```
import pandas as pd
import random

# Step 1: Define a list of sample contexts, questions, and answers
sample_contexts = [
    "Our store is open from 9 AM to 9 PM, Monday through Saturday. On Sundays,
    "Our main branch is located at 123 Main Street, Springfield, USA, near the
    "You can reach out to our customer support via email at support@example.com
    "Customers can return any product within 30 days of purchase, provided they
    "Yes, we offer free shipping on all orders over $50. Orders below this amou
    "After placing an order, a tracking link will be sent to your registered em
    "We accept various payment options, including Visa, MasterCard, American Ex
    "Yes, we ship to select countries. Visit our website's shipping policy page
]

sample_questions = [
    "What are your store hours?",
    "Where is your store located?",
    "How can I contact support?",
    "What is your return policy?",
    "Do you offer free shipping?",
    "How can I track my order?",
    "What payment methods are accepted?",
    "Do you offer international shipping?",
]

sample_answers = [
    "9 AM to 9 PM (Mon-Sat), 10 AM to 6 PM (Sun).",
    "123 Main Street, Springfield, USA.",
    "support@example.com or +1-800-123-4567.",
    "Return within 30 days with a receipt.",
    "Free shipping on orders over $50.",
    "Check the tracking link sent via email.",
    "Visa, MasterCard, American Express, PayPal.",
    "Yes, to select countries.",
]

# Step 2: Generate 100 random entries
data = []
for _ in range(100):
    idx = random.randint(0, len(sample_contexts) - 1)
    data.append({
        "Question": sample_questions[idx],
        "Context": sample_contexts[idx],
        "Answer": sample_answers[idx],
    })

# Step 3: Convert to a DataFrame
df = pd.DataFrame(data)
```

```
# Step 4: Save as a CSV File
csv_filename = "QA_FineTuning_100_Entries.csv"
df.to_csv(csv_filename, index=False)
print(f"Dataset with 100 entries saved as {csv_filename}.")

# If running in Google Colab, include the download option
try:
    from google.colab import files
    files.download(csv_filename)
except ImportError:
    print("If running locally, find the file in the script's directory.")
```

```
Dataset with 100 entries saved as QA_FineTuning_100_Entries.csv.
```

```
!pip install datasets
```

```
Collecting datasets
  Downloading datasets-3.2.0-py3-none-any.whl.metadata (20 kB)
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-packages (from
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.10/dist-packages (f
Requirement already satisfied: pyarrow>=15.0.0 in /usr/local/lib/python3.10/dist-package
Collecting dill<0.3.9,>=0.3.0 (from datasets)
  Downloading dill-0.3.8-py3-none-any.whl.metadata (10 kB)
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-packages (from d
Requirement already satisfied: requests>=2.32.2 in /usr/local/lib/python3.10/dist-packag
Requirement already satisfied: tqdm>=4.66.3 in /usr/local/lib/python3.10/dist-packages (
Collecting xxhash (from datasets)
  Downloading xxhash-3.5.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.me
Collecting multiprocess<0.70.17 (from datasets)
  Downloading multiprocess-0.70.16-py310-none-any.whl.metadata (7.2 kB)
Collecting fsspec<=2024.9.0,>=2023.1.0 (from fsspec[http]<=2024.9.0,>=2023.1.0->datasets
  Downloading fsspec-2024.9.0-py3-none-any.whl.metadata (11 kB)
Requirement already satisfied: aiohttp in /usr/local/lib/python3.10/dist-packages (from
Requirement already satisfied: huggingface-hub>=0.23.0 in /usr/local/lib/python3.10/dist
Requirement already satisfied: packaging in /usr/local/lib/python3.10/dist-packages (fro
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-packages (f
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in /usr/local/lib/python3.10/dist
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.10/dist-packag
Requirement already satisfied: async-timeout<6.0,>=4.0 in /usr/local/lib/python3.10/dist
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.10/dist-packa
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.10/dist-pac
Requirement already satisfied: propcache>=0.2.0 in /usr/local/lib/python3.10/dist-packag
Requirement already satisfied: yarl<2.0,>=1.17.0 in /usr/local/lib/python3.10/dist-packa
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.10/d
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dis
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-pack
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-pack
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.10/dist-
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.10/dist-packages
```

```
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from
Downloading datasets-3.2.0-py3-none-any.whl (480 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 480.6/480.6 kB 9.1 MB/s eta 0:00:00
Downloading dill-0.3.8-py3-none-any.whl (116 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 116.3/116.3 kB 9.5 MB/s eta 0:00:00
Downloading fsspec-2024.9.0-py3-none-any.whl (179 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 179.3/179.3 kB 13.1 MB/s eta 0:00:00
Downloading multiprocess-0.70.16-py310-none-any.whl (134 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 134.8/134.8 kB 10.5 MB/s eta 0:00:00
Downloading xxhash-3.5.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (194
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 194.1/194.1 kB 13.5 MB/s eta 0:00:00
ERROR: Operation cancelled by user
```

```
from google.colab import files

# Upload the file
uploaded = files.upload()
```

Choose Files QA_FineTu…_Entries.csv

- **QA_FineTuning_100_Entries.csv**(text/csv) - 18932 bytes, last modified: 12/29/2024 - 100% done

```
Saving QA_FineTuning_100_Entries.csv to QA_FineTuning_100_Entries (1).csv
```

```
import pandas as pd

# Replace 'your_file.csv' with the uploaded file name
df = pd.read_csv("QA_FineTuning_100_Entries.csv")
print(df.head())  # Display the first few rows
print(df.tail())
```

```
                            Question  \
0  What payment methods are accepted?
1         Do you offer free shipping?
2          What are your store hours?
3         Do you offer free shipping?
4           How can I track my order?

                                             Context  \
0  We accept various payment options, including V...
1  Yes, we offer free shipping on all orders over...
2  Our store is open from 9 AM to 9 PM, Monday th...
3  Yes, we offer free shipping on all orders over...
4  After placing an order, a tracking link will b...

                                      Answer
0   Visa, MasterCard, American Express, PayPal.
1            Free shipping on orders over $50.
2  9 AM to 9 PM (Mon-Sat), 10 AM to 6 PM (Sun).
3            Free shipping on orders over $50.
4      Check the tracking link sent via email.
```

```
                                Question  \
95        Where is your store located?
96          What are your store hours?
97          How can I contact support?
98         What is your return policy?
99  What payment methods are accepted?

                                                 Context  \
95  Our main branch is located at 123 Main Street,...
96  Our store is open from 9 AM to 9 PM, Monday th...
97  You can reach out to our customer support via ...
98  Customers can return any product within 30 day...
99  We accept various payment options, including V...

                                          Answer
95           123 Main Street, Springfield, USA.
96  9 AM to 9 PM (Mon-Sat), 10 AM to 6 PM (Sun).
97       support@example.com or +1-800-123-4567.
98         Return within 30 days with a receipt.
99   Visa, MasterCard, American Express, PayPal.
```

```
import pandas as pd

# Load the dataset
csv_filename = "QA_FineTuning_100_Entries.csv"
df = pd.read_csv(csv_filename)

# Check for missing or empty values
missing_values = df.isnull().sum()
print("Missing values in each column:")
print(missing_values)

# Validate contextual relevance between Question, Context, and Answer
for index, row in df.iterrows():
    question, context, answer = row["Question"], row["Context"], row["Answer"]
    if answer not in context:
        print(f"Issue found in row {index}:")
        print(f"Question: {question}")
        print(f"Context: {context}")
        print(f"Answer: {answer}")
```

```
Missing values in each column:
Question    0
Context     0
Answer      0
dtype: int64
Issue found in row 0:
Question: What payment methods are accepted?
Context: We accept various payment options, including Visa, MasterCard, American Expr
Answer: Visa, MasterCard, American Express, PayPal.
Issue found in row 1:
Question: Do you offer free shipping?
```

```
Context: Yes, we offer free shipping on all orders over $50. Orders below this amount
Answer: Free shipping on orders over $50.
Issue found in row 2:
Question: What are your store hours?
Context: Our store is open from 9 AM to 9 PM, Monday through Saturday. On Sundays, it
Answer: 9 AM to 9 PM (Mon-Sat), 10 AM to 6 PM (Sun).
Issue found in row 3:
Question: Do you offer free shipping?
Context: Yes, we offer free shipping on all orders over $50. Orders below this amount
Answer: Free shipping on orders over $50.
Issue found in row 4:
Question: How can I track my order?
Context: After placing an order, a tracking link will be sent to your registered emai
Answer: Check the tracking link sent via email.
Issue found in row 5:
Question: What is your return policy?
Context: Customers can return any product within 30 days of purchase, provided they h
Answer: Return within 30 days with a receipt.
Issue found in row 6:
Question: Do you offer international shipping?
Context: Yes, we ship to select countries. Visit our website's shipping policy page t
Answer: Yes, to select countries.
Issue found in row 7:
Question: What are your store hours?
Context: Our store is open from 9 AM to 9 PM, Monday through Saturday. On Sundays, it
Answer: 9 AM to 9 PM (Mon-Sat), 10 AM to 6 PM (Sun).
Issue found in row 8:
Question: Do you offer free shipping?
Context: Yes, we offer free shipping on all orders over $50. Orders below this amount
Answer: Free shipping on orders over $50.
Issue found in row 9:
Question: Where is your store located?
Context: Our main branch is located at 123 Main Street, Springfield, USA, near the ce
Answer: 123 Main Street, Springfield, USA.
Issue found in row 10:
Question: Do you offer international shipping?
Context: Yes, we ship to select countries. Visit our website's shipping policy page t
Answer: Yes, to select countries.
Issue found in row 11:
Question: How can I contact support?
Context: You can reach out to our customer support via email at support@example.com c
Answer: support@example.com or +1-800-123-4567.
Issue found in row 12:
Question: Where is your store located?
Context: Our main branch is located at 123 Main Street, Springfield, USA, near the ce
```

```
!pip install datasets
```

```
Collecting datasets
  Using cached datasets-3.2.0-py3-none-any.whl.metadata (20 kB)
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-packages (from
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.10/dist-packages (f
Requirement already satisfied: pyarrow>=15.0.0 in /usr/local/lib/python3.10/dist-package
```

```
Collecting dill<0.3.9,>=0.3.0 (from datasets)
  Using cached dill-0.3.8-py3-none-any.whl.metadata (10 kB)
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-packages (from d
Requirement already satisfied: requests>=2.32.2 in /usr/local/lib/python3.10/dist-packag
Requirement already satisfied: tqdm>=4.66.3 in /usr/local/lib/python3.10/dist-packages (
Collecting xxhash (from datasets)
  Using cached xxhash-3.5.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.m
Collecting multiprocess<0.70.17 (from datasets)
  Using cached multiprocess-0.70.16-py310-none-any.whl.metadata (7.2 kB)
Collecting fsspec<=2024.9.0,>=2023.1.0 (from fsspec[http]<=2024.9.0,>=2023.1.0->datasets
  Using cached fsspec-2024.9.0-py3-none-any.whl.metadata (11 kB)
Requirement already satisfied: aiohttp in /usr/local/lib/python3.10/dist-packages (from
Requirement already satisfied: huggingface-hub>=0.23.0 in /usr/local/lib/python3.10/dist
Requirement already satisfied: packaging in /usr/local/lib/python3.10/dist-packages (fro
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-packages (f
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in /usr/local/lib/python3.10/dist
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.10/dist-packag
Requirement already satisfied: async-timeout<6.0,>=4.0 in /usr/local/lib/python3.10/dist
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.10/dist-packa
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.10/dist-pac
Requirement already satisfied: propcache>=0.2.0 in /usr/local/lib/python3.10/dist-packag
Requirement already satisfied: yarl<2.0,>=1.17.0 in /usr/local/lib/python3.10/dist-packa
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.10/d
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dis
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-pack
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-pack
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.10/dist-
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.10/dist-packages
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from
Using cached datasets-3.2.0-py3-none-any.whl (480 kB)
Using cached dill-0.3.8-py3-none-any.whl (116 kB)
Using cached fsspec-2024.9.0-py3-none-any.whl (179 kB)
Using cached multiprocess-0.70.16-py310-none-any.whl (134 kB)
Using cached xxhash-3.5.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (19
Installing collected packages: xxhash, fsspec, dill, multiprocess, datasets
  Attempting uninstall: fsspec
    Found existing installation: fsspec 2024.10.0
    Uninstalling fsspec-2024.10.0:
      Successfully uninstalled fsspec-2024.10.0
ERROR: pip's dependency resolver does not currently take into account all the packages t
gcsfs 2024.10.0 requires fsspec==2024.10.0, but you have fsspec 2024.9.0 which is incomp
Successfully installed datasets-3.2.0 dill-0.3.8 fsspec-2024.9.0 multiprocess-0.70.16 xx
```

```
from datasets import Dataset
```

```
!pip show datasets
```

```
Name: datasets
Version: 3.2.0
Summary: HuggingFace community-driven open-source library of datasets
Home-page: https://github.com/huggingface/datasets
Author: HuggingFace Inc.
Author-email: thomas@huggingface.co
License: Apache 2.0
Location: /usr/local/lib/python3.10/dist-packages
Requires: aiohttp, dill, filelock, fsspec, huggingface-hub, multiprocess, numpy, packagi
Required-by:
```

```
from transformers import pipeline, AutoTokenizer, AutoModelForQuestionAnswering
from sklearn.model_selection import train_test_split
from datasets import Dataset
import torch

# Load dataset
csv_filename = "QA_FineTuning_100_Entries.csv"
df = pd.read_csv(csv_filename)

# Split into train and validation sets
train_df, val_df = train_test_split(df, test_size=0.2, random_state=42)

# Convert to Hugging Face Dataset
train_dataset = Dataset.from_pandas(train_df)
val_dataset = Dataset.from_pandas(val_df)

# Load pre-trained QA model and tokenizer
model_name = "distilbert-base-cased-distilled-squad"
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForQuestionAnswering.from_pretrained(model_name)

# Fine-tune the model (simplified training example)
qa_pipeline = pipeline("question-answering", model=model, tokenizer=tokenizer)

# Validation loop
correct_answers = 0
for index, row in val_df.iterrows():
    question = row["Question"]
    context = row["Context"]
    expected_answer = row["Answer"]

    # Get model prediction
    prediction = qa_pipeline({"question": question, "context": context})
    predicted_answer = prediction["answer"]

    # Check accuracy
    if expected_answer.strip().lower() == predicted_answer.strip().lower():
        correct_answers += 1
```

```
accuracy = correct_answers / len(val_df) * 100
print(f"Validation Accuracy: {accuracy:.2f}%")
```

```
Device set to use cpu
/usr/local/lib/python3.10/dist-packages/transformers/pipelines/question_answering.py:391
  warnings.warn(
Validation Accuracy: 0.00%
```

```
import pandas as pd
from sklearn.model_selection import train_test_split

# Step 1: Load the dataset
# Replace 'your_file.csv' with the actual file name
file_name = "QA_FineTuning_100_Entries.csv"
df = pd.read_csv(file_name)

# Display initial dataset info
print("Dataset Information:")
print(df.info())
print("\nSample Data:")
print(df.head())

# Step 2: Split the dataset into Training, Validation, and Test sets
# 70% Training, 15% Validation, 15% Test
train_df, temp_df = train_test_split(df, test_size=0.3, random_state=42)  # Ini
val_df, test_df = train_test_split(temp_df, test_size=0.5, random_state=42)  #

print("\nData Split:")
print(f"Training Set: {len(train_df)} entries")
print(f"Validation Set: {len(val_df)} entries")
print(f"Test Set: {len(test_df)} entries")

# Step 3: Check Data Quality
print("\nChecking for Missing Values:")
print(df.isnull().sum())  # Check for missing values

# Optional: Check for duplicate rows
duplicate_rows = df.duplicated().sum()
print(f"Duplicate Rows Found: {duplicate_rows}")

# Step 4: Evaluate Data Diversity
# Count unique values in each column
print("\nData Diversity Evaluation:")
for column in df.columns:
    unique_count = df[column].nunique()
    print(f"{column}: {unique_count} unique values")

# Optional: Check balance across classes for classification tasks
# For QA tasks, examine the distribution of questions
```

```
question_distribution = df['Question'].value_counts()
print("\nQuestion Distribution:")
print(question_distribution)

# Step 5: Save the split datasets
train_df.to_csv("Training_Set.csv", index=False)
val_df.to_csv("Validation_Set.csv", index=False)
test_df.to_csv("Test_Set.csv", index=False)

print("\nDatasets saved as Training_Set.csv, Validation_Set.csv, and Test_Set.c
```

```
Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 3 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Question  100 non-null    object
 1   Context   100 non-null    object
 2   Answer    100 non-null    object
dtypes: object(3)
memory usage: 2.5+ KB
None

Sample Data:
                            Question  \
0  What payment methods are accepted?
1         Do you offer free shipping?
2          What are your store hours?
3         Do you offer free shipping?
4           How can I track my order?

                                             Context  \
0  We accept various payment options, including V...
1  Yes, we offer free shipping on all orders over...
2  Our store is open from 9 AM to 9 PM, Monday th...
3  Yes, we offer free shipping on all orders over...
4  After placing an order, a tracking link will b...

                                        Answer
0   Visa, MasterCard, American Express, PayPal.
1             Free shipping on orders over $50.
2  9 AM to 9 PM (Mon-Sat), 10 AM to 6 PM (Sun).
3             Free shipping on orders over $50.
4       Check the tracking link sent via email.

Data Split:
Training Set: 70 entries
Validation Set: 15 entries
Test Set: 15 entries

Checking for Missing Values:
Question    0
Context     0
```

```
Answer      0
dtype: int64
Duplicate Rows Found: 92

Data Diversity Evaluation:
Question: 8 unique values
Context: 8 unique values
Answer: 8 unique values

Question Distribution:
Question
What are your store hours?              18
Where is your store located?            18
Do you offer free shipping?             13
How can I track my order?               13
```