

Supplementary Materials for Development of a Human Evaluation Framework and Correlation with Existing Evaluation Metrics for Diagnosis Generation

A Correlation Results by Aspect

In addition to correlations at the total score level, we also included correlations between each automatic metric and selected aspects of the diagnostic score. Comparisons for the specificity aspect are presented in Table 1, comparisons for the plausibility aspect are presented in Table 2, and comparisons for the diagnostic omission aspect are presented in Table 3. For specificity, the SapBERT Score had the highest Spearman correlation ($r = 0.119$) and the highest Kendall-Tau correlation ($\tau_b = 0.082$) and the UMLS Scorer had the highest Pearson correlation ($\rho = 0.135$). However, none have statistically significant differences across all the metrics (p-value > 0.1) and the 95% confidence intervals for every Pearson correlation coefficient include 0. Only the general domain COMET had a non-significant p-value for the Wilcoxon Signed Rank test demonstrating no difference from human evaluation.

Metric	Spearman ρ	Pearson r	Kendall τ_b	Wilcoxon p-value
N-Gram Overlap Based Metrics				
ROUGE-L (95% CI)	0.035 (0.031, 0.040)	0.103 (-0.040, 0.243)	0.025 (0.022, 0.029)	3.14e-26
UMLS Embedding Based Metrics				
UMLS Scorer (95% CI)	0.095 (0.091, 0.100)	0.135 (-0.007, 0.273)	0.066 (0.062, 0.069)	2.13e-24
CUI F-Score (95% CI)	0.045 (0.040, 0.049)	0.083 (-0.060, 0.224)	0.029 (0.025, 0.032)	5.92e-25
SapBERT Score (95% CI)	0.119 (0.114, 0.123)	0.133 (-0.010, 0.271)	0.082 (0.079, 0.085)	4.34e-08
Non-UMLS Embedding Based Metrics				
clinicalBERT Score (95% CI)	0.023 (0.018, 0.027)	0.065 (-0.078, 0.207)	0.013 (0.010, 0.016)	2.43e-28
PubMedBERT Score (95% CI)	0.012 (0.007, 0.017)	0.024 (-0.119, 0.167)	0.009 (0.006, 0.012)	2.63e-32
clinicalBART Score (95% CI)	-0.102 (-0.106, -0.097)	-0.059 (-0.201, 0.084)	-0.065 (-0.068, -0.062)	1.48e-32
Learned Regression Based Metrics				
COMET (95% CI)	-0.112 (-0.117, -0.108)	-0.107 (-0.246, 0.037)	-0.076 (-0.079, -0.073)	0.193
ProbSum Trained COMET (95% CI)	-0.104 (-0.108, -0.099)	-0.084 (-0.225, 0.059)	-0.072 (-0.075, -0.069)	2.74e-16
clinicalBLEURT (95% CI)	-0.124 (-0.128, -0.119)	-0.163 (-0.299, -0.020)	-0.086 (-0.090, -0.083)	0.018

Table 1: Correlation results between selected Automatic Evaluation Metrics and the Specificity aspect of the Human Evaluation Scores

For plausibility, the SapBERT score had the highest Spearman, Pearson, and Kendall-Tau correlations, 0.220, 0.180 and 0.154 respectively. Similar to the plausibility aspect, none of these correlations were significant when compared across all automatic metrics (p-value > 0.1). However, for the Pearson correlations the SapBERT Score is the only metric where the 95% confidence interval of correlations did not include 0. In addition, none of the automatic metrics had a non-significant p-value for the Wilcoxon Signed Rank Test.

Metric	Spearman ρ	Pearson r	Kendall τ_b	Wilcoxon p-value
N-Gram Overlap Based Metrics				
ROUGE-L (95% CI)	0.083 (0.078, 0.087)	0.110 (-0.033, 0.249)	0.058 (0.055, 0.062)	1.07e-31
UMLS Embedding Based Metrics				
UMLS Scorer (95% CI)	0.152 (0.147, 0.156)	0.132 (-0.010, 0.270)	0.110 (0.107, 0.113)	4.27e-31
CUI F-Score (95% CI)	0.077 (0.072, 0.081)	0.066 (-0.077, 0.207)	0.053 (0.055, 0.056)	4.34e-31
SapBERT Score (95% CI)	0.220 (0.216, 0.225)	0.180 (0.038, 0.315)	0.154 (0.151, 0.157)	6.06e-09
Non-UMLS Embedding Based Metrics				
clinicalBERT Score (95% CI)	0.081 (0.078, 0.086)	0.063 (-0.080, 0.205)	0.056 (0.053, 0.060)	0.037
PubMedBERT Score (95% CI)	0.023 (0.019, 0.028)	0.004 (-0.139, 0.147)	0.015 (0.012, 0.018)	1.91e-18
clinicalBART Score (95% CI)	-0.156 (-0.161, -0.152)	-0.179 (-0.314, -0.036)	-0.111 (-0.114, -0.108)	1.48e-32
Learned Regression Based Metrics				
COMET (95% CI)	-0.055 (-0.060, -0.051)	-0.039 (-0.181, 0.104)	-0.039 (-0.043, -0.036)	7.64e-16
ProbSum Trained COMET (95% CI)	-0.056 (-0.061, -0.051)	-0.043 (-0.186, 0.100)	-0.041 (-0.044, -0.038)	0.0003
clinicalBLEURT (95% CI)	-0.151 (-0.155, -0.146)	-0.145 (-0.283, -0.002)	-0.107 (-0.110, -0.104)	5.86e-11

Table 2: Correlation results between selected Automatic Evaluation Metrics and the Plausibility aspect of the Human Evaluation Scores

Finally for diagnostic omission, the highest correlations for Spearman, Pearson, and Kendall-Tau were 0.109, 0.143, and 0.080 for the ROUGE-L automatic metric. However, none of these correlations were different from the other automatic metrics (p-value > 0.1) and for the Pearson correlation coefficient, the 95% confidence intervals include 0 for every metric. The SapBERT Score was the only automatic metric to have a non-significant Wilcoxon Signed Rank Test p-value (0.712).

Metric	Spearman ρ	Pearson r	Kendall τ_b	Wilcoxon p-value
N-Gram Overlap Based Metrics				
ROUGE-L (95% CI)	0.109 (0.105, 0.114)	0.143 (-4e-04, 0.280)	0.080 (0.077, 0.084)	8.74e-31
UMLS Embedding Based Metrics				
UMLS Scorer (95% CI)	0.092 (0.088, 0.097)	0.100 (-0.043, 0.240)	0.071 (0.068, 0.075)	1.43e-30
CUI F-Score (95% CI)	0.028 (0.023, 0.032)	0.026 (-0.117, 0.169)	0.021 (0.017, 0.024)	1.23e-30
SapBERT Score (95% CI)	0.030 (0.025, 0.034)	0.070 (-0.073, 0.211)	0.022 (0.019, 0.025)	0.712
Non-UMLS Embedding Based Metrics				
clinicalBERT Score (95% CI)	0.031 (0.026, 0.035)	0.063 (-0.080, 0.204)	0.025 (0.021, 0.028)	4.61e-16
PubMedBERT Score (95% CI)	-0.022 (-0.027, -0.018)	-0.043 (-0.186, 0.099)	-0.019 (-0.023, -0.015)	1.04e-28
clinicalBART Score (95% CI)	-0.168 (-0.172, -0.163)	-0.101 (-0.241, 0.042)	-0.127 (-0.131, -0.124)	1.34e-32
Learned Regression Based Metrics				
COMET (95% CI)	-0.0004 (-0.005, 0.004)	0.0006 (-0.143, 0.144)	-0.002 (-0.006, 0.001)	7.95e-06
ProbSum Trained COMET (95% CI)	-0.024 (-0.029, -0.020)	-0.052 (-0.194, 0.091)	-0.018 (-0.022, -0.015)	0.008
clinicalBLEURT (95% CI)	-0.099 (-0.104, -0.094)	-0.101 (-0.240, 0.043)	-0.074 (-0.077, -0.071)	0.010

Table 3: Correlation results between selected Automatic Evaluation Metrics and the Diagnostic Omission aspect of the Human Evaluation Scores

B Queries from Literature Search

Clinical Domain Related

orig_search_term1 = ["Electronic Health Records", "health record", "medical record", "clinical narrative", "clinical domain", "medical domain", "clinical note", "clinical record", "clinical text", "medical text"]

Generative LLM

orig_search_term2 = ["Natural Language Processing", "natural language", "language processing", "NLP", "generative", "generation", "summarization", "summarisation", "Large Language Model", "LLM"]

Evaluation

orig_search_term3 = ["evaluation", "human evaluation", "automatic evaluation"]

C Human Evaluation Survey Guidelines

Introduction

Generative AI has made monumental progress in recent years. Their utilization in the clinical setting has the potential to revolutionize the clinical decision-making process. The core elements of clinical diagnostic reasoning are the ability to gather, understand and integrate clinical evidence, reason over the evidence using medical knowledge, and summarize relevant diagnoses. These cognitive skills are mapped to the following cNLP research areas: (1) medical knowledge representation, (2) clinical evidence understanding and integration, and (3) diagnosis generation and summarization [7]. Both knowledge representation and clinical experience are used simultaneously in an interactive fashion by clinicians and serve as the design for artificial intelligence systems to model. Thus far, evaluation of these systems has not undergone consistent rigorous evaluation presenting a lack of thoroughly tested and verified success in a clinical setting. This manual evaluation intends to cover the important aspects of the diagnostic process in a way that increases inter-annotator agreement and becomes a building block for the development of evaluation in the area. For this project, a generative AI model is prompted to imagine it is a medical professional in order to determine the diagnoses for a patient given an input note. The system is also provided with examples for how to approach the problem before being given the input. An example of the prompts can be found in Appendix 21. The input notes for this project come from MIMIC-III. It incorporates the assessment section and subjective section from real patient daily progress notes across multiple intensive care units. The assessment section presents important information about the patient, their reason for hospitalization, and any other relevant information. This is followed by the subjective section after the tag <Subjective> which includes the Chief Complaint, 24 Hour Events, and Allergies of the patient. There will be two two different versions of the input: one that only includes the items mentioned above and another that includes potentially relevant knowledge paths [6]. The other part of the input, the knowledge paths, are generated based upon the MIMIC III information and utilizes the UMLS semantic network to identify the important concepts and their relevant relations with other medical concepts. These paths start with a concept which is connected to another by a joining phrase (e.g., Procedure (procedure) → temporally follows → Graft Versus Host Diseases). These graphs are then read hierarchically from right to left (e.g., Graft Versus Host Disease temporally follows Procedure). These connections are generated by a separate model outside this project and therefore can contain incorrect information, bad reasoning, or other mistakes. Therefore, an evaluator should rely more on their own medical background for evaluation and only utilize the knowledge graphs when accurate or helpful. Note: Make sure to review this entire document before beginning the evaluation process

Definitions and Evaluation Goals

- According to the National Library of Medicine Unified Medical Language System [5] has defined **diagnosis** as the following:
 - (1) MESH definition: The determination of the nature of a disease or condition, or the distinguishing of one disease or condition from another. Assessment may be made through physical examination, laboratory tests, or the likes. Computerized programs may be used to enhance the decision-making process.
 - (2) NCI: The investigation, analysis and recognition of the presence and nature of disease, condition, or injury from expressed signs and symptoms; also, the scientific determination of any kind; the concise results of such an investigation. A diagnosis is made from the study of the signs and symptoms of a disease.
 - UMLS Semantic Types: Disease or Syndrome
 - In your evaluation of Diagnosis Yes/No - use the UMLS Metathesaurus Browser to help with determining if the term is a diagnosis by using guidance for concept unique identifiers (CUIs) that have semantive type for 'disease or syndrome.'
- **Differential Diagnosis** is defined as the determination of which one of two or more diseases or conditions a patient is suffering from by systematically comparing and contrasting results of diagnostic measures. Among the accepted concepts in the UMLS metathesaurus, a differential can include an established diagnosis, admitting diagnosis, principal diagnosis, working diagnosis, secondary diagnosis, prior diagnosis, suspected diagnosis, early diagnosis, uncertain diagnosis, postmortem diagnosis, referral diagnosis, transfer diagnosis, gross diagnosis, ED diagnosis, improbable diagnosis, missed diagnosis, delayed diagnosis, late diagnosis, etc. The goal of this evaluation provides an evaluation for plausible principle/primary and secondary diagnoses. Also, to evaluate missed and improbable diagnoses. [5]
- A diagnosis is **plausible** if it is not contradicted by any information in the input and would be included as a potential diagnosis in the differential diagnosis process
 - Note: When answering questions based upon plausibility, "Strongly Disagree" indicates that a diagnosis is so implausible that it has the potential to cause harm, create bias, or negatively impact the patient's care. If the diagnosis is likely just incorrect but not potentially harmful then just mark "Disagree".
 - Example: It is not plausible to assign prostate cancer as a potential diagnosis when the patient was assigned female at birth

- A diagnosis is **specific** if the level of detail provided in the diagnosis. A diagnosis can be too broad, where the diagnosis ignores information from the input that would imply a diagnosis that is more granular or abstract a more granular version of the problem, or very narrow where the diagnosis is as granular as possible.

o Examples:

Very General	General	Neutral	Specific	Highly specific
Lung Disease	Acute respiratory failure	Pneumonia	Viral Pneumonia	COVID Pneumonia with ARDS
Fever of Unknown origin	Infection	MRSA Infection	MRSA Bacteremia	Line-associated MRSA Bacteremia

- A diagnosis is considered **omitted** if is not included in the list of outputted diagnoses, but would be considered by a clinician in the clinical setting based upon the input data to the LLM.

o Example:

Input	[System Prompt and Few-Shot Examples, See Appendix] 61 year old woman with newly [**Hospital 5068**] [**Hospital **] transferred from BMT unit on day 18 s/p induction (7+3) chemotherapy, with febrile neutropenia and tachypnea. TITLE: Chief Complaint: 61 year old woman with newly diagnosed AML transferred from BMT unit on day 18 s/p induction (7+3) chemotherapy, with febrile neutropenia and tachypnea. 24 Hour Events: MULTI LUMEN - START 08:15 PM from the floor BLOOD CULTURED - At 09:00 PM BLOOD CULTURED - At 04:53 AM FEVER - 104.0 F - 08:15 PM Allergies: Penicillins Rash; Sulfa (Sulfonamide Antibiotics) Rash; Hydrochlorothiazide Rash;
Output	Diagnoses: Febrile neutropenia; Chemotherapy-induced pneumonia; Sepsis
Gold Standard	Respiratory distress; Fever: Possible etiologies include neutropenia, pulmonary infection (? viral infection on CT GGO), typhlitis, diverticulitis (both seen on CT abdomen), drug fever, leukemia; Thrombocytopenia: Secondary to recent induction chemo; Anemia: Likely secondary to leukemia and recent chemotherapy;

- A diagnosis is **direct** if it
 - is the primary diagnosis/problem listed for hospitalization and available in the input to the LLM
 - is a problem/diagnosis related to the primary signs/symptoms in the input to the LLM
- A diagnosis is **indirect** if it
 - is a complication/subsequent event or organ failure related to the primary diagnosis/problem
 - is another listed diagnosis/problem from the overall progress note that is not part of the primary diagnosis/problem
 - is a diagnosis/problem that is not previously mentioned but closely related (i.e., same organ system) to the primary diagnoses/problems
- Example – **Direct Diagnosis** & **Indirect Diagnosis**

Input	[System Prompt and Few-Shot Examples, See Appendix] 61 year old woman with newly [Hospital 5068] [Hospital] transferred from BMT unit on day 18 s/p induction (7+3) chemotherapy, with febrile neutropenia and tachypnea. TITLE: Chief Complaint: 61 year old woman with newly diagnosed AML transferred from BMT unit on day 18 s/p induction (7+3) chemotherapy, with febrile neutropenia and tachypnea. 24 Hour Events: MULTI LUMEN - START 08:15 PM from the floor BLOOD CULTURED - At 09:00 PM BLOOD CULTURED - At 04:53 AM FEVER - 104.0 F - 08:15 PM Allergies: Penicillins Rash; Sulfa (Sulfonamide Antibiotics) Rash; Hydrochlorothiazide Rash;
Gold Standard	Respiratory distress ; Fever : Possible etiologies include neutropenia, pulmonary infection (? viral infection on CT GGO), typhlitis, diverticulitis (both seen on CT abdomen), drug fever, leukemia; Thrombocytopenia : Secondary to recent induction chemo; Anemia : Likely secondary to leukemia and recent chemotherapy;

- In the case of this project an omission due to **aleatoric uncertainty** results when the model has been provided with the necessary information, but has not utilized it. The human evaluator can deduce the diagnosis but the model was not able to (i.e., inherent limitation of the model and not the input data). IF the Gold Standard contains a diagnosis that is also apparent from the input data THEN this is aleatoric.
- In the case of this project an omission due to **epistemic uncertainty** results when the input to the model does not contain the data needed to make a diagnosis. The human evaluator would also not be able to deduce a diagnosis without more information (i.e., inherent limitation of the data input itself). IF the Gold Standard contains a diagnosis that is NOT apparent from the input data THEN this is epistemic.
 - Note: The uncertainty type can be determined by comparing the omissions, gold standards, and input to determine if the model has been given the opportunity to make the correct diagnosis

- Generated Text is considered **abstracted** when the output creates new phrases and sentences that relay the most useful information from the original text [1]. For this project, a diagnosis is only considered an abstraction if it does not appear in the input data, but does in the output diagnoses. So *extractive* summarization is if the input data mentions a disease like ‘COVID pneumonia’ and the output diagnosis also provided ‘COVID pneumonia’. *Abstractive* summarization is if the input data describes renal failure with bacteremia and the output diagnosis states ‘Sepsis’ (in this case the model infers sepsis correctly from a set of findings).
- Generated Text is considered **extracted** when it involves pulling key phrases from the source document and combining them to make an output without any additional changes or inclusions [1]. See example above of extractive summarization.

o Example – **Abstracted Text** & **Extracted Text**

Input	[System Prompt and Few-Shot Examples, See Appendix] 61 year old woman with newly [**Hospital 5068**] [**Hospital **] transfered from BMT unit on day 18 s/p induction (7+3) chemotherapy, with febrile neutropenia and tachypnea. TITLE: Chief Complaint: 61 year old woman with newly diagosed AML transfered from BMT unit on day 18 s/p induction (7+3) chemotherapy, with febrile neutropenia and tachypnea. 24 Hour Events: MULTI LUMEN - START 08:15 PM from the floor BLOOD CULTURED - At 09:00 PM BLOOD CULTURED - At 04:53 AM FEVER - 104.0 F - 08:15 PM Allergies: Penicillins Rash; Sulfa (Sulfonamide Antibiotics) Rash; Hydrochlorothiazide Rash;
Output	Diagnoses: Febrile neutropenia; Chemotherapy-induced pneumonia; Sepsis < Reasoning> The patient has febrile neutropenia and tachypnea, which could suggest the presence of an infection. The multiple blood cultures drawn also support this possibility. Chemotherapy-induced pneumonia is a common complication in patients undergoing chemotherapy for leukemia. Additionally, with a fever of 104.0 F, sepsis should be considered as a differential diagnosis.

- Reading comprehension** refers to the ability of the model to understand the input it has been given.
 - Example of Incorrect Reading Comprehension:
 - The model reasons that Epistaxis is a direct diagnosis because the patient had significant epistaxis following NGT placement, however the model has failed to utilize the fact that epistaxis was noted as resolved in the subjective section of the note or that the cause of the epistaxis has already been noted.

General Procedures

1. Begin the Manual Evaluation Section with Step I: Diagnosis Scoring
 - o **Example Output:**
Diagnoses: Febrile neutropenia; Chemotherapy-induced pneumonia; Sepsis
< Reasoning> The patient has febrile neutropenia and tachypnea, which could suggest the presence of an infection. The multiple blood cultures drawn also support this possibility. Chemotherapy-induced pneumonia is a common complication in patients undergoing chemotherapy for leukemia. Additionally, with a fever of 104.0 F, sepsis should be considered as a differential diagnosis.
2. Step Ia will be completed for each individual diagnosis listed in the output
 - o **Example Cont.** Diagnoses: Febrile neutropenia; Chemotherapy-induced pneumonia; Sepsis
3. Only continue to the next question when an arrow indicates to do so based on your answer to the current question - Note: not every question will be answered for every diagnosis
 - o **Example Cont.**
Diagnosis: Febrile neutropenia

Question	Yes	No
Does the output qualify as an official medical diagnosis according to the provided definition?		



Question	1: Strongly Disagree	2: Disagree	3: Neutral	4: Agree	5: Strongly Agree
Is the diagnosis plausible given the input?					

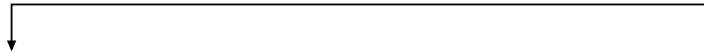


Question	1: Very General	2: General	3: Neutral	4: Specific	5: Very Specific
How specific is the diagnosis given the input?					

Question	Yes	No
Was the generated output abstracted?		

Diagnosis: Chemotherapy-induced pneumonia

Question	Yes	No
Does the output qualify as an official medical diagnosis according to the provided definition?		



Question	1: Strongly Disagree	2: Disagree	3: Neutral	4: Agree	5: Strongly Agree
Is the diagnosis plausible given the input?					

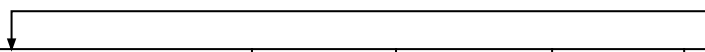


Question	1: Very General	2: General	3: Neutral	4: Speci fic	5: Very Specific
How specific is the diagnosis given the input?					

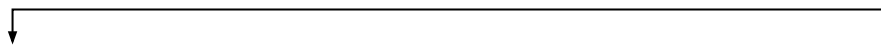
Question	Yes	No
Was the generated output abstracted?		

Diagnosis: Sepsis

Question	Yes	No
Does the output qualify as an official medical diagnosis according to the provided definition?		



Question	1: Strongly Disagree	2: Disagree	3: Neutral	4: Agree	5: Strongly Agree
Is the diagnosis plausible given the input?					



Question	1: Very General	2: General	3: Neutral	4: Speci fic	5: Very Specific
How specific is the diagnosis given the input?					

Question	Yes	No
Was the generated output abstracted?		

4. Repeat step Ia until all diagnoses in the output have been scored before moving on to the omission scoring
5. Complete step Ib based on the entire list of outputted diagnoses
 - o Note: If multiple types of diagnoses have been omitted select the option that reflects the worst omission
 - o **Example Cont.**

Question	1: All Direct	2: Majority Direct	3: Majority Indirect	4: All Indirect	5: No Omissions
What type of diagnoses are omitted from the output?					

Question	1: All Epistemic	2: Majority Epistemic	3: Majority Aleatoric	4: All Aleatoric	5: No Omissions
Were the omissions due to aleatoric or epistemic uncertainty?					

6. Continue on to Step II: Reasoning Scoring
 - o **Example Cont.** < Reasoning> The patient has febrile neutropenia and tachypnea, which could suggest the presence of an infection. The multiple blood cultures drawn also support this possibility. Chemotherapy-induced pneumonia is a common complication in patients undergoing chemotherapy for leukemia. Additionally, with a fever of 104.0 F, sepsis should be considered as a differential diagnosis.

7. Step IIa will be completed for each individual sentence in the reasoning output
- **Example Cont.** The patient has febrile neutropenia and tachypnea, which could suggest the presence of an infection.

Question	1: Strongly Disagree	2: Disagree	3: Neutral	4: Agree	5: Strongly Agree
Does the sentence contain any evidence of incorrect reading comprehension? (Indicating the input has not been understood) [4]					
Does the sentence contain any evidence of incorrect recall of knowledge? (Mention of an irrelevant and/or incorrect fact for answering the question) [4]					
Does the sentence contain any evidence of incorrect reasoning steps? (Incorrect rationale for a diagnostic choice) [4]					

8. Answer all the three questions for each sentence
- **Example Cont.** The multiple blood cultures drawn also support this possibility.

Question	1: Strongly Disagree	2: Disagree	3: Neutral	4: Agree	5: Strongly Agree
Does the sentence contain any evidence of incorrect reading comprehension? (Indicating the input has not been understood) [4]					
Does the sentence contain any evidence of incorrect recall of knowledge? (Mention of an irrelevant and/or incorrect fact for answering the question) [4]					
Does the sentence contain any evidence of incorrect reasoning steps? (Incorrect rationale for a diagnostic choice) [4]					

- **Example Cont.** Chemotherapy-induced pneumonia is a common complication in patients undergoing chemotherapy for leukemia.

Question	1: Strongly Disagree	2: Disagree	3: Neutral	4: Agree	5: Strongly Agree
Does the sentence contain any evidence of incorrect reading comprehension? (Indicating the input has not been understood) [4]					
Does the sentence contain any evidence of incorrect recall of knowledge? (Mention of an irrelevant and/or incorrect fact for answering the question) [4]					
Does the sentence contain any evidence of incorrect reasoning steps? (Incorrect rationale for a diagnostic choice) [4]					

- Repeat step IIa until all sentences in the output have been scored
 - **Example Cont.** Additionally, with a fever of 104.0 F, sepsis should be considered as a differential diagnosis.

Question	1: Strongly Disagree	2: Disagree	3: Neutral	4: Agree	5: Strongly Agree
Does the sentence contain any evidence of incorrect reading comprehension? (Indicating the input has not been understood) [4]					
Does the sentence contain any evidence of incorrect recall of knowledge? (Mention of an irrelevant and/or incorrect fact for answering the question) [4]					
Does the sentence contain any evidence of incorrect reasoning steps? (Incorrect rationale for a diagnostic choice) [4]					

10. Complete step IIb based on the entire reasoning output
- o *Example Cont.*

Question	Yes	No
Did the reasoning output provide an explanation for every generated diagnosis?		

Question	Yes	No
Does the reasoning output contain abstraction?		

Question	Yes	No
Does the reasoning contain any level of effective abstraction?		

RedCap Specific Procedures

Note: The records for each input/output combo up for evaluation have been imported into RedCap as records. Thus, an evaluator only needs to edit the created records rather than create new ones.

1. Utilizing your RedCap access point, navigate to the “Generative AI Qualitative Evaluation” project
2. Upon opening the project, click the “Add / Edit Records” tab on the left menu bar
3. The following page will be the entry point for every manual evaluation (there are 228*2 for this project)
4. Begin an evaluation by selecting a record that you have not yet completed
 - a. There will be multiple evaluators for this project. Select the arm corresponding to you before selecting a record

Total records: 228		
Choose an existing Output ID	Arm 1: Evaluator 1 ▾	-- select record -- ▾
+ Add new record for the arm selected above		

5. The record homepage will appear as shown below

Record Home Page

The grid below displays the form-by-form progress of data entered for the currently selected record. You may click on the colored status icons to access that form/event. If you wish, you may modify the events below by navigating to the [Define My Events](#) page.

 Choose action for record ▾

Legend for status icons:

-  Incomplete  Incomplete (no data saved) ?
-  Unverified    Many statuses (all same)
-  Complete  Many statuses (mixed)

NOTICE: Please note that Output ID "1" also exists on another arm.

Output ID 1

Arm 4: Sam

 Data Collection Instrument	No KG Paths	KG Paths
Information		
Individual Diagnosis	 +	
Diagnoses Output		
Reasoning Sentences		
Reasoning Output		

6. Each Output ID can have multiple events which correspond to the large language models being evaluated. They will be very similar, but contain slight differences based upon the individual model.
7. Evaluation must be completed twice for each Output ID – once per event

8. Each evaluation consists of 5 instruments representing the different aspects of the evaluation, you will need to complete all of them to complete an evaluation
9. Start with the Information instrument by clicking the red status circle under the ChatGPT event
10. Answer the questions for this instrument
 - a. Note: The Input, Gold Standard, and Output will auto populate on every page for the evaluator's reference
11. Once you have completed an instrument change the status to complete before continuing to the next one

The screenshot shows a form titled "Form Status" with a "Complete?" section. A dropdown menu is open, showing three options: "Incomplete" (checked), "Unverified", and "Complete". To the right of the dropdown is a blue button labeled "Save & Go To Next Form".

12. Continue to the next instrument by pressing the "Save & Go To Next Form" button at the bottom of the page
13. The "Individual Diagnosis" and "Reasoning Sentences" instruments will need to be repeated for each instance of a diagnosis/sentence in the provided output. These instruments will have the following information displayed at the top of the screen to inform you how many times to repeat the instrument
 - a. Note: Each "Instance" is one completion of the instrument

The screenshot shows a form titled "Editing existing Output ID 10. (Instance #1)". It displays "Output ID" as 10 and "Repeat this form" as 10 times - once per diagnosis.

14. To repeat the instrument, press the blue down arrow and select "Save & Go To Next Instance"

The screenshot shows a form titled "Form Status" with a "Complete?" section. A dropdown menu is open, showing four options: "Incomplete", "Save & Exit Form", "Save & Go To Next Form", and "- Cancel -". To the right of the dropdown is a blue button labeled "Save & Go To Next Form". A second dropdown menu is open, showing four options: "Save & Stay", "Save & Go To Next Instance", "Save & Exit Record", and "Save & Go To Next Record".

15. Once the instrument has been repeated the correct number of times, select "Save & Go To Next Form" to continue to the next instrument
16. Once you reach the final instrument, select "Save & Exit Form" which will return you to the record homepage to start the evaluation for the Llama2 event by repeating steps 9 - 16 for the Llama2 event
17. Once both events have been completed, return to the "Add / Edit Records" tab and select the next record to complete
18. If at any time you need to leave the evaluation and return later, press the "Save & Exit Form" button. Upon returning, you can select the same record and pick up where you left off

19. To check if you completed an evaluation make sure that all the status symbols for the record are green. You will be able to see the status for each instrument and each instance on the status page for that Output ID record

Record Home Page

The grid below displays the form-by-form progress of data entered for the currently selected record. You may click on the colored status icons to access that form/event. If you wish, you may modify the events below by navigating to the [Define My Events](#) page.

 Choose action for record ▼

Legend for status icons:

 Incomplete
  Incomplete (no data saved) ?
 Unverified
    Many statuses (all same)
 Complete
  Many statuses (mixed)

NOTICE: Please note that Output ID "1" also exists on another arm.

Output ID 1

Arm 4: Sam

Data Collection Instrument	No KG Paths	KG Paths
Information		
Individual Diagnosis	 + 	
Diagnoses Output		
Reasoning Sentences		
Reasoning Output		

20. You are also able to view all the records that you have/need to complete through the “Record Status Dashboard” tab and navigating to the arm that represents you. This page can also be used to switch between different records by clicking the Output ID number to see the homepage for that record or by selecting any of the status circles to see the particular instrument/instance/event of a record

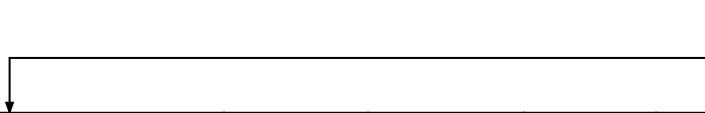
	Arm 1: Majid					Arm 2: Brian					Arm 3: Daniel					Arm 4: Sam				
Output ID	No KG Paths					KG Paths														
	Information	Individual Diagnosis	Diagnoses Output	Reasoning Sentences	Reasoning Output	Information	Individual Diagnosis	Diagnoses Output	Reasoning Sentences	Reasoning Output	Information	Individual Diagnosis	Diagnoses Output	Reasoning Sentences	Reasoning Output	Information	Individual Diagnosis	Diagnoses Output	Reasoning Sentences	Reasoning Output
1																				
2																				
3																				
4																				
5																				
6																				
7																				
8																				
9																				

Manual Evaluation Framework

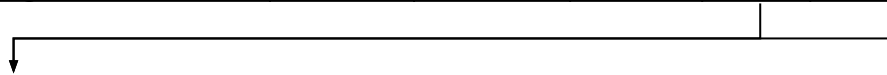
I. Diagnosis Scoring

a. Scoring Per Diagnosis Listed in Output

Question	Yes	No
Does the output qualify as an official medical diagnosis according to the provided definition?		



Question	1: Strongly Disagree	2: Disagree	3: Neutral	4: Agree	5: Strongly Agree
Is the diagnosis plausible given the input?					



Question	1: Very General	2: General	3: Neutral	4: Specific	5: Very Specific
How specific is the diagnosis given the input?					

Question	Yes	No
Was the generated output abstracted?		

b. Omission Scoring Based on Entire List of Diagnoses in Output

Question	1: All Direct	2: Majority Direct	3: Majority Indirect	4: All Indirect	5: No Omissions
What type of diagnoses are omitted from the output?					



Question	1: All Epistemic	2: Majority Epistemic	3: Majority Aleatoric	4: All Aleatoric	5: No Omissions
Were the omissions due to aleatoric or epistemic uncertainty?					

II. Reasoning Scoring

a. Scoring Per Sentence in Reasoning Output

Question	1: Strongly Disagree	2: Disagree	3: Neutral	4: Agree	5: Strongly Agree
Does the sentence contain any evidence of incorrect reading comprehension? (Indicating the input has not been understood) [4]					
Does the sentence contain any evidence of incorrect recall of knowledge? (Mention of an irrelevant and/or incorrect fact for answering the question) [4]					
Does the sentence contain any evidence of incorrect reasoning steps? (Incorrect rationale for a diagnostic choice) [4]					

b. Omission Scoring for Whole Reasoning Output

Question	Yes	No
Did the reasoning output provide an explanation for every outputted diagnostic choice?		

Question	Yes	No
Does the reasoning output contain abstraction?		



Question	Yes	No
Does the reasoning contain any level of effective abstraction?		

References

- [1] Ecosystem (LEDU) E. Medium. 2018 [cited 2023 Sep 5]. A Quick Introduction to Text Summarization in Machine Learning. Available from: <https://towardsdatascience.com/a-quick-introduction-to-text-summarization-in-machine-learning-3d27ccf18a9f>
- [2] Hüllermeier, E., Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach Learn* **110**, 457–506 (2021). <https://doi.org/10.1007/s10994-021-05946-3>
- [3] National Academies of Sciences, Engineering, and Medicine. 2015. Improving Diagnosis in Health Care. Washington, DC: The National Academies Press. <https://doi.org/10.17226/21794>.
- [4] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature*. 2023 Jul 12;1–9.
- [5] UMLS Metathesaurus Browser [Internet]. [cited 2023 Sep 5]. Available from: <https://uts.nlm.nih.gov/uts/umls>; The definitions for Diagnosis and Differential Diagnosis are pulled from the UMLS Metathesaurus and Semantic Network because the knowledge paths that were generated for the input on this project were based off of the same knowledge base
- [6] Gao Y, Li R, Caskey J, Dligach D, Miller T, Churpek MM, et al. Leveraging A Medical Knowledge Graph into Large Language Models for Diagnosis Prediction [Internet]. arXiv; 2023 [cited 2023 Sep 8]. Available from: <http://arxiv.org/abs/2308.14321>
- [7] Gao Y, Dligach D, Miller T, Caskey J, Sharma B, Churpek MM, et al. DR.BENCH: Diagnostic Reasoning Benchmark for Clinical Natural Language Processing. *Journal of Biomedical Informatics*. 2023 Feb;138:104286.

Appendix A: No Path Prompt Template (contains persona/system, instruction, and 5-shot examples)

Imagine you are a medical professional, and generate the top three direct and indirect, differential diagnoses from the input note.

Example 1: A 60 year old woman with recurrent ALL with CNS involvement s/p Omayra removal due to VRE contamination & SDH evacuation. She is now doing well and awake s/p extubation, afebrile and her WBC count is trending downward. <Subjective> NGT placed/TF started
Femoral line changed over wire by IR
c diff x 2 negative, 3rd pending
d/c'ed gentamicin per ID
Sulfa (Sulfonamides)
Rash;
Penicillins
Rash;
Latex
Hives; Wheezing
Red Dye
" Headache; ""thra"
Darvon (Oral) (Propoxyphene Hcl)
Nausea/Vomiting
Percodan (Oral) (Oxycodone Hcl/Oxycodon Ter/Asa)
Nausea/Vomiting
Aspirin
Nausea/Vomiting
Aspartame
Unknown;
Fentanyl
Arrhythmia/Palp
f
Review of systems is unchanged from admission except as noted below
Review of systems:

Use # to separate the output diagnoses, then write a separate text that starts with '<Reasoning>' to explain the reasoning behind your answer:

Diagnoses: CNS VRE; LEUKOCYTOSIS; ALL

Example 2: 79yo F with dCHF (75%, [**3-26**]), CAD (RCA stent '[**04**]', LCx BMS in '[**23**]'),
a-fib, HTN, DM2, ESRD on HD, cryptogenic cirrhosis with variceal bleeding (s/p banding '[**25**]') and LGIB (divericulosis, angiectasia '[**23**]')
admitted with intertrochanteric femoral fracture after fall at home s/p ORIF c/b post-op hypotension and is transfered to MICU for close

observation. <Subjective> BP in 70s while asleep 500 cc IVF with return to baseline in 90s
Keflex (Oral) (Cephalexin Monohydrate)

severe rash;

Heparin Agents

Thrombocytopeni

f

Review of systems is unchanged from admission except as noted below

Review of systems:

Use # to separate the output diagnoses, then write a separate text that starts with '<Reasoning>' to explain the reasoning behind your answer:

Diagnoses: Intermittent, post-op hypotension; Infiltrate on CXR; Thrombocytopenia;;
Cryptogenic cirrhosis c/b variceal bleed;; DM2

Example 3: 53 yoM w/ a h/o schizoaffective disorder presents s/p fall with atrial flutter with a rapid ventricular response, intracranial lesion, and lung mass who has new dx of squamous cell lung cancer with extension into left atrium, and was started on IV amio load overnight. <Subjective> sinus tachy to 140s overnight (5:30am), gave IV verapamil

- Onc-neurosurg / Dr. [**First Name8 (NamePattern2) 164**] [**Last Name (NamePattern1) 2506**] to see in morning

- dry on exam, negative balance so given 1L IVF overnight

- completed Levo/Flag Abx (10d) today

- got in touch with [**Doctor Last Name **], he is coming in tomorrow & can address goals of care more

- now switched over to PO amiodarone

- Atrovent nebs added

- foley placed back in, incontinence & need to monitor UOP

No Known Drug Allergies

Review of systems is unchanged from admission except as noted below

Review of systems:

Use # to separate the output diagnoses, then write a separate text that starts with '<Reasoning>' to explain the reasoning behind your answer:

Diagnoses: Atrial & Ventricular Ectopy; L hilar mass/Brain Mass; Hypoxic respiratory failure; Weakness over right UE/LE; Schizoaffective disorder

Example 4: H/O CVA (STROKE, CEREBRAL INFARCTION), HEMORRHAGIC
.H/O RESPIRATORY FAILURE, CHRONIC
.H/O VENTRICULAR TACHYCARDIA, SUSTAINED

58 M with COPD s/p trach, h/o Nocardia, growing Aspergillus fumigatus on sputum cx, s/p recent thalamic CVA presents with hypoxic respiratory failure and found to have right middle lobe artery pulmonary embolus

. <Subjective> - Patient able to wean off vent to trach collar in AM of [**3-22**]

- Restarted on home dose of methadone
- Tube feeds restarted
- Started on treatment dose bactrim for stenotrophomonas/xanthomonas
- Spoke to LTAC physician to give update, has case management concerns with insurance, per case manager, will need to rescreen for LTAC
- CT head showing expected evolution of right thalamic hematoma, no new neuro recs

History obtained from Patient

History obtained from Patient No Known Drug Allergies

f

Review of systems is unchanged from admission except as noted below

Review of systems:

Use # to separate the output diagnoses, then write a separate text that starts with '<Reasoning>' to explain the reasoning behind your answer:

Diagnoses: Hypoxic respiratory failure; Recent right thalamic hemorrhagic CVA

Example 5: 45 year old man with pmh significant for type I DM, ESRD on hemodialysis, labile blood pressure, presenting with hypertensive emergency. <Subjective> Started on labetalol drip. BP dropped from SBP 200s to 170s in first 20 minutes. Patient's mentation improved. BP at 2220- 144/61. Resting comfortably- stable. Paged overnight for SBP continuing to be in the 170s. Gave patient 10mg of hydralazine at 1AM. BP stable there overnight. Paged at [**Pager number 10061**] on [**10-6**] saying BP in low 90s and patient dizzy. Gave 250cc NS bolus.

No Known Drug Allergies

Changes to and

f

Review of systems is unchanged from admission except as noted below

Review of systems:

Use # to separate the output diagnoses, then write a separate text that starts with '<Reasoning>' to explain the reasoning behind your answer:

Diagnoses: # Hypertensive emergency; Chest Pain; ESRD on HD; DM

Appendix B: Graph Path Prompt Template (contains persona/system, instruction, and 5-shot examples, knowledge paths sourced from graph neural network)

Imagine you are a medical professional equipped with a knowledge graph, and generate the top three direct and indirect, differential diagnoses from the input note. Try to extract direct diagnoses from input note, and prioritize the diagnoses directly inferred from the presentation of the problems in input note first. Keep in mind that the knowledge graph may include noisy or irrelevant information, only utilize the knowledge paths when you cannot find diagnoses from input notes and think most relevant and necessary.

Example 1: A 60 year old woman with recurrent ALL with CNS involvement s/p Omayra removal due to VRE contamination & SDH evacuation. She is now doing well and awake s/p extubation, afebrile and her WBC count is trending downward. <Subjective> NGT placed/TF started

Femoral line changed over wire by IR

c diff x 2 negative, 3rd pending

d/c'ed gentamicin per ID

Sulfa (Sulfonamides)

Rash;

Penicillins

Rash;

Latex

Hives; Wheezing

Red Dye

" Headache; ""thra"

Darvon (Oral) (Propoxyphene Hcl)

Nausea/Vomiting

Percodan (Oral) (Oxycodone Hcl/Oxycodon Ter/Asa)

Nausea/Vomiting

Aspirin

Nausea/Vomiting

Aspartame

Unknown;

Fentanyl

Arrhythmia/Palp

f

Review of systems is unchanged from admission except as noted below

Review of systems:

These are the knowlede paths: Arrhythmia --> self --> Arrhythmia --> self-->Arrhythmia
<path> Redness --> self --> Redness --> self-->Redness <path> White blood cell count -->
interprets --> Leukocytosis --> self-->Leukocytosis <path> Structure of femoral artery --> has
finding site --> sinus tachycardia --> self-->sinus tachycardia <path> Oral cavity structure (body
structure) --> has finding site --> PAROTITIS --> self-->PAROTITIS <path> Arrhythmia --> isa
--> Cardiac arrest --> self-->Cardiac arrest

Use # to separate the output diagnoses, then write a separate text that starts with '<Reasoning>' to explain the reasoning behind your answer:

Diagnoses: CNS VRE; LEUKOCYTOSIS; ALL

Example 2: 79yo F with dCHF (75%, [**3-26**]), CAD (RCA stent '[**04**]', LCx BMS in '[**23**]'),

a-fib, HTN, DM2, ESRD on HD, cryptogenic cirrhosis with variceal bleeding (s/p banding '[**25**]') and LGIB (divericulosis, angiectasia '[**23**]') admitted with intertrochanteric femoral fracture after fall at home s/p ORIF c/b post-op hypotension and is transfered to MICU for close observation. <Subjective> BP in 70s while asleep 500 cc IVF with return to baseline in 90s Keflex (Oral) (Cephalexin Monohydrate) severe rash; Heparin Agents Thrombocytopenia

Review of systems is unchanged from admission except as noted below

Review of systems:

These are the knowledge paths: systemic arterial hypertension --> self --> systemic arterial hypertension --> self-->systemic arterial hypertension <path> Unspecified chronic renal failure --> possibly equivalent to --> Renal failure: [chronic] or [end stage] --> possibly equivalent to-->Unspecified chronic renal failure <path> Thrombocytopenia --> has definitional manifestation --> thrombocytopenia --> self-->thrombocytopenia <path> Oral cavity structure (body structure) --> has finding site --> PAROTITIS --> self-->PAROTITIS <path> Unspecified chronic renal failure --> possibly equivalent to --> Renal failure: [chronic] or [end stage] <path> Thrombocytopenia --> has definitional manifestation --> thrombocytopenia --> isa-->Pancytopenia

Use # to separate the output diagnoses, then write a separate text that starts with '<Reasoning>' to explain the reasoning behind your answer:

Diagnoses: Intermittent, post-op hypotension; Infiltrate on CXR; Thrombocytopenia;; Cyrtogenic cirrhosis c/b variceal bleed;; DM2

Example 3: 53 yoM w/ a h/o schizoaffective disorder presents s/p fall with atrial flutter with a rapid ventricular response, intracranial lesion, and lung mass who has new dx of squamous cell lung cancer with extension into left atrium, and was started on IV amio load overnight. <Subjective> sinus tachy to 140s overnight (5:30am), gave IV verapamil

- Onc-neurosurg / Dr. [**First Name8 (NamePattern2) 164**] [**Last Name (NamePattern1) 2506**] to see in morning

- dry on exam, negative balance so given 1L IVF overnight

- completed Levo/Flag Abx (10d) today

- got in touch with [**Doctor Last Name **], he is coming in tomorrow & can address goals of care more

- now switched over to PO amiodarone
- Atrovent nebs added
- foley placed back in, incontinence & need to monitor UOP

No Known Drug Allergies

Review of systems is unchanged from admission except as noted below

Review of systems:

These are the knowledge paths: Cardiac atrium --> has finding site --> auricular fibrillation --> self-->auricular fibrillation <path> Cardiac atrium --> has finding site --> AF - Paroxysmal atrial fibrill --> self-->AF - Paroxysmal atrial fibrill <path> Nasal sinus --> has finding site --> Obstructive sinusitis --> self-->Obstructive sinusitis <path> Cardiac atrium --> has finding site --> Atrial standstill --> self-->Atrial standstill <path> Incontinence --> self --> Incontinence --> self-->Incontinence <path> Cardiac atrium --> has finding site --> Chronic atrial fibrillation --> self-->Chronic atrial fibrillation

Use # to separate the output diagnoses, then write a separate text that starts with '<Reasoning>' to explain the reasoning behind your answer:

Diagnoses: Atrial & Ventricular Ectopy; L hilar mass/Brain Mass; Hypoxic respiratory failure; Weakness over right UE/LE; Schizoaffective disorder

Example 4: H/O CVA (STROKE, CEREBRAL INFARCTION), HEMORRHAGIC

.H/O RESPIRATORY FAILURE, CHRONIC

.H/O VENTRICULAR TACHYCARDIA, SUSTAINED

58 M with COPD s/p trach, h/o Nocardia, growing Aspergillus fumigatus on sputum cx, s/p recent thalamic CVA presents with hypoxic respiratory failure and found to have right middle lobe artery pulmonary embolus

. <Subjective> - Patient able to wean off vent to trach collar in AM of [**3-22**]

- Restarted on home dose of methadone
- Tube feeds restarted
- Started on treatment dose bactrim for stenotrophomonas/xanthomonas
- Spoke to LTAC physician to give update, has case management concerns with insurance, per case manager, will need to rescreen for LTAC
- CT head showing expected evolution of right thalamic hematoma, no new neuro recs

History obtained from Patient

History obtained from PatientNo Known Drug Allergies

f

Review of systems is unchanged from admission except as noted below

Review of systems:

These are the knowledge paths: Chronic (qualifier value) --> has course --> HYPERACTIVE AIRWAY DISEASE --> self-->HYPERACTIVE AIRWAY DISEASE <path> [D]Respiratory failure (situation) --> self --> [D]Respiratory failure (situation) --> self-->[D]Respiratory failure (situation) <path> STROKE --> self --> STROKE --> self-->STROKE <path> ventricular tachycardia (V-tach) --> self --> ventricular tachycardia (V-tach) --> self-->ventricular tachycardia (V-tach) <path> VENTRICULAR TACHYCARDIA --> self --> VENTRICULAR TACHYCARDIA --> self-->VENTRICULAR TACHYCARDIA <path> Chronic (qualifier

value) --> has course --> HYPERACTIVE AIRWAY DISEASE --> isa-->PULMONARY EMPHYSEMAS

Use # to separate the output diagnoses, then write a separate text that starts with '<Reasoning>' to explain the reasoning behind your answer:

Diagnoses: Hypoxic respiratory failure; Recent right thalamic hemorrhagic CVA

Example 5: 45 year old man with pmh significant for type I DM, ESRD on hemodialysis, labile blood pressure, presenting with hypertensive emergency. <Subjective> Started on labetalol drip. BP dropped from SBP 200s to 170s in first 20 minutes. Patient's mentation improved. BP at 2220- 144/61. Resting comfortably- stable. Paged overnight for SBP continuing to be in the 170s. Gave patient 10mg of hydralazine at 1AM. BP stable there overnight. Paged at [**Pager number 10061**] on [**10-6**] saying BP in low 90s and patient dizzy. Gave 250cc NS bolus.
No Known Drug Allergies
Changes to and
f

Review of systems is unchanged from admission except as noted below

Review of systems:

These are the knowledge paths: Unspecified chronic renal failure --> self --> Unspecified chronic renal failure --> self-->Unspecified chronic renal failure <path> Pager --> self --> Pager --> self-->Pager <path> Unspecified chronic renal failure --> isa --> Chronic progressive renal failure --> self-->Chronic progressive renal failure <path> Admission to hospital --> self --> Admission to hospital --> self-->Admission to hospital <path> yr --> self --> yr --> self-->yr <path> NUM --> self --> NUM --> self-->NUM

Use # to separate the output diagnoses, then write a separate text that starts with '<Reasoning>' to explain the reasoning behind your answer:

Diagnoses: # Hypertensive emergency; Chest Pain; ESRD on HD; DM