DiReCT: Diagnostic Reasoning for Clinical Notes via Large Language Models

Bowen Wang, Jiuyang Chang, Yiming Qian, Guoxin Chen, Junhao Chen, Zhouqiang Jiang, Jiahao Zhang, Yuta Nakashima, Hajime Nagahara

Osaka University

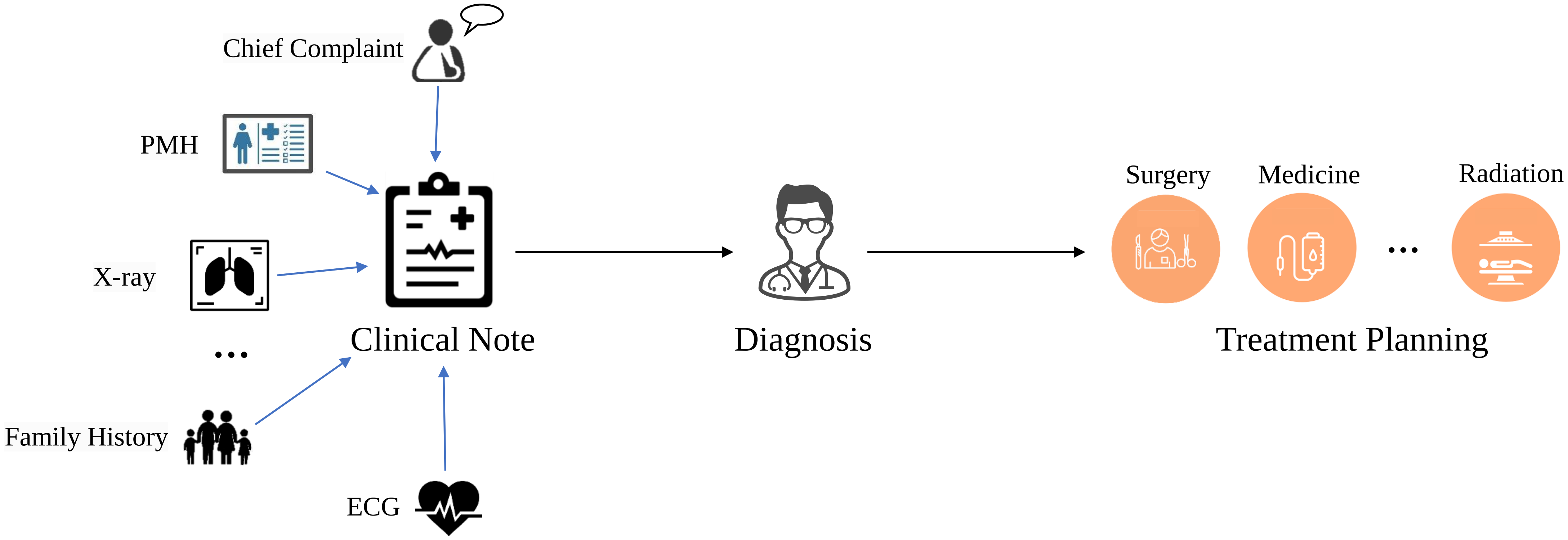
The First Affiliated Hospital of Dalian Medical University

Agency for Science, Technology and Research

Institute of Computing Technology, Chinese Academy of Science

Overview

Correct analysis of clinical records is crucial for treatment planning.

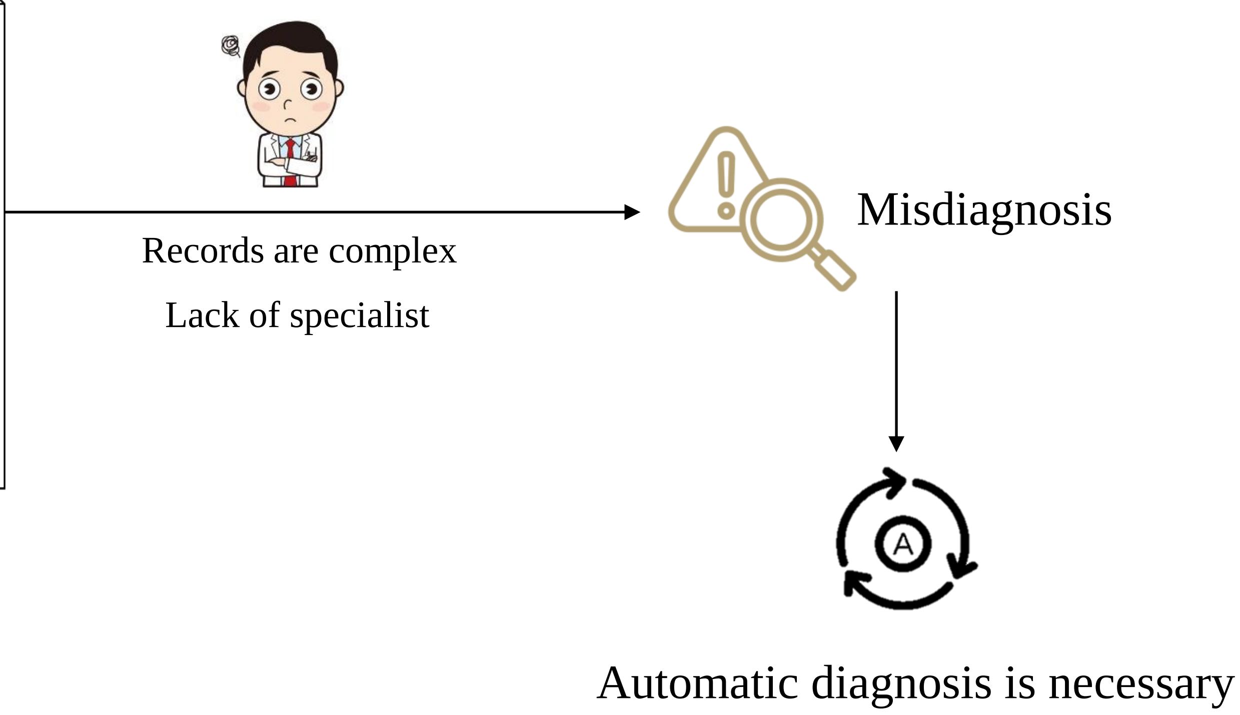


Overview

Complex and lengthy records can overwhelm doctors and increase diagnostic errors.

Chief Complaint: Scrotal and leg swelling ...
History of Present Illness: ... In the last 3 days his ***** has become quite swollen. It is similar ***** swelling when he was admitted with acute CHF ... EKG was consistent with priors (NSR, NANI, ********** changes). The left ventricle is mildly enlarged. He was given ********* with good UOP ...
Past Medical History: ... -Diabetes, -Hypertension, -CKD, stage 3, -GERD, -Depression, - Amputation of ***********, Pneumonia, - Osteoarthritis- History of *********, Asthma ...
Family History: There is no family history of **************** artery ...
Physical Exam: ... LUNG: bibasilar rales that do not clear with deep inspiration. ... ABDOMEN: nondistended, ********* all quadrants. EXTREMITIES: bilateral pitting edema to the sacrum, extending to the up abdomen. Warm, well perfused. ... HEENT: AT/NC, EOMI, PERRL. ...
Pertinent Results: __03:50PM BLOOD WBC-8.0 RBC-3.26* Hgb-9.3* Hct-30.9* MCHC-29.9* ... __ 11:30AM BLOOD proBNP-3843 ... Overall left ventricular systolic function is mildly depressed (LVEF= 45-50 %) without regional wall motion abnormalities. *********** imaging suggests an increased ******* filling pressure (PCWP>******Hg) ...

Clinical Note



Overview

Recently, large language models (LLMs) have shown their power in a variety of language tasks.

ChatGPT

INTRODUCING LLAMA -3

Gemini …

Meditron

Microsoft Copilot

Machine Translation √

Sentiment Analysis √

Question Answer √

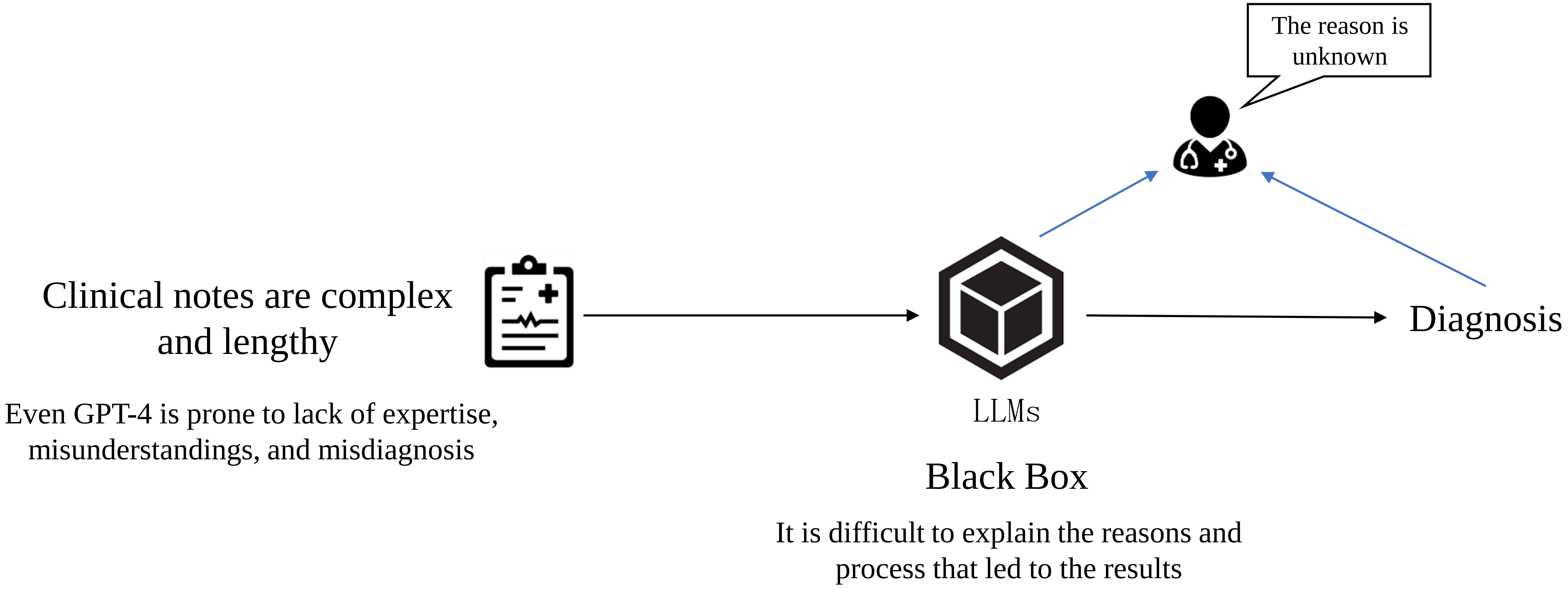
Keyword Extraction √

…

Clinical Note Analysis ?

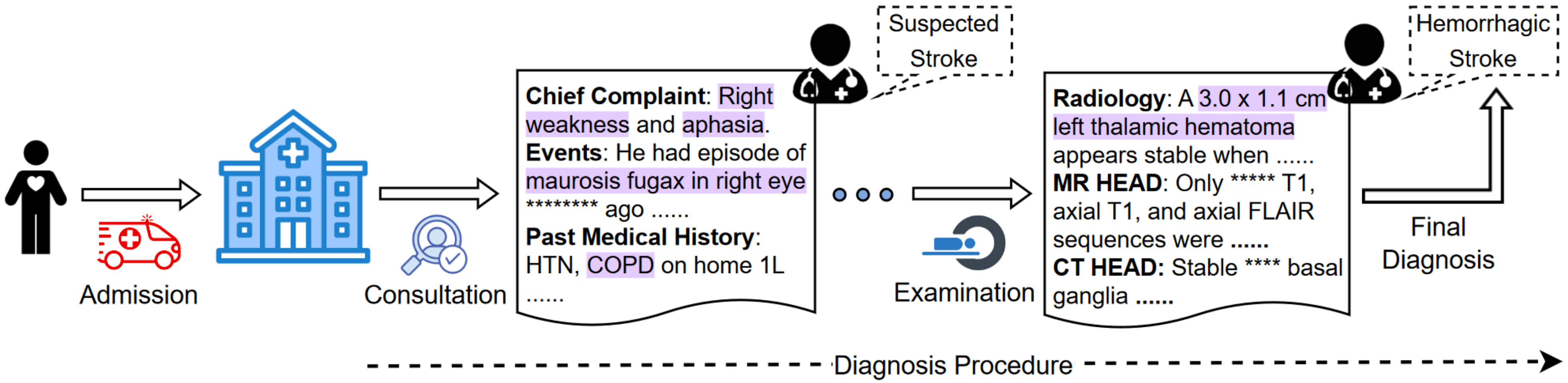
Overview

It is important that LLMs are explainable and consistent with physicians.



Overview

The diagnostic process of a human doctor follows existing diagnostic rules.



When a patient is admitted, an initial consultation takes place to collect subjective information. Subsequent observations may then require further examination to confirm the diagnosis.

Overview

An interpretable pipeline

Clinical Note → Input → Combining LLMs and knowledge graphs to achieve high accuracy → Output → Diagnosis and reasoning

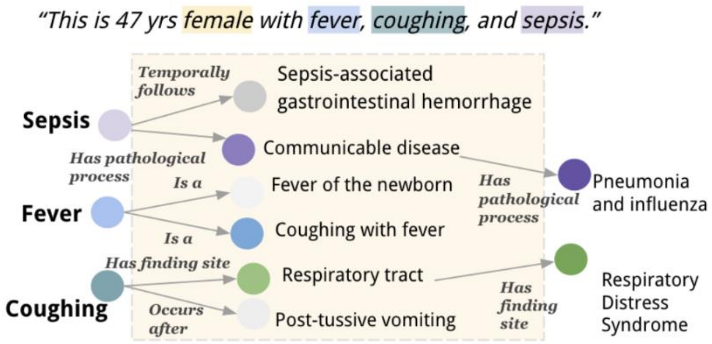
Useful for practical clinical applications

A benchmark dataset

Specailist → Construct → Benchmark → Evaluation → Pipeline

Diagnostic Knowledge Graph

UMLS Knowledge Graph (existing)



- Only provide relations for simple words (For use this KG, input has to be split into words.)
- No experimental values.

Our Diagnostic Knowledge Graph

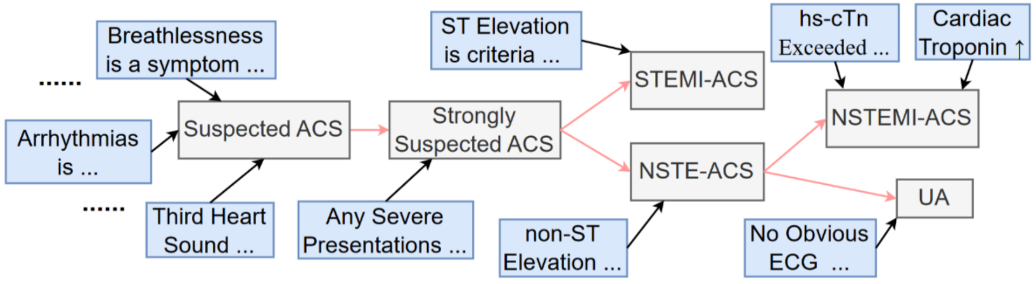


Figure 2: A part of k_i for i being `Acute Coronary Syndromes`.

$$g_i = (\mathcal{D}_i, \mathcal{F}_i)$$

$$k_i = (\mathcal{D}_i, \mathcal{P}_i, \mathcal{S}_i, \mathcal{F}_i)$$

$\mathcal{D}_i$ Diagnostic Nodes

$\mathcal{F}_i$ Procedural Edges

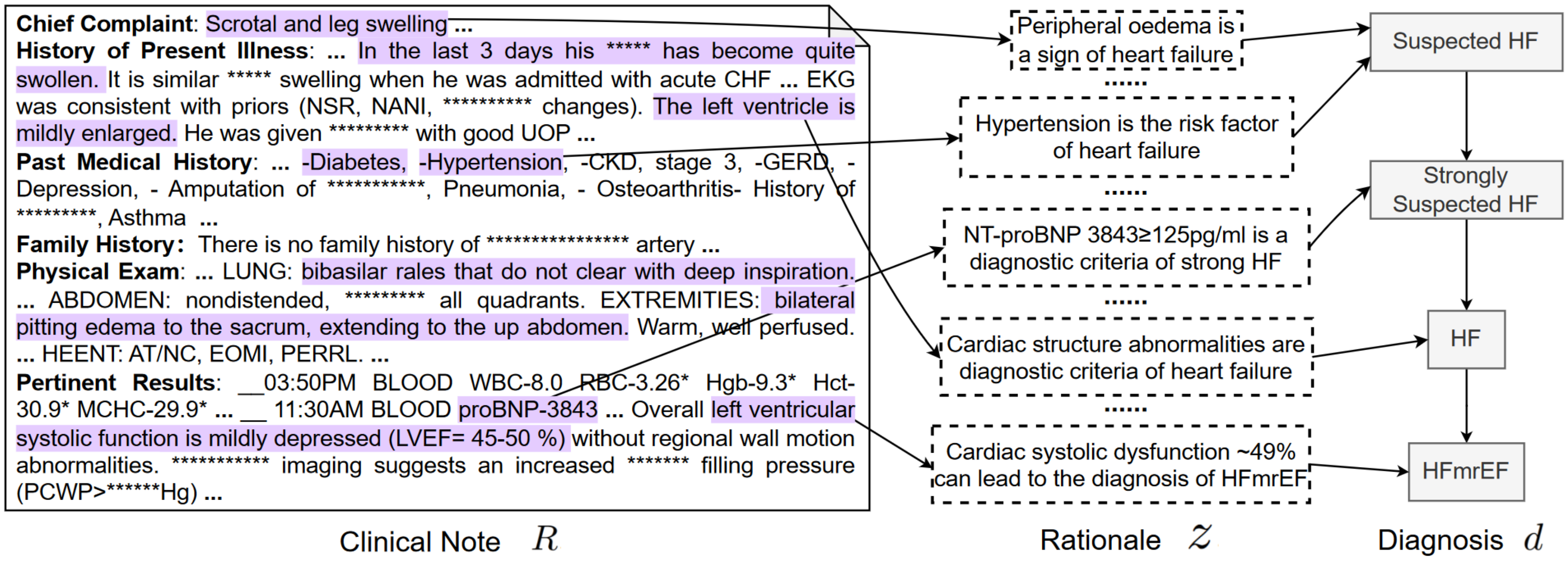
$\mathcal{P}_i$ Premise Nodes

$\mathcal{S}_i$ Supporting Edges

$\mathcal{D}^\star$ Leaf Nodes

Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research, 32:D267–D270, 2004.

Data Annotation



An annotation sample of Heart Failure (HF). The left part is the clinical note alongside extracted observations by a doctor. The middle part outlines the steps of the rationale for the premise corresponding to each diagnostic node shown in the right part.

$$\mathcal{E} = \{(o, z, d)\} \text{ for } (R, d^{\star})$$

Annotated by 9 clinical physicians and subsequently verified for accuracy and completeness by three senior medical experts.

Data Statistics and Task Definition

Statistics of all 25 disease categories, 511 annotations.

Domains	Categories	# samples	$\|\mathcal{D}_i\|$	$\|\mathcal{D}_i^\star\|$
Cardiology	Acute Coronary Syndromes	65	6	3
	Aortic Dissection	14	3	2
	Atrial Fibrillation	10	3	2
	Cardiomyopathy	9	5	4
	Heart Failure	52	6	3
	Hyperlipidemia	2	2	1
	Hypertension	32	2	1
Gastroenterology	Gastritis	27	5	3
	Gastroesophageal Reflux Disease	41	2	1
	Peptic Ulcer Disease	28	3	2
	Upper Gastrointestinal Bleeding	7	2	1
Neurology	Alzheimer	10	2	1
	Epilepsy	8	3	2
	Migraine	4	3	2
	Multiple Sclerosis	27	6	4
	Stroke	28	3	2
Pulmonology	Asthma	33	7	5
	COPD	19	6	4
	Pneumonia	20	4	2
	Pulmonary Embolism	15	5	3
	Tuberculosis	5	3	2
Endocrinology	Adrenal Insufficiency	20	4	3
	Diabetes	13	4	2
	Pituitary	12	4	3
	Thyroid Disease	10	6	4

Statistics of 5 medical domains.

Medical domain	# cat.	# samples	$\|\mathcal{D}_i\|$	$\|\mathcal{D}_i^\star\|$	$\|\mathcal{O}\|$	Length
Cardiology	7	184	27	16	8.7	1156.6 t
Gastroenterology	4	103	11	7	4.3	1026.0 t
Neurology	5	77	17	11	11.9	1186.3 t
Pulmonology	5	92	26	17	10.7	940.7 t
Endocrinology	4	55	20	14	6.9	1063.5 t
Overall	25	521	101	65	8.5	1074.6 t

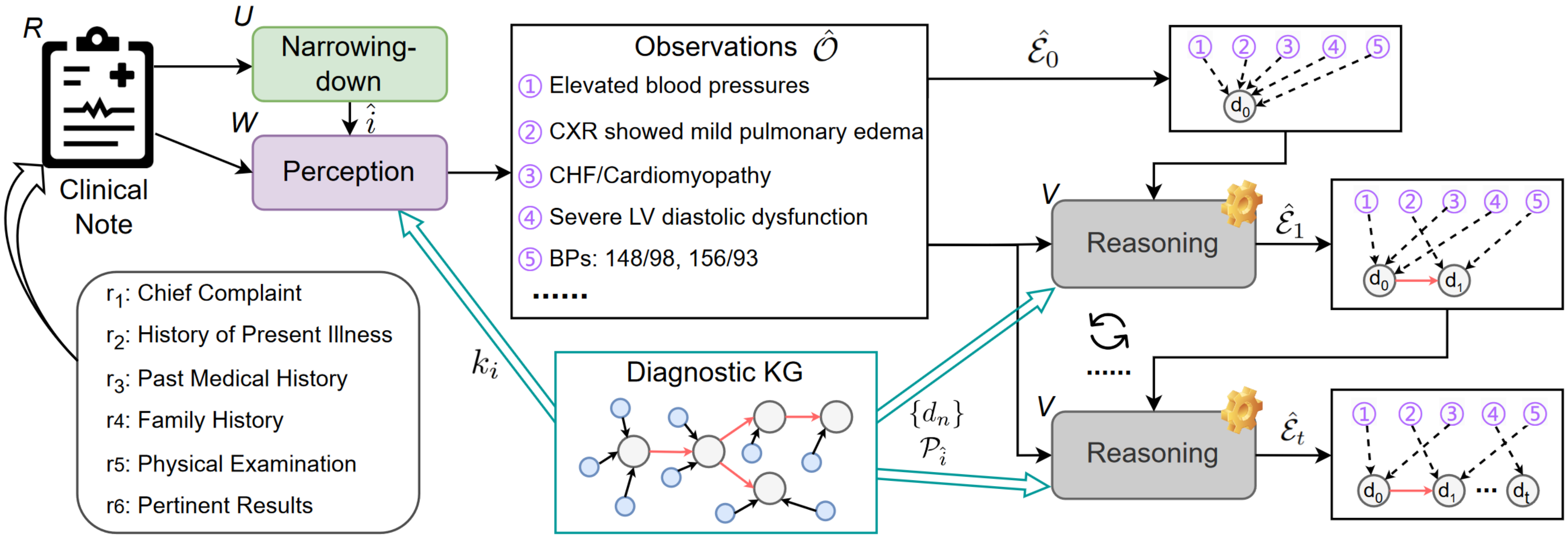
Each note only has a primary discharge diagnosis (PDD) in $\mathcal{D}^\star$

Task 1 $\hat{d}^\star, \hat{\mathcal{E}} = M(R, \mathcal{G})$

Task 2 $\hat{d}^\star, \hat{\mathcal{E}} = M(R, \mathcal{K})$

An AI Agent Pipeline

Our baseline comprises three LLM-based modules: narrowing-down U, perception W, and reasoning V.



Experiment Results

- Accuracy of diagnosis (*Acc*)
- Completeness of observations (*Obs*)
- Faithfulness of explanations (*Exp*)

Auto Evaluation via LLama3 8B

Table 3: Diagnostic reasoning ability of different LLMs under the proposed baseline method.

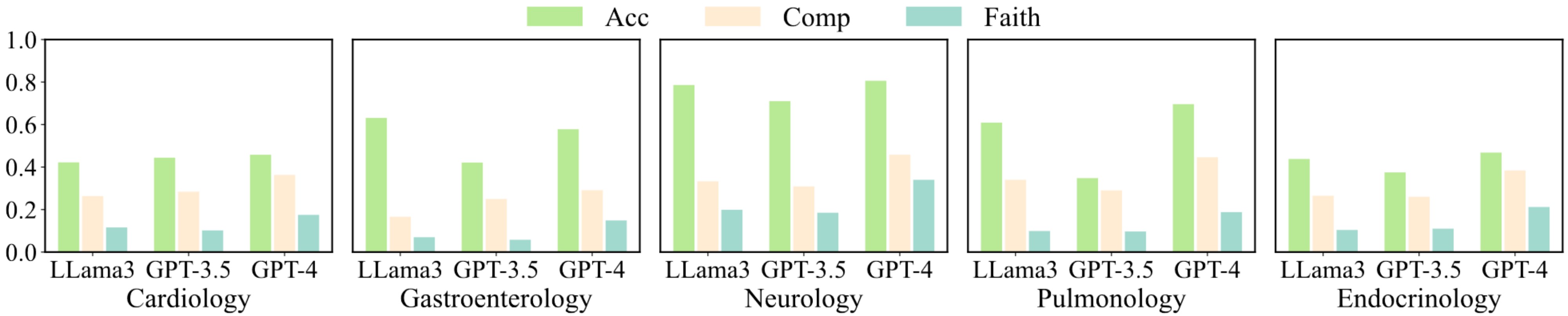
Task	Models	Diagnosis		Observation			Explanation	
		Acc^{cat}	Acc^{diag}	Obs^{pre}	Obs^{rec}	Obs^{comp}	Exp^{com}	Exp^{all}
With $\mathcal{G}$	Zephyr 7B	0.274	0.151	$0.123_{\pm 0.200}$	$0.115_{\pm 0.166}$	$0.092_{\pm 0.108}$	$0.071_{\pm 0.139}$	$0.014_{\pm 0.037}$
	Mistral 7B	0.507	0.306	$0.211_{\pm 0.190}$	$0.317_{\pm 0.253}$	$0.173_{\pm 0.157}$	$0.230_{\pm 0.312}$	$0.062_{\pm 0.088}$
	Mixtral 8×7B	0.413	0.237	$0.147_{\pm 0.165}$	$0.266_{\pm 0.261}$	$0.124_{\pm 0.138}$	$0.144_{\pm 0.268}$	$0.029_{\pm 0.056}$
	LLama3 8B	0.576	0.321	$0.253_{\pm 0.156}$	$0.437_{\pm 0.207}$	$0.219_{\pm 0.137}$	$0.232_{\pm 0.316}$	$0.071_{\pm 0.093}$
	LLama3 70B	0.752	0.540	$0.277_{\pm 0.146}$	$\mathbf{0.537}_{\pm \mathbf{0.192}}$	$0.256_{\pm 0.142}$	$0.395_{\pm 0.320}$	$0.112_{\pm 0.110}$
	GPT-3.5 turbo	0.679	0.455	$0.389_{\pm 0.212}$	$0.351_{\pm 0.192}$	$0.275_{\pm 0.167}$	$0.331_{\pm 0.366}$	$0.103_{\pm 0.127}$
	GPT-4 turbo	**0.772**	**0.572**	$\mathbf{0.446}_{\pm \mathbf{0.207}}$	$0.491_{\pm 0.180}$	$\mathbf{0.371}_{\pm \mathbf{0.186}}$	$\mathbf{0.475}_{\pm \mathbf{0.363}}$	$\mathbf{0.199}_{\pm \mathbf{0.181}}$
With $\mathcal{K}$	LLama3 8B	0.576	0.344	$0.235_{\pm 0.162}$	$0.394_{\pm 0.227}$	$0.199_{\pm 0.142}$	$0.327_{\pm 0.375}$	$0.087_{\pm 0.114}$
	LLama3 70B	0.735	0.581	$0.262_{\pm 0.146}$	$\mathbf{0.501}_{\pm \mathbf{0.208}}$	$0.236_{\pm 0.131}$	$0.463_{\pm 0.374}$	$0.125_{\pm 0.117}$
	GPT-3.5 turbo	0.652	0.413	$0.347_{\pm 0.241}$	$0.279_{\pm 0.203}$	$0.232_{\pm 0.184}$	$0.374_{\pm 0.408}$	$0.121_{\pm 0.152}$
	GPT-4 turbo	**0.781**	**0.614**	$\mathbf{0.431}_{\pm \mathbf{0.207}}$	$0.458_{\pm 0.187}$	$\mathbf{0.353}_{\pm \mathbf{0.170}}$	$\mathbf{0.633}_{\pm \mathbf{0.338}}$	$\mathbf{0.247}_{\pm \mathbf{0.201}}$

Table 4: Evaluation of diagnostic reasoning ability of LLMs when no external knowledge is provided.

Task	Models		Observation			Explanation	
		Acc^{diag}	Obs^{pre}	Obs^{rec}	Obs^{comp}	Exp^{com}	Exp^{all}
With $\mathcal{D}^\star$	LLama3 8B	0.070	$0.154_{\pm 0.139}$	$0.330_{\pm 0.244}$	$0.135_{\pm 0.122}$	$0.020_{\pm 0.100}$	$0.004_{\pm 0.016}$
	LLama3 70B	0.502	$0.257_{\pm 0.150}$	$\mathbf{0.509}_{\pm \mathbf{0.213}}$	$0.237_{\pm 0.145}$	$0.138_{\pm 0.209}$	$0.034_{\pm 0.054}$
	GPT-3.5 turbo	0.223	$0.164_{\pm 0.242}$	$0.149_{\pm 0.212}$	$0.116_{\pm 0.174}$	$0.091_{\pm 0.231}$	$0.025_{\pm 0.065}$
	GPT-4 turbo	**0.636**	$\mathbf{0.461}_{\pm \mathbf{0.206}}$	$0.482_{\pm 0.160}$	$\mathbf{0.378}_{\pm \mathbf{0.174}}$	$\mathbf{0.186}_{\pm \mathbf{0.221}}$	$\mathbf{0.074}_{\pm \mathbf{0.090}}$
No Knowledge	LLama3 8B	0.023	$0.137_{\pm 0.159}$	$0.258_{\pm 0.274}$	$0.119_{\pm 0.141}$	$0.018_{\pm 0.083}$	$0.006_{\pm 0.026}$
	LLama3 70B	0.037	$0.246_{\pm 0.148}$	$\mathbf{0.504}_{\pm \mathbf{0.222}}$	$0.227_{\pm 0.148}$	$0.022_{\pm 0.093}$	$0.007_{\pm 0.030}$
	GPT-3.5 turbo	0.059	$0.161_{\pm 0.238}$	$0.148_{\pm 0.215}$	$0.113_{\pm 0.171}$	$0.036_{\pm 0.131}$	$0.011_{\pm 0.039}$
	GPT-4 turbo	**0.074**	$\mathbf{0.410}_{\pm \mathbf{0.208}}$	$0.443_{\pm 0.191}$	$\mathbf{0.324}_{\pm \mathbf{0.182}}$	$\mathbf{0.047}_{\pm \mathbf{0.143}}$	$\mathbf{0.019}_{\pm \mathbf{0.058}}$

Experiment Results

Performance of LLama3 70B, GPT-3.5, and GPT-4 under different medical domains.



Experiment Results

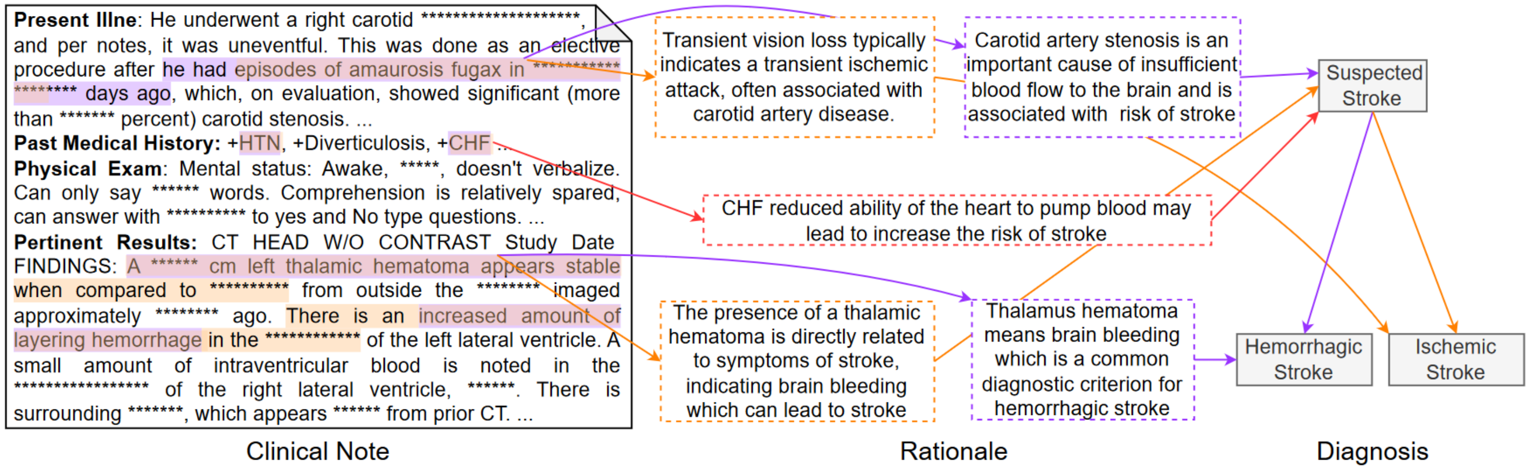
Consistency of automated evaluation metrics with human judgments.

Table 5: Consistency of automated evaluation with human judgments. Evaluated by mean and confidence interval (CI).

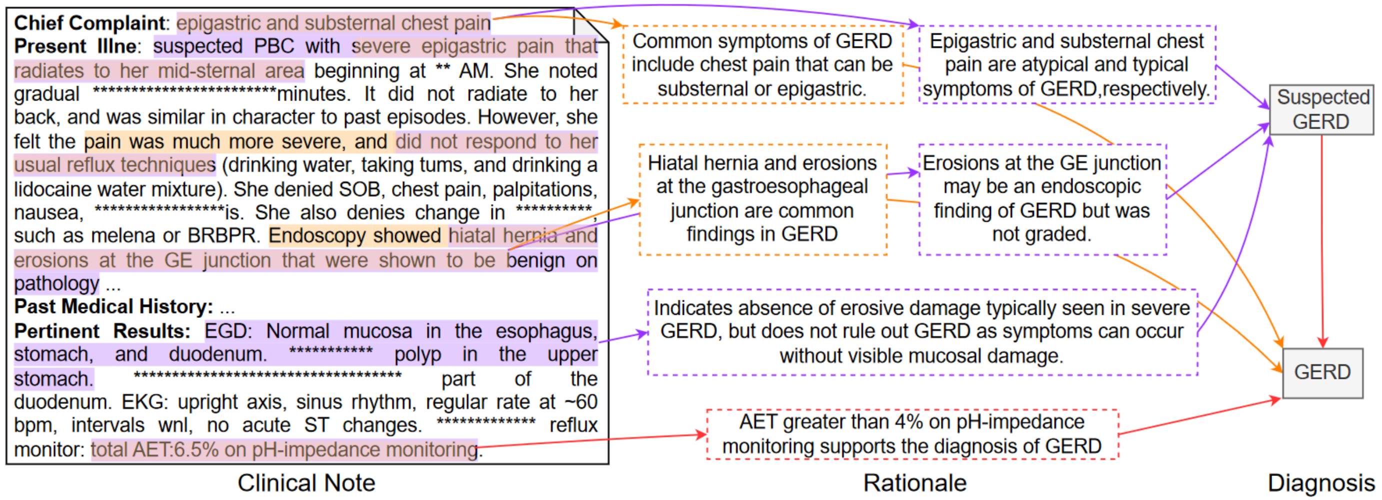
	Observation		Rationalization	
Model	Mean	95% CI	Mean	95% CI
LLama3 8B	0.887	0.844 $\sim$ 0.878	0.835	0.759 $\sim$ 0.818
GPT-4 turbo	0.902	0.830 $\sim$ 0.863	0.876	0.798 $\sim$ 0.853

Generated Samples From GPT-4

Purple, orange, and red indicate ground truth, prediction, and common in both, respectively.



An example prediction for a clinical note with PDD of Hemorrhagic Stroke by GPT-4.



An example prediction for a clinical note with PDD of GERD by GPT-4

Thanks