Amber Healthcare

Take Home Assignment

Prediction Model

1. Why LLM based approach?

   Size of the dataset

2. Why GPT-4o?

   Best in the class for medical text understanding

3. Prompt Engineering Strategies -

   Input Prompt - chain of thought reasoning

   Output - Structured prompt

4. Evaluation metrics -

   Precision, Recall, F1

Structure of the Prompt

[Case note]: {medical_transcript}

[Task]: Medical coding specialist AI...

[Instructions]:

- Identify symptoms, findings, assessments

- Maximum 5 highest probability conditions

- Use ICD-10 hierarchy

[Output Format]:

- ICD-10 Code: [specific code]

- Description: [condition name]

- Evidence from Transcript: [supporting quotes]

- Probability: [numerical confidence]

- Confidence Level: [High/Medium/Low]

Experimentation

	Precision	*Recall*	*F1*
GPT 3.5 turbo - zero shot prompt	0.11	0.19	0.14
GPT 3.5 turbo - prompt w reasoning	0.16	0.21	0.18
GPT 4o - prompt w reasoning	**0.23**	**0.32**	**0.27**
GPT 4o - few shot prompt	0.18	0.25	0.21
O3-mini - prompt w reasoning	0.0	0.0	0.0

Uncertainty Estimation & Model Calibration

What is uncertainty estimation?

Quantifying how confident an ML model is about its prediction

What is model calibration?

Measures whether a model’s prediction matches its reality.

Why Monte carlo estimation?

Easy to implement, popular in the medical domain

What did I try for model calibration?

Temperature, top_p

Uncertainty metrics & Calibration parameters

Code consistency - frequency across samples

Confidence Score - average consistency

Reliable codes - above >50% consistency

Risk level - high/medium/low

Temperature - scales the distribution to make it more or less confident. 0 is more confident, and 2 is less confident

top_p - controls the no of tokens considered during generation

Live Demo & Design Explanation

Future Works

LLM guided tree-search - As explained in https://openreview.net/forum?id=mqnR8rGWkn

Alternative evaluation method - distance based on the hierarchical structure of ICD10 codes

Manual Analysis of the results - each of the codes, their accuracy metrics and confidence

Post training of LLM - Could be better at handling contradicting sentences in conversations; also better at correlations between medical conditions

ClinicalBERT/BioBERT - Would be good for benchmarking and for ensemble