This assignment is quite open-ended: you will build a solution based on reading comprehension QA to extract specific information from medical transcriptions written by attending physicians. We are primarily interested in the following information:

- How old is the patient?
- Does the patient have any complaints?
- What is the reason for this consultation?
- Gender of Patient?

However, you are encouraged to extend this list and extract any other type of information that you find valuable. You may use pre-tuned general purpose QA models like RoBERT-QA or BioBERT pre-tuned on SQuAD or any LLM api, hugging face, haystack, langchain etc. However, your final solution must satisfy 2 core requirements: (1) Exact match > 55% across all information types; and (2) F1 > 55% across all information types.

The complete dataset(without annotations/ground truth), smaller annotated train and test dataset can be found on this link :

Assignment

You should work with smaller annotated json data to train your model and test it on the test.json or if you are ambitious enough you can try to manually annotate some part from the complete data's transcriptions (you will have to add questions by yourself in annotators).

You are free to use anything available on the internet, including GPT etc.
Submissions will be evaluated based on uniqueness and accuracy scores.
You Should be submitting a colab notebook by 10th EOD.