from datasets import load\_dataset

common\_voice = load\_dataset("mozilla-foundation/common\_voice\_11", "en", use\_auth\_token=True)

def get\_audio\_data(dataset, split="train"):

audio\_data = dataset[split]

return audio\_data

if \_\_name\_\_ == "\_\_main\_\_":

split = "train" # You can change this to "validation" or "test" if needed

audio\_data = get\_audio\_data(common\_voice, split)

print(f"Number of samples in {split} split: {len(audio\_data)}")

print("Sample data:")

for sample in audio\_data[:5]:

print(f"Audio Path: {sample['path']}")

print(f"Text: {sample['sentence']}")

print("\n")