



## Assignment: Clinical NLP Challenge (No LLMs)



### Objective

Extract structured clinical information from a multi-report PDF using **only traditional NLP / ML / NER methods** (no OpenAI/GPT or any LLMs).

You are given a PDF with real-world endoscopy reports. Your task is to:

- Extract clinical entities and relevant codes
  - Organize output in structured JSON format
  - **Do not use LLMs or prompt engineering**
- 



### Input



**File:** [Input Data for assignment.pdf](#)

Contains **4 clinical reports**, each including:

- Diagnosis and procedure sections
- 



### Your Task

For **each report**, extract and output the following:



### Required Fields


- **Clinical Terms:** Symptoms, conditions, procedures, medical findings (e.g., "diverticulosis", "rectal bleeding", "colonoscopy")
- **Anatomical Locations:** Organs or body parts mentioned (e.g., "rectum", "sigmoid colon")
- **Diagnosis:** Natural language summary or extraction of diagnoses (e.g., "internal hemorrhoids", "Barrett's esophagus")
- **Procedures:** Clinical procedures performed (e.g., "colonoscopy", "EGD with biopsy")
- **ICD-10:** International Classification of Diseases codes
- **CPT:** Current Procedural Terminology codes
- **HCPCS:** Healthcare Common Procedure Coding System codes (if applicable)
- **Modifiers:** Any medical billing modifiers (if applicable)

## Output Format (per report)

```
json
{
  "ReportID": "Report 1",
  "Clinical Terms": [],
  "Anatomical Locations": [],
  "Diagnosis": [],
  "Procedures": [],
  "ICD-10": [],
  "CPT": [],
  "HCPCS": [],
  "Modifiers": []
}
```

---

## Constraints

-  Do **not** use LLMs ( Prompt Engineering)(e.g., OpenAI, GPT-4, Claude, Mistral, etc.)