```
# Step 0: Install necessary libraries (if not already installed)
!pip install -U transformers datasets pandas scikit-learn

# Step 1: Imports
from transformers import AutoTokenizer, AutoModelForSequenceClassification, Trainer, TrainingArguments, DataCollatorWithPadding
from datasets import load_dataset
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split


# Step 2: Load and Split CSV Data
df = pd.read_csv("/content/chatbot_data.csv")
# Assuming 'Question' is the text column and 'Answer' is the label column in your CSV
train_texts, test_texts, train_labels, test_labels = train_test_split(df['Question'], df['Answer'], test_size=0.2)

# Step 3: Save as Separate Files (for Hugging Face Datasets)
# Ensure column names match the expected 'text' and 'label' for the Hugging Face dataset loading later
train_df = pd.DataFrame({'text': train_texts, 'label': train_labels})
test_df = pd.DataFrame({'text': test_texts, 'label': test_labels})
train_df.to_csv("train.csv", index=False)
test_df.to_csv("test.csv", index=False)

# Step 4: Load datasets from CSV
data_files = {"train": "train.csv", "test": "test.csv"}
dataset = load_dataset("csv", data_files=data_files)

# Step 5: Load Pretrained Tokenizer & Model
model_name = "bert-base-uncased"
tokenizer = AutoTokenizer.from_pretrained(model_name)
# Assuming the labels in your CSV are integers representing class IDs.
# If your labels are strings, you'll need to map them to integers.
# Also, make sure num_labels matches the number of unique classes in your 'Answer' column.
# For demonstration, assuming 2 classes. You might need to adjust num_labels.
label_map = {label: i for i, label in enumerate(df['Answer'].unique())}
num_labels = len(label_map)
model = AutoModelForSequenceClassification.from_pretrained(model_name, num_labels=num_labels)

# Map labels to integers in the datasets
def map_labels_to_integers(example):
    example['label'] = label_map[example['label']]
    return example

dataset = dataset.map(map_labels_to_integers)

# Step 6: Tokenize Texts
def tokenize_function(example):
    return tokenizer(example["text"], truncation=True)

tokenized_datasets = dataset.map(tokenize_function, batched=True)
data_collator = DataCollatorWithPadding(tokenizer=tokenizer)

# Step 7: Training Arguments
training_args = TrainingArguments(
    output_dir="/content/results",
    eval_strategy="epoch", # Corrected the parameter name
    learning_rate=2e-5,
    per_device_train_batch_size=8,
    per_device_eval_batch_size=8,
    num_train_epochs=3,
    weight_decay=0.01,
)
# Step 8: Trainer Setup
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=tokenized_datasets["train"],
    eval_dataset=tokenized_datasets["test"],
    tokenizer=tokenizer,
    data_collator=data_collator,
)

# Step 9: Train the Model
trainer.train()

# Step 10: Save Model
```

```
trainer.save_model("fine_tuned_bert_model")
```

```
Requirement already satisfied: transformers in /usr/local/lib/python3.11/dist-packages (4.52.4)
Requirement already satisfied: datasets in /usr/local/lib/python3.11/dist-packages (3.6.0)
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages (2.3.0)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.11/dist-packages (1.7.0)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from transformers) (3.18.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.30.0 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.33.0)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-packages (from transformers) (2.0.2)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from transformers) (24.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-packages (from transformers) (6.0.2)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.11/dist-packages (from transformers) (2024.11.6)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages (from transformers) (2.32.3)
Requirement already satisfied: tokenizers<0.22,>=0.21 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.21.1)
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: pyarrow>=15.0.0 in /usr/local/lib/python3.11/dist-packages (from datasets) (18.1.0)
Requirement already satisfied: dill<0.3.9,>=0.3.0 in /usr/local/lib/python3.11/dist-packages (from datasets) (0.3.7)
Requirement already satisfied: xxhash in /usr/local/lib/python3.11/dist-packages (from datasets) (3.5.0)
Requirement already satisfied: multiprocess<0.70.17 in /usr/local/lib/python3.11/dist-packages (from datasets) (0.70.15)
Requirement already satisfied: fsspec<=2025.3.0,>=2023.1.0 in /usr/local/lib/python3.11/dist-packages (from fsspec[http]<=2025.3.0,>=
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas) (2.9.0.post0)
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: scipy>=1.8.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn) (1.15.3)
Requirement already satisfied: joblib>=1.2.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn) (1.5.1)
Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn) (3.6.0)
Requirement already satisfied: aiohttp!=4.0.0a0,!=4.0.0a1 in /usr/local/lib/python3.11/dist-packages (from fsspec[http]<=2025.3.0,>=2
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0,>=0.30
Requirement already satisfied: hf-xet<2.0.0,>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0,>=0.30.0->tr
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->pandas) (1.17.0)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (3.4
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (2.4.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (2025.6.15
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1->f
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1->fsspec[h
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1->fsspec[http
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1->fsspec[
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1->fsspe
Requirement already satisfied: propcache>=0.2.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp!=4.0.0a0,!=4.0.0a1->fsspec[h
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[
Generating train split:    4/0 [00:00<00:00, 142.23 examples/s]

Generating test split:    1/0 [00:00<00:00, 40.19 examples/s]

Some weights of BertForSequenceClassification were not initialized from the model checkpoint at bert-base-uncased and are newly initi
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Map: 100%                                        4/4 [00:00<00:00, 154.03 examples/s]

Map: 100%                                        1/1 [00:00<00:00, 41.18 examples/s]

Map: 100%                                        4/4 [00:00<00:00, 160.34 examples/s]

Map: 100%                                        1/1 [00:00<00:00, 41.69 examples/s]

/tmp/ipython-input-8-1906840728.py:64: FutureWarning: `tokenizer` is deprecated and will be removed in version 5.0.0 for `Trainer.__i
  trainer = Trainer(
[3/3 00:15, Epoch 3/3]
```

Epoch	Training Loss	Validation Loss
1	No log	1.662523
2	No log	1.699944
3	No log	1.717896