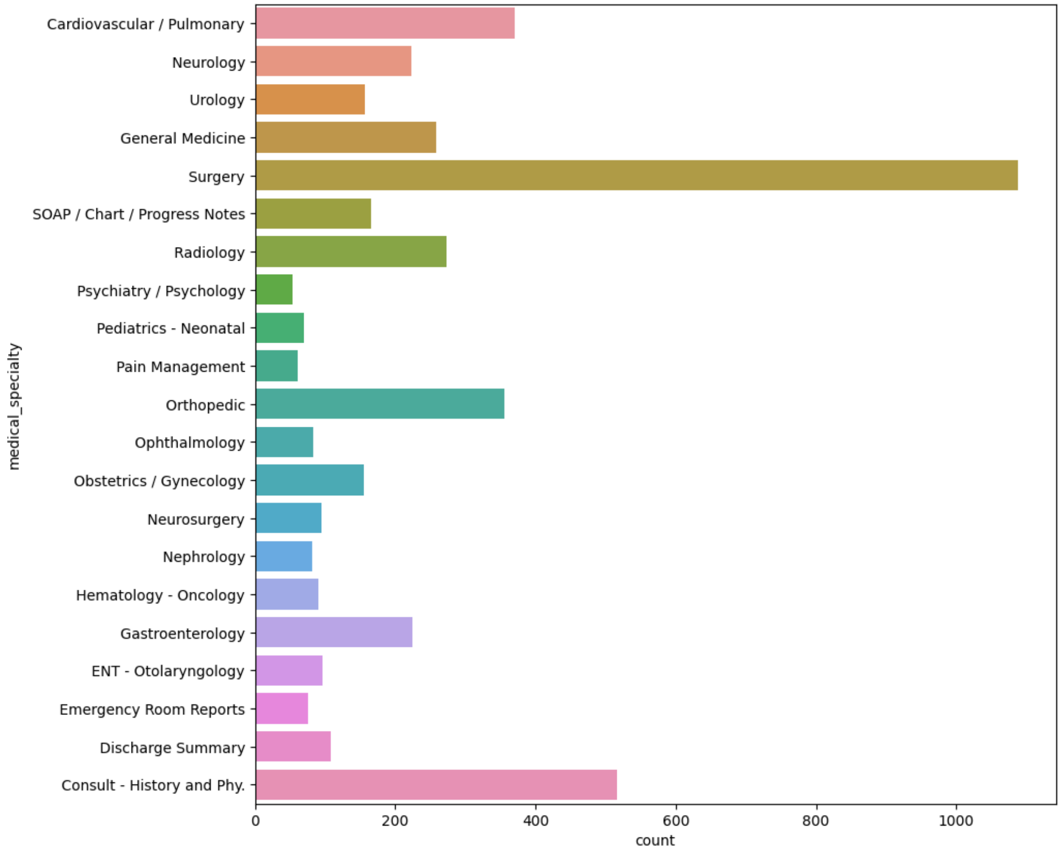
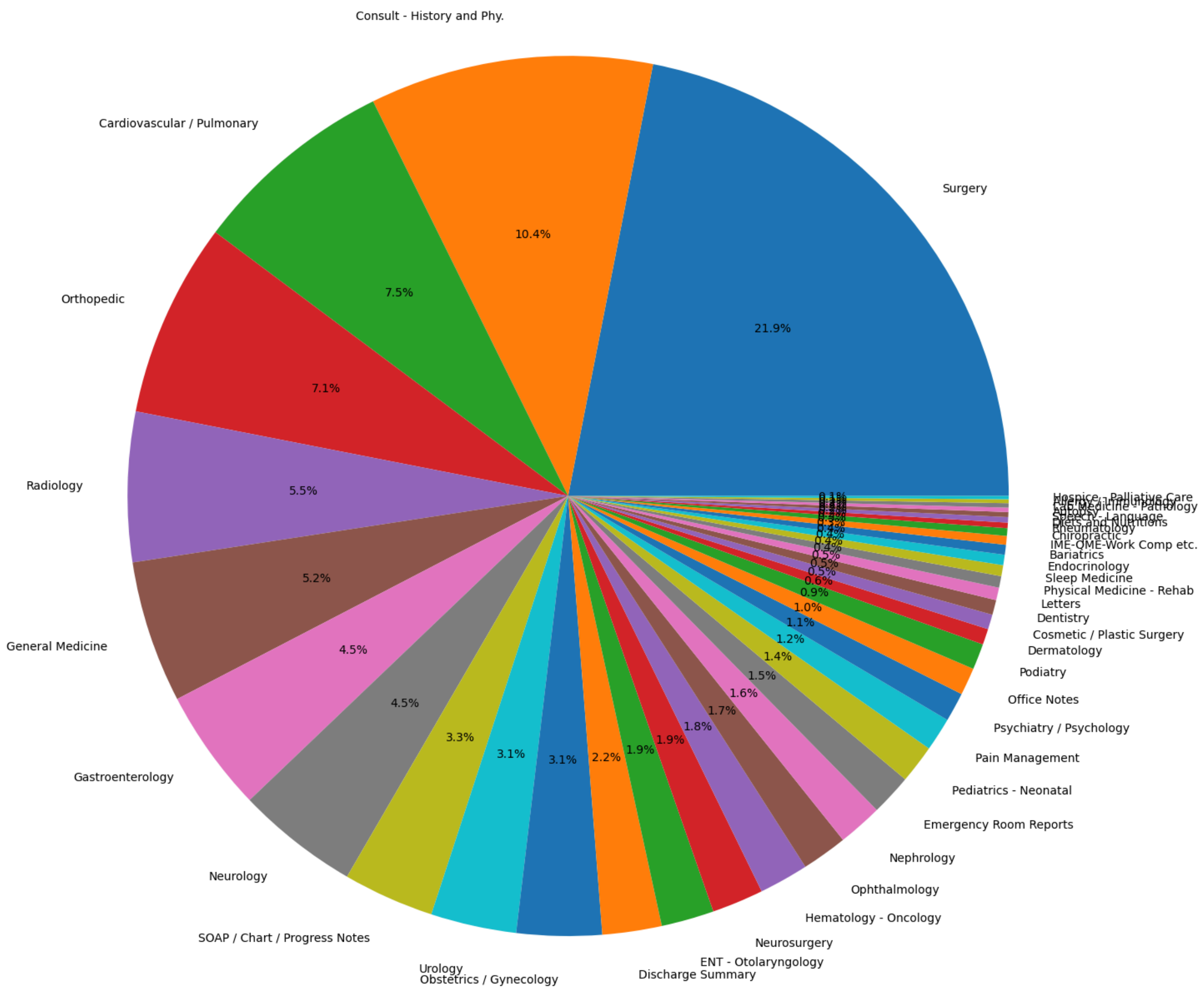
Assignment : Classification

- I have followed an LLM based approach.
- I have played with various BERT based models and tried to generate inference using zero shot, many shot & model finetuning.
- Models used → DistilBert, BioBert, BioClinicalBert & bert-based-uncased
- There are a total of 40 classes.
- Another approach which I thought of was using Sentence Transformers to generate embedding for the transcription, we can then use the train set. Now for each sentence in the test set we will again generate the embedding and use KNN method to determine the medical_speciality for the given test transcription.
- Classical ML based algos can also be applied but given the explosion of LLMs and their vast understanding of the world, I have tried this approach.
- Statistical approaches such as co-occurrence matrices can be used. We could further apply SVD to generate dense representations. But this approach however wil not be feasible due to large vocab size.
- We can also merge the label set i.e. medical_speciality, but it will rely on the medical domain knowledge.

Data Exploration:




Consult - History and Phy.
Cardiovascular / Pulmonary
Surgery
10.4%
7.5%
Orthopedic
21.9%
7.1%
Radiology
5.5%
5.2%
General Medicine
4.5%
Gastroenterology
4.5%
3.3%
3.1%
3.1%
2.2%
1.9%
1.9%
1.8%
1.7%
1.6%
1.5%
1.4%
1.2%
1.1%
1.0%
0.9%
0.6%
0.5%
0.5%
0.5%
0.4%
0.4%
0.4%
0.4%
0.1%
Hospice - Palliative Care
Allergy / Immunology
Lab Medicine - Pathology
Autopsy
Speech - Language
Diets and Nutritions
Rheumatology
Chiropractic
IME-QME-Work Comp etc.
Bariatrics
Endocrinology
Sleep Medicine
Physical Medicine - Rehab
Letters
Dentistry
Cosmetic / Plastic Surgery
Dermatology
Podiatry
Office Notes
Psychiatry / Psychology
Pain Management
Pediatrics - Neonatal
Emergency Room Reports
Nephrology
Ophthalmology
Hematology - Oncology
Neurosurgery
ENT - Otolaryngology
Discharge Summary
Neurology
SOAP / Chart / Progress Notes
Urology
Obstetrics / Gynecology