Pipeline

```
!pip install metaflow pandas spacy transformers datasets evaluate seqeval matplotlib
!python -m spacy download en_core_web_sm
```

Show hidden output

```
%%writefile medical_extraction_flow.py
from metaflow import FlowSpec, step, Parameter
import os, json
import pandas as pd
import numpy as np
import spacy
import matplotlib.pyplot as plt
from datasets import Dataset
from transformers import AutoTokenizer, pipeline
import evaluate

class MedicalExtractionFlow(FlowSpec):
    data_path = Parameter("data_path", default="open_ave_data.csv")
    @step
    def start(self):
        self.df = pd.read_csv(self.data_path)
        self.df = self.df.rename(columns={
            "ReportText":"text","ExamName":"Examination",
            "clinicaldata":"Clinical","findings":"Findings",
            "impression":"Impression"
        }).dropna(subset=["text"]).reset_index(drop=True)
        self.df["clean"] = (self.df["text"]
                            .str.replace(r"\s+"," ",regex=True)
                            .str.strip().str.strip(":.;"))
        self.df["clean_match"] = self.df["clean"].str.lower().str.strip(":.; ")
        self.next(self.preprocess)

    @step
    def preprocess(self):
        nlp = spacy.blank("en")
        self.df["tokens"] = self.df["clean"].apply(lambda t: [tok.text for tok in nlp(t)])
        lengths = self.df["tokens"].str.len()
        plt.hist(lengths, bins=30); plt.xlabel("tokens"); plt.ylabel("reports")
        plt.savefig("lengths.png"); plt.close()
        self.next(self.infer)

    @step
    def infer(self):
        ds = Dataset.from_pandas(self.df[["clean"]], preserve_index=False)
        MODEL = "Qwen/Qwen3-0.6B"
        tokenizer = AutoTokenizer.from_pretrained(MODEL, trust_remote_code=True)
        ner = pipeline("token-classification", model=MODEL, tokenizer=tokenizer,
                       aggregation_strategy="simple", device=0)
        fields = ["Examination","Clinical","Findings","Impression"]
        def extract_batch(batch):
            texts = batch["clean"]; ents = ner(texts)
            out = {f"pred_{f}":[] for f in fields}
            for e_list in ents:
                groups = {f:[] for f in fields}
                for e in e_list:
                    if e["entity_group"] in groups:
                        groups[e["entity_group"]].append(e["word"].strip())
                for f in fields:
                    out[f"pred_{f}"].append(" ".join(groups[f]).strip())
            return out
        ds = ds.map(extract_batch, batched=True, batch_size=16, remove_columns=["clean"])
        self.df = pd.concat([self.df, ds.to_pandas()], axis=1)
        self.next(self.evaluate)

    @step
    def evaluate(self):
        fields = ["Examination","Clinical","Findings","Impression"]
        metrics = {}
        for f in fields:
            true = self.df[f].fillna("").str.lower().str.strip(":.; ")
            pred = self.df[f"pred_{f}"].fillna("").str.lower().str.strip(":.; ")
            match = true == pred
            self.df[f"{f}_match"] = match
            metrics[f] = float(match.mean())
        metrics["overall_exact_match"] = float(np.mean(list(metrics.values())))
        self.metrics = metrics
        print("Metrics:", metrics)
        self.next(self.save)

    @step
    def save(self):
        self.df.to_csv("results.csv", index=False)
        with open("metrics.json","w") as f: json.dump(self.metrics, f, indent=2)
        self.next(self.end)

    @step
    def end(self):
        print("Done. Artifacts: results.csv, metrics.json, lengths.png")

if __name__=="__main__":
    MedicalExtractionFlow()
```

```
Writing medical_extraction_flow.py
```

```
import os
os.environ["METAFLOW_USER"] = "suhasi"
```

```
!python medical_extraction_flow.py run
```

```
2025-08-04 21:14:37.959769: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register cuFFT factory: Attempting to register factory for plug
WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
E0000 00:00:1754342077.979131    6292 cuda_dnn.cc:8310] Unable to register cuDNN factory: Attempting to register factory for plugin cuDNN when one has already been
E0000 00:00:1754342077.985260    6292 cuda_blas.cc:1418] Unable to register cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has already be
Metaflow 2.16.8 executing MedicalExtractionFlow for user:suhas
Validating your flow...
    The graph looks good!
Running pylint...
    Pylint not found, so extra checks are disabled.
2025-08-04 21:14:41.240 Workflow starting (run-id 1754342081239626):
2025-08-04 21:14:41.250 [1754342081239626/start/1 (pid 6353)] Task is starting.
2025-08-04 21:14:48.969 [1754342081239626/start/1 (pid 6353)] 2025-08-04 21:14:48.969633: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to r
2025-08-04 21:14:48.989 [1754342081239626/start/1 (pid 6353)] WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
2025-08-04 21:14:48.995 [1754342081239626/start/1 (pid 6353)] E0000 00:00:1754342088.989393    6353 cuda_dnn.cc:8310] Unable to register cuDNN factory: Attempting t
2025-08-04 21:14:48.995 [1754342081239626/start/1 (pid 6353)] E0000 00:00:1754342088.995482    6353 cuda_blas.cc:1418] Unable to register cuBLAS factory: Attempting
2025-08-04 21:14:53.940 [1754342081239626/start/1 (pid 6353)] Task finished successfully.
2025-08-04 21:14:53.944 [1754342081239626/preprocess/2 (pid 6418)] Task is starting.
2025-08-04 21:15:01.637 [1754342081239626/preprocess/2 (pid 6418)] 2025-08-04 21:15:01.637504: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable
2025-08-04 21:15:01.657 [1754342081239626/preprocess/2 (pid 6418)] WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
2025-08-04 21:15:01.663 [1754342081239626/preprocess/2 (pid 6418)] E0000 00:00:1754342101.657211    6418 cuda_dnn.cc:8310] Unable to register cuDNN factory: Attempt
2025-08-04 21:15:01.663 [1754342081239626/preprocess/2 (pid 6418)] E0000 00:00:1754342101.663301    6418 cuda_blas.cc:1418] Unable to register cuBLAS factory: Atte
2025-08-04 21:15:07.175 [1754342081239626/preprocess/2 (pid 6418)] Task finished successfully.
2025-08-04 21:15:07.179 [1754342081239626/infer/3 (pid 6485)] Task is starting.
2025-08-04 21:15:17.347 [1754342081239626/infer/3 (pid 6485)] 2025-08-04 21:15:17.347349: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to r
2025-08-04 21:15:17.368 [1754342081239626/infer/3 (pid 6485)] WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
2025-08-04 21:15:17.374 [1754342081239626/infer/3 (pid 6485)] E0000 00:00:1754342117.367954    6485 cuda_dnn.cc:8310] Unable to register cuDNN factory: Attempting t
2025-08-04 21:15:17.374 [1754342081239626/infer/3 (pid 6485)] E0000 00:00:1754342117.374099    6485 cuda_blas.cc:1418] Unable to register cuBLAS factory: Attempting
2025-08-04 21:15:25.084 [1754342081239626/infer/3 (pid 6485)] Some weights of Qwen3ForTokenClassification were not initialized from the model checkpoint at Qwen/Qwe
2025-08-04 21:15:25.088 [1754342081239626/infer/3 (pid 6485)] You should probably TRAIN this model on a down-stream task to be able to use it for predictions and in
2025-08-04 21:15:25.088 [1754342081239626/infer/3 (pid 6485)] Device set to use cuda:0
Map:  17%|██          | 160/954 [00:12<01:00, 13.05 examples/s]You seem to be using the pipelines sequentially on GPU. In order to maximize efficiency please use a d
Map: 100%|██████████| 954/954 [01:16<00:00, 12.53 examples/s]
2025-08-04 21:16:57.268 [1754342081239626/infer/3 (pid 6485)] Task finished successfully.
2025-08-04 21:16:57.284 [1754342081239626/evaluate/4 (pid 6966)] Task is starting.
2025-08-04 21:17:10.909 [1754342081239626/evaluate/4 (pid 6966)] 2025-08-04 21:17:10.909288: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable t
2025-08-04 21:17:10.948 [1754342081239626/evaluate/4 (pid 6966)] WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
2025-08-04 21:17:10.948 [1754342081239626/evaluate/4 (pid 6966)] E0000 00:00:1754342230.947829    6966 cuda_dnn.cc:8310] Unable to register cuDNN factory: Attemptin
2025-08-04 21:17:10.960 [1754342081239626/evaluate/4 (pid 6966)] E0000 00:00:1754342230.960016    6966 cuda_blas.cc:1418] Unable to register cuBLAS factory: Attempt
2025-08-04 21:17:16.030 [1754342081239626/evaluate/4 (pid 6966)] Metrics: {'Examination': 0.0, 'Clinical': 0.0041928721174004195, 'Findings': 0.0, 'Impression': 0.0
2025-08-04 21:17:17.868 [1754342081239626/evaluate/4 (pid 6966)] Task finished successfully.
2025-08-04 21:17:17.872 [1754342081239626/save/5 (pid 7067)] Task is starting.
2025-08-04 21:17:25.658 [1754342081239626/save/5 (pid 7067)] 2025-08-04 21:17:25.658696: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to re
2025-08-04 21:17:25.678 [1754342081239626/save/5 (pid 7067)] WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
2025-08-04 21:17:25.684 [1754342081239626/save/5 (pid 7067)] E0000 00:00:1754342245.678245    7067 cuda_dnn.cc:8310] Unable to register cuDNN factory: Attempting to
2025-08-04 21:17:25.684 [1754342081239626/save/5 (pid 7067)] E0000 00:00:1754342245.684180    7067 cuda_blas.cc:1418] Unable to register cuBLAS factory: Attempting
2025-08-04 21:17:30.872 [1754342081239626/save/5 (pid 7067)] Task finished successfully.
2025-08-04 21:17:30.876 [1754342081239626/end/6 (pid 7132)] Task is starting.
2025-08-04 21:17:38.744 [1754342081239626/end/6 (pid 7132)] 2025-08-04 21:17:38.744794: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to reg
2025-08-04 21:17:38.764 [1754342081239626/end/6 (pid 7132)] WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
2025-08-04 21:17:38.771 [1754342081239626/end/6 (pid 7132)] E0000 00:00:1754342258.764852    7132 cuda_dnn.cc:8310] Unable to register cuDNN factory: Attempting to
2025-08-04 21:17:38.771 [1754342081239626/end/6 (pid 7132)] E0000 00:00:1754342258.771043    7132 cuda_blas.cc:1418] Unable to register cuBLAS factory: Attempting t
2025-08-04 21:17:42.099 [1754342081239626/end/6 (pid 7132)] Done. Artifacts: results.csv, metrics.json, lengths.png
2025-08-04 21:17:43.894 [1754342081239626/end/6 (pid 7132)] Task finished successfully.
2025-08-04 21:17:43.895 Done!
```

Start coding or generate with AI.