from sagemaker.async\_inference import AsyncInferenceConfig

async\_config = AsyncInferenceConfig(output\_path="s3://bucket/output/")

predictor = model.deploy(instance\_count=1, instance\_type="ml.m5.large", async\_inference\_config=async\_config)