_norm): T5LayerNorm()
  (dropout): Dropout(p=0.1, inplace=False)
)
(decoder): T5Stack(
  (embed_tokens): Embedding(32128, 512)
  (block): ModuleList(
    (0): T5Block(
      (layer): ModuleList(
        (0): T5LayerSelfAttention(
          (SelfAttention): T5Attention(
            (q): Linear(in_features=512, out_features=384, bias=False)
            (k): Linear(in_features=512, out_features=384, bias=False)
            (v): Linear(in_features=512, out_features=384, bias=False)
            (o): Linear(in_features=384, out_features=512, bias=False)
            (relative_attention_bias): Embedding(32, 6)
          )
          (layer_norm): T5LayerNorm()
          (dropout): Dropout(p=0.1, inplace=False)
        )
        (1): T5LayerCrossAttention(
          (EncDecAttention): T5Attention(
            (q): Linear(in_features=512, out_features=384, bias=False)
            (k): Linear(in_features=512, out_features=384, bias=False)
            (v): Linear(in_features=512, out_features=384, bias=False)
            (o): Linear(in_features=384, out_features=512, bias=False)
          )
          (layer_norm): T5LayerNorm()
          (dropout): Dropout(p=0.1, inplace=False)
        )
        (2): T5LayerFF(
          (DenseReluDense): T5DenseGatedActDense(
            (wi_0): Linear(in_features=512, out_features=1024, bias=False)
            (wi_1): Linear(in_features=512, out_features=1024, bias=False)
            (wo): Linear(in_features=1024, out_features=512, bias=False)
            (dropout): Dropout(p=0.1, inplace=False)
            (act): NewGELUActivation()
          )
          (layer_norm): T5LayerNorm()
```

```
            (dropout): Dropout(p=0.1, inplace=False)
          )
        )
      )
      (1-7): 7 x T5Block(
        (layer): ModuleList(
          (0): T5LayerSelfAttention(
            (SelfAttention): T5Attention(
              (q): Linear(in_features=512, out_features=384, bias=False)
              (k): Linear(in_features=512, out_features=384, bias=False)
              (v): Linear(in_features=512, out_features=384, bias=False)
              (o): Linear(in_features=384, out_features=512, bias=False)
            )
            (layer_norm): T5LayerNorm()
            (dropout): Dropout(p=0.1, inplace=False)
          )
          (1): T5LayerCrossAttention(
            (EncDecAttention): T5Attention(
              (q): Linear(in_features=512, out_features=384, bias=False)
              (k): Linear(in_features=512, out_features=384, bias=False)
              (v): Linear(in_features=512, out_features=384, bias=False)
              (o): Linear(in_features=384, out_features=512, bias=False)
            )
            (layer_norm): T5LayerNorm()
            (dropout): Dropout(p=0.1, inplace=False)
          )
          (2): T5LayerFF(
            (DenseReluDense): T5DenseGatedActDense(
              (wi_0): Linear(in_features=512, out_features=1024, bias=False)
              (wi_1): Linear(in_features=512, out_features=1024, bias=False)
              (wo): Linear(in_features=1024, out_features=512, bias=False)
              (dropout): Dropout(p=0.1, inplace=False)
              (act): NewGELUActivation()
            )
            (layer_norm): T5LayerNorm()
            (dropout): Dropout(p=0.1, inplace=False)
          )
        )
      )
    )
    (final_layer_norm): T5LayerNorm()
    (dropout): Dropout(p=0.1, inplace=False)
  )
  (lm_head): Linear(in_features=512, out_features=32128, bias=False)
)
```

2.3 - Evaluate the Model Qualitatively (Human Evaluation) (2 points)

As with many GenAI applications, a qualitative approach where you ask yourself the question "Is my model behaving the way it is supposed to?" is usually a good starting point. In the example below (the same one we started this notebook with), you can see how the fine-tuned model is able to create a reasonable summary of the dialogue compared to the original inability to understand what is being asked of the model.

In [69]:
```
# Get random test dialogue and label from test dataset
index = 200
dialogue = dataset['test'][index]['dialogue']
human_baseline_summary = dataset['test'][index]['summary']

# Construct prompt
prompt = "Summarize the following conversation.\n\n" + dialogue + "\n\nSummary:"

# Tokenize the prompt
input_ids = tokenizer(prompt, return_tensors="pt").input_ids

# Move input_ids to the same device as the model
input_ids = input_ids.to(device)

# Get text output from original model
original_tokenizer = tokenizer
original_model_response = original_model.generate(input_ids)
original_model_outputs = original_model_response[0]
original_model_text_output = original_tokenizer.decode(
    original_model_outputs,
    skip_special_tokens=True
)

# Get text output from instruct finetuned model
instruct_model_response = instruct_model.generate(input_ids)
instruct_model_outputs = instruct_model_response[0]
instruct_model_text_output = instruct_tokenizer.decode(
    instruct_model_outputs,
    skip_special_tokens=True
)

print(dash_line)
print(f'BASELINE HUMAN SUMMARY:\n{human_baseline_summary}')
print(dash_line)
print(f'ORIGINAL MODEL:\n{original_model_text_output}')
print(dash_line)
print(f'INSTRUCT MODEL:\n{instruct_model_text_output}')
```

```
---------------------------------------------------------------------------------------------------
BASELINE HUMAN SUMMARY:
#Person1# teaches #Person2# how to upgrade software and hardware in #Person2#'s system.
---------------------------------------------------------------------------------------------------
ORIGINAL MODEL:
Share this with the others.
---------------------------------------------------------------------------------------------------
INSTRUCT MODEL:
#Person1# You could consider adding a painting program to your software.
```

2.3 Compare outputs across models using the same test examples, Analyze improvements in summary quality, coherence, and relevance

As you can see in the output above, the finetuned model captures the conversation much better than the original model which ended up giving a different response to the test prompt this time. The original model doesn't really capture the purpose of the conversation well at all while the finetuned model at least captures some key points made in the conversation.

2.4 - Evaluate the Model Quantitatively (with ROUGE Metric) (3 points)

The ROUGE metric helps quantify the validity of summarizations produced by models. It compares summarizations to a "baseline" summary which is usually created by a human. While not perfect, it does indicate the overall increase in summarization effectiveness that we have accomplished by fine-tuning.

In [70]:
```
# Load rouge evaluator
rouge = evaluate.load("rouge")
```

```
Downloading builder script: 0.00B [00:00, ?B/s]
```

Generate the outputs for the sample of the test dataset (only 10 dialogues and summaries to save time), and save the results.

In [77]:
```
# Get 10 test set dialogues and their labels
dialogues = dataset['test'][40:50]['dialogue']
human_baseline_summaries = dataset['test'][40:50]['summary']

original_model_summaries = []
instruct_model_summaries = []

for _, dialogue in enumerate(dialogues):
    prompt = "Summarize the following conversation.\n\n" + dialogue + "\n\nSummary:"
    # Tokenize the prompt
    input_ids = tokenizer(prompt, return_tensors="pt").input_ids

    # Move input_ids to the same device as the model
    input_ids = input_ids.to(device)
```

```
    # Get text output from original model
    original_model_response = original_model.generate(input_ids)
    original_model_outputs = original_model_response[0]
    original_model_text_output = original_tokenizer.decode(
        original_model_outputs,
        skip_special_tokens=True
    )

    # Get text output from instruct finetuned model
    instruct_model_response = instruct_model.generate(input_ids)
    instruct_model_outputs = instruct_model_response[0]
    instruct_model_text_output = instruct_tokenizer.decode(
        instruct_model_outputs,
        skip_special_tokens=True
    )

    # Append model outputs to appropriate lists
    original_model_summaries.append(original_model_text_output)
    instruct_model_summaries.append(instruct_model_text_output)

# Display human baseline, original model, and finetuned model summaries in a dataframe
zipped_summaries = list(zip(human_baseline_summaries, original_model_summaries, instruct_model_summaries))
df = pd.DataFrame(zipped_summaries, columns = ['human_baseline_summaries', 'original_model_summaries', 'instruct_model_
df
```

Out[77]:

	human_baseline_summaries	original_model_summaries	instruct_model_summaries
0	#Person1# is in a hurry to catch a train. Tom ...	@Person1#Person1#Person2#It's a minute	@Person, I'm not a fan of the ten to nine by m...
1	#Person1# is rushing to catch a train but Tom ...	@PresidentSony_Persons are not the same.	@Person, I'm not a fan of the ten to nine by m...
2	#Person1# wants to adjust #Person1#'s life and...	You should not do this.	#Person1#
3	#Person1# has a bad lifestyle. #Person2# kindl...	It's a good idea.	#Person1#
4	#Person2# hopes #Person1# will become healthy ...	#Person1#	#Person1#
5	#Person1# tells #Person2# that Ruojia is marri...	#Person, #Person.	#Person! #Person! #Person! #Person! #Person!
6	#Person2# is surprised to know from #Person1# ...	Share your ideas.	#Person! #Person! #Person! #Person! #Person!
7	#Person2# is surprised that Ruojia's married. ...	@Person_____	#Person! #Person! #Person! #Person! #Person!
8	#Person2# at first thinks #Person1#'s behaviou...	#Person1#Person2#	#Person1# You might make a few enemies.
9	#Person1# plans on playing a trick to others. ...	'I'm not a fan of being rude to your friends.	#Person1# You might make a few enemies.

Evaluate the models computing ROUGE metrics. Notice the improvement in the results!

In [79]:
```
# Get rouge score metric for original model generated summaries
original_model_results = rouge.compute(
    predictions=original_model_summaries, references=human_baseline_summaries
)

# Get rouge score metric for finetuned model generated summaries
instruct_model_results = rouge.compute(
    predictions=instruct_model_summaries, references=human_baseline_summaries
)

print('ORIGINAL MODEL:')
print(original_model_results)
print('INSTRUCT MODEL:')
print(instruct_model_results)
```

```
ORIGINAL MODEL:
{'rouge1': 0.10210084033613445, 'rouge2': 0.0, 'rougeL': 0.08015390064071605, 'rougeLsum': 0.07840561511961107}
INSTRUCT MODEL:
{'rouge1': 0.11086793517633425, 'rouge2': 0.0, 'rougeL': 0.09812110581030985, 'rougeLsum': 0.09455990521235702}
```

2.4 Analyze and compare performance metrics between models

As you can clearly see from the output above, the finetuned model slightly beats the original model in every rouge metric except rouge2 which is 0.0 for both the original and finetuned model generated summaries. When you analyze the generated summaries qualitatively the outputs from the original model and finetuned model both look pretty bad though. The original model I guess allows for a more creative output than the finetuned model which is maybe why the original model generated summaries that have more of a diverse vocabulary.

The file `data/dialogue-summary-training-results.csv` contains a pre-populated list of all model results which you can use to evaluate on a larger section of data. Let's do that for each of the models:

In [80]:
```
# Update the path to the CSV file
results_path = "../input/pre-populated-list/dialogue-summary-training-results.csv"
results = pd.read_csv(results_path)

human_baseline_summaries = results['human_baseline_summaries'].values
original_model_summaries = results['original_model_summaries'].values
instruct_model_summaries = results['instruct_model_summaries'].values

# Get rouge score metric for original model generated summaries
original_model_results = rouge.compute(
    predictions=original_model_summaries, references=human_baseline_summaries
)

# Get rouge score metric for finetuned model generated summaries
instruct_model_results = rouge.compute(
    predictions=instruct_model_summaries, references=human_baseline_summaries
)


print('ORIGINAL MODEL:')
print(original_model_results)
print('INSTRUCT MODEL:')
print(instruct_model_results)
```

```
ORIGINAL MODEL:
{'rouge1': 0.2216686882994889, 'rouge2': 0.0707492488737373, 'rougeL': 0.19245630286595683, 'rougeLsum': 0.192409231638
204}
INSTRUCT MODEL:
{'rouge1': 0.4041959932817219, 'rouge2': 0.17064828985299663, 'rougeL': 0.3267557101191949, 'rougeLsum': 0.326676672517
1105}
```

The results show substantial improvement in all ROUGE metrics:

In [83]:
```
print("Absolute percentage improvement of INSTRUCT MODEL over ORIGINAL MODEL")
```

```
improvement = (np.array(list(instruct_model_results.values())) - np.array(list(original_model_results.values())))
for key, value in zip(instruct_model_results.keys(), improvement):
    print(f'{key}: {value*100:.2f}%')
```

```
Absolute percentage improvement of INSTRUCT MODEL over ORIGINAL MODEL
rouge1: 18.25%
rouge2: 9.99%
rougeL: 13.43%
rougeLsum: 13.43%
```

2.4 Analyze and compare performance metrics between models

As you can clearly see from the output above, when evaluated across a larger section of data, the finetuned model beats the original model in every rouge metric with a much wider margin than previously seen when comparing with only 10 test samples.

3 - Perform Parameter Efficient Fine-Tuning (PEFT) (10 points)

Now, let's perform **Parameter Efficient Fine-Tuning (PEFT)** fine-tuning as opposed to "full fine-tuning" as you did above. PEFT is a form of instruction fine-tuning that is much more efficient than full fine-tuning - with comparable evaluation results as you will see soon.

PEFT is a generic term that includes **Low-Rank Adaptation (LoRA)** and prompt tuning (which is NOT THE SAME as prompt engineering!). In most cases, when someone says PEFT, they typically mean LoRA. LoRA, at a very high level, allows the user to fine-tune their model using fewer compute resources (in some cases, a single GPU). After fine-tuning for a specific task, use case, or tenant with LoRA, the result is that the original LLM remains unchanged and a newly-trained “LoRA adapter” emerges. This LoRA adapter is much, much smaller than the original LLM - on the order of a single-digit % of the original LLM size (MBs vs GBs).

That said, at inference time, the LoRA adapter needs to be reunited and combined with its original LLM to serve the inference request. The benefit, however, is that many LoRA adapters can re-use the original LLM which reduces overall memory requirements when serving multiple tasks and use cases.

3.1 - Setup the PEFT/LoRA model for Fine-Tuning (2 points)

You need to set up the PEFT/LoRA model for fine-tuning with a new layer/parameter adapter. Using PEFT/LoRA, you are freezing the underlying LLM and only training the adapter. Have a look at the LoRA configuration below. Note the rank (`r`) hyper-parameter, which defines the rank/dimension of the adapter to be trained.

In [84]:
```
from peft import LoraConfig, get_peft_model, TaskType

# Configure LoRA parameters using LoraConfig with appropriate rank, alpha, and target modules
lora_config = LoraConfig(
```

```
    r = 16, # TODO
    lora_alpha=32,
    target_modules=["q", "v"],
    lora_dropout=0.05,
    bias="none",
    task_type=TaskType.SEQ_2_SEQ_LM # FLAN-T5
)
```

Add LoRA adapter layers/parameters to the original LLM to be trained.

In [86]:
```
# Initialize the PEFT model
peft_model = get_peft_model(original_model, lora_config)

# Verify the reduction in trainable parameters compared to full fine tuning
print_number_of_trainable_model_parameters(peft_model)
```

```
Total Number of Parameters: 77649280
Total Number of Trainable Parameters: 688128
```

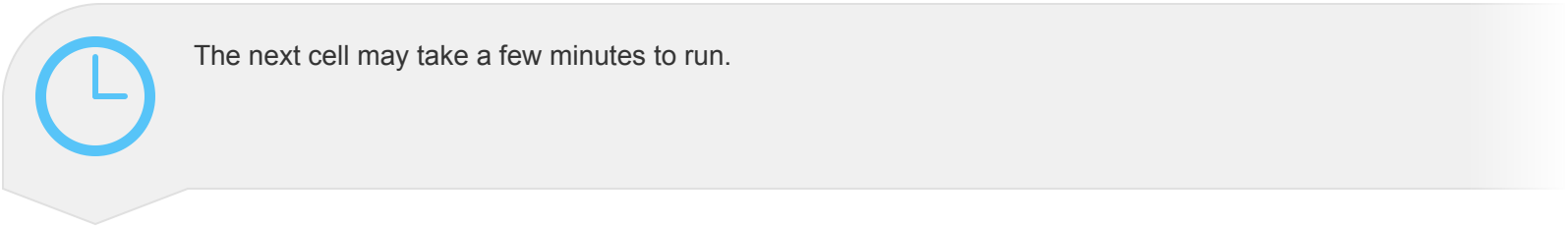
3.2 - Train PEFT Adapter (3 points)

Define training arguments and create `Trainer` instance.

In [87]:
```
output_dir = f'./peft-dialogue-summary-training-{str(int(time.time()))}'

# Set up training arguments specific to PEFT, including higher learning rate
peft_training_args = TrainingArguments(
    output_dir=output_dir,
    learning_rate=1e-3,
    num_train_epochs=3,
    auto_find_batch_size=True,
    logging_steps=100
)

# Initialize training using the Hugging Face Trainer
peft_trainer = Trainer(
    model=peft_model,
    args=peft_training_args,
    train_dataset=tokenized_datasets['train'],
    eval_dataset=tokenized_datasets['validation']
)
```

```
No label_names provided for model class `PeftModelForSeq2SeqLM`. Since `PeftModel` hides base models input arguments, i
f label_names is not given, label_names can't be set automatically within `Trainer`. Note that empty label_names list w
ill be used instead.
```

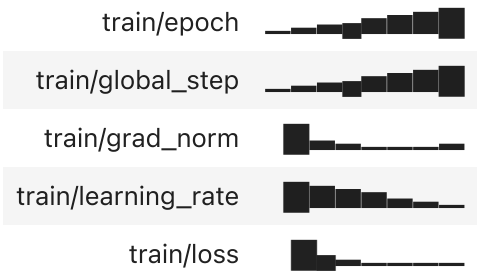
Now everything is ready to train the PEFT adapter and save the model.



In [88]:
```
# Login to wandb and initialize it
wandb.init(project="hw4_problem1", name="3.2")

peft_model_path="./flan-t5-peft-finetuned"

# Execute the training process using peft_trainer.train()
peft_trainer.train()

# Save the fine-tuned model checkpoint
peft_trainer.save_model(peft_model_path)
perf_tokenizer = original_tokenizer
perf_tokenizer.save_pretrained(peft_model_path)
```

Finishing previous runs because reinit is set to 'default'.

Run history:



Run summary:

total_flos	2316567469621248.0
train/epoch	3
train/global_step	780
train/grad_norm	15.75
train/learning_rate	1e-05
train/loss	4.7306
train_loss	6.20044
train_runtime	427.1481
train_samples_per_second	29.175
train_steps_per_second	1.826

View run **2.2** at: https://wandb.ai/pkh2120-columbia-university/hw4_problem1/runs/z4f57jnq
View project at: https://wandb.ai/pkh2120-columbia-university/hw4_problem1
Synced 5 W&B file(s), 0 media file(s), 0 artifact file(s) and 0 other file(s)

Find logs at: `./wandb/run-20251115_173629-z4f57jnq/logs`

Tracking run with wandb version 0.21.0

Run data is saved locally in `/kaggle/working/wandb/run-20251115_182757-1lupmg03`

Syncing run **3.2** to Weights & Biases (docs)

View project at https://wandb.ai/pkh2120-columbia-university/hw4_problem1

View run at https://wandb.ai/pkh2120-columbia-university/hw4_problem1/runs/1lupmg03

```
/usr/local/lib/python3.11/dist-packages/torch/nn/parallel/_functions.py:70: UserWarning: Was asked to gather along dime
nsion 0, but all input tensors were scalars; will instead unsqueeze and return a vector.
  warnings.warn(
```

[780/780 06:14, Epoch 3/3]

Step	Training Loss
100	3.149200
200	2.164500
300	1.990500
400	1.906600
500	1.856700
600	1.826900
700	1.811200


```
/usr/local/lib/python3.11/dist-packages/torch/nn/parallel/_functions.py:70: UserWarning: Was asked to gather along dime
nsion 0, but all input tensors were scalars; will instead unsqueeze and return a vector.
  warnings.warn(
```


Out[88]:
```
('./flan-t5-peft-finetuned/tokenizer_config.json',
 './flan-t5-peft-finetuned/special_tokens_map.json',
 './flan-t5-peft-finetuned/spiece.model',
 './flan-t5-peft-finetuned/added_tokens.json',
 './flan-t5-peft-finetuned/tokenizer.json')
```

That training was performed on a subset of data. To load a fully trained PEFT model, read a checkpoint of a PEFT model from Google Drive.

Prepare this model by adding an adapter to the original FLAN-T5 model. You are setting `is_trainable=False` because the plan is only to perform inference with this PEFT model. If you were preparing the model for further training, you would set `is_trainable=True`.

In [94]:
```
from peft import PeftModel, PeftConfig

# Get PeftConfig from peft_model_path
peft_config = PeftConfig.from_pretrained(peft_model_path)

# Get base peft model from peft_config
peft_model_base = AutoModelForSeq2SeqLM.from_pretrained(peft_config.base_model_name_or_path) # TODO
# Get tokenizer from peft_model_path
peft_tokenizer = AutoTokenizer.from_pretrained(peft_model_path) # TODO

# Generate peft_model
peft_model = PeftModel.from_pretrained(
    peft_model_base,
    "./flan-t5-peft-finetuned",
    is_trainable=False  # For inference only
)


# Move the entire peft_model to the device
peft_model = peft_model.to(device)
```

The number of trainable parameters will be `0` due to `is_trainable=False` setting:

In [95]:
```
print_number_of_trainable_model_parameters(peft_model)
```

```
Total Number of Parameters: 77649280
Total Number of Trainable Parameters: 0
```

3.3 - Evaluate the Model Qualitatively (Human Evaluation) (2 points)

Make inferences for the same example as in sections 1.3 and 2.3, with the original model, fully fine-tuned and PEFT model.

In [96]:
```
# Get random dialogue and label from test dataset
index = 200
dialogue = dataset['test'][index]['dialogue']
human_baseline_summary = dataset['test'][index]['summary']

prompt = "Summarize the following conversation.\n\n" + dialogue + "\n\nSummary:"

# Tokenize the prompt
input_ids = tokenizer(prompt, return_tensors="pt").input_ids

# Move input_ids to the same device as the model
input_ids = input_ids.to(device)
```

```
# Get text output from original model
original_model_response = original_model.generate(input_ids)
original_model_outputs = original_model_response[0]
original_model_text_output = original_tokenizer.decode(
    original_model_outputs,
    skip_special_tokens=True
)

# Get text output from instruct finetuned model
instruct_model_response = instruct_model.generate(input_ids)
instruct_model_outputs = instruct_model_response[0]
instruct_model_text_output = tokenizer.decode(
    instruct_model_outputs,
    skip_special_tokens=True
)

# Get text output from peft model
peft_model_response = peft_model.generate(input_ids=input_ids)
peft_model_outputs = peft_model_response[0]
peft_model_text_output = peft_tokenizer.decode(
    peft_model_outputs,
    skip_special_tokens=True
)

print(dash_line)
print(f'BASELINE HUMAN SUMMARY:\n{human_baseline_summary}')
print(dash_line)
print(f'ORIGINAL MODEL:\n{original_model_text_output}')
print(dash_line)
print(f'INSTRUCT MODEL:\n{instruct_model_text_output}')
print(dash_line)
print(f'PEFT MODEL: {peft_model_text_output}')
```

```
---------------------------------------------------------------------------------------------------
BASELINE HUMAN SUMMARY:
#Person1# teaches #Person2# how to upgrade software and hardware in #Person2#'s system.
---------------------------------------------------------------------------------------------------
ORIGINAL MODEL:
#Person1# thinks #Person2# should upgrade to the system because it is
---------------------------------------------------------------------------------------------------
INSTRUCT MODEL:
#Person1# You could consider adding a painting program to your software.
---------------------------------------------------------------------------------------------------
PEFT MODEL: #Person1# thinks adding a painting program to the software would allow #Person
```

3.3 Analyze the quality of summaries considering different aspects

As you can see in the output above, the finetuned model captures the conversation much better than the original model which ended up giving a different response to the test prompt this time. The original model doesn't really capture the purpose of the conversation well while the finetuned model at least captures some key points made in the conversation. In addition, the peft model surprisingly captures the conversation better than both the finetuned model and the original model. Unlike the finetuned model, the peft model is able to identify that the conversation is a third-party conversation. Maybe the fully finetuned model was overfitting and the peft model is able to generalize much better than the fully finetuned model due to the reduction in model size.

3.4 - Evaluate the Model Quantitatively (with ROUGE Metric) (3 points)

Perform inferences for the sample of the test dataset (only 10 dialogues and summaries to save time).

In [104…

```
# Get 10 test set dialogues and their labels
dialogues = dataset['test'][40:50]['dialogue']
human_baseline_summaries = dataset['test'][40:50]['summary']

original_model_summaries = []
instruct_model_summaries = []
peft_model_summaries = []

for idx, dialogue in enumerate(dialogues):
    prompt = "Summarize the following conversation.\n\n" + dialogue + "\n\nSummary:"

    # Move input_ids to the same device as the model
    input_ids = tokenizer(prompt, return_tensors="pt").input_ids.to(device)

    # Get text output from original model
    original_model_response = original_model.generate(input_ids)
    original_model_outputs = original_model_response[0]
    original_model_text_output = original_tokenizer.decode(
        original_model_outputs,
        skip_special_tokens=True
    )

    # Get text output from instruct finetuned model
    instruct_model_response = instruct_model.generate(input_ids)
    instruct_model_outputs = instruct_model_response[0]
    instruct_model_text_output = instruct_tokenizer.decode(
        instruct_model_outputs,
        skip_special_tokens=True
    )

    # Get text output from peft model
    peft_model_response = peft_model.generate(input_ids=input_ids)
```

```
    peft_model_outputs = peft_model_response[0]
    peft_model_text_output = peft_tokenizer.decode(
        peft_model_outputs,
        skip_special_tokens=True
    )

    # append the model outputs to their respective lists
    original_model_summaries.append(original_model_text_output)
    instruct_model_summaries.append(instruct_model_text_output)
    peft_model_summaries.append(peft_model_text_output)

zipped_summaries = list(zip(human_baseline_summaries, original_model_summaries, instruct_model_summaries, peft_model_su

# display the outputs of all the models as a pandas dataframe
df = pd.DataFrame(zipped_summaries, columns = ['human_baseline_summaries', 'original_model_summaries', 'instruct_model_
df
```

Out[104…

	human_baseline_summaries	original_model_summaries	instruct_model_summaries	peft_model_summaries
0	#Person1# is in a hurry to catch a train. Tom ...	#P1# tells Tom Tom is ten to nine by his watch...	@Person, I'm not a fan of the ten to nine by m...	Tom is waiting for the train to arrive. He has...
1	#Person1# is rushing to catch a train but Tom ...	#Person2# is off now and will catch the train.	@Person, I'm not a fan of the ten to nine by m...	Tom is waiting for the train to arrive. He has...
2	#Person1# wants to adjust #Person1#'s life and...	#Person2# tells #Person1# #Person2# can't	#Person1#	#Person1# tells #Person2# that #Person2# can'
3	#Person1# has a bad lifestyle. #Person2# kindl...	#Person2# tells #Person2# #Person1# can't	#Person1#	#Person1# tells #Person2# that #Person2# can'
4	#Person2# hopes #Person1# will become healthy ...	#Person1# tells #Person1# #Person1# can't	#Person1#	#Person1# tells #Person2# that #Person2# can'
5	#Person1# tells #Person2# that Ruojia is marri...	#Person1# invites #Person2# to the party tonig...	#Person! #Person! #Person! #Person! #Person!	Ruojia's party is going to be a party tonight....
6	#Person2# is surprised to know from #Person1# ...	Ruojia wants to go to the party tonight. #Pers...	#Person! #Person! #Person! #Person! #Person!	Ruojia's party is going to be a party tonight....
7	#Person2# is surprised that Ruojia's married. ...	#Person1# wants to go to the party. #Person1# ...	#Person! #Person! #Person! #Person! #Person!	Ruojia's party is going to be a party tonight....
8	#Person2# at first thinks #Person1#'s behaviou...	#Person1# tells #Person1# that #Person22# is	#Person1# You might make a few enemies.	#Person1# tells #Person2# that the two ugly ol...
9	#Person1# plans on playing a trick to others. ...	#Person1# tells #Person1# #Person1#'s friends	#Person1# You might make a few enemies.	#Person1# tells #Person2# that the two ugly ol...

Compute ROUGE score for this subset of the data.

In [105…

```
# Load the rouge evaluator
rouge = evaluate.load('rouge')

# Get rouge score metric for original model generated summaries
original_model_results = rouge.compute(
    predictions=original_model_summaries, references=human_baseline_summaries
)

# Get rouge score metric for finetuned model generated summaries
instruct_model_results = rouge.compute(
    predictions=instruct_model_summaries, references=human_baseline_summaries
)

# Get rouge score metric for peft model generated summaries
peft_model_results = rouge.compute(
    predictions=peft_model_summaries, references=human_baseline_summaries
)

print('ORIGINAL MODEL:')
print(original_model_results)
print('INSTRUCT MODEL:')
print(instruct_model_results)
print('PEFT MODEL:')
print(peft_model_results)
```

```
ORIGINAL MODEL:
{'rouge1': 0.2297754317843419, 'rouge2': 0.016856060606060607, 'rougeL': 0.19614187176847026, 'rougeLsum': 0.1962943161
0291276}
INSTRUCT MODEL:
{'rouge1': 0.11086793517633425, 'rouge2': 0.0, 'rougeL': 0.09812110581030985, 'rougeLsum': 0.09455990521235702}
PEFT MODEL:
{'rouge1': 0.27515248972164164, 'rouge2': 0.05506175640250051, 'rougeL': 0.20027310328222514, 'rougeLsum': 0.1992112306
3254126}
```

3.4 Analyze PEFT vs. original model metrics Compare PEFT vs. full fine-tuning results

As you can clearly see from the output above, the finetuned model performed the worst, followed by the original model, and then the peft model which performed the best across all metrics. The fact that the finetuned model performed the worst indicates that finetuning the entire model caused overfitting. Surprisingly, the peft model performed the best which indicates that the model has way more parameters than it actually needs and reducing the model size can help it generalize better/produce better results on unseen data.

Notice, that PEFT model results are not too bad, while the training process was much easier!

You already computed ROUGE score on the full dataset, after loading the results from the `data/dialogue-summary-training-results.csv` file. Load the values for the PEFT model now and check its performance compared to other models.

In [107…

```
# Update the path to the CSV file
results_path = "../input/pre-populated-list/dialogue-summary-training-results.csv"
results = pd.read_csv(results_path)

human_baseline_summaries = results['human_baseline_summaries'].values
original_model_summaries = results['original_model_summaries'].values
instruct_model_summaries = results['instruct_model_summaries'].values
peft_model_summaries     = results['peft_model_summaries'].values

# Get rouge score metric for original model generated summaries
original_model_results = rouge.compute(
    predictions=original_model_summaries, references=human_baseline_summaries
)

# Get rouge score metric for finetuned model generated summaries
instruct_model_results = rouge.compute(
    predictions=instruct_model_summaries, references=human_baseline_summaries
)

# Get rouge score metric for peft model generated summaries
peft_model_results = rouge.compute(
    predictions=peft_model_summaries, references=human_baseline_summaries
)

print('ORIGINAL MODEL:')
print(original_model_results)
print('INSTRUCT MODEL:')
print(instruct_model_results)
print('PEFT MODEL:')
print(peft_model_results)
```

```
ORIGINAL MODEL:
{'rouge1': 0.2216686882994889, 'rouge2': 0.0707492488737373, 'rougeL': 0.19245630286595683, 'rougeLsum': 0.192409231638
204}
INSTRUCT MODEL:
{'rouge1': 0.4041959932817219, 'rouge2': 0.17064828985299663, 'rougeL': 0.3267557101191949, 'rougeLsum': 0.326676672517
1105}
PEFT MODEL:
{'rouge1': 0.39119098357131776, 'rouge2': 0.15459808342905274, 'rougeL': 0.31367299251500014, 'rougeLsum': 0.3136061516
8633016}
```

The results show less of an improvement over full fine-tuning, but the benefits of PEFT typically outweigh the slightly-lower performance metrics.

Calculate the improvement of PEFT over the original model:

In [108…
```
print("Absolute percentage improvement of PEFT MODEL over ORIGINAL MODEL")

improvement = (np.array(list(peft_model_results.values())) - np.array(list(original_model_results.values())))
for key, value in zip(peft_model_results.keys(), improvement):
    print(f'{key}: {value*100:.2f}%')
```

```
Absolute percentage improvement of PEFT MODEL over ORIGINAL MODEL
rouge1: 16.95%
rouge2: 8.38%
rougeL: 12.12%
rougeLsum: 12.12%
```

Now calculate the improvement of PEFT over a full fine-tuned model:

In [109…
```
print("Absolute percentage improvement of PEFT MODEL over INSTRUCT MODEL")

improvement = (np.array(list(peft_model_results.values())) - np.array(list(instruct_model_results.values())))
for key, value in zip(peft_model_results.keys(), improvement):
    print(f'{key}: {value*100:.2f}%')
```

```
Absolute percentage improvement of PEFT MODEL over INSTRUCT MODEL
rouge1: -1.30%
rouge2: -1.61%
rougeL: -1.31%
rougeLsum: -1.31%
```

3.4 Analyze and compare performance metrics between models

Evaluate the trade-off between performance and computational efficiency

As you can clearly see from the output above, when evaluated across a larger section of data, the finetuned model beats both the original model and the peft model in every rouge metric. However, the peft model rouge metrics are very close to the finetuned model metrics indicating that reducing the model size with LoRA still allows it to produce comparable results that are pretty close to what we can achieve with full model finetuning. Finetuning the peft model is also way more computationally efficient due to the significantly less amount of parameters that it handles as shown in the previous output that prints the number of trainable parameters for the peft model. As a result, it might be worth it to just finetune the peft model for more epochs than finetuning the full model if there's a computational budget to be met. In this case the peft model might actually produce better results than finetuning the full model.

Here you see a small percentage decrease in the ROUGE metrics vs. full fine-tuned. However, the training requires much less computing and memory resources (often just a single GPU).

Limitations Encountered During the Fine-Tuning Process: I was not really able to get good performance when finetuning the full model here on Kaggle no matter what reasonable values I tried for number of epochs(1-5) and learning rates I tried. I settled on this after around 2 hours of trying to figure this out.

I've executed this entire notebook on Kaggle where some of the necessary dependencies and libraries are already installed and don't need to be explicitly installed using pip install.

LLM Inference Benchmarking Assignment (25 points)

Overview

In this assignment, you will learn how to benchmark Large Language Model (LLM) inference using vLLM, a high-performance inference engine. You will use a small model (OPT-125M) to understand the basics of throughput measurement and the impact of different parameters on inference speed.

Environment Setup

- Successfully install all required packages (2 points)

First, install the required packages:

In [1]:
```
!pip install "numpy<2"
!pip install transformers torch pandas matplotlib seaborn vllm
```

```
Requirement already satisfied: numpy<2 in /usr/local/lib/python3.11/dist-packages (1.26.4)
Requirement already satisfied: transformers in /usr/local/lib/python3.11/dist-packages (4.57.1)
Requirement already satisfied: torch in /usr/local/lib/python3.11/dist-packages (2.8.0)
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages (2.3.3)
Requirement already satisfied: matplotlib in /usr/local/lib/python3.11/dist-packages (3.10.7)
Requirement already satisfied: seaborn in /usr/local/lib/python3.11/dist-packages (0.13.2)
Requirement already satisfied: vllm in /usr/local/lib/python3.11/dist-packages (0.11.0)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from transformers) (3.20.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.34.0 in /usr/local/lib/python3.11/dist-packages (from transforme
rs) (0.36.0)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-packages (from transformers) (1.26.4)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from transformers) (25.0)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-packages (from transformers) (6.0.3)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.11/dist-packages (from transformers) (2025.1
1.3)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages (from transformers) (2.32.5)
Requirement already satisfied: tokenizers<=0.23.0,>=0.22.0 in /usr/local/lib/python3.11/dist-packages (from transformer
s) (0.22.1)
Requirement already satisfied: safetensors>=0.4.3 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.5.
3)
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.11/dist-packages (from transformers) (4.67.1)
Requirement already satisfied: typing-extensions>=4.10.0 in /usr/local/lib/python3.11/dist-packages (from torch) (4.15.
0)
Requirement already satisfied: sympy>=1.13.3 in /usr/local/lib/python3.11/dist-packages (from torch) (1.14.0)
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-packages (from torch) (3.5)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.11/dist-packages (from torch) (3.1.6)
Requirement already satisfied: fsspec in /usr/local/lib/python3.11/dist-packages (from torch) (2025.10.0)
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.8.93 in /usr/local/lib/python3.11/dist-packages (from torch)
(12.8.93)
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.8.90 in /usr/local/lib/python3.11/dist-packages (from torc
h) (12.8.90)
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.8.90 in /usr/local/lib/python3.11/dist-packages (from torch)
(12.8.90)
Requirement already satisfied: nvidia-cudnn-cu12==9.10.2.21 in /usr/local/lib/python3.11/dist-packages (from torch) (9.
10.2.21)
Requirement already satisfied: nvidia-cublas-cu12==12.8.4.1 in /usr/local/lib/python3.11/dist-packages (from torch) (1
2.8.4.1)
Requirement already satisfied: nvidia-cufft-cu12==11.3.3.83 in /usr/local/lib/python3.11/dist-packages (from torch) (1
1.3.3.83)
Requirement already satisfied: nvidia-curand-cu12==10.3.9.90 in /usr/local/lib/python3.11/dist-packages (from torch) (1
0.3.9.90)
Requirement already satisfied: nvidia-cusolver-cu12==11.7.3.90 in /usr/local/lib/python3.11/dist-packages (from torch)
(11.7.3.90)
Requirement already satisfied: nvidia-cusparse-cu12==12.5.8.93 in /usr/local/lib/python3.11/dist-packages (from torch)
(12.5.8.93)
Requirement already satisfied: nvidia-cusparselt-cu12==0.7.1 in /usr/local/lib/python3.11/dist-packages (from torch)
(0.7.1)
Requirement already satisfied: nvidia-nccl-cu12==2.27.3 in /usr/local/lib/python3.11/dist-packages (from torch) (2.27.
```

```
3)
Requirement already satisfied: nvidia-nvtx-cu12==12.8.90 in /usr/local/lib/python3.11/dist-packages (from torch) (12.8.
90)
Requirement already satisfied: nvidia-nvjitlink-cu12==12.8.93 in /usr/local/lib/python3.11/dist-packages (from torch)
(12.8.93)
Requirement already satisfied: nvidia-cufile-cu12==1.13.1.3 in /usr/local/lib/python3.11/dist-packages (from torch) (1.
13.1.3)
Requirement already satisfied: triton==3.4.0 in /usr/local/lib/python3.11/dist-packages (from torch) (3.4.0)
Requirement already satisfied: setuptools>=40.8.0 in /usr/local/lib/python3.11/dist-packages (from triton==3.4.0->torc
h) (75.2.0)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas) (2.9.0.p
ost0)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from pandas) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas) (2025.2)
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (1.3.2)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (4.59.0)
Requirement already satisfied: kiwisolver>=1.3.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (1.4.8)
Requirement already satisfied: pillow>=8 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (11.3.0)
Requirement already satisfied: pyparsing>=3 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (3.0.9)
Requirement already satisfied: cachetools in /usr/local/lib/python3.11/dist-packages (from vllm) (6.2.1)
Requirement already satisfied: psutil in /usr/local/lib/python3.11/dist-packages (from vllm) (7.1.3)
Requirement already satisfied: sentencepiece in /usr/local/lib/python3.11/dist-packages (from vllm) (0.2.0)
Requirement already satisfied: blake3 in /usr/local/lib/python3.11/dist-packages (from vllm) (1.0.8)
Requirement already satisfied: py-cpuinfo in /usr/local/lib/python3.11/dist-packages (from vllm) (9.0.0)
Requirement already satisfied: protobuf in /usr/local/lib/python3.11/dist-packages (from vllm) (6.33.0)
Requirement already satisfied: fastapi>=0.115.0 in /usr/local/lib/python3.11/dist-packages (from fastapi[standard]>=0.1
15.0->vllm) (0.116.1)
Requirement already satisfied: aiohttp in /usr/local/lib/python3.11/dist-packages (from vllm) (3.13.2)
Requirement already satisfied: openai>=1.99.1 in /usr/local/lib/python3.11/dist-packages (from vllm) (2.7.1)
Requirement already satisfied: pydantic>=2.11.7 in /usr/local/lib/python3.11/dist-packages (from vllm) (2.12.4)
Requirement already satisfied: prometheus_client>=0.18.0 in /usr/local/lib/python3.11/dist-packages (from vllm) (0.22.
1)
Requirement already satisfied: prometheus-fastapi-instrumentator>=7.0.0 in /usr/local/lib/python3.11/dist-packages (fro
m vllm) (7.1.0)
Requirement already satisfied: tiktoken>=0.6.0 in /usr/local/lib/python3.11/dist-packages (from vllm) (0.9.0)
Requirement already satisfied: lm-format-enforcer==0.11.3 in /usr/local/lib/python3.11/dist-packages (from vllm) (0.11.
3)
Requirement already satisfied: llguidance<0.8.0,>=0.7.11 in /usr/local/lib/python3.11/dist-packages (from vllm) (0.7.3
0)
Requirement already satisfied: outlines_core==0.2.11 in /usr/local/lib/python3.11/dist-packages (from vllm) (0.2.11)
Requirement already satisfied: diskcache==5.6.3 in /usr/local/lib/python3.11/dist-packages (from vllm) (5.6.3)
Requirement already satisfied: lark==1.2.2 in /usr/local/lib/python3.11/dist-packages (from vllm) (1.2.2)
Requirement already satisfied: xgrammar==0.1.25 in /usr/local/lib/python3.11/dist-packages (from vllm) (0.1.25)
Requirement already satisfied: partial-json-parser in /usr/local/lib/python3.11/dist-packages (from vllm) (0.2.1.1.post
7)
Requirement already satisfied: pyzmq>=25.0.0 in /usr/local/lib/python3.11/dist-packages (from vllm) (26.2.1)
Requirement already satisfied: msgspec in /usr/local/lib/python3.11/dist-packages (from vllm) (0.19.0)
```

```
Requirement already satisfied: gguf>=0.13.0 in /usr/local/lib/python3.11/dist-packages (from vllm) (0.17.1)
Requirement already satisfied: mistral_common>=1.8.2 in /usr/local/lib/python3.11/dist-packages (from mistral_common[au
dio,image]>=1.8.2->vllm) (1.8.5)
Requirement already satisfied: opencv-python-headless>=4.11.0 in /usr/local/lib/python3.11/dist-packages (from vllm)
(4.11.0.86)
Requirement already satisfied: einops in /usr/local/lib/python3.11/dist-packages (from vllm) (0.8.1)
Requirement already satisfied: compressed-tensors==0.11.0 in /usr/local/lib/python3.11/dist-packages (from vllm) (0.11.
0)
Requirement already satisfied: depyf==0.19.0 in /usr/local/lib/python3.11/dist-packages (from vllm) (0.19.0)
Requirement already satisfied: cloudpickle in /usr/local/lib/python3.11/dist-packages (from vllm) (3.1.2)
Requirement already satisfied: watchfiles in /usr/local/lib/python3.11/dist-packages (from vllm) (1.1.1)
Requirement already satisfied: python-json-logger in /usr/local/lib/python3.11/dist-packages (from vllm) (4.0.0)
Requirement already satisfied: scipy in /usr/local/lib/python3.11/dist-packages (from vllm) (1.15.3)
Requirement already satisfied: ninja in /usr/local/lib/python3.11/dist-packages (from vllm) (1.13.0)
Requirement already satisfied: pybase64 in /usr/local/lib/python3.11/dist-packages (from vllm) (1.4.2)
Requirement already satisfied: cbor2 in /usr/local/lib/python3.11/dist-packages (from vllm) (5.7.1)
Requirement already satisfied: setproctitle in /usr/local/lib/python3.11/dist-packages (from vllm) (1.3.7)
Requirement already satisfied: openai-harmony>=0.0.3 in /usr/local/lib/python3.11/dist-packages (from vllm) (0.0.8)
Requirement already satisfied: numba==0.61.2 in /usr/local/lib/python3.11/dist-packages (from vllm) (0.61.2)
Requirement already satisfied: ray>=2.48.0 in /usr/local/lib/python3.11/dist-packages (from ray[cgraph]>=2.48.0->vllm)
(2.51.1)
Requirement already satisfied: torchaudio==2.8.0 in /usr/local/lib/python3.11/dist-packages (from vllm) (2.8.0)
Requirement already satisfied: torchvision==0.23.0 in /usr/local/lib/python3.11/dist-packages (from vllm) (0.23.0)
Requirement already satisfied: xformers==0.0.32.post1 in /usr/local/lib/python3.11/dist-packages (from vllm) (0.0.32.po
st1)
Requirement already satisfied: frozendict in /usr/local/lib/python3.11/dist-packages (from compressed-tensors==0.11.0->
vllm) (2.4.6)
Requirement already satisfied: astor in /usr/local/lib/python3.11/dist-packages (from depyf==0.19.0->vllm) (0.8.1)
Requirement already satisfied: dill in /usr/local/lib/python3.11/dist-packages (from depyf==0.19.0->vllm) (0.4.0)
Requirement already satisfied: interegular>=0.3.2 in /usr/local/lib/python3.11/dist-packages (from lm-format-enforcer==
0.11.3->vllm) (0.3.3)
Requirement already satisfied: llvmlite<0.45,>=0.44.0dev0 in /usr/local/lib/python3.11/dist-packages (from numba==0.61.
2->vllm) (0.44.0)
Requirement already satisfied: starlette<0.48.0,>=0.40.0 in /usr/local/lib/python3.11/dist-packages (from fastapi>=0.11
5.0->fastapi[standard]>=0.115.0->vllm) (0.47.2)
Requirement already satisfied: fastapi-cli>=0.0.8 in /usr/local/lib/python3.11/dist-packages (from fastapi-cli[standar
d]>=0.0.8; extra == "standard"->fastapi[standard]>=0.115.0->vllm) (0.0.16)
Requirement already satisfied: httpx>=0.23.0 in /usr/local/lib/python3.11/dist-packages (from fastapi[standard]>=0.115.
0->vllm) (0.28.1)
Requirement already satisfied: python-multipart>=0.0.18 in /usr/local/lib/python3.11/dist-packages (from fastapi[standa
rd]>=0.115.0->vllm) (0.0.20)
Requirement already satisfied: email-validator>=2.0.0 in /usr/local/lib/python3.11/dist-packages (from fastapi[standar
d]>=0.115.0->vllm) (2.3.0)
Requirement already satisfied: uvicorn>=0.12.0 in /usr/local/lib/python3.11/dist-packages (from uvicorn[standard]>=0.1
2.0; extra == "standard"->fastapi[standard]>=0.115.0->vllm) (0.35.0)
Requirement already satisfied: hf-xet<2.0.0,>=1.1.3 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.
0,>=0.34.0->transformers) (1.2.0)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from jinja2->torch) (3.0.3)
```

```
Requirement already satisfied: jsonschema>=4.21.1 in /usr/local/lib/python3.11/dist-packages (from mistral_common>=1.8.
2->mistral_common[audio,image]>=1.8.2->vllm) (4.25.0)
Requirement already satisfied: pydantic-extra-types>=2.10.5 in /usr/local/lib/python3.11/dist-packages (from pydantic-e
xtra-types[pycountry]>=2.10.5->mistral_common>=1.8.2->mistral_common[audio,image]>=1.8.2->vllm) (2.10.6)
Requirement already satisfied: anyio<5,>=3.5.0 in /usr/local/lib/python3.11/dist-packages (from openai>=1.99.1->vllm)
(4.11.0)
Requirement already satisfied: distro<2,>=1.7.0 in /usr/local/lib/python3.11/dist-packages (from openai>=1.99.1->vllm)
(1.9.0)
Requirement already satisfied: jiter<1,>=0.10.0 in /usr/local/lib/python3.11/dist-packages (from openai>=1.99.1->vllm)
(0.10.0)
Requirement already satisfied: sniffio in /usr/local/lib/python3.11/dist-packages (from openai>=1.99.1->vllm) (1.3.1)
Requirement already satisfied: annotated-types>=0.6.0 in /usr/local/lib/python3.11/dist-packages (from pydantic>=2.11.7
->vllm) (0.7.0)
Requirement already satisfied: pydantic-core==2.41.5 in /usr/local/lib/python3.11/dist-packages (from pydantic>=2.11.7-
>vllm) (2.41.5)
Requirement already satisfied: typing-inspection>=0.4.2 in /usr/local/lib/python3.11/dist-packages (from pydantic>=2.1
1.7->vllm) (0.4.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->panda
s) (1.17.0)
Requirement already satisfied: click!=8.3.0,>=7.0 in /usr/local/lib/python3.11/dist-packages (from ray>=2.48.0->ray[cgr
aph]>=2.48.0->vllm) (8.3.1)
Requirement already satisfied: msgpack<2.0.0,>=1.0.0 in /usr/local/lib/python3.11/dist-packages (from ray>=2.48.0->ray
[cgraph]>=2.48.0->vllm) (1.1.2)
Requirement already satisfied: cupy-cuda12x in /usr/local/lib/python3.11/dist-packages (from ray[cgraph]>=2.48.0->vllm)
(13.6.0)
Requirement already satisfied: charset_normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests->tran
sformers) (3.4.4)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests->transformers)
(3.11)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests->transforme
rs) (2.5.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests->transforme
rs) (2025.10.5)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.11/dist-packages (from sympy>=1.13.3->torc
h) (1.3.0)
Requirement already satisfied: aiohappyeyeballs>=2.5.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->vllm)
(2.6.1)
Requirement already satisfied: aiosignal>=1.4.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->vllm) (1.4.0)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->vllm) (25.4.0)
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.11/dist-packages (from aiohttp->vllm) (1.8.
0)
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.11/dist-packages (from aiohttp->vllm) (6.
7.0)
Requirement already satisfied: propcache>=0.2.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->vllm) (0.4.1)
Requirement already satisfied: yarl<2.0,>=1.17.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->vllm) (1.22.
0)
Requirement already satisfied: dnspython>=2.0.0 in /usr/local/lib/python3.11/dist-packages (from email-validator>=2.0.0
->fastapi[standard]>=0.115.0->vllm) (2.8.0)
```

```
Requirement already satisfied: typer>=0.15.1 in /usr/local/lib/python3.11/dist-packages (from fastapi-cli>=0.0.8->fasta
pi-cli[standard]>=0.0.8; extra == "standard"->fastapi[standard]>=0.115.0->vllm) (0.16.0)
Requirement already satisfied: rich-toolkit>=0.14.8 in /usr/local/lib/python3.11/dist-packages (from fastapi-cli>=0.0.8
->fastapi-cli[standard]>=0.0.8; extra == "standard"->fastapi[standard]>=0.115.0->vllm) (0.15.1)
Requirement already satisfied: fastapi-cloud-cli>=0.1.1 in /usr/local/lib/python3.11/dist-packages (from fastapi-cli[st
andard]>=0.0.8; extra == "standard"->fastapi[standard]>=0.115.0->vllm) (0.3.1)
Requirement already satisfied: httpcore==1.* in /usr/local/lib/python3.11/dist-packages (from httpx>=0.23.0->fastapi[st
andard]>=0.115.0->vllm) (1.0.9)
Requirement already satisfied: h11>=0.16 in /usr/local/lib/python3.11/dist-packages (from httpcore==1.*->httpx>=0.23.0-
>fastapi[standard]>=0.115.0->vllm) (0.16.0)
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in /usr/local/lib/python3.11/dist-packages (from js
onschema>=4.21.1->mistral_common>=1.8.2->mistral_common[audio,image]>=1.8.2->vllm) (2025.4.1)
Requirement already satisfied: referencing>=0.28.4 in /usr/local/lib/python3.11/dist-packages (from jsonschema>=4.21.1-
>mistral_common>=1.8.2->mistral_common[audio,image]>=1.8.2->vllm) (0.36.2)
Requirement already satisfied: rpds-py>=0.7.1 in /usr/local/lib/python3.11/dist-packages (from jsonschema>=4.21.1->mist
ral_common>=1.8.2->mistral_common[audio,image]>=1.8.2->vllm) (0.26.0)
Requirement already satisfied: pycountry>=23 in /usr/local/lib/python3.11/dist-packages (from pydantic-extra-types[pyco
untry]>=2.10.5->mistral_common>=1.8.2->mistral_common[audio,image]>=1.8.2->vllm) (24.6.1)
Requirement already satisfied: httptools>=0.6.3 in /usr/local/lib/python3.11/dist-packages (from uvicorn[standard]>=0.1
2.0; extra == "standard"->fastapi[standard]>=0.115.0->vllm) (0.7.1)
Requirement already satisfied: python-dotenv>=0.13 in /usr/local/lib/python3.11/dist-packages (from uvicorn[standard]>=
0.12.0; extra == "standard"->fastapi[standard]>=0.115.0->vllm) (1.2.1)
Requirement already satisfied: uvloop>=0.15.1 in /usr/local/lib/python3.11/dist-packages (from uvicorn[standard]>=0.12.
0; extra == "standard"->fastapi[standard]>=0.115.0->vllm) (0.22.1)
Requirement already satisfied: websockets>=10.4 in /usr/local/lib/python3.11/dist-packages (from uvicorn[standard]>=0.1
2.0; extra == "standard"->fastapi[standard]>=0.115.0->vllm) (15.0.1)
Requirement already satisfied: fastrlock>=0.5 in /usr/local/lib/python3.11/dist-packages (from cupy-cuda12x->ray[cgrap
h]>=2.48.0->vllm) (0.8.3)
Requirement already satisfied: soundfile>=0.12.1 in /usr/local/lib/python3.11/dist-packages (from mistral_common>=1.8.2
->mistral_common[audio,image]>=1.8.2->vllm) (0.13.1)
Requirement already satisfied: soxr>=0.5.0 in /usr/local/lib/python3.11/dist-packages (from mistral_common>=1.8.2->mist
ral_common[audio,image]>=1.8.2->vllm) (0.5.0.post1)
Requirement already satisfied: rignore>=0.5.1 in /usr/local/lib/python3.11/dist-packages (from fastapi-cloud-cli>=0.1.1
->fastapi-cli[standard]>=0.0.8; extra == "standard"->fastapi[standard]>=0.115.0->vllm) (0.7.6)
Requirement already satisfied: sentry-sdk>=2.20.0 in /usr/local/lib/python3.11/dist-packages (from fastapi-cloud-cli>=
0.1.1->fastapi-cli[standard]>=0.0.8; extra == "standard"->fastapi[standard]>=0.115.0->vllm) (2.33.2)
Requirement already satisfied: rich>=13.7.1 in /usr/local/lib/python3.11/dist-packages (from rich-toolkit>=0.14.8->fast
api-cli>=0.0.8->fastapi-cli[standard]>=0.0.8; extra == "standard"->fastapi[standard]>=0.115.0->vllm) (14.2.0)
Requirement already satisfied: cffi>=1.0 in /usr/local/lib/python3.11/dist-packages (from soundfile>=0.12.1->mistral_co
mmon>=1.8.2->mistral_common[audio,image]>=1.8.2->vllm) (2.0.0)
Requirement already satisfied: shellingham>=1.3.0 in /usr/local/lib/python3.11/dist-packages (from typer>=0.15.1->fasta
pi-cli>=0.0.8->fastapi-cli[standard]>=0.0.8; extra == "standard"->fastapi[standard]>=0.115.0->vllm) (1.5.4)
Requirement already satisfied: pycparser in /usr/local/lib/python3.11/dist-packages (from cffi>=1.0->soundfile>=0.12.1-
>mistral_common>=1.8.2->mistral_common[audio,image]>=1.8.2->vllm) (2.23)
Requirement already satisfied: markdown-it-py>=2.2.0 in /usr/local/lib/python3.11/dist-packages (from rich>=13.7.1->ric
h-toolkit>=0.14.8->fastapi-cli>=0.0.8->fastapi-cli[standard]>=0.0.8; extra == "standard"->fastapi[standard]>=0.115.0->v
llm) (4.0.0)
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in /usr/local/lib/python3.11/dist-packages (from rich>=13.7.1->r
```

```
ich-toolkit>=0.14.8->fastapi-cli>=0.0.8->fastapi-cli[standard]>=0.0.8; extra == "standard"->fastapi[standard]>=0.115.0-
>vllm) (2.19.2)
Requirement already satisfied: mdurl~=0.1 in /usr/local/lib/python3.11/dist-packages (from markdown-it-py>=2.2.0->rich>
=13.7.1->rich-toolkit>=0.14.8->fastapi-cli>=0.0.8->fastapi-cli[standard]>=0.0.8; extra == "standard"->fastapi[standard]
>=0.115.0->vllm) (0.1.2)
```

Import the necessary libraries:

In [2]:

```python
# Import necessary libraries
import dataclasses
import time
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from typing import List
from vllm import LLM, SamplingParams
from transformers import AutoTokenizer
import gc
import torch

print("All libraries imported successfully!")
```

```
INFO 11-18 06:52:03 [__init__.py:216] Automatically detected platform cuda.
2025-11-18 06:52:04.539482: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register cuFFT fac
tory: Attempting to register factory for plugin cuFFT when one has already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
E0000 00:00:1763448724.563899    6422 cuda_dnn.cc:8310] Unable to register cuDNN factory: Attempting to register factor
y for plugin cuDNN when one has already been registered
E0000 00:00:1763448724.573948    6422 cuda_blas.cc:1418] Unable to register cuBLAS factory: Attempting to register fact
ory for plugin cuBLAS when one has already been registered

---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'

---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'

---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'

---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'

---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
All libraries imported successfully!
```

Part 1: Creating Sample Requests

- Implement the create_synthetic_requests function correctly (5 points)

We'll create a function to generate sample requests for benchmarking:

In [3]:
```python
import random

@dataclasses.dataclass
class SampleRequest:
    """A class representing a single inference request for benchmarking."""
    prompt: str
    prompt_len: int
    expected_output_len: int

def create_synthetic_requests(
    tokenizer,
    num_requests: int,
    input_len: int,
    output_len: int
) -> List[SampleRequest]:
    """Create synthetic requests for benchmarking.

    Args:
        tokenizer: The tokenizer to use
        num_requests: Number of requests to generate
        input_len: Desired input length in tokens
        output_len: Desired output length in tokens

    Returns:
        List of SampleRequest objects
    """
    requests = []

    prompts = [
        "Tell me about the difference between recurrent neural networks and bidirectional long short term memory networ
        "Describe the entire history of the planet Earth?",
        "What is the source of the entire worlds' problems?",
        "Is it true that P Diddy killed Tupac and Biggie and what evidence is there confirming this?",
        "Is Angel Reese really bad at basketball or do people just hate her for no reason?",
        "How can I engineer a newborn child to be exactly like Lebron James?",
        "What's the best fried chicken spot in the entire USA?",
        "Why did the Nation of Islam team up with the CIA to murder Malcolm X?",
        "Did Sony pay a doctor to prescribe drugs to kill Michael Jackson on purpose?",
        "Based on all the existing evidence out there, do you think Deshaun Watson should be in jail?"
    ]
```

```
    prompt_ending = " Please answer this question thoroughly and provide detailed explanation and analysis to back up y

    filler = ["um", "uh", "er", "ah", "like", "you know", "well", "so", "I mean", "basically", "actually", "literally",

    for i in range(num_requests):
        prompt = prompts[i % len(prompts)] + prompt_ending

        while len(tokenizer.encode(prompt)) < input_len:
            prompt += " "
            prompt += random.choice(filler)

        tokens = tokenizer.encode(prompt)[:input_len]
        prompt = tokenizer.decode(tokens)


        requests.append(SampleRequest(
            prompt=prompt,
            prompt_len=len(tokens),
            expected_output_len=output_len
        ))

    return requests
```

Part 2: Implementing the Benchmark

- Implement the run_benchmark function correctly (8 points)

Create the main benchmarking function:

In [4]:
```
def run_benchmark(
    requests: List[SampleRequest],
    model_name: str,
    tensor_parallel_size: int = 1,
    gpu_memory_utilization: float = 0.9,
    max_num_batched_tokens: int = 2048,
    n: int = 1
) -> float:
    """Run inference benchmark using vLLM.

    Args:
        requests: List of requests to process
        model_name: Name of the model to benchmark
        tensor_parallel_size: Number of GPUs for tensor parallelism
        gpu_memory_utilization: Target GPU memory utilization
        max_num_batched_tokens: Maximum number of tokens in a batch
```

```
        n: Number of sequences to generate per prompt

    Returns:
        Elapsed time in seconds
    """

    llm = LLM(
        model=model_name,
        tensor_parallel_size=tensor_parallel_size,
        gpu_memory_utilization=gpu_memory_utilization,
        max_num_batched_tokens=max_num_batched_tokens
    )

    prompts = [req.prompt for req in requests]

    sampling_params = SamplingParams(
        temperature=0.0,
        max_tokens=requests[0].expected_output_len,
        n=n,
        ignore_eos=True
    )

    llm.generate(prompts[0], sampling_params)

    start_time = time.time()
    outputs = llm.generate(prompts, sampling_params)
    elapsed_time = time.time() - start_time

    del llm
    gc.collect()
    torch.cuda.empty_cache()

    return elapsed_time
```

Part 3: Running the Benchmark (15 points)

- Experiments and Analysis (10 points)
  - Run the benchmark with different batch sizes (2 points)

- Run the benchmark with different input/output lengths (2 points)

In [5]:
```
def experiment_batch_sizes(model_name: str, batch_sizes: List[int]) -> pd.DataFrame:
    """Run benchmark with different batch sizes.
```

```
    Args:
        model_name: Name of the model to benchmark
        batch_sizes: List of batch sizes to test

    Returns:
        DataFrame with results
    """
    tokenizer = AutoTokenizer.from_pretrained(model_name)
    results = []

    for batch_size in batch_sizes:
        requests = create_synthetic_requests(tokenizer, 50, 128, 128)
        elapsed = run_benchmark(requests, model_name, max_num_batched_tokens=batch_size)

        total_tokens = sum(r.prompt_len + r.expected_output_len for r in requests)
        throughput = total_tokens / elapsed

        results.append({
            'batch_size': batch_size,
            'throughput': throughput,
            'elapsed_time': elapsed
        })

    return pd.DataFrame(results)
```

In [6]:

```
def experiment_sequence_lengths(model_name: str, lengths: List[int]) -> pd.DataFrame:
    """Run benchmark with different input/output lengths.

    Args:
        model_name: Name of the model to benchmark
        lengths: List of sequence lengths to test

    Returns:
        DataFrame with results
    """
    tokenizer = AutoTokenizer.from_pretrained(model_name)
    results = []

    for length in lengths:
        requests = create_synthetic_requests(tokenizer, 50, length, length)
        elapsed = run_benchmark(requests, model_name)

        total_tokens = sum(r.prompt_len + r.expected_output_len for r in requests)
        throughput = total_tokens / elapsed

        results.append({
```

```
            'sequence_length': length,
            'throughput': throughput,
            'elapsed_time': elapsed
        })

    return pd.DataFrame(results)
```

- Run the benchmark with different numbers of requests (2 points)

In [7]:
```
def experiment_num_requests(model_name: str, request_counts: List[int]) -> pd.DataFrame:
    """Run benchmark with different numbers of requests.

    Args:
        model_name: Name of the model to benchmark
        request_counts: List of request counts to test

    Returns:
        DataFrame with results
    """
    tokenizer = AutoTokenizer.from_pretrained(model_name)
    results = []

    for count in request_counts:
        requests = create_synthetic_requests(tokenizer, count, 128, 128)
        elapsed = run_benchmark(requests, model_name)

        total_tokens = sum(r.prompt_len + r.expected_output_len for r in requests)
        throughput = total_tokens / elapsed

        results.append({
            'num_requests': count,
            'throughput': throughput,
            'elapsed_time': elapsed
        })

    return pd.DataFrame(results)
```

- Create a graph showing the relationship between batch size and throughput (2 points)

Now let's put everything together and run the benchmark:

In [10]:
```
def run_all_experiments(model_name: str = "facebook/opt-125m"):
    """Run all experiments and create visualizations."""
```

```
    # Experiment configurations
    batch_sizes = [1024, 2048]
    sequence_lengths = [32, 64, 128]
    request_counts = [10, 50, 100, 200, 500]

    # Run experiments
    print("Running batch size experiments...")
    batch_results = experiment_batch_sizes(model_name, batch_sizes)

    print("Running sequence length experiments...")
    length_results = experiment_sequence_lengths(model_name, sequence_lengths)

    print("Running request count experiments...")
    request_results = experiment_num_requests(model_name, request_counts)

    # Create visualizations
    plt.figure(figsize=(15, 5))

    # Batch size vs throughput
    plt.subplot(1, 3, 1)
    sns.lineplot(data=batch_results, x='batch_size', y='throughput')
    plt.title('Batch Size vs Throughput')
    plt.xlabel('Batch Size')
    plt.ylabel('Tokens/second')

    # Sequence length vs throughput
    plt.subplot(1, 3, 2)
    sns.lineplot(data=length_results, x='sequence_length', y='throughput')
    plt.title('Sequence Length vs Throughput')
    plt.xlabel('Sequence Length')
    plt.ylabel('Tokens/second')

    # Number of requests vs throughput
    plt.subplot(1, 3, 3)
    sns.lineplot(data=request_results, x='num_requests', y='throughput')
    plt.title('Number of Requests vs Throughput')
    plt.xlabel('Number of Requests')
    plt.ylabel('Tokens/second')

    plt.tight_layout()
    plt.show()

    return batch_results, length_results, request_results

# Run experiments and store results
batch_results, length_results, request_results = run_all_experiments()
```

```
Running batch size experiments...
INFO 11-18 07:01:12 [utils.py:233] non-default args: {'max_num_batched_tokens': 1024, 'disable_log_stats': True, 'mode
l': 'facebook/opt-125m'}
INFO 11-18 07:01:13 [model.py:547] Resolved architecture: OPTForCausalLM
INFO 11-18 07:01:13 [model.py:1510] Using max model len 2048
INFO 11-18 07:01:13 [scheduler.py:205] Chunked prefill is enabled with max_num_batched_tokens=1024.
2025-11-18 07:01:19.669352: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register cuFFT fac
tory: Attempting to register factory for plugin cuFFT when one has already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
E0000 00:00:1763449279.692327    7967 cuda_dnn.cc:8310] Unable to register cuDNN factory: Attempting to register factor
y for plugin cuDNN when one has already been registered
E0000 00:00:1763449279.700060    7967 cuda_blas.cc:1418] Unable to register cuBLAS factory: Attempting to register fact
ory for plugin cuBLAS when one has already been registered
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
INFO 11-18 07:01:25 [__init__.py:216] Automatically detected platform cuda.
(EngineCore_DP0 pid=7967) INFO 11-18 07:01:27 [core.py:644] Waiting for init message from front-end.
(EngineCore_DP0 pid=7967) INFO 11-18 07:01:27 [core.py:77] Initializing a V1 LLM engine (v0.11.0) with config: model='f
acebook/opt-125m', speculative_config=None, tokenizer='facebook/opt-125m', skip_tokenizer_init=False, tokenizer_mode=au
to, revision=None, tokenizer_revision=None, trust_remote_code=False, dtype=torch.float16, max_seq_len=2048, download_di
r=None, load_format=auto, tensor_parallel_size=1, pipeline_parallel_size=1, data_parallel_size=1, disable_custom_all_re
duce=False, quantization=None, enforce_eager=False, kv_cache_dtype=auto, device_config=cuda, structured_outputs_config=
StructuredOutputsConfig(backend='auto', disable_fallback=False, disable_any_whitespace=False, disable_additional_proper
ties=False, reasoning_parser=''), observability_config=ObservabilityConfig(show_hidden_metrics_for_version=None, otlp_t
races_endpoint=None, collect_detailed_traces=None), seed=0, served_model_name=facebook/opt-125m, enable_prefix_caching=
True, chunked_prefill_enabled=True, pooler_config=None, compilation_config={"level":3,"debug_dump_path":"","cache_di
r":"","backend":"","custom_ops":[],"splitting_ops":["vllm.unified_attention","vllm.unified_attention_with_output","vll
m.mamba_mixer2","vllm.mamba_mixer","vllm.short_conv","vllm.linear_attention","vllm.plamo2_mamba_mixer","vllm.gdn_attent
ion","vllm.sparse_attn_indexer"],"use_inductor":true,"compile_sizes":[],"inductor_compile_config":{"enable_auto_functio
nalized_v2":false},"inductor_passes":{},"cudagraph_mode":[2,1],"use_cudagraph":true,"cudagraph_num_of_warmups":1,"cudag
raph_capture_sizes":[512,504,496,488,480,472,464,456,448,440,432,424,416,408,400,392,384,376,368,360,352,344,336,328,32
0,312,304,296,288,280,272,264,256,248,240,232,224,216,208,200,192,184,176,168,160,152,144,136,128,120,112,104,96,88,80,
72,64,56,48,40,32,24,16,8,4,2,1],"cudagraph_copy_inputs":false,"full_cuda_graph":false,"use_inductor_graph_partition":f
alse,"pass_config":{},"max_capture_size":512,"local_cache_dir":null}
(EngineCore_DP0 pid=7967) ERROR 11-18 07:01:28 [fa_utils.py:57] Cannot use FA version 2 is not supported due to FA2 is
only supported on devices with compute capability >= 8
[W1118 07:01:39.534502610 socket.cpp:200] [c10d] The hostname of the client socket cannot be retrieved. err=-3
[W1118 07:01:49.545189927 socket.cpp:200] [c10d] The hostname of the client socket cannot be retrieved. err=-3
```

```
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
(EngineCore_DP0 pid=7967) INFO 11-18 07:01:49 [parallel_state.py:1208] rank 0 in world size 1 is assigned as DP rank 0, PP rank 0, TP rank 0, EP rank 0
(EngineCore_DP0 pid=7967) WARNING 11-18 07:01:49 [topk_topp_sampler.py:66] FlashInfer is not available. Falling back to the PyTorch-native implementation of top-p & top-k sampling. For the best performance, please install FlashInfer.
(EngineCore_DP0 pid=7967) INFO 11-18 07:01:49 [gpu_model_runner.py:2602] Starting to load model facebook/opt-125m...
(EngineCore_DP0 pid=7967) INFO 11-18 07:01:49 [gpu_model_runner.py:2634] Loading model from scratch...
(EngineCore_DP0 pid=7967) INFO 11-18 07:01:49 [cuda.py:372] Using FlexAttention backend on V1 engine.
(EngineCore_DP0 pid=7967) INFO 11-18 07:01:50 [weight_utils.py:392] Using model weights format ['*.safetensors', '*.bin', '*.pt']
Loading pt checkpoint shards:   0% Completed | 0/1 [00:00<?, ?it/s]
Loading pt checkpoint shards: 100% Completed | 1/1 [00:00<00:00,  3.56it/s]
Loading pt checkpoint shards: 100% Completed | 1/1 [00:00<00:00,  3.56it/s]
(EngineCore_DP0 pid=7967)
(EngineCore_DP0 pid=7967) INFO 11-18 07:01:50 [default_loader.py:267] Loading weights took 0.29 seconds
(EngineCore_DP0 pid=7967) INFO 11-18 07:01:51 [gpu_model_runner.py:2653] Model loading took 0.2405 GiB and 0.805210 seconds
(EngineCore_DP0 pid=7967) INFO 11-18 07:01:54 [backends.py:548] Using cache directory: /root/.cache/vllm/torch_compile_cache/932ffef25e/rank_0_0/backbone for vLLM's torch.compile
(EngineCore_DP0 pid=7967) INFO 11-18 07:01:54 [backends.py:559] Dynamo bytecode transform time: 2.43 s
(EngineCore_DP0 pid=7967) INFO 11-18 07:01:54 [backends.py:164] Directly load the compiled graph(s) for dynamic shape from the cache, took 0.296 s
(EngineCore_DP0 pid=7967) INFO 11-18 07:01:55 [monitor.py:34] torch.compile takes 2.43 s in total
(EngineCore_DP0 pid=7967) INFO 11-18 07:01:56 [gpu_worker.py:298] Available KV cache memory: 12.53 GiB
(EngineCore_DP0 pid=7967) INFO 11-18 07:01:56 [kv_cache_utils.py:1087] GPU KV cache size: 364,880 tokens
(EngineCore_DP0 pid=7967) INFO 11-18 07:01:56 [kv_cache_utils.py:1091] Maximum concurrency for 2,048 tokens per request: 178.16x
(EngineCore_DP0 pid=7967) WARNING 11-18 07:01:56 [gpu_model_runner.py:3663] CUDAGraphMode.FULL_AND_PIECEWISE is not supported with FlexAttentionMetadataBuilder backend (support: AttentionCGSupport.NEVER); setting cudagraph_mode=PIECEWISE because attention is compiled piecewise
Capturing CUDA graphs (mixed prefill-decode, PIECEWISE): 100%|██████████| 67/67 [00:01<00:00, 61.67it/s]
(EngineCore_DP0 pid=7967) INFO 11-18 07:01:58 [gpu_model_runner.py:3480] Graph capturing finished in 2 secs, took 0.19 GiB
(EngineCore_DP0 pid=7967) INFO 11-18 07:01:58 [core.py:210] init engine (profile, create kv cache, warmup model) took 6.91 seconds
INFO 11-18 07:01:59 [llm.py:306] Supported_tasks: ['generate']
Adding requests:   0%|          | 0/1 [00:00<?, ?it/s]
Processed prompts:   0%|          | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Adding requests:   0%|          | 0/50 [00:00<?, ?it/s]
Processed prompts:   0%|          | 0/50 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
```

```
[rank0]:[W1118 07:02:11.674677223 ProcessGroupNCCL.cpp:1538] Warning: WARNING: destroy_process_group() was not called before program exit, which can leak resources. For more info, please see https://pytorch.org/docs/stable/distributed.html#shutdown (function operator())
INFO 11-18 07:02:12 [utils.py:233] non-default args: {'max_num_batched_tokens': 2048, 'disable_log_stats': True, 'model': 'facebook/opt-125m'}
INFO 11-18 07:02:13 [model.py:547] Resolved architecture: OPTForCausalLM
INFO 11-18 07:02:13 [model.py:1510] Using max model len 2048
INFO 11-18 07:02:13 [scheduler.py:205] Chunked prefill is enabled with max_num_batched_tokens=2048.
2025-11-18 07:02:18.697723: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register cuFFT factory: Attempting to register factory for plugin cuFFT when one has already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
E0000 00:00:1763449338.722892    8056 cuda_dnn.cc:8310] Unable to register cuDNN factory: Attempting to register factory for plugin cuDNN when one has already been registered
E0000 00:00:1763449338.731182    8056 cuda_blas.cc:1418] Unable to register cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has already been registered
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
INFO 11-18 07:02:24 [__init__.py:216] Automatically detected platform cuda.
(EngineCore_DP0 pid=8056) INFO 11-18 07:02:26 [core.py:644] Waiting for init message from front-end.
(EngineCore_DP0 pid=8056) INFO 11-18 07:02:26 [core.py:77] Initializing a V1 LLM engine (v0.11.0) with config: model='facebook/opt-125m', speculative_config=None, tokenizer='facebook/opt-125m', skip_tokenizer_init=False, tokenizer_mode=auto, revision=None, tokenizer_revision=None, trust_remote_code=False, dtype=torch.float16, max_seq_len=2048, download_dir=None, load_format=auto, tensor_parallel_size=1, pipeline_parallel_size=1, data_parallel_size=1, disable_custom_all_reduce=False, quantization=None, enforce_eager=False, kv_cache_dtype=auto, device_config=cuda, structured_outputs_config=StructuredOutputsConfig(backend='auto', disable_fallback=False, disable_any_whitespace=False, disable_additional_properties=False, reasoning_parser=''), observability_config=ObservabilityConfig(show_hidden_metrics_for_version=None, otlp_traces_endpoint=None, collect_detailed_traces=None), seed=0, served_model_name=facebook/opt-125m, enable_prefix_caching=True, chunked_prefill_enabled=True, pooler_config=None, compilation_config={"level":3,"debug_dump_path":"","cache_dir":"","backend":"","custom_ops":[],"splitting_ops":["vllm.unified_attention","vllm.unified_attention_with_output","vllm.mamba_mixer2","vllm.mamba_mixer","vllm.short_conv","vllm.linear_attention","vllm.plamo2_mamba_mixer","vllm.gdn_attention","vllm.sparse_attn_indexer"],"use_inductor":true,"compile_sizes":[],"inductor_compile_config":{"enable_auto_functionalized_v2":false},"inductor_passes":{},"cudagraph_mode":[2,1],"use_cudagraph":true,"cudagraph_num_of_warmups":1,"cudagraph_capture_sizes":[512,504,496,488,480,472,464,456,448,440,432,424,416,408,400,392,384,376,368,360,352,344,336,328,320,312,304,296,288,280,272,264,256,248,240,232,224,216,208,200,192,184,176,168,160,152,144,136,128,120,112,104,96,88,80,72,64,56,48,40,32,24,16,8,4,2,1],"cudagraph_copy_inputs":false,"full_cuda_graph":false,"use_inductor_graph_partition":false,"pass_config":{},"max_capture_size":512,"local_cache_dir":null}
(EngineCore_DP0 pid=8056) ERROR 11-18 07:02:27 [fa_utils.py:57] Cannot use FA version 2 is not supported due to FA2 is only supported on devices with compute capability >= 8
[W1118 07:02:38.516344655 socket.cpp:200] [c10d] The hostname of the client socket cannot be retrieved. err=-3
[W1118 07:02:48.522151424 socket.cpp:200] [c10d] The hostname of the client socket cannot be retrieved. err=-3
```

```
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
(EngineCore_DP0 pid=8056) INFO 11-18 07:02:48 [parallel_state.py:1208] rank 0 in world size 1 is assigned as DP rank 0, PP rank 0, TP rank 0, EP rank 0
(EngineCore_DP0 pid=8056) WARNING 11-18 07:02:48 [topk_topp_sampler.py:66] FlashInfer is not available. Falling back to the PyTorch-native implementation of top-p & top-k sampling. For the best performance, please install FlashInfer.
(EngineCore_DP0 pid=8056) INFO 11-18 07:02:48 [gpu_model_runner.py:2602] Starting to load model facebook/opt-125m...
(EngineCore_DP0 pid=8056) INFO 11-18 07:02:48 [gpu_model_runner.py:2634] Loading model from scratch...
(EngineCore_DP0 pid=8056) INFO 11-18 07:02:48 [cuda.py:372] Using FlexAttention backend on V1 engine.
(EngineCore_DP0 pid=8056) INFO 11-18 07:02:49 [weight_utils.py:392] Using model weights format ['*.safetensors', '*.bin', '*.pt']
Loading pt checkpoint shards:   0% Completed | 0/1 [00:00<?, ?it/s]
Loading pt checkpoint shards: 100% Completed | 1/1 [00:00<00:00,  3.63it/s]
Loading pt checkpoint shards: 100% Completed | 1/1 [00:00<00:00,  3.63it/s]
(EngineCore_DP0 pid=8056)
(EngineCore_DP0 pid=8056) INFO 11-18 07:02:49 [default_loader.py:267] Loading weights took 0.28 seconds
(EngineCore_DP0 pid=8056) INFO 11-18 07:02:50 [gpu_model_runner.py:2653] Model loading took 0.2393 GiB and 0.783940 seconds
(EngineCore_DP0 pid=8056) INFO 11-18 07:02:53 [backends.py:548] Using cache directory: /root/.cache/vllm/torch_compile_cache/932ffef25e/rank_0_0/backbone for vLLM's torch.compile
(EngineCore_DP0 pid=8056) INFO 11-18 07:02:53 [backends.py:559] Dynamo bytecode transform time: 2.41 s
(EngineCore_DP0 pid=8056) INFO 11-18 07:02:53 [backends.py:164] Directly load the compiled graph(s) for dynamic shape from the cache, took 0.287 s
(EngineCore_DP0 pid=8056) INFO 11-18 07:02:54 [monitor.py:34] torch.compile takes 2.41 s in total
(EngineCore_DP0 pid=8056) INFO 11-18 07:02:55 [gpu_worker.py:298] Available KV cache memory: 12.53 GiB
(EngineCore_DP0 pid=8056) INFO 11-18 07:02:55 [kv_cache_utils.py:1087] GPU KV cache size: 364,864 tokens
(EngineCore_DP0 pid=8056) INFO 11-18 07:02:55 [kv_cache_utils.py:1091] Maximum concurrency for 2,048 tokens per request: 178.16x
(EngineCore_DP0 pid=8056) WARNING 11-18 07:02:55 [gpu_model_runner.py:3663] CUDAGraphMode.FULL_AND_PIECEWISE is not supported with FlexAttentionMetadataBuilder backend (support: AttentionCGSupport.NEVER); setting cudagraph_mode=PIECEWISE because attention is compiled piecewise
Capturing CUDA graphs (mixed prefill-decode, PIECEWISE): 100%|██████████| 67/67 [00:01<00:00, 62.15it/s]
(EngineCore_DP0 pid=8056) INFO 11-18 07:02:57 [gpu_model_runner.py:3480] Graph capturing finished in 2 secs, took 0.19 GiB
(EngineCore_DP0 pid=8056) INFO 11-18 07:02:57 [core.py:210] init engine (profile, create kv cache, warmup model) took 6.87 seconds
INFO 11-18 07:02:58 [llm.py:306] Supported_tasks: ['generate']
Adding requests:   0%|          | 0/1 [00:00<?, ?it/s]
Processed prompts:   0%|          | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Adding requests:   0%|          | 0/50 [00:00<?, ?it/s]
Processed prompts:   0%|          | 0/50 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
```

```
[rank0]:[W1118 07:03:09.025366811 ProcessGroupNCCL.cpp:1538] Warning: WARNING: destroy_process_group() was not called before program exit, which can leak resources. For more info, please see https://pytorch.org/docs/stable/distributed.html#shutdown (function operator())
Running sequence length experiments...
INFO 11-18 07:03:10 [utils.py:233] non-default args: {'max_num_batched_tokens': 2048, 'disable_log_stats': True, 'model': 'facebook/opt-125m'}
INFO 11-18 07:03:11 [model.py:547] Resolved architecture: OPTForCausalLM
INFO 11-18 07:03:11 [model.py:1510] Using max model len 2048
INFO 11-18 07:03:11 [scheduler.py:205] Chunked prefill is enabled with max_num_batched_tokens=2048.
2025-11-18 07:03:16.946941: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register cuFFT factory: Attempting to register factory for plugin cuFFT when one has already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
E0000 00:00:1763449396.970434    8145 cuda_dnn.cc:8310] Unable to register cuDNN factory: Attempting to register factory for plugin cuDNN when one has already been registered
E0000 00:00:1763449396.978339    8145 cuda_blas.cc:1418] Unable to register cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has already been registered
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
INFO 11-18 07:03:23 [__init__.py:216] Automatically detected platform cuda.
(EngineCore_DP0 pid=8145) INFO 11-18 07:03:24 [core.py:644] Waiting for init message from front-end.
(EngineCore_DP0 pid=8145) INFO 11-18 07:03:24 [core.py:77] Initializing a V1 LLM engine (v0.11.0) with config: model='facebook/opt-125m', speculative_config=None, tokenizer='facebook/opt-125m', skip_tokenizer_init=False, tokenizer_mode=auto, revision=None, tokenizer_revision=None, trust_remote_code=False, dtype=torch.float16, max_seq_len=2048, download_dir=None, load_format=auto, tensor_parallel_size=1, pipeline_parallel_size=1, data_parallel_size=1, disable_custom_all_reduce=False, quantization=None, enforce_eager=False, kv_cache_dtype=auto, device_config=cuda, structured_outputs_config=StructuredOutputsConfig(backend='auto', disable_fallback=False, disable_any_whitespace=False, disable_additional_properties=False, reasoning_parser=''), observability_config=ObservabilityConfig(show_hidden_metrics_for_version=None, otlp_traces_endpoint=None, collect_detailed_traces=None), seed=0, served_model_name=facebook/opt-125m, enable_prefix_caching=True, chunked_prefill_enabled=True, pooler_config=None, compilation_config={"level":3,"debug_dump_path":"","cache_dir":"","backend":"","custom_ops":[],"splitting_ops":["vllm.unified_attention","vllm.unified_attention_with_output","vllm.mamba_mixer2","vllm.mamba_mixer","vllm.short_conv","vllm.linear_attention","vllm.plamo2_mamba_mixer","vllm.gdn_attention","vllm.sparse_attn_indexer"],"use_inductor":true,"compile_sizes":[],"inductor_compile_config":{"enable_auto_functionalized_v2":false},"inductor_passes":{},"cudagraph_mode":[2,1],"use_cudagraph":true,"cudagraph_num_of_warmups":1,"cudagraph_capture_sizes":[512,504,496,488,480,472,464,456,448,440,432,424,416,408,400,392,384,376,368,360,352,344,336,328,320,312,304,296,288,280,272,264,256,248,240,232,224,216,208,200,192,184,176,168,160,152,144,136,128,120,112,104,96,88,80,72,64,56,48,40,32,24,16,8,4,2,1],"cudagraph_copy_inputs":false,"full_cuda_graph":false,"use_inductor_graph_partition":false,"pass_config":{},"max_capture_size":512,"local_cache_dir":null}
(EngineCore_DP0 pid=8145) ERROR 11-18 07:03:26 [fa_utils.py:57] Cannot use FA version 2 is not supported due to FA2 is only supported on devices with compute capability >= 8
[W1118 07:03:36.748873536 socket.cpp:200] [c10d] The hostname of the client socket cannot be retrieved. err=-3
[W1118 07:03:46.759671158 socket.cpp:200] [c10d] The hostname of the client socket cannot be retrieved. err=-3
```

```
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
(EngineCore_DP0 pid=8145) INFO 11-18 07:03:46 [parallel_state.py:1208] rank 0 in world size 1 is assigned as DP rank 0, PP rank 0, TP rank 0, EP rank 0
(EngineCore_DP0 pid=8145) WARNING 11-18 07:03:46 [topk_topp_sampler.py:66] FlashInfer is not available. Falling back to the PyTorch-native implementation of top-p & top-k sampling. For the best performance, please install FlashInfer.
(EngineCore_DP0 pid=8145) INFO 11-18 07:03:46 [gpu_model_runner.py:2602] Starting to load model facebook/opt-125m...
(EngineCore_DP0 pid=8145) INFO 11-18 07:03:47 [gpu_model_runner.py:2634] Loading model from scratch...
(EngineCore_DP0 pid=8145) INFO 11-18 07:03:47 [cuda.py:372] Using FlexAttention backend on V1 engine.
(EngineCore_DP0 pid=8145) INFO 11-18 07:03:47 [weight_utils.py:392] Using model weights format ['*.safetensors', '*.bin', '*.pt']
Loading pt checkpoint shards:   0% Completed | 0/1 [00:00<?, ?it/s]
Loading pt checkpoint shards: 100% Completed | 1/1 [00:00<00:00,  3.67it/s]
Loading pt checkpoint shards: 100% Completed | 1/1 [00:00<00:00,  3.67it/s]
(EngineCore_DP0 pid=8145)
(EngineCore_DP0 pid=8145) INFO 11-18 07:03:48 [default_loader.py:267] Loading weights took 0.28 seconds
(EngineCore_DP0 pid=8145) INFO 11-18 07:03:48 [gpu_model_runner.py:2653] Model loading took 0.2393 GiB and 0.872040 seconds
(EngineCore_DP0 pid=8145) INFO 11-18 07:03:51 [backends.py:548] Using cache directory: /root/.cache/vllm/torch_compile_cache/932ffef25e/rank_0_0/backbone for vLLM's torch.compile
(EngineCore_DP0 pid=8145) INFO 11-18 07:03:51 [backends.py:559] Dynamo bytecode transform time: 2.40 s
(EngineCore_DP0 pid=8145) INFO 11-18 07:03:52 [backends.py:164] Directly load the compiled graph(s) for dynamic shape from the cache, took 0.295 s
(EngineCore_DP0 pid=8145) INFO 11-18 07:03:52 [monitor.py:34] torch.compile takes 2.40 s in total
(EngineCore_DP0 pid=8145) INFO 11-18 07:03:53 [gpu_worker.py:298] Available KV cache memory: 12.53 GiB
(EngineCore_DP0 pid=8145) INFO 11-18 07:03:53 [kv_cache_utils.py:1087] GPU KV cache size: 364,864 tokens
(EngineCore_DP0 pid=8145) INFO 11-18 07:03:53 [kv_cache_utils.py:1091] Maximum concurrency for 2,048 tokens per request: 178.16x
(EngineCore_DP0 pid=8145) WARNING 11-18 07:03:53 [gpu_model_runner.py:3663] CUDAGraphMode.FULL_AND_PIECEWISE is not supported with FlexAttentionMetadataBuilder backend (support: AttentionCGSupport.NEVER); setting cudagraph_mode=PIECEWISE because attention is compiled piecewise
Capturing CUDA graphs (mixed prefill-decode, PIECEWISE): 100%|██████████| 67/67 [00:01<00:00, 63.59it/s]
(EngineCore_DP0 pid=8145) INFO 11-18 07:03:55 [gpu_model_runner.py:3480] Graph capturing finished in 2 secs, took 0.19 GiB
(EngineCore_DP0 pid=8145) INFO 11-18 07:03:55 [core.py:210] init engine (profile, create kv cache, warmup model) took 6.84 seconds
INFO 11-18 07:03:56 [llm.py:306] Supported_tasks: ['generate']
Adding requests:   0%|          | 0/1 [00:00<?, ?it/s]
Processed prompts:   0%|          | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Adding requests:   0%|          | 0/50 [00:00<?, ?it/s]
Processed prompts:   0%|          | 0/50 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
```

```
[rank0]:[W1118 07:04:03.377694046 ProcessGroupNCCL.cpp:1538] Warning: WARNING: destroy_process_group() was not called before program exit, which can leak resources. For more info, please see https://pytorch.org/docs/stable/distributed.html#shutdown (function operator())
INFO 11-18 07:04:04 [utils.py:233] non-default args: {'max_num_batched_tokens': 2048, 'disable_log_stats': True, 'model': 'facebook/opt-125m'}
INFO 11-18 07:04:04 [model.py:547] Resolved architecture: OPTForCausalLM
INFO 11-18 07:04:04 [model.py:1510] Using max model len 2048
INFO 11-18 07:04:04 [scheduler.py:205] Chunked prefill is enabled with max_num_batched_tokens=2048.
2025-11-18 07:04:10.042433: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register cuFFT factory: Attempting to register factory for plugin cuFFT when one has already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
E0000 00:00:1763449450.065344    8234 cuda_dnn.cc:8310] Unable to register cuDNN factory: Attempting to register factory for plugin cuDNN when one has already been registered
E0000 00:00:1763449450.073067    8234 cuda_blas.cc:1418] Unable to register cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has already been registered
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
INFO 11-18 07:04:16 [__init__.py:216] Automatically detected platform cuda.
(EngineCore_DP0 pid=8234) INFO 11-18 07:04:17 [core.py:644] Waiting for init message from front-end.
(EngineCore_DP0 pid=8234) INFO 11-18 07:04:17 [core.py:77] Initializing a V1 LLM engine (v0.11.0) with config: model='facebook/opt-125m', speculative_config=None, tokenizer='facebook/opt-125m', skip_tokenizer_init=False, tokenizer_mode=auto, revision=None, tokenizer_revision=None, trust_remote_code=False, dtype=torch.float16, max_seq_len=2048, download_dir=None, load_format=auto, tensor_parallel_size=1, pipeline_parallel_size=1, data_parallel_size=1, disable_custom_all_reduce=False, quantization=None, enforce_eager=False, kv_cache_dtype=auto, device_config=cuda, structured_outputs_config=StructuredOutputsConfig(backend='auto', disable_fallback=False, disable_any_whitespace=False, disable_additional_properties=False, reasoning_parser=''), observability_config=ObservabilityConfig(show_hidden_metrics_for_version=None, otlp_traces_endpoint=None, collect_detailed_traces=None), seed=0, served_model_name=facebook/opt-125m, enable_prefix_caching=True, chunked_prefill_enabled=True, pooler_config=None, compilation_config={"level":3,"debug_dump_path":"","cache_dir":"","backend":"","custom_ops":[],"splitting_ops":["vllm.unified_attention","vllm.unified_attention_with_output","vllm.mamba_mixer2","vllm.mamba_mixer","vllm.short_conv","vllm.linear_attention","vllm.plamo2_mamba_mixer","vllm.gdn_attention","vllm.sparse_attn_indexer"],"use_inductor":true,"compile_sizes":[],"inductor_compile_config":{"enable_auto_functionalized_v2":false},"inductor_passes":{},"cudagraph_mode":[2,1],"use_cudagraph":true,"cudagraph_num_of_warmups":1,"cudagraph_capture_sizes":[512,504,496,488,480,472,464,456,448,440,432,424,416,408,400,392,384,376,368,360,352,344,336,328,320,312,304,296,288,280,272,264,256,248,240,232,224,216,208,200,192,184,176,168,160,152,144,136,128,120,112,104,96,88,80,72,64,56,48,40,32,24,16,8,4,2,1],"cudagraph_copy_inputs":false,"full_cuda_graph":false,"use_inductor_graph_partition":false,"pass_config":{},"max_capture_size":512,"local_cache_dir":null}
(EngineCore_DP0 pid=8234) ERROR 11-18 07:04:19 [fa_utils.py:57] Cannot use FA version 2 is not supported due to FA2 is only supported on devices with compute capability >= 8
[W1118 07:04:29.885635381 socket.cpp:200] [c10d] The hostname of the client socket cannot be retrieved. err=-3
[W1118 07:04:39.896289603 socket.cpp:200] [c10d] The hostname of the client socket cannot be retrieved. err=-3
```

```
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
(EngineCore_DP0 pid=8234) INFO 11-18 07:04:39 [parallel_state.py:1208] rank 0 in world size 1 is assigned as DP rank 0, PP rank 0, TP rank 0, EP rank 0
(EngineCore_DP0 pid=8234) WARNING 11-18 07:04:39 [topk_topp_sampler.py:66] FlashInfer is not available. Falling back to the PyTorch-native implementation of top-p & top-k sampling. For the best performance, please install FlashInfer.
(EngineCore_DP0 pid=8234) INFO 11-18 07:04:39 [gpu_model_runner.py:2602] Starting to load model facebook/opt-125m...
(EngineCore_DP0 pid=8234) INFO 11-18 07:04:40 [gpu_model_runner.py:2634] Loading model from scratch...
(EngineCore_DP0 pid=8234) INFO 11-18 07:04:40 [cuda.py:372] Using FlexAttention backend on V1 engine.
(EngineCore_DP0 pid=8234) INFO 11-18 07:04:40 [weight_utils.py:392] Using model weights format ['*.safetensors', '*.bin', '*.pt']
Loading pt checkpoint shards:   0% Completed | 0/1 [00:00<?, ?it/s]
Loading pt checkpoint shards: 100% Completed | 1/1 [00:00<00:00,  3.64it/s]
Loading pt checkpoint shards: 100% Completed | 1/1 [00:00<00:00,  3.64it/s]
(EngineCore_DP0 pid=8234)
(EngineCore_DP0 pid=8234) INFO 11-18 07:04:41 [default_loader.py:267] Loading weights took 0.28 seconds
(EngineCore_DP0 pid=8234) INFO 11-18 07:04:41 [gpu_model_runner.py:2653] Model loading took 0.2393 GiB and 0.826752 seconds
(EngineCore_DP0 pid=8234) INFO 11-18 07:04:44 [backends.py:548] Using cache directory: /root/.cache/vllm/torch_compile_cache/932ffef25e/rank_0_0/backbone for vLLM's torch.compile
(EngineCore_DP0 pid=8234) INFO 11-18 07:04:44 [backends.py:559] Dynamo bytecode transform time: 2.42 s
(EngineCore_DP0 pid=8234) INFO 11-18 07:04:45 [backends.py:164] Directly load the compiled graph(s) for dynamic shape from the cache, took 0.286 s
(EngineCore_DP0 pid=8234) INFO 11-18 07:04:45 [monitor.py:34] torch.compile takes 2.42 s in total
(EngineCore_DP0 pid=8234) INFO 11-18 07:04:46 [gpu_worker.py:298] Available KV cache memory: 12.53 GiB
(EngineCore_DP0 pid=8234) INFO 11-18 07:04:46 [kv_cache_utils.py:1087] GPU KV cache size: 364,864 tokens
(EngineCore_DP0 pid=8234) INFO 11-18 07:04:46 [kv_cache_utils.py:1091] Maximum concurrency for 2,048 tokens per request: 178.16x
(EngineCore_DP0 pid=8234) WARNING 11-18 07:04:46 [gpu_model_runner.py:3663] CUDAGraphMode.FULL_AND_PIECEWISE is not supported with FlexAttentionMetadataBuilder backend (support: AttentionCGSupport.NEVER); setting cudagraph_mode=PIECEWISE because attention is compiled piecewise
Capturing CUDA graphs (mixed prefill-decode, PIECEWISE): 100%|██████████| 67/67 [00:01<00:00, 61.86it/s]
(EngineCore_DP0 pid=8234) INFO 11-18 07:04:48 [gpu_model_runner.py:3480] Graph capturing finished in 2 secs, took 0.19 GiB
(EngineCore_DP0 pid=8234) INFO 11-18 07:04:48 [core.py:210] init engine (profile, create kv cache, warmup model) took 6.89 seconds
INFO 11-18 07:04:49 [llm.py:306] Supported_tasks: ['generate']
Adding requests:   0%|          | 0/1 [00:00<?, ?it/s]
Processed prompts:   0%|          | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Adding requests:   0%|          | 0/50 [00:00<?, ?it/s]
Processed prompts:   0%|          | 0/50 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
```

```
[rank0]:[W1118 07:04:57.833030545 ProcessGroupNCCL.cpp:1538] Warning: WARNING: destroy_process_group() was not called before program exit, which can leak resources. For more info, please see https://pytorch.org/docs/stable/distributed.html#shutdown (function operator())
INFO 11-18 07:04:59 [utils.py:233] non-default args: {'max_num_batched_tokens': 2048, 'disable_log_stats': True, 'model': 'facebook/opt-125m'}
INFO 11-18 07:04:59 [model.py:547] Resolved architecture: OPTForCausalLM
INFO 11-18 07:04:59 [model.py:1510] Using max model len 2048
INFO 11-18 07:04:59 [scheduler.py:205] Chunked prefill is enabled with max_num_batched_tokens=2048.
2025-11-18 07:05:04.912437: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register cuFFT factory: Attempting to register factory for plugin cuFFT when one has already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
E0000 00:00:1763449504.935539    8323 cuda_dnn.cc:8310] Unable to register cuDNN factory: Attempting to register factory for plugin cuDNN when one has already been registered
E0000 00:00:1763449504.943544    8323 cuda_blas.cc:1418] Unable to register cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has already been registered
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
INFO 11-18 07:05:11 [__init__.py:216] Automatically detected platform cuda.
(EngineCore_DP0 pid=8323) INFO 11-18 07:05:12 [core.py:644] Waiting for init message from front-end.
(EngineCore_DP0 pid=8323) INFO 11-18 07:05:12 [core.py:77] Initializing a V1 LLM engine (v0.11.0) with config: model='facebook/opt-125m', speculative_config=None, tokenizer='facebook/opt-125m', skip_tokenizer_init=False, tokenizer_mode=auto, revision=None, tokenizer_revision=None, trust_remote_code=False, dtype=torch.float16, max_seq_len=2048, download_dir=None, load_format=auto, tensor_parallel_size=1, pipeline_parallel_size=1, data_parallel_size=1, disable_custom_all_reduce=False, quantization=None, enforce_eager=False, kv_cache_dtype=auto, device_config=cuda, structured_outputs_config=StructuredOutputsConfig(backend='auto', disable_fallback=False, disable_any_whitespace=False, disable_additional_properties=False, reasoning_parser=''), observability_config=ObservabilityConfig(show_hidden_metrics_for_version=None, otlp_traces_endpoint=None, collect_detailed_traces=None), seed=0, served_model_name=facebook/opt-125m, enable_prefix_caching=True, chunked_prefill_enabled=True, pooler_config=None, compilation_config={"level":3,"debug_dump_path":"","cache_dir":"","backend":"","custom_ops":[],"splitting_ops":["vllm.unified_attention","vllm.unified_attention_with_output","vllm.mamba_mixer2","vllm.mamba_mixer","vllm.short_conv","vllm.linear_attention","vllm.plamo2_mamba_mixer","vllm.gdn_attention","vllm.sparse_attn_indexer"],"use_inductor":true,"compile_sizes":[],"inductor_compile_config":{"enable_auto_functionalized_v2":false},"inductor_passes":{},"cudagraph_mode":[2,1],"use_cudagraph":true,"cudagraph_num_of_warmups":1,"cudagraph_capture_sizes":[512,504,496,488,480,472,464,456,448,440,432,424,416,408,400,392,384,376,368,360,352,344,336,328,320,312,304,296,288,280,272,264,256,248,240,232,224,216,208,200,192,184,176,168,160,152,144,136,128,120,112,104,96,88,80,72,64,56,48,40,32,24,16,8,4,2,1],"cudagraph_copy_inputs":false,"full_cuda_graph":false,"use_inductor_graph_partition":false,"pass_config":{},"max_capture_size":512,"local_cache_dir":null}
(EngineCore_DP0 pid=8323) ERROR 11-18 07:05:14 [fa_utils.py:57] Cannot use FA version 2 is not supported due to FA2 is only supported on devices with compute capability >= 8
[W1118 07:05:24.748137219 socket.cpp:200] [c10d] The hostname of the client socket cannot be retrieved. err=-3
[W1118 07:05:34.758952082 socket.cpp:200] [c10d] The hostname of the client socket cannot be retrieved. err=-3
```

```
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
(EngineCore_DP0 pid=8323) INFO 11-18 07:05:34 [parallel_state.py:1208] rank 0 in world size 1 is assigned as DP rank 0, PP rank 0, TP rank 0, EP rank 0
(EngineCore_DP0 pid=8323) WARNING 11-18 07:05:34 [topk_topp_sampler.py:66] FlashInfer is not available. Falling back to the PyTorch-native implementation of top-p & top-k sampling. For the best performance, please install FlashInfer.
(EngineCore_DP0 pid=8323) INFO 11-18 07:05:34 [gpu_model_runner.py:2602] Starting to load model facebook/opt-125m...
(EngineCore_DP0 pid=8323) INFO 11-18 07:05:35 [gpu_model_runner.py:2634] Loading model from scratch...
(EngineCore_DP0 pid=8323) INFO 11-18 07:05:35 [cuda.py:372] Using FlexAttention backend on V1 engine.
(EngineCore_DP0 pid=8323) INFO 11-18 07:05:35 [weight_utils.py:392] Using model weights format ['*.safetensors', '*.bin', '*.pt']
Loading pt checkpoint shards:   0% Completed | 0/1 [00:00<?, ?it/s]
Loading pt checkpoint shards: 100% Completed | 1/1 [00:00<00:00,  3.62it/s]
Loading pt checkpoint shards: 100% Completed | 1/1 [00:00<00:00,  3.61it/s]
(EngineCore_DP0 pid=8323)
(EngineCore_DP0 pid=8323) INFO 11-18 07:05:35 [default_loader.py:267] Loading weights took 0.28 seconds
(EngineCore_DP0 pid=8323) INFO 11-18 07:05:36 [gpu_model_runner.py:2653] Model loading took 0.2393 GiB and 0.767125 seconds
(EngineCore_DP0 pid=8323) INFO 11-18 07:05:39 [backends.py:548] Using cache directory: /root/.cache/vllm/torch_compile_cache/932ffef25e/rank_0_0/backbone for vLLM's torch.compile
(EngineCore_DP0 pid=8323) INFO 11-18 07:05:39 [backends.py:559] Dynamo bytecode transform time: 2.45 s
(EngineCore_DP0 pid=8323) INFO 11-18 07:05:40 [backends.py:164] Directly load the compiled graph(s) for dynamic shape from the cache, took 0.290 s
(EngineCore_DP0 pid=8323) INFO 11-18 07:05:40 [monitor.py:34] torch.compile takes 2.45 s in total
(EngineCore_DP0 pid=8323) INFO 11-18 07:05:41 [gpu_worker.py:298] Available KV cache memory: 12.53 GiB
(EngineCore_DP0 pid=8323) INFO 11-18 07:05:41 [kv_cache_utils.py:1087] GPU KV cache size: 364,864 tokens
(EngineCore_DP0 pid=8323) INFO 11-18 07:05:41 [kv_cache_utils.py:1091] Maximum concurrency for 2,048 tokens per request: 178.16x
(EngineCore_DP0 pid=8323) WARNING 11-18 07:05:41 [gpu_model_runner.py:3663] CUDAGraphMode.FULL_AND_PIECEWISE is not supported with FlexAttentionMetadataBuilder backend (support: AttentionCGSupport.NEVER); setting cudagraph_mode=PIECEWISE because attention is compiled piecewise
Capturing CUDA graphs (mixed prefill-decode, PIECEWISE): 100%|██████████| 67/67 [00:01<00:00, 62.74it/s]
(EngineCore_DP0 pid=8323) INFO 11-18 07:05:43 [gpu_model_runner.py:3480] Graph capturing finished in 2 secs, took 0.19 GiB
(EngineCore_DP0 pid=8323) INFO 11-18 07:05:43 [core.py:210] init engine (profile, create kv cache, warmup model) took 6.93 seconds
INFO 11-18 07:05:44 [llm.py:306] Supported_tasks: ['generate']
Adding requests:   0%|          | 0/1 [00:00<?, ?it/s]
Processed prompts:   0%|          | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Adding requests:   0%|          | 0/50 [00:00<?, ?it/s]
Processed prompts:   0%|          | 0/50 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
```

```
[rank0]:[W1118 07:05:56.501768524 ProcessGroupNCCL.cpp:1538] Warning: WARNING: destroy_process_group() was not called before program exit, which can leak resources. For more info, please see https://pytorch.org/docs/stable/distributed.html#shutdown (function operator())
Running request count experiments...
INFO 11-18 07:05:57 [utils.py:233] non-default args: {'max_num_batched_tokens': 2048, 'disable_log_stats': True, 'model': 'facebook/opt-125m'}
INFO 11-18 07:05:57 [model.py:547] Resolved architecture: OPTForCausalLM
INFO 11-18 07:05:57 [model.py:1510] Using max model len 2048
INFO 11-18 07:05:57 [scheduler.py:205] Chunked prefill is enabled with max_num_batched_tokens=2048.
2025-11-18 07:06:03.249311: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register cuFFT factory: Attempting to register factory for plugin cuFFT when one has already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
E0000 00:00:1763449563.272089    8412 cuda_dnn.cc:8310] Unable to register cuDNN factory: Attempting to register factory for plugin cuDNN when one has already been registered
E0000 00:00:1763449563.279675    8412 cuda_blas.cc:1418] Unable to register cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has already been registered
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
INFO 11-18 07:06:09 [__init__.py:216] Automatically detected platform cuda.
(EngineCore_DP0 pid=8412) INFO 11-18 07:06:10 [core.py:644] Waiting for init message from front-end.
(EngineCore_DP0 pid=8412) INFO 11-18 07:06:10 [core.py:77] Initializing a V1 LLM engine (v0.11.0) with config: model='facebook/opt-125m', speculative_config=None, tokenizer='facebook/opt-125m', skip_tokenizer_init=False, tokenizer_mode=auto, revision=None, tokenizer_revision=None, trust_remote_code=False, dtype=torch.float16, max_seq_len=2048, download_dir=None, load_format=auto, tensor_parallel_size=1, pipeline_parallel_size=1, data_parallel_size=1, disable_custom_all_reduce=False, quantization=None, enforce_eager=False, kv_cache_dtype=auto, device_config=cuda, structured_outputs_config=StructuredOutputsConfig(backend='auto', disable_fallback=False, disable_any_whitespace=False, disable_additional_properties=False, reasoning_parser=''), observability_config=ObservabilityConfig(show_hidden_metrics_for_version=None, otlp_traces_endpoint=None, collect_detailed_traces=None), seed=0, served_model_name=facebook/opt-125m, enable_prefix_caching=True, chunked_prefill_enabled=True, pooler_config=None, compilation_config={"level":3,"debug_dump_path":"","cache_dir":"","backend":"","custom_ops":[],"splitting_ops":["vllm.unified_attention","vllm.unified_attention_with_output","vllm.mamba_mixer2","vllm.mamba_mixer","vllm.short_conv","vllm.linear_attention","vllm.plamo2_mamba_mixer","vllm.gdn_attention","vllm.sparse_attn_indexer"],"use_inductor":true,"compile_sizes":[],"inductor_compile_config":{"enable_auto_functionalized_v2":false},"inductor_passes":{},"cudagraph_mode":[2,1],"use_cudagraph":true,"cudagraph_num_of_warmups":1,"cudagraph_capture_sizes":[512,504,496,488,480,472,464,456,448,440,432,424,416,408,400,392,384,376,368,360,352,344,336,328,320,312,304,296,288,280,272,264,256,248,240,232,224,216,208,200,192,184,176,168,160,152,144,136,128,120,112,104,96,88,80,72,64,56,48,40,32,24,16,8,4,2,1],"cudagraph_copy_inputs":false,"full_cuda_graph":false,"use_inductor_graph_partition":false,"pass_config":{},"max_capture_size":512,"local_cache_dir":null}
(EngineCore_DP0 pid=8412) ERROR 11-18 07:06:12 [fa_utils.py:57] Cannot use FA version 2 is not supported due to FA2 is only supported on devices with compute capability >= 8
[W1118 07:06:22.051050995 socket.cpp:200] [c10d] The hostname of the client socket cannot be retrieved. err=-3
[W1118 07:06:32.061707194 socket.cpp:200] [c10d] The hostname of the client socket cannot be retrieved. err=-3
```

```
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
(EngineCore_DP0 pid=8412) INFO 11-18 07:06:32 [parallel_state.py:1208] rank 0 in world size 1 is assigned as DP rank 0, PP rank 0, TP rank 0, EP rank 0
(EngineCore_DP0 pid=8412) WARNING 11-18 07:06:33 [topk_topp_sampler.py:66] FlashInfer is not available. Falling back to the PyTorch-native implementation of top-p & top-k sampling. For the best performance, please install FlashInfer.
(EngineCore_DP0 pid=8412) INFO 11-18 07:06:33 [gpu_model_runner.py:2602] Starting to load model facebook/opt-125m...
(EngineCore_DP0 pid=8412) INFO 11-18 07:06:33 [gpu_model_runner.py:2634] Loading model from scratch...
(EngineCore_DP0 pid=8412) INFO 11-18 07:06:33 [cuda.py:372] Using FlexAttention backend on V1 engine.
(EngineCore_DP0 pid=8412) INFO 11-18 07:06:33 [weight_utils.py:392] Using model weights format ['*.safetensors', '*.bin', '*.pt']
Loading pt checkpoint shards:   0% Completed | 0/1 [00:00<?, ?it/s]
Loading pt checkpoint shards: 100% Completed | 1/1 [00:00<00:00,  3.63it/s]
Loading pt checkpoint shards: 100% Completed | 1/1 [00:00<00:00,  3.63it/s]
(EngineCore_DP0 pid=8412)
(EngineCore_DP0 pid=8412) INFO 11-18 07:06:34 [default_loader.py:267] Loading weights took 0.28 seconds
(EngineCore_DP0 pid=8412) INFO 11-18 07:06:35 [gpu_model_runner.py:2653] Model loading took 0.2393 GiB and 0.758835 seconds
(EngineCore_DP0 pid=8412) INFO 11-18 07:06:37 [backends.py:548] Using cache directory: /root/.cache/vllm/torch_compile_cache/932ffef25e/rank_0_0/backbone for vLLM's torch.compile
(EngineCore_DP0 pid=8412) INFO 11-18 07:06:37 [backends.py:559] Dynamo bytecode transform time: 2.43 s
(EngineCore_DP0 pid=8412) INFO 11-18 07:06:38 [backends.py:164] Directly load the compiled graph(s) for dynamic shape from the cache, took 0.287 s
(EngineCore_DP0 pid=8412) INFO 11-18 07:06:38 [monitor.py:34] torch.compile takes 2.43 s in total
(EngineCore_DP0 pid=8412) INFO 11-18 07:06:39 [gpu_worker.py:298] Available KV cache memory: 12.53 GiB
(EngineCore_DP0 pid=8412) INFO 11-18 07:06:40 [kv_cache_utils.py:1087] GPU KV cache size: 364,864 tokens
(EngineCore_DP0 pid=8412) INFO 11-18 07:06:40 [kv_cache_utils.py:1091] Maximum concurrency for 2,048 tokens per request: 178.16x
(EngineCore_DP0 pid=8412) WARNING 11-18 07:06:40 [gpu_model_runner.py:3663] CUDAGraphMode.FULL_AND_PIECEWISE is not supported with FlexAttentionMetadataBuilder backend (support: AttentionCGSupport.NEVER); setting cudagraph_mode=PIECEWISE because attention is compiled piecewise
Capturing CUDA graphs (mixed prefill-decode, PIECEWISE): 100%|██████████| 67/67 [00:01<00:00, 62.56it/s]
(EngineCore_DP0 pid=8412) INFO 11-18 07:06:41 [gpu_model_runner.py:3480] Graph capturing finished in 2 secs, took 0.19 GiB
(EngineCore_DP0 pid=8412) INFO 11-18 07:06:41 [core.py:210] init engine (profile, create kv cache, warmup model) took 6.94 seconds
INFO 11-18 07:06:42 [llm.py:306] Supported_tasks: ['generate']
Adding requests:   0%|          | 0/1 [00:00<?, ?it/s]
Processed prompts:   0%|          | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Adding requests:   0%|          | 0/10 [00:00<?, ?it/s]
Processed prompts:   0%|          | 0/10 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
```

```
[rank0]:[W1118 07:06:50.178353701 ProcessGroupNCCL.cpp:1538] Warning: WARNING: destroy_process_group() was not called before program exit, which can leak resources. For more info, please see https://pytorch.org/docs/stable/distributed.html#shutdown (function operator())
INFO 11-18 07:06:52 [utils.py:233] non-default args: {'max_num_batched_tokens': 2048, 'disable_log_stats': True, 'model': 'facebook/opt-125m'}
INFO 11-18 07:06:52 [model.py:547] Resolved architecture: OPTForCausalLM
INFO 11-18 07:06:52 [model.py:1510] Using max model len 2048
INFO 11-18 07:06:52 [scheduler.py:205] Chunked prefill is enabled with max_num_batched_tokens=2048.
2025-11-18 07:06:58.161247: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register cuFFT factory: Attempting to register factory for plugin cuFFT when one has already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
E0000 00:00:1763449618.184488    8501 cuda_dnn.cc:8310] Unable to register cuDNN factory: Attempting to register factory for plugin cuDNN when one has already been registered
E0000 00:00:1763449618.192179    8501 cuda_blas.cc:1418] Unable to register cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has already been registered
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
INFO 11-18 07:07:04 [__init__.py:216] Automatically detected platform cuda.
(EngineCore_DP0 pid=8501) INFO 11-18 07:07:05 [core.py:644] Waiting for init message from front-end.
(EngineCore_DP0 pid=8501) INFO 11-18 07:07:05 [core.py:77] Initializing a V1 LLM engine (v0.11.0) with config: model='facebook/opt-125m', speculative_config=None, tokenizer='facebook/opt-125m', skip_tokenizer_init=False, tokenizer_mode=auto, revision=None, tokenizer_revision=None, trust_remote_code=False, dtype=torch.float16, max_seq_len=2048, download_dir=None, load_format=auto, tensor_parallel_size=1, pipeline_parallel_size=1, data_parallel_size=1, disable_custom_all_reduce=False, quantization=None, enforce_eager=False, kv_cache_dtype=auto, device_config=cuda, structured_outputs_config=StructuredOutputsConfig(backend='auto', disable_fallback=False, disable_any_whitespace=False, disable_additional_properties=False, reasoning_parser=''), observability_config=ObservabilityConfig(show_hidden_metrics_for_version=None, otlp_traces_endpoint=None, collect_detailed_traces=None), seed=0, served_model_name=facebook/opt-125m, enable_prefix_caching=True, chunked_prefill_enabled=True, pooler_config=None, compilation_config={"level":3,"debug_dump_path":"","cache_dir":"","backend":"","custom_ops":[],"splitting_ops":["vllm.unified_attention","vllm.unified_attention_with_output","vllm.mamba_mixer2","vllm.mamba_mixer","vllm.short_conv","vllm.linear_attention","vllm.plamo2_mamba_mixer","vllm.gdn_attention","vllm.sparse_attn_indexer"],"use_inductor":true,"compile_sizes":[],"inductor_compile_config":{"enable_auto_functionalized_v2":false},"inductor_passes":{},"cudagraph_mode":[2,1],"use_cudagraph":true,"cudagraph_num_of_warmups":1,"cudagraph_capture_sizes":[512,504,496,488,480,472,464,456,448,440,432,424,416,408,400,392,384,376,368,360,352,344,336,328,320,312,304,296,288,280,272,264,256,248,240,232,224,216,208,200,192,184,176,168,160,152,144,136,128,120,112,104,96,88,80,72,64,56,48,40,32,24,16,8,4,2,1],"cudagraph_copy_inputs":false,"full_cuda_graph":false,"use_inductor_graph_partition":false,"pass_config":{},"max_capture_size":512,"local_cache_dir":null}
(EngineCore_DP0 pid=8501) ERROR 11-18 07:07:07 [fa_utils.py:57] Cannot use FA version 2 is not supported due to FA2 is only supported on devices with compute capability >= 8
[W1118 07:07:17.065260943 socket.cpp:200] [c10d] The hostname of the client socket cannot be retrieved. err=-3
[W1118 07:07:27.076145303 socket.cpp:200] [c10d] The hostname of the client socket cannot be retrieved. err=-3
```

```
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
(EngineCore_DP0 pid=8501) INFO 11-18 07:07:27 [parallel_state.py:1208] rank 0 in world size 1 is assigned as DP rank 0, PP rank 0, TP rank 0, EP rank 0
(EngineCore_DP0 pid=8501) WARNING 11-18 07:07:28 [topk_topp_sampler.py:66] FlashInfer is not available. Falling back to the PyTorch-native implementation of top-p & top-k sampling. For the best performance, please install FlashInfer.
(EngineCore_DP0 pid=8501) INFO 11-18 07:07:28 [gpu_model_runner.py:2602] Starting to load model facebook/opt-125m...
(EngineCore_DP0 pid=8501) INFO 11-18 07:07:28 [gpu_model_runner.py:2634] Loading model from scratch...
(EngineCore_DP0 pid=8501) INFO 11-18 07:07:28 [cuda.py:372] Using FlexAttention backend on V1 engine.
(EngineCore_DP0 pid=8501) INFO 11-18 07:07:28 [weight_utils.py:392] Using model weights format ['*.safetensors', '*.bin', '*.pt']
Loading pt checkpoint shards:   0% Completed | 0/1 [00:00<?, ?it/s]
Loading pt checkpoint shards: 100% Completed | 1/1 [00:00<00:00,  3.61it/s]
Loading pt checkpoint shards: 100% Completed | 1/1 [00:00<00:00,  3.61it/s]
(EngineCore_DP0 pid=8501)
(EngineCore_DP0 pid=8501) INFO 11-18 07:07:29 [default_loader.py:267] Loading weights took 0.28 seconds
(EngineCore_DP0 pid=8501) INFO 11-18 07:07:30 [gpu_model_runner.py:2653] Model loading took 0.2393 GiB and 0.880920 seconds
(EngineCore_DP0 pid=8501) INFO 11-18 07:07:33 [backends.py:548] Using cache directory: /root/.cache/vllm/torch_compile_cache/932ffef25e/rank_0_0/backbone for vLLM's torch.compile
(EngineCore_DP0 pid=8501) INFO 11-18 07:07:33 [backends.py:559] Dynamo bytecode transform time: 2.45 s
(EngineCore_DP0 pid=8501) INFO 11-18 07:07:33 [backends.py:164] Directly load the compiled graph(s) for dynamic shape from the cache, took 0.299 s
(EngineCore_DP0 pid=8501) INFO 11-18 07:07:33 [monitor.py:34] torch.compile takes 2.45 s in total
(EngineCore_DP0 pid=8501) INFO 11-18 07:07:34 [gpu_worker.py:298] Available KV cache memory: 12.53 GiB
(EngineCore_DP0 pid=8501) INFO 11-18 07:07:35 [kv_cache_utils.py:1087] GPU KV cache size: 364,864 tokens
(EngineCore_DP0 pid=8501) INFO 11-18 07:07:35 [kv_cache_utils.py:1091] Maximum concurrency for 2,048 tokens per request: 178.16x
(EngineCore_DP0 pid=8501) WARNING 11-18 07:07:35 [gpu_model_runner.py:3663] CUDAGraphMode.FULL_AND_PIECEWISE is not supported with FlexAttentionMetadataBuilder backend (support: AttentionCGSupport.NEVER); setting cudagraph_mode=PIECEWISE because attention is compiled piecewise
Capturing CUDA graphs (mixed prefill-decode, PIECEWISE): 100%|██████████| 67/67 [00:01<00:00, 62.88it/s]
(EngineCore_DP0 pid=8501) INFO 11-18 07:07:37 [gpu_model_runner.py:3480] Graph capturing finished in 2 secs, took 0.19 GiB
(EngineCore_DP0 pid=8501) INFO 11-18 07:07:37 [core.py:210] init engine (profile, create kv cache, warmup model) took 6.95 seconds
INFO 11-18 07:07:38 [llm.py:306] Supported_tasks: ['generate']
Adding requests:   0%|          | 0/1 [00:00<?, ?it/s]
Processed prompts:   0%|          | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Adding requests:   0%|          | 0/50 [00:00<?, ?it/s]
Processed prompts:   0%|          | 0/50 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
```

```
[rank0]:[W1118 07:07:49.082965363 ProcessGroupNCCL.cpp:1538] Warning: WARNING: destroy_process_group() was not called before program exit, which can leak resources. For more info, please see https://pytorch.org/docs/stable/distributed.html#shutdown (function operator())
INFO 11-18 07:07:51 [utils.py:233] non-default args: {'max_num_batched_tokens': 2048, 'disable_log_stats': True, 'model': 'facebook/opt-125m'}
INFO 11-18 07:07:52 [model.py:547] Resolved architecture: OPTForCausalLM
INFO 11-18 07:07:52 [model.py:1510] Using max model len 2048
INFO 11-18 07:07:52 [scheduler.py:205] Chunked prefill is enabled with max_num_batched_tokens=2048.
2025-11-18 07:07:57.640620: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register cuFFT factory: Attempting to register factory for plugin cuFFT when one has already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
E0000 00:00:1763449677.664006    8590 cuda_dnn.cc:8310] Unable to register cuDNN factory: Attempting to register factory for plugin cuDNN when one has already been registered
E0000 00:00:1763449677.671609    8590 cuda_blas.cc:1418] Unable to register cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has already been registered
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
INFO 11-18 07:08:03 [__init__.py:216] Automatically detected platform cuda.
(EngineCore_DP0 pid=8590) INFO 11-18 07:08:05 [core.py:644] Waiting for init message from front-end.
(EngineCore_DP0 pid=8590) INFO 11-18 07:08:05 [core.py:77] Initializing a V1 LLM engine (v0.11.0) with config: model='facebook/opt-125m', speculative_config=None, tokenizer='facebook/opt-125m', skip_tokenizer_init=False, tokenizer_mode=auto, revision=None, tokenizer_revision=None, trust_remote_code=False, dtype=torch.float16, max_seq_len=2048, download_dir=None, load_format=auto, tensor_parallel_size=1, pipeline_parallel_size=1, data_parallel_size=1, disable_custom_all_reduce=False, quantization=None, enforce_eager=False, kv_cache_dtype=auto, device_config=cuda, structured_outputs_config=StructuredOutputsConfig(backend='auto', disable_fallback=False, disable_any_whitespace=False, disable_additional_properties=False, reasoning_parser=''), observability_config=ObservabilityConfig(show_hidden_metrics_for_version=None, otlp_traces_endpoint=None, collect_detailed_traces=None), seed=0, served_model_name=facebook/opt-125m, enable_prefix_caching=True, chunked_prefill_enabled=True, pooler_config=None, compilation_config={"level":3,"debug_dump_path":"","cache_dir":"","backend":"","custom_ops":[],"splitting_ops":["vllm.unified_attention","vllm.unified_attention_with_output","vllm.mamba_mixer2","vllm.mamba_mixer","vllm.short_conv","vllm.linear_attention","vllm.plamo2_mamba_mixer","vllm.gdn_attention","vllm.sparse_attn_indexer"],"use_inductor":true,"compile_sizes":[],"inductor_compile_config":{"enable_auto_functionalized_v2":false},"inductor_passes":{},"cudagraph_mode":[2,1],"use_cudagraph":true,"cudagraph_num_of_warmups":1,"cudagraph_capture_sizes":[512,504,496,488,480,472,464,456,448,440,432,424,416,408,400,392,384,376,368,360,352,344,336,328,320,312,304,296,288,280,272,264,256,248,240,232,224,216,208,200,192,184,176,168,160,152,144,136,128,120,112,104,96,88,80,72,64,56,48,40,32,24,16,8,4,2,1],"cudagraph_copy_inputs":false,"full_cuda_graph":false,"use_inductor_graph_partition":false,"pass_config":{},"max_capture_size":512,"local_cache_dir":null}
(EngineCore_DP0 pid=8590) ERROR 11-18 07:08:06 [fa_utils.py:57] Cannot use FA version 2 is not supported due to FA2 is only supported on devices with compute capability >= 8
[W1118 07:08:17.470442149 socket.cpp:200] [c10d] The hostname of the client socket cannot be retrieved. err=-3
[W1118 07:08:27.481195701 socket.cpp:200] [c10d] The hostname of the client socket cannot be retrieved. err=-3
```

```
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
(EngineCore_DP0 pid=8590) INFO 11-18 07:08:27 [parallel_state.py:1208] rank 0 in world size 1 is assigned as DP rank 0, PP rank 0, TP rank 0, EP rank 0
(EngineCore_DP0 pid=8590) WARNING 11-18 07:08:27 [topk_topp_sampler.py:66] FlashInfer is not available. Falling back to the PyTorch-native implementation of top-p & top-k sampling. For the best performance, please install FlashInfer.
(EngineCore_DP0 pid=8590) INFO 11-18 07:08:27 [gpu_model_runner.py:2602] Starting to load model facebook/opt-125m...
(EngineCore_DP0 pid=8590) INFO 11-18 07:08:27 [gpu_model_runner.py:2634] Loading model from scratch...
(EngineCore_DP0 pid=8590) INFO 11-18 07:08:27 [cuda.py:372] Using FlexAttention backend on V1 engine.
(EngineCore_DP0 pid=8590) INFO 11-18 07:08:28 [weight_utils.py:392] Using model weights format ['*.safetensors', '*.bin', '*.pt']
Loading pt checkpoint shards:   0% Completed | 0/1 [00:00<?, ?it/s]
Loading pt checkpoint shards: 100% Completed | 1/1 [00:00<00:00,  3.64it/s]
Loading pt checkpoint shards: 100% Completed | 1/1 [00:00<00:00,  3.63it/s]
(EngineCore_DP0 pid=8590)
(EngineCore_DP0 pid=8590) INFO 11-18 07:08:28 [default_loader.py:267] Loading weights took 0.28 seconds
(EngineCore_DP0 pid=8590) INFO 11-18 07:08:29 [gpu_model_runner.py:2653] Model loading took 0.2393 GiB and 0.774182 seconds
(EngineCore_DP0 pid=8590) INFO 11-18 07:08:32 [backends.py:548] Using cache directory: /root/.cache/vllm/torch_compile_cache/932ffef25e/rank_0_0/backbone for vLLM's torch.compile
(EngineCore_DP0 pid=8590) INFO 11-18 07:08:32 [backends.py:559] Dynamo bytecode transform time: 2.41 s
(EngineCore_DP0 pid=8590) INFO 11-18 07:08:32 [backends.py:164] Directly load the compiled graph(s) for dynamic shape from the cache, took 0.290 s
(EngineCore_DP0 pid=8590) INFO 11-18 07:08:33 [monitor.py:34] torch.compile takes 2.41 s in total
(EngineCore_DP0 pid=8590) INFO 11-18 07:08:34 [gpu_worker.py:298] Available KV cache memory: 12.53 GiB
(EngineCore_DP0 pid=8590) INFO 11-18 07:08:34 [kv_cache_utils.py:1087] GPU KV cache size: 364,864 tokens
(EngineCore_DP0 pid=8590) INFO 11-18 07:08:34 [kv_cache_utils.py:1091] Maximum concurrency for 2,048 tokens per request: 178.16x
(EngineCore_DP0 pid=8590) WARNING 11-18 07:08:34 [gpu_model_runner.py:3663] CUDAGraphMode.FULL_AND_PIECEWISE is not supported with FlexAttentionMetadataBuilder backend (support: AttentionCGSupport.NEVER); setting cudagraph_mode=PIECEWISE because attention is compiled piecewise
Capturing CUDA graphs (mixed prefill-decode, PIECEWISE): 100%|██████████| 67/67 [00:01<00:00, 63.14it/s]
(EngineCore_DP0 pid=8590) INFO 11-18 07:08:36 [gpu_model_runner.py:3480] Graph capturing finished in 2 secs, took 0.19 GiB
(EngineCore_DP0 pid=8590) INFO 11-18 07:08:36 [core.py:210] init engine (profile, create kv cache, warmup model) took 6.85 seconds
INFO 11-18 07:08:37 [llm.py:306] Supported_tasks: ['generate']
Adding requests:   0%|          | 0/1 [00:00<?, ?it/s]
Processed prompts:   0%|          | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Adding requests:   0%|          | 0/100 [00:00<?, ?it/s]
Processed prompts:   0%|          | 0/100 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
```

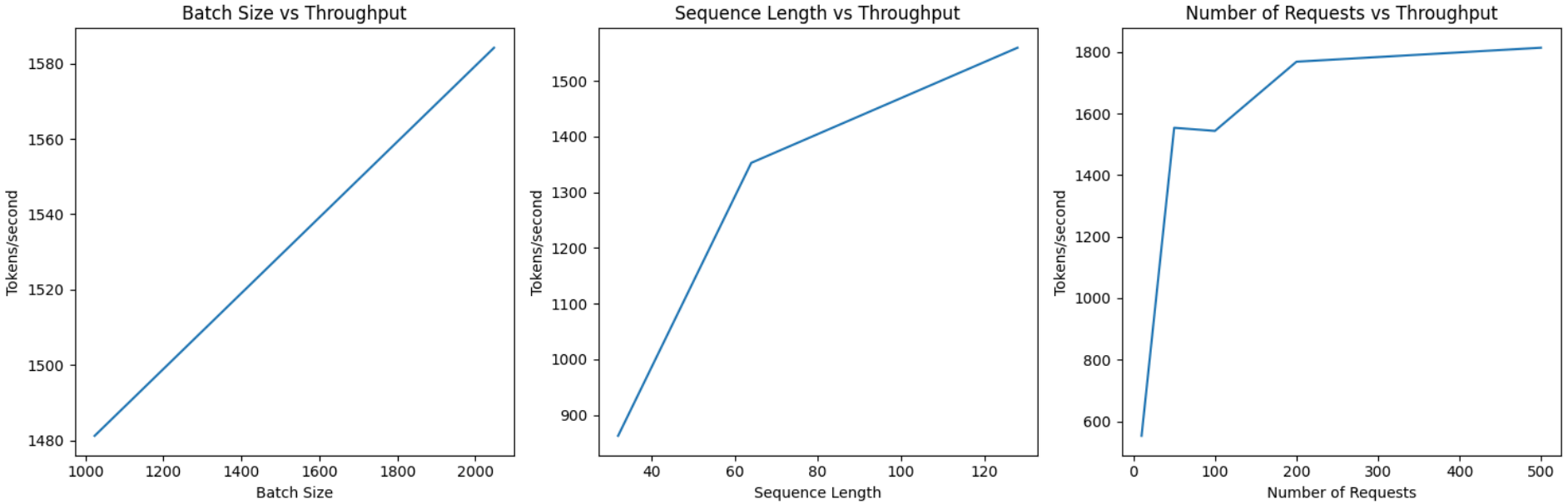
```
[rank0]:[W1118 07:08:57.552513379 ProcessGroupNCCL.cpp:1538] Warning: WARNING: destroy_process_group() was not called before program exit, which can leak resources. For more info, please see https://pytorch.org/docs/stable/distributed.html#shutdown (function operator())
INFO 11-18 07:09:00 [utils.py:233] non-default args: {'max_num_batched_tokens': 2048, 'disable_log_stats': True, 'model': 'facebook/opt-125m'}
INFO 11-18 07:09:00 [model.py:547] Resolved architecture: OPTForCausalLM
INFO 11-18 07:09:00 [model.py:1510] Using max model len 2048
INFO 11-18 07:09:00 [scheduler.py:205] Chunked prefill is enabled with max_num_batched_tokens=2048.
2025-11-18 07:09:06.359321: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register cuFFT factory: Attempting to register factory for plugin cuFFT when one has already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
E0000 00:00:1763449746.383026    8679 cuda_dnn.cc:8310] Unable to register cuDNN factory: Attempting to register factory for plugin cuDNN when one has already been registered
E0000 00:00:1763449746.390753    8679 cuda_blas.cc:1418] Unable to register cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has already been registered
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
INFO 11-18 07:09:12 [__init__.py:216] Automatically detected platform cuda.
(EngineCore_DP0 pid=8679) INFO 11-18 07:09:14 [core.py:644] Waiting for init message from front-end.
(EngineCore_DP0 pid=8679) INFO 11-18 07:09:14 [core.py:77] Initializing a V1 LLM engine (v0.11.0) with config: model='facebook/opt-125m', speculative_config=None, tokenizer='facebook/opt-125m', skip_tokenizer_init=False, tokenizer_mode=auto, revision=None, tokenizer_revision=None, trust_remote_code=False, dtype=torch.float16, max_seq_len=2048, download_dir=None, load_format=auto, tensor_parallel_size=1, pipeline_parallel_size=1, data_parallel_size=1, disable_custom_all_reduce=False, quantization=None, enforce_eager=False, kv_cache_dtype=auto, device_config=cuda, structured_outputs_config=StructuredOutputsConfig(backend='auto', disable_fallback=False, disable_any_whitespace=False, disable_additional_properties=False, reasoning_parser=''), observability_config=ObservabilityConfig(show_hidden_metrics_for_version=None, otlp_traces_endpoint=None, collect_detailed_traces=None), seed=0, served_model_name=facebook/opt-125m, enable_prefix_caching=True, chunked_prefill_enabled=True, pooler_config=None, compilation_config={"level":3,"debug_dump_path":"","cache_dir":"","backend":"","custom_ops":[],"splitting_ops":["vllm.unified_attention","vllm.unified_attention_with_output","vllm.mamba_mixer2","vllm.mamba_mixer","vllm.short_conv","vllm.linear_attention","vllm.plamo2_mamba_mixer","vllm.gdn_attention","vllm.sparse_attn_indexer"],"use_inductor":true,"compile_sizes":[],"inductor_compile_config":{"enable_auto_functionalized_v2":false},"inductor_passes":{},"cudagraph_mode":[2,1],"use_cudagraph":true,"cudagraph_num_of_warmups":1,"cudagraph_capture_sizes":[512,504,496,488,480,472,464,456,448,440,432,424,416,408,400,392,384,376,368,360,352,344,336,328,320,312,304,296,288,280,272,264,256,248,240,232,224,216,208,200,192,184,176,168,160,152,144,136,128,120,112,104,96,88,80,72,64,56,48,40,32,24,16,8,4,2,1],"cudagraph_copy_inputs":false,"full_cuda_graph":false,"use_inductor_graph_partition":false,"pass_config":{},"max_capture_size":512,"local_cache_dir":null}
(EngineCore_DP0 pid=8679) ERROR 11-18 07:09:15 [fa_utils.py:57] Cannot use FA version 2 is not supported due to FA2 is only supported on devices with compute capability >= 8
[W1118 07:09:25.226413113 socket.cpp:200] [c10d] The hostname of the client socket cannot be retrieved. err=-3
[W1118 07:09:35.237144579 socket.cpp:200] [c10d] The hostname of the client socket cannot be retrieved. err=-3
```

```
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
(EngineCore_DP0 pid=8679) INFO 11-18 07:09:35 [parallel_state.py:1208] rank 0 in world size 1 is assigned as DP rank 0, PP rank 0, TP rank 0, EP rank 0
(EngineCore_DP0 pid=8679) WARNING 11-18 07:09:36 [topk_topp_sampler.py:66] FlashInfer is not available. Falling back to the PyTorch-native implementation of top-p & top-k sampling. For the best performance, please install FlashInfer.
(EngineCore_DP0 pid=8679) INFO 11-18 07:09:36 [gpu_model_runner.py:2602] Starting to load model facebook/opt-125m...
(EngineCore_DP0 pid=8679) INFO 11-18 07:09:36 [gpu_model_runner.py:2634] Loading model from scratch...
(EngineCore_DP0 pid=8679) INFO 11-18 07:09:36 [cuda.py:372] Using FlexAttention backend on V1 engine.
(EngineCore_DP0 pid=8679) INFO 11-18 07:09:37 [weight_utils.py:392] Using model weights format ['*.safetensors', '*.bin', '*.pt']
Loading pt checkpoint shards:   0% Completed | 0/1 [00:00<?, ?it/s]
Loading pt checkpoint shards: 100% Completed | 1/1 [00:00<00:00,  3.62it/s]
Loading pt checkpoint shards: 100% Completed | 1/1 [00:00<00:00,  3.61it/s]
(EngineCore_DP0 pid=8679)
(EngineCore_DP0 pid=8679) INFO 11-18 07:09:37 [default_loader.py:267] Loading weights took 0.29 seconds
(EngineCore_DP0 pid=8679) INFO 11-18 07:09:38 [gpu_model_runner.py:2653] Model loading took 0.2393 GiB and 0.771048 seconds
(EngineCore_DP0 pid=8679) INFO 11-18 07:09:41 [backends.py:548] Using cache directory: /root/.cache/vllm/torch_compile_cache/932ffef25e/rank_0_0/backbone for vLLM's torch.compile
(EngineCore_DP0 pid=8679) INFO 11-18 07:09:41 [backends.py:559] Dynamo bytecode transform time: 2.44 s
(EngineCore_DP0 pid=8679) INFO 11-18 07:09:41 [backends.py:164] Directly load the compiled graph(s) for dynamic shape from the cache, took 0.292 s
(EngineCore_DP0 pid=8679) INFO 11-18 07:09:41 [monitor.py:34] torch.compile takes 2.44 s in total
(EngineCore_DP0 pid=8679) INFO 11-18 07:09:42 [gpu_worker.py:298] Available KV cache memory: 12.53 GiB
(EngineCore_DP0 pid=8679) INFO 11-18 07:09:43 [kv_cache_utils.py:1087] GPU KV cache size: 364,864 tokens
(EngineCore_DP0 pid=8679) INFO 11-18 07:09:43 [kv_cache_utils.py:1091] Maximum concurrency for 2,048 tokens per request: 178.16x
(EngineCore_DP0 pid=8679) WARNING 11-18 07:09:43 [gpu_model_runner.py:3663] CUDAGraphMode.FULL_AND_PIECEWISE is not supported with FlexAttentionMetadataBuilder backend (support: AttentionCGSupport.NEVER); setting cudagraph_mode=PIECEWISE because attention is compiled piecewise
Capturing CUDA graphs (mixed prefill-decode, PIECEWISE): 100%|██████████| 67/67 [00:01<00:00, 62.42it/s]
(EngineCore_DP0 pid=8679) INFO 11-18 07:09:45 [gpu_model_runner.py:3480] Graph capturing finished in 2 secs, took 0.19 GiB
(EngineCore_DP0 pid=8679) INFO 11-18 07:09:45 [core.py:210] init engine (profile, create kv cache, warmup model) took 6.95 seconds
INFO 11-18 07:09:45 [llm.py:306] Supported_tasks: ['generate']
Adding requests:   0%|          | 0/1 [00:00<?, ?it/s]
Processed prompts:   0%|          | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Adding requests:   0%|          | 0/200 [00:00<?, ?it/s]
Processed prompts:   0%|          | 0/200 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
```

```
[rank0]:[W1118 07:10:18.783044149 ProcessGroupNCCL.cpp:1538] Warning: WARNING: destroy_process_group() was not called before program exit, which can leak resources. For more info, please see https://pytorch.org/docs/stable/distributed.html#shutdown (function operator())
INFO 11-18 07:10:25 [utils.py:233] non-default args: {'max_num_batched_tokens': 2048, 'disable_log_stats': True, 'model': 'facebook/opt-125m'}
INFO 11-18 07:10:25 [model.py:547] Resolved architecture: OPTForCausalLM
INFO 11-18 07:10:25 [model.py:1510] Using max model len 2048
INFO 11-18 07:10:25 [scheduler.py:205] Chunked prefill is enabled with max_num_batched_tokens=2048.
2025-11-18 07:10:31.214035: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register cuFFT factory: Attempting to register factory for plugin cuFFT when one has already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
E0000 00:00:1763449831.237091    8768 cuda_dnn.cc:8310] Unable to register cuDNN factory: Attempting to register factory for plugin cuDNN when one has already been registered
E0000 00:00:1763449831.244741    8768 cuda_blas.cc:1418] Unable to register cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has already been registered
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
INFO 11-18 07:10:37 [__init__.py:216] Automatically detected platform cuda.
(EngineCore_DP0 pid=8768) INFO 11-18 07:10:39 [core.py:644] Waiting for init message from front-end.
(EngineCore_DP0 pid=8768) INFO 11-18 07:10:39 [core.py:77] Initializing a V1 LLM engine (v0.11.0) with config: model='facebook/opt-125m', speculative_config=None, tokenizer='facebook/opt-125m', skip_tokenizer_init=False, tokenizer_mode=auto, revision=None, tokenizer_revision=None, trust_remote_code=False, dtype=torch.float16, max_seq_len=2048, download_dir=None, load_format=auto, tensor_parallel_size=1, pipeline_parallel_size=1, data_parallel_size=1, disable_custom_all_reduce=False, quantization=None, enforce_eager=False, kv_cache_dtype=auto, device_config=cuda, structured_outputs_config=StructuredOutputsConfig(backend='auto', disable_fallback=False, disable_any_whitespace=False, disable_additional_properties=False, reasoning_parser=''), observability_config=ObservabilityConfig(show_hidden_metrics_for_version=None, otlp_traces_endpoint=None, collect_detailed_traces=None), seed=0, served_model_name=facebook/opt-125m, enable_prefix_caching=True, chunked_prefill_enabled=True, pooler_config=None, compilation_config={"level":3,"debug_dump_path":"","cache_dir":"","backend":"","custom_ops":[],"splitting_ops":["vllm.unified_attention","vllm.unified_attention_with_output","vllm.mamba_mixer2","vllm.mamba_mixer","vllm.short_conv","vllm.linear_attention","vllm.plamo2_mamba_mixer","vllm.gdn_attention","vllm.sparse_attn_indexer"],"use_inductor":true,"compile_sizes":[],"inductor_compile_config":{"enable_auto_functionalized_v2":false},"inductor_passes":{},"cudagraph_mode":[2,1],"use_cudagraph":true,"cudagraph_num_of_warmups":1,"cudagraph_capture_sizes":[512,504,496,488,480,472,464,456,448,440,432,424,416,408,400,392,384,376,368,360,352,344,336,328,320,312,304,296,288,280,272,264,256,248,240,232,224,216,208,200,192,184,176,168,160,152,144,136,128,120,112,104,96,88,80,72,64,56,48,40,32,24,16,8,4,2,1],"cudagraph_copy_inputs":false,"full_cuda_graph":false,"use_inductor_graph_partition":false,"pass_config":{},"max_capture_size":512,"local_cache_dir":null}
(EngineCore_DP0 pid=8768) ERROR 11-18 07:10:40 [fa_utils.py:57] Cannot use FA version 2 is not supported due to FA2 is only supported on devices with compute capability >= 8
[W1118 07:10:50.120223089 socket.cpp:200] [c10d] The hostname of the client socket cannot be retrieved. err=-3
[W1118 07:11:00.130929409 socket.cpp:200] [c10d] The hostname of the client socket cannot be retrieved. err=-3
```

```
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
[Gloo] Rank 0 is connected to 0 peer ranks. Expected number of connected peer ranks is : 0
(EngineCore_DP0 pid=8768) INFO 11-18 07:11:00 [parallel_state.py:1208] rank 0 in world size 1 is assigned as DP rank 0, PP rank 0, TP rank 0, EP rank 0
(EngineCore_DP0 pid=8768) WARNING 11-18 07:11:01 [topk_topp_sampler.py:66] FlashInfer is not available. Falling back to the PyTorch-native implementation of top-p & top-k sampling. For the best performance, please install FlashInfer.
(EngineCore_DP0 pid=8768) INFO 11-18 07:11:01 [gpu_model_runner.py:2602] Starting to load model facebook/opt-125m...
(EngineCore_DP0 pid=8768) INFO 11-18 07:11:01 [gpu_model_runner.py:2634] Loading model from scratch...
(EngineCore_DP0 pid=8768) INFO 11-18 07:11:01 [cuda.py:372] Using FlexAttention backend on V1 engine.
(EngineCore_DP0 pid=8768) INFO 11-18 07:11:01 [weight_utils.py:392] Using model weights format ['*.safetensors', '*.bin', '*.pt']
Loading pt checkpoint shards:   0% Completed | 0/1 [00:00<?, ?it/s]
Loading pt checkpoint shards: 100% Completed | 1/1 [00:00<00:00,  3.58it/s]
Loading pt checkpoint shards: 100% Completed | 1/1 [00:00<00:00,  3.58it/s]
(EngineCore_DP0 pid=8768)
(EngineCore_DP0 pid=8768) INFO 11-18 07:11:02 [default_loader.py:267] Loading weights took 0.29 seconds
(EngineCore_DP0 pid=8768) INFO 11-18 07:11:03 [gpu_model_runner.py:2653] Model loading took 0.2393 GiB and 0.761851 seconds
(EngineCore_DP0 pid=8768) INFO 11-18 07:11:05 [backends.py:548] Using cache directory: /root/.cache/vllm/torch_compile_cache/932ffef25e/rank_0_0/backbone for vLLM's torch.compile
(EngineCore_DP0 pid=8768) INFO 11-18 07:11:05 [backends.py:559] Dynamo bytecode transform time: 2.42 s
(EngineCore_DP0 pid=8768) INFO 11-18 07:11:06 [backends.py:164] Directly load the compiled graph(s) for dynamic shape from the cache, took 0.288 s
(EngineCore_DP0 pid=8768) INFO 11-18 07:11:06 [monitor.py:34] torch.compile takes 2.42 s in total
(EngineCore_DP0 pid=8768) INFO 11-18 07:11:07 [gpu_worker.py:298] Available KV cache memory: 12.53 GiB
(EngineCore_DP0 pid=8768) INFO 11-18 07:11:08 [kv_cache_utils.py:1087] GPU KV cache size: 364,864 tokens
(EngineCore_DP0 pid=8768) INFO 11-18 07:11:08 [kv_cache_utils.py:1091] Maximum concurrency for 2,048 tokens per request: 178.16x
(EngineCore_DP0 pid=8768) WARNING 11-18 07:11:08 [gpu_model_runner.py:3663] CUDAGraphMode.FULL_AND_PIECEWISE is not supported with FlexAttentionMetadataBuilder backend (support: AttentionCGSupport.NEVER); setting cudagraph_mode=PIECEWISE because attention is compiled piecewise
Capturing CUDA graphs (mixed prefill-decode, PIECEWISE): 100%|██████████| 67/67 [00:01<00:00, 61.90it/s]
(EngineCore_DP0 pid=8768) INFO 11-18 07:11:10 [gpu_model_runner.py:3480] Graph capturing finished in 2 secs, took 0.19 GiB
(EngineCore_DP0 pid=8768) INFO 11-18 07:11:10 [core.py:210] init engine (profile, create kv cache, warmup model) took 6.93 seconds
INFO 11-18 07:11:10 [llm.py:306] Supported_tasks: ['generate']
Adding requests:   0%|          | 0/1 [00:00<?, ?it/s]
Processed prompts:   0%|          | 0/1 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
Adding requests:   0%|          | 0/500 [00:00<?, ?it/s]
Processed prompts:   0%|          | 0/500 [00:00<?, ?it/s, est. speed input: 0.00 toks/s, output: 0.00 toks/s]
```

```
[rank0]:[W1118 07:12:24.190660733 ProcessGroupNCCL.cpp:1538] Warning: WARNING: destroy_process_group() was not called b
efore program exit, which can leak resources. For more info, please see https://pytorch.org/docs/stable/distributed.htm
l#shutdown (function operator())
```



In [11]:
```
print(batch_results)
print(length_results)
print(request_results)
```

```
   batch_size   throughput  elapsed_time
0        1024  1481.249009      8.641356
1        2048  1584.172501      8.079928
   sequence_length   throughput  elapsed_time
0               32   862.707103      3.709254
1               64  1352.743424      4.731126
2              128  1559.473363      8.207899
   num_requests   throughput  elapsed_time
0            10   552.875013      4.630341
1            50  1553.687925      8.238463
2           100  1543.557289     16.585066
3           200  1768.359138     28.953395
4           500  1813.848725     70.568178
```

- Write a clear analysis of the results (2 points)

2.3.3 Experiments and Analysis

Parameter Studies

Batch size experiments: To vary the batch size, I just varied the max_num_batched_tokens argument passed when intializing the LLM. Since the sequence length was set to 128 for input and output in the batch size experiments, when max_num_batched_tokens = 1024 then the number of requests the LLM processes together is 1024 / 128 = 8 and when max_num_batched_tokens = 2048 then the number of requests the LLM processes together is 2048 / 128 = 16. From the above visualization of batch size vs. throughput and the above display of the dataframe for the batch size experiment results, we can see that for the two batch sizes tested(1024, 2048) the throughput increases slightly. Maybe the throughput will increase more rapidly for larger batch sizes but unfortunately whenever I tested larger batch sizes, I kept getting Out of Memory errors.

Input/output length experiments: From the above visualization of sequence length vs. throughput and the above display of the dataframe for the sequence length experiment results, we can clearly see that throughput increases with sequence length which means that the GPUs are getting utilized better as the sequence length increases as well. Although the elapsed time of each run increases as sequence length increases due to the LLM having to handle more tokens, the GPU is utilized more efficiently as seen in the increases in throughput.

Request count experiments: From the above visualization of number of requests vs. throughput and the above display of the dataframe for the request count experiment results, we can clearly see that throughput generally improves as the number of requests increases. However, it's also clearly evident that the throughput increase slows down after number of requests = 50. So the LLM definitely is able to scale efficiently with more requests up to a certain point after which the throughput gains kind of slow down since the GPUs are probably already being fully utilized.

Question 4: Implement ReAct Agent with Multiple Tools (25 points)

Implement a ReAct (Reasoning and Acting) agent as described by Yao et al. [1], incorporating three main tools: search, compare, and analyze. This agent should be able to handle complex queries by reasoning about which tool to use and when.

a) (4 points) Implement the search tool using the SerpAPI integration from previous questions. Ensure it can be easily used by the ReAct agent.

- Proper integration with SerpAPI
- Formatting the search results for use by the ReAct agent

b) (5 points) Create a custom comparison tool using LangChain's `Tool` class. The tool should accept multiple items and a category as input and return a comparison result.

- Implementing the comparison logic
- Creating an appropriate prompt template for the comparison
- Proper error handling for invalid inputs

c) (5 points) Implement an analysis tool that can summarize and extract key information from search results or comparisons. This tool should use the OpenAI model to generate insightful analyses.

- Implementing the analysis logic
- Creating an appropriate prompt template for the analysis
- Ensuring the analysis output is concise and relevant

d) (6 points) Integrate these tools with a ReAct agent using LangChain. Your implementation should:

- Use LangChain's `initialize_agent` function with the `AgentType.ZERO_SHOT_REACT_DESCRIPTION` agent type
- Include all three tools (search, compare, analyze) as available actions for the agent
- Implement proper error handling and fallback strategies
- Ensure smooth transitions between tools in the agent's reasoning process

e) (5 points) Implement a simple Streamlit user interface for your ReAct agent. Your implementation should include:

- A text input field for users to enter their queries
- A button to submit the query and trigger the ReAct agent
- A display area for showing the final results
- A section to display the step-by-step reasoning process of the ReAct agent

In [2]:
```
# Install required packages
!pip install -q "langchain>=0.2.10" "langchain-openai>=0.2.10" "langchain-community>=0.2.10" "google-search-results" "s
```

```
  Preparing metadata (setup.py) ... done
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 76.0/76.0 kB 3.0 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2.5/2.5 MB 38.0 MB/s eta 0:00:00a 0:00:01
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 10.2/10.2 MB 94.5 MB/s eta 0:00:00:00:010:01
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 449.8/449.8 kB 31.3 MB/s eta 0:00:00
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 6.9/6.9 MB 103.5 MB/s eta 0:00:0000:01
  Building wheel for google-search-results (setup.py) ... done
```

In [3]:
```
# Import necessary libraries
import os
from dotenv import load_dotenv
from langchain_openai import OpenAI
from langchain.agents import load_tools, initialize_agent, AgentType
from langchain.tools import Tool
from langchain.chains import LLMChain
from langchain.prompts import PromptTemplate

# Set API keys
os.environ["OPENAI_API_KEY"] = 'sk-proj-VnNa9DbCMbMUtNqVmu610nzfYZ5LPcgBH-kHKBgefL5FuFHGcfFUQALsd5B69CeLm-N76GlYNQT3Blb
os.environ["SERPAPI_API_KEY"] = 'bdd56e9df6bb98875430f428df033decd2286bed67432865a1a1213314345f80'
load_dotenv()

# Initialize the OpenAI language model
llm = OpenAI(temperature=0)
```

a) Implement the search tool

In [4]:
```
# Load the search tool using SerpAPI
search_tool = load_tools(["serpapi"], llm=llm)
```

b) Create a custom comparison tool

In [5]:
```
import json
def compare_items(query: str) -> str:
    # Extract items and category from the query
    extraction_prompt = PromptTemplate(
        input_variables=["query"],
        template="""Reason about the query and come up with at least two items to compare and the comparison category f
        Query: {query}
        CRITICAL: You must respond with ONLY valid JSON. Do not include any explanatory text, markdown formatting, or c
```

```
        Your response must be exactly in this format: {{"items": ["item1", "item2"], "category": "category description"}}
        Rules:
        - Return ONLY the JSON object, nothing else
        - Do not use markdown code blocks (no ```)
        - Do not add any text before or after the JSON
        - Items should be in a list
        - Category should be a short descriptive string

        JSON:"""
    )

    extraction_result = llm(extraction_prompt.format(query=query))
    # Parse the result
    try:
        input_data = json.loads(extraction_result)
        items = input_data.get("items", [])
        category = input_data.get("category", "")
    except json.JSONDecodeError:
        return f"Error: Could not parse query properly. Please specify items and category in the query clearly."

    if not items or len(items) < 2:
        return "Error: Insufficient number of items found to compare"
    if not category:
        return "Error: Category not specified"

    # Run the comparison
    comparison_prompt = PromptTemplate(
        input_variables=["items", "category", "query"],
        template="""You are an expert analyst. Compare the following items in the context of {category}.
        Items to compare: {items}
        Keep in mind that the ultimate goal is to generate information to answer the query: {query}
        Provide a comprehensive comparison covering all the below listed points in detail:
        1. Similarities between the items
        2. Differences between the items
        3. Strengths and weaknesses of each item
        4. Summary recommendation or conclusion

        Be objective, detailed, and well-structured in your analysis."""
    )
    items_str = ", ".join(items)
    prompt = comparison_prompt.format(items=items_str, category=category, query=query)
    result = llm(prompt)
    return result

compare_tool = Tool(
    name="Compare",
    func=compare_items,
```

```
    description='Reason about the query, come up with multiple items and a specific category to describe the query, com
)
```

c) Implement an analysis tool

In [6]:
```
# New function to analyze search results and perform comparisons
def analyze_results(results: str) -> str:
    analysis_prompt = PromptTemplate(
        input_variables=["results"],
        template="""You are an expert data analyst. Analyze the following results and provide key insights to answer th
        Results to analyze:{results}
        Keep in mind that the ultimate goal is to generate information to answer the user query
        Provide a concise analysis including:
        1. Main findings and key points
        2. Important patterns or trends identified
        3. Critical takeaways
        4. Actionable insights or recommendations
        Keep your analysis focused, relevant, and well-organized. Be concise but comprehensive."""
    )

    prompt = analysis_prompt.format(results=results)
    result = llm(prompt)
    return result

analyze_tool = Tool(
    name="Analyze",
    func=analyze_results,
    description="Analyze information to extract insights, patterns, and conclusions. Input: results to analyze. Summari
)
```

d) Integrate tools with a ReAct agent

In [7]:
```
# Integrate tools with ReAct agent
all_tools = search_tool + [compare_tool, analyze_tool]
agent = initialize_agent(
    tools=all_tools,
    llm=llm,
    agent=AgentType.ZERO_SHOT_REACT_DESCRIPTION,
    verbose=True,
    early_stopping_method="generate",
    handle_parsing_errors=True,
    return_intermediate_steps=True
)
```

```
/tmp/ipykernel_48/3543538376.py:3: LangChainDeprecationWarning: LangChain agents will continue to be supported, but it
is recommended for new use cases to be built with LangGraph. LangGraph offers a more flexible and full-featured framewo
rk for building agents, including support for tool-calling, persistence of state, and human-in-the-loop workflows. For
details, refer to the `LangGraph documentation <https://langchain-ai.github.io/langgraph/>`_ as well as guides for `Mig
rating from AgentExecutor <https://python.langchain.com/docs/how_to/migrate_agent/>`_ and LangGraph's `Pre-built ReAct
agent <https://langchain-ai.github.io/langgraph/how-tos/create-react-agent/>`_.
  agent = initialize_agent(
```

In [8]:
```python
def process_query(query: str, max_steps: int = 100) -> str:
    try:
        return agent({"input": query, "max_iterations": max_steps})["output"]
    except RecursionError:
        return "The query was too complex and exceeded the maximum number of steps. Please try a simpler query."
    except Exception as e:
        return f"An error occurred: {str(e)}"
```

Test Your Implementation

Use the cell below to test your implementation with a sample query.

In [9]:
```python
# Test your implementation
sample_query = "What are the top 3 smartphones in 2023, and how do they compare in terms of camera quality and battery

result = process_query(sample_query)
print(result)
```

```
/tmp/ipykernel_48/629042221.py:3: LangChainDeprecationWarning: The method `Chain.__call__` was deprecated in langchain
0.1.0 and will be removed in 1.0. Use :meth:`~invoke` instead.
  return agent({"input": query, "max_iterations": max_steps})["output"]
```

```
> Entering new AgentExecutor chain...
 I should use the Search tool to find information about the top smartphones in 2023.
Action: Search
Action Input: "Top smartphones 2023"
Observation: {'type': 'organic_result', 'title': 'Smartphones 2023'}
Thought: I should use the Compare tool to compare the top smartphones in terms of camera quality and battery life.
Action: Compare
Action Input: "Smartphones 2023"
```

```
/tmp/ipykernel_48/4215994792.py:20: LangChainDeprecationWarning: The method `BaseLLM.__call__` was deprecated in langch
ain-core 0.1.7 and will be removed in 1.0. Use :meth:`~invoke` instead.
  extraction_result = llm(extraction_prompt.format(query=query))
```

Observation:

1. Similarities between Smartphones and 2023:
- Both are related to technology and advancements in the digital world.
- Both have a significant impact on our daily lives and are constantly evolving.
- Both are constantly being improved and upgraded to meet the changing needs and demands of consumers.
- Both have a wide range of features and capabilities that make them essential tools for communication, entertainment, and productivity.

2. Differences between Smartphones and 2023:
- Smartphones are physical devices, while 2023 is a specific point in time.
- Smartphones are tangible products that can be purchased and used, while 2023 is a concept that represents the future.
- Smartphones have a specific set of features and capabilities, while 2023 is open to interpretation and can bring about a variety of changes and advancements in technology.

3. Strengths and weaknesses of Smartphones:
- Strengths:
  - Portability: Smartphones are small and lightweight, making them easy to carry around and use on the go.
  - Connectivity: Smartphones allow us to stay connected with others through various means such as calls, texts, and social media.
  - Multi-functionality: Smartphones have a wide range of features and capabilities, making them versatile tools for communication
Thought: I should use the Analyze tool to extract key information and insights from the comparison.
Action: Analyze
Action Input: "Comparison of top smartphones in 2023"
Observation:

1. Main findings and key points:
- The top smartphones in 2023 are expected to be the iPhone 13, Samsung Galaxy S23, and Google Pixel 6.
- These smartphones are expected to have advanced features such as 5G connectivity, improved cameras, and longer battery life.
- The iPhone 13 is projected to have the highest sales, followed by the Samsung Galaxy S23 and Google Pixel 6.

2. Important patterns or trends identified:
- The top smartphones in 2023 are all from well-established brands with a strong reputation in the market.
- The focus of these smartphones is on improving connectivity, camera quality, and battery life, indicating a growing demand for these features among consumers.
- The projected sales of these smartphones suggest that there will be a high demand for them in the market.

3. Critical takeaways:
- The competition among smartphone brands is expected to be intense in 2023, with the top brands vying for the top spot.
- Consumers are increasingly looking for advanced features in their smartphones, and brands need to keep up with these demands to stay competitive.
- The success of these top smartphones will depend on their ability to deliver on their promised features and meet consumer expectations.

4. Actionable insights or recommendations:

Thought: **I now know the final answer.**
Final Answer: The top 3 smartphones in 2023 are expected to be the iPhone 13, Samsung Galaxy S23, and Google Pixel 6. They will have advanced features such as 5G connectivity, improved cameras, and longer battery life. The competition among smartphone brands will be intense, and brands need to focus on delivering on consumer demands for advanced features to stay competitive.

> Finished chain.
The top 3 smartphones in 2023 are expected to be the iPhone 13, Samsung Galaxy S23, and Google Pixel 6. They will have advanced features such as 5G connectivity, improved cameras, and longer battery life. The competition among smartphone brands will be intense, and brands need to focus on delivering on consumer demands for advanced features to stay competitive.

In [13]:
```python
import streamlit as st

st.title("ReAct Agent")

query = st.text_area(
    "Enter your query:"
)

if st.button("Submit Query"):
    if query:
        result = process_query(query)

        st.subheader("Final Answer")
        st.write(result)
    else:
        st.warning("Please enter a query")
```

```
2025-11-19 00:49:23.634 Thread 'MainThread': missing ScriptRunContext! This warning can be ignored when running in bare mode.
2025-11-19 00:49:23.635 Thread 'MainThread': missing ScriptRunContext! This warning can be ignored when running in bare mode.
2025-11-19 00:49:23.636 Thread 'MainThread': missing ScriptRunContext! This warning can be ignored when running in bare mode.
2025-11-19 00:49:23.636 Thread 'MainThread': missing ScriptRunContext! This warning can be ignored when running in bare mode.
2025-11-19 00:49:23.637 Thread 'MainThread': missing ScriptRunContext! This warning can be ignored when running in bare mode.
2025-11-19 00:49:23.638 Thread 'MainThread': missing ScriptRunContext! This warning can be ignored when running in bare mode.
2025-11-19 00:49:23.639 Thread 'MainThread': missing ScriptRunContext! This warning can be ignored when running in bare mode.
2025-11-19 00:49:23.639 Thread 'MainThread': missing ScriptRunContext! This warning can be ignored when running in bare mode.
2025-11-19 00:49:23.640 Thread 'MainThread': missing ScriptRunContext! This warning can be ignored when running in bare mode.
2025-11-19 00:49:23.641 Thread 'MainThread': missing ScriptRunContext! This warning can be ignored when running in bare mode.
2025-11-19 00:49:23.642 Thread 'MainThread': missing ScriptRunContext! This warning can be ignored when running in bare mode.
2025-11-19 00:49:23.642 Thread 'MainThread': missing ScriptRunContext! This warning can be ignored when running in bare mode.
2025-11-19 00:49:23.643 Thread 'MainThread': missing ScriptRunContext! This warning can be ignored when running in bare mode.
2025-11-19 00:49:23.644 Thread 'MainThread': missing ScriptRunContext! This warning can be ignored when running in bare mode.
2025-11-19 00:49:23.644 Thread 'MainThread': missing ScriptRunContext! This warning can be ignored when running in bare mode.
2025-11-19 00:49:23.645 Thread 'MainThread': missing ScriptRunContext! This warning can be ignored when running in bare mode.
```

In [14]:
```
!streamlit run /usr/local/lib/python3.11/dist-packages/colab_kernel_launcher.py
```

```
Collecting usage statistics. To deactivate, set browser.gatherUsageStats to false.


  You can now view your Streamlit app in your browser.

  Local URL: http://localhost:8501
  Network URL: http://172.19.2.2:8501
  External URL: http://104.154.77.242:8501

^C
  Stopping...
```

Submission Requirements

Please submit the following items as part of your solution:

1. Your complete code implementation for the ReAct agent and its tools.
2. A sample question that you used to test your tool (make it complex enough to demonstrate the use of multiple tools).
3. The final answer provided by your ReAct agent for the sample question.
4. The complete history traces of the ReAct agent for your sample question, showing its thought process, actions, and observations. Your traces should follow a format similar to this example:

```
Thought: I need to find information about top smartphones first
Action: Search[top smartphones 2023]
Observation: [Search results about top smartphones]
Thought: Now I should compare the top two options
Action: Compare[iPhone 14 Pro, Samsung Galaxy S23 Ultra, smartphones]
Observation: [Comparison result]
Thought: I should analyze this comparison for the user
Action: Analyze[comparison result]
Observation: [Analysis of the comparison]
Final Answer: [Your agent's final response to the user's query]
```

Ensure that your submission clearly demonstrates the agent's ability to reason about which tool to use and how to interpret the results from each tool. Your history traces should show a logical flow of thoughts, actions, and observations, culminating in a final answer that addresses the initial query.

Note: Ensure that your ReAct agent can seamlessly switch between these tools based on the task at hand. The agent should be able to reason about which tool to use next and how to interpret the results from each tool.

References

[1] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). ReAct: Synergizing reasoning and acting in language models. arXiv preprint arXiv:2210.03629. https://arxiv.org/pdf/2210.03629

In [ ]:

Problem 4

Getting started with TinyTimeMixer (TTM)

This notebooke demonstrates the usage of a pre-trained `TinyTimeMixer` model for several multivariate time series forecasting tasks. For details related to model architecture, refer to the TTM paper.

We will use a pre-trained TTM-1024-96 model. That means the TTM model can take an input of 1024 time points (`context_length`), and can forecast upto 96 time points (`forecast_length`) in the future. We will use the pre-trained TTM in two settings:

1. **Zero-shot**: The pre-trained TTM will be directly used to evaluate on the `test` split of the target data. Note that the TTM was NOT pre-trained on the target data.
2. **Few-shot**: The pre-trained TTM will be quickly fine-tuned on only 5% of the `train` split of the target data, and subsequently, evaluated on the `test` part of the target data.

Pre-trained TTM models will be fetched from the Hugging Face TTM Model Repository.

1. TTM-R1 pre-trained models can be found here: TTM-R1 Model Card
   A. For 512-96 model set `TTM_MODEL_REVISION="main"`
   B. For 1024-96 model set `TTM_MODEL_REVISION="1024_96_v1"`
2. TTM-R2 pre-trained models can be found here: TTM-R2 Model Card
   A. For 512-96 model set `TTM_MODEL_REVISION="main"`
   B. For 1024-96 model set `TTM_MODEL_REVISION="1024-96-r2"`
   C. For 1536-96 model set `TTM_MODEL_REVISION="1536-96-r2"`

Details about the revisions (R1 and R2) can be found here.

4.1 Environment Setup - Installation

In [1]:
```
# Install the tsfm library
! pip install "tsfm_public[notebooks] @ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18"
```

```
Collecting tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18 (from tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18)
  Cloning https://github.com/ibm-granite/granite-tsfm.git (to revision v0.2.18) to /tmp/pip-install-atqsf0yg/tsfm-public_61bee257b11c4b7089ae80fad97c2d82
  Running command git clone --filter=blob:none --quiet https://github.com/ibm-granite/granite-tsfm.git /tmp/pip-install-atqsf0yg/tsfm-public_61bee257b11c4b7089ae80fad97c2d82
  Running command git checkout -q 4db1cfd157767c8de39bedb5c1f90a8d7d6e5850
  Resolved https://github.com/ibm-granite/granite-tsfm.git to commit 4db1cfd157767c8de39bedb5c1f90a8d7d6e5850
  Installing build dependencies ... done
  Getting requirements to build wheel ... done
  Preparing metadata (pyproject.toml) ... done
Requirement already satisfied: pandas>=2.2.0 in /usr/local/lib/python3.11/dist-packages (from tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (2.2.3)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.11/dist-packages (from tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (1.2.2)
Requirement already satisfied: transformers>=4.38.0 in /usr/local/lib/python3.11/dist-packages (from transformers[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (4.53.3)
Requirement already satisfied: datasets in /usr/local/lib/python3.11/dist-packages (from tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (4.4.1)
Requirement already satisfied: deprecated in /usr/local/lib/python3.11/dist-packages (from tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (1.3.1)
Collecting urllib3<2,>=1.26.19 (from tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18)
  Downloading urllib3-1.26.20-py2.py3-none-any.whl.metadata (50 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 50.1/50.1 kB 1.6 MB/s eta 0:00:00
Requirement already satisfied: numpy<2 in /usr/local/lib/python3.11/dist-packages (from tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (1.26.4)
Collecting jupyter (from tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18)
  Downloading jupyter-1.1.1-py2.py3-none-any.whl.metadata (2.0 kB)
Requirement already satisfied: matplotlib in /usr/local/lib/python3.11/dist-packages (from tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (3.7.2)
Requirement already satisfied: ipywidgets in /usr/local/lib/python3.11/dist-packages (from tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (8.1.5)
Requirement already satisfied: plotly in /usr/local/lib/python3.11/dist-packages (from tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (5.24.1)
Collecting kaleido (from tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18)
  Downloading kaleido-1.2.0-py3-none-any.whl.metadata (5.6 kB)
```

```
Requirement already satisfied: tensorboard in /usr/local/lib/python3.11/dist-packages (from tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (2.18.0)
Requirement already satisfied: mkl_fft in /usr/local/lib/python3.11/dist-packages (from numpy<2->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (1.3.8)
Requirement already satisfied: mkl_random in /usr/local/lib/python3.11/dist-packages (from numpy<2->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (1.2.4)
Requirement already satisfied: mkl_umath in /usr/local/lib/python3.11/dist-packages (from numpy<2->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (0.1.1)
Requirement already satisfied: mkl in /usr/local/lib/python3.11/dist-packages (from numpy<2->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (2025.3.0)
Requirement already satisfied: tbb4py in /usr/local/lib/python3.11/dist-packages (from numpy<2->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (2022.3.0)
Requirement already satisfied: mkl-service in /usr/local/lib/python3.11/dist-packages (from numpy<2->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (2.4.1)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas>=2.2.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from pandas>=2.2.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas>=2.2.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (2025.2)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from transformers>=4.38.0->transformers[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (3.20.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.30.0 in /usr/local/lib/python3.11/dist-packages (from transformers>=4.38.0->transformers[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (0.36.0)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from transformers>=4.38.0->transformers[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (25.0)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-packages (from transformers>=4.38.0->transformers[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (6.0.3)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.11/dist-packages (from transformers>=4.38.0->transformers[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (2025.11.3)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages (from transformers>=4.38.0->transformers[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (2.32.5)
```

```
Requirement already satisfied: tokenizers<0.22,>=0.21 in /usr/local/lib/python3.11/dist-packages (from transformers>=4.
38.0->transformers[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_publi
c[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (0.21.2)
Requirement already satisfied: safetensors>=0.4.3 in /usr/local/lib/python3.11/dist-packages (from transformers>=4.38.0
->transformers[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[no
tebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (0.5.3)
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.11/dist-packages (from transformers>=4.38.0->transf
ormers[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]
@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (4.67.1)
Requirement already satisfied: torch>=2.1 in /usr/local/lib/python3.11/dist-packages (from transformers[torch]>=4.38.0-
>tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.c
om/ibm-granite/granite-tsfm.git@v0.2.18) (2.6.0+cu124)
Requirement already satisfied: accelerate>=0.26.0 in /usr/local/lib/python3.11/dist-packages (from transformers[torch]>
=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://
github.com/ibm-granite/granite-tsfm.git@v0.2.18) (1.9.0)
Collecting pyarrow>=21.0.0 (from datasets->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->ts
fm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18)
  Downloading pyarrow-22.0.0-cp311-cp311-manylinux_2_28_x86_64.whl.metadata (3.2 kB)
Requirement already satisfied: dill<0.4.1,>=0.3.0 in /usr/local/lib/python3.11/dist-packages (from datasets->tsfm_publi
c@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-gran
ite/granite-tsfm.git@v0.2.18) (0.4.0)
Requirement already satisfied: httpx<1.0.0 in /usr/local/lib/python3.11/dist-packages (from datasets->tsfm_public@ git+
https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/gra
nite-tsfm.git@v0.2.18) (0.28.1)
Requirement already satisfied: xxhash in /usr/local/lib/python3.11/dist-packages (from datasets->tsfm_public@ git+https
://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite
-tsfm.git@v0.2.18) (3.6.0)
Requirement already satisfied: multiprocess<0.70.19 in /usr/local/lib/python3.11/dist-packages (from datasets->tsfm_pub
lic@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-gr
anite/granite-tsfm.git@v0.2.18) (0.70.18)
Requirement already satisfied: fsspec<=2025.10.0,>=2023.1.0 in /usr/local/lib/python3.11/dist-packages (from fsspec[htt
p]<=2025.10.0,>=2023.1.0->datasets->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_publ
ic[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (2025.10.0)
Requirement already satisfied: wrapt<3,>=1.10 in /usr/local/lib/python3.11/dist-packages (from deprecated->tsfm_public@
git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granit
e/granite-tsfm.git@v0.2.18) (1.17.2)
Requirement already satisfied: comm>=0.1.3 in /usr/local/lib/python3.11/dist-packages (from ipywidgets->tsfm_public@ gi
t+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/g
ranite-tsfm.git@v0.2.18) (0.2.3)
Requirement already satisfied: ipython>=6.1.0 in /usr/local/lib/python3.11/dist-packages (from ipywidgets->tsfm_public@
git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granit
e/granite-tsfm.git@v0.2.18) (7.34.0)
Requirement already satisfied: traitlets>=4.3.1 in /usr/local/lib/python3.11/dist-packages (from ipywidgets->tsfm_publi
c@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-gran
ite/granite-tsfm.git@v0.2.18) (5.7.1)
Requirement already satisfied: widgetsnbextension~=4.0.12 in /usr/local/lib/python3.11/dist-packages (from ipywidgets->
tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.co
m/ibm-granite/granite-tsfm.git@v0.2.18) (4.0.15)
```

```
Requirement already satisfied: jupyterlab-widgets~=3.0.12 in /usr/local/lib/python3.11/dist-packages (from ipywidgets->
tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.co
m/ibm-granite/granite-tsfm.git@v0.2.18) (3.0.15)
Requirement already satisfied: notebook in /usr/local/lib/python3.11/dist-packages (from jupyter->tsfm_public@ git+http
s://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite
-tsfm.git@v0.2.18) (6.5.4)
Requirement already satisfied: jupyter-console in /usr/local/lib/python3.11/dist-packages (from jupyter->tsfm_public@ g
it+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/
granite-tsfm.git@v0.2.18) (6.1.0)
Requirement already satisfied: nbconvert in /usr/local/lib/python3.11/dist-packages (from jupyter->tsfm_public@ git+htt
ps://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granit
e-tsfm.git@v0.2.18) (6.4.5)
Requirement already satisfied: ipykernel in /usr/local/lib/python3.11/dist-packages (from jupyter->tsfm_public@ git+htt
ps://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granit
e-tsfm.git@v0.2.18) (6.17.1)
Requirement already satisfied: jupyterlab in /usr/local/lib/python3.11/dist-packages (from jupyter->tsfm_public@ git+ht
tps://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/grani
te-tsfm.git@v0.2.18) (3.6.8)
Collecting choreographer>=1.1.1 (from kaleido->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18
->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18)
  Downloading choreographer-1.2.1-py3-none-any.whl.metadata (6.8 kB)
Collecting logistro>=1.0.8 (from kaleido->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsf
m_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18)
  Downloading logistro-2.0.1-py3-none-any.whl.metadata (3.9 kB)
Requirement already satisfied: orjson>=3.10.15 in /usr/local/lib/python3.11/dist-packages (from kaleido->tsfm_public@ g
it+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/
granite-tsfm.git@v0.2.18) (3.11.0)
Collecting pytest-timeout>=2.4.0 (from kaleido->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.1
8->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18)
  Downloading pytest_timeout-2.4.0-py3-none-any.whl.metadata (20 kB)
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib->tsfm_publi
c@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-gran
ite/granite-tsfm.git@v0.2.18) (1.3.2)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.11/dist-packages (from matplotlib->tsfm_public@ g
it+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/
granite-tsfm.git@v0.2.18) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.11/dist-packages (from matplotlib->tsfm_publ
ic@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-gra
nite/granite-tsfm.git@v0.2.18) (4.59.0)
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib->tsfm_publ
ic@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-gra
nite/granite-tsfm.git@v0.2.18) (1.4.8)
Requirement already satisfied: pillow>=6.2.0 in /usr/local/lib/python3.11/dist-packages (from matplotlib->tsfm_public@
git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granit
e/granite-tsfm.git@v0.2.18) (11.3.0)
Requirement already satisfied: pyparsing<3.1,>=2.3.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib->tsfm_
public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm
-granite/granite-tsfm.git@v0.2.18) (3.0.9)
```

```
Requirement already satisfied: tenacity>=6.2.0 in /usr/local/lib/python3.11/dist-packages (from plotly->tsfm_public@ gi
t+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/g
ranite-tsfm.git@v0.2.18) (9.1.2)
Requirement already satisfied: scipy>=1.3.2 in /usr/local/lib/python3.11/dist-packages (from scikit-learn->tsfm_public@
git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granit
e/granite-tsfm.git@v0.2.18) (1.15.3)
Requirement already satisfied: joblib>=1.1.1 in /usr/local/lib/python3.11/dist-packages (from scikit-learn->tsfm_public
@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-grani
te/granite-tsfm.git@v0.2.18) (1.5.2)
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn->tsfm
_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ib
m-granite/granite-tsfm.git@v0.2.18) (3.6.0)
Requirement already satisfied: absl-py>=0.4 in /usr/local/lib/python3.11/dist-packages (from tensorboard->tsfm_public@
git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granit
e/granite-tsfm.git@v0.2.18) (1.4.0)
Requirement already satisfied: grpcio>=1.48.2 in /usr/local/lib/python3.11/dist-packages (from tensorboard->tsfm_public
@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-grani
te/granite-tsfm.git@v0.2.18) (1.74.0)
Requirement already satisfied: markdown>=2.6.8 in /usr/local/lib/python3.11/dist-packages (from tensorboard->tsfm_publi
c@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-gran
ite/granite-tsfm.git@v0.2.18) (3.8.2)
Requirement already satisfied: protobuf!=4.24.0,>=3.19.6 in /usr/local/lib/python3.11/dist-packages (from tensorboard->
tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.co
m/ibm-granite/granite-tsfm.git@v0.2.18) (6.33.0)
Requirement already satisfied: setuptools>=41.0.0 in /usr/local/lib/python3.11/dist-packages (from tensorboard->tsfm_pu
blic@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-g
ranite/granite-tsfm.git@v0.2.18) (75.2.0)
Requirement already satisfied: six>1.9 in /usr/local/lib/python3.11/dist-packages (from tensorboard->tsfm_public@ git+h
ttps://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/gran
ite-tsfm.git@v0.2.18) (1.17.0)
Requirement already satisfied: tensorboard-data-server<0.8.0,>=0.7.0 in /usr/local/lib/python3.11/dist-packages (from t
ensorboard->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+http
s://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (0.7.2)
Requirement already satisfied: werkzeug>=1.0.1 in /usr/local/lib/python3.11/dist-packages (from tensorboard->tsfm_publi
c@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-gran
ite/granite-tsfm.git@v0.2.18) (3.1.3)
Requirement already satisfied: psutil in /usr/local/lib/python3.11/dist-packages (from accelerate>=0.26.0->transformers
[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+
https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (7.1.3)
Requirement already satisfied: simplejson>=3.19.3 in /usr/local/lib/python3.11/dist-packages (from choreographer>=1.1.1
->kaleido->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+http
s://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (3.20.1)
Requirement already satisfied: aiohttp!=4.0.0a0,!=4.0.0a1 in /usr/local/lib/python3.11/dist-packages (from fsspec[http]
<=2025.10.0,>=2023.1.0->datasets->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public
[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (3.13.2)
Requirement already satisfied: anyio in /usr/local/lib/python3.11/dist-packages (from httpx<1.0.0->datasets->tsfm_publi
c@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-gran
ite/granite-tsfm.git@v0.2.18) (4.11.0)
```

```
Requirement already satisfied: certifi in /usr/local/lib/python3.11/dist-packages (from httpx<1.0.0->datasets->tsfm_pub
lic@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-gr
anite/granite-tsfm.git@v0.2.18) (2025.10.5)
Requirement already satisfied: httpcore==1.* in /usr/local/lib/python3.11/dist-packages (from httpx<1.0.0->datasets->ts
fm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/
ibm-granite/granite-tsfm.git@v0.2.18) (1.0.9)
Requirement already satisfied: idna in /usr/local/lib/python3.11/dist-packages (from httpx<1.0.0->datasets->tsfm_public
@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-grani
te/granite-tsfm.git@v0.2.18) (3.11)
Requirement already satisfied: h11>=0.16 in /usr/local/lib/python3.11/dist-packages (from httpcore==1.*->httpx<1.0.0->d
atasets->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://
github.com/ibm-granite/granite-tsfm.git@v0.2.18) (0.16.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.11/dist-packages (from huggingface-
hub<1.0,>=0.30.0->transformers>=4.38.0->transformers[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/gr
anite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (4.15.0)
Requirement already satisfied: hf-xet<2.0.0,>=1.1.3 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.
0,>=0.30.0->transformers>=4.38.0->transformers[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-
tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (1.2.0)
Requirement already satisfied: jedi>=0.16 in /usr/local/lib/python3.11/dist-packages (from ipython>=6.1.0->ipywidgets->
tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.co
m/ibm-granite/granite-tsfm.git@v0.2.18) (0.19.2)
Requirement already satisfied: decorator in /usr/local/lib/python3.11/dist-packages (from ipython>=6.1.0->ipywidgets->t
sfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.co
m/ibm-granite/granite-tsfm.git@v0.2.18) (4.4.2)
Requirement already satisfied: pickleshare in /usr/local/lib/python3.11/dist-packages (from ipython>=6.1.0->ipywidgets-
>tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.c
om/ibm-granite/granite-tsfm.git@v0.2.18) (0.7.5)
Requirement already satisfied: prompt-toolkit!=3.0.0,!=3.0.1,<3.1.0,>=2.0.0 in /usr/local/lib/python3.11/dist-packages
(from ipython>=6.1.0->ipywidgets->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public
[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (3.0.51)
Requirement already satisfied: pygments in /usr/local/lib/python3.11/dist-packages (from ipython>=6.1.0->ipywidgets->ts
fm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/
ibm-granite/granite-tsfm.git@v0.2.18) (2.19.2)
Requirement already satisfied: backcall in /usr/local/lib/python3.11/dist-packages (from ipython>=6.1.0->ipywidgets->ts
fm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/
ibm-granite/granite-tsfm.git@v0.2.18) (0.2.0)
Requirement already satisfied: matplotlib-inline in /usr/local/lib/python3.11/dist-packages (from ipython>=6.1.0->ipywi
dgets->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://gi
thub.com/ibm-granite/granite-tsfm.git@v0.2.18) (0.1.7)
Requirement already satisfied: pexpect>4.3 in /usr/local/lib/python3.11/dist-packages (from ipython>=6.1.0->ipywidgets-
>tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.c
om/ibm-granite/granite-tsfm.git@v0.2.18) (4.9.0)
Requirement already satisfied: pytest>=7.0.0 in /usr/local/lib/python3.11/dist-packages (from pytest-timeout>=2.4.0->ka
leido->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://gi
thub.com/ibm-granite/granite-tsfm.git@v0.2.18) (8.4.1)
Requirement already satisfied: charset_normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests->tran
sformers>=4.38.0->transformers[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18
->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (3.4.4)
```

```
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-packages (from torch>=2.1->transformers[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (3.5)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.11/dist-packages (from torch>=2.1->transformers[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (3.1.6)
Collecting nvidia-cuda-nvrtc-cu12==12.4.127 (from torch>=2.1->transformers[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18)
  Downloading nvidia_cuda_nvrtc_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cuda-runtime-cu12==12.4.127 (from torch>=2.1->transformers[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18)
  Downloading nvidia_cuda_runtime_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cuda-cupti-cu12==12.4.127 (from torch>=2.1->transformers[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18)
  Downloading nvidia_cuda_cupti_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cudnn-cu12==9.1.0.70 (from torch>=2.1->transformers[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18)
  Downloading nvidia_cudnn_cu12-9.1.0.70-py3-none-manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cublas-cu12==12.4.5.8 (from torch>=2.1->transformers[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18)
  Downloading nvidia_cublas_cu12-12.4.5.8-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cufft-cu12==11.2.1.3 (from torch>=2.1->transformers[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18)
  Downloading nvidia_cufft_cu12-11.2.1.3-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-curand-cu12==10.3.5.147 (from torch>=2.1->transformers[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18)
  Downloading nvidia_curand_cu12-10.3.5.147-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cusolver-cu12==11.6.1.9 (from torch>=2.1->transformers[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18)
  Downloading nvidia_cusolver_cu12-11.6.1.9-py3-none-manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cusparse-cu12==12.3.1.170 (from torch>=2.1->transformers[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18)
  Downloading nvidia_cusparse_cu12-12.3.1.170-py3-none-manylinux2014_x86_64.whl.metadata (1.6 kB)
Requirement already satisfied: nvidia-cusparselt-cu12==0.6.2 in /usr/local/lib/python3.11/dist-packages (from torch>=2.1->transformers[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (0.6.2)
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in /usr/local/lib/python3.11/dist-packages (from torch>=2.1->transformers[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (2.21.5)
```

```
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=2.1->
transformers[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[note
books]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (12.4.127)
Collecting nvidia-nvjitlink-cu12==12.4.127 (from torch>=2.1->transformers[torch]>=4.38.0->tsfm_public@ git+https://gith
ub.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.gi
t@v0.2.18)
  Downloading nvidia_nvjitlink_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Requirement already satisfied: triton==3.2.0 in /usr/local/lib/python3.11/dist-packages (from torch>=2.1->transformers
[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+
https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (3.2.0)
Requirement already satisfied: sympy==1.13.1 in /usr/local/lib/python3.11/dist-packages (from torch>=2.1->transformers
[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+
https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (1.13.1)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.11/dist-packages (from sympy==1.13.1->torch
>=2.1->transformers[torch]>=4.38.0->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_publ
ic[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (1.3.0)
Requirement already satisfied: MarkupSafe>=2.1.1 in /usr/local/lib/python3.11/dist-packages (from werkzeug>=1.0.1->tens
orboard->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://
github.com/ibm-granite/granite-tsfm.git@v0.2.18) (3.0.3)
Requirement already satisfied: debugpy>=1.0 in /usr/local/lib/python3.11/dist-packages (from ipykernel->jupyter->tsfm_p
ublic@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-
granite/granite-tsfm.git@v0.2.18) (1.8.15)
Requirement already satisfied: jupyter-client>=6.1.12 in /usr/local/lib/python3.11/dist-packages (from ipykernel->jupyt
er->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://githu
b.com/ibm-granite/granite-tsfm.git@v0.2.18) (8.6.3)
Requirement already satisfied: nest-asyncio in /usr/local/lib/python3.11/dist-packages (from ipykernel->jupyter->tsfm_p
ublic@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-
granite/granite-tsfm.git@v0.2.18) (1.6.0)
Requirement already satisfied: pyzmq>=17 in /usr/local/lib/python3.11/dist-packages (from ipykernel->jupyter->tsfm_publ
ic@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-gra
nite/granite-tsfm.git@v0.2.18) (26.2.1)
Requirement already satisfied: tornado>=6.1 in /usr/local/lib/python3.11/dist-packages (from ipykernel->jupyter->tsfm_p
ublic@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-
granite/granite-tsfm.git@v0.2.18) (6.5.2)
Requirement already satisfied: jupyter-core in /usr/local/lib/python3.11/dist-packages (from jupyterlab->jupyter->tsfm_
public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm
-granite/granite-tsfm.git@v0.2.18) (5.8.1)
Requirement already satisfied: jupyterlab-server~=2.19 in /usr/local/lib/python3.11/dist-packages (from jupyterlab->jup
yter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://git
hub.com/ibm-granite/granite-tsfm.git@v0.2.18) (2.28.0)
Requirement already satisfied: jupyter-server<3,>=1.16.0 in /usr/local/lib/python3.11/dist-packages (from jupyterlab->j
upyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://g
ithub.com/ibm-granite/granite-tsfm.git@v0.2.18) (2.12.5)
Requirement already satisfied: jupyter-ydoc~=0.2.4 in /usr/local/lib/python3.11/dist-packages (from jupyterlab->jupyter
->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.
com/ibm-granite/granite-tsfm.git@v0.2.18) (0.2.5)
Requirement already satisfied: jupyter-server-ydoc~=0.8.0 in /usr/local/lib/python3.11/dist-packages (from jupyterlab->
jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://
```

```
github.com/ibm-granite/granite-tsfm.git@v0.2.18) (0.8.0)
Requirement already satisfied: nbclassic in /usr/local/lib/python3.11/dist-packages (from jupyterlab->jupyter->tsfm_pub
lic@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-gr
anite/granite-tsfm.git@v0.2.18) (1.3.1)
Requirement already satisfied: argon2-cffi in /usr/local/lib/python3.11/dist-packages (from notebook->jupyter->tsfm_pub
lic@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-gr
anite/granite-tsfm.git@v0.2.18) (25.1.0)
Requirement already satisfied: ipython-genutils in /usr/local/lib/python3.11/dist-packages (from notebook->jupyter->tsf
m_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/i
bm-granite/granite-tsfm.git@v0.2.18) (0.2.0)
Requirement already satisfied: nbformat in /usr/local/lib/python3.11/dist-packages (from notebook->jupyter->tsfm_public
@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-grani
te/granite-tsfm.git@v0.2.18) (5.10.4)
Requirement already satisfied: Send2Trash>=1.8.0 in /usr/local/lib/python3.11/dist-packages (from notebook->jupyter->ts
fm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/
ibm-granite/granite-tsfm.git@v0.2.18) (1.8.3)
Requirement already satisfied: terminado>=0.8.3 in /usr/local/lib/python3.11/dist-packages (from notebook->jupyter->tsf
m_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/i
bm-granite/granite-tsfm.git@v0.2.18) (0.18.1)
Requirement already satisfied: prometheus-client in /usr/local/lib/python3.11/dist-packages (from notebook->jupyter->ts
fm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/
ibm-granite/granite-tsfm.git@v0.2.18) (0.22.1)
Requirement already satisfied: mistune<2,>=0.8.1 in /usr/local/lib/python3.11/dist-packages (from nbconvert->jupyter->t
sfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.co
m/ibm-granite/granite-tsfm.git@v0.2.18) (0.8.4)
Requirement already satisfied: jupyterlab-pygments in /usr/local/lib/python3.11/dist-packages (from nbconvert->jupyter-
>tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.c
om/ibm-granite/granite-tsfm.git@v0.2.18) (0.3.0)
Requirement already satisfied: entrypoints>=0.2.2 in /usr/local/lib/python3.11/dist-packages (from nbconvert->jupyter->
tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.co
m/ibm-granite/granite-tsfm.git@v0.2.18) (0.4)
Requirement already satisfied: bleach in /usr/local/lib/python3.11/dist-packages (from nbconvert->jupyter->tsfm_public@
git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granit
e/granite-tsfm.git@v0.2.18) (6.2.0)
Requirement already satisfied: pandocfilters>=1.4.1 in /usr/local/lib/python3.11/dist-packages (from nbconvert->jupyter
->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.
com/ibm-granite/granite-tsfm.git@v0.2.18) (1.5.1)
Requirement already satisfied: testpath in /usr/local/lib/python3.11/dist-packages (from nbconvert->jupyter->tsfm_publi
c@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-gran
ite/granite-tsfm.git@v0.2.18) (0.6.0)
Requirement already satisfied: defusedxml in /usr/local/lib/python3.11/dist-packages (from nbconvert->jupyter->tsfm_pub
lic@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-gr
anite/granite-tsfm.git@v0.2.18) (0.7.1)
Requirement already satisfied: beautifulsoup4 in /usr/local/lib/python3.11/dist-packages (from nbconvert->jupyter->tsfm
_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ib
m-granite/granite-tsfm.git@v0.2.18) (4.13.4)
Requirement already satisfied: nbclient<0.6.0,>=0.5.0 in /usr/local/lib/python3.11/dist-packages (from nbconvert->jupyt
er->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://githu
```

```
b.com/ibm-granite/granite-tsfm.git@v0.2.18) (0.5.13)
Requirement already satisfied: onemkl-license==2025.3.0 in /usr/local/lib/python3.11/dist-packages (from mkl->numpy<2->
tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.co
m/ibm-granite/granite-tsfm.git@v0.2.18) (2025.3.0)
Requirement already satisfied: intel-openmp<2026,>=2024 in /usr/local/lib/python3.11/dist-packages (from mkl->numpy<2->
tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.co
m/ibm-granite/granite-tsfm.git@v0.2.18) (2024.2.0)
Requirement already satisfied: tbb==2022.* in /usr/local/lib/python3.11/dist-packages (from mkl->numpy<2->tsfm_public@
git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granit
e/granite-tsfm.git@v0.2.18) (2022.3.0)
Requirement already satisfied: tcmlib==1.* in /usr/local/lib/python3.11/dist-packages (from tbb==2022.*->mkl->numpy<2->
tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.co
m/ibm-granite/granite-tsfm.git@v0.2.18) (1.4.0)
Requirement already satisfied: intel-cmplr-lib-rt in /usr/local/lib/python3.11/dist-packages (from mkl_umath->numpy<2->
tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.co
m/ibm-granite/granite-tsfm.git@v0.2.18) (2024.2.0)
Requirement already satisfied: aiohappyeyeballs>=2.5.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a
0,!=4.0.0a1->fsspec[http]<=2025.10.0,>=2023.1.0->datasets->tsfm_public@ git+https://github.com/ibm-granite/granite-tsf
m.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (2.6.1)
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]<=2025.10.0,>=2023.1.0->datasets->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v
0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (1.4.0)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a
1->fsspec[http]<=2025.10.0,>=2023.1.0->datasets->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.
18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (25.4.0)
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]<=2025.10.0,>=2023.1.0->datasets->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v
0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (1.8.0)
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]<=2025.10.0,>=2023.1.0->datasets->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git
@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (6.7.0)
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]<=2025.10.0,>=2023.1.0->datasets->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v
0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (0.4.1)
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]<=2025.10.0,>=2023.1.0->datasets->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v
0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (1.22.0)
Requirement already satisfied: intel-cmplr-lib-ur==2024.2.0 in /usr/local/lib/python3.11/dist-packages (from intel-open
mp<2026,>=2024->mkl->numpy<2->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[not
ebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (2024.2.0)
Requirement already satisfied: parso<0.9.0,>=0.8.4 in /usr/local/lib/python3.11/dist-packages (from jedi>=0.16->ipython
>=6.1.0->ipywidgets->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@
git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (0.8.4)
Requirement already satisfied: platformdirs>=2.5 in /usr/local/lib/python3.11/dist-packages (from jupyter-core->jupyter
lab->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+htt
ps://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (4.5.0)
Requirement already satisfied: jupyter-events>=0.9.0 in /usr/local/lib/python3.11/dist-packages (from jupyter-server<3,
>=1.16.0->jupyterlab->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[no
```

```
tebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (0.12.0)
Requirement already satisfied: jupyter-server-terminals in /usr/local/lib/python3.11/dist-packages (from jupyter-server
<3,>=1.16.0->jupyterlab->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public
[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (0.5.3)
Requirement already satisfied: overrides in /usr/local/lib/python3.11/dist-packages (from jupyter-server<3,>=1.16.0->ju
pyterlab->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ gi
t+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (7.7.0)
Requirement already satisfied: websocket-client in /usr/local/lib/python3.11/dist-packages (from jupyter-server<3,>=1.1
6.0->jupyterlab->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[noteboo
ks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (1.8.0)
Requirement already satisfied: sniffio>=1.1 in /usr/local/lib/python3.11/dist-packages (from anyio->httpx<1.0.0->datase
ts->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://githu
b.com/ibm-granite/granite-tsfm.git@v0.2.18) (1.3.1)
Requirement already satisfied: jupyter-server-fileid<1,>=0.6.0 in /usr/local/lib/python3.11/dist-packages (from jupyter
-server-ydoc~=0.8.0->jupyterlab->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsf
m_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (0.9.3)
Requirement already satisfied: ypy-websocket<0.9.0,>=0.8.2 in /usr/local/lib/python3.11/dist-packages (from jupyter-ser
ver-ydoc~=0.8.0->jupyterlab->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_pu
blic[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (0.8.4)
Requirement already satisfied: y-py<0.7.0,>=0.6.0 in /usr/local/lib/python3.11/dist-packages (from jupyter-ydoc~=0.2.4-
>jupyterlab->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@
git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (0.6.2)
Requirement already satisfied: babel>=2.10 in /usr/local/lib/python3.11/dist-packages (from jupyterlab-server~=2.19->ju
pyterlab->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ gi
t+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (2.17.0)
Requirement already satisfied: json5>=0.9.0 in /usr/local/lib/python3.11/dist-packages (from jupyterlab-server~=2.19->j
upyterlab->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ g
it+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (0.12.1)
Requirement already satisfied: jsonschema>=4.18.0 in /usr/local/lib/python3.11/dist-packages (from jupyterlab-server~=
2.19->jupyterlab->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[noteboo
oks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (4.25.0)
Requirement already satisfied: notebook-shim>=0.2.3 in /usr/local/lib/python3.11/dist-packages (from nbclassic->jupyter
lab->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+htt
ps://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (0.2.4)
Requirement already satisfied: fastjsonschema>=2.15 in /usr/local/lib/python3.11/dist-packages (from nbformat->notebook
->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+http
s://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (2.21.1)
Requirement already satisfied: ptyprocess>=0.5 in /usr/local/lib/python3.11/dist-packages (from pexpect>4.3->ipython>=
6.1.0->ipywidgets->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ gi
t+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (0.7.0)
Requirement already satisfied: wcwidth in /usr/local/lib/python3.11/dist-packages (from prompt-toolkit!=3.0.0,!=3.0.1,<
3.1.0,>=2.0.0->ipython>=6.1.0->ipywidgets->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->ts
fm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (0.2.13)
Requirement already satisfied: iniconfig>=1 in /usr/local/lib/python3.11/dist-packages (from pytest>=7.0.0->pytest-time
out>=2.4.0->kaleido->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@
git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (2.1.0)
Requirement already satisfied: pluggy<2,>=1.5 in /usr/local/lib/python3.11/dist-packages (from pytest>=7.0.0->pytest-ti
meout>=2.4.0->kaleido->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]
```

```
@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (1.6.0)
Requirement already satisfied: argon2-cffi-bindings in /usr/local/lib/python3.11/dist-packages (from argon2-cffi->noteb
ook->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+htt
ps://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (21.2.0)
Requirement already satisfied: soupsieve>1.2 in /usr/local/lib/python3.11/dist-packages (from beautifulsoup4->nbconvert
->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+http
s://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (2.7)
Requirement already satisfied: webencodings in /usr/local/lib/python3.11/dist-packages (from bleach->nbconvert->jupyter
->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.
com/ibm-granite/granite-tsfm.git@v0.2.18) (0.5.1)
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in /usr/local/lib/python3.11/dist-packages (from js
onschema>=4.18.0->jupyterlab-server~=2.19->jupyterlab->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite
-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (2025.4.1)
Requirement already satisfied: referencing>=0.28.4 in /usr/local/lib/python3.11/dist-packages (from jsonschema>=4.18.0-
>jupyterlab-server~=2.19->jupyterlab->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18
->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (0.36.2)
Requirement already satisfied: rpds-py>=0.7.1 in /usr/local/lib/python3.11/dist-packages (from jsonschema>=4.18.0->jupy
terlab-server~=2.19->jupyterlab->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsf
m_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (0.26.0)
Requirement already satisfied: python-json-logger>=2.0.4 in /usr/local/lib/python3.11/dist-packages (from jupyter-event
s>=0.9.0->jupyter-server<3,>=1.16.0->jupyterlab->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.
git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (4.0.0)
Requirement already satisfied: rfc3339-validator in /usr/local/lib/python3.11/dist-packages (from jupyter-events>=0.9.0
->jupyter-server<3,>=1.16.0->jupyterlab->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.
2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (0.1.4)
Requirement already satisfied: rfc3986-validator>=0.1.1 in /usr/local/lib/python3.11/dist-packages (from jupyter-events
>=0.9.0->jupyter-server<3,>=1.16.0->jupyterlab->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.g
it@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (0.1.1)
Requirement already satisfied: aiofiles<23,>=22.1.0 in /usr/local/lib/python3.11/dist-packages (from ypy-websocket<0.9.
0,>=0.8.2->jupyter-server-ydoc~=0.8.0->jupyterlab->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsf
m.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (22.1.0)
Requirement already satisfied: aiosqlite<1,>=0.17.0 in /usr/local/lib/python3.11/dist-packages (from ypy-websocket<0.9.
0,>=0.8.2->jupyter-server-ydoc~=0.8.0->jupyterlab->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsf
m.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (0.21.0)
Requirement already satisfied: cffi>=1.0.1 in /usr/local/lib/python3.11/dist-packages (from argon2-cffi-bindings->argon
2-cffi->notebook->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[noteb
ooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (2.0.0)
Requirement already satisfied: pycparser in /usr/local/lib/python3.11/dist-packages (from cffi>=1.0.1->argon2-cffi-bind
ings->argon2-cffi->notebook->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_pu
blic[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (2.23)
Requirement already satisfied: fqdn in /usr/local/lib/python3.11/dist-packages (from jsonschema[format-nongpl]>=4.18.0-
>jupyter-events>=0.9.0->jupyter-server<3,>=1.16.0->jupyterlab->jupyter->tsfm_public@ git+https://github.com/ibm-granit
e/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (1.5.
1)
Requirement already satisfied: isoduration in /usr/local/lib/python3.11/dist-packages (from jsonschema[format-nongpl]>=
4.18.0->jupyter-events>=0.9.0->jupyter-server<3,>=1.16.0->jupyterlab->jupyter->tsfm_public@ git+https://github.com/ibm-
granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18)
(20.11.0)
```

```
Requirement already satisfied: jsonpointer>1.13 in /usr/local/lib/python3.11/dist-packages (from jsonschema[format-nongpl]>=4.18.0->jupyter-events>=0.9.0->jupyter-server<3,>=1.16.0->jupyterlab->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (3.0.0)
Requirement already satisfied: rfc3987-syntax>=1.1.0 in /usr/local/lib/python3.11/dist-packages (from jsonschema[format-nongpl]>=4.18.0->jupyter-events>=0.9.0->jupyter-server<3,>=1.16.0->jupyterlab->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (1.1.0)
Requirement already satisfied: uri-template in /usr/local/lib/python3.11/dist-packages (from jsonschema[format-nongpl]>=4.18.0->jupyter-events>=0.9.0->jupyter-server<3,>=1.16.0->jupyterlab->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (1.3.0)
Requirement already satisfied: webcolors>=24.6.0 in /usr/local/lib/python3.11/dist-packages (from jsonschema[format-nongpl]>=4.18.0->jupyter-events>=0.9.0->jupyter-server<3,>=1.16.0->jupyterlab->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (24.11.1)
Requirement already satisfied: lark>=1.2.2 in /usr/local/lib/python3.11/dist-packages (from rfc3987-syntax>=1.1.0->jsonschema[format-nongpl]>=4.18.0->jupyter-events>=0.9.0->jupyter-server<3,>=1.16.0->jupyterlab->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (1.3.1)
Requirement already satisfied: arrow>=0.15.0 in /usr/local/lib/python3.11/dist-packages (from isoduration->jsonschema[format-nongpl]>=4.18.0->jupyter-events>=0.9.0->jupyter-server<3,>=1.16.0->jupyterlab->jupyter->tsfm_public@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18->tsfm_public[notebooks]@ git+https://github.com/ibm-granite/granite-tsfm.git@v0.2.18) (1.4.0)
Downloading urllib3-1.26.20-py2.py3-none-any.whl (144 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 144.2/144.2 kB 5.1 MB/s eta 0:00:00
Downloading jupyter-1.1.1-py2.py3-none-any.whl (2.7 kB)
Downloading kaleido-1.2.0-py3-none-any.whl (68 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 69.0/69.0 kB 3.3 MB/s eta 0:00:00
Downloading choreographer-1.2.1-py3-none-any.whl (49 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 49.3/49.3 kB 2.8 MB/s eta 0:00:00
Downloading logistro-2.0.1-py3-none-any.whl (8.6 kB)
Downloading pyarrow-22.0.0-cp311-cp311-manylinux_2_28_x86_64.whl (47.7 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 47.7/47.7 MB 37.7 MB/s eta 0:00:00:00:0100:01
Downloading pytest_timeout-2.4.0-py3-none-any.whl (14 kB)
Downloading nvidia_cublas_cu12-12.4.5.8-py3-none-manylinux2014_x86_64.whl (363.4 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 363.4/363.4 MB 4.7 MB/s eta 0:00:00:00:0100:01
Downloading nvidia_cuda_cupti_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl (13.8 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 13.8/13.8 MB 94.8 MB/s eta 0:00:00:00:0100:01
Downloading nvidia_cuda_nvrtc_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl (24.6 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 24.6/24.6 MB 74.7 MB/s eta 0:00:00:00:0100:01
Downloading nvidia_cuda_runtime_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl (883 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 883.7/883.7 kB 41.2 MB/s eta 0:00:00
Downloading nvidia_cudnn_cu12-9.1.0.70-py3-none-manylinux2014_x86_64.whl (664.8 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 664.8/664.8 MB 2.3 MB/s eta 0:00:00:00:0100:01
Downloading nvidia_cufft_cu12-11.2.1.3-py3-none-manylinux2014_x86_64.whl (211.5 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 211.5/211.5 MB 3.6 MB/s eta 0:00:00:00:0100:01
Downloading nvidia_curand_cu12-10.3.5.147-py3-none-manylinux2014_x86_64.whl (56.3 MB)
```

```
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 56.3/56.3 MB 3.4 MB/s eta 0:00:000:00:0100:01
Downloading nvidia_cusolver_cu12-11.6.1.9-py3-none-manylinux2014_x86_64.whl (127.9 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 127.9/127.9 MB 9.8 MB/s eta 0:00:00:00:0100:01
Downloading nvidia_cusparse_cu12-12.3.1.170-py3-none-manylinux2014_x86_64.whl (207.5 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 207.5/207.5 MB 8.2 MB/s eta 0:00:00:00:0100:01
Downloading nvidia_nvjitlink_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl (21.1 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 21.1/21.1 MB 81.7 MB/s eta 0:00:00:00:0100:01
Building wheels for collected packages: tsfm_public
  Building wheel for tsfm_public (pyproject.toml) ... done
  Created wheel for tsfm_public: filename=tsfm_public-0.2.18-py3-none-any.whl size=2338110 sha256=524371d936346e4777158
6b46f1169b19fb47401cddb09e40aa5dc6913d6a992
  Stored in directory: /tmp/pip-ephem-wheel-cache-2wym_8ub/wheels/47/e5/96/874ab391fafe9e1f7ff69603c654a39ad27a2e4ae2d0
9eebf6
Successfully built tsfm_public
Installing collected packages: urllib3, pyarrow, nvidia-nvjitlink-cu12, nvidia-curand-cu12, nvidia-cufft-cu12, nvidia-c
uda-runtime-cu12, nvidia-cuda-nvrtc-cu12, nvidia-cuda-cupti-cu12, nvidia-cublas-cu12, logistro, pytest-timeout, nvidia-
cusparse-cu12, nvidia-cudnn-cu12, choreographer, nvidia-cusolver-cu12, kaleido, jupyter, tsfm_public
  Attempting uninstall: urllib3
    Found existing installation: urllib3 2.5.0
    Uninstalling urllib3-2.5.0:
      Successfully uninstalled urllib3-2.5.0
  Attempting uninstall: pyarrow
    Found existing installation: pyarrow 19.0.1
    Uninstalling pyarrow-19.0.1:
      Successfully uninstalled pyarrow-19.0.1
  Attempting uninstall: nvidia-nvjitlink-cu12
    Found existing installation: nvidia-nvjitlink-cu12 12.5.82
    Uninstalling nvidia-nvjitlink-cu12-12.5.82:
      Successfully uninstalled nvidia-nvjitlink-cu12-12.5.82
  Attempting uninstall: nvidia-curand-cu12
    Found existing installation: nvidia-curand-cu12 10.3.6.82
    Uninstalling nvidia-curand-cu12-10.3.6.82:
      Successfully uninstalled nvidia-curand-cu12-10.3.6.82
  Attempting uninstall: nvidia-cufft-cu12
    Found existing installation: nvidia-cufft-cu12 11.2.3.61
    Uninstalling nvidia-cufft-cu12-11.2.3.61:
      Successfully uninstalled nvidia-cufft-cu12-11.2.3.61
  Attempting uninstall: nvidia-cuda-runtime-cu12
    Found existing installation: nvidia-cuda-runtime-cu12 12.5.82
    Uninstalling nvidia-cuda-runtime-cu12-12.5.82:
      Successfully uninstalled nvidia-cuda-runtime-cu12-12.5.82
  Attempting uninstall: nvidia-cuda-nvrtc-cu12
    Found existing installation: nvidia-cuda-nvrtc-cu12 12.5.82
    Uninstalling nvidia-cuda-nvrtc-cu12-12.5.82:
      Successfully uninstalled nvidia-cuda-nvrtc-cu12-12.5.82
  Attempting uninstall: nvidia-cuda-cupti-cu12
    Found existing installation: nvidia-cuda-cupti-cu12 12.5.82
    Uninstalling nvidia-cuda-cupti-cu12-12.5.82:
```

```
    Successfully uninstalled nvidia-cuda-cupti-cu12-12.5.82
  Attempting uninstall: nvidia-cublas-cu12
    Found existing installation: nvidia-cublas-cu12 12.5.3.2
    Uninstalling nvidia-cublas-cu12-12.5.3.2:
      Successfully uninstalled nvidia-cublas-cu12-12.5.3.2
  Attempting uninstall: nvidia-cusparse-cu12
    Found existing installation: nvidia-cusparse-cu12 12.5.1.3
    Uninstalling nvidia-cusparse-cu12-12.5.1.3:
      Successfully uninstalled nvidia-cusparse-cu12-12.5.1.3
  Attempting uninstall: nvidia-cudnn-cu12
    Found existing installation: nvidia-cudnn-cu12 9.3.0.75
    Uninstalling nvidia-cudnn-cu12-9.3.0.75:
      Successfully uninstalled nvidia-cudnn-cu12-9.3.0.75
  Attempting uninstall: nvidia-cusolver-cu12
    Found existing installation: nvidia-cusolver-cu12 11.6.3.83
    Uninstalling nvidia-cusolver-cu12-11.6.3.83:
      Successfully uninstalled nvidia-cusolver-cu12-11.6.3.83
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behavio
ur is the source of the following dependency conflicts.
bigframes 2.12.0 requires google-cloud-bigquery-storage<3.0.0,>=2.30.0, which is not installed.
pylibcudf-cu12 25.2.2 requires pyarrow<20.0.0a0,>=14.0.0; platform_machine == "x86_64", but you have pyarrow 22.0.0 whi
ch is incompatible.
cudf-cu12 25.2.2 requires pyarrow<20.0.0a0,>=14.0.0; platform_machine == "x86_64", but you have pyarrow 22.0.0 which is
incompatible.
ray 2.51.1 requires click!=8.3.0,>=7.0, but you have click 8.3.0 which is incompatible.
google-colab 1.0.0 requires notebook==6.5.7, but you have notebook 6.5.4 which is incompatible.
google-colab 1.0.0 requires pandas==2.2.2, but you have pandas 2.2.3 which is incompatible.
google-colab 1.0.0 requires requests==2.32.3, but you have requests 2.32.5 which is incompatible.
google-colab 1.0.0 requires tornado==6.4.2, but you have tornado 6.5.2 which is incompatible.
bigframes 2.12.0 requires rich<14,>=12.4.4, but you have rich 14.2.0 which is incompatible.
libcugraph-cu12 25.6.0 requires libraft-cu12==25.6.*, but you have libraft-cu12 25.2.0 which is incompatible.
google-ai-generativelanguage 0.6.15 requires protobuf!=4.21.0,!=4.21.1,!=4.21.2,!=4.21.3,!=4.21.4,!=4.21.5,<6.0.0dev,>=
3.20.2, but you have protobuf 6.33.0 which is incompatible.
cudf-polars-cu12 25.6.0 requires pylibcudf-cu12==25.6.*, but you have pylibcudf-cu12 25.2.2 which is incompatible.
pydrive2 1.21.3 requires cryptography<44, but you have cryptography 46.0.3 which is incompatible.
pydrive2 1.21.3 requires pyOpenSSL<=24.2.1,>=19.1.0, but you have pyopenssl 25.3.0 which is incompatible.
tensorflow 2.18.0 requires protobuf!=4.21.0,!=4.21.1,!=4.21.2,!=4.21.3,!=4.21.4,!=4.21.5,<6.0.0dev,>=3.20.3, but you ha
ve protobuf 6.33.0 which is incompatible.
pylibcugraph-cu12 25.6.0 requires pylibraft-cu12==25.6.*, but you have pylibraft-cu12 25.2.0 which is incompatible.
pylibcugraph-cu12 25.6.0 requires rmm-cu12==25.6.*, but you have rmm-cu12 25.2.0 which is incompatible.
jupyter-kernel-gateway 2.5.2 requires jupyter-client<8.0,>=5.2.0, but you have jupyter-client 8.6.3 which is incompatib
le.
gcsfs 2025.3.0 requires fsspec==2025.3.0, but you have fsspec 2025.10.0 which is incompatible.
Successfully installed choreographer-1.2.1 jupyter-1.1.1 kaleido-1.2.0 logistro-2.0.1 nvidia-cublas-cu12-12.4.5.8 nvidi
a-cuda-cupti-cu12-12.4.127 nvidia-cuda-nvrtc-cu12-12.4.127 nvidia-cuda-runtime-cu12-12.4.127 nvidia-cudnn-cu12-9.1.0.70
nvidia-cufft-cu12-11.2.1.3 nvidia-curand-cu12-10.3.5.147 nvidia-cusolver-cu12-11.6.1.9 nvidia-cusparse-cu12-12.3.1.170
nvidia-nvjitlink-cu12-12.4.127 pyarrow-22.0.0 pytest-timeout-2.4.0 tsfm_public-0.2.18 urllib3-1.26.20
```

4.1 Environment Setup - Imports

In [2]:

```
import math
import os
import tempfile

import pandas as pd
from torch.optim import AdamW
from torch.optim.lr_scheduler import OneCycleLR
from transformers import EarlyStoppingCallback, Trainer, TrainingArguments, set_seed

from tsfm_public import (
    TimeSeriesPreprocessor,
    TinyTimeMixerForPrediction,
    TrackingCallback,
    count_parameters,
    get_datasets,
)
from tsfm_public.toolkit.visualization import plot_predictions
```

```
2025-11-16 05:44:09.883973: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register cuFFT fac
tory: Attempting to register factory for plugin cuFFT when one has already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
E0000 00:00:1763271850.289908      48 cuda_dnn.cc:8310] Unable to register cuDNN factory: Attempting to register factor
y for plugin cuDNN when one has already been registered
E0000 00:00:1763271850.389705      48 cuda_blas.cc:1418] Unable to register cuBLAS factory: Attempting to register fact
ory for plugin cuBLAS when one has already been registered

---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'

---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'

---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'

---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'

---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
AttributeError: 'MessageFactory' object has no attribute 'GetPrototype'
```

4.1 Environment Setup - Important arguments

In [29]:
```
# Set seed for reproducibility
SEED = 42
set_seed(SEED)

# DATA ROOT PATH
# Make sure to download the target data (here etth1) on the `DATA_ROOT_PATH` folder.
# ETT is available at: https://github.com/zhouhaoyi/ETDataset/tree/main
target_dataset = "ettm2"
DATA_ROOT_PATH = "https://raw.githubusercontent.com/zhouhaoyi/ETDataset/main/ETT-small/ETTh1.csv"

# Results dir
OUT_DIR = "ttm_finetuned_models/"

# Forecasting parameters
context_length = 1024
forecast_length = 96
fewshot_fraction = 0.05

# ----- TTM model path -----
TTM_MODEL_PATH = "ibm-granite/granite-timeseries-ttm-r2"


# ----- TTM model branch -----
# For R2 models
TTM_MODEL_REVISION="1024-96-r2"
```

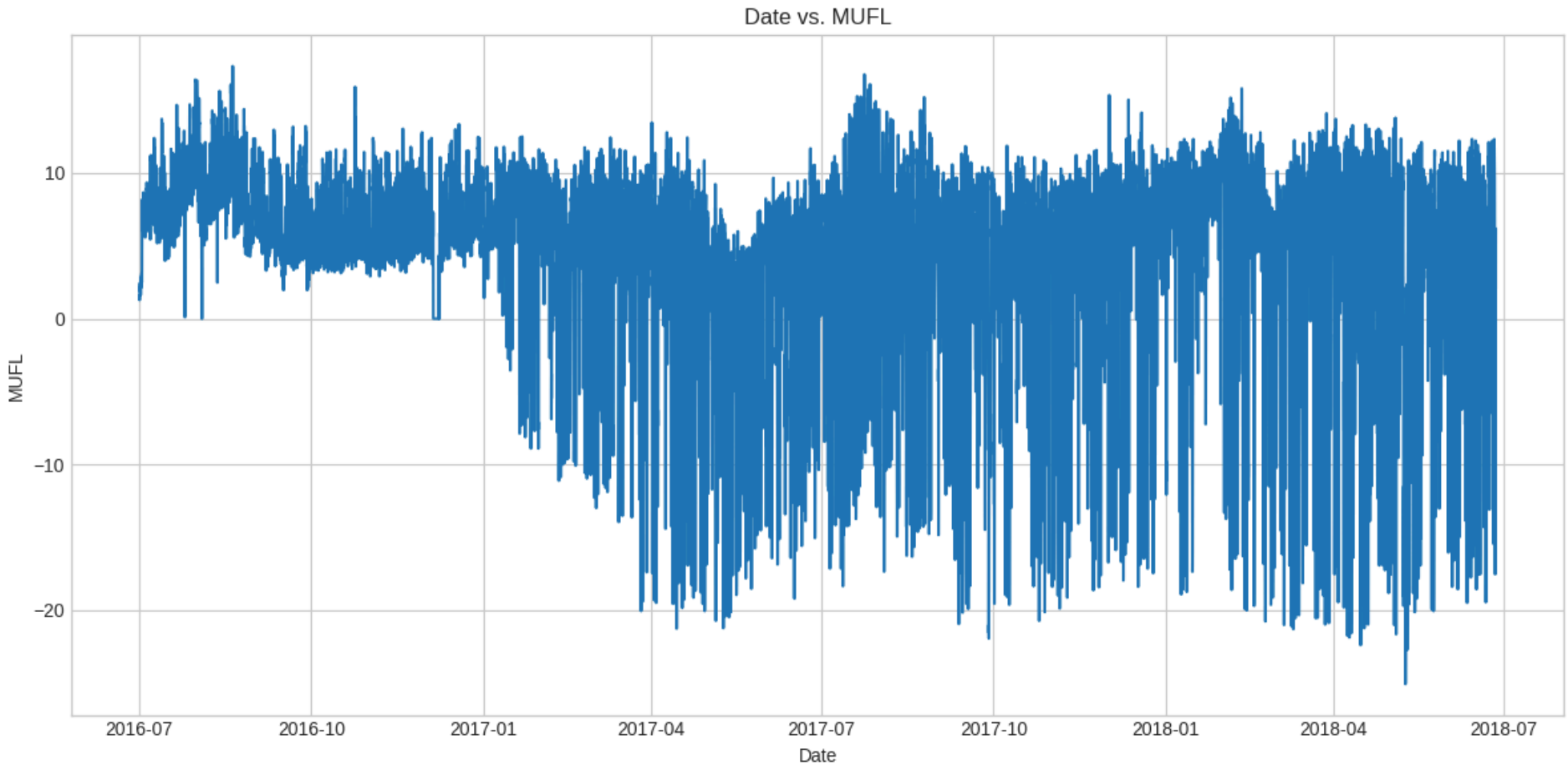
4.2 Data processing pipeline

Note: Here we use the TimeSeriesPreprocessor (TSP) module for data preparation. For standard datasets, TSP can quickly prepare the dataloaders using YAML files defined here. Refer to the TTM Getting Started for example usage. Similar YAML file can be written for any new dataset as well.

In [30]:
```
# Load the data file and see the columns
df = pd.read_csv(DATA_ROOT_PATH)

# View the first 20 rows
print(df.head(20))
```

	date	HUFL	HULL	MUFL	MULL	LUFL	LULL	OT
0	2016-07-01 00:00:00	5.827	2.009	1.599	0.462	4.203	1.340	30.531000
1	2016-07-01 01:00:00	5.693	2.076	1.492	0.426	4.142	1.371	27.787001
2	2016-07-01 02:00:00	5.157	1.741	1.279	0.355	3.777	1.218	27.787001
3	2016-07-01 03:00:00	5.090	1.942	1.279	0.391	3.807	1.279	25.044001
4	2016-07-01 04:00:00	5.358	1.942	1.492	0.462	3.868	1.279	21.948000
5	2016-07-01 05:00:00	5.626	2.143	1.528	0.533	4.051	1.371	21.174000
6	2016-07-01 06:00:00	7.167	2.947	2.132	0.782	5.026	1.858	22.792000
7	2016-07-01 07:00:00	7.435	3.282	2.310	1.031	5.087	2.224	23.143999
8	2016-07-01 08:00:00	5.559	3.014	2.452	1.173	2.955	1.432	21.667000
9	2016-07-01 09:00:00	4.555	2.545	1.919	0.817	2.680	1.371	17.445999
10	2016-07-01 10:00:00	4.957	2.545	1.990	0.853	2.955	1.492	19.979000
11	2016-07-01 11:00:00	5.760	2.545	2.203	0.853	3.442	1.492	20.118999
12	2016-07-01 12:00:00	4.689	2.545	1.812	0.853	2.833	1.523	19.205000
13	2016-07-01 13:00:00	4.689	2.679	1.777	1.244	3.107	1.614	18.572001
14	2016-07-01 14:00:00	5.090	2.947	2.452	1.350	2.559	1.432	19.556000
15	2016-07-01 15:00:00	5.090	3.148	2.487	1.350	2.589	1.523	17.305000
16	2016-07-01 16:00:00	4.220	2.411	1.706	0.782	2.619	1.492	19.486000
17	2016-07-01 17:00:00	4.756	2.344	1.635	0.711	3.076	1.492	19.134001
18	2016-07-01 18:00:00	5.626	2.880	2.523	1.208	3.076	1.492	20.681999
19	2016-07-01 19:00:00	5.492	3.014	2.452	1.208	3.015	1.553	18.712000

In [31]:
```
# Plot date vs "MUFL"
import matplotlib.pyplot as plt
import pandas as pd

df['date'] = pd.to_datetime(df['date'])

plt.figure(figsize=(12, 6))
plt.plot(df['date'], df['MUFL'])
plt.xlabel('Date')
plt.ylabel('MUFL')
plt.title('Date vs. MUFL')
plt.tight_layout()
plt.show()
```



In [32]:

```
dataset_path = DATA_ROOT_PATH
timestamp_column = "date"
id_columns = []
target_columns = ["HUFL", "HULL", "MUFL", "MULL", "LUFL", "LULL", "OT"]
split_config = {
    "train": [0, 12 * 30 * 24],
    "valid": [12 * 30 * 24, 12 * 30 * 24 + 4 * 30 * 24],
    "test": [
        12 * 30 * 24 + 4 * 30 * 24,
        12 * 30 * 24 + 8 * 30 * 24,
    ],
}

data = pd.read_csv(
    dataset_path,
    parse_dates=[timestamp_column],
)

column_specifiers = {
```

```
    "timestamp_column": timestamp_column,
    "id_columns": id_columns,
    "target_columns": target_columns,
    "control_columns": [],
}

tsp = TimeSeriesPreprocessor(
    **column_specifiers,
    context_length=context_length,
    prediction_length=forecast_length,
    scaling=True,
    encode_categorical=False,
    scaler_type="standard",
)

# Obtain the train_dataset, valid_dataset, test_dataset using tsp get_dataset() by specifying the right arguments
train_dataset, valid_dataset, test_dataset = get_datasets(
    tsp, data, split_config
)

# Print the length of train_dataset, valid_dataset, test_dataset
print("Length of train_dataset: " + str(len(train_dataset)))
print("Length of valid_dataset: " + str(len(valid_dataset)))
print("Length of test_dataset: " + str(len(test_dataset)))
```

```
Length of train_dataset: 7521
Length of valid_dataset: 2785
Length of test_dataset: 2785
```

In [33]:
```
train_dataset[3]
```

Out[33]:

```
{'past_values': tensor([[-0.4899, -0.0378, -0.6887,  ...,  1.0017,  0.7784,  0.8626],
        [-0.4438, -0.0378, -0.6501,  ...,  1.0613,  0.7784,  0.5252],
        [-0.3977,  0.0584, -0.6436,  ...,  1.2401,  0.9243,  0.4409],
        ...,
        [ 0.8122,  1.3726,  0.6442,  ...,  1.2108,  0.9243,  1.9895],
        [ 0.8468,  1.4683,  0.6056,  ...,  1.3886,  1.0211,  1.8822],
        [ 0.6509,  1.4683,  0.4317,  ...,  1.3886,  1.1163,  1.8132]]),
 'future_values': tensor([[0.7661, 1.3726, 0.5411, 1.1990, 1.5078, 0.7784, 1.6139],
        [1.7342, 0.9880, 1.1914, 0.5351, 3.5312, 0.9719, 1.5755],
        [1.6996, 0.7315, 1.1079, 0.0554, 3.4130, 1.0687, 1.5525],
        [2.3218, 2.3659, 1.9126, 1.9185, 2.7877, 0.8275, 1.5678],
        [2.0567, 1.8209, 1.7388, 1.2914, 2.2220, 0.8275, 1.5372],
        [2.2412, 1.8851, 1.7388, 1.2359, 3.7100, 1.4067, 1.5678],
        [2.2066, 2.1736, 1.6551, 1.1990, 3.5019, 1.3591, 1.4835],
        [1.7687, 1.2765, 1.3139, 0.7562, 3.0261, 1.2131, 1.4835],
        [1.8377, 1.9492, 1.4554, 1.3283, 2.9362, 1.3099, 1.4682],
        [1.8838, 1.6607, 1.4040, 1.0148, 2.7877, 1.6971, 1.3992],
        [1.4575, 1.4363, 1.1336, 0.9592, 2.5200, 1.3099, 1.3532],
        [1.2962, 1.0200, 0.8374, 0.4427, 2.8170, 1.4559, 1.3149],
        [1.2846, 1.6927, 0.8310, 0.8668, 2.7281, 1.8414, 1.4758],
        [1.0773, 1.5324, 0.9854, 1.1253, 0.4663, 1.2131, 1.4912],
        [0.7315, 0.9880, 0.5991, 1.7342, 0.8835, 1.1655, 1.5372],
        [1.0197, 2.0133, 0.7343, 1.6787, 1.4189, 0.9243, 1.6905],
        [1.1695, 2.0133, 0.7602, 1.4944, 2.0432, 0.9719, 1.6675],
        [1.1695, 1.8209, 0.8439, 1.5863, 1.9250, 0.7784, 1.6521],
        [1.2385, 1.8530, 0.9276, 1.6050, 2.4301, 1.0211, 1.7671],
        [1.2616, 1.9812, 0.8760, 1.4389, 2.3412, 1.2131, 1.9358],
        [1.2041, 1.6607, 0.8116, 1.2728, 1.8654, 0.9243, 2.0355],
        [0.9044, 1.5645, 0.6699, 1.2914, 1.6270, 0.6816, 2.1275],
        [0.8814, 1.5966, 0.6636, 1.3283, 1.5977, 0.8751, 2.1427],
        [0.8353, 1.4683, 0.5799, 1.3465, 1.7755, 1.1163, 1.9971],
        [1.0197, 1.7568, 0.7859, 1.4570, 2.1331, 0.8751, 1.8132],
        [2.0567, 2.3659, 1.3654, 1.7342, 4.3050, 1.3591, 1.7365],
        [1.8954, 2.5903, 1.2494, 1.8261, 4.0676, 1.2623, 1.6905],
        [1.9645, 1.5324, 1.4234, 0.7749, 3.1443, 1.1163, 1.6751],
        [1.7457, 1.0521, 1.3460, 0.8855, 2.7574, 1.0687, 1.6521],
        [1.7342, 1.8209, 1.5328, 1.1435, 1.8947, 0.6340, 1.6139],
        [1.9299, 1.8851, 1.7839, 1.5313, 1.5078, 0.6340, 1.5908],
        [1.6535, 1.8209, 1.4942, 1.2546, 1.3290, 0.6816, 1.4988],
        [1.3884, 2.5903, 1.2623, 2.5088, 0.9421, 0.5848, 1.4452],
        [1.2041, 2.0774, 1.1914, 1.9185, 0.6744, 0.9243, 1.4222],
        [1.0543, 1.6927, 1.0177, 1.4757, 0.6744, 0.5848, 1.3608],
        [0.9966, 1.9812, 0.9211, 1.7711, 0.4966, 0.4896, 1.2842],
        [1.0428, 2.4300, 0.9082, 1.8630, 0.3471, 0.8275, 1.4758],
        [0.4666, 2.1415, 0.4574, 1.6050, 0.7340, 1.2131, 1.5065],
        [0.6970, 2.0133, 0.5411, 1.6232, 0.8532, 1.1655, 1.5295],
        [0.9966, 1.1803, 0.7408, 0.5901, 1.5977, 1.2131, 1.6215],
```

```
[1.1234, 1.0200, 0.8374, 0.5533, 2.1331, 1.1163, 1.6829],
[1.1926, 0.9559, 0.8953, 0.4982, 2.4897, 1.4559, 1.6675],
[1.4691, 1.0200, 0.9725, 0.3872, 2.8473, 1.1655, 1.7671],
[1.3999, 1.2124, 0.9468, 0.4245, 2.3412, 0.9243, 1.9204],
[1.2616, 1.1803, 0.8439, 0.4245, 1.9543, 0.9243, 2.0431],
[1.0543, 1.5966, 0.7602, 1.2359, 1.9250, 1.3591, 2.1658],
[0.9736, 1.8530, 0.6828, 1.6232, 1.8654, 1.0211, 2.0967],
[0.7892, 1.4683, 0.5734, 1.2914, 1.4482, 0.6340, 1.9971],
[0.9851, 1.6607, 0.7085, 1.4202, 1.8058, 0.8751, 1.7365],
[1.9184, 1.3085, 1.2817, 0.5720, 4.0373, 1.1655, 1.6905],
[1.7918, 0.9880, 1.2045, 0.4245, 3.7999, 1.1163, 1.7059],
[1.6074, 1.4363, 1.1657, 0.8300, 3.2342, 1.0687, 1.6445],
[1.4575, 1.2444, 1.0948, 0.7931, 2.5200, 0.9243, 1.6061],
[1.9991, 1.9812, 1.3268, 1.0516, 4.3949, 1.7938, 1.6061],
[1.6535, 1.8530, 1.1079, 1.0698, 3.7999, 1.5527, 1.5295],
[1.4921, 2.3659, 0.9405, 1.8998, 3.3534, 1.5527, 1.5142],
[1.3307, 2.4300, 0.8760, 1.8448, 2.8766, 1.4067, 1.4452],
[1.1811, 2.0453, 0.5991, 1.5681, 3.0554, 1.5527, 1.4068],
[1.3999, 2.3659, 0.8116, 1.8448, 2.9958, 1.5527, 1.3838],
[1.3768, 2.4621, 0.8180, 1.5681, 3.0554, 1.6003, 1.3149],
[1.2385, 2.7506, 0.7022, 1.8448, 2.9958, 2.1318, 1.4835],
[0.5703, 2.1736, 0.5154, 1.6232, 0.7047, 1.0687, 1.4912],
[0.6854, 2.2377, 0.6248, 1.6600, 0.6148, 0.9243, 1.5525],
[1.0889, 2.2697, 0.7473, 1.5863, 1.5674, 0.9243, 1.6905],
[1.1580, 2.1736, 0.7922, 1.6787, 1.8947, 0.8275, 1.6829],
[1.2385, 1.7248, 0.8631, 1.6232, 2.1624, 1.0211, 1.6675],
[1.4345, 1.1803, 1.1014, 0.6457, 2.4301, 1.0211, 1.7671],
[1.4345, 0.9559, 1.1142, 0.8118, 2.1624, 0.9719, 1.9895],
[1.2846, 0.9880, 0.9017, 0.3872, 2.1331, 1.1163, 2.0431],
[1.0658, 1.1803, 0.8310, 0.7194, 1.6866, 1.4067, 2.0737],
[0.9736, 1.8530, 0.7408, 1.6050, 1.6563, 0.9719, 2.1581],
[0.8238, 1.6286, 0.5734, 1.3652, 1.6270, 0.9719, 1.9434],
[1.1004, 1.7248, 0.7473, 1.3652, 2.1624, 0.9243, 1.7441],
[1.9530, 1.1803, 1.3331, 0.7194, 4.1272, 1.1163, 1.6061],
[1.8838, 1.0200, 1.3396, 0.6638, 3.5312, 0.9243, 1.6061],
[2.1144, 1.9492, 1.6745, 1.4570, 2.7574, 0.8751, 1.6675],
[1.8954, 1.8209, 1.6551, 1.2177, 2.4604, 0.9719, 1.6751],
[2.0913, 1.5004, 1.7002, 1.2914, 2.1028, 0.9243, 1.5908],
[1.6189, 1.6286, 1.5006, 1.2359, 1.4189, 0.5848, 1.5295],
[1.6074, 2.2377, 1.4813, 2.0846, 0.9421, 0.5372, 1.5218],
[1.2846, 1.3085, 1.2623, 0.8668, 0.8835, 0.4896, 1.4758],
[1.4575, 1.5004, 1.3589, 1.2728, 0.5259, 0.4404, 1.4758],
[1.1004, 1.7248, 1.1014, 1.7524, 0.4370, 0.4404, 1.3915],
[0.5934, 1.6607, 0.5605, 1.2914, 0.4966, 0.4896, 1.3838],
[0.4320, 1.4047, 0.4574, 1.1622, 0.2582, 0.8275, 1.4528],
[0.1439, 1.1162, 0.0776, 1.0885, 0.5259, 0.7784, 1.4988],
[0.4090, 1.3726, 0.2642, 1.0516, 0.8239, 0.8275, 1.4988],
[0.7085, 1.3406, 0.4188, 0.9961, 1.3290, 1.2131, 1.5908],
```

```
        [0.8699, 1.7248, 0.4639, 1.1253, 1.7755, 1.1163, 1.6061],
        [0.8814, 1.5966, 0.5348, 0.9224, 2.0735, 1.0211, 1.6599],
        [1.1580, 1.5324, 0.6056, 1.0329, 2.8766, 1.2131, 1.7671],
        [0.9736, 1.3726, 0.5670, 0.9961, 2.5200, 1.2131, 1.9512],
        [0.8929, 1.5004, 0.4445, 0.8486, 2.0735, 0.8751, 1.9971],
        [0.6739, 1.0841, 0.4188, 0.7562, 1.6563, 1.0687, 2.0967],
        [0.5703, 0.8597, 0.3030, 0.6270, 1.6563, 0.8751, 2.0967],
        [0.5127, 0.8277, 0.2514, 0.6270, 1.6270, 1.2131, 1.8974]]),
'past_observed_mask': tensor([[True, True, True,  ..., True, True, True],
        [True, True, True,  ..., True, True, True],
        [True, True, True,  ..., True, True, True],
        ...,
        [True, True, True,  ..., True, True, True],
        [True, True, True,  ..., True, True, True],
        [True, True, True,  ..., True, True, True]]),
'future_observed_mask': tensor([[True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
        [True, True, True, True, True, True, True],
```

```
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
[True, True, True, True, True, True, True],
```

```
          [True, True, True, True, True, True, True],
          [True, True, True, True, True, True, True],
          [True, True, True, True, True, True, True],
          [True, True, True, True, True, True, True],
          [True, True, True, True, True, True, True],
          [True, True, True, True, True, True, True],
          [True, True, True, True, True, True, True],
          [True, True, True, True, True, True, True],
          [True, True, True, True, True, True, True],
          [True, True, True, True, True, True, True],
          [True, True, True, True, True, True, True],
          [True, True, True, True, True, True, True],
          [True, True, True, True, True, True, True],
          [True, True, True, True, True, True, True]]),
 'timestamp': Timestamp('2016-08-12 18:00:00'),
 'id': (0,)}
```

4.3 Zero-shot evaluation method

In [34]:
```
zeroshot_model = TinyTimeMixerForPrediction.from_pretrained(TTM_MODEL_PATH, revision=TTM_MODEL_REVISION)
zeroshot_model
```

```
Out[34]: TinyTimeMixerForPrediction(
           (backbone): TinyTimeMixerModel(
             (encoder): TinyTimeMixerEncoder(
               (patcher): Linear(in_features=128, out_features=384, bias=True)
               (mlp_mixer_encoder): TinyTimeMixerBlock(
                 (mixers): ModuleList(
                   (0): TinyTimeMixerAdaptivePatchingBlock(
                     (mixer_layers): ModuleList(
                       (0-1): 2 x TinyTimeMixerLayer(
                         (patch_mixer): PatchMixerBlock(
                           (norm): TinyTimeMixerNormLayer(
                             (norm): LayerNorm((96,), eps=1e-05, elementwise_affine=True)
                           )
                           (mlp): TinyTimeMixerMLP(
                             (fc1): Linear(in_features=32, out_features=64, bias=True)
                             (dropout1): Dropout(p=0.4, inplace=False)
                             (fc2): Linear(in_features=64, out_features=32, bias=True)
                             (dropout2): Dropout(p=0.4, inplace=False)
                           )
                           (gating_block): TinyTimeMixerGatedAttention(
                             (attn_layer): Linear(in_features=32, out_features=32, bias=True)
                             (attn_softmax): Softmax(dim=-1)
                           )
                         )
                         (feature_mixer): FeatureMixerBlock(
                           (norm): TinyTimeMixerNormLayer(
                             (norm): LayerNorm((96,), eps=1e-05, elementwise_affine=True)
                           )
                           (mlp): TinyTimeMixerMLP(
                             (fc1): Linear(in_features=96, out_features=192, bias=True)
                             (dropout1): Dropout(p=0.4, inplace=False)
                             (fc2): Linear(in_features=192, out_features=96, bias=True)
                             (dropout2): Dropout(p=0.4, inplace=False)
                           )
                           (gating_block): TinyTimeMixerGatedAttention(
                             (attn_layer): Linear(in_features=96, out_features=96, bias=True)
                             (attn_softmax): Softmax(dim=-1)
                           )
                         )
                       )
                     )
                   )
                   (1): TinyTimeMixerAdaptivePatchingBlock(
                     (mixer_layers): ModuleList(
                       (0-1): 2 x TinyTimeMixerLayer(
                         (patch_mixer): PatchMixerBlock(
                           (norm): TinyTimeMixerNormLayer(
```

```
            (norm): LayerNorm((192,), eps=1e-05, elementwise_affine=True)
          )
          (mlp): TinyTimeMixerMLP(
            (fc1): Linear(in_features=16, out_features=32, bias=True)
            (dropout1): Dropout(p=0.4, inplace=False)
            (fc2): Linear(in_features=32, out_features=16, bias=True)
            (dropout2): Dropout(p=0.4, inplace=False)
          )
          (gating_block): TinyTimeMixerGatedAttention(
            (attn_layer): Linear(in_features=16, out_features=16, bias=True)
            (attn_softmax): Softmax(dim=-1)
          )
        )
        (feature_mixer): FeatureMixerBlock(
          (norm): TinyTimeMixerNormLayer(
            (norm): LayerNorm((192,), eps=1e-05, elementwise_affine=True)
          )
          (mlp): TinyTimeMixerMLP(
            (fc1): Linear(in_features=192, out_features=384, bias=True)
            (dropout1): Dropout(p=0.4, inplace=False)
            (fc2): Linear(in_features=384, out_features=192, bias=True)
            (dropout2): Dropout(p=0.4, inplace=False)
          )
          (gating_block): TinyTimeMixerGatedAttention(
            (attn_layer): Linear(in_features=192, out_features=192, bias=True)
            (attn_softmax): Softmax(dim=-1)
          )
        )
      )
    )
  )
  (2): TinyTimeMixerAdaptivePatchingBlock(
    (mixer_layers): ModuleList(
      (0-1): 2 x TinyTimeMixerLayer(
        (patch_mixer): PatchMixerBlock(
          (norm): TinyTimeMixerNormLayer(
            (norm): LayerNorm((384,), eps=1e-05, elementwise_affine=True)
          )
          (mlp): TinyTimeMixerMLP(
            (fc1): Linear(in_features=8, out_features=16, bias=True)
            (dropout1): Dropout(p=0.4, inplace=False)
            (fc2): Linear(in_features=16, out_features=8, bias=True)
            (dropout2): Dropout(p=0.4, inplace=False)
          )
          (gating_block): TinyTimeMixerGatedAttention(
            (attn_layer): Linear(in_features=8, out_features=8, bias=True)
            (attn_softmax): Softmax(dim=-1)
          )
```

```
              )
              (feature_mixer): FeatureMixerBlock(
                (norm): TinyTimeMixerNormLayer(
                  (norm): LayerNorm((384,), eps=1e-05, elementwise_affine=True)
                )
                (mlp): TinyTimeMixerMLP(
                  (fc1): Linear(in_features=384, out_features=768, bias=True)
                  (dropout1): Dropout(p=0.4, inplace=False)
                  (fc2): Linear(in_features=768, out_features=384, bias=True)
                  (dropout2): Dropout(p=0.4, inplace=False)
                )
                (gating_block): TinyTimeMixerGatedAttention(
                  (attn_layer): Linear(in_features=384, out_features=384, bias=True)
                  (attn_softmax): Softmax(dim=-1)
                )
              )
            )
          )
        )
      )
    )
  )
  (patching): TinyTimeMixerPatchify()
  (scaler): TinyTimeMixerStdScaler()
)
(decoder): TinyTimeMixerDecoder(
  (adapter): Linear(in_features=384, out_features=256, bias=True)
  (decoder_block): TinyTimeMixerBlock(
    (mixers): ModuleList(
      (0-1): 2 x TinyTimeMixerLayer(
        (patch_mixer): PatchMixerBlock(
          (norm): TinyTimeMixerNormLayer(
            (norm): LayerNorm((256,), eps=1e-05, elementwise_affine=True)
          )
          (mlp): TinyTimeMixerMLP(
            (fc1): Linear(in_features=8, out_features=16, bias=True)
            (dropout1): Dropout(p=0.4, inplace=False)
            (fc2): Linear(in_features=16, out_features=8, bias=True)
            (dropout2): Dropout(p=0.4, inplace=False)
          )
          (gating_block): TinyTimeMixerGatedAttention(
            (attn_layer): Linear(in_features=8, out_features=8, bias=True)
            (attn_softmax): Softmax(dim=-1)
          )
        )
        (feature_mixer): FeatureMixerBlock(
          (norm): TinyTimeMixerNormLayer(
            (norm): LayerNorm((256,), eps=1e-05, elementwise_affine=True)
```

```
            )
            (mlp): TinyTimeMixerMLP(
              (fc1): Linear(in_features=256, out_features=512, bias=True)
              (dropout1): Dropout(p=0.4, inplace=False)
              (fc2): Linear(in_features=512, out_features=256, bias=True)
              (dropout2): Dropout(p=0.4, inplace=False)
            )
            (gating_block): TinyTimeMixerGatedAttention(
              (attn_layer): Linear(in_features=256, out_features=256, bias=True)
              (attn_softmax): Softmax(dim=-1)
            )
          )
        )
      )
    )
    (head): TinyTimeMixerForPredictionHead(
      (dropout_layer): Dropout(p=0.4, inplace=False)
      (base_forecast_block): Linear(in_features=2048, out_features=96, bias=True)
      (flatten): Flatten(start_dim=-2, end_dim=-1)
    )
  )
```

In [36]:

```python
#Perform zero-shot evaluation using the pre-trained model on the test dataset and calculate the evaluation error.
import wandb

# Login to wandb and initialize it
wandb.login(key = "43d56b51a7a89074cdecf6fd4ef49d1d5256762e")
wandb.init(project="hw4_problem4", name="zero_shot_eval_1")

temp_dir = tempfile.mkdtemp()
# zeroshot_trainer
zeroshot_trainer = Trainer(
    model=zeroshot_model,
    args=TrainingArguments(
        output_dir=temp_dir,
        per_device_eval_batch_size=64,
        seed=SEED,
    ),
)

zeroshot_results = zeroshot_trainer.evaluate(test_dataset)
print("Evaluation Loss: " + str(zeroshot_results["eval_loss"]))
```

```
wandb: WARNING If you're specifying your api key in code, ensure this code is not shared publicly.
wandb: WARNING Consider setting the WANDB_API_KEY environment variable, or running `wandb login` from the command line.
wandb: Appending key for api.wandb.ai to your netrc file: /root/.netrc
```

Tracking run with wandb version 0.21.0

Run data is saved locally in `/kaggle/working/wandb/run-20251116_064311-zy7m9e4j`

Syncing run **zero_shot_eval_1** to Weights & Biases (docs)

View project at https://wandb.ai/pkh2120-columbia-university/hw4_problem4

View run at https://wandb.ai/pkh2120-columbia-university/hw4_problem4/runs/zy7m9e4j

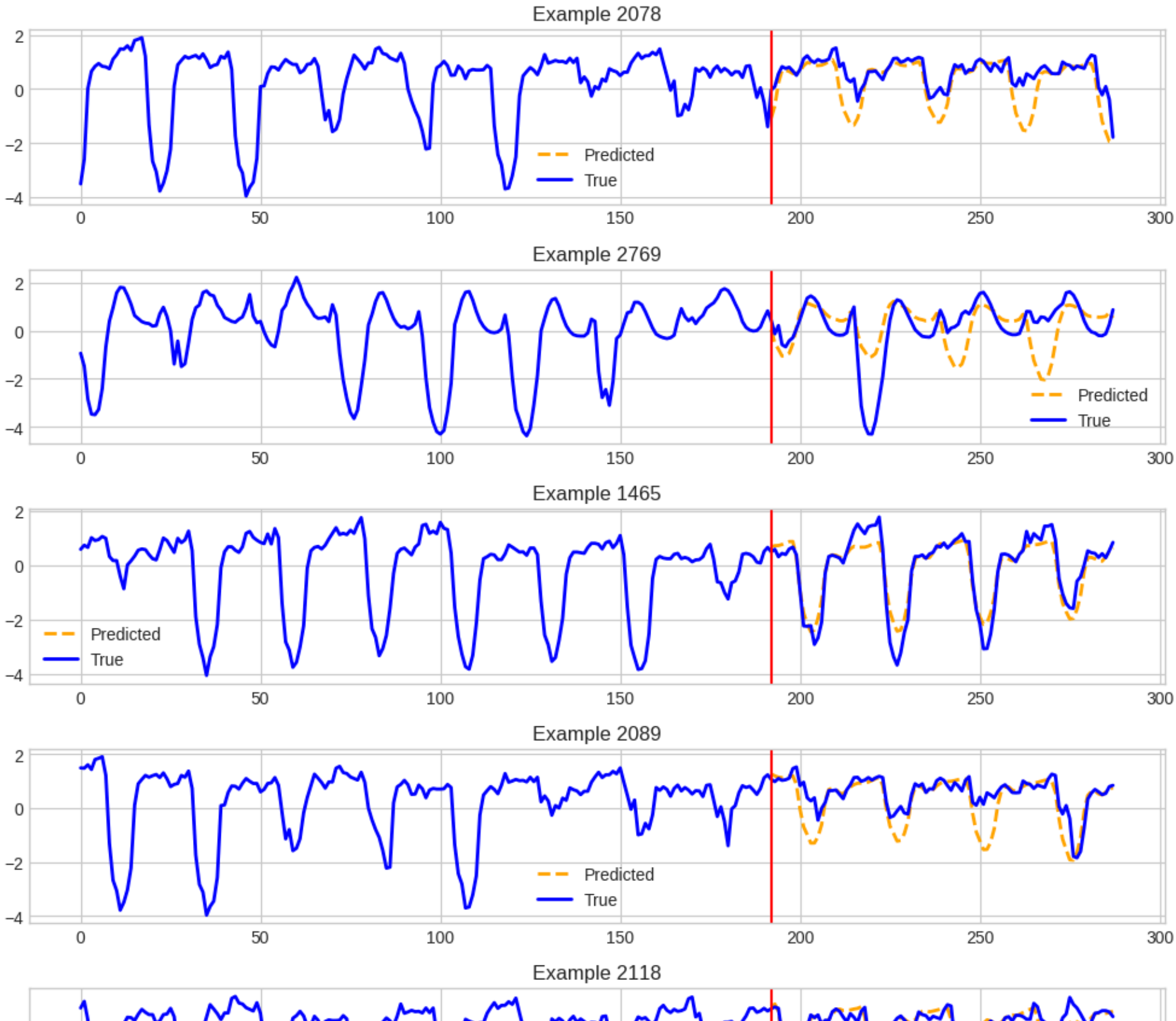
```
/usr/local/lib/python3.11/dist-packages/torch/nn/parallel/_functions.py:70: UserWarning: Was asked to gather along dime
nsion 0, but all input tensors were scalars; will instead unsqueeze and return a vector.
  warnings.warn(
```

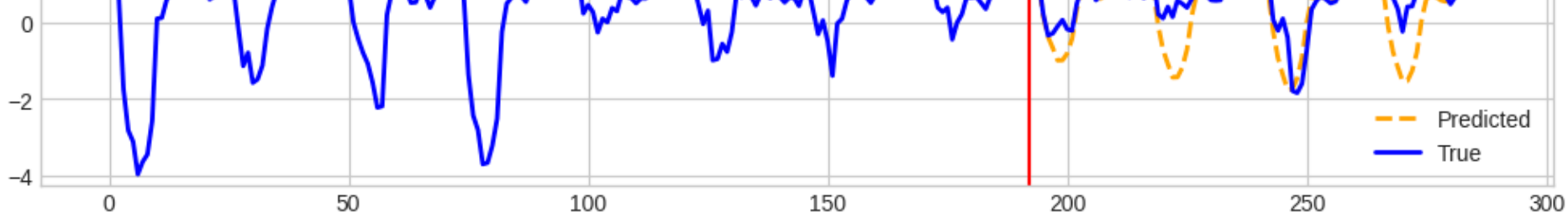
[22/22 00:01]

```
Evaluation Loss: 0.35860303044319153
```

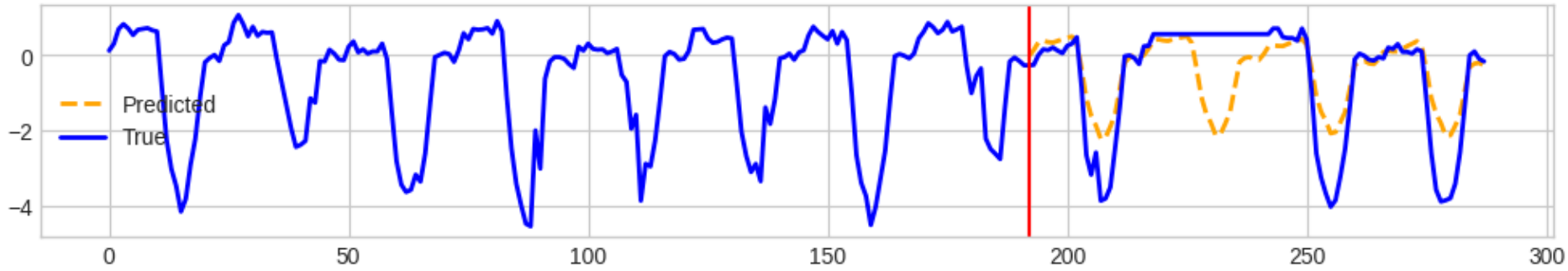
In [37]:

```python
# # plot the zero-shot predictions
# The following code plots predictions for 10 samples for every target column
# target_columns = ["HUFL", "HULL", "MUFL", "MULL", "LUFL", "LULL", "OT"]
# 70 plots in total

for i, col in enumerate(target_columns):
    plot_predictions(
        model=zeroshot_model,
        dset=test_dataset,
        plot_dir=OUT_DIR,
        plot_prefix="test_zeroshot",
        channel=i
    )
```

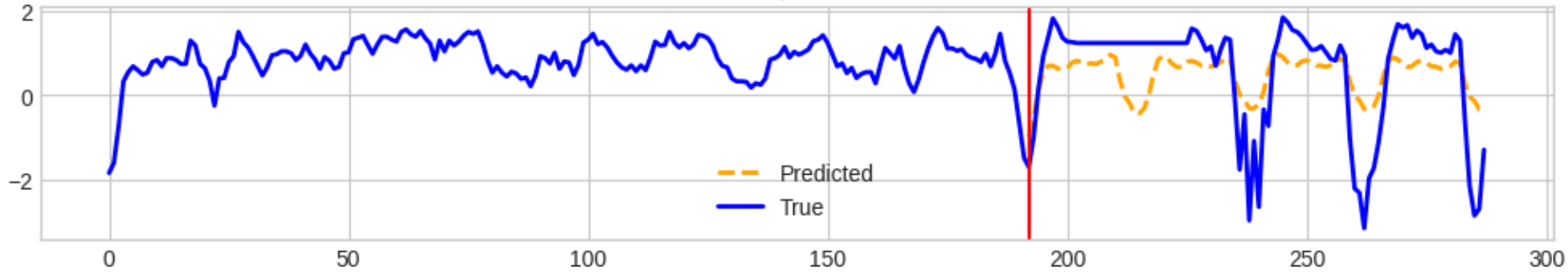

Example 2078
Predicted
True
Example 2769
Predicted
True
Example 1465
Predicted
True
Example 2089
Predicted
True
Example 2118
2
0
−2
−4
0
50
100
150
200
250
300



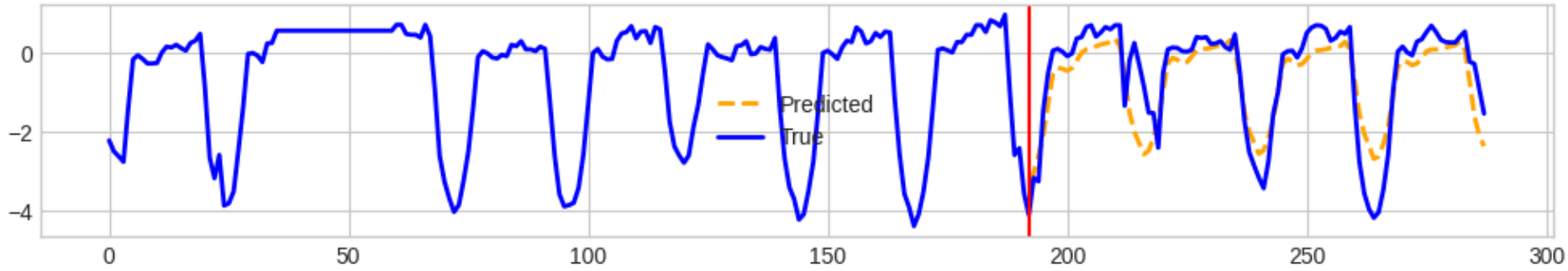
Example 141



Example 2366

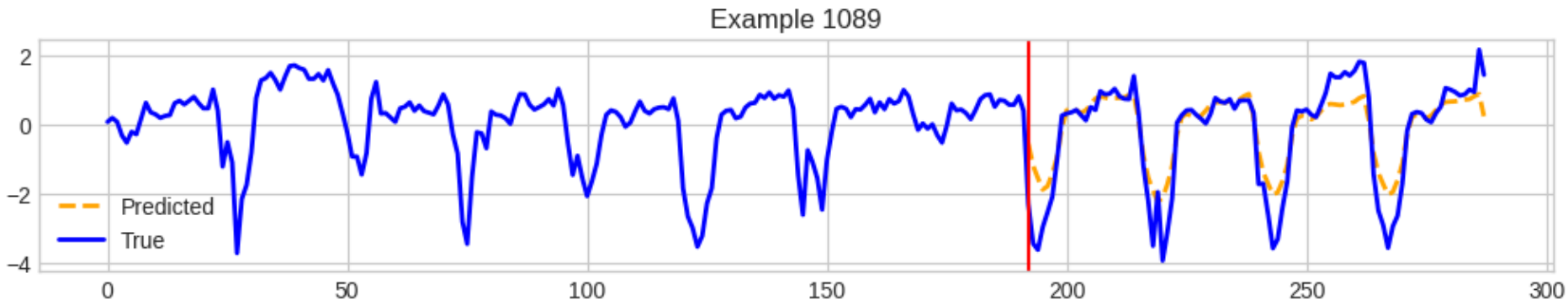


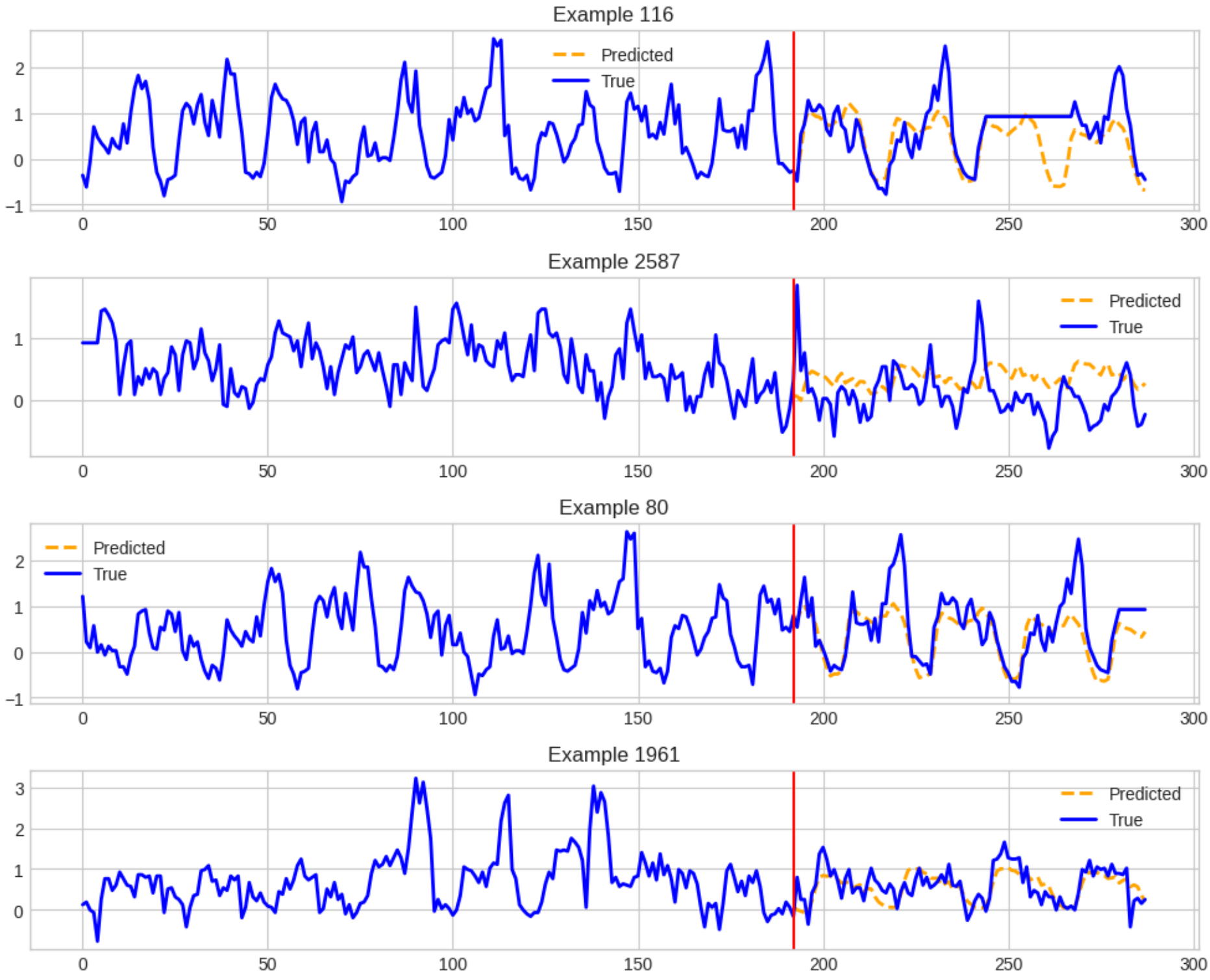
Example 324

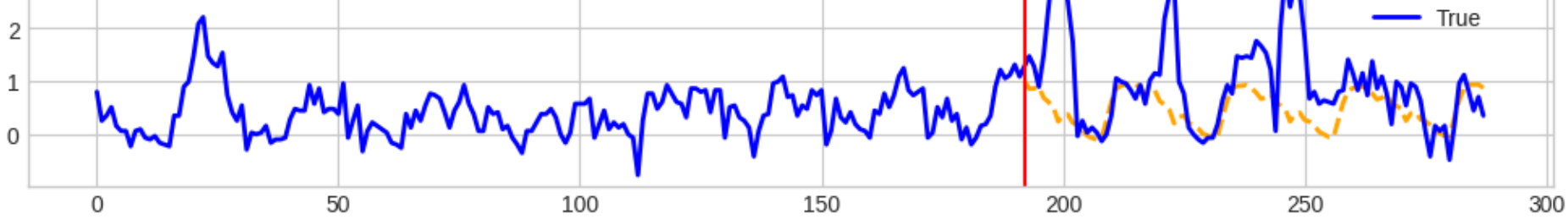


Example 1580

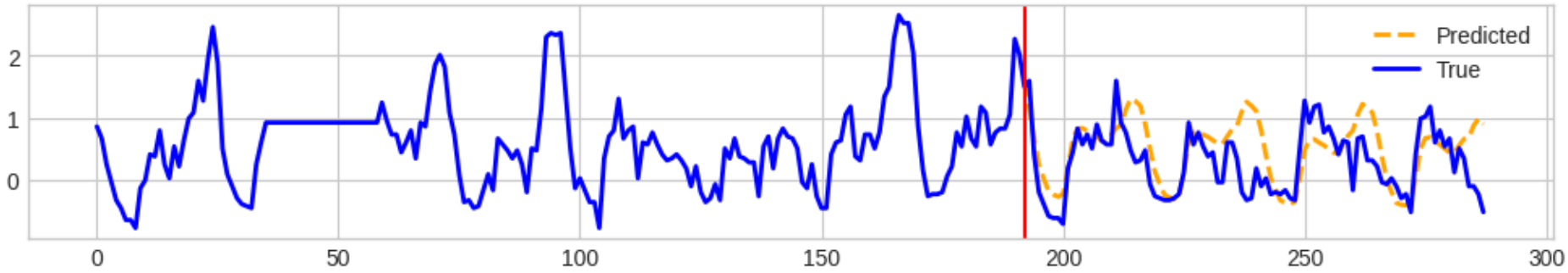

−2
−4
0
50
100
150
200
250
300


Example 1089
2
0
−2
−4
Predicted
True
0
50
100
150
200
250
300

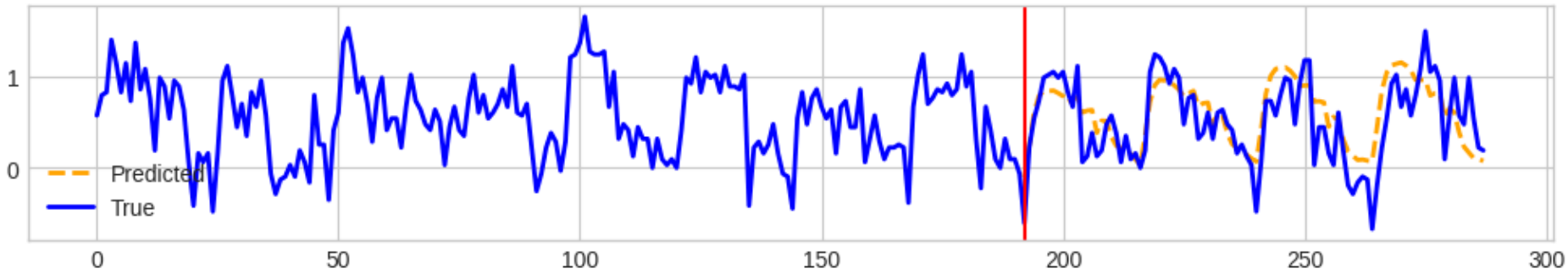

Example 116
Predicted
True
Example 2587
Predicted
True
Example 80
Predicted
True
Example 1961
Predicted
True
Example 1853
Predicted



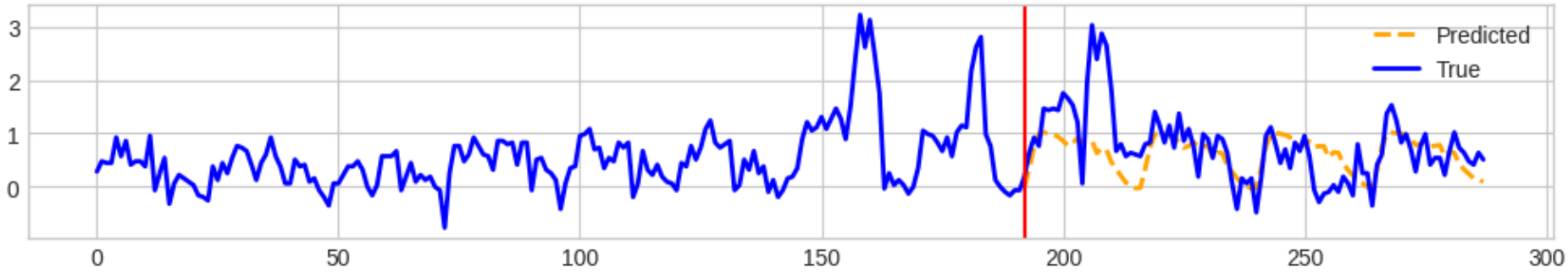
Example 325



Example 2109

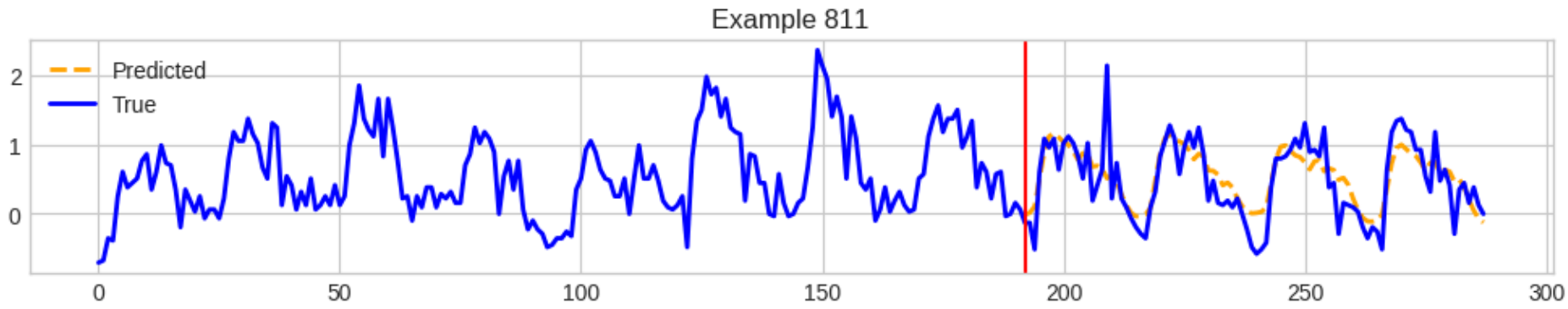


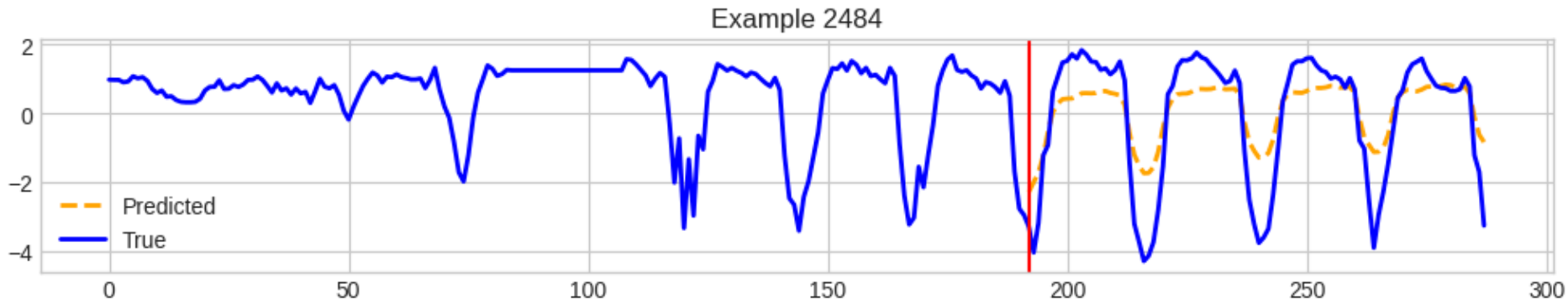
Example 1893

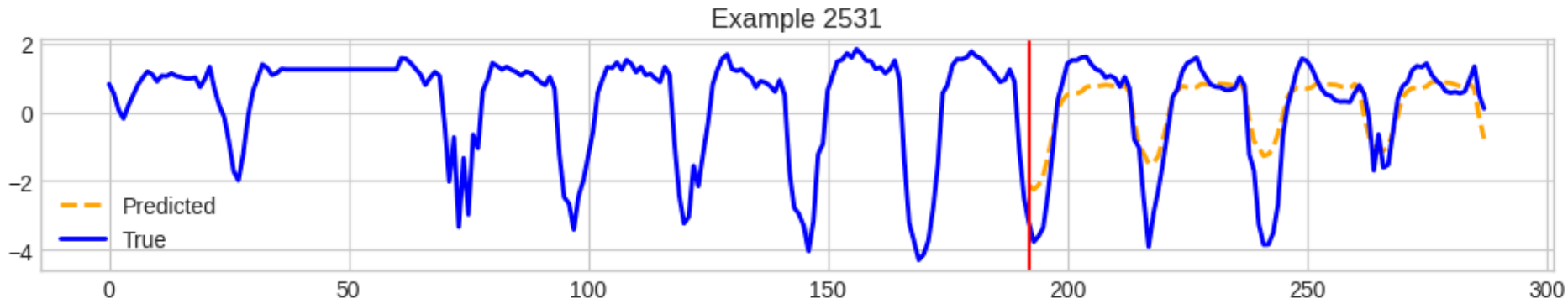


Example 1730

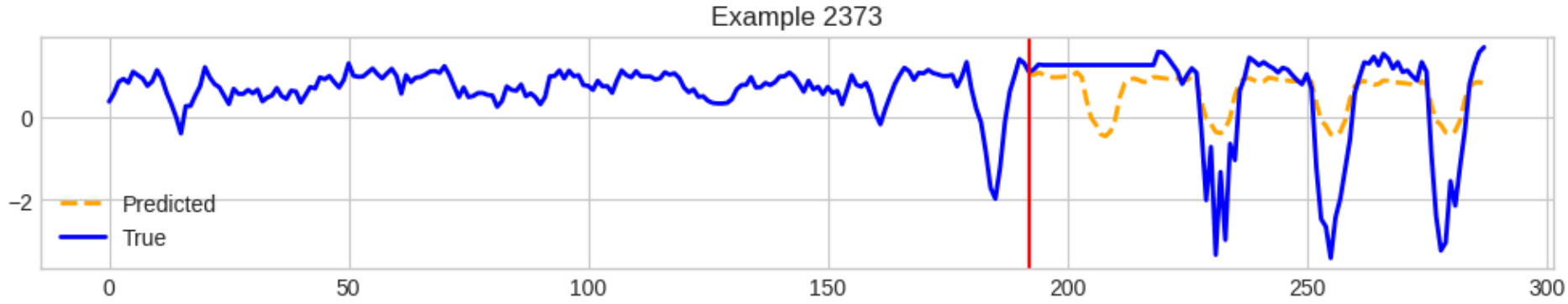

0
0
50
100
150
200
250
300

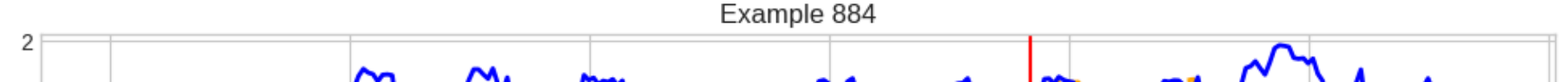

Example 811
Predicted
True
2
1
0
0
50
100
150
200
250
300

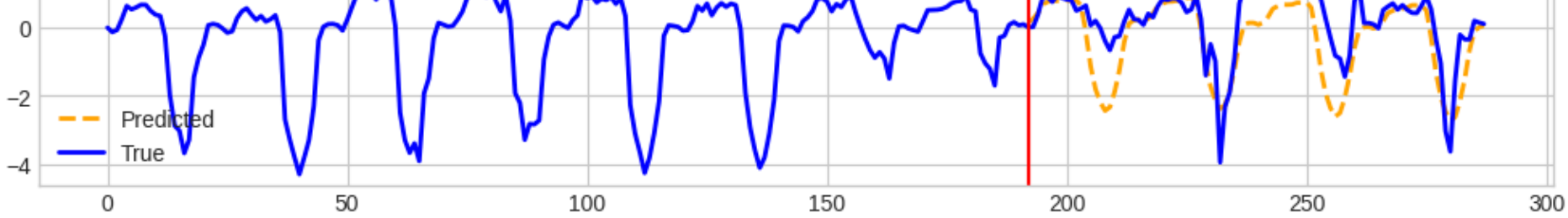

Example 2484
Predicted
True


Example 2531
Predicted
True

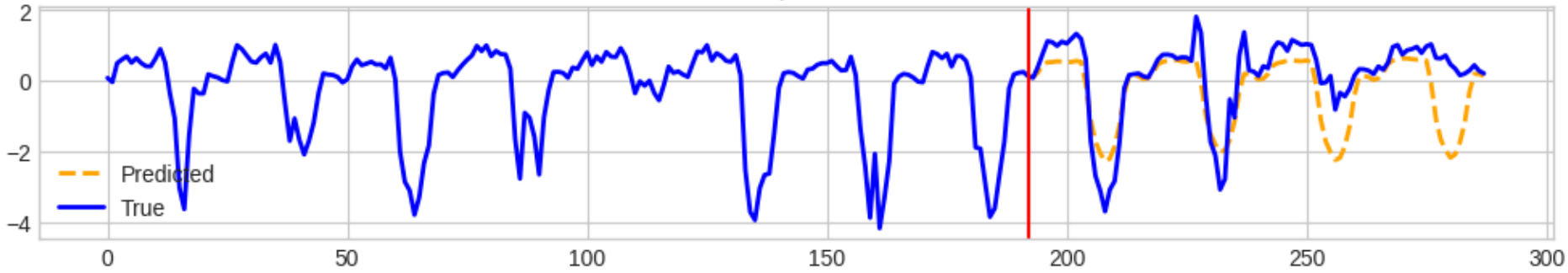

Example 1699
Predicted
True


Example 2373
Predicted
True

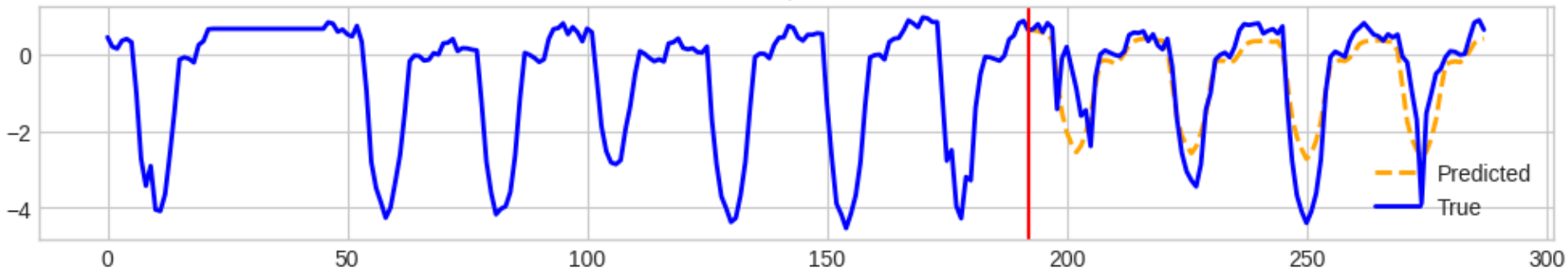

Example 884



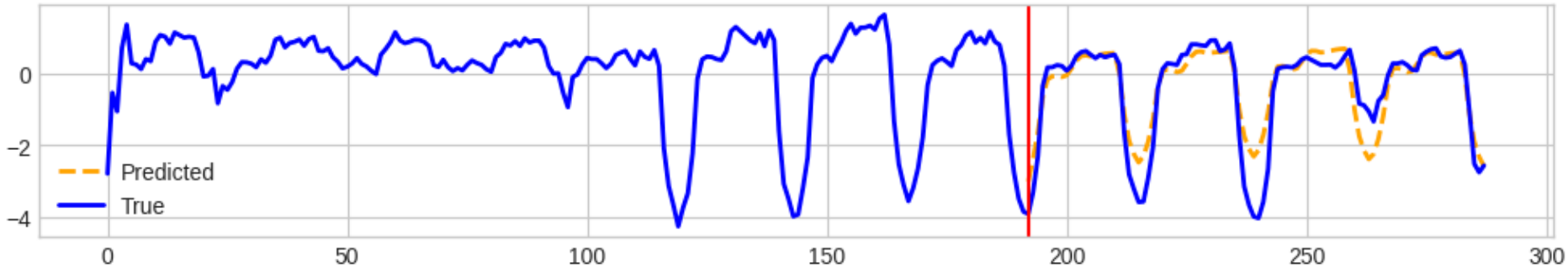
Example 1148



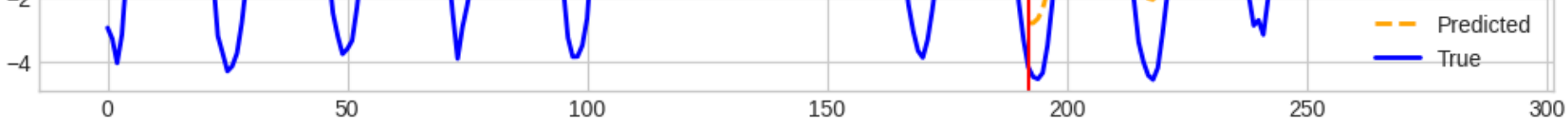
Example 338

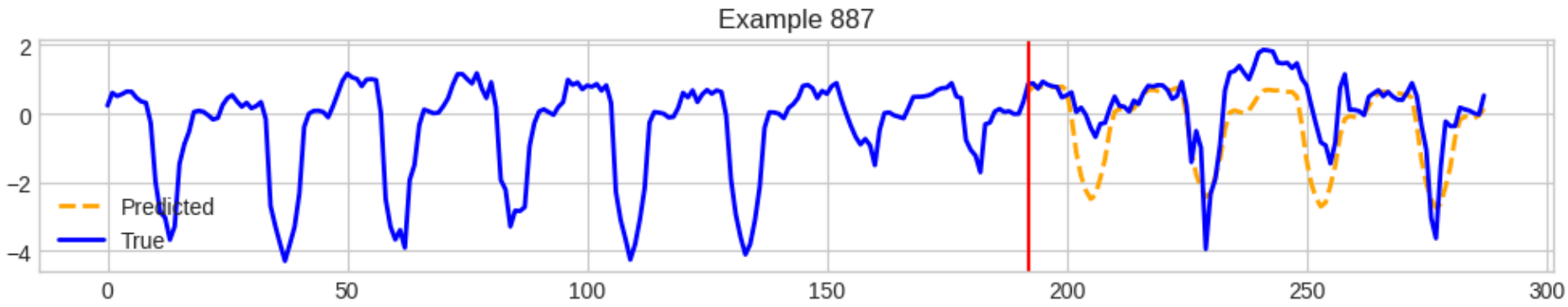


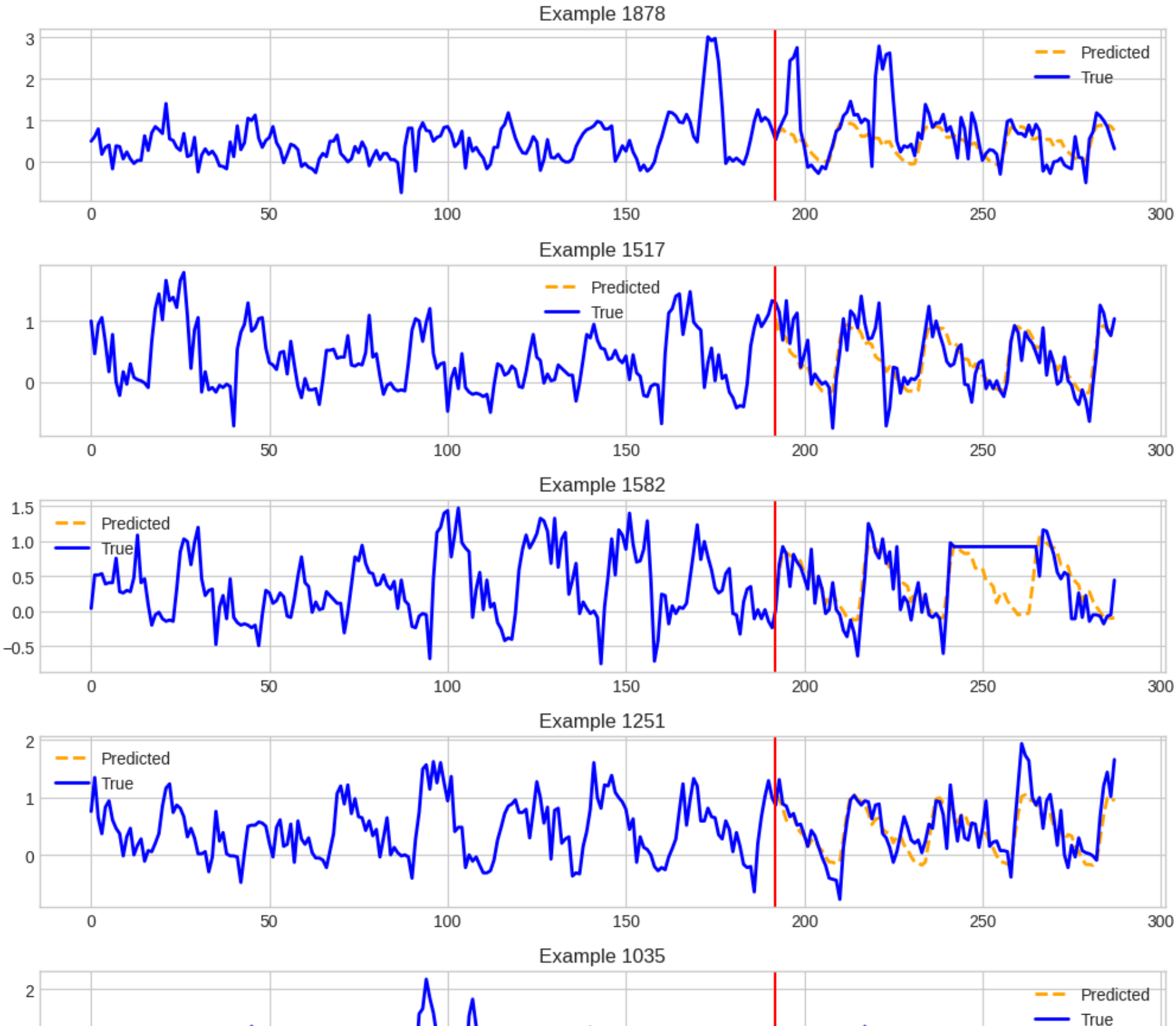
Example 1381

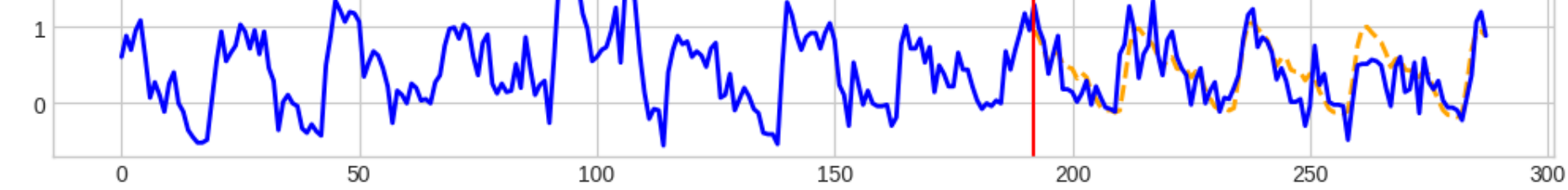


Example 2675

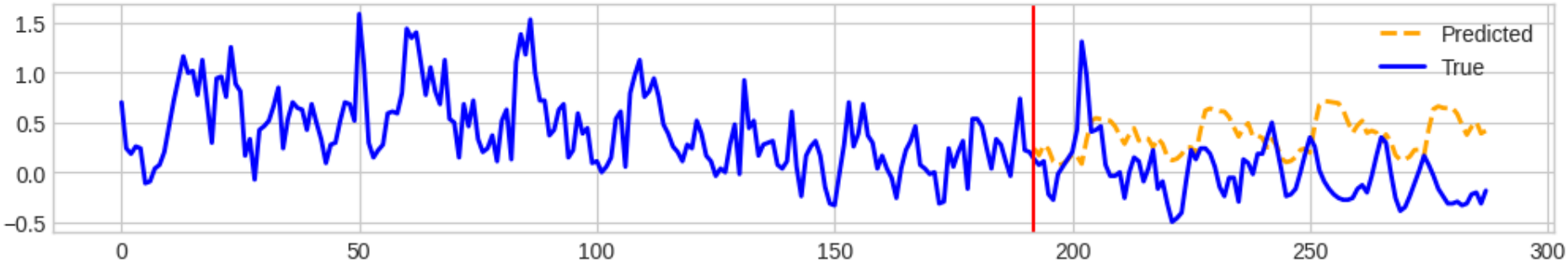

−2
−4
0
50
100
150
200
250
300
Predicted
True


Example 887
2
0
−2
−4
0
50
100
150
200
250
300
Predicted
True

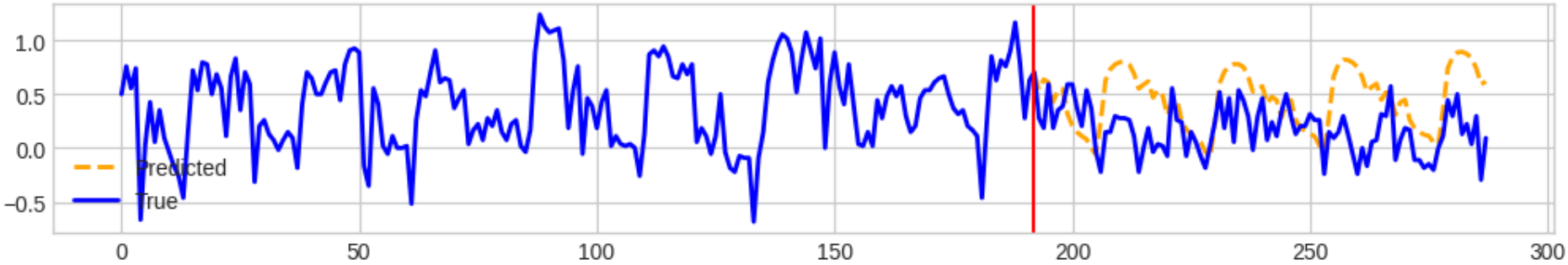

Example 1878
Predicted
True
Example 1517
Predicted
True
Example 1582
Predicted
True
Example 1251
Predicted
True
Example 1035
Predicted
True



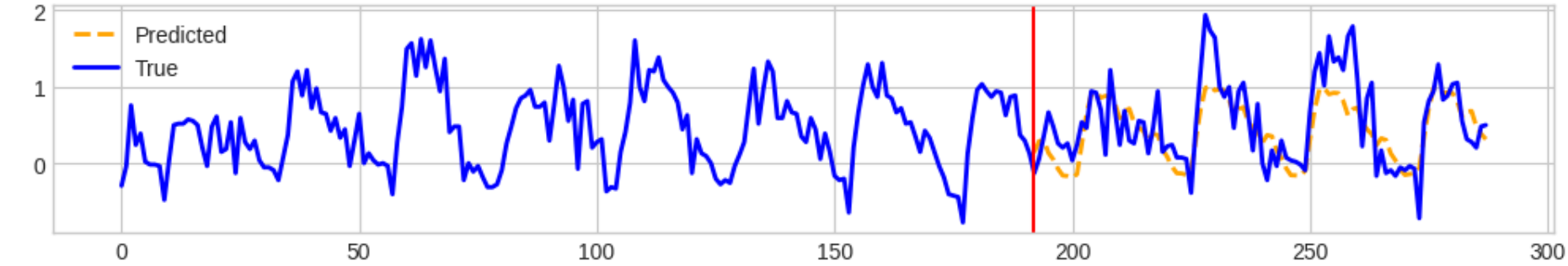
Example 2627



Example 2240

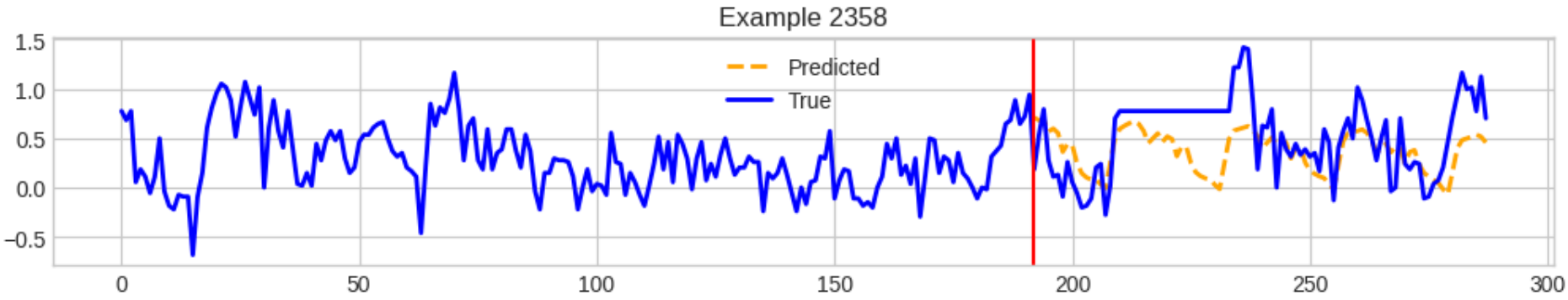


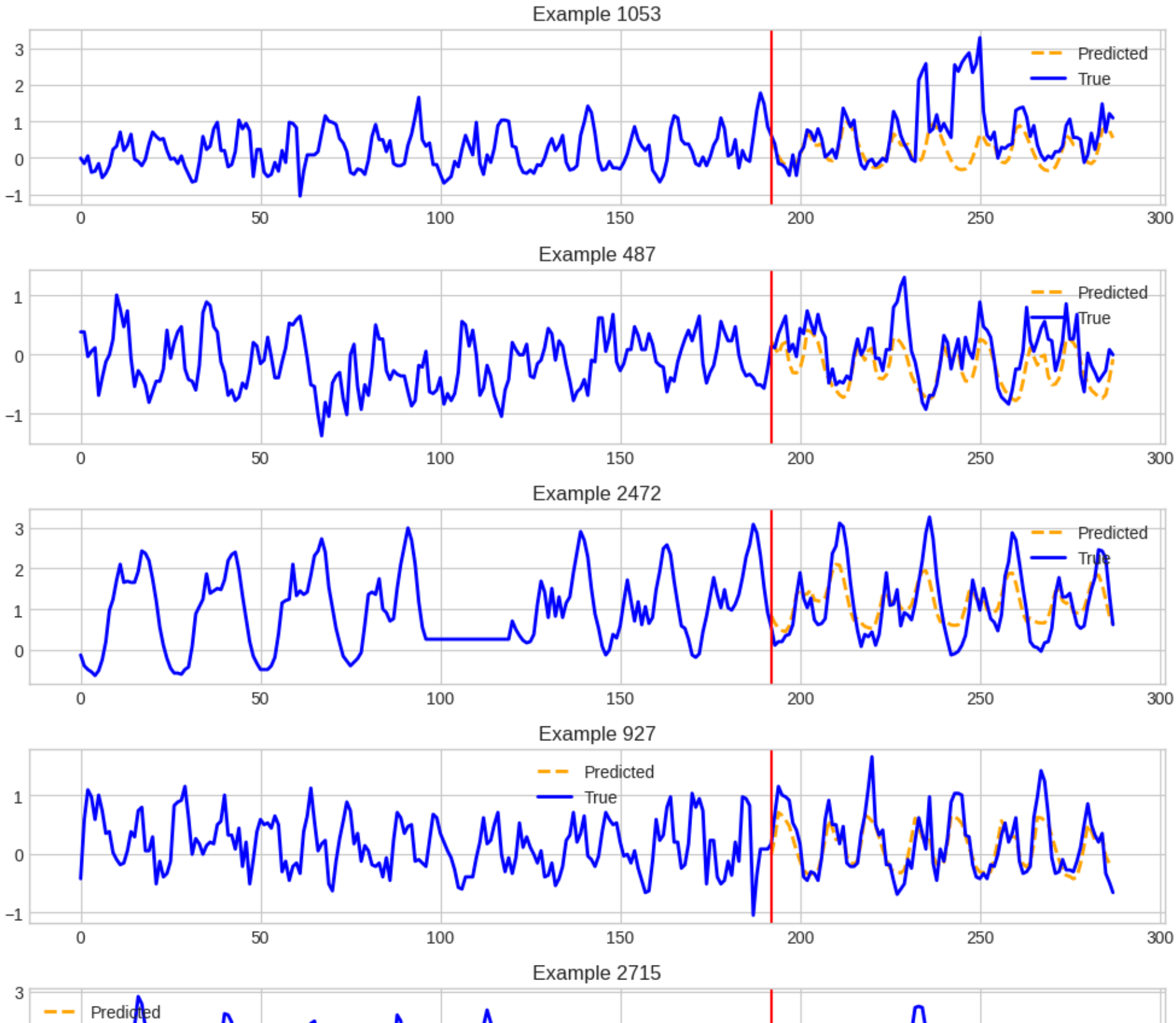
Example 1284

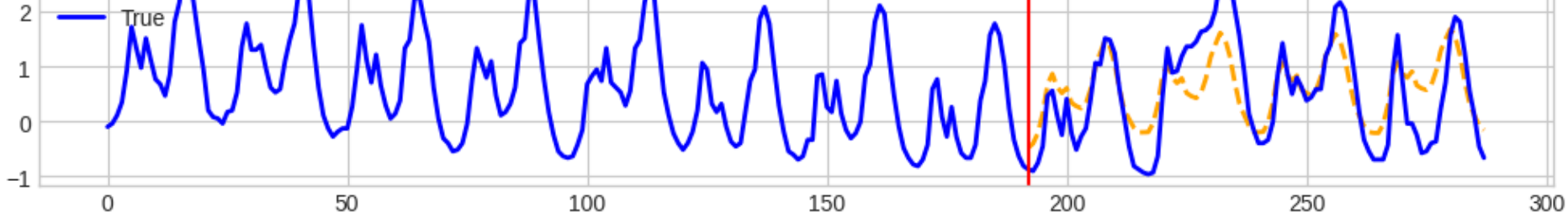


Example 2488

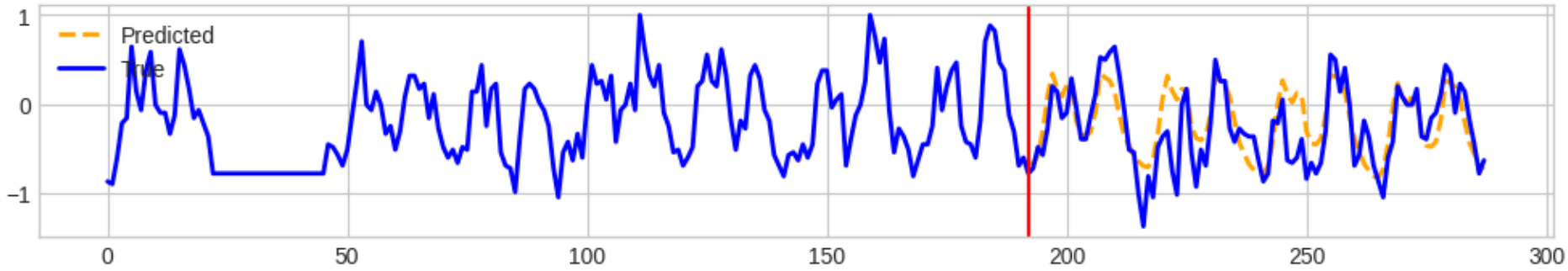

0
50
100
150
200
250
300


Example 2358
Predicted
True
1.5
1.0
0.5
0.0
−0.5
0
50
100
150
200
250
300

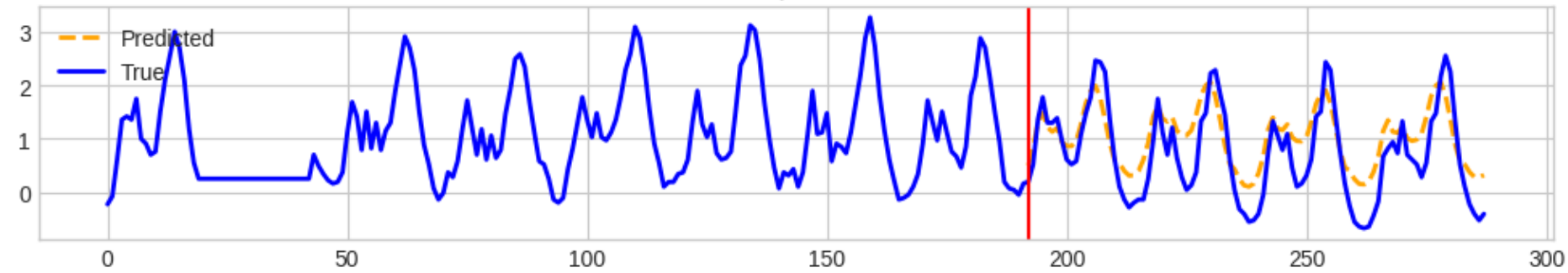

Example 1053
Predicted
True
Example 487
Predicted
True
Example 2472
Predicted
True
Example 927
Predicted
True
Example 2715
Predicted



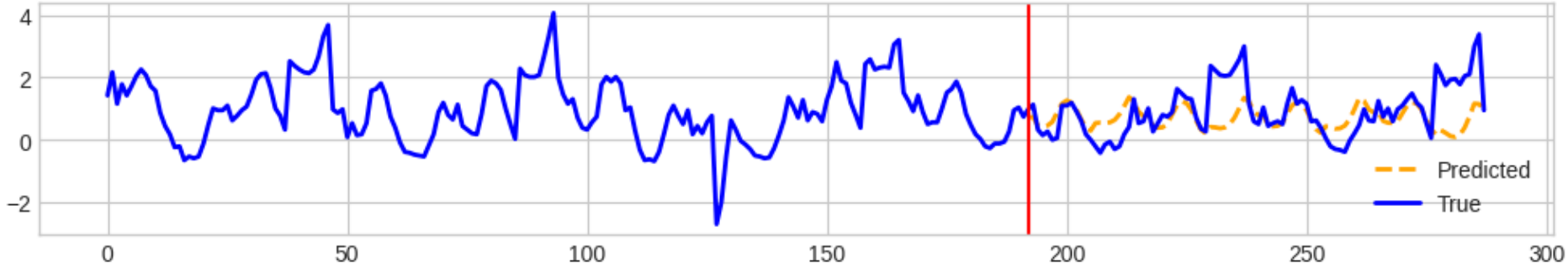
Example 338



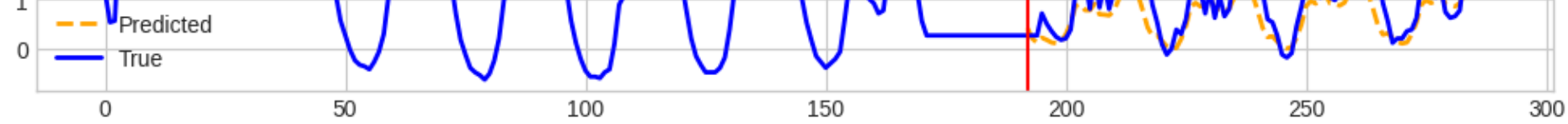
Example 2549



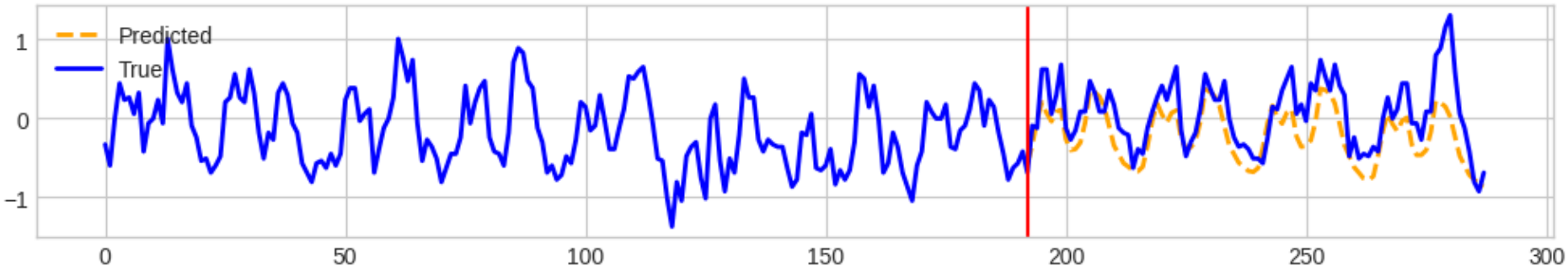
Example 2002

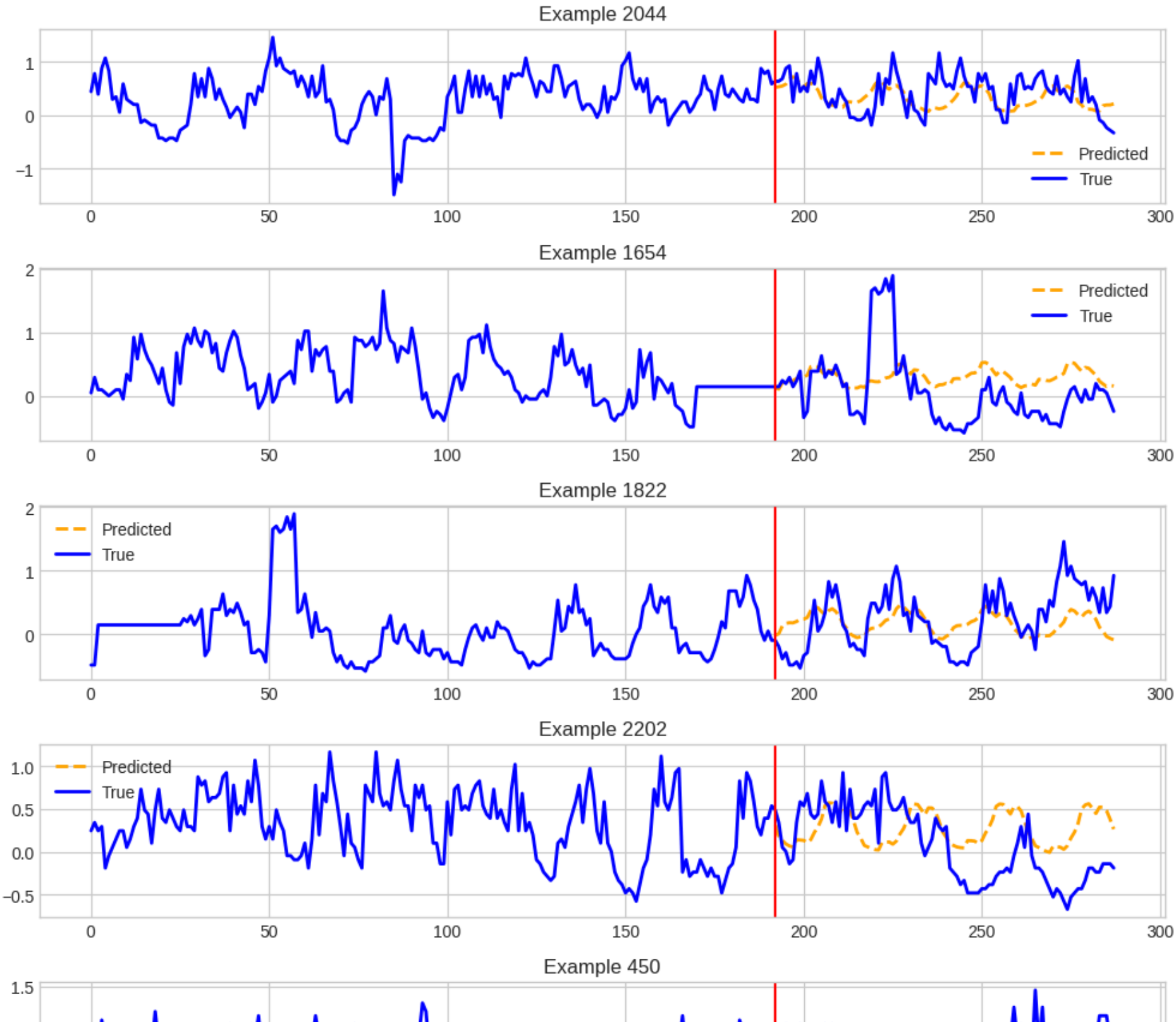


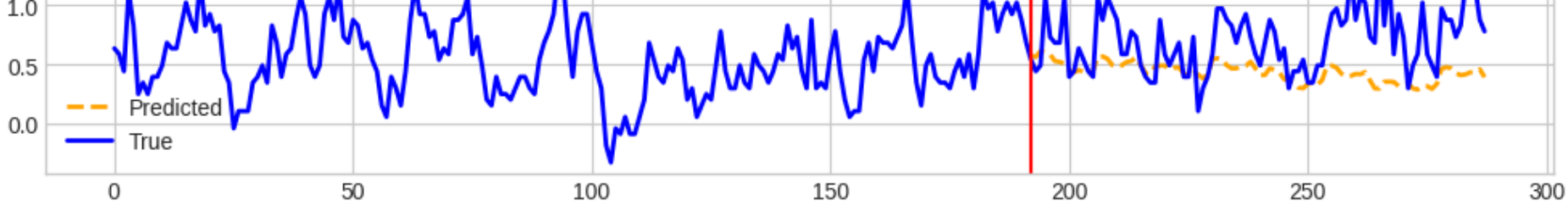
Example 2397


Predicted
True
1
0
0
50
100
150
200
250
300

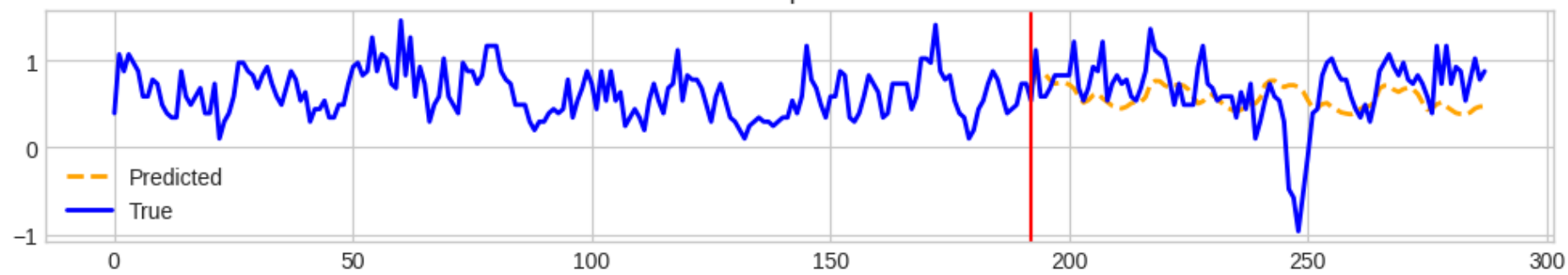
Example 436


Predicted
True
1
0
−1
0
50
100
150
200
250
300

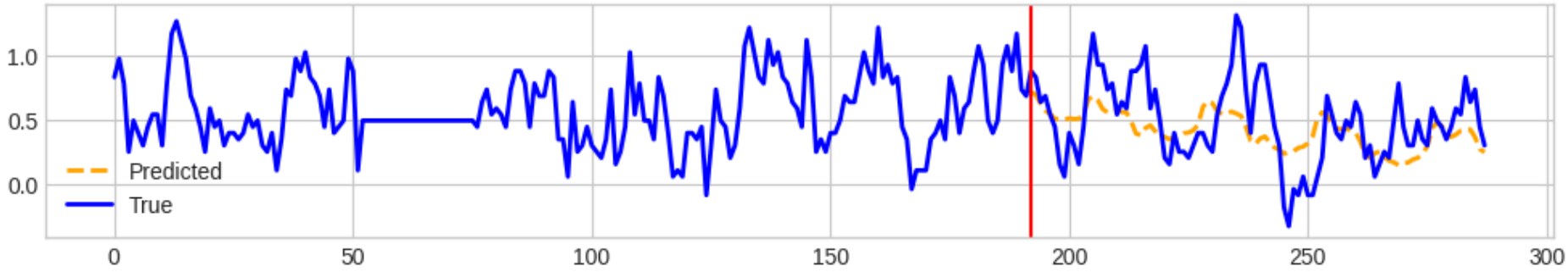

Example 2044
Predicted
True
Example 1654
Predicted
True
Example 1822
Predicted
True
Example 2202
Predicted
True
Example 450



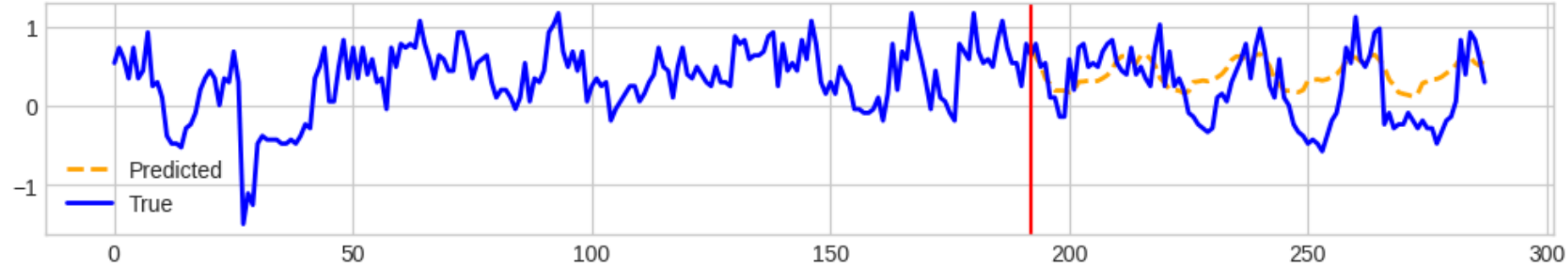
Example 655



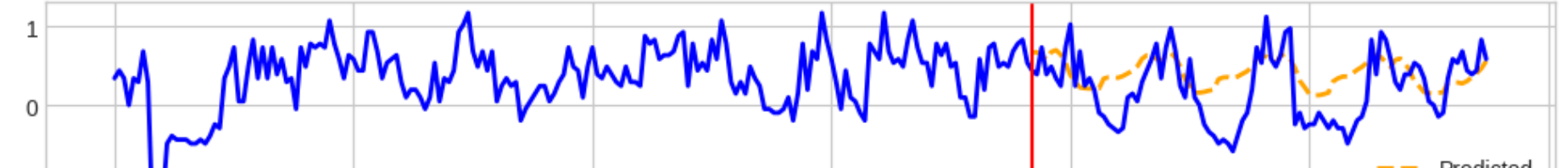
Example 308

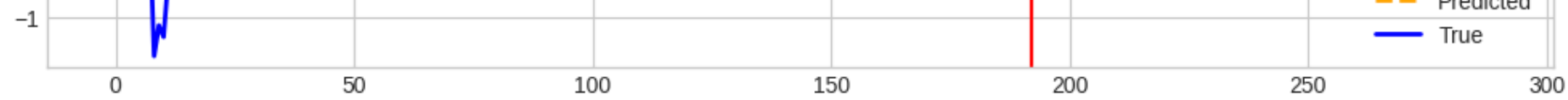


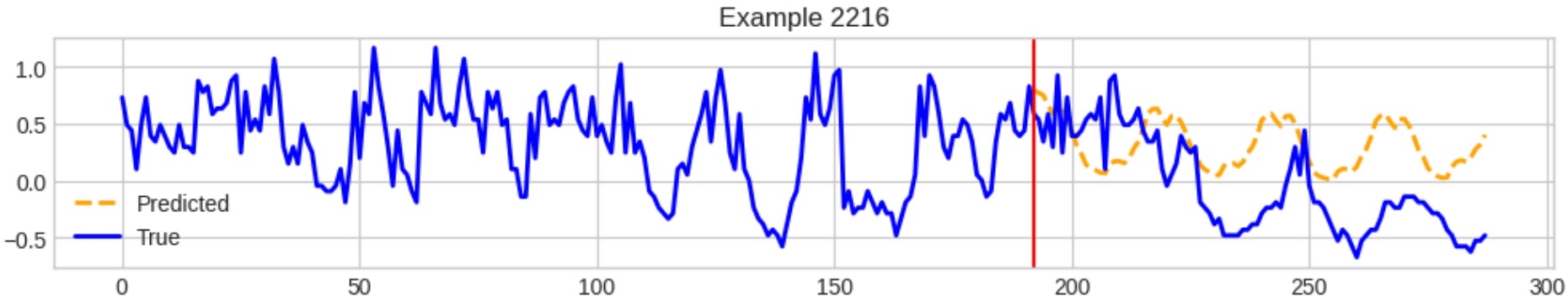
Example 2102

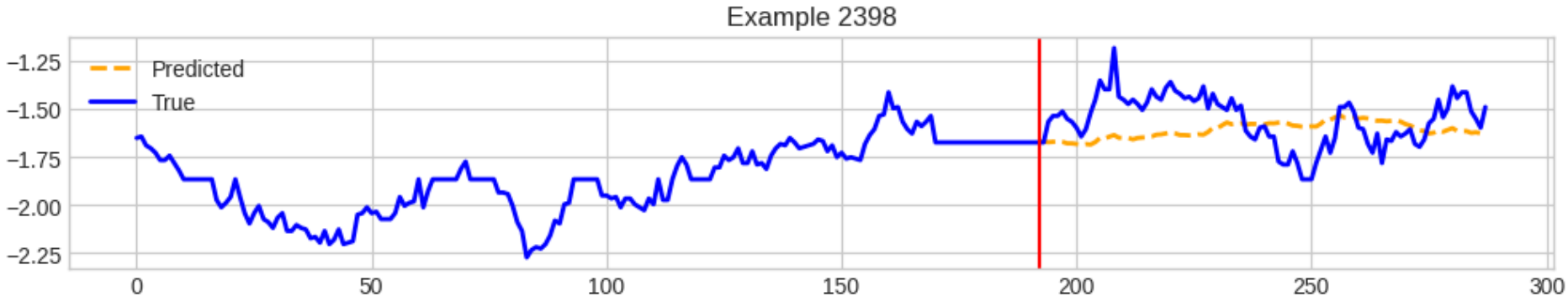


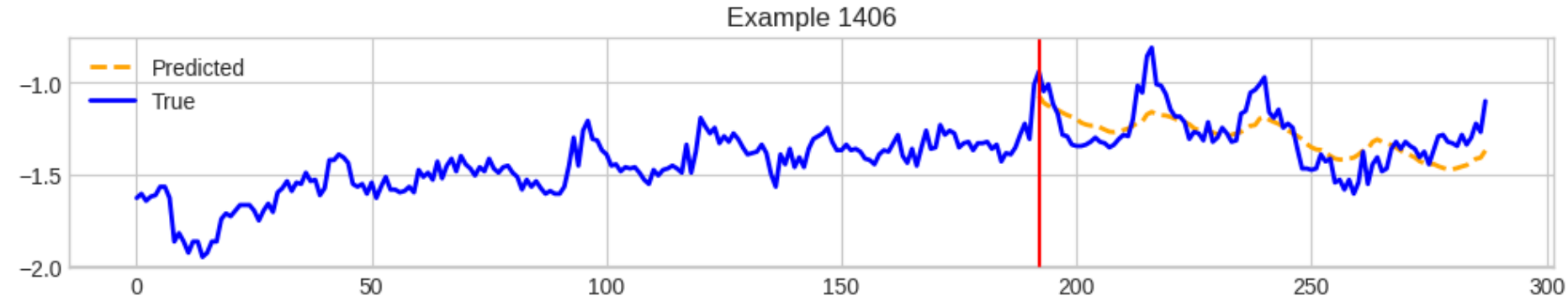
Example 2121




−1
0
50
100
150
200
250
300
Predicted
True

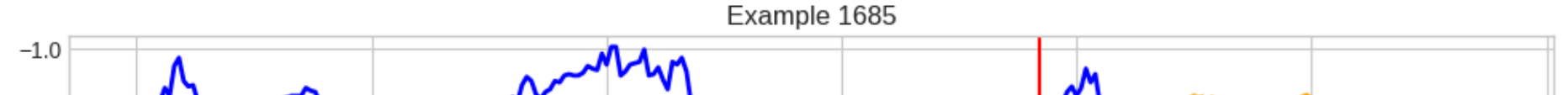

Example 2216
1.0
0.5
0.0
−0.5
0
50
100
150
200
250
300
Predicted
True


Example 2398
Predicted
True
−1.25
−1.50
−1.75
−2.00
−2.25
0
50
100
150
200
250
300

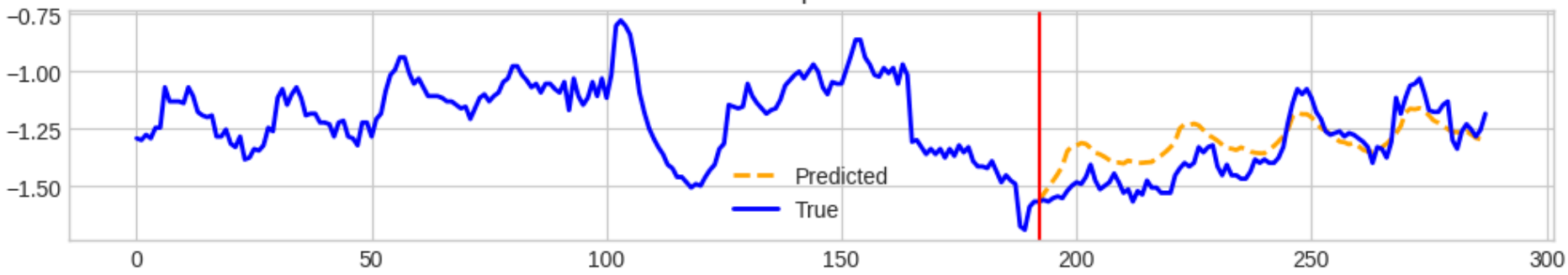

Example 1406
Predicted
True
−1.0
−1.5
−2.0
0
50
100
150
200
250
300


Example 1182
Predicted
True
−1.00
−1.25
−1.50
−1.75
−2.00
0
50
100
150
200
250
300

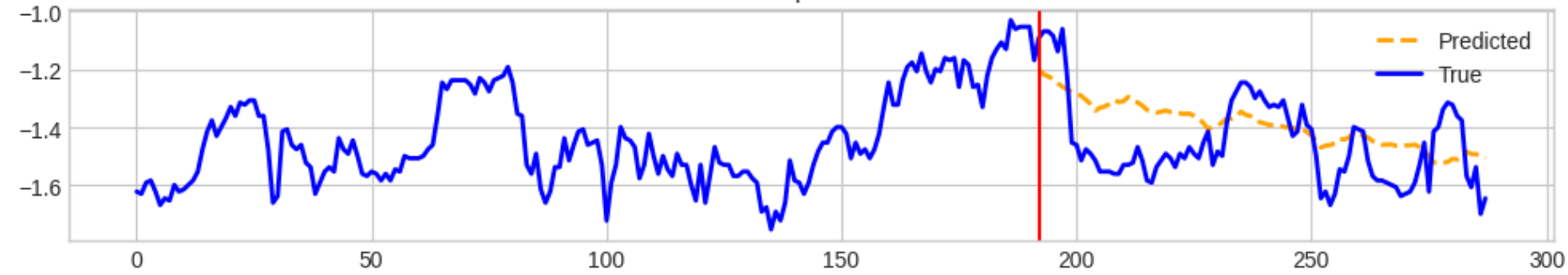

Example 1856
Predicted
True
−1.25
−1.50
−1.75
−2.00
0
50
100
150
200
250
300


Example 1685
−1.0

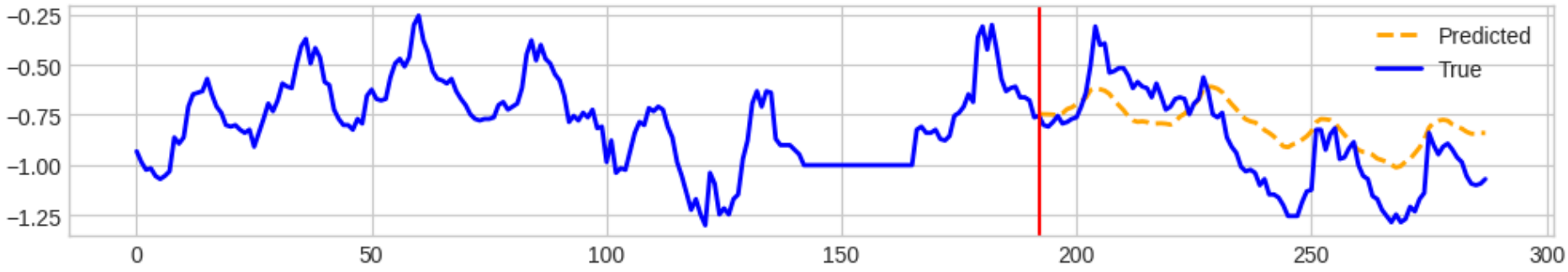
Example 895



Example 2731



Example 218



Example 667

4.4 Zero-shot evaluation by truncating the length

In [38]:

```
#Forecast 24 hrs in the future using the TTM-1024-96 model

tsp = TimeSeriesPreprocessor(
    **column_specifiers,
    context_length=context_length,
    prediction_length=24,
    scaling=True,
    encode_categorical=False,
    scaler_type="standard",
)

train_dataset, valid_dataset, test_dataset = get_datasets(
    tsp, data, split_config
)

zeroshot_model = TinyTimeMixerForPrediction.from_pretrained(
    TTM_MODEL_PATH,
    revision=TTM_MODEL_REVISION,
    prediction_filter_length=24
)

temp_dir = tempfile.mkdtemp()
zeroshot_trainer = Trainer(
    model=zeroshot_model,
    args=TrainingArguments(
        output_dir=temp_dir,
        per_device_eval_batch_size=64,
        seed=SEED,
```

```
    ),
)

wandb.init(project="hw4_problem4", name="zero_shot_eval_2")

zeroshot_results = zeroshot_trainer.evaluate(test_dataset)
print("Evaluation Loss: " + str(zeroshot_results["eval_loss"]))
```

Finishing previous runs because reinit is set to 'default'.

Run history:

eval/loss	▁
eval/runtime	▁
eval/samples_per_second	▁
eval/steps_per_second	▁
train/global_step	▁

Run summary:

eval/loss	0.3586
eval/runtime	1.4864
eval/samples_per_second	1873.666
eval/steps_per_second	14.801
train/global_step	0

View run **zero_shot_eval_1** at: https://wandb.ai/pkh2120-columbia-university/hw4_problem4/runs/zy7m9e4j

View project at: https://wandb.ai/pkh2120-columbia-university/hw4_problem4

Synced 5 W&B file(s), 0 media file(s), 0 artifact file(s) and 0 other file(s)

Find logs at: `./wandb/run-20251116_064311-zy7m9e4j/logs`

Tracking run with wandb version 0.21.0

Run data is saved locally in `/kaggle/working/wandb/run-20251116_065050-sn49ex98`

Syncing run **zero_shot_eval_2** to Weights & Biases (docs)

View project at https://wandb.ai/pkh2120-columbia-university/hw4_problem4

View run at https://wandb.ai/pkh2120-columbia-university/hw4_problem4/runs/sn49ex98

```
/usr/local/lib/python3.11/dist-packages/torch/nn/parallel/_functions.py:70: UserWarning: Was asked to gather along dime
nsion 0, but all input tensors were scalars; will instead unsqueeze and return a vector.
  warnings.warn(
```

[23/23 00:01]

```
Evaluation Loss: 0.3100925385951996
```

4.5 Few-shot finetune and evaluation method

Load model

Optionally, we can change some parameters of the model, e.g., dropout of the head.

In [39]:

```
finetune_forecast_model = TinyTimeMixerForPrediction.from_pretrained(
    TTM_MODEL_PATH, revision=TTM_MODEL_REVISION, head_dropout=0.7
)
finetune_forecast_model
```

Out[39]:
```
TinyTimeMixerForPrediction(
  (backbone): TinyTimeMixerModel(
    (encoder): TinyTimeMixerEncoder(
      (patcher): Linear(in_features=128, out_features=384, bias=True)
      (mlp_mixer_encoder): TinyTimeMixerBlock(
        (mixers): ModuleList(
          (0): TinyTimeMixerAdaptivePatchingBlock(
            (mixer_layers): ModuleList(
              (0-1): 2 x TinyTimeMixerLayer(
                (patch_mixer): PatchMixerBlock(
                  (norm): TinyTimeMixerNormLayer(
                    (norm): LayerNorm((96,), eps=1e-05, elementwise_affine=True)
                  )
                  (mlp): TinyTimeMixerMLP(
                    (fc1): Linear(in_features=32, out_features=64, bias=True)
                    (dropout1): Dropout(p=0.4, inplace=False)
                    (fc2): Linear(in_features=64, out_features=32, bias=True)
                    (dropout2): Dropout(p=0.4, inplace=False)
                  )
                  (gating_block): TinyTimeMixerGatedAttention(
                    (attn_layer): Linear(in_features=32, out_features=32, bias=True)
                    (attn_softmax): Softmax(dim=-1)
                  )
                )
                (feature_mixer): FeatureMixerBlock(
                  (norm): TinyTimeMixerNormLayer(
                    (norm): LayerNorm((96,), eps=1e-05, elementwise_affine=True)
                  )
                  (mlp): TinyTimeMixerMLP(
                    (fc1): Linear(in_features=96, out_features=192, bias=True)
                    (dropout1): Dropout(p=0.4, inplace=False)
                    (fc2): Linear(in_features=192, out_features=96, bias=True)
                    (dropout2): Dropout(p=0.4, inplace=False)
                  )
                  (gating_block): TinyTimeMixerGatedAttention(
                    (attn_layer): Linear(in_features=96, out_features=96, bias=True)
                    (attn_softmax): Softmax(dim=-1)
                  )
                )
              )
            )
          )
          (1): TinyTimeMixerAdaptivePatchingBlock(
            (mixer_layers): ModuleList(
              (0-1): 2 x TinyTimeMixerLayer(
                (patch_mixer): PatchMixerBlock(
                  (norm): TinyTimeMixerNormLayer(
```

```
            (norm): LayerNorm((192,), eps=1e-05, elementwise_affine=True)
          )
          (mlp): TinyTimeMixerMLP(
            (fc1): Linear(in_features=16, out_features=32, bias=True)
            (dropout1): Dropout(p=0.4, inplace=False)
            (fc2): Linear(in_features=32, out_features=16, bias=True)
            (dropout2): Dropout(p=0.4, inplace=False)
          )
          (gating_block): TinyTimeMixerGatedAttention(
            (attn_layer): Linear(in_features=16, out_features=16, bias=True)
            (attn_softmax): Softmax(dim=-1)
          )
        )
        (feature_mixer): FeatureMixerBlock(
          (norm): TinyTimeMixerNormLayer(
            (norm): LayerNorm((192,), eps=1e-05, elementwise_affine=True)
          )
          (mlp): TinyTimeMixerMLP(
            (fc1): Linear(in_features=192, out_features=384, bias=True)
            (dropout1): Dropout(p=0.4, inplace=False)
            (fc2): Linear(in_features=384, out_features=192, bias=True)
            (dropout2): Dropout(p=0.4, inplace=False)
          )
          (gating_block): TinyTimeMixerGatedAttention(
            (attn_layer): Linear(in_features=192, out_features=192, bias=True)
            (attn_softmax): Softmax(dim=-1)
          )
        )
      )
    )
  )
  (2): TinyTimeMixerAdaptivePatchingBlock(
    (mixer_layers): ModuleList(
      (0-1): 2 x TinyTimeMixerLayer(
        (patch_mixer): PatchMixerBlock(
          (norm): TinyTimeMixerNormLayer(
            (norm): LayerNorm((384,), eps=1e-05, elementwise_affine=True)
          )
          (mlp): TinyTimeMixerMLP(
            (fc1): Linear(in_features=8, out_features=16, bias=True)
            (dropout1): Dropout(p=0.4, inplace=False)
            (fc2): Linear(in_features=16, out_features=8, bias=True)
            (dropout2): Dropout(p=0.4, inplace=False)
          )
          (gating_block): TinyTimeMixerGatedAttention(
            (attn_layer): Linear(in_features=8, out_features=8, bias=True)
            (attn_softmax): Softmax(dim=-1)
          )
```

```
              )
              (feature_mixer): FeatureMixerBlock(
                (norm): TinyTimeMixerNormLayer(
                  (norm): LayerNorm((384,), eps=1e-05, elementwise_affine=True)
                )
                (mlp): TinyTimeMixerMLP(
                  (fc1): Linear(in_features=384, out_features=768, bias=True)
                  (dropout1): Dropout(p=0.4, inplace=False)
                  (fc2): Linear(in_features=768, out_features=384, bias=True)
                  (dropout2): Dropout(p=0.4, inplace=False)
                )
                (gating_block): TinyTimeMixerGatedAttention(
                  (attn_layer): Linear(in_features=384, out_features=384, bias=True)
                  (attn_softmax): Softmax(dim=-1)
                )
              )
            )
          )
        )
      )
    )
  )
  (patching): TinyTimeMixerPatchify()
  (scaler): TinyTimeMixerStdScaler()
)
(decoder): TinyTimeMixerDecoder(
  (adapter): Linear(in_features=384, out_features=256, bias=True)
  (decoder_block): TinyTimeMixerBlock(
    (mixers): ModuleList(
      (0-1): 2 x TinyTimeMixerLayer(
        (patch_mixer): PatchMixerBlock(
          (norm): TinyTimeMixerNormLayer(
            (norm): LayerNorm((256,), eps=1e-05, elementwise_affine=True)
          )
          (mlp): TinyTimeMixerMLP(
            (fc1): Linear(in_features=8, out_features=16, bias=True)
            (dropout1): Dropout(p=0.7, inplace=False)
            (fc2): Linear(in_features=16, out_features=8, bias=True)
            (dropout2): Dropout(p=0.7, inplace=False)
          )
          (gating_block): TinyTimeMixerGatedAttention(
            (attn_layer): Linear(in_features=8, out_features=8, bias=True)
            (attn_softmax): Softmax(dim=-1)
          )
        )
        (feature_mixer): FeatureMixerBlock(
          (norm): TinyTimeMixerNormLayer(
            (norm): LayerNorm((256,), eps=1e-05, elementwise_affine=True)
```

```
            )
            (mlp): TinyTimeMixerMLP(
              (fc1): Linear(in_features=256, out_features=512, bias=True)
              (dropout1): Dropout(p=0.7, inplace=False)
              (fc2): Linear(in_features=512, out_features=256, bias=True)
              (dropout2): Dropout(p=0.7, inplace=False)
            )
            (gating_block): TinyTimeMixerGatedAttention(
              (attn_layer): Linear(in_features=256, out_features=256, bias=True)
              (attn_softmax): Softmax(dim=-1)
            )
          )
        )
      )
    )
    (head): TinyTimeMixerForPredictionHead(
      (dropout_layer): Dropout(p=0.7, inplace=False)
      (base_forecast_block): Linear(in_features=2048, out_features=96, bias=True)
      (flatten): Flatten(start_dim=-2, end_dim=-1)
    )
  )
```

Freeze the TTM backbone

In [40]:
```python
print(
    "Number of params before freezing backbone",
    count_parameters(finetune_forecast_model),
)

# Freeze the backbone of the model
for name, param in finetune_forecast_model.named_parameters():
    if "head" not in name:
        param.requires_grad = False


# Count params
print(
    "Number of params after freezing the backbone",
    count_parameters(finetune_forecast_model),
)
```

```
Number of params before freezing backbone 2964960
Number of params after freezing the backbone 196704
```

Finetune few-shot 5%

In [41]:
```
# Important parameters
learning_rate = 0.001
num_epochs = 1  # Ideally, we need more epochs (try offline preferably in a gpu for faster computation)
batch_size = 64
```

In [48]:
```
print(f"Using learning rate = {learning_rate}")
finetune_forecast_args = TrainingArguments(
    output_dir=os.path.join(OUT_DIR, "output"),
    overwrite_output_dir=True,
    learning_rate=learning_rate,
    num_train_epochs=num_epochs,
    do_eval=True,
    eval_strategy="epoch",
    per_device_train_batch_size=batch_size,
    per_device_eval_batch_size=batch_size,
    dataloader_num_workers=8,
    report_to=None,
    save_strategy="epoch",
    logging_strategy="epoch",
    save_total_limit=1,
    logging_dir=os.path.join(OUT_DIR, "logs"),  # Make sure to specify a logging directory
    load_best_model_at_end=True,  # Load the best model when training ends
    metric_for_best_model="eval_loss",  # Metric to monitor for early stopping
    greater_is_better=False,  # For loss
    seed=SEED,
)

# Create the early stopping callback
early_stopping_callback = EarlyStoppingCallback(
    early_stopping_patience=10,  # Number of epochs with no improvement after which to stop
    early_stopping_threshold=0.0,  # Minimum improvement required to consider as improvement
)
tracking_callback = TrackingCallback()

# Optimizer and scheduler
optimizer = AdamW(finetune_forecast_model.parameters(), lr=learning_rate)
scheduler = OneCycleLR(
    optimizer,
    learning_rate,
    epochs=num_epochs,
    steps_per_epoch=math.ceil(len(train_dataset) / (batch_size)),
)

tsp = TimeSeriesPreprocessor(
    **column_specifiers,
    context_length=context_length,
    prediction_length=forecast_length,
```

```
    scaling=True,
    encode_categorical=False,
    scaler_type="standard",
)

train_dataset, valid_dataset, test_dataset = get_datasets(
    tsp, data, split_config
)

def create_few_shot_dataset(dataset, percentage):
    import torch
    total_samples = len(dataset)
    num_samples = int(total_samples * percentage)
    indices = list(range(num_samples))
    return torch.utils.data.Subset(dataset, indices)

train_dataset_5 = create_few_shot_dataset(train_dataset, 0.05)

finetune_forecast_trainer = Trainer(
    model=finetune_forecast_model,
    args=finetune_forecast_args,
    train_dataset=train_dataset_5,
    eval_dataset=valid_dataset,
    callbacks=[early_stopping_callback, tracking_callback],
    optimizers=(optimizer, scheduler),
)

# Fine tune
finetune_forecast_trainer.train()
```

```
Using learning rate = 0.001
/usr/local/lib/python3.11/dist-packages/torch/utils/data/dataloader.py:624: UserWarning: This DataLoader will create 8 worker processes in total. Our suggested max number of worker in current system is 4, which is smaller than what this DataLoader is going to create. Please be aware that excessive worker creation might get DataLoader running slow or even freeze, lower the worker number to avoid potential slowness/freeze if necessary.
  warnings.warn(
/usr/local/lib/python3.11/dist-packages/torch/nn/parallel/_functions.py:70: UserWarning: Was asked to gather along dimension 0, but all input tensors were scalars; will instead unsqueeze and return a vector.
  warnings.warn(
```

[3/3 00:01, Epoch 1/1]

Epoch	Training Loss	Validation Loss
1	0.559500	0.675480

```
[TrackingCallback] Mean Epoch Time = 0.6967084407806396 seconds, Total Train Time = 2.377359390258789
```

Out[48]:
```
TrainOutput(global_step=3, training_loss=0.5594596068064371, metrics={'train_runtime': 2.415, 'train_samples_per_second': 155.693, 'train_steps_per_second': 1.242, 'total_flos': 47946391879680.0, 'train_loss': 0.5594596068064371, 'epoch': 1.0})
```

In [49]:
```
# Evaluate the fine-tuned model
finetune_forecast_model_results = finetune_forecast_trainer.evaluate(test_dataset)
print("Evaluation Loss: " + str(finetune_forecast_model_results["eval_loss"]))
```

```
/usr/local/lib/python3.11/dist-packages/torch/utils/data/dataloader.py:624: UserWarning: This DataLoader will create 8 worker processes in total. Our suggested max number of worker in current system is 4, which is smaller than what this DataLoader is going to create. Please be aware that excessive worker creation might get DataLoader running slow or even freeze, lower the worker number to avoid potential slowness/freeze if necessary.
  warnings.warn(
/usr/local/lib/python3.11/dist-packages/torch/nn/parallel/_functions.py:70: UserWarning: Was asked to gather along dimension 0, but all input tensors were scalars; will instead unsqueeze and return a vector.
  warnings.warn(
```


[22/22 00:00]

```
Evaluation Loss: 0.3596187233924866
```

Finetune few-shot 10%

In [50]:
```
tsp = TimeSeriesPreprocessor(
    **column_specifiers,
    context_length=context_length,
    prediction_length=forecast_length,
    scaling=True,
    encode_categorical=False,
    scaler_type="standard",
)

train_dataset, valid_dataset, test_dataset = get_datasets(
    tsp, data, split_config
)

def create_few_shot_dataset(dataset, percentage):
    import torch
    total_samples = len(dataset)
    num_samples = int(total_samples * percentage)
    indices = list(range(num_samples))
    return torch.utils.data.Subset(dataset, indices)

train_dataset_10 = create_few_shot_dataset(train_dataset, 0.10)

finetune_forecast_trainer = Trainer(
    model=finetune_forecast_model,
    args=finetune_forecast_args,
```

```
    train_dataset=train_dataset_10,
    eval_dataset=valid_dataset,
    callbacks=[early_stopping_callback, tracking_callback],
    optimizers=(optimizer, scheduler),
)

# Fine tune
finetune_forecast_trainer.train()
```

```
/usr/local/lib/python3.11/dist-packages/torch/utils/data/dataloader.py:624: UserWarning: This DataLoader will create 8
worker processes in total. Our suggested max number of worker in current system is 4, which is smaller than what this D
ataLoader is going to create. Please be aware that excessive worker creation might get DataLoader running slow or even
freeze, lower the worker number to avoid potential slowness/freeze if necessary.
  warnings.warn(
/usr/local/lib/python3.11/dist-packages/torch/nn/parallel/_functions.py:70: UserWarning: Was asked to gather along dime
nsion 0, but all input tensors were scalars; will instead unsqueeze and return a vector.
  warnings.warn(
```


[6/6 00:02, Epoch 1/1]

Epoch	Training Loss	Validation Loss
1	0.507200	0.675216

```
[TrackingCallback] Mean Epoch Time = 0.8784408569335938 seconds, Total Train Time = 2.6249022483825684
```

Out[50]:
```
TrainOutput(global_step=6, training_loss=0.5072402954101562, metrics={'train_runtime': 2.6367, 'train_samples_per_seco
nd': 285.208, 'train_steps_per_second': 2.276, 'total_flos': 95892783759360.0, 'train_loss': 0.5072402954101562, 'epoc
h': 1.0})
```

In [51]:
```
# Evaluate the fine-tuned model
finetune_forecast_model_results = finetune_forecast_trainer.evaluate(test_dataset)
print("Evaluation Loss: " + str(finetune_forecast_model_results["eval_loss"]))
```

```
/usr/local/lib/python3.11/dist-packages/torch/utils/data/dataloader.py:624: UserWarning: This DataLoader will create 8
worker processes in total. Our suggested max number of worker in current system is 4, which is smaller than what this D
ataLoader is going to create. Please be aware that excessive worker creation might get DataLoader running slow or even
freeze, lower the worker number to avoid potential slowness/freeze if necessary.
  warnings.warn(
/usr/local/lib/python3.11/dist-packages/torch/nn/parallel/_functions.py:70: UserWarning: Was asked to gather along dime
nsion 0, but all input tensors were scalars; will instead unsqueeze and return a vector.
  warnings.warn(
```


[22/22 00:01]

```
Evaluation Loss: 0.3602588474750519
```

4.6 Few-shot evaluation by changing loss function

Try few-shot 5% forecasting on etth1 by changing the `loss` to `mae` (mean absolute error). Freeze the backbone and fine-tune for only 1 epoch. What is the evaluation error you get?

In [52]:

```
tsp = TimeSeriesPreprocessor(
    **column_specifiers,
    context_length=context_length,
    prediction_length=forecast_length,
    scaling=True,
    encode_categorical=False,
    scaler_type="standard",
)

train_dataset, valid_dataset, test_dataset = get_datasets(
    tsp, data, split_config
)

def create_few_shot_dataset(dataset, percentage):
    import torch
    total_samples = len(dataset)
    num_samples = int(total_samples * percentage)
    indices = list(range(num_samples))
    return torch.utils.data.Subset(dataset, indices)

train_dataset_5 = create_few_shot_dataset(train_dataset, 0.05)

finetune_forecast_model_mae = TinyTimeMixerForPrediction.from_pretrained(
    TTM_MODEL_PATH,
    revision=TTM_MODEL_REVISION,
    loss="mae"  # Change loss to MAE
)

finetune_forecast_trainer = Trainer(
    model=finetune_forecast_model_mae,
    args=finetune_forecast_args,
    train_dataset=train_dataset_5,
    eval_dataset=valid_dataset,
    callbacks=[early_stopping_callback, tracking_callback],
    optimizers=(optimizer, scheduler),
)

# Fine tune
finetune_forecast_trainer.train()
```

```
/usr/local/lib/python3.11/dist-packages/torch/utils/data/dataloader.py:624: UserWarning: This DataLoader will create 8
worker processes in total. Our suggested max number of worker in current system is 4, which is smaller than what this D
ataLoader is going to create. Please be aware that excessive worker creation might get DataLoader running slow or even
freeze, lower the worker number to avoid potential slowness/freeze if necessary.
  warnings.warn(
/usr/local/lib/python3.11/dist-packages/torch/nn/parallel/_functions.py:70: UserWarning: Was asked to gather along dime
nsion 0, but all input tensors were scalars; will instead unsqueeze and return a vector.
  warnings.warn(
```

[3/3 00:02, Epoch 1/1]

Epoch	Training Loss	Validation Loss
1	0.434600	0.556280

```
[TrackingCallback] Mean Epoch Time = 0.8676722049713135 seconds, Total Train Time = 2.599323272705078
```

Out[52]:
```
TrainOutput(global_step=3, training_loss=0.43460126717885333, metrics={'train_runtime': 2.6116, 'train_samples_per_sec
ond': 143.973, 'train_steps_per_second': 1.149, 'total_flos': 47946391879680.0, 'train_loss': 0.43460126717885333, 'ep
och': 1.0})
```

In [53]:
```
# Evaluate the fine-tuned model
finetune_forecast_model_results = finetune_forecast_trainer.evaluate(test_dataset)
print("Evaluation Loss: " + str(finetune_forecast_model_results["eval_loss"]))
```

```
/usr/local/lib/python3.11/dist-packages/torch/utils/data/dataloader.py:624: UserWarning: This DataLoader will create 8
worker processes in total. Our suggested max number of worker in current system is 4, which is smaller than what this D
ataLoader is going to create. Please be aware that excessive worker creation might get DataLoader running slow or even
freeze, lower the worker number to avoid potential slowness/freeze if necessary.
  warnings.warn(
/usr/local/lib/python3.11/dist-packages/torch/nn/parallel/_functions.py:70: UserWarning: Was asked to gather along dime
nsion 0, but all input tensors were scalars; will instead unsqueeze and return a vector.
  warnings.warn(
```

[22/22 00:00]

```
Evaluation Loss: 0.39422306418418884
```

4.7 Zero-shot on channel 0 and 2

In your notebook, add `prediction_channel_indices=[0,2]` during model loading to forecast only 0th and 2nd channels. In this case, execute the following code and note the output shape.

```
zeroshot_model = TinyTimeMixerForPrediction.from_pretrained(TTM_MODEL_PATH,
revision=TTM_MODEL_REVISION, prediction_channel_indices=[0,2])
output = zeroshot_model.forward(test_dataset[0]['past_values'].unsqueeze(0), return_loss=False)
output.prediction_outputs.shape
```

In [54]:
```
zeroshot_model = TinyTimeMixerForPrediction.from_pretrained(TTM_MODEL_PATH, revision=TTM_MODEL_REVISION, prediction_cha
output = zeroshot_model.forward(test_dataset[0]['past_values'].unsqueeze(0), return_loss=False)
output.prediction_outputs.shape
```

Out[54]:
```
torch.Size([1, 96, 2])
```