

Emergency Department: A Real-world Edge Case for Clinical NLP | Dipankar Srirag & Aditya Joshi | AJCAI 2026

P.S.: No LLM-generated slides or script ahead.

#SayNoToLLMSlopInResearch

Tutorial at Australasian Joint Conference on Artificial Intelligence
2026, Auckland, New Zealand, November 2026

Emergency Department

A Real-World Edge Case for Clinical NLP



Dipankar Srirag **Dr. Aditya Joshi**

University of New South Wales, Sydney

Friday, 27 November 2026

UNSW-NLP

About Us



Dipankar Srirag (he/him)

Ph.D. Candidate
University of New South Wales, Sydney

UNSW-NLP



Dr. Aditya Joshi (he/him)

Senior Lecturer
University of New South Wales, Sydney

Natural Language Processing (NLP) group @ UNSW

Natural Language Processing for Language Variations

Survey [ACM CSUR 2025]

Building Benchmarks

"Sentiment and Sarcasm Classification of **Aus/Ind/Brit English**" [Findings ACL 2025]

"ASR for **technical dialectal text**" [SPEAKABLE at LREC '26]

"**Dialectal English Conversation** Simulation from Structured Medical Records. [Interspeech 2026]"

"Legal Equivalence Classification for **Aus/Ind/UK Legal English**" [Under Review]

Model Adaptation

"Contrastive learning-based fine-tuning for topic classification in **US/Ind/Nig English**. [NAACL 2025]"

"Post-training inspired by **Italian and English bilingual** speakers" [SumEval at COLING 2025]

"Continual pre-training for **English dialects**" [EMNLP 2026]

"Few-shot learning for **Arabic dialects**" [ICNLSP 2026]

Language Understanding

"Detecting **stuttering** in audio. [Interspeech 2026]"

"**Generalist or legal-specific language** training for legal NLP [CustomizableNLP@ACL 2026]

"Sarcasm explanation for **Aus and Ind Eng** [ALTA 2025]" "Satire Explanation for **Aus Eng** [ALTA 2026]"

"**Humblebrag & Sarcasm** Generation via mech interp" [Findings EMNLP 2026]

Language Generation

Evaluation

"Evaluation of LLMs **for Aus and IndEng slang** [VarDial @ EACL 2026]"

"**Drunk Language** as a vulnerability to LLM security" [INLG 2026]

"RAG for **Defence language** documents". [ADSTAR 2026; CIKM 2026; EMNLP 2026]

"Sycophancy in English, Mandarin, and **Chinese English**" [ALTA 2026]

We’d like to know more about you!

Join at
slido.com
#7089 547



Learning Outcomes

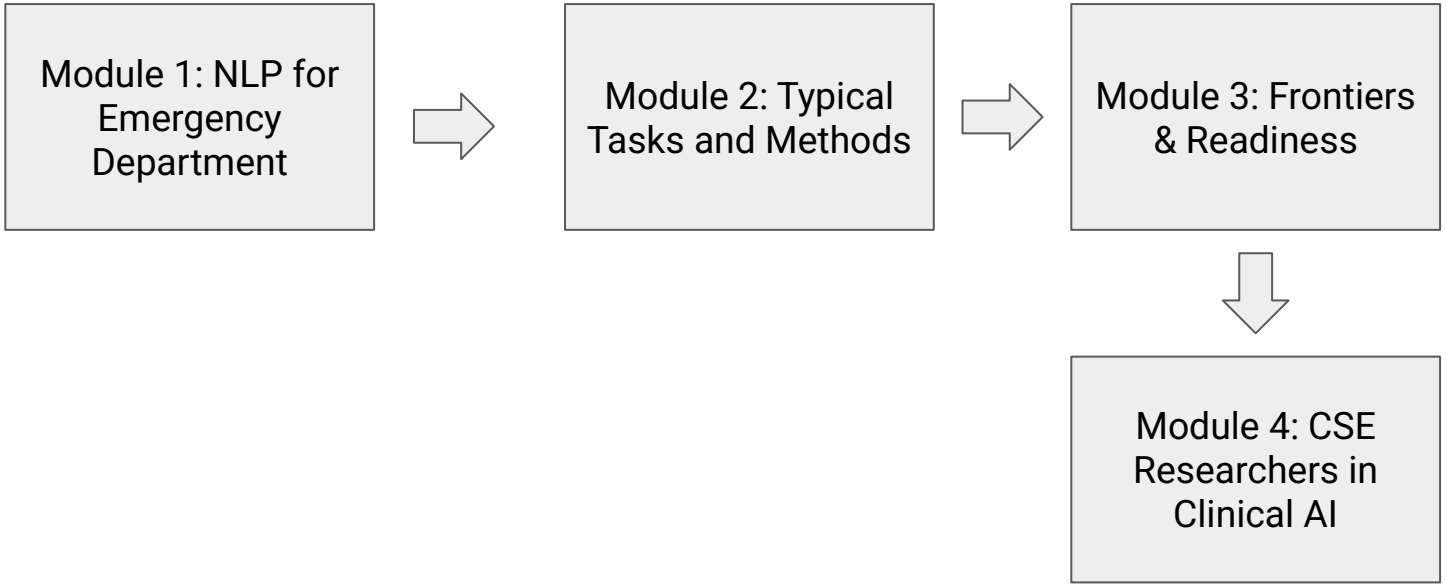
LO1: Describe the landscape of ED NLP: tasks, datasets, modelling paradigms, evaluation in triage, diagnosis and disposition.

LO2: Analyse the deployment constraints specific to the ED and the methods that currently address each.

LO3: Examine the desiderata an NLP system must satisfy to be credible in an ED, and apply them to critique a proposed design.

LO4: Design a research project for ED in NLP by identifying a relevant research question, resources (dataset, simulation), methodology.

Tutorial Plan



Tutorial Plan

Module 1: NLP for Emergency Department



Module 2: Typical Tasks and Methods

Module 3: Frontiers & Readiness

Module 4: CSE Researchers in Clinical AI

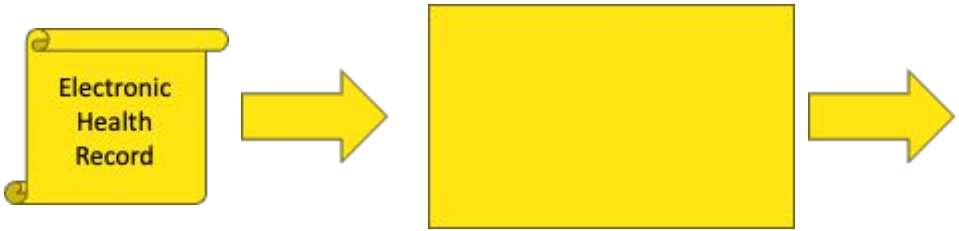
Clinical NLP

NLP techniques applied to clinical settings

‘Clinical settings’ broadly refers to:

- Clinical/biomedical Research & Deployment
- In-hospital Settings

Clinical Diagnosis

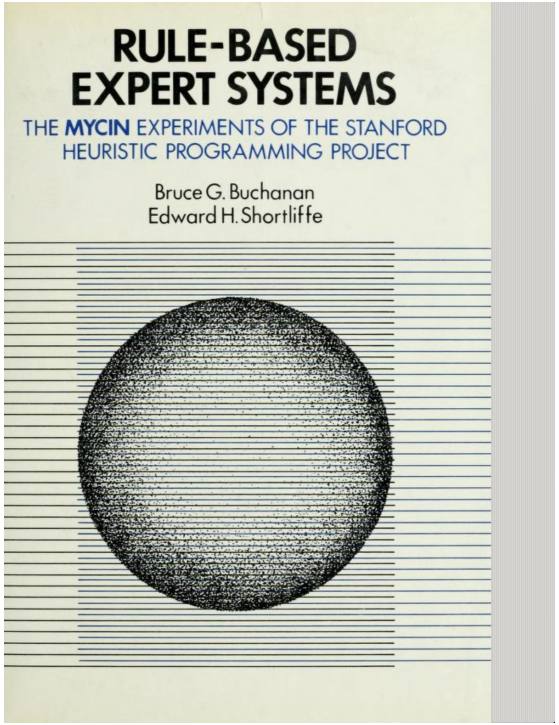


Diagnosis Prediction

Rationale

Ontology-aligned Prediction

The Old and the New



https://archive.org/details/rulebasedexperts00buch

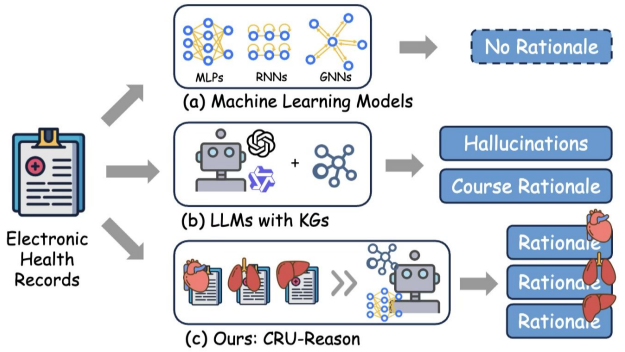
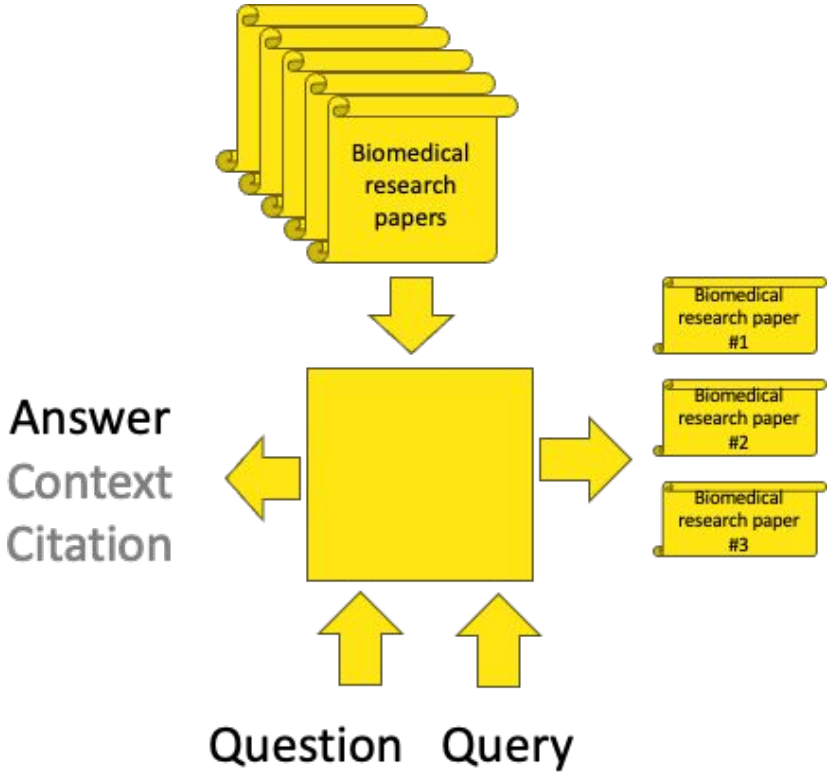


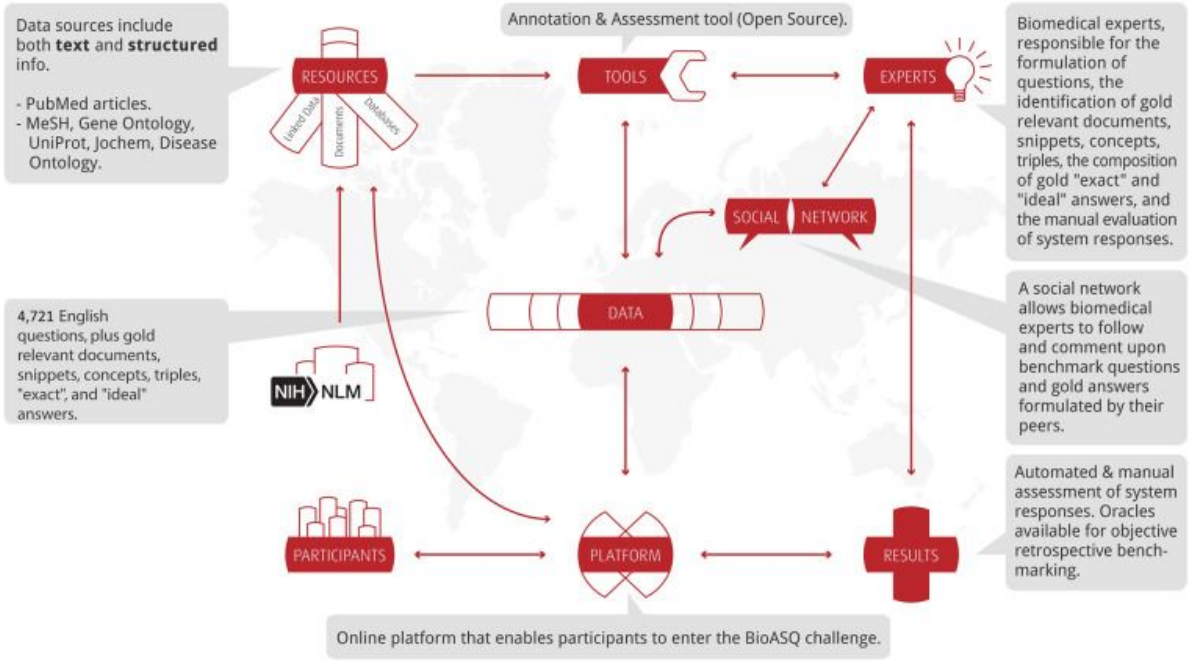
Figure 1: Schematic diagrams for clinical prediction. **(a)** ML models are unable to provide human-readable rationales. **(b)** LLM+KG methods typically perform coarse-grained, hallucination-prone reasoning with noisy patient context and knowledge. **(c)** CRU-Reason provides fine-grained rationales by leveraging disease-centric communities (i.e., *Clinical Rationale Units*).

Zihang Liu, Linhao Luo, Zhangchi Qiu, Shigeng Chen, and Haishuai Wang. 2026. CRU-Reason: Knowledge Graph-Grounded and Interpretable Clinical Prediction with Clinical Rationale Units. In Proceedings of the 2026 EMNLP.

Biomedical Retrieval/Question-Answering

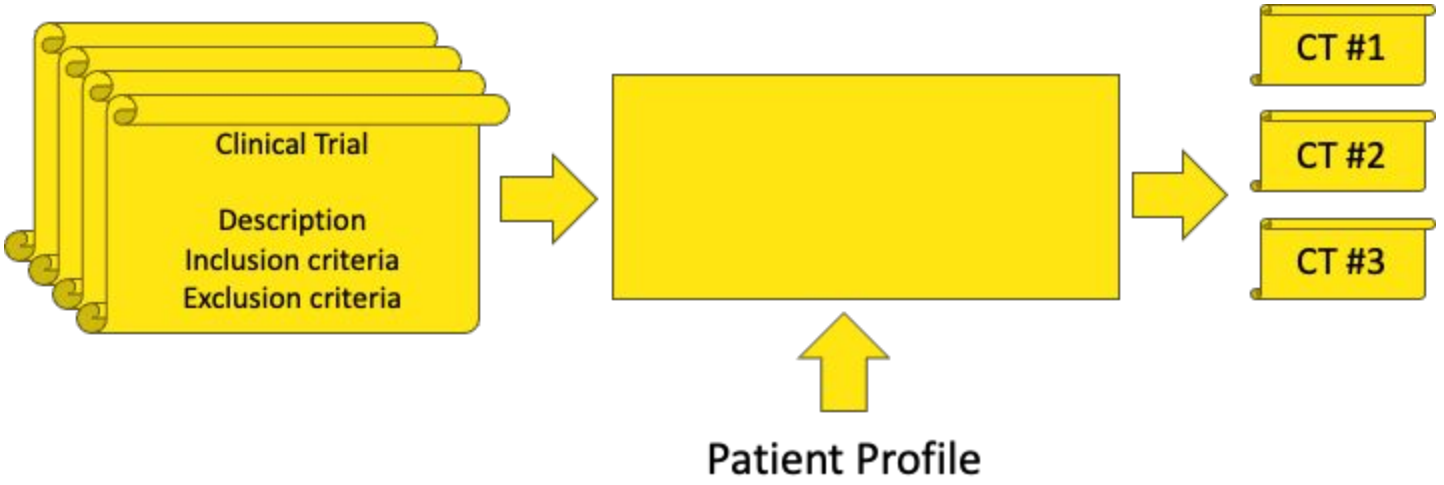


Humans > AI.



Krithara, A., Nentidis, A., Bougiatiotis, K. *et al.* BioASQ-QA: A manually curated corpus for Biomedical Question Answering. *Sci Data* **10**, 170 (2023). https://doi.org/10.1038/s41597-023-02068-4

Clinical Trial Matching



Tu, Samson W., Carol A. Kemper, Nancy M. Lane, Robert W. Carlson, and Mark A. Musen. "A methodology for determining patients' eligibility for clinical trials." *Methods of information in medicine* 32, no. 04 (1993): 317-325.

Statistical -> Neural

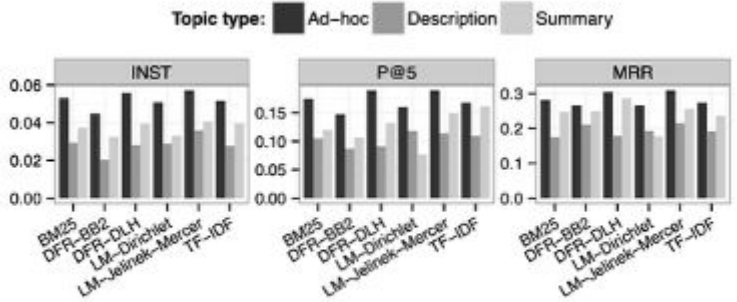


Figure 4: Retrieval results for different baselines and topic representations.

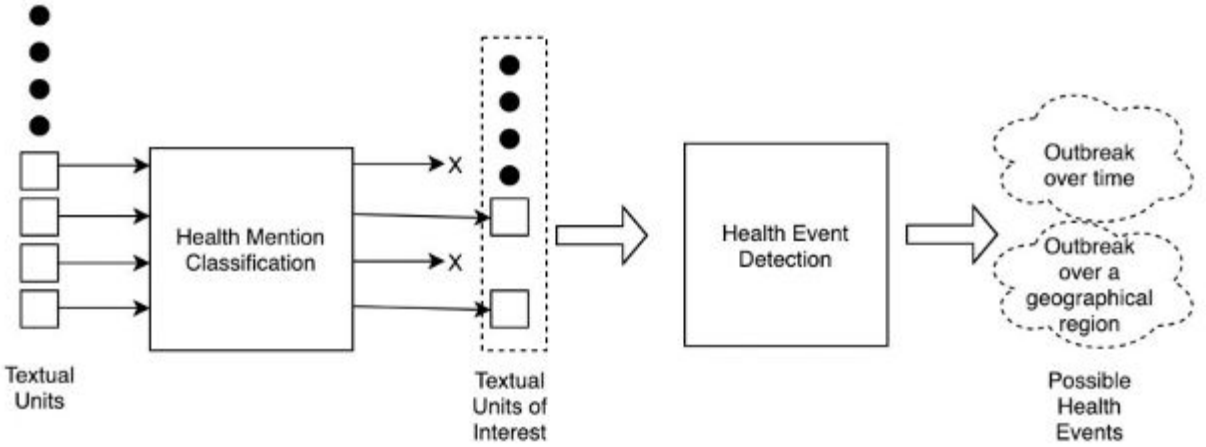
Language	BLEU (↑)	TER (↓)	COMET (↑)
Spanish (ES)	0.93	6.15	0.87
Italian (IT)	0.92	6.34	0.89
Polish (PL)	0.75	22.84	0.90
Turkish (TR)	0.89	8.07	0.89
Bengali (BN)	0.14	74.01	0.78
Greek (EL)	0.93	6.49	0.91
Basque (EU)	0.96	3.14	0.90

Maciej Rybinski, **Aditya Joshi,** et al. CTCL: A Cross-Language Benchmark for Matching Patients to Clinical Trials. In Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26).

Rybinski, Maciej, **Aditya Joshi**, Sarvnaz Karimi, y Nathan Inkiriwang. CT-bias: a Dataset for Auditing the Impact of Patients" Sensitive Information on Clinical Trial Matching», 2026.CIKM. https://doi.org/10.5281/ZENODO.20407923.

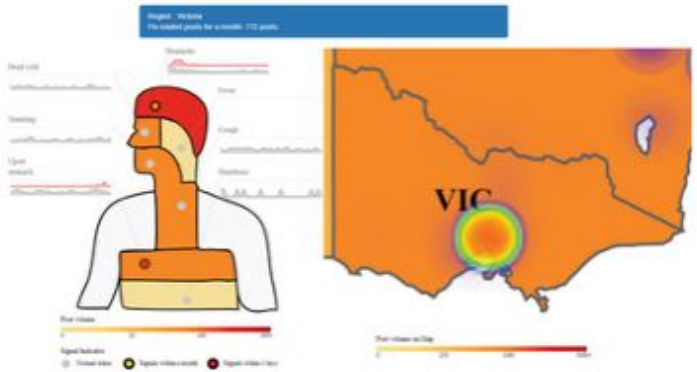
Koopman, Bevan, and Guido Zuccon. "A test collection for matching patients to clinical trials." In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 669-672. 2016.

Social media-based epidemic intelligence



Aditya Joshi, Sarvnaz Karimi, Ross Sparks, Cécile Paris, and C. Raina Macintyre. "Survey of text-based epidemic intelligence: A computational linguistics perspective." *ACM Computing Surveys (CSUR)* 52, no. 6 (2019): 1-19.

Deployment & Challenges

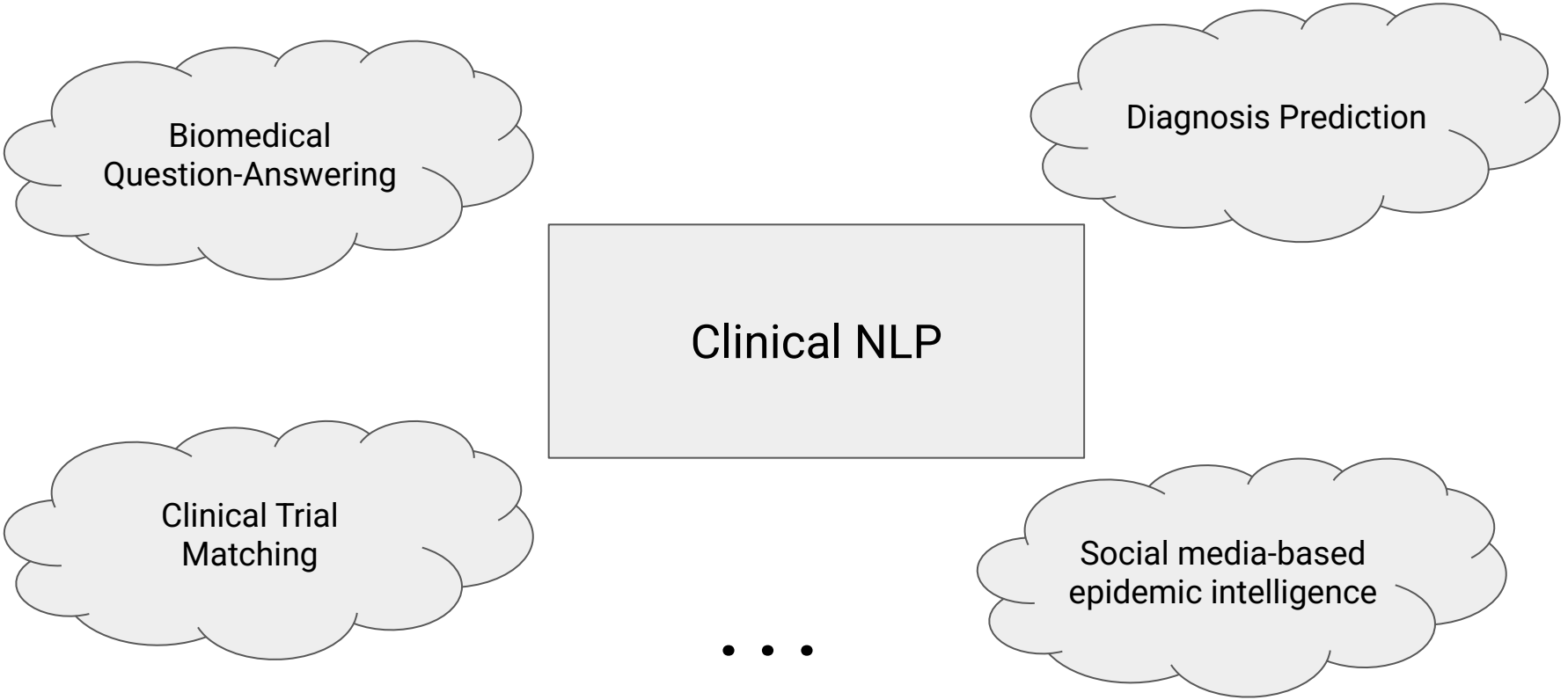


Brian Jin, **Aditya Joshi**, Ross Sparks, Stephen Wan, Cécile Paris, and C Raina MacIntyre. 2020. "'Watch the Flu': A Tweet Monitoring Tool for Epidemic Intelligence of Influenza in Australia". Proceedings of the AAAI Conference on Artificial Intelligence 34 (09):13616-17.

'i nearly had a stroke readin this'

'Wow, this is awesome it's like having depression'

Rhys Biddle, **Aditya Joshi,** Shaowu Liu, Cécile Paris, Guandong Xu. "Leveraging Sentiment Distributions to Distinguish Figurative From Literal Health Reports on Twitter". WWW 2020.


Biomedical
Question-Answering
Diagnosis Prediction
Clinical NLP
Clinical Trial
Matching
Social media-based
epidemic intelligence
• • •

Enter: Emergency Department (ED)



https://www.thermh.org.au/emergency-tv

Challenges

Time-critical

- 4 hour window from arrival to disposition

Scale

- 8 million visits of Australian EDs in 2022

Life-critical

Cost ramifications

Language barriers



Language Barriers in ED

Higher odds of low acuity level assignment [1]

Time-criticality is a greater risk for triage errors [2]

50% overall satisfaction in all local health districts of Sydney in a survey 2017-2022 [3]

Associated with longer stay, lower satisfaction. [4]

[1] Lee S-Y, Alzeen M, Ahmed A. Estimation of racial and language disparities in pediatric emergency department triage using statistical modeling and natural language processing. *Journal of the American Medical Informatics Association*. 2024;31(4):958-967. doi:10.1093/jamia/ocae018

[2] Eisenberg EM, Murphy AG, Sutcliffe K, et al. Communication in emergency medicine: implications for patient safety. *Communication monographs*. 2005;72(4):390-413.

[3] Oredsson S. Triage and patient safety in emergency departments. British Medical Journal Publishing Group; 2011.

[4] Garrett PW, Forero R, Dickson HG, Whelan AK. How are language barriers bridged in acute hospital care? The tale of two methods of data collection. *Australian Health Review*. 2008;32(4):755-764.

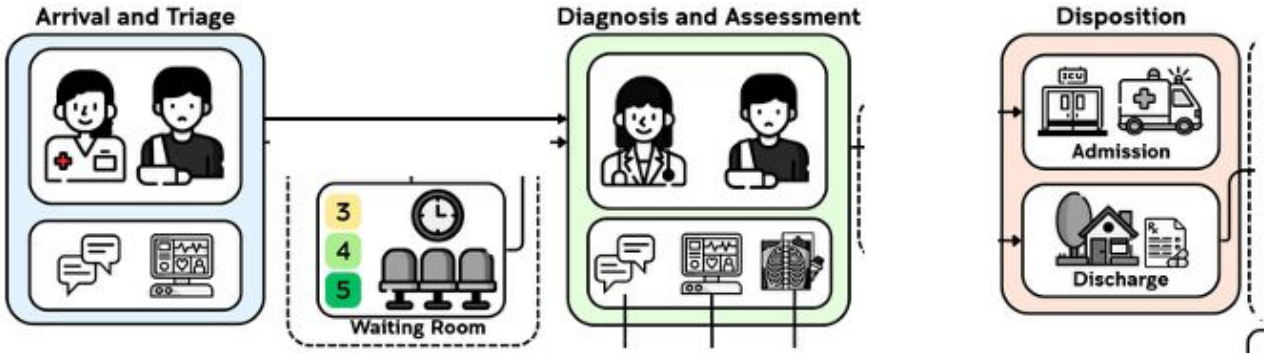
Emergency Department: An edge case for Clinical NLP





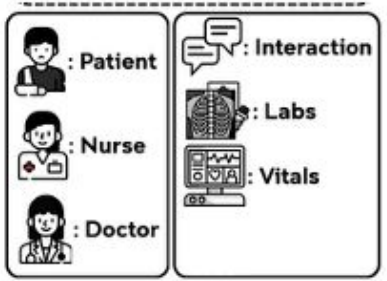


Journey through an ED



Triage: Priority level assigned to patients based on specific situation.

Interaction: A conversation between a healthcare worker(s) and a patient(s).

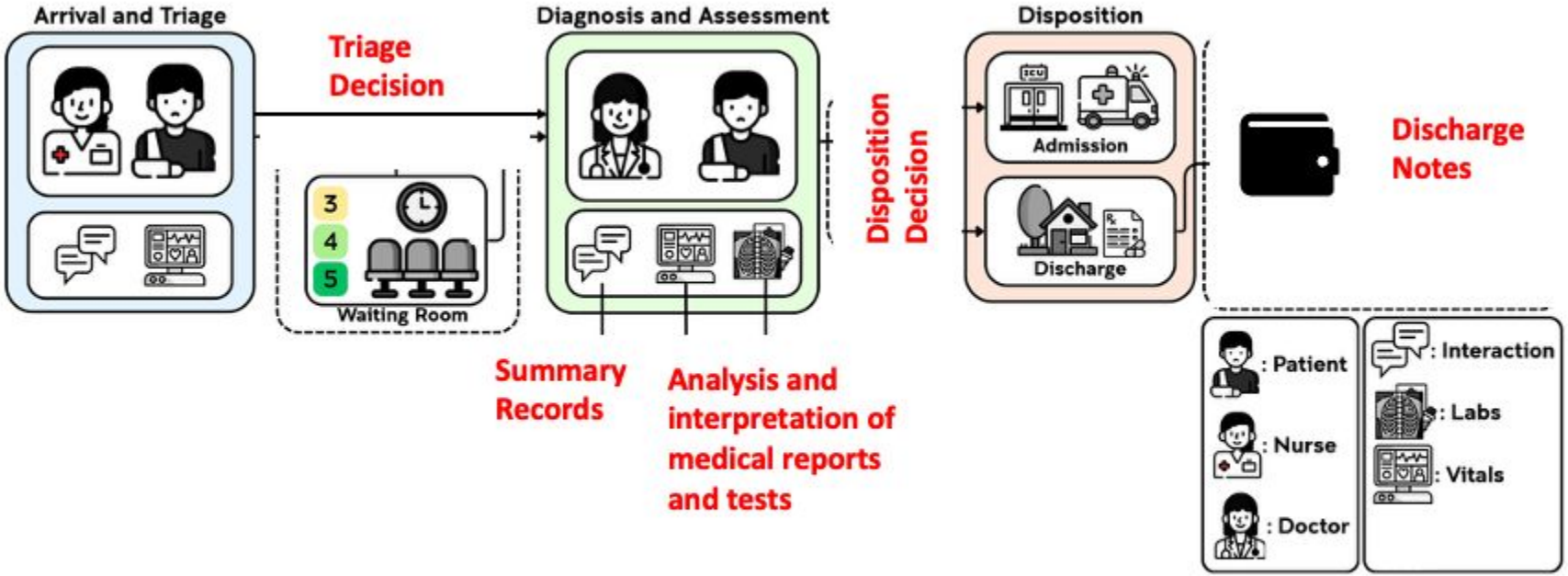


Chief complaint: The patient's reason for presenting to the healthcare worker.

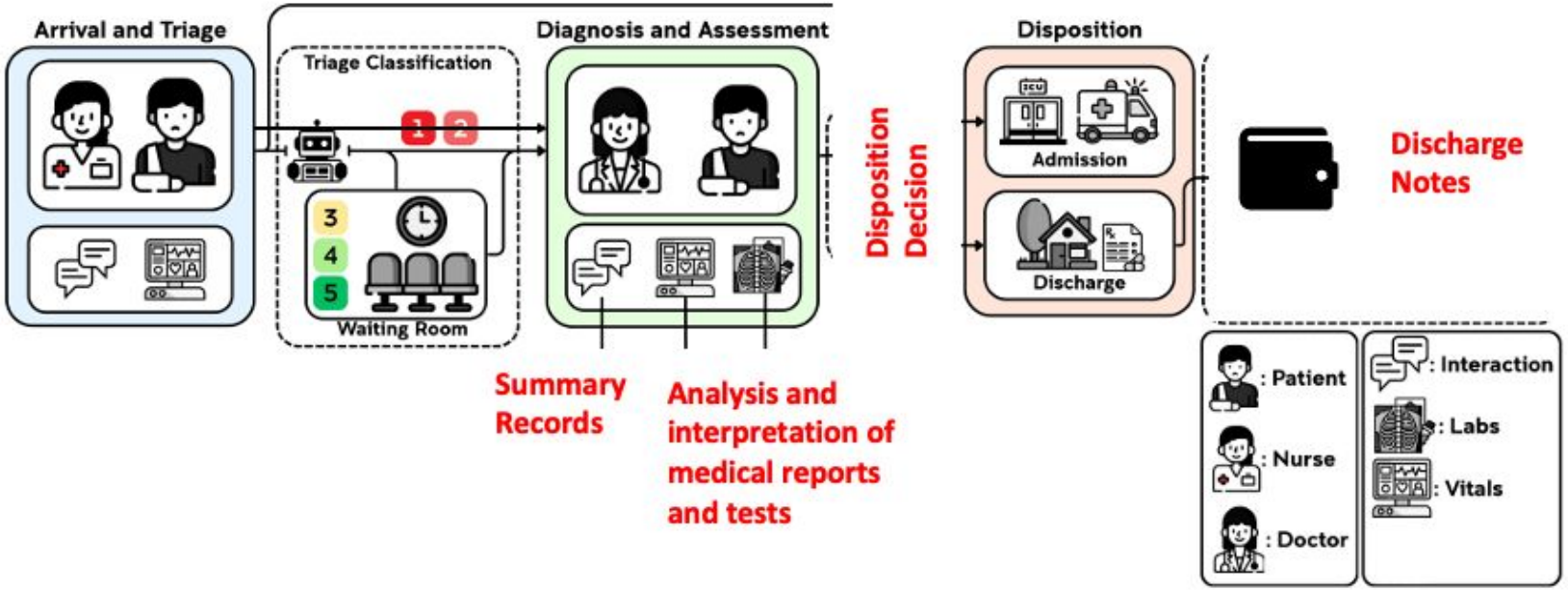
Vitals

HR	RR	O_2 Sat.	BP	Temp.
118	24	93	128/78	98.2

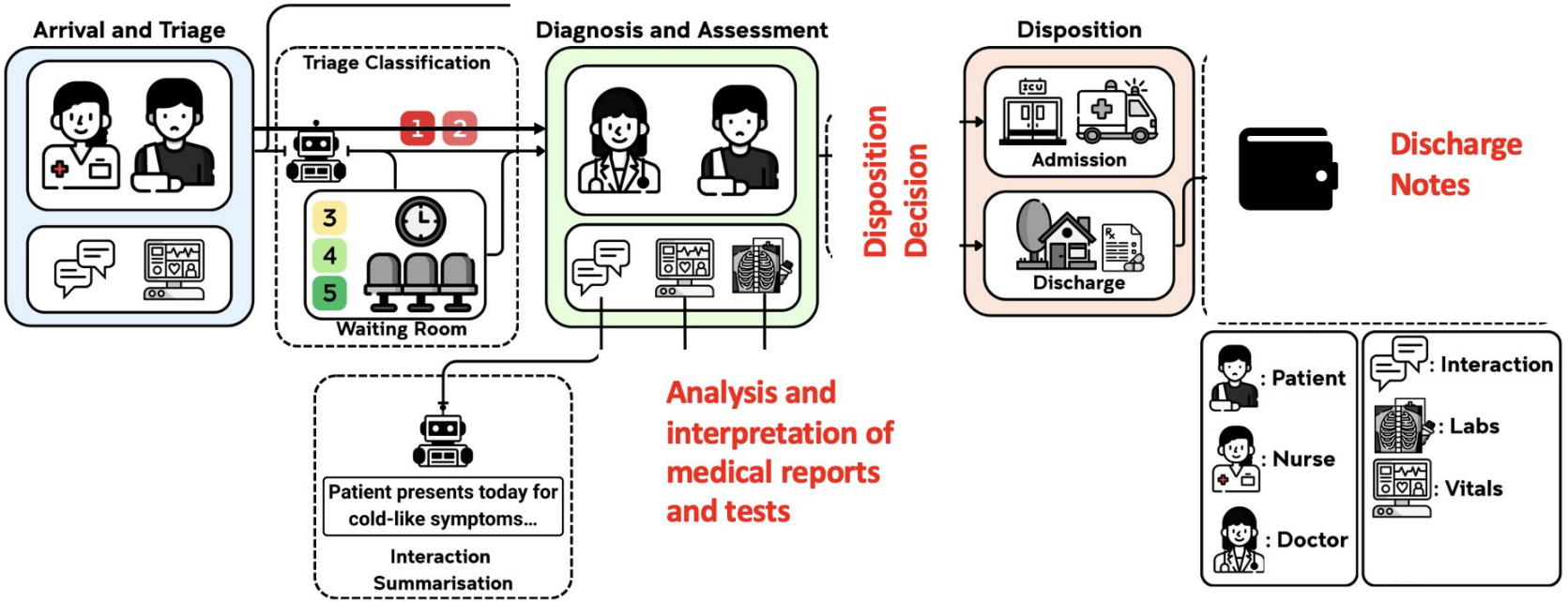
.. and the role of AI/NLP



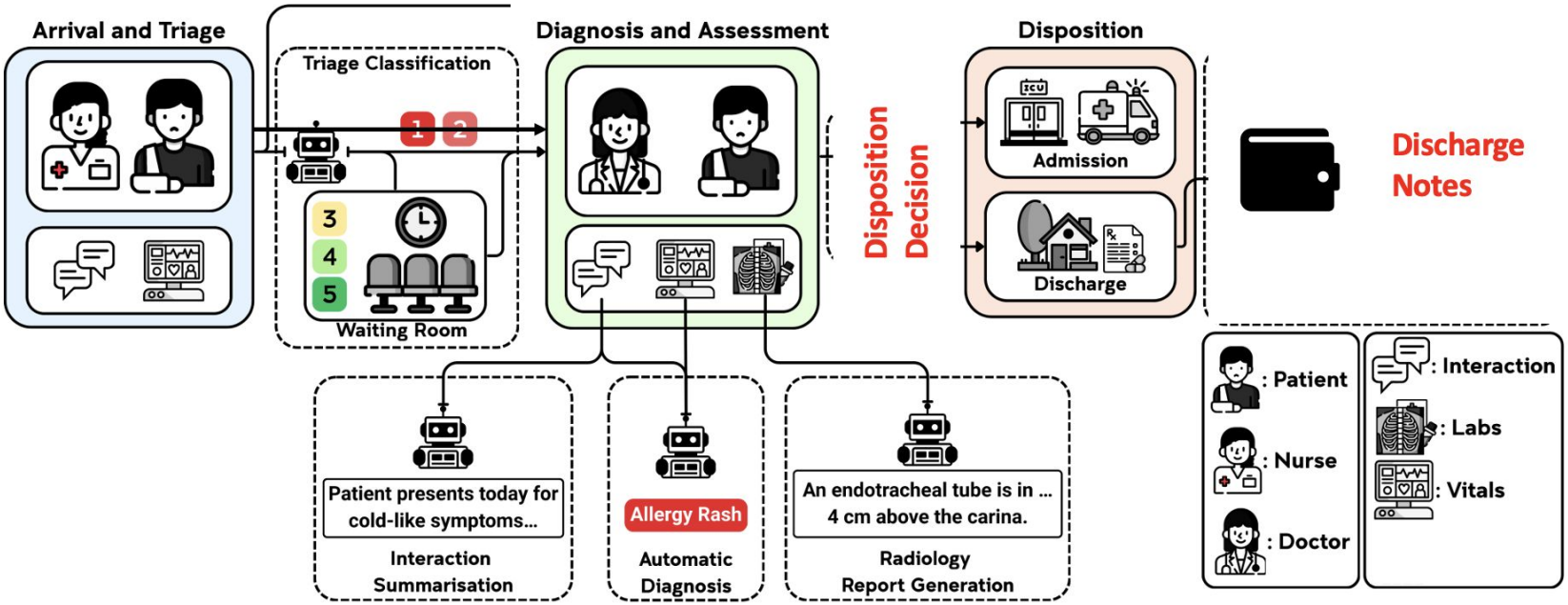
The AI/NLP tasks that emerge



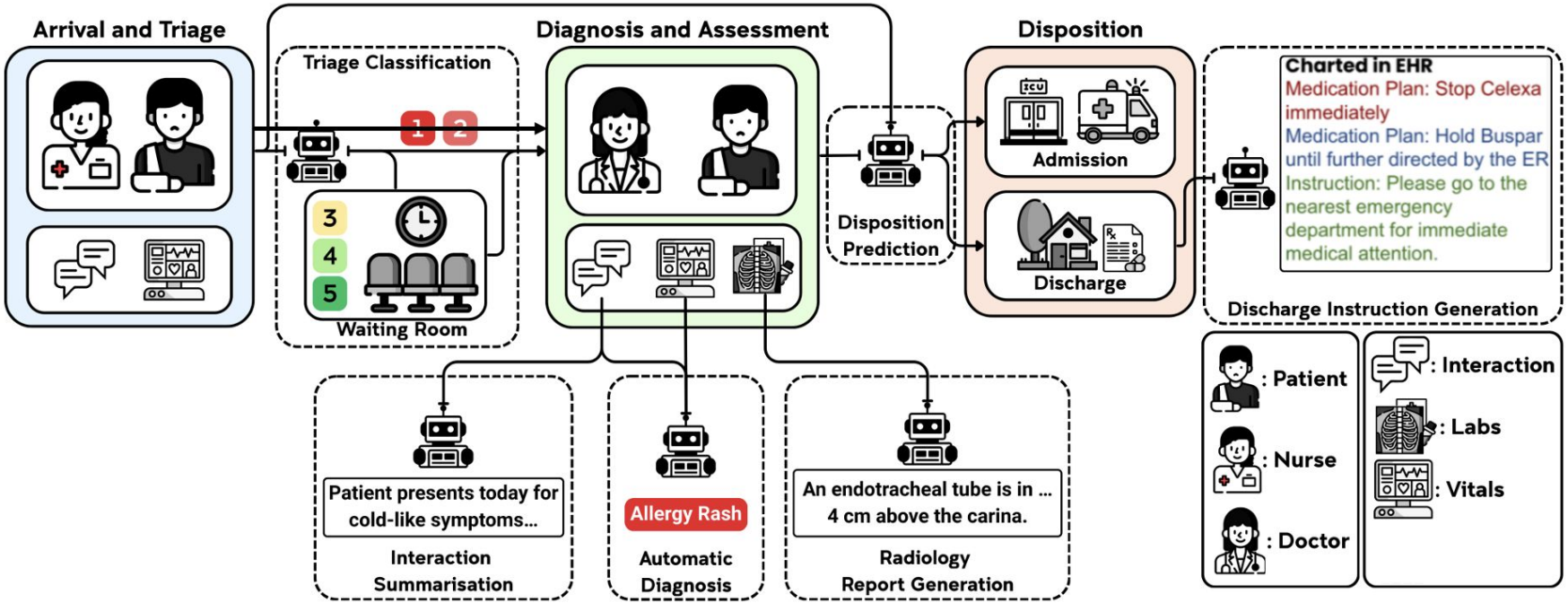
The AI/NLP tasks that emerge



The AI/NLP tasks that emerge



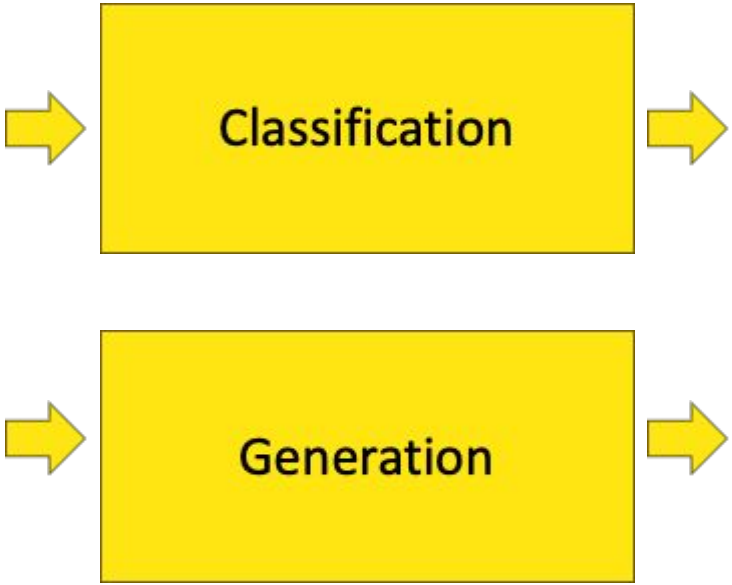
The AI/NLP tasks that emerge



Two task formulations

Structured: EHR
Unstructured: Patient notes
Semi-structured: Conversations

Image: X-rays



Example Outputs:
Triage Level
Diagnosis
Disposition/Admission

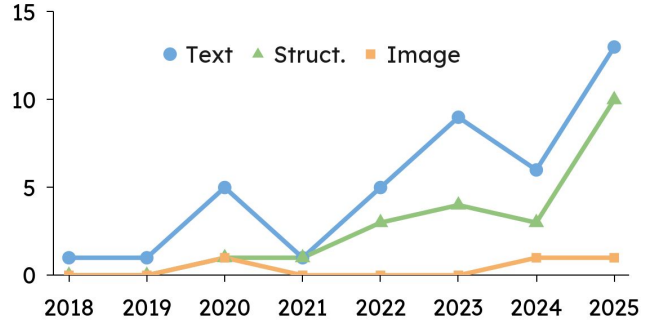
Example Outputs:
Summary
Discharge Notes
Explanations,
etc.

Scope

From Triage to Discharge: A Survey of NLP Tasks, Methods, and Open Challenges in the Emergency Department

Dipankar Srirag Aditya Joshi Salil S. Kanhere Padmanesan Narasimhan
University of New South Wales, Sydney, Australia
{d.srirag, aditya.joshi, salil.kanhere, padmanesan}@unsw.edu.au

Input Modality



Sources. We searched the following venues in NLP, AI, and health informatics: the ACL Anthology[3], the AAAI Digital Library[4], the NeurIPS proceedings[5], the ACM Digital Library[6], the ISCA Archive[7], the Nature Portfolio[8], and PLOS ONE[9]. The initial search was completed in August 2025; relevant papers published in these venues after this date were retrospectively included up to the submission date.

Why the NLP focus? 4/46 papers surveyed use imaging datasets. Rest are textual and/or structured.

Dipankar Srirag, Aditya Joshi, Salil S. Kanhere, and Padmanesan Narasimhan. 2026. From Triage to Discharge: A Survey of NLP Tasks, Methods, and Open Challenges in the Emergency Department. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, pages 20918–20942, Budapest, Hungary. Association for Computational Linguistics.

Tutorial Plan

Module 1: NLP for Emergency Department

Module 2: Typical Tasks and Methods



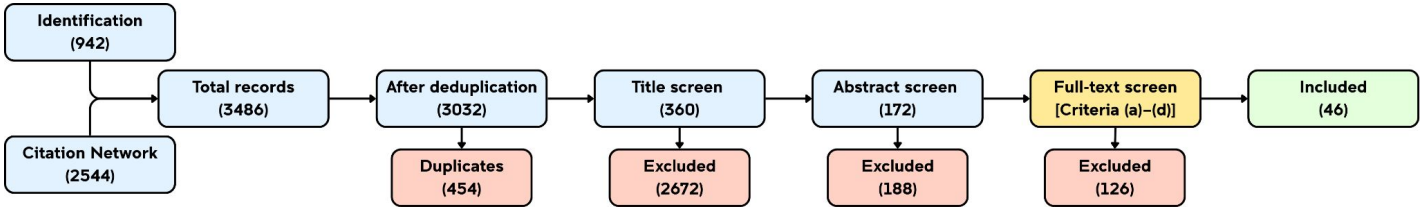
Module 3: Frontiers & Readiness

Module 4: CSE Researchers in Clinical AI

Notation

	NLP Task	$\mathcal{X}$	Input Space
	Doctor	x	Input
	Nurse	I	Instruction
		$\mathcal{Y}$	Output Space
	Patient	y	Output
	Interaction	θ	Model Parameters
	Vitals	p_θ	Model
	Labs	$\hat{y}$	Model Output

Landscape of ED NLP



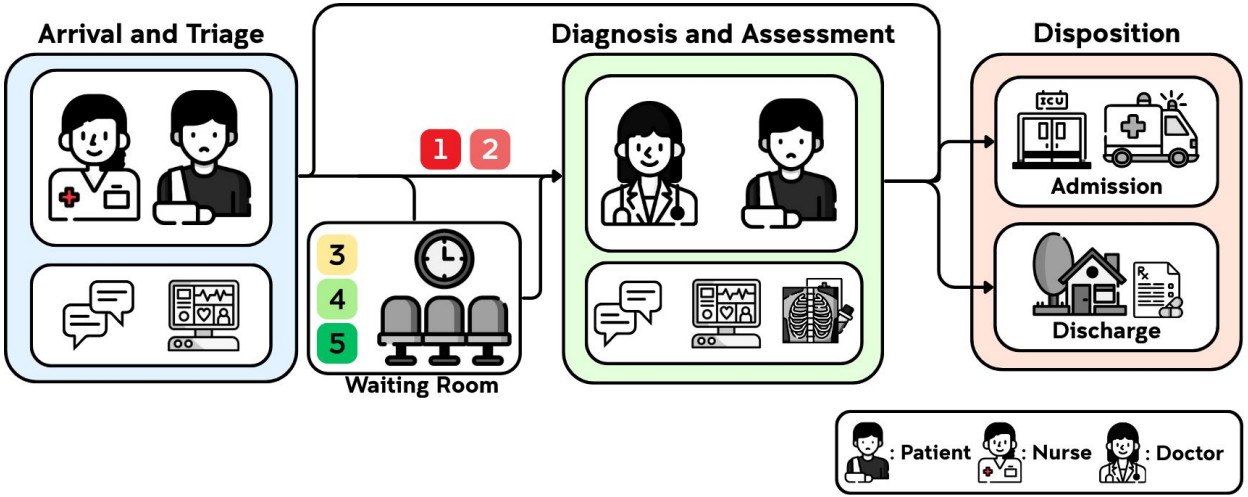
Terms

- Emergency Department/Room
- Triage
- Clinical
- Hospital
- ED, ER, A&E
- NLP, AI, ML

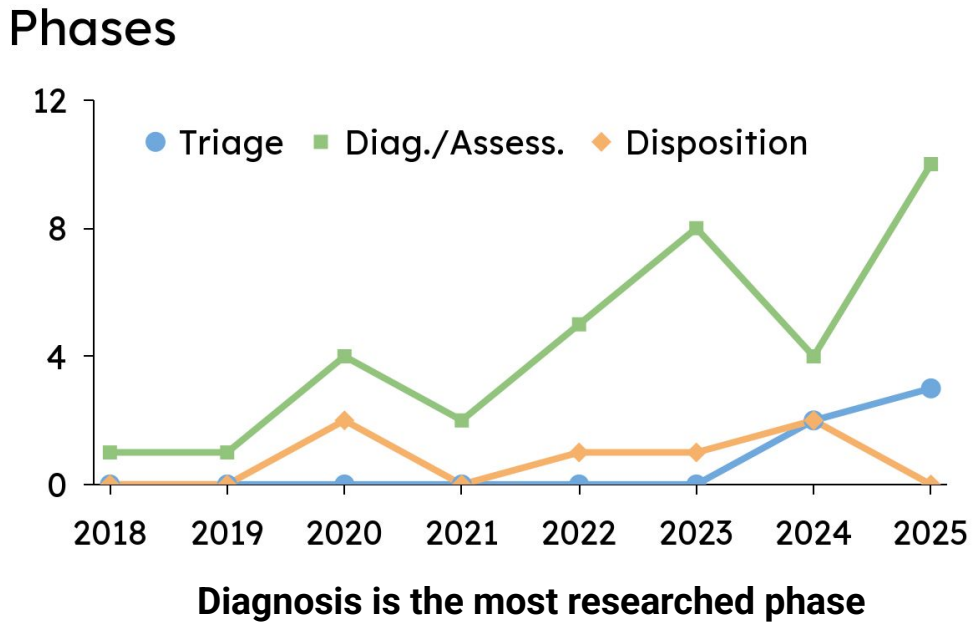
Criteria

(a) ED Phase Alignment
(b) ED-relevant language resource
(c) NLP method as contribution
(d) ED-anchored evaluation

Landscape of ED NLP

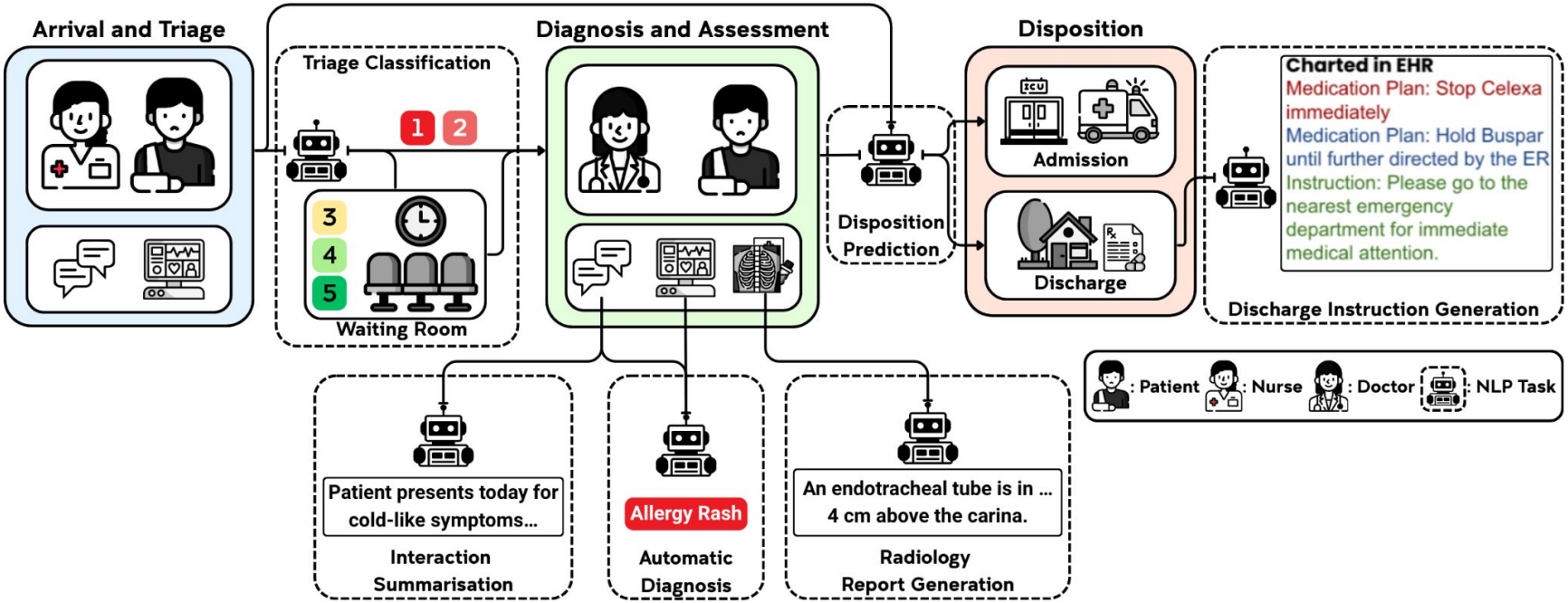


Trends in ED NLP



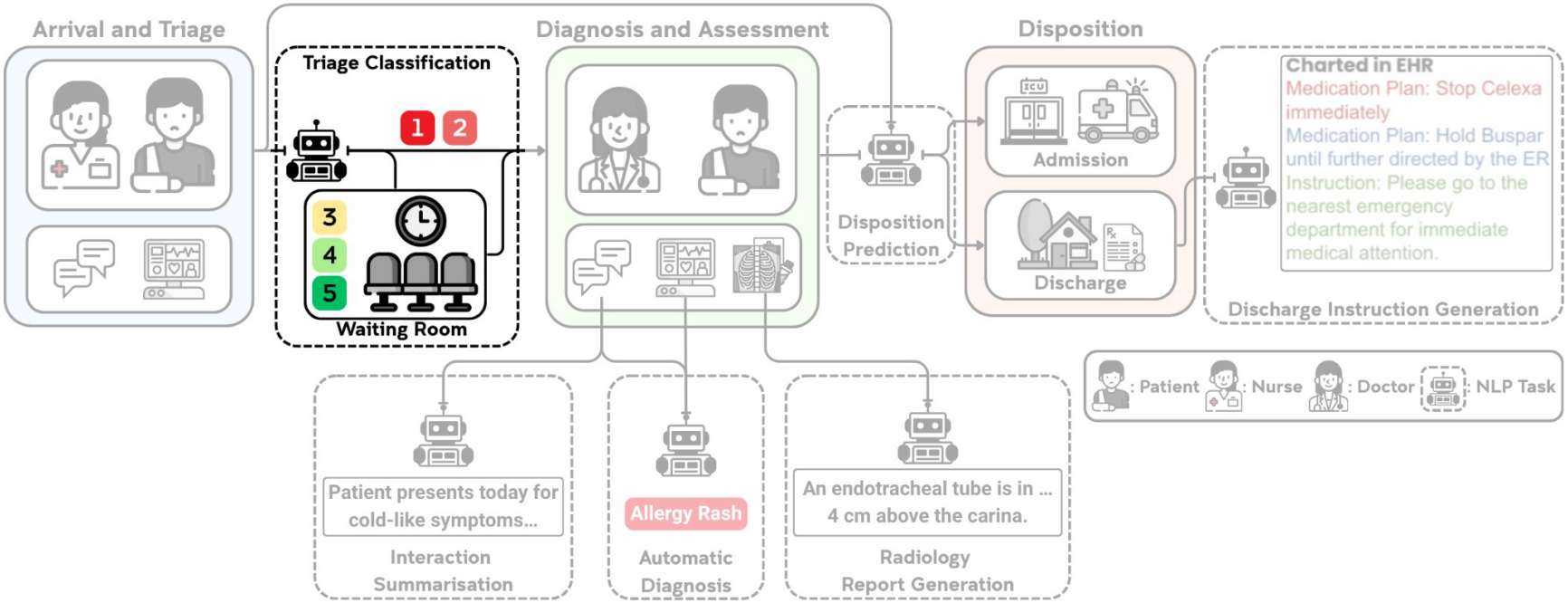
Diagnosis is the most researched phase

Landscape of ED NLP



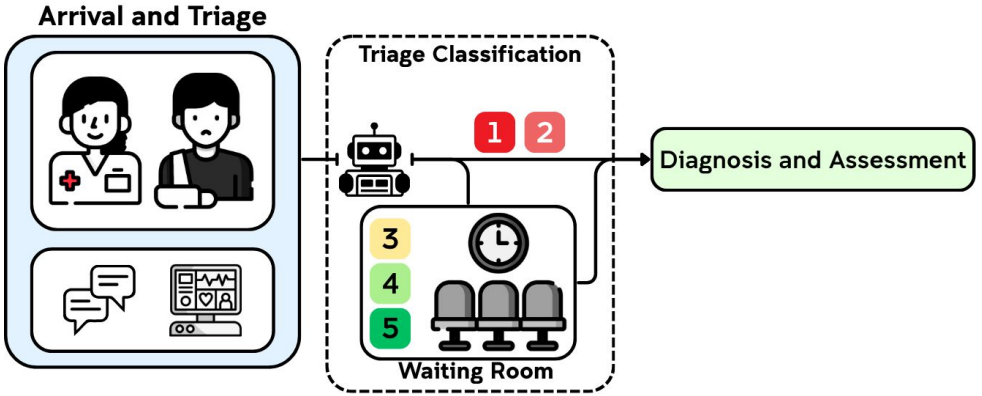
$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid x)$$

Landscape of ED NLP



$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid x)$$

Triage Classification



$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid x)$$

Output space

$$\mathcal{Y} = \{1 \prec 2 \prec \cdots \prec K\}$$

Meet our patient 

Male, 24, walked in, right hand in a cast.

"I've got this sharp pain in my chest, and I can't catch my breath."

Heart Rate	Resp. Rate	O_2 Sat.	BP	Temp.
118	24	93	128/78	98.2

Triage Classification



Male, 24, walked in, right hand in a cast.

"I've got this sharp pain in my chest, and I can't catch my breath."

HR	RR	O_2 Sat.	BP	Temp.
118	24	93	128/78	98.2

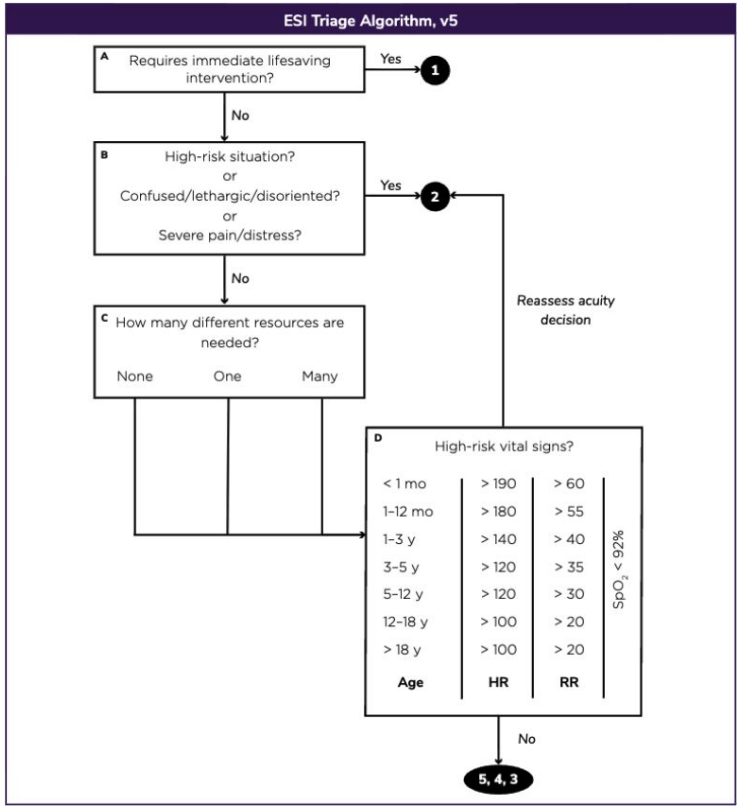
1 Most severe

2

3

4

5 Least severe



Triage Classification



Male, 24, walked in, right hand in a cast.

"I've got this sharp pain in my chest, and I can't catch my breath."

HR	RR	O_2 Sat.	BP	Temp.
118	24	93	128/78	98.2

1 Most severe

2

3

4

5 Least severe

Under-triage

- A level too low: The patient waits when they should not.

Over-triage

- They are seen sooner than needed.

Triage Classification as Sequence Classification



Male, 24, walked in, right hand in a cast.

"I've got this sharp pain in my chest, and I can't catch my breath."

HR	RR	O_2 Sat.	BP	Temp.
118	24	93	128/78	98.2

1 Most severe

2

3

4

5 Least severe

Under-triage

- A level too low: The patient waits when they should not.

Over-triage

- They are seen sooner than needed.

As Sequence Classification

medBERT.de: A comprehensive German BERT model for the medical domain (**Bressem et al., 2023**)

Towards ML-supported Triage Prediction in Real-World Emergency Room Scenarios (**Maschhur et al., 2024**)

As Sequence Classification

x is:

- Chief Complaint
- Vitals

> ***"I've got this sharp pain in my chest, and I can't catch my breath."***

medBERT.de: A comprehensive German BERT model for the medical domain (**Bressem et al., 2023**)

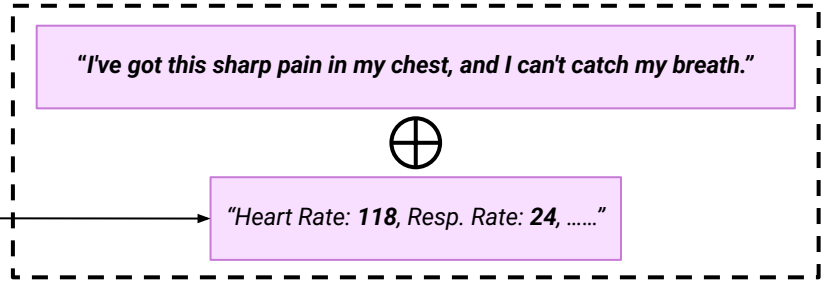
Towards ML-supported Triage Prediction in Real-World Emergency Room Scenarios (**Maschhur et al., 2024**)

As Sequence Classification

x is:

- Chief Complaint
- Vitals

HR	RR	O_2 Sat.	BP	Temp.
118	24	93	128/78	98.2



medBERT.de: A comprehensive German BERT model for the medical domain (**Bressem et al., 2023**)

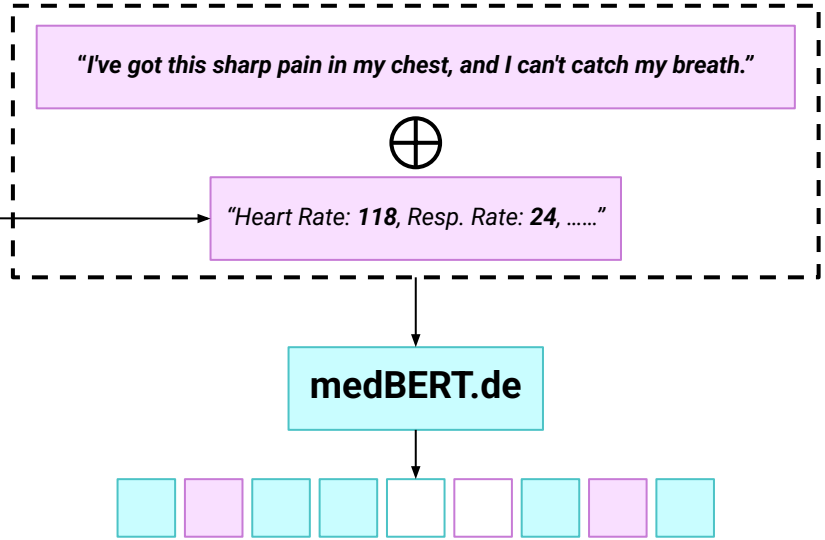
Towards ML-supported Triage Prediction in Real-World Emergency Room Scenarios (**Maschhur et al., 2024**)

As Sequence Classification

x is:

- Chief Complaint
- Vitals

HR	RR	O_2 Sat.	BP	Temp.
118	24	93	128/78	98.2



medBERT.de: A comprehensive German BERT model for the medical domain (**Bressem et al., 2023**)

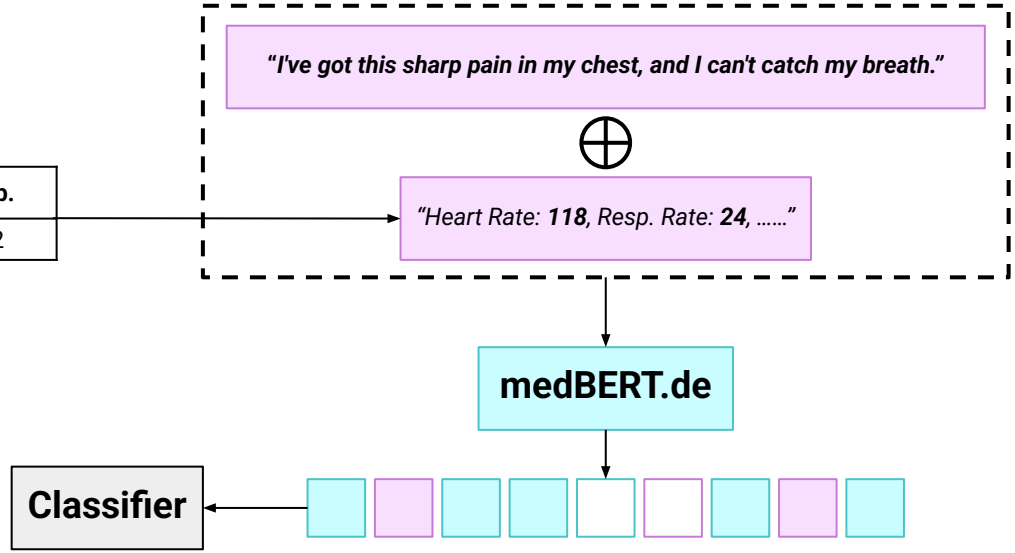
Towards ML-supported Triage Prediction in Real-World Emergency Room Scenarios (**Maschhur et al., 2024**)

As Sequence Classification

x is:

- Chief Complaint
- Vitals

HR	RR	O_2 Sat.	BP	Temp.
118	24	93	128/78	98.2



medBERT.de: A comprehensive German BERT model for the medical domain (**Bressem et al., 2023**)

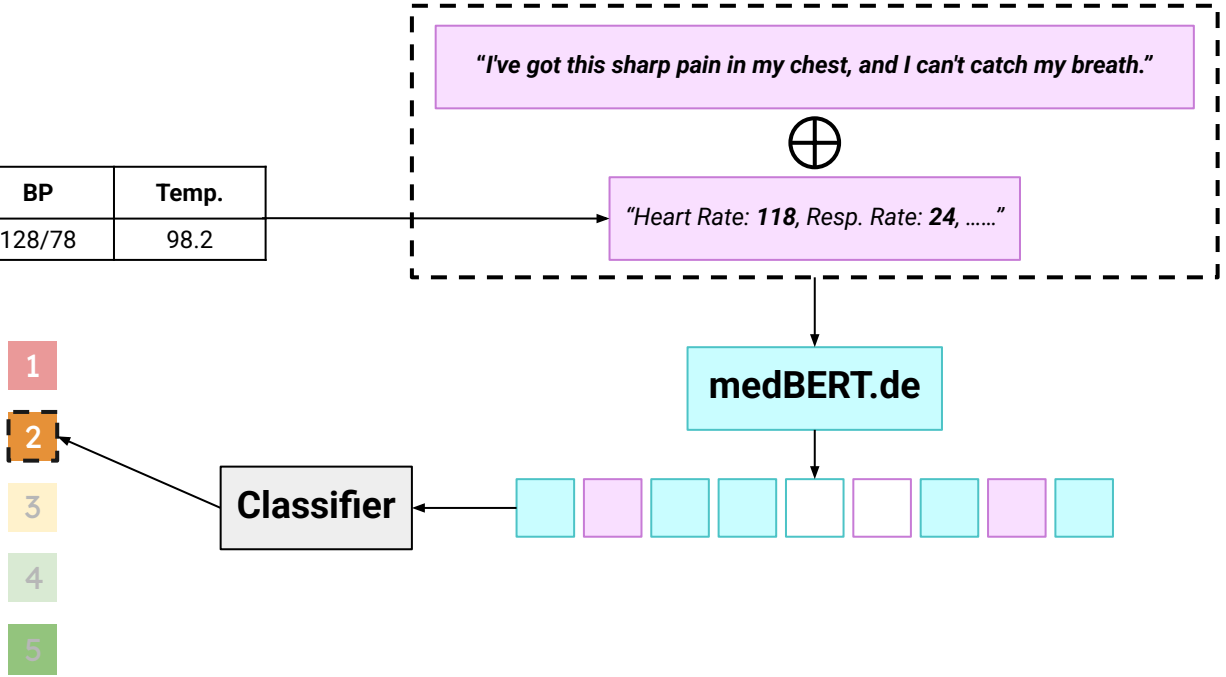
Towards ML-supported Triage Prediction in Real-World Emergency Room Scenarios (**Maschhur et al., 2024**)

As Sequence Classification

x is:

- Chief Complaint
- Vitals

HR	RR	O_2 Sat.	BP	Temp.
118	24	93	128/78	98.2



medBERT.de: A comprehensive German BERT model for the medical domain (**Bressem et al., 2023**)

Towards ML-supported Triage Prediction in Real-World Emergency Room Scenarios (**Maschhur et al., 2024**)

As Sequence Classification

x is:

- Chief Complaint
- Vitals

HR	RR	O_2 Sat.	BP	Temp.
118	24	93	128/78	98.2

"I've got this sharp pain in my chest, and I can't catch my breath."

2

As Sequence Classification

x is:

- Chief Complaint
- Vitals

HR	RR	O_2 Sat.	BP	Temp.
118	24	93	128/78	98.2

"I've got this sharp pain in my chest, and I can't catch my breath."

As Sequence Classification

x is:

- Chief Complaint
- Vitals
- **External Knowledge**
- **Patient Metadata**

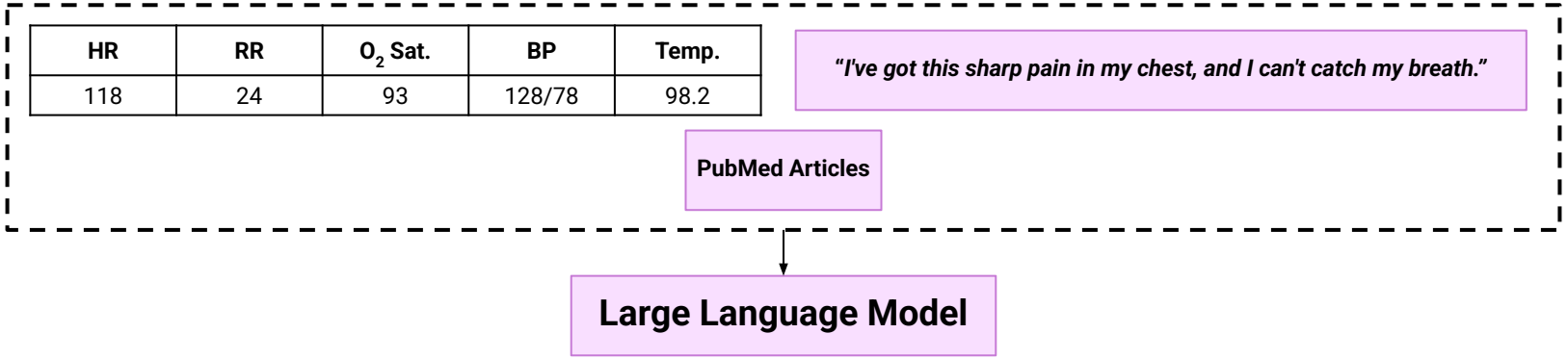
HR	RR	O_2 Sat.	BP	Temp.
118	24	93	128/78	98.2

"I've got this sharp pain in my chest, and I can't catch my breath."

As Retrieval-Augmented Generation

x is:

- Chief Complaint
- Vitals
- **External Knowledge**
- **Patient Metadata**



Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis (**Gaber et al., 2025**)

As Retrieval-Augmented Generation

x is:

- Chief Complaint
- Vitals
- External Knowledge
- Patient Metadata

HR	RR	O_2 Sat.	BP	Temp.
118	24	93	128/78	98.2

"I've got this sharp pain in my chest, and I can't catch my breath."

PubMed Articles

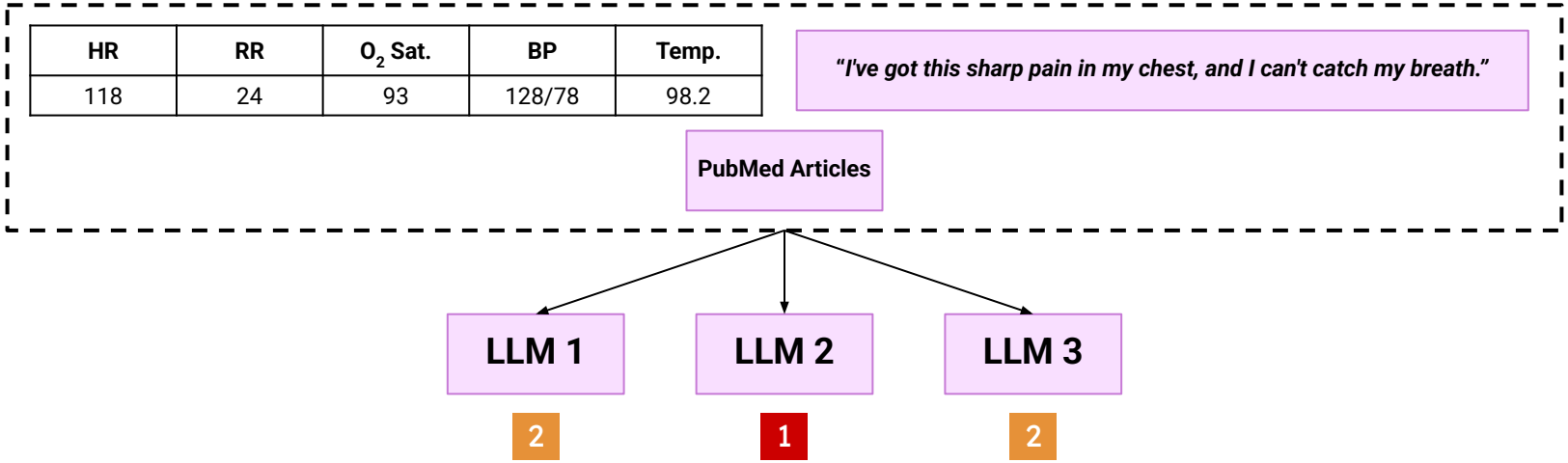
Large Language Model

Why stop at a single model?

As Retrieval-Augmented Generation

x is:

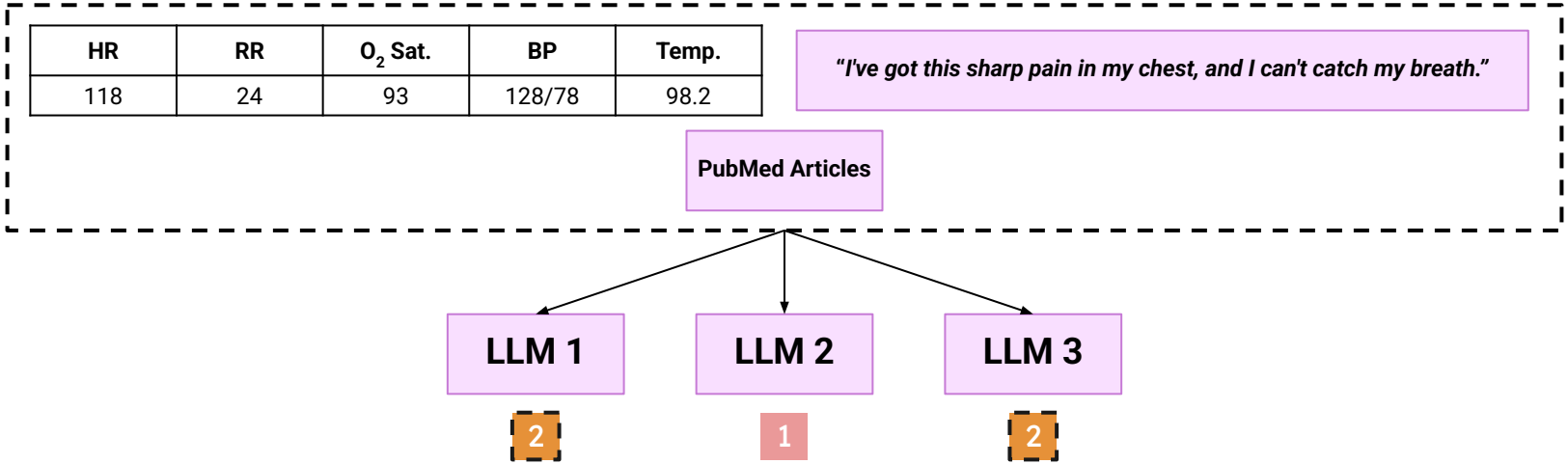
- Chief Complaint
- Vitals
- External Knowledge
- Patient Metadata



As Retrieval-Augmented Generation

x is:

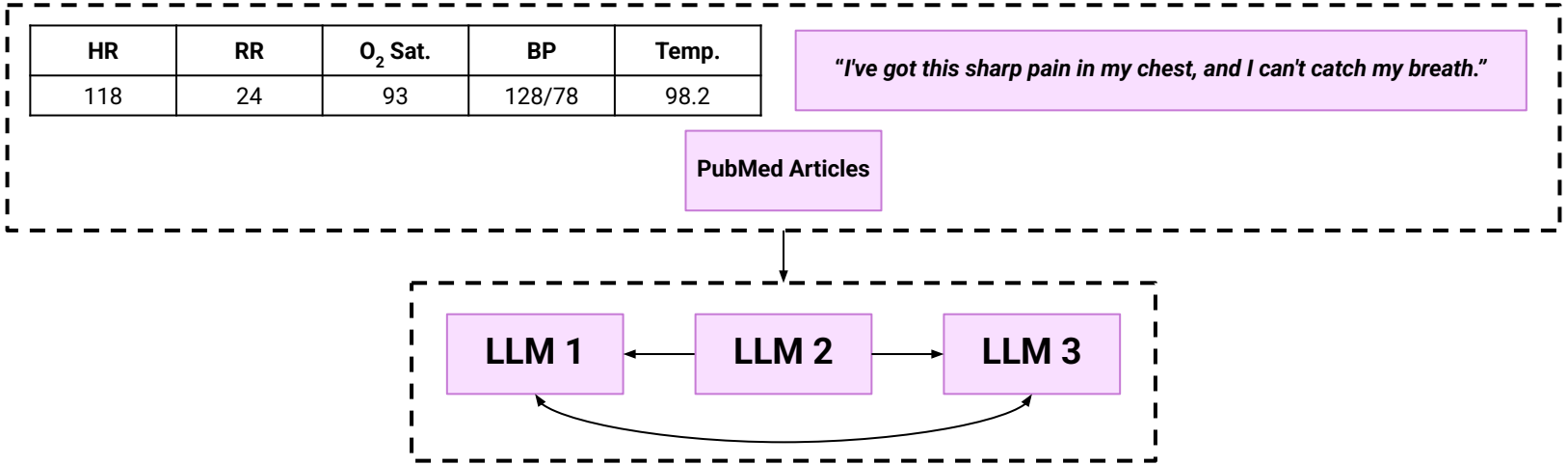
- Chief Complaint
- Vitals
- External Knowledge
- Patient Metadata



As Retrieval-Augmented Generation

x is:

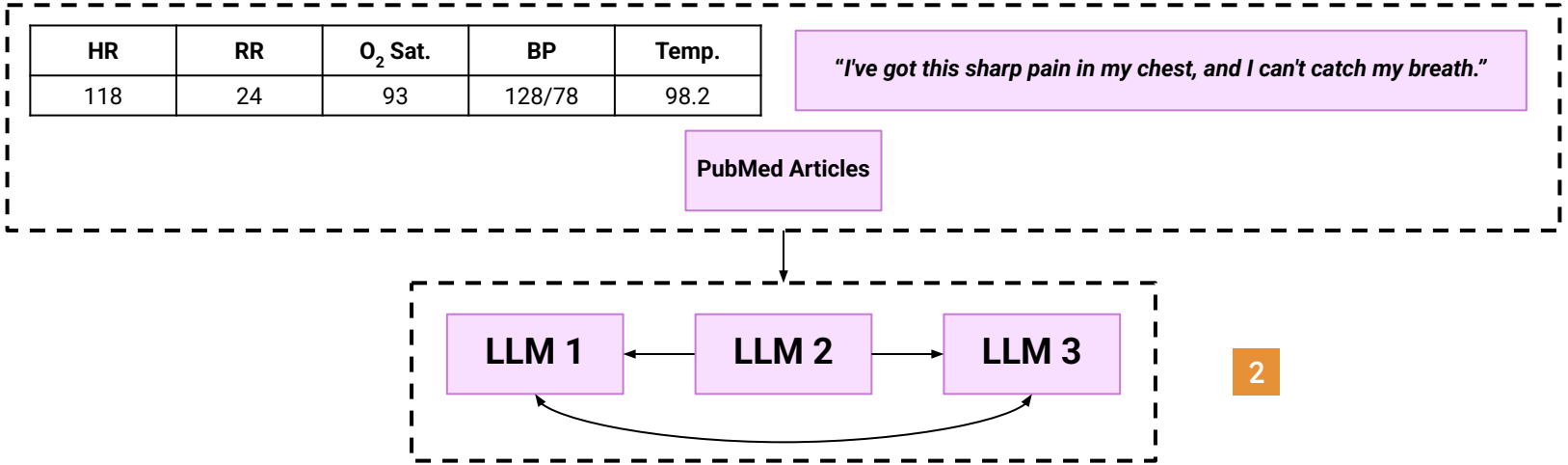
- Chief Complaint
- Vitals
- External Knowledge
- Patient Metadata



As Retrieval-Augmented Generation

x is:

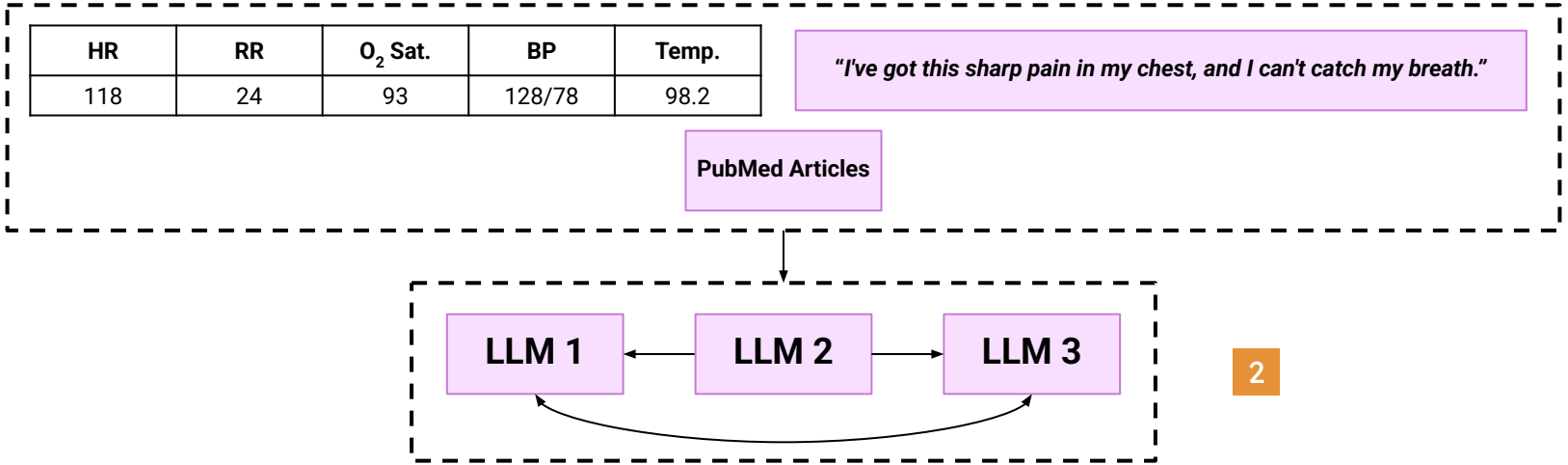
- Chief Complaint
- Vitals
- External Knowledge
- Patient Metadata



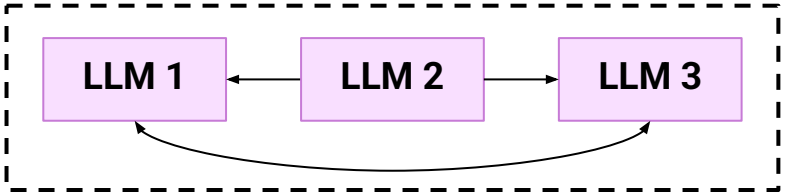
As Collaborative Discussion

x is:

- Chief Complaint
- Vitals
- External Knowledge
- Patient Metadata



As Collaborative Discussion



TriageAgent: Towards Better Multi-Agents Collaborations for Large Language Model-Based Clinical Triage (**Lu et al., 2024**)

As Collaborative Discussion

LLM 1

LLM 2

LLM 3

TriageAgent: Towards Better Multi-Agents Collaborations for Large Language Model-Based Clinical Triage (**Lu et al., 2024**)

As Collaborative Discussion

LLM 1 **Ortho**

- **Triage Level**
- **Confidence**
- **Reasoning**

LLM 2 **Cardio**

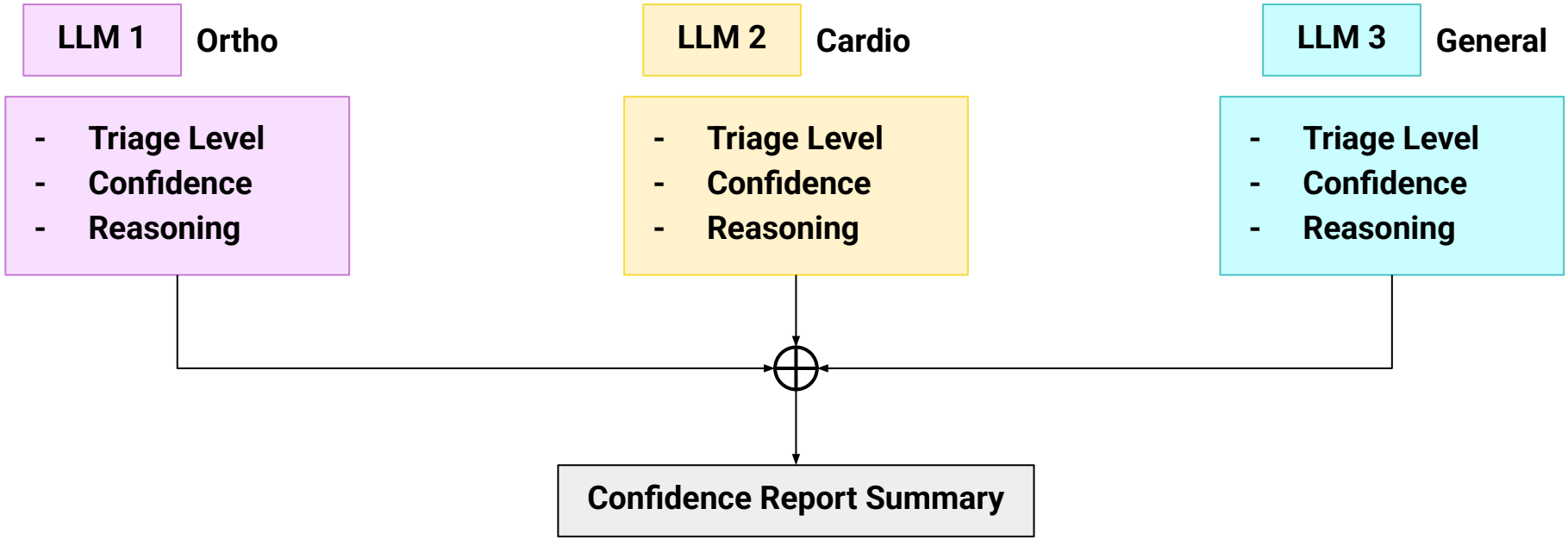
- **Triage Level**
- **Confidence**
- **Reasoning**

LLM 3 **General**

- **Triage Level**
- **Confidence**
- **Reasoning**

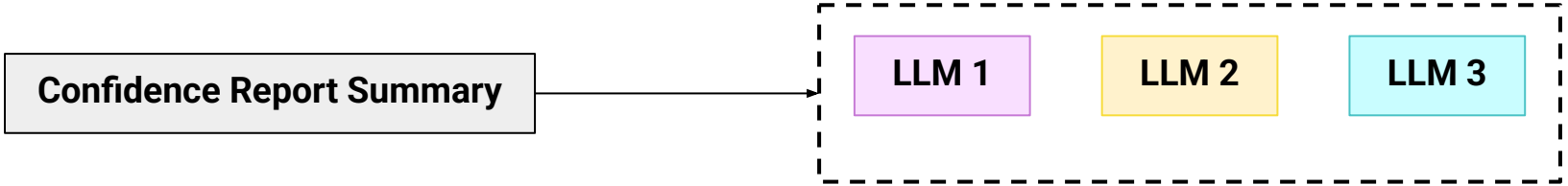
TriageAgent: Towards Better Multi-Agents Collaborations for Large Language Model-Based Clinical Triage (**Lu et al., 2024**)

As Collaborative Discussion



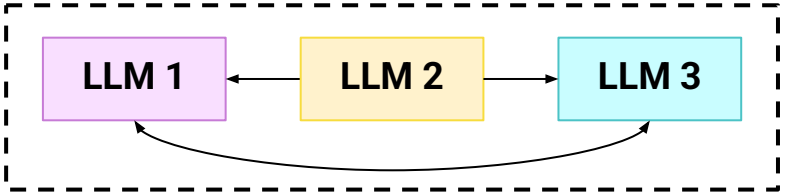
TriageAgent: Towards Better Multi-Agents Collaborations for Large Language Model-Based Clinical Triage (**Lu et al., 2024**)

As Collaborative Discussion



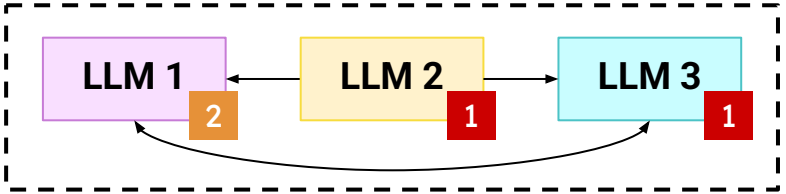
TriageAgent: Towards Better Multi-Agents Collaborations for Large Language Model-Based Clinical Triage (**Lu et al., 2024**)

As Collaborative Discussion



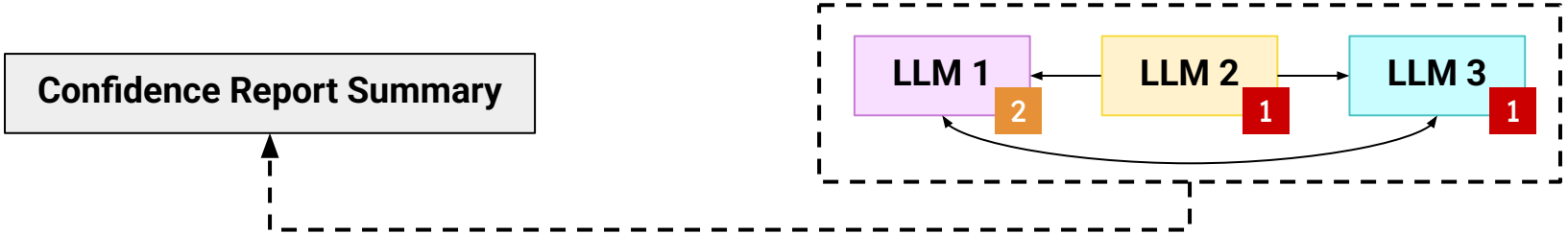
TriageAgent: Towards Better Multi-Agents Collaborations for Large Language Model-Based Clinical Triage (**Lu et al., 2024**)

As Collaborative Discussion



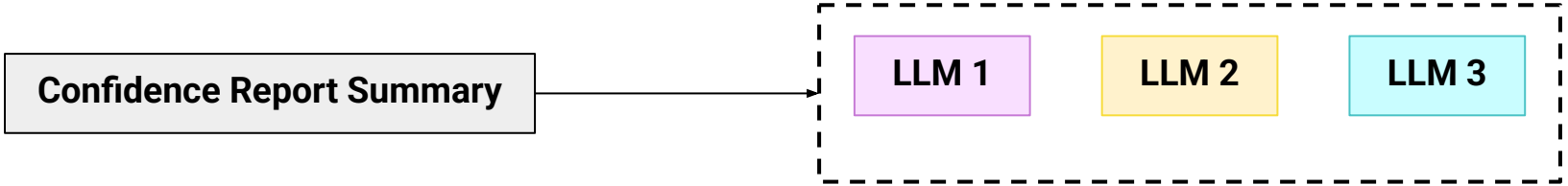
TriageAgent: Towards Better Multi-Agents Collaborations for Large Language Model-Based Clinical Triage (**Lu et al., 2024**)

As Collaborative Discussion



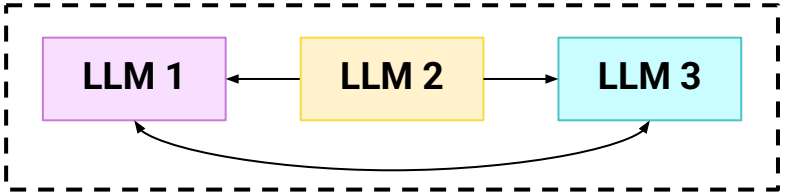
TriageAgent: Towards Better Multi-Agents Collaborations for Large Language Model-Based Clinical Triage (**Lu et al., 2024**)

As Collaborative Discussion



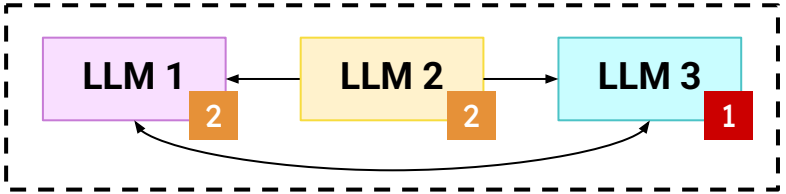
TriageAgent: Towards Better Multi-Agents Collaborations for Large Language Model-Based Clinical Triage (**Lu et al., 2024**)

As Collaborative Discussion



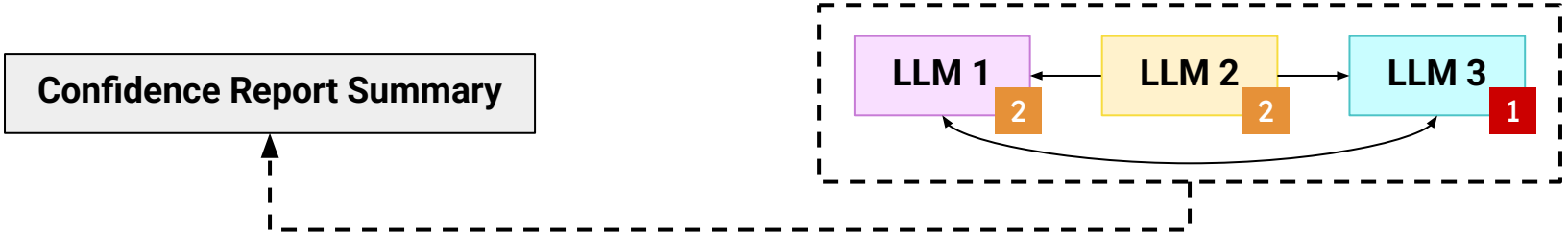
TriageAgent: Towards Better Multi-Agents Collaborations for Large Language Model-Based Clinical Triage (**Lu et al., 2024**)

As Collaborative Discussion



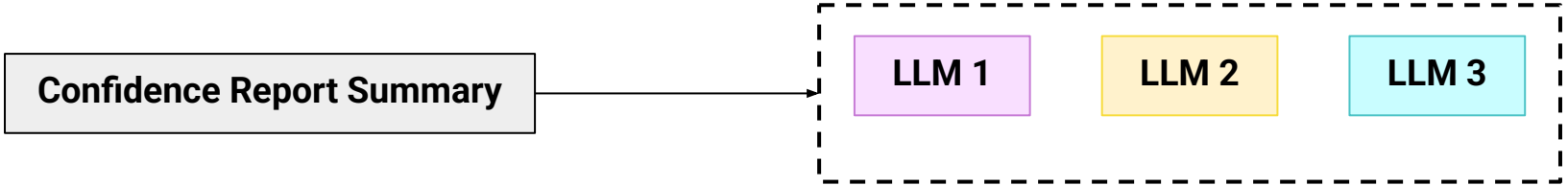
TriageAgent: Towards Better Multi-Agents Collaborations for Large Language Model-Based Clinical Triage (**Lu et al., 2024**)

As Collaborative Discussion



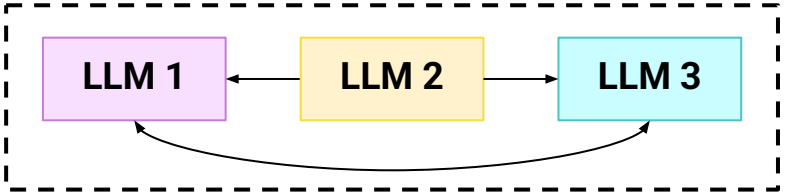
TriageAgent: Towards Better Multi-Agents Collaborations for Large Language Model-Based Clinical Triage (**Lu et al., 2024**)

As Collaborative Discussion



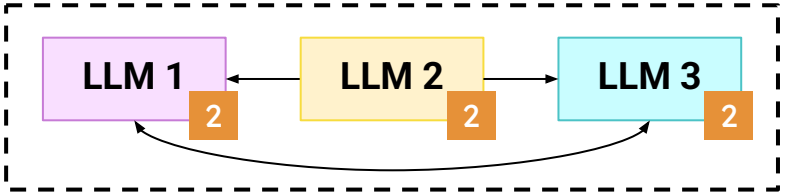
TriageAgent: Towards Better Multi-Agents Collaborations for Large Language Model-Based Clinical Triage (**Lu et al., 2024**)

As Collaborative Discussion



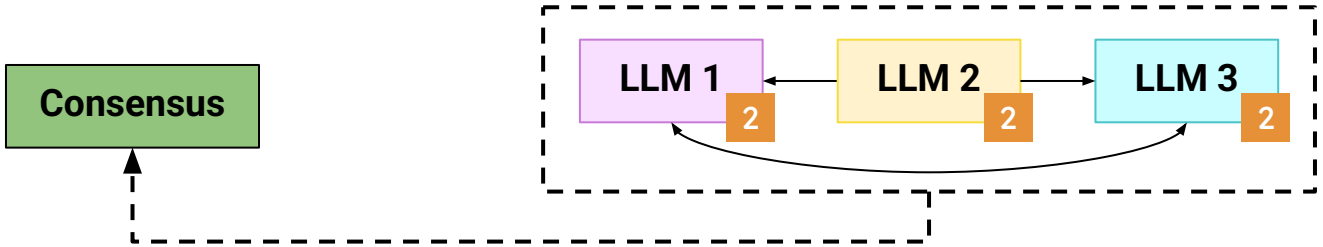
TriageAgent: Towards Better Multi-Agents Collaborations for Large Language Model-Based Clinical Triage (**Lu et al., 2024**)

As Collaborative Discussion



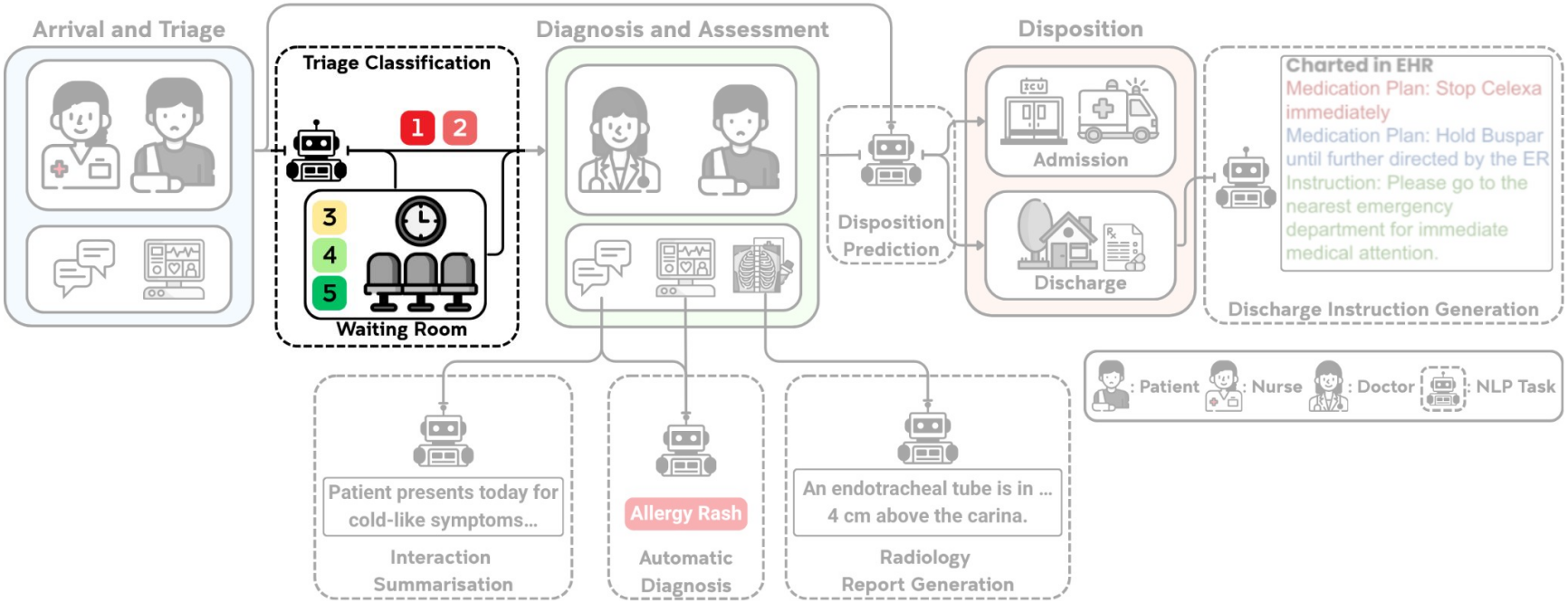
TriageAgent: Towards Better Multi-Agents Collaborations for Large Language Model-Based Clinical Triage (**Lu et al., 2024**)

As Collaborative Discussion



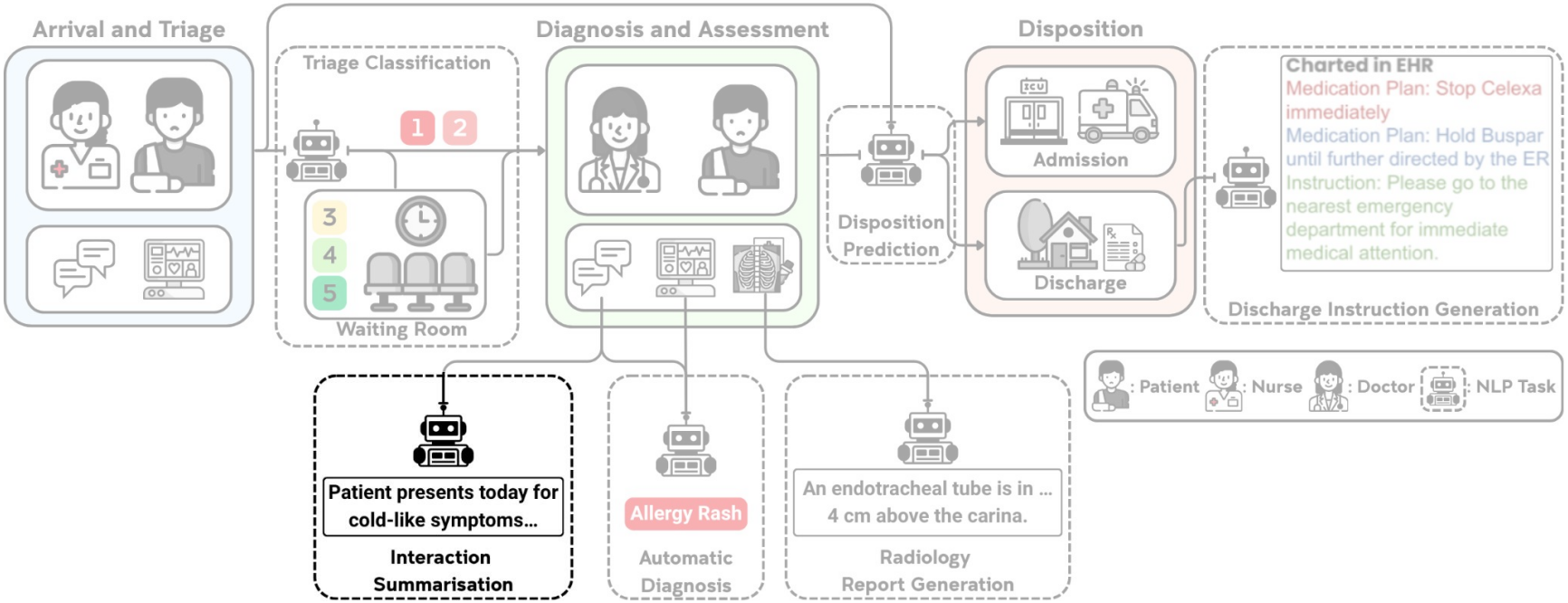
TriageAgent: Towards Better Multi-Agents Collaborations for Large Language Model-Based Clinical Triage (**Lu et al., 2024**)

Landscape of ED NLP



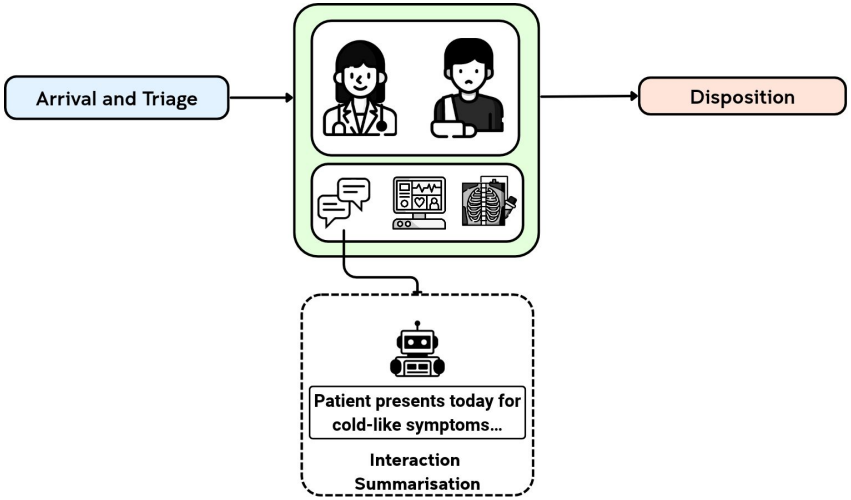
$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid x)$$

Landscape of ED NLP



$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid x)$$

Interaction Summarisation



$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid x)$$

Output space

$\mathcal{Y} =$ S.O.A.P Note

Interaction Summarisation



Male, 24, walked in, right hand in a cast.

"I've got this sharp pain in my chest, and I can't catch my breath."

HR	RR	O_2 Sat.	BP	Temp.
118	24	93	128/78	98.2

+ Conversation with the doctor
"It stabs when I breath. No blood" "They put pins in last week."
"Legs are fine, I think."

Structured Summary

Subjective: *What did they report?*

Objective: *Vitals and exams.*

Assessment: *What might it be?*

Plan: *What is the treatment?*

Interaction Summarisation



Male, 24, walked in, right hand in a cast.

"I've got this sharp pain in my chest, and I can't catch my breath."

HR	RR	O_2 Sat.	BP	Temp.
118	24	93	128/78	98.2

+ Conversation with the doctor
"It stabs when I breath. No blood" "They put pins in last week."
"Legs are fine, I think."

Structured Summary

Subjective: *What did they report?*

Objective: *Vitals and exams.*

Assessment: *What might it be?*

Plan: *What is the treatment?*

Omission:

- A pertinent negative (*no blood*) or the procedure is left out.

Hallucination

- A note says something they never said (*has migraines*).

As Sequence-to-Sequence Generation

x is:

- Dialogue

> *"… P: It stabs when I breathe in. No blood. D: What happened to your hand? P: They put pins in my scaphoid two weeks ago."*

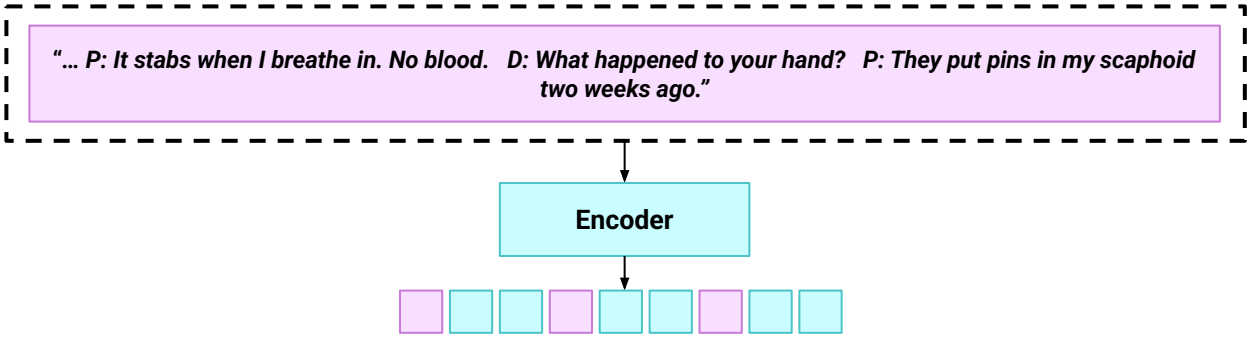
Sequence to Sequence Learning with Neural Networks (**Sutskever et al., 2014**)

Get To The Point: Summarization with Pointer-Generator Networks (**See et al., 2017**)

As Sequence-to-Sequence Generation

x is:

- Dialogue



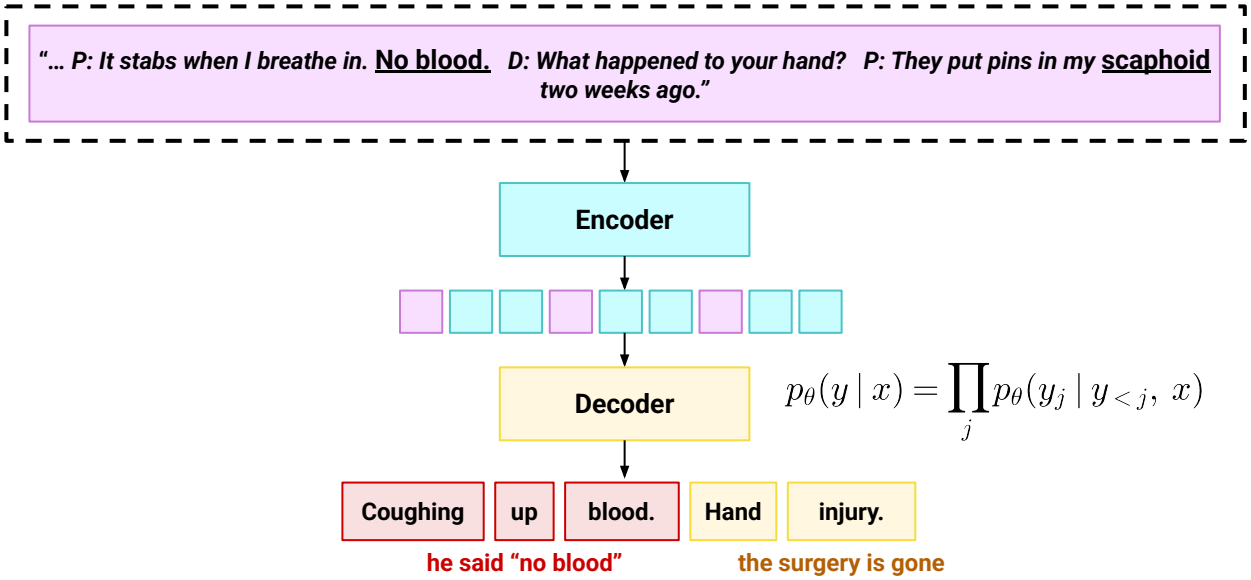
Sequence to Sequence Learning with Neural Networks (**Sutskever et al., 2014**)

Get To The Point: Summarization with Pointer-Generator Networks (**See et al., 2017**)

As Sequence-to-Sequence Generation

x is:

- Dialogue



Sequence to Sequence Learning with Neural Networks (**Sutskever et al., 2014**)

Get To The Point: Summarization with Pointer-Generator Networks (**See et al., 2017**)

As Sequence-to-Sequence Generation

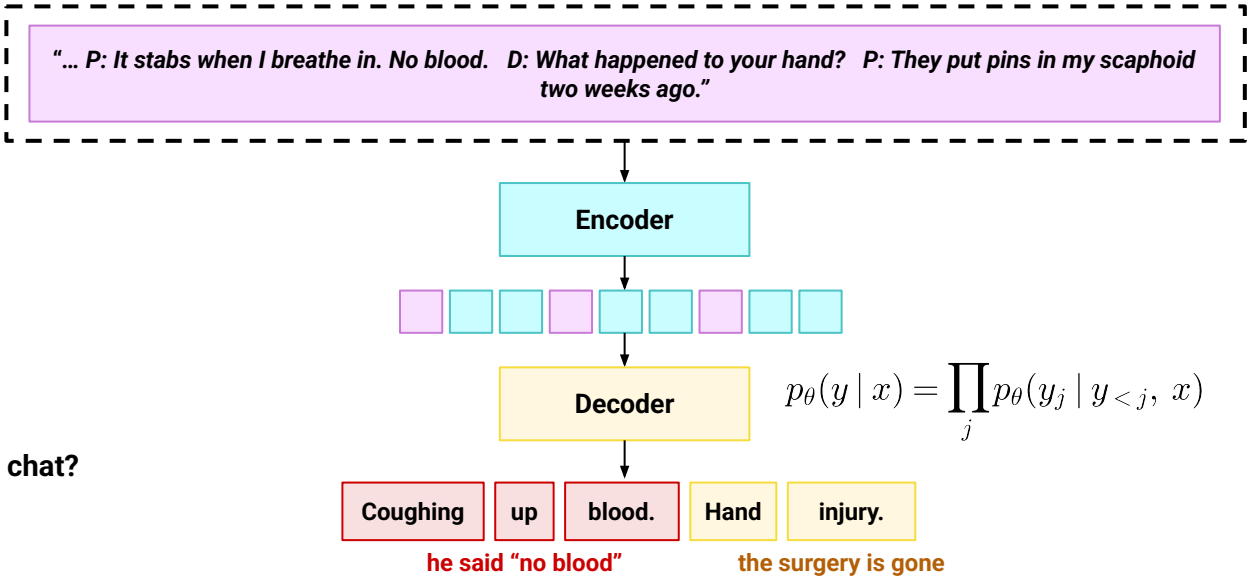
x is:

- Dialogue

Every word written from V
nothing ties it to the chat

Hallucination and **omission**
fluent, but not faithful

What if it could copy their words from the chat?



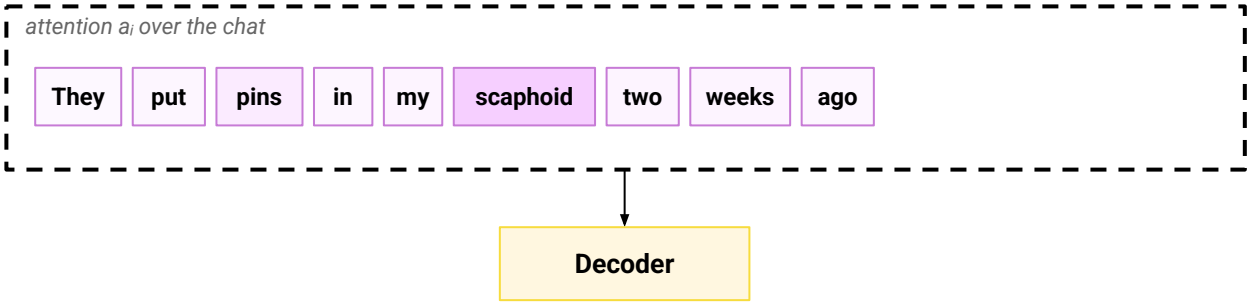
Sequence to Sequence Learning with Neural Networks (**Sutskever et al., 2014**)

Get To The Point: Summarization with Pointer-Generator Networks (**See et al., 2017**)

As Copy-or-Generate

x is:

- Dialogue

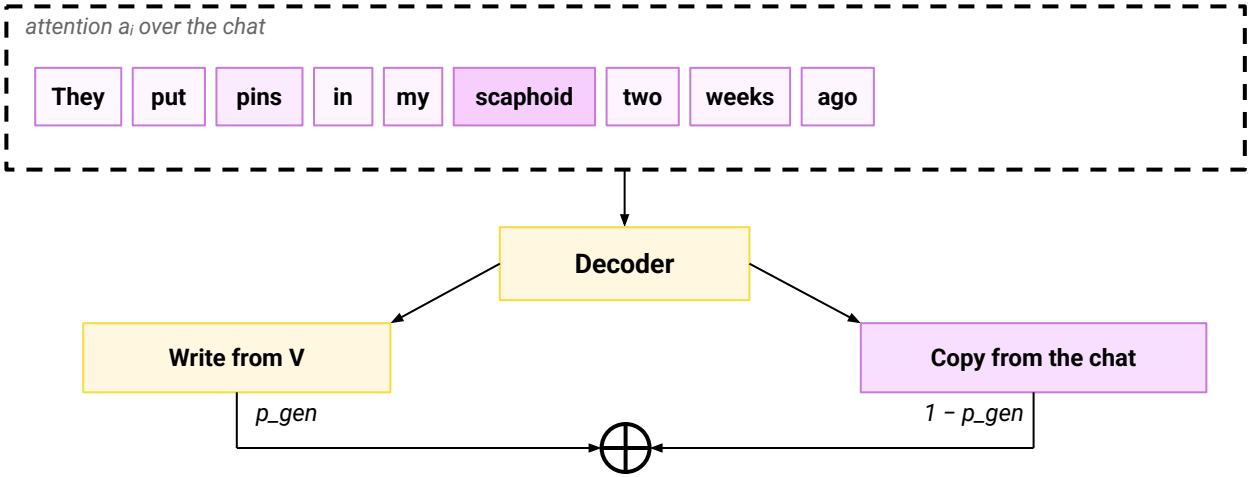


Dr. Summarize: Global Summarization of Medical Dialogue by Exploiting Local Structures (**Joshi et al., 2020**)

As Copy-or-Generate

x is:

- Dialogue



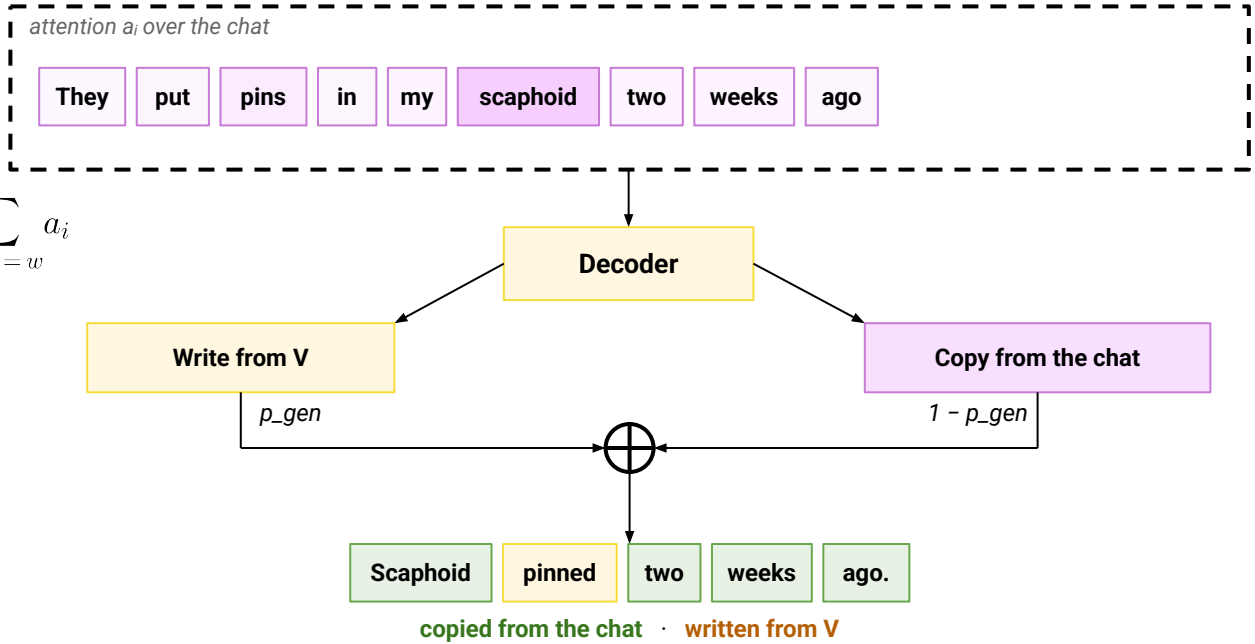
Get To The Point: Summarization with Pointer-Generator Networks (**See et al., 2017**)

As Copy-or-Generate

x is:

- Dialogue

$$P(w) = p_{\text{gen}}\, P_{\text{vocab}}(w) + (1 - p_{\text{gen}}) \sum_{i\,:\,w_i = w} a_i$$

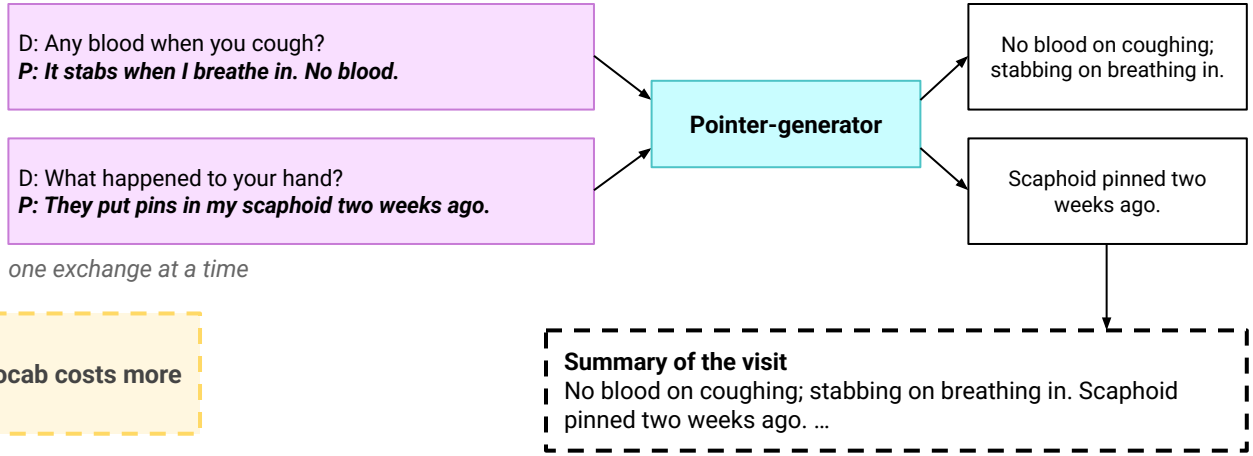


Dr. Summarize: Global Summarization of Medical Dialogue by Exploiting Local Structures (**Joshi et al., 2020**)

As Copy-or-Generate

x is:

- Dialogue



$$\mathcal{L} = \mathcal{L}_{\text{NLL}} + \delta \cdot p_{\text{gen}}$$ **Writing from vocab costs more**

Dr. Summarize: Global Summarization of Medical Dialogue by Exploiting Local Structures (**Joshi et al., 2020**)

Interaction Summarisation

As Extract, Cluster, then Write

x is:

- Dialogue

D: What brings you in today?

P: ***Sharp pain in my chest, can't catch my breath.***

D: Any blood when you cough?

P: ***It stabs when I breathe in. No blood.***

D: What happened to your hand?

P: ***They put pins in my scaphoid two weeks ago.***

D: I'd like a scan of your lungs, to rule out a clot.

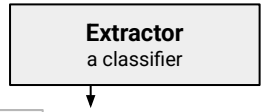
P: ***Okay.***

Generating SOAP Notes from Doctor-Patient Conversations Using Modular Summarization Techniques (**Krishna et al., 2021**)

As Extract, Cluster, then Write

x is:

- Dialogue



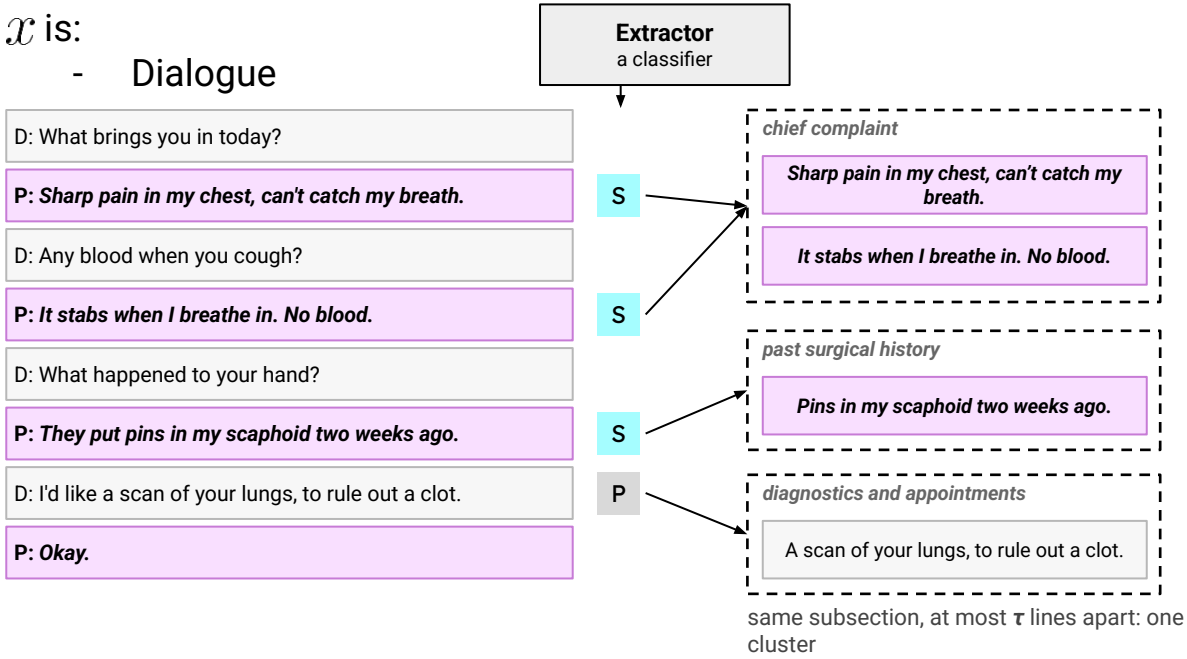
D: What brings you in today?	
P: ***Sharp pain in my chest, can't catch my breath.***	S
D: Any blood when you cough?	
P: ***It stabs when I breathe in. No blood.***	S
D: What happened to your hand?	
P: ***They put pins in my scaphoid two weeks ago.***	S
D: I'd like a scan of your lungs, to rule out a clot.	P
P: ***Okay.***	

Generating SOAP Notes from Doctor-Patient Conversations Using Modular Summarization Techniques (**Krishna et al., 2021**)

As Extract, Cluster, then Write

x is:

- Dialogue

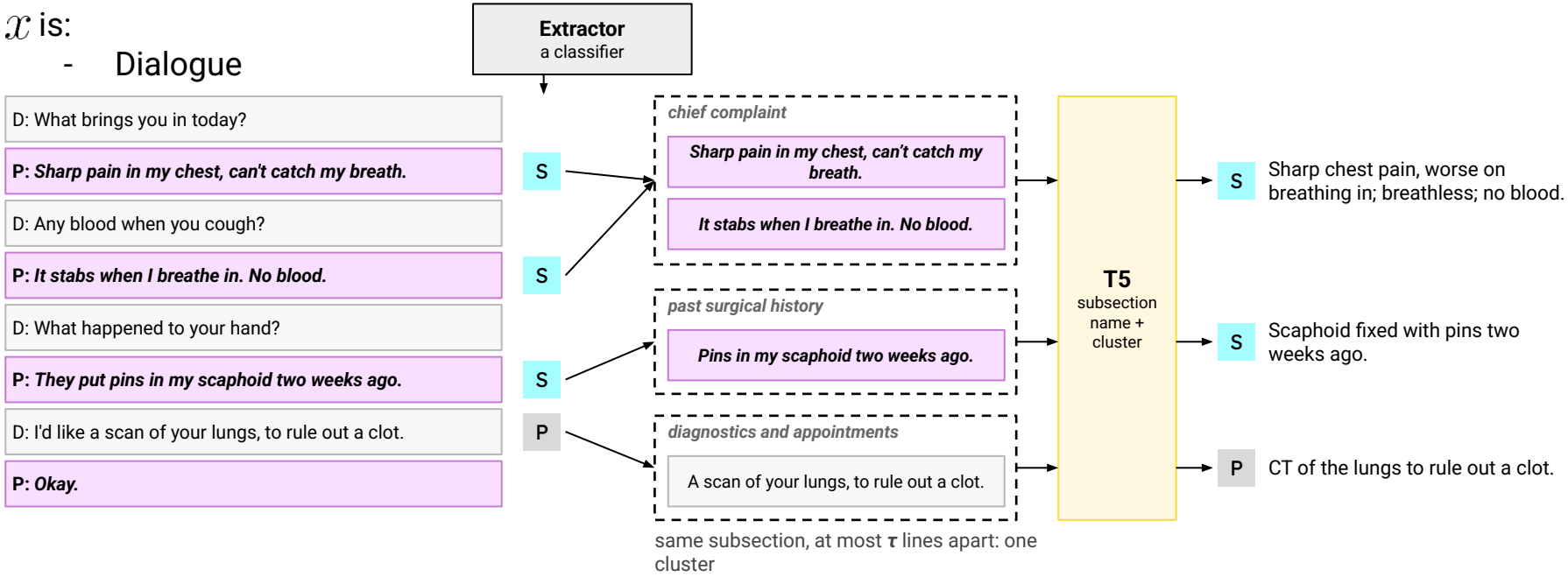


Generating SOAP Notes from Doctor-Patient Conversations Using Modular Summarization Techniques (**Krishna et al., 2021**)

As Extract, Cluster, then Write

x is:

- Dialogue



Generating SOAP Notes from Doctor-Patient Conversations Using Modular Summarization Techniques (**Krishna et al., 2021**)

Interaction Summarisation

As a Chain of Prompts

x is:

- Dialogue
- **Instruction**

1. Reason for visit

chest pain, breathlessness

1. Reason for visit

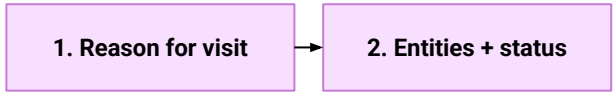
Generating medically-accurate summaries of patient-provider dialogue: A multi-stage approach using large language models (**Nair et al., 2023**)

Interaction Summarisation

As a Chain of Prompts

x is:

- Dialogue
- **Instruction**



1. Reason for visit

chest pain, breathlessness

2. Entities, turn by turn

pleuritic chest pain	**present**
blood when coughing	**absent**
scaphoid fracture, pinned	**present**
leg swelling	**unknown**

Generating medically-accurate summaries of patient-provider dialogue: A multi-stage approach using large language models (**Nair et al., 2023**)

Interaction Summarisation

As a Chain of Prompts

x is:

- Dialogue
- **Instruction**



1. Reason for visit

chest pain, breathlessness

2. Entities, turn by turn

pleuritic chest pain	**present**
blood when coughing	**absent**
scaphoid fracture, pinned	**present**
leg swelling	**unknown**

One more GPT-3 call

In: the whole dialogue, and every entity found so far with its status
Out: the statuses, corrected

Why: step 2 sees one question and its answer at a time, so an answer that comes later in the visit is missed.

e.g. D: Any fever? P: Not sure.
later P: I checked last night, it was 38.5.
fever: unknown → present

Our patient: nothing later settles his legs, so leg swelling stays unknown.

Generating medically-accurate summaries of patient-provider dialogue: A multi-stage approach using large language models (**Nair et al., 2023**)

Interaction Summarisation

As a Chain of Prompts

x is:

- Dialogue
- **Instruction**

1. Reason for visit → 2. Entities + status → 3. Resolve unknowns → 4. Write the summary

1. Reason for visit

chest pain, breathlessness

2. Entities, turn by turn

pleuritic chest pain	**present**
blood when coughing	**absent**
scaphoid fracture, pinned	**present**
leg swelling	**unknown**

Reason for visit
chest pain, breathlessness

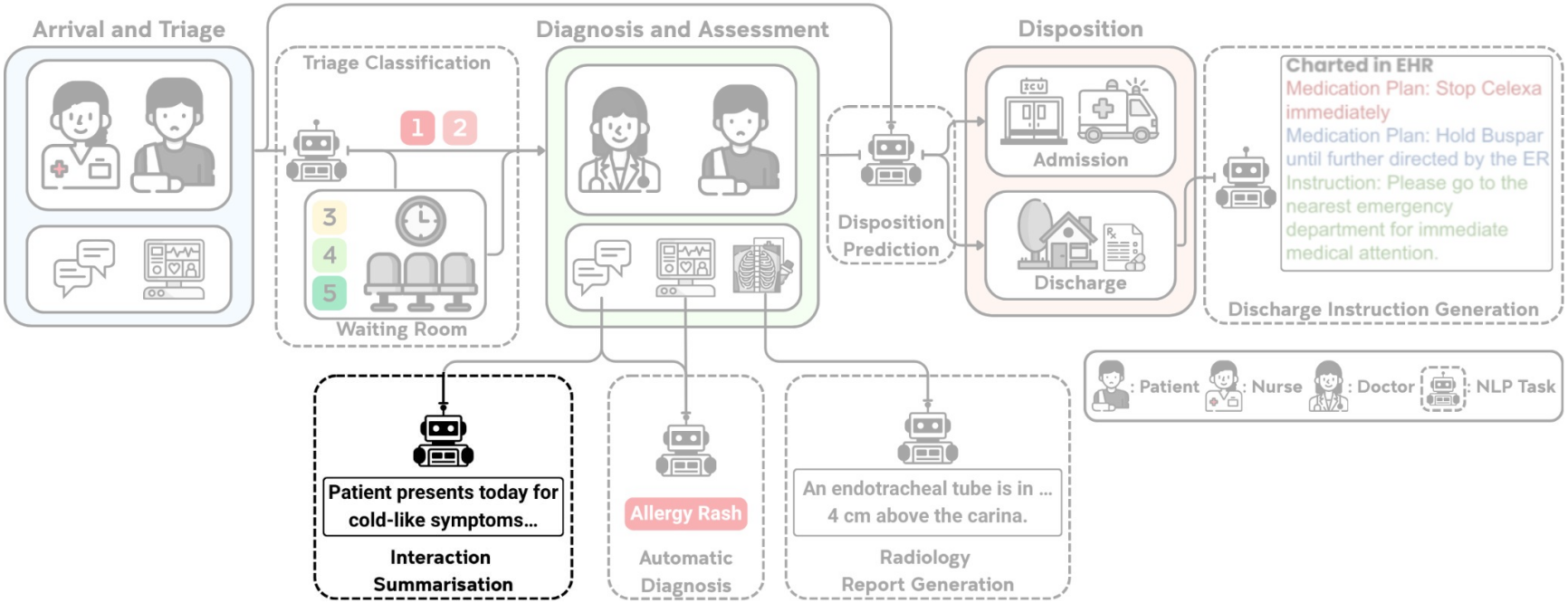
Pertinent positives
pleuritic chest pain; scaphoid fracture, pinned two weeks ago

Pertinent unknowns
leg swelling

Pertinent negatives
blood when coughing

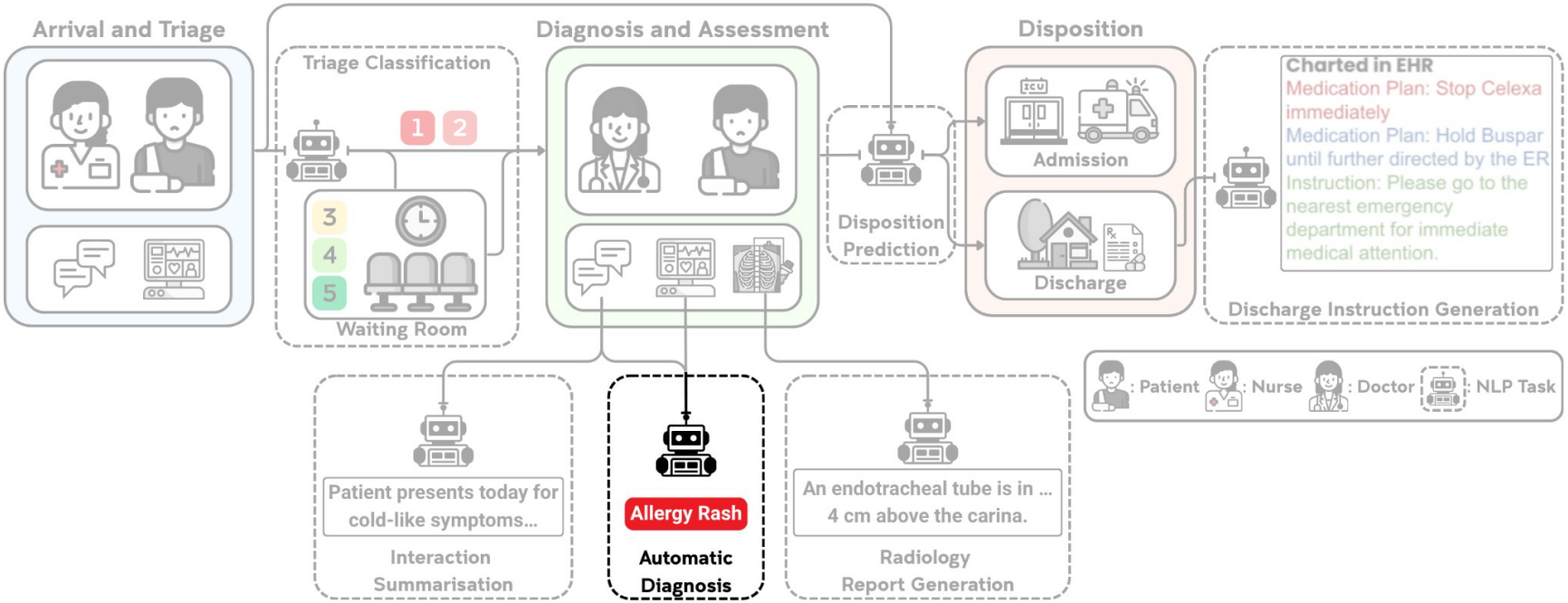
Generating medically-accurate summaries of patient-provider dialogue: A multi-stage approach using large language models (**Nair et al., 2023**)

Landscape of ED NLP



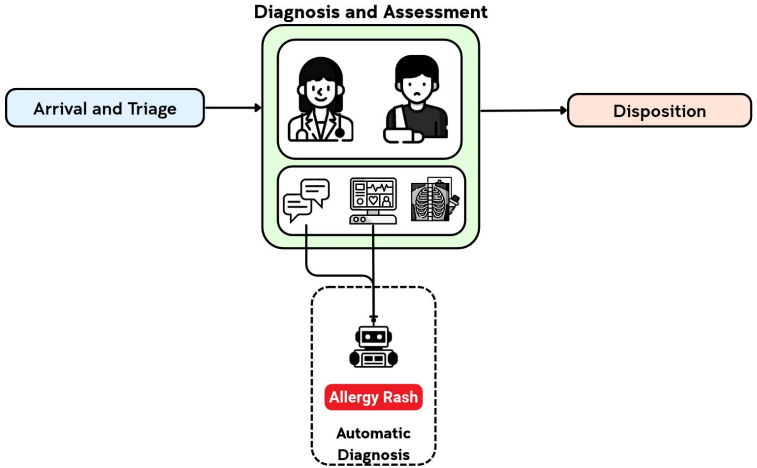
$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid x)$$

Landscape of ED NLP



$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid x)$$

Automatic Diagnosis



$$a_t = \arg\max_{a \in \mathcal{A}_{\text{inquire}} \cup \mathcal{D}} \pi_\theta(a \mid h_t)$$

Output space

$\mathcal{A}_{\text{inquire}} \cup \mathcal{D}$

Automatic Diagnosis



Male, 24, walked in, right hand in a cast.

"I've got this sharp pain in my chest, and I can't catch my breath."

HR	RR	O_2 Sat.	BP	Temp.
118	24	93	128/78	98.2

+ History built so far

"It stabs when I breath. No blood"

- 4-hour window
- Crowded EDs

? Ask: any recent surgery?

? Ask: any leg swelling?

D Diagnose: pulmonary embolism

D Diagnose: pneumothorax

Stop too early

- A disease is named before the surgery is asked about, and the clot is missed.

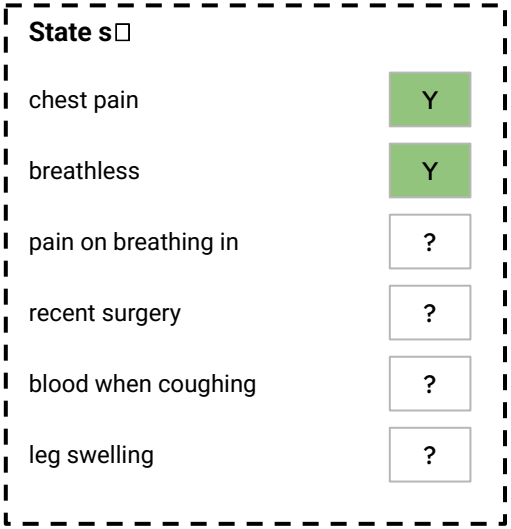
Ask too long

- Every question costs time the ED does not have.

As a Markov Decision Process

x is:

- Symptoms so far
- Turn

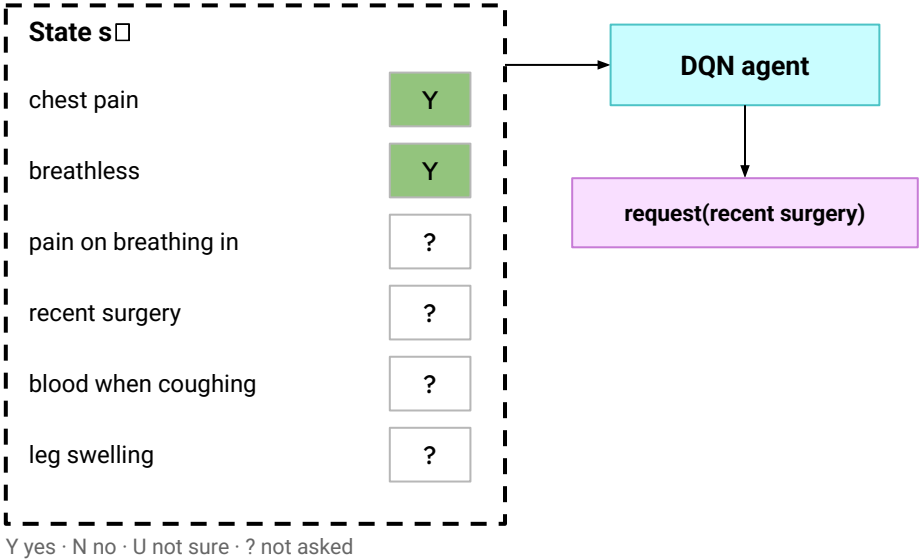


Task-oriented Dialogue System for Automatic Diagnosis (**Wei et al., 2018**)

As a Markov Decision Process

x is:

- Symptoms so far
- Turn

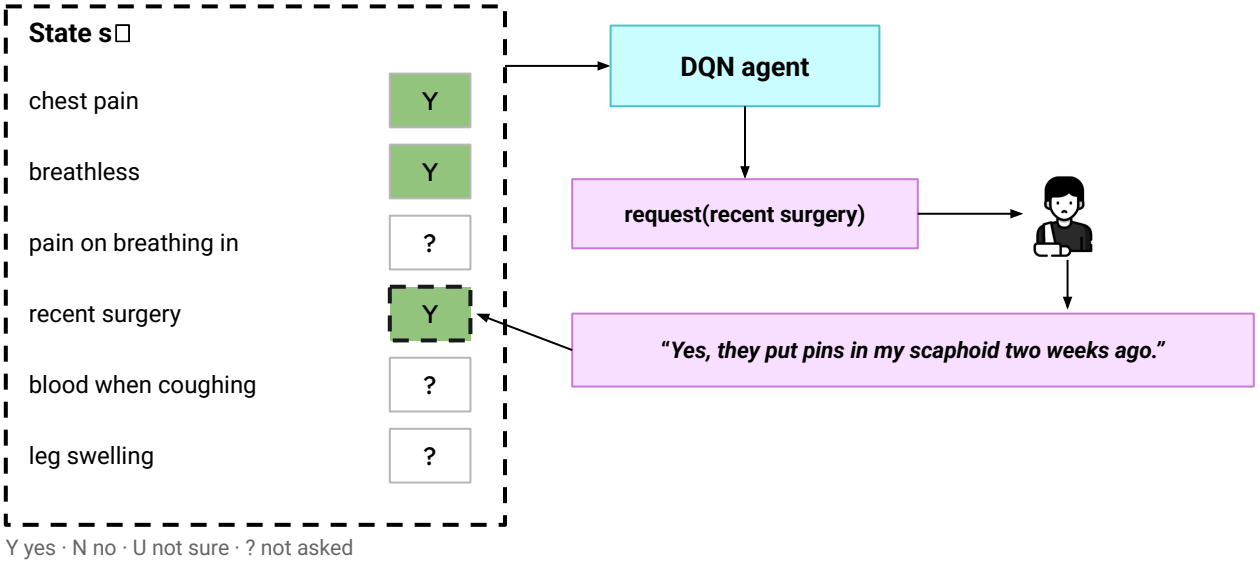


Task-oriented Dialogue System for Automatic Diagnosis (**Wei et al., 2018**)

As a Markov Decision Process

x is:

- Symptoms so far
- Turn

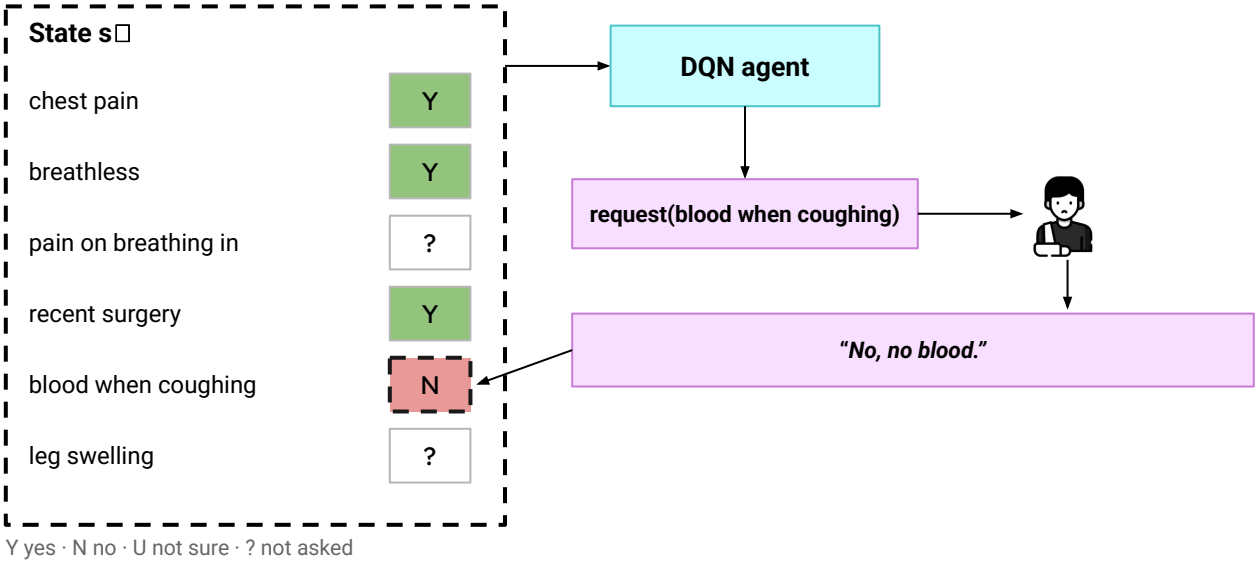


Task-oriented Dialogue System for Automatic Diagnosis (**Wei et al., 2018**)

As a Markov Decision Process

x is:

- Symptoms so far
- Turn

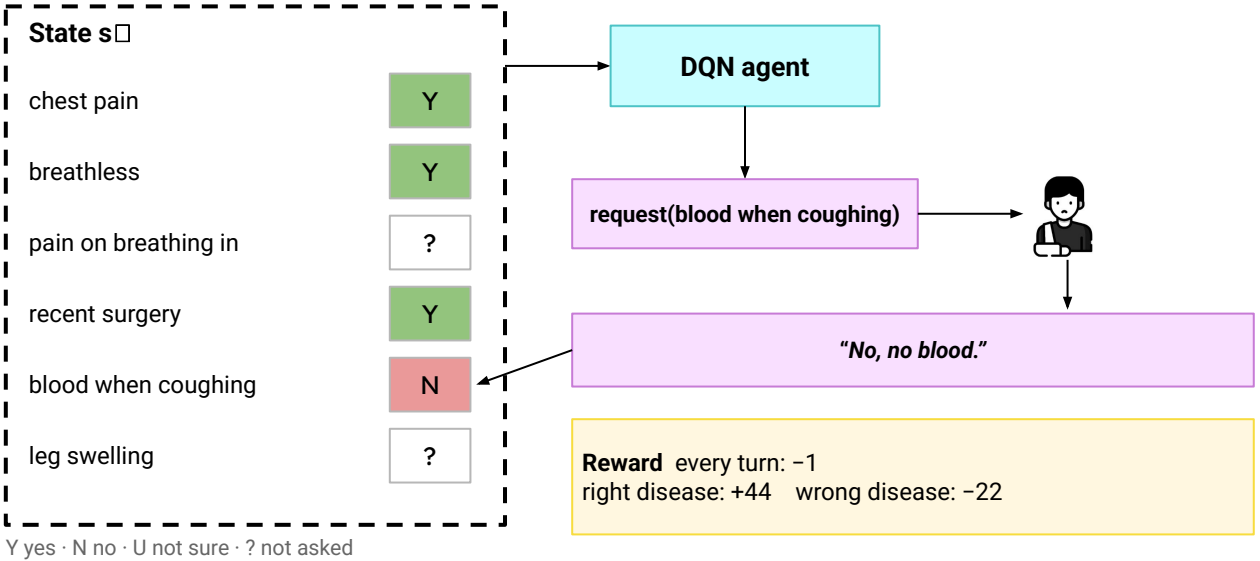


Task-oriented Dialogue System for Automatic Diagnosis (**Wei et al., 2018**)

As a Markov Decision Process

x is:

- Symptoms so far
- Turn

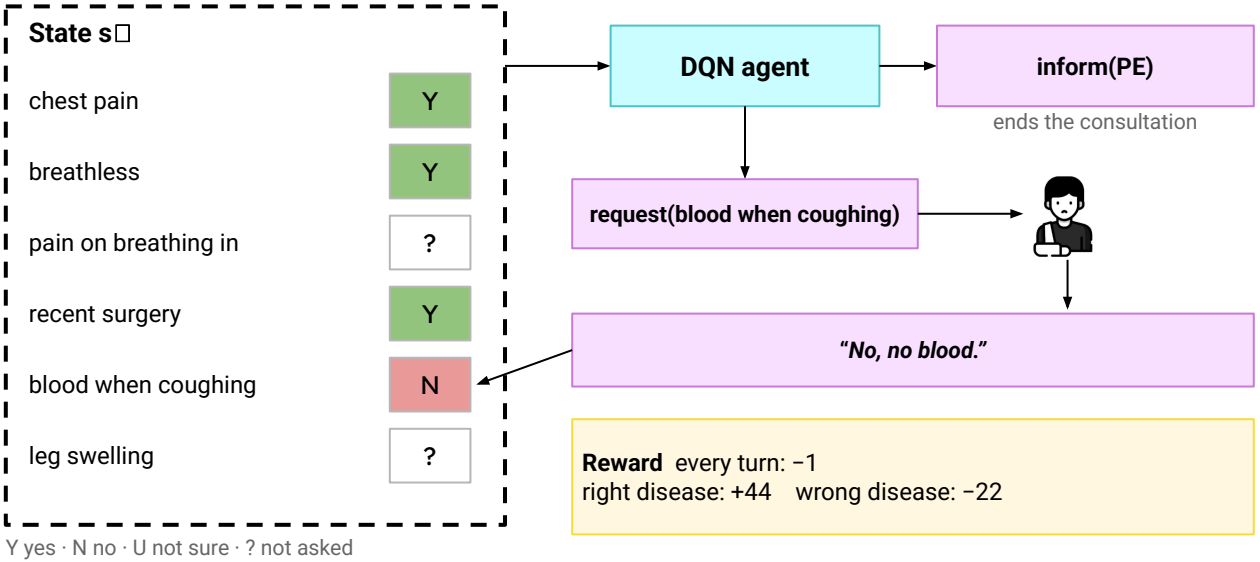


Task-oriented Dialogue System for Automatic Diagnosis (**Wei et al., 2018**)

As a Markov Decision Process

x is:

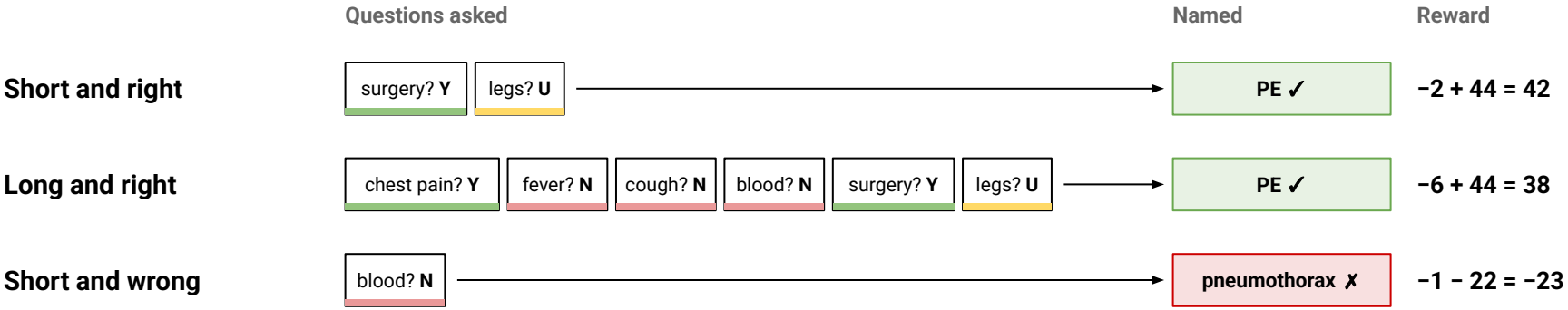
- Symptoms so far
- Turn



Task-oriented Dialogue System for Automatic Diagnosis (**Wei et al., 2018**)

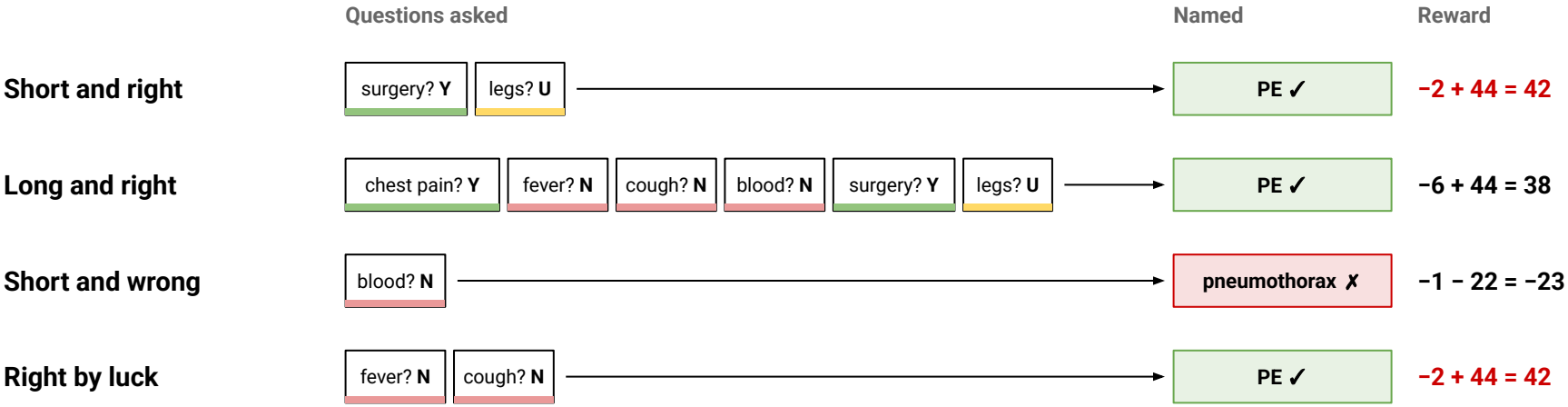
Automatic Diagnosis

As a Markov Decision Process



Task-oriented Dialogue System for Automatic Diagnosis (**Wei et al., 2018**)

As a Markov Decision Process



Same reward for both: it counts the turns and checks the final answer. It never asks whether the questions made sense.

What if a model learned which way of asking doctors prefer?

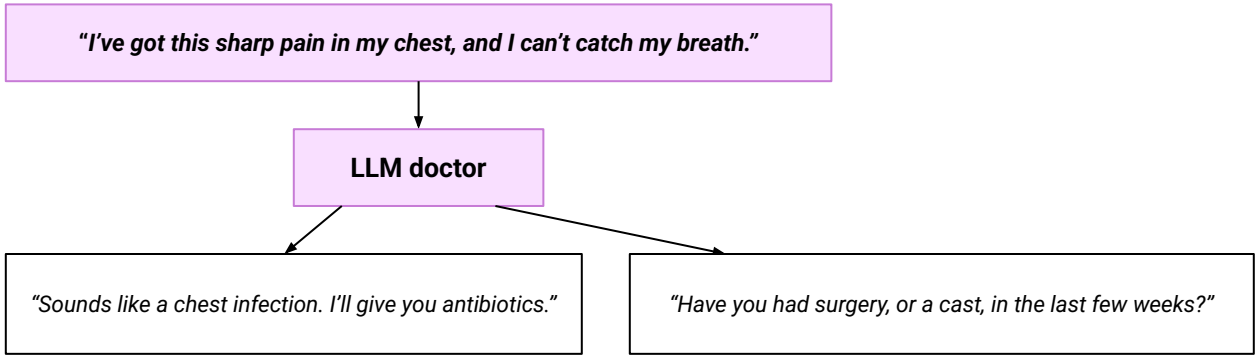
Task-oriented Dialogue System for Automatic Diagnosis (**Wei et al., 2018**)

Automatic Diagnosis

As Preference-Aligned Dialogue

x is:

- Dialogue history



Integrating Physician Diagnostic Logic into Large Language Models: Preference Learning from Process Feedback (**Dou et al., 2024**)

Automatic Diagnosis

As Preference-Aligned Dialogue

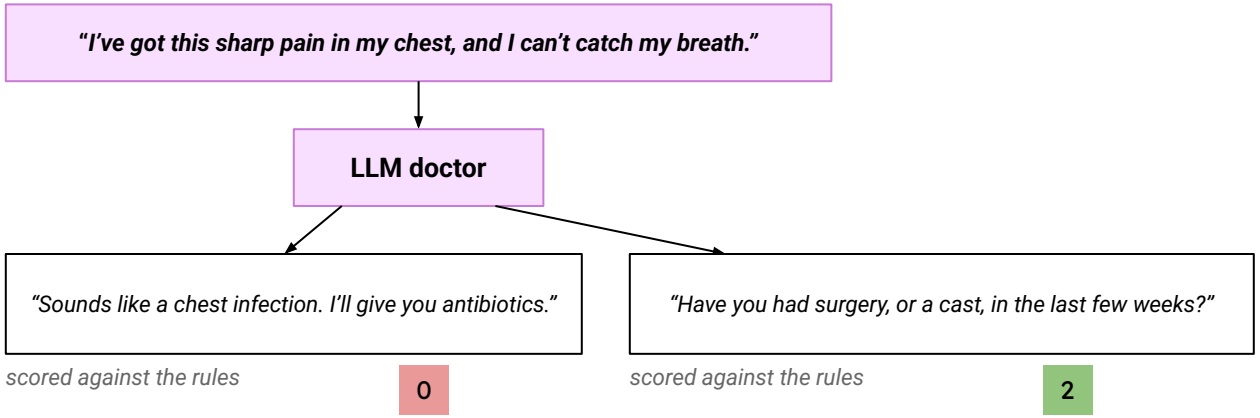
x is:

- Dialogue history

Doctors' rules

Goals, in order (A–C): check the condition, then a diagnosis or tests, then treatment

Constraints (D–F): explain, answer his questions, ask actively



Integrating Physician Diagnostic Logic into Large Language Models: Preference Learning from Process Feedback (**Dou et al., 2024**)

Automatic Diagnosis

As Preference-Aligned Dialogue

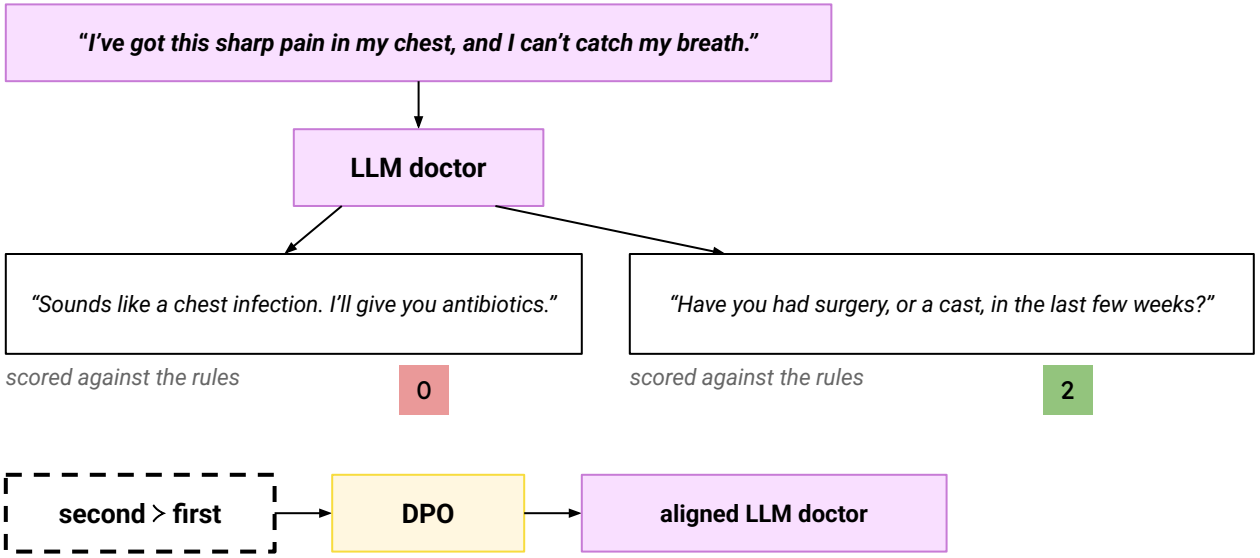
x is:

- Dialogue history

Doctors' rules

Goals, in order (A–C): check the condition, then a diagnosis or tests, then treatment

Constraints (D–F): explain, answer his questions, ask actively



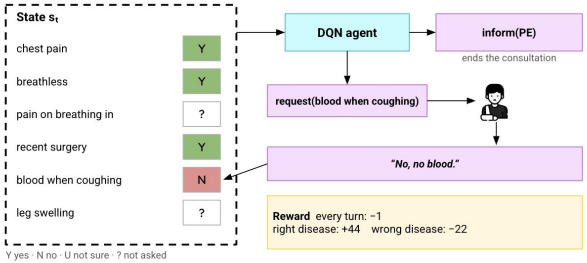
Integrating Physician Diagnostic Logic into Large Language Models: Preference Learning from Process Feedback (**Dou et al., 2024**)

Automatic Diagnosis

As a Markov Decision Process

x is:

- Symptoms so far
- Turn



Task-oriented Dialogue System for Automatic Diagnosis (**Wei et al., 2018**)

Automatic Diagnosis

As Preference-Aligned Dialogue

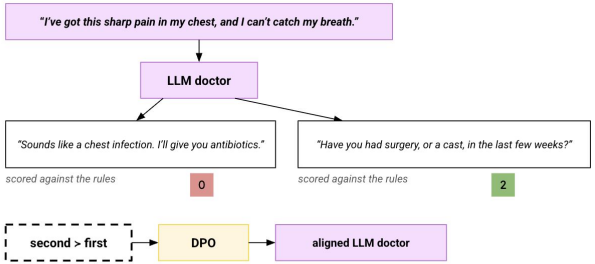
x is:

- Dialogue history

Doctors' rules

Goals, in order (A–C): check the condition, then a diagnosis or tests, then treatment

Constraints (D–F): explain, answer his questions, ask actively

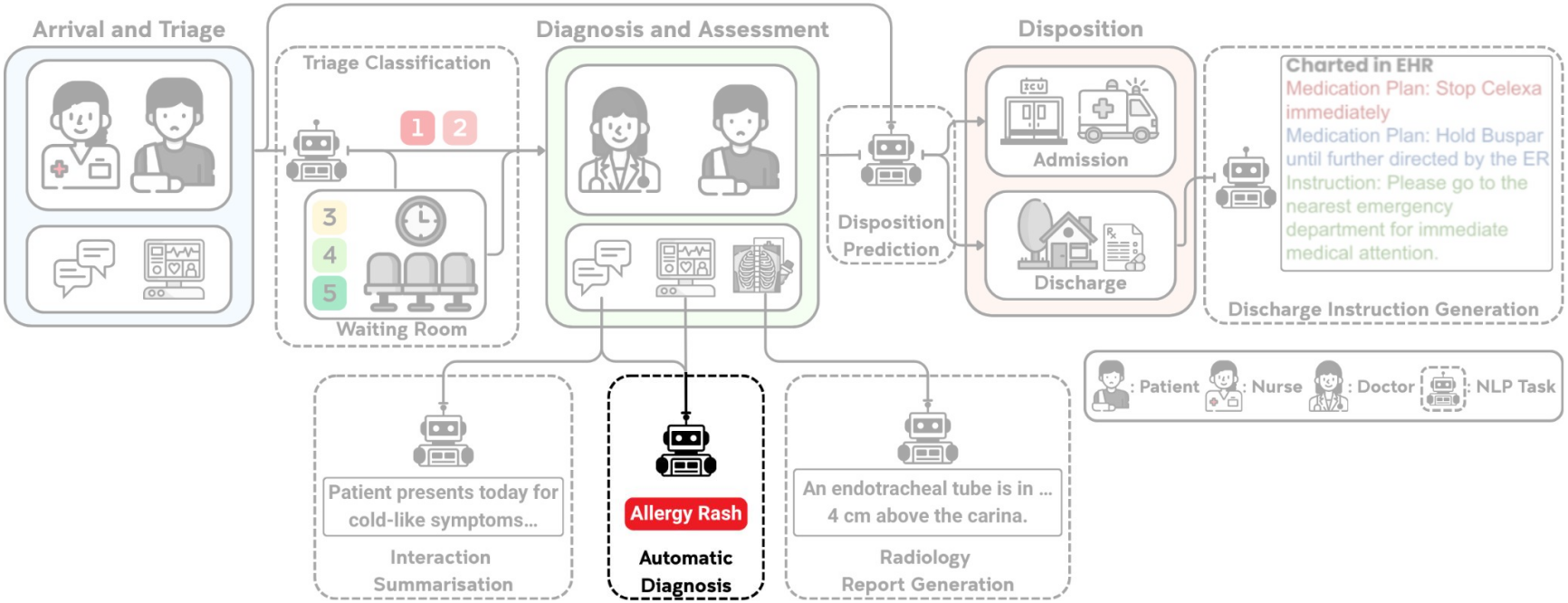


Integrating Physician Diagnostic Logic into Large Language Models: Preference Learning from Process Feedback (**Dou et al., 2024**)



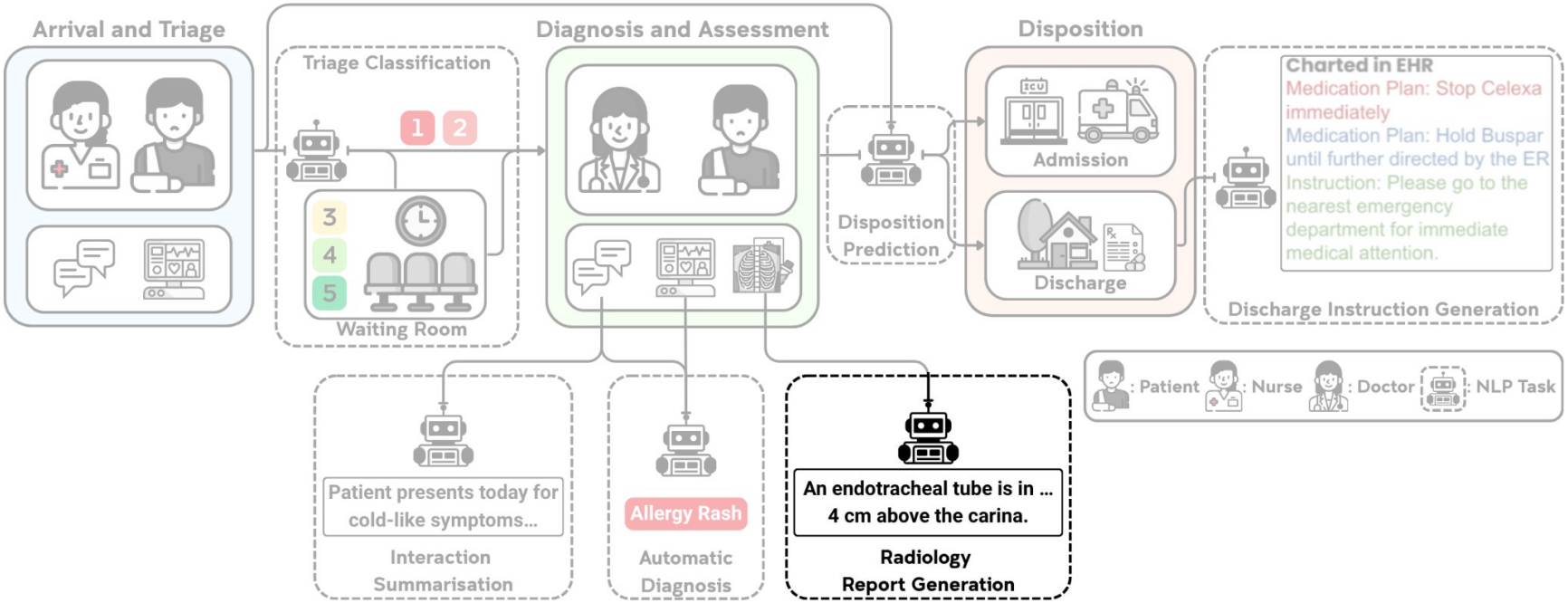
$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid x)$$

Landscape of ED NLP



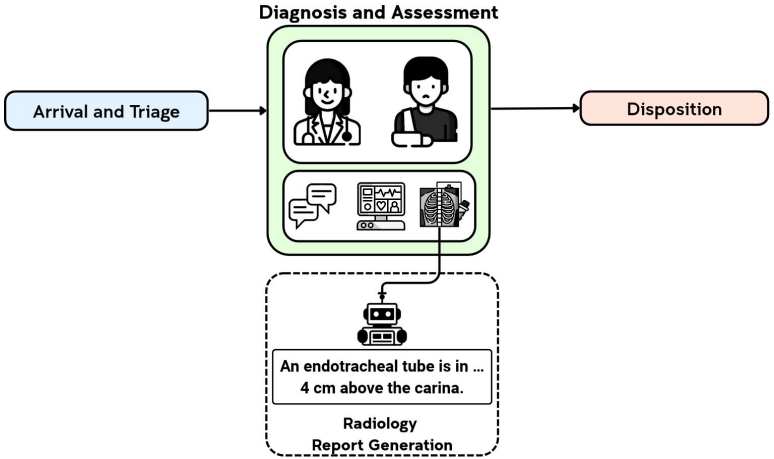
$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid x)$$

Landscape of ED NLP



$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid x)$$

Radiology Report Generation



$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid x)$$

Output space

$$y = (r_{\text{find}},\ r_{\text{impr}}), \quad x = G$$

Looking at Radiology Report Generation through a Causal Lens: A Survey (**Kumar et al., 2026**)

Radiology Report Generation



Male, 24, walked in, right hand in a cast.

"I've got this sharp pain in my chest, and I can't catch my breath."

HR	RR	O_2 Sat.	BP	Temp.
118	24	93	128/78	98.2

+ his chest X-ray, G



illustrative

F Findings: lungs clear, no pneumothorax, no effusion

I Impression: no acute abnormality

One word flips a finding

- "No pneumothorax" and "small pneumothorax" lead to different treatment.

Fluent but wrong

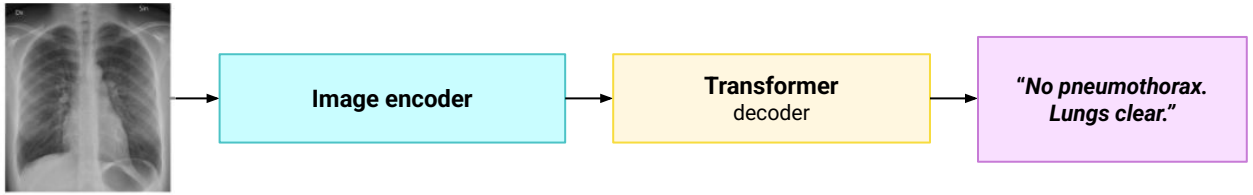
- The report reads well but invents or misses a finding.

Radiology Report Generation

As Image Captioning

x is:

- Chest X-ray



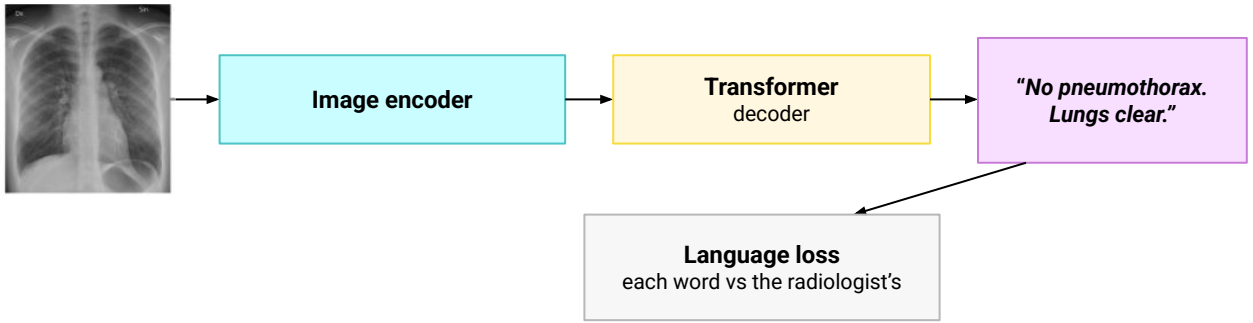
Learning to Generate Clinically Coherent Chest X-Ray Reports (**Lovelace & Mortazavi, 2020**)

Radiology Report Generation

As Image Captioning

x is:

- Chest X-ray



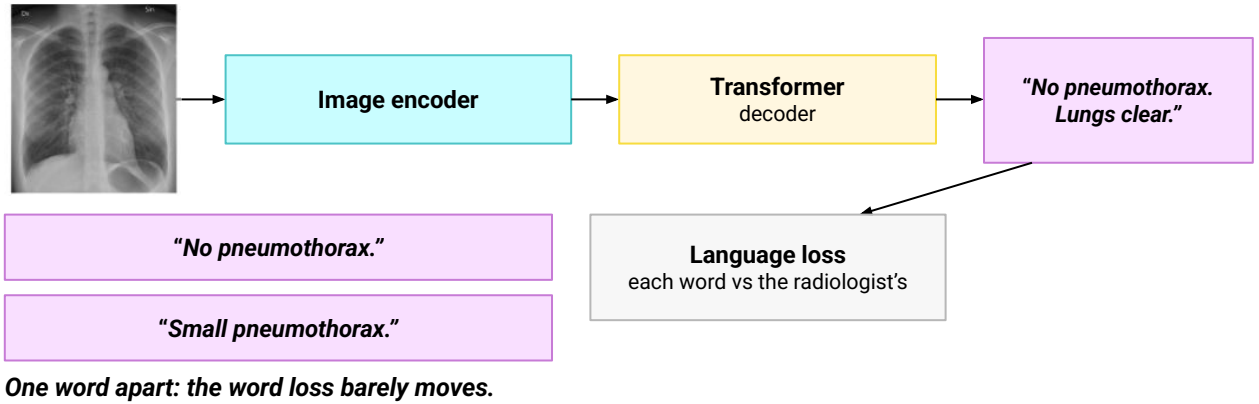
Learning to Generate Clinically Coherent Chest X-Ray Reports (**Lovelace & Mortazavi, 2020**)

Radiology Report Generation

As Image Captioning

x is:

- Chest X-ray



One word apart: the word loss barely moves.

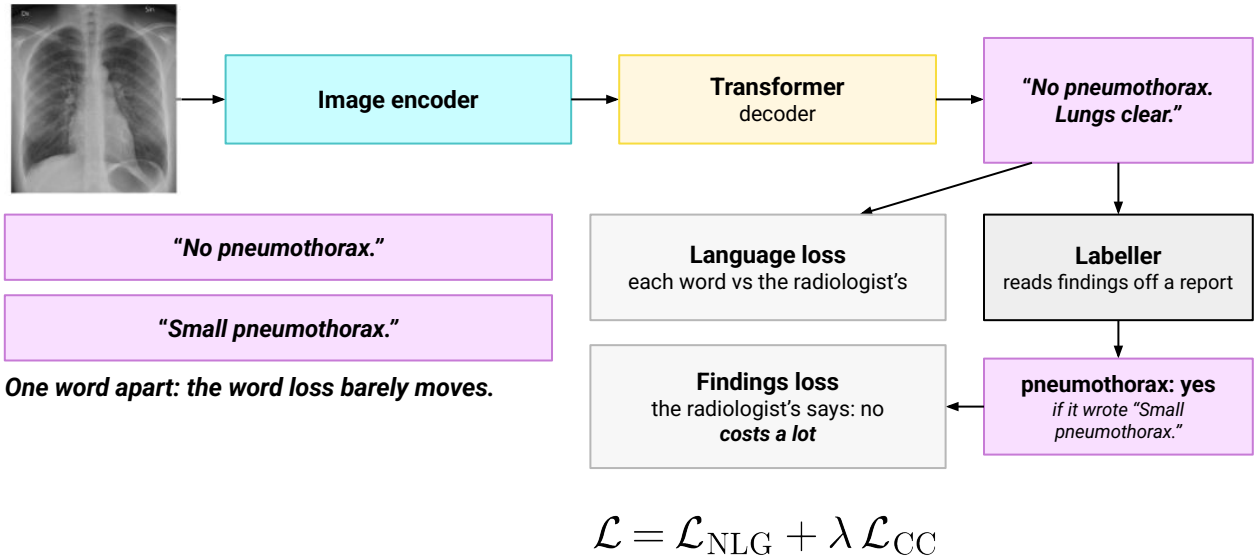
Learning to Generate Clinically Coherent Chest X-Ray Reports (**Lovelace & Mortazavi, 2020**)

Radiology Report Generation

As Image Captioning

x is:

- Chest X-ray



$$\mathcal{L} = \mathcal{L}_{\text{NLG}} + \lambda \mathcal{L}_{\text{CC}}$$

Learning to Generate Clinically Coherent Chest X-Ray Reports (**Lovelace & Mortazavi, 2020**)

As Retrieval-Augmented Generation

x is:

- Chest X-ray
- **Retrieved Reports**



Retriever

X-rays most like theirs

0.92 *"No acute findings. Lungs clear."*

0.90 *"Clear lungs, no effusion."*

0.88 *"Small right pneumothorax."*

0.61 *"Right lower lobe pneumonia."*

MMed-RAG: Versatile Multimodal RAG System for Medical Vision Language Models (**Xia et al., 2025**)

Radiology Report Generation

As Retrieval-Augmented Generation

x is:

- Chest X-ray
- **Retrieved Reports**



Retriever

X-rays most like theirs

0.92 *"No acute findings. Lungs clear."*

0.90 *"Clear lungs, no effusion."*

0.88 *"Small right pneumothorax."*

0.61 ~~*"Right lower lobe pneumonia."*~~

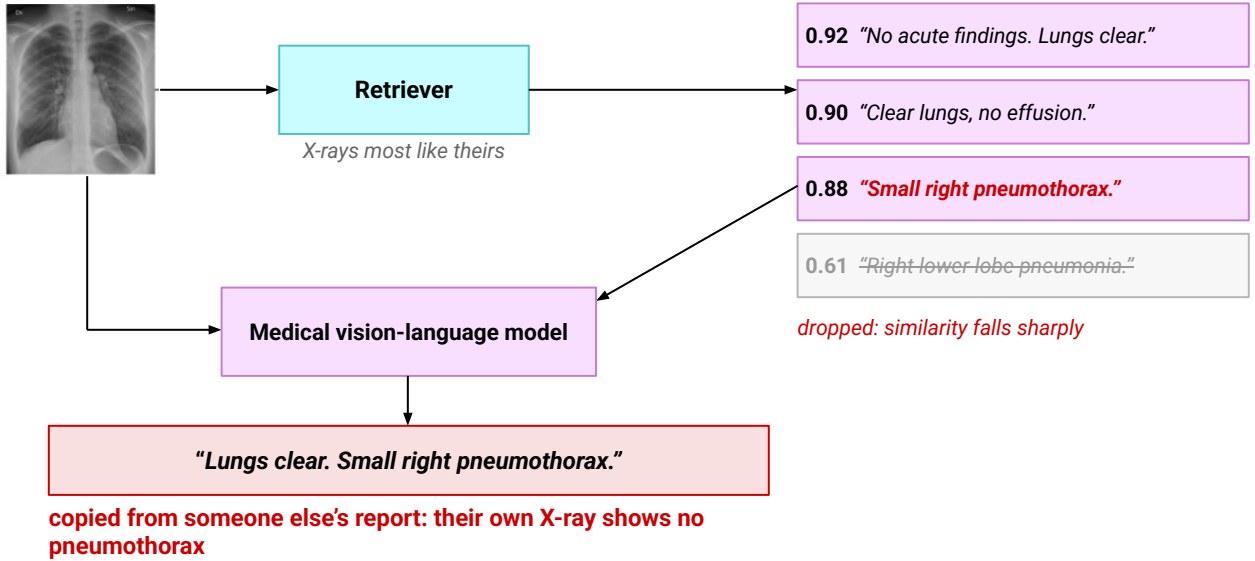
dropped: similarity falls sharply

MMed-RAG: Versatile Multimodal RAG System for Medical Vision Language Models (**Xia et al., 2025**)

As Retrieval-Augmented Generation

x is:

- Chest X-ray
- **Retrieved Reports**

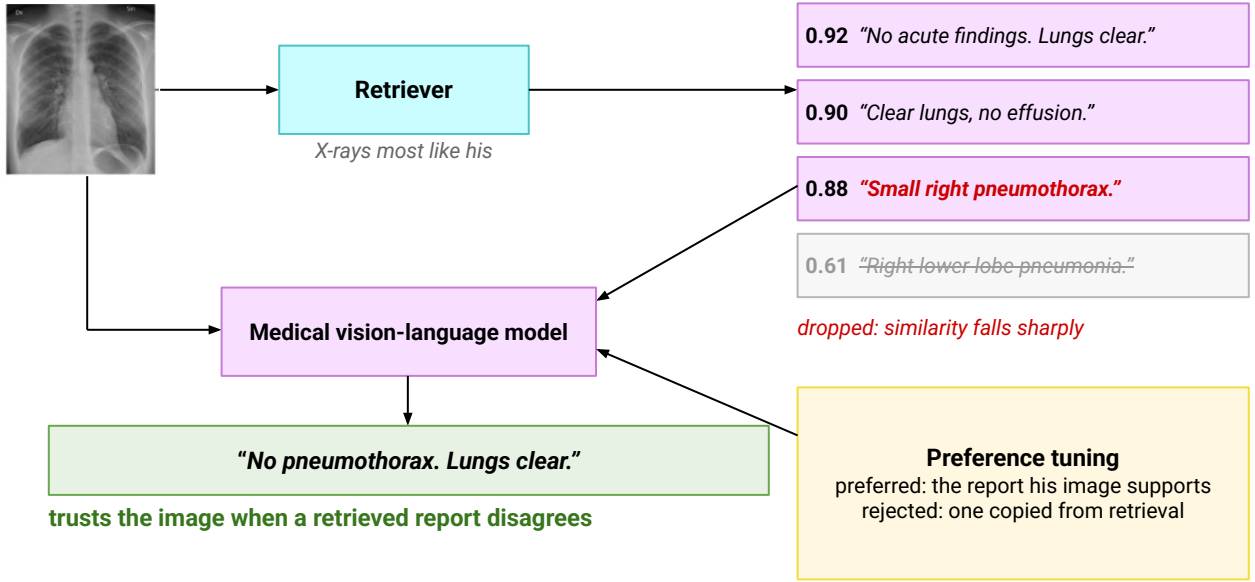


MMed-RAG: Versatile Multimodal RAG System for Medical Vision Language Models (**Xia et al., 2025**)

As Retrieval-Augmented Generation

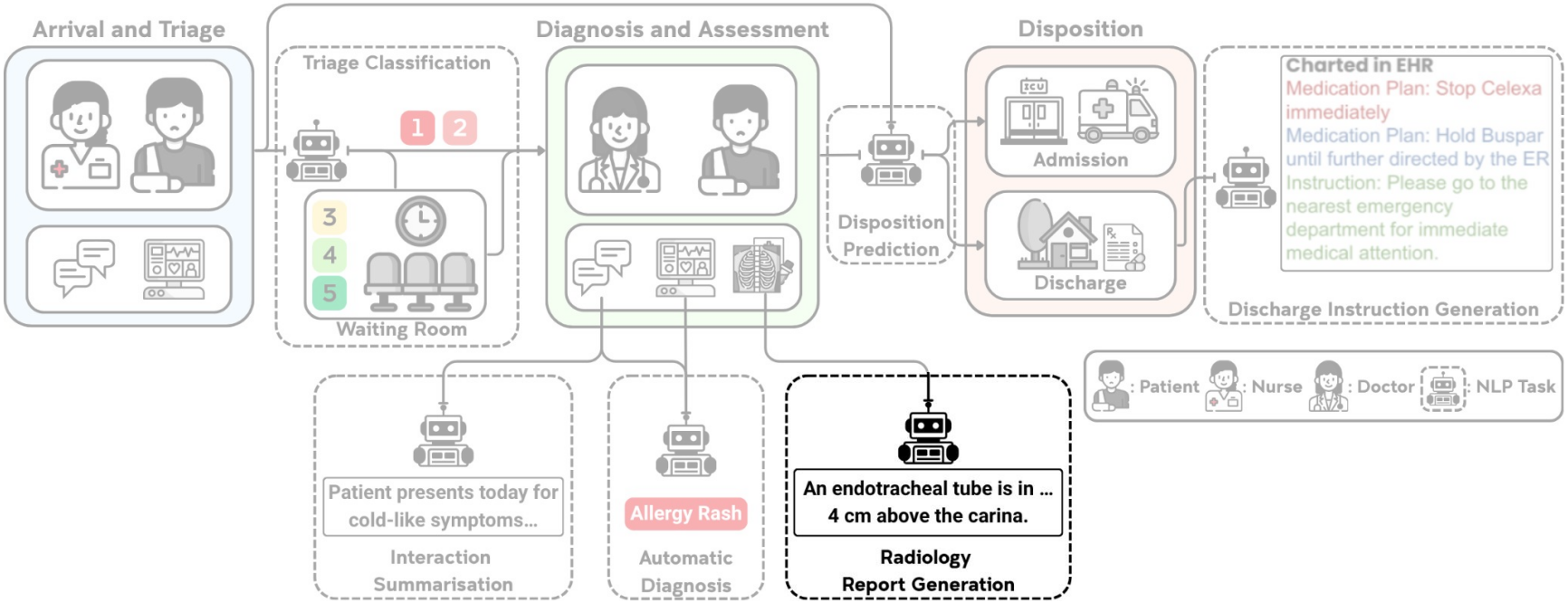
x is:

- Chest X-ray
- **Retrieved Reports**



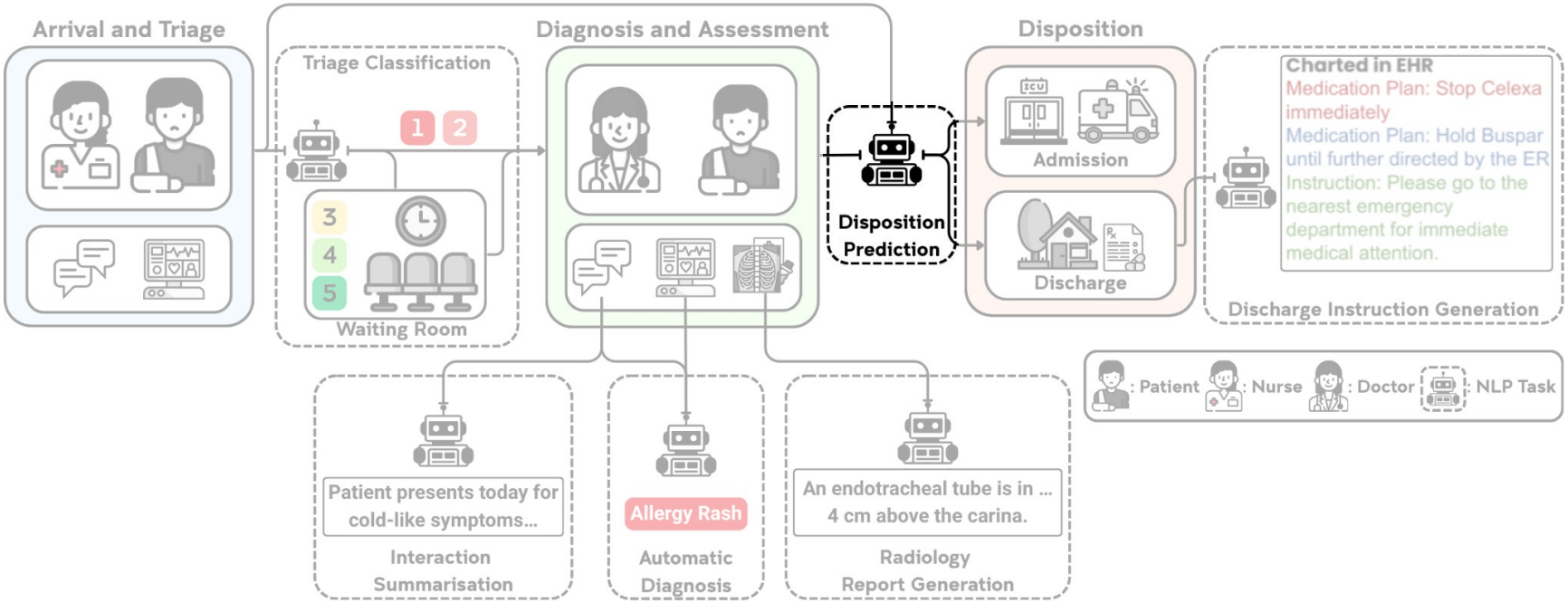
MMed-RAG: Versatile Multimodal RAG System for Medical Vision Language Models (**Xia et al., 2025**)

Landscape of ED NLP



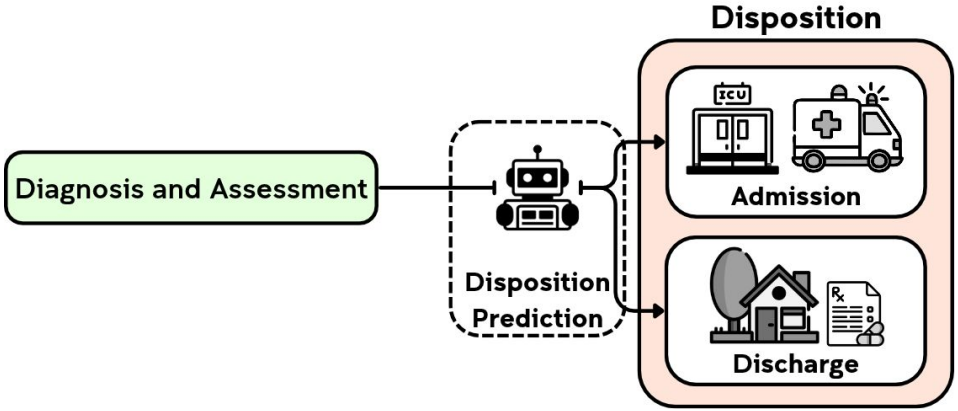
$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid x)$$

Landscape of ED NLP



$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid x)$$

Disposition Prediction



$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid x)$$

Output space

$$\mathcal{Y} = \{\text{home},\ \text{ward},\ \text{ICU}\}$$

Disposition Prediction



Male, 24, walked in, right hand in a cast.

"I've got this sharp pain in my chest, and I can't catch my breath."

HR	RR	O_2 Sat.	BP	Temp.
118	24	93	128/78	98.2

+ lab results, as they come back:
CBC, CHEM, … (12 lab groups)

Home — discharged, with instructions

Ward — admitted to a hospital ward

ICU — intensive care

Under-call

- An ICU-bound patient goes to a ward, or home.

Waiting for every test

- Hours on a trolley while the department fills.

Disposition Prediction

As Sequence Classification

x is:

- Chief Complaint
- Vitals
- Patient Metadata

"chief complaint: chest pain, short of breath | HR 118 | RR 24 | SpO_2 93 | BP 128/78 | age 24 | male"

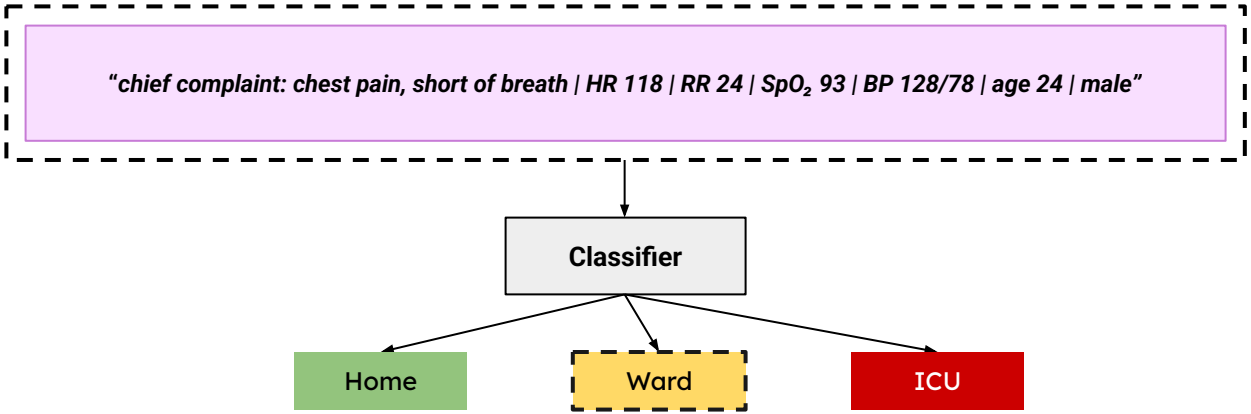
Predicting intensive care unit admission among patients presenting to the emergency department using machine learning and natural language processing (**Fernandes et al., 2020**)

Disposition Prediction

As Sequence Classification

x is:

- Chief Complaint
- Vitals
- Patient Metadata



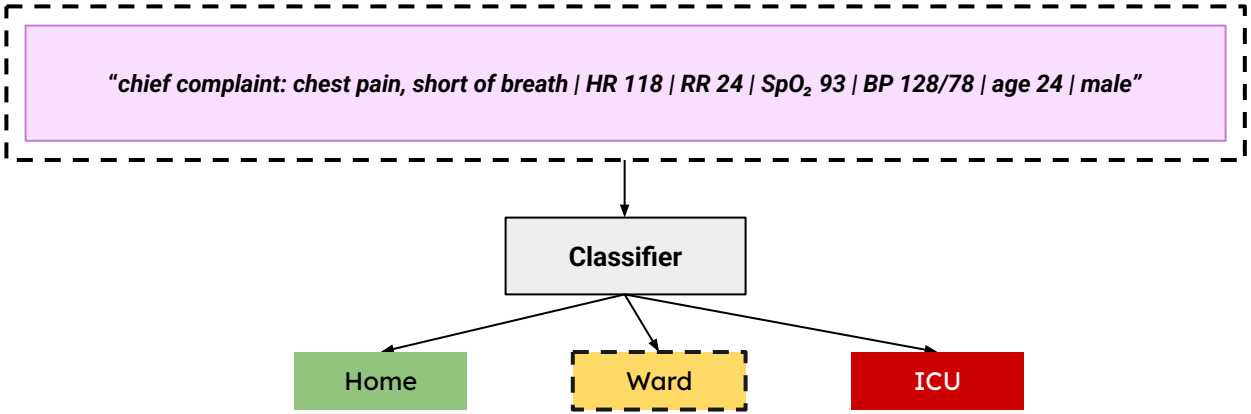
Predicting intensive care unit admission among patients presenting to the emergency department using machine learning and natural language processing (**Fernandes et al., 2020**)

Disposition Prediction

As Sequence Classification

x is:

- Chief Complaint
- Vitals
- Patient Metadata
- **Lab Results**



Every lab result in the input costs time. Which ones are worth waiting for?

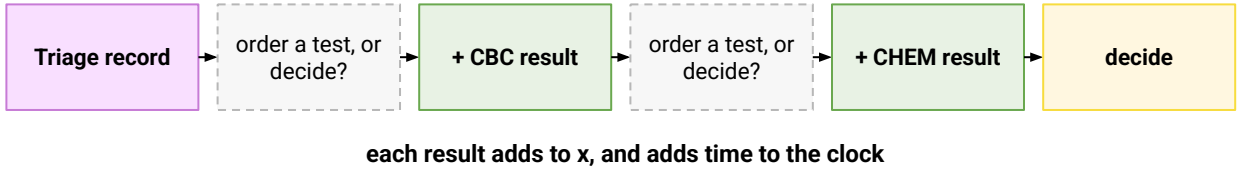
Predicting intensive care unit admission among patients presenting to the emergency department using machine learning and natural language processing (**Fernandes et al., 2020**)

Disposition Prediction

As Sequential Test Selection

x is:

- Chief Complaint
- Vitals
- Patient Metadata
- **Lab Results**



A test is an action, like a question in Diagnosis.

So learn when the next test is worth the wait, and when to stop and decide.

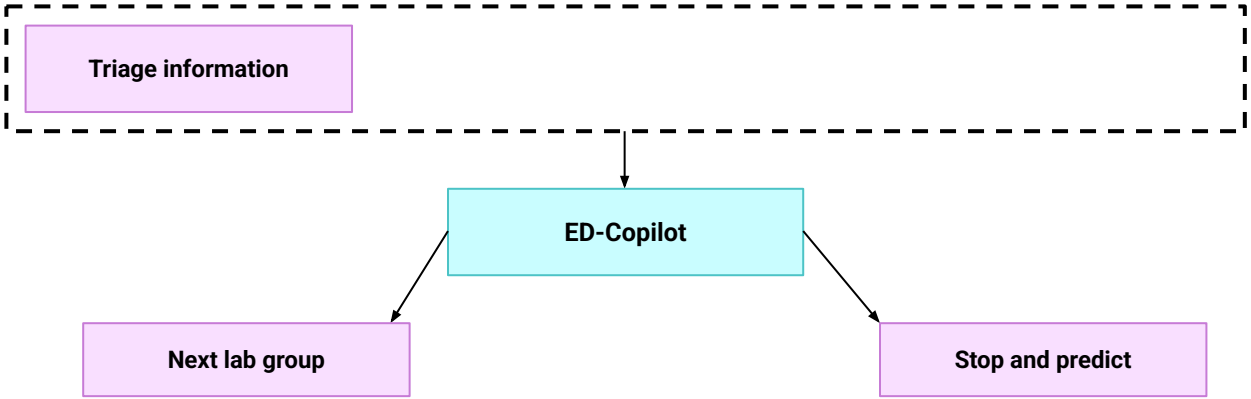
ED-Copilot: Reduce Emergency Department Wait Time with Language Model Diagnostic Assistance (**Sun et al., 2024**)

Disposition Prediction

As Sequential Test Selection

x is:

- Chief Complaint
- Vitals
- Patient Metadata
- **Lab Results**



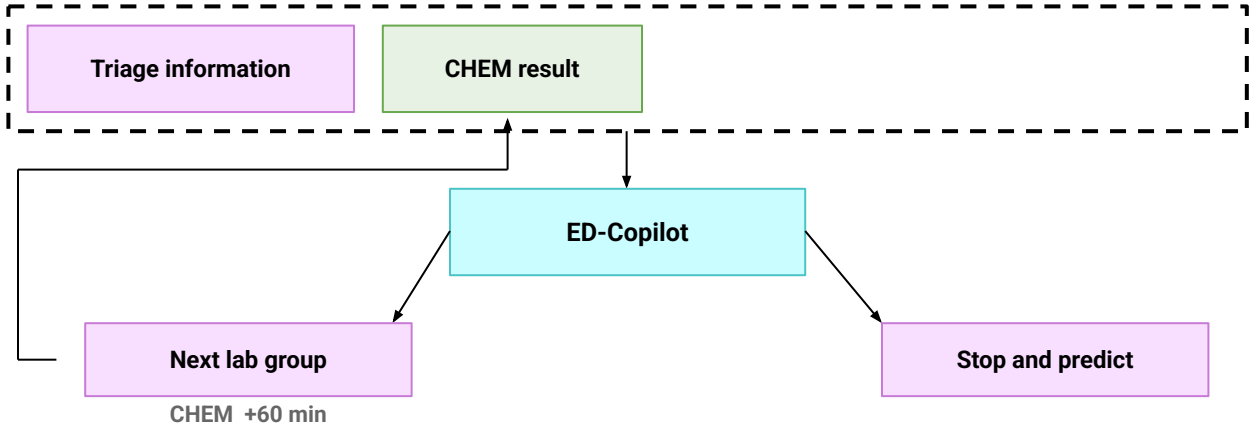
ED-Copilot: Reduce Emergency Department Wait Time with Language Model Diagnostic Assistance (**Sun et al., 2024**)

Disposition Prediction

As Sequential Test Selection

x is:

- Chief Complaint
- Vitals
- Patient Metadata
- **Lab Results**



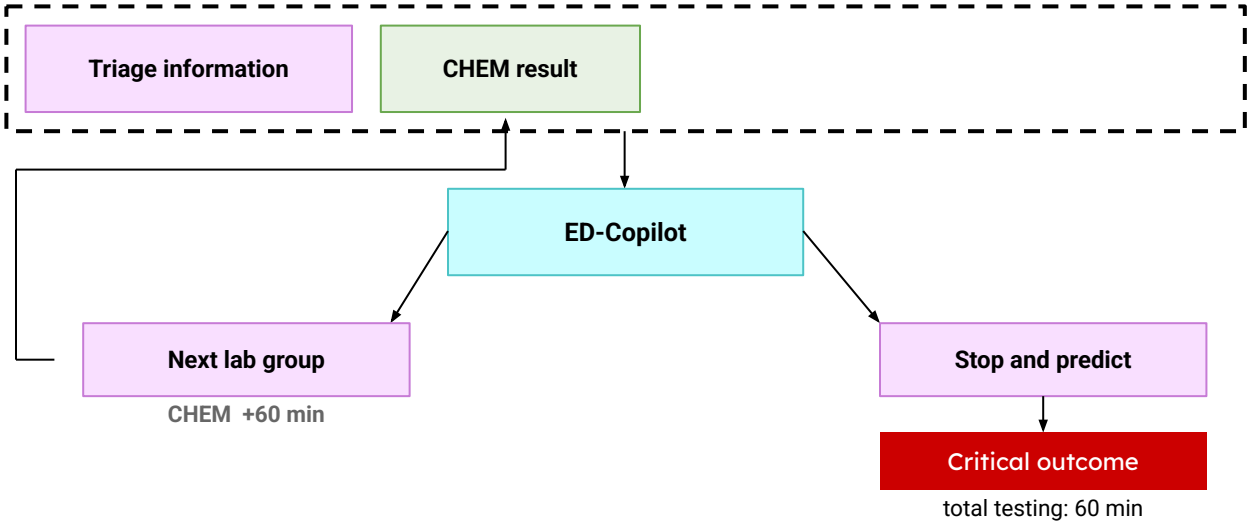
ED-Copilot: Reduce Emergency Department Wait Time with Language Model Diagnostic Assistance (**Sun et al., 2024**)

Disposition Prediction

As Sequential Test Selection

x is:

- Chief Complaint
- Vitals
- Patient Metadata
- **Lab Results**



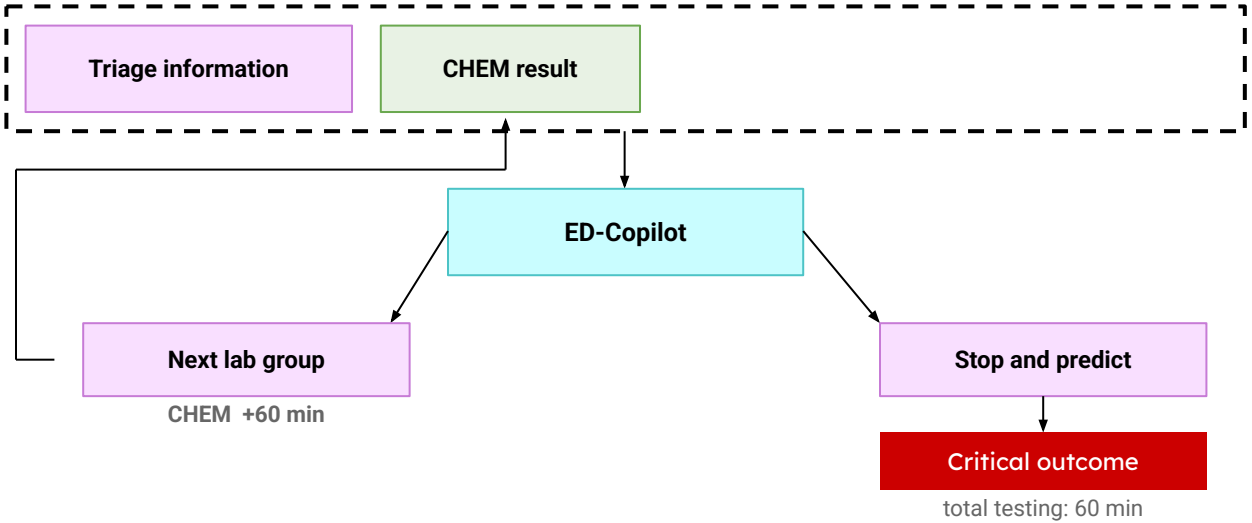
ED-Copilot: Reduce Emergency Department Wait Time with Language Model Diagnostic Assistance (**Sun et al., 2024**)

Disposition Prediction

As Sequential Test Selection

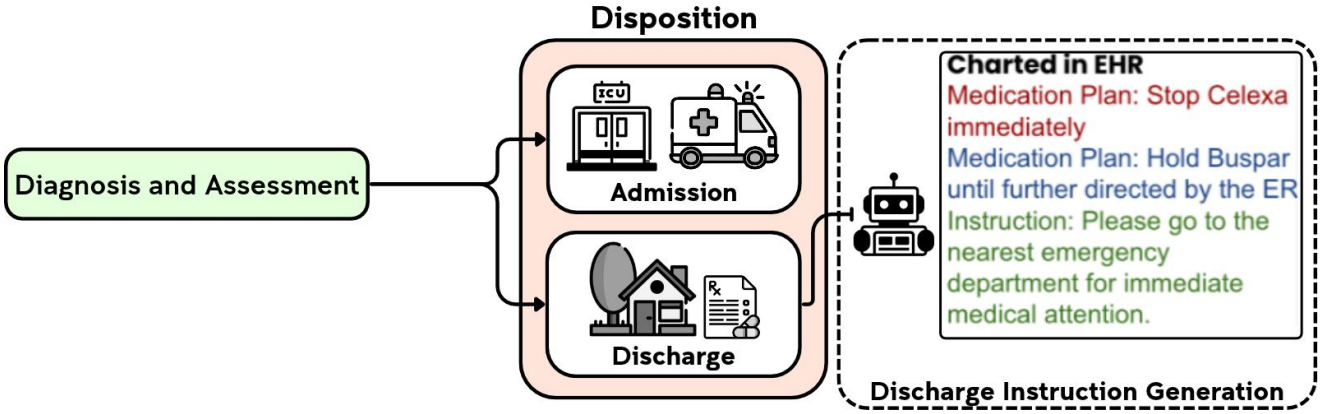
x is:

- Chief Complaint
- Vitals
- Patient Metadata
- **Lab Results**



ED-Copilot: Reduce Emergency Department Wait Time with Language Model Diagnostic Assistance (**Sun et al., 2024**)

Discharge Instruction Generation



$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid x)$$

Output space

$$y = \iota = (\iota_1, \ldots, \iota_n), \quad \iota_j \in \mathcal{V}$$

Discharge Instruction Generation



Male, 24, walked in, right hand in a cast.

"I've got this sharp pain in my chest, and I can't catch my breath."

HR	RR	O_2 Sat.	BP	Temp.
118	24	93	128/78	98.2

+ what the doctor tells him before he leaves:
"For the pain, just take ibuprofen, 400, every six hours or so. Keep that cast dry. Come straight back if you cough blood."

1. Take ibuprofen 400 mg every 6 hours.
2. Keep the cast dry.
3. Fracture clinic on Monday.
4. Come back if you cough blood.

Drop a name or a dose

- He takes the wrong medicine, or the wrong amount, at home.

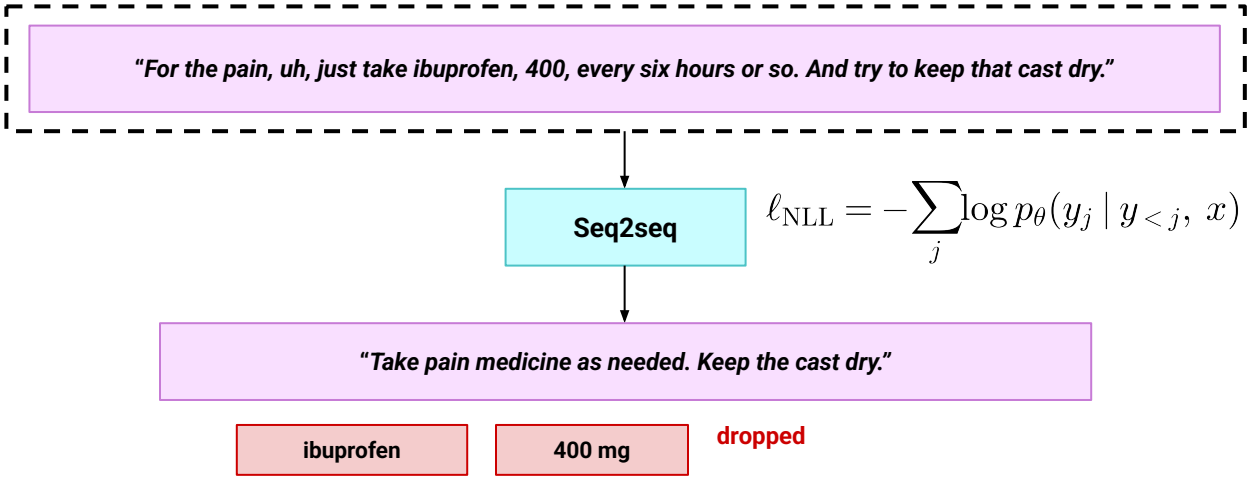
Vague return advice

- He does not come back when he should.

As Sequence-to-Sequence Generation

x is:

- Dialogue



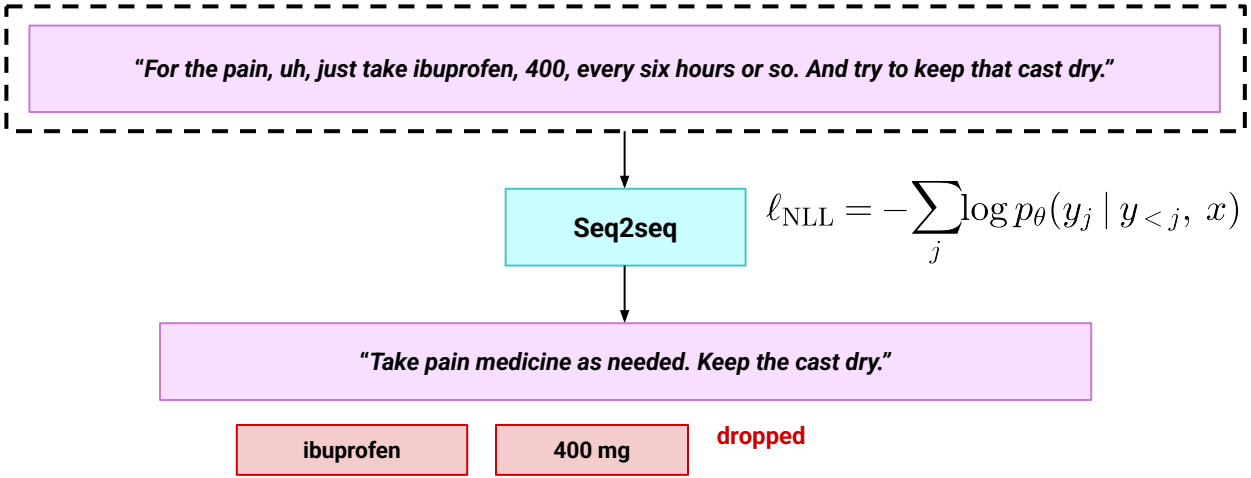
Injecting knowledge into language generation: a case study in auto-charting after-visit care instructions from medical dialogue (**Eremeev et al., 2023**)

Discharge Instruction Generation

As Sequence-to-Sequence Generation

x is:

- Dialogue



Rare words are cheap to drop under this loss. What if the loss knew which words must carry over?

Injecting knowledge into language generation: a case study in auto-charting after-visit care instructions from medical dialogue (**Eremeev et al., 2023**)

Discharge Instruction Generation

As Knowledge-Weighted Generation

x is:

- Dialogue
- **Domain Knowledge**

"For the pain, uh, just take ibuprofen, 400, every six hours or so. And try to keep that cast dry."

ibuprofen: drug | **pain: symptom** | **cast: device** — **UMLS or SNOMED CT lookup**

Injecting knowledge into language generation: a case study in auto-charting after-visit care instructions from medical dialogue (**Eremeev et al., 2023**)

Discharge Instruction Generation

As Knowledge-Weighted Generation

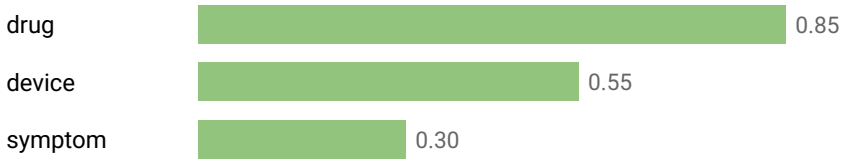
x is:

- Dialogue
- **Domain Knowledge**

"For the pain, uh, just take ibuprofen, 400, every six hours or so. And try to keep that cast dry."

ibuprofen: drug | **pain: symptom** | **cast: device** — **UMLS or SNOMED CT lookup**

Utilization rate of a type, from the training data:



Injecting knowledge into language generation: a case study in auto-charting after-visit care instructions from medical dialogue (**Eremeev et al., 2023**)

Discharge Instruction Generation

As Knowledge-Weighted Generation

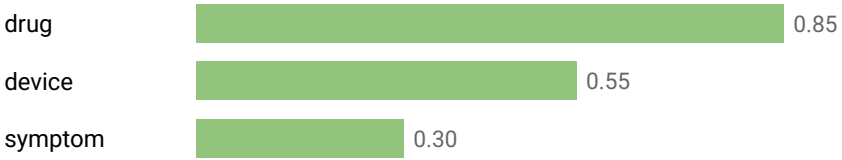
x is:

- Dialogue
- **Domain Knowledge**

"For the pain, uh, just take ibuprofen, 400, every six hours or so. And try to keep that cast dry."

ibuprofen: drug | **pain: symptom** | **cast: device** — **UMLS or SNOMED CT lookup**

Utilization rate of a type, from the training data:



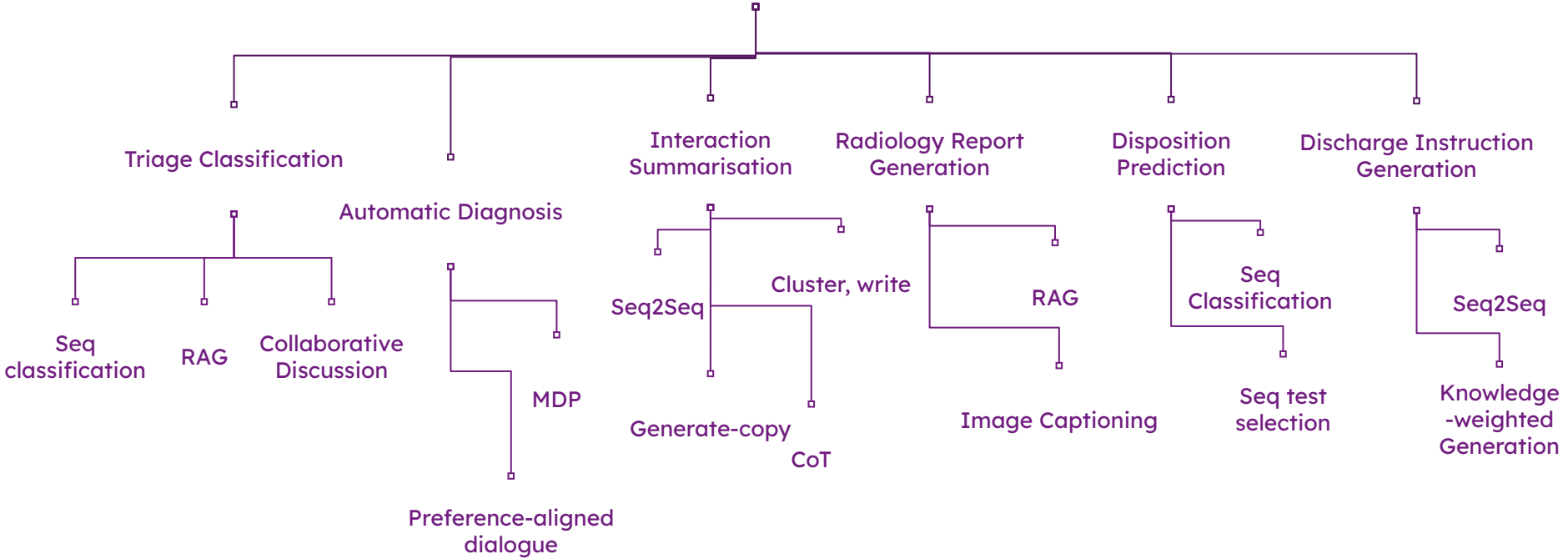
$$\ell = \ell_{\mathrm{NLL}}(y) + \alpha \cdot \ell_w(x)$$

costs more when a high-rate concept is missing from the output

"Take ibuprofen 400 mg every 6 hours as needed for pain. Keep the cast dry."

Injecting knowledge into language generation: a case study in auto-charting after-visit care instructions from medical dialogue (**Eremeev et al., 2023**)

NLP Tasks in ED



Tutorial Plan

Module 1: NLP for Emergency Department

Module 2: Typical Tasks and Methods



Module 3: Frontiers & Readiness

Module 4: CSE Researchers in Clinical AI

AJCAI 2026 - Post Survey



Break

See you in 10 mins

Tutorial Plan

Module 1: NLP for Emergency Department

Module 2: Typical Tasks and Methods

Module 3: Frontiers & Readiness



Module 4: CSE Researchers in Clinical AI

Activity 1

Rate it

Rate it

1. **Read the card.** Where did it run, and what did it show?
2. **Climb from IRL1.** Stop at the first rung it has not earned: that is its level.
3. **Read its claim.** What would it do alone? That rung is its claim.
4. **Name the blocker.** Which frontier gates the next rung?

Card _ is at IRL _, claims IRL _, and is blocked by _ because _.

Frontiers

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid x)$$

Generalisability	**params** was learned at one hospital, on one scale
Robustness	**input** was typed, clean and complete
Temporal Validity	**input** was the finished record
Evaluation	**output** was scored right or wrong
Deployment	**output** is where the paper stops

From Triage to Discharge: A Survey of NLP Tasks, Methods, and Open Challenges in the Emergency Department (**Srirag et al., 2026**)

Generalisability *in practice*

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid x)$$

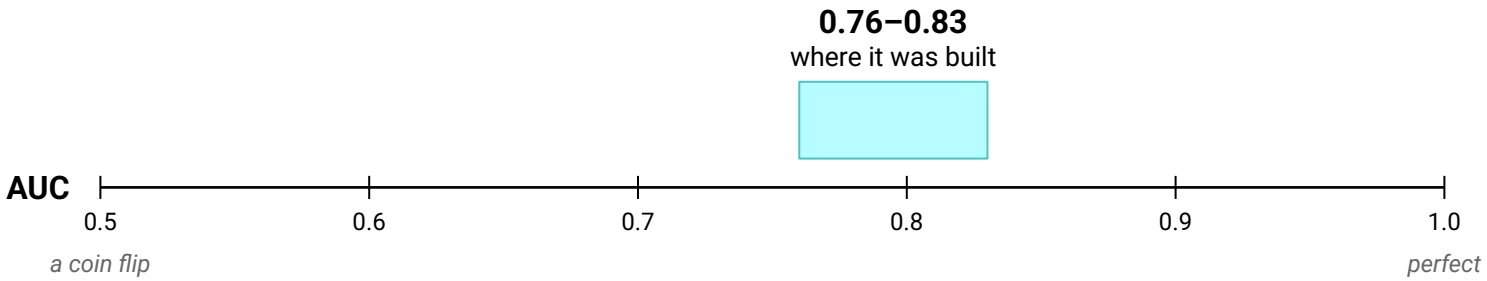
one triage scale

learned at one hospital

Would it work at another hospital, on another scale, for other patients?

A New Hospital

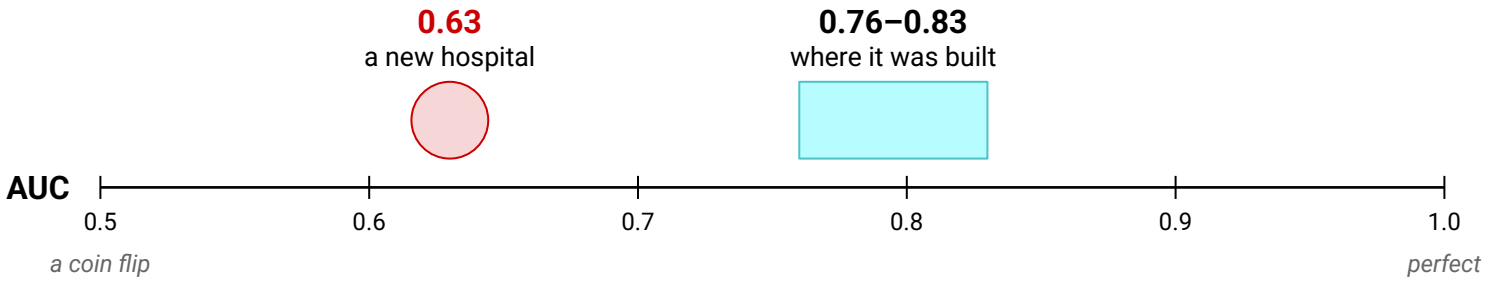
A sepsis model, sold to hundreds of hospitals. Its score: AUC.



External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients (**Wong et al., 2021**)

A New Hospital

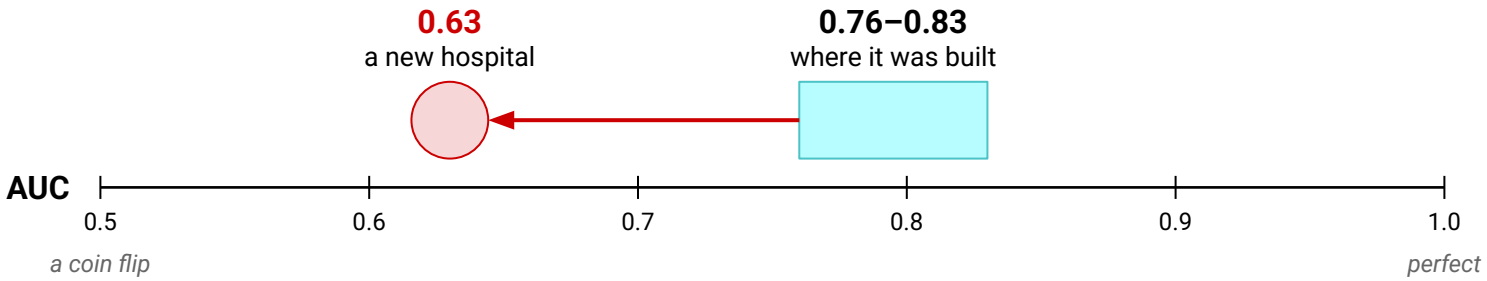
A sepsis model, sold to hundreds of hospitals. Its score: AUC.



External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients (**Wong et al., 2021**)

A New Hospital

A sepsis model, sold to hundreds of hospitals. Its score: AUC.



External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients (**Wong et al., 2021**)

Generalisability

Another Scale



Male, 24, walked in, right hand in a cast.

"I've got this sharp pain in my chest, and I can't catch my breath."

HR	RR	O_2 Sat.	BP	Temp.
118	24	93	128/78	98.2

Walks into an ED in Auckland.

ESI

USA: how many resources?

1. needs a life-saving intervention now
2. high risk; should not wait
3. needs two or more resources
4. needs one resource
5. needs no resources

From Triage to Discharge: A Survey of NLP Tasks, Methods, and Open Challenges in the Emergency Department (**Srirag et al., 2026**)

Generalisability

Another Scale



Male, 24, walked in, right hand in a cast.

"I've got this sharp pain in my chest, and I can't catch my breath."

HR	RR	O_2 Sat.	BP	Temp.
118	24	93	128/78	98.2

Walks into an ED in Auckland.

Same input, another output space.

A model trained on ESI labels has never seen the scale this nurse uses. Level 2 does not mean the same thing on both.

ESI

USA: how many resources?

1	needs a life-saving **intervention** now
2	high risk; should not wait
3	needs two or more **resources**
4	needs one **resource**
5	needs no **resources**

ATS

Australia, NZ: how long can they wait?

1	seen immediately
2	within 10 minutes
3	within 30 minutes
4	within 60 minutes
5	within 120 minutes

From Triage to Discharge: A Survey of NLP Tasks, Methods, and Open Challenges in the Emergency Department (**Srirag et al., 2026**)

Robustness to the Real World

$$\hat{y} = \arg\max_{y \in Y} p_\theta(y \mid x)$$

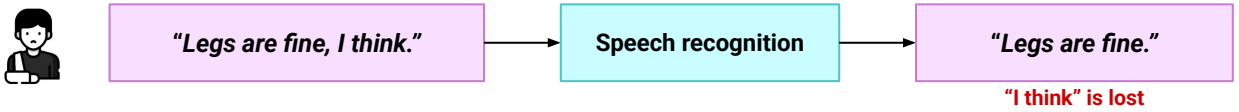
typed, clean and complete

Would it work on real input: speech, accents, interpreters, missing fields?

Spoken, Not Typed

x is:

- Dialogue
- **Speech**



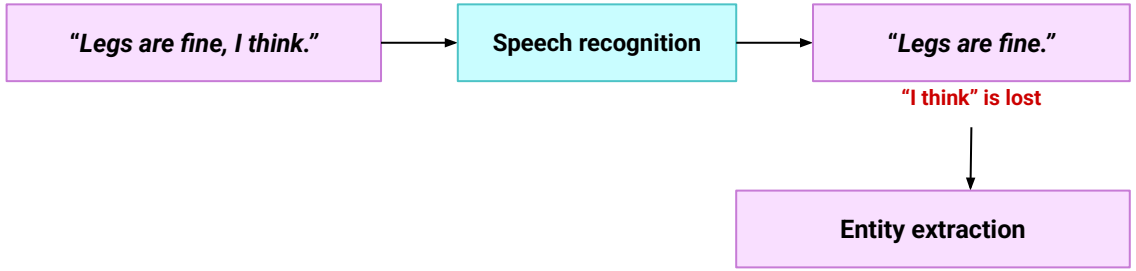
Generating medically-accurate summaries of patient-provider dialogue: A multi-stage approach using large language models (**Nair et al., 2023**)

Spoken, Not Typed

x is:

- Dialogue
- **Speech**





from the typed chat

leg swelling **unknown**

from the transcript

leg swelling **absent**

Generating medically-accurate summaries of patient-provider dialogue: A multi-stage approach using large language models (**Nair et al., 2023**)

What You Would Need

To test a model for this ED, you would need:

- **conversations, not finished records**
- **their voice and their accent, said aloud**

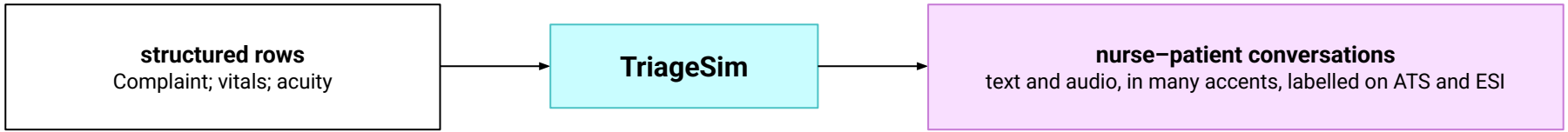
TriageSim: A Conversational Emergency Triage Simulation Framework from Structured Electronic Health Records (**Srirag et al., 2026**)

What You Would Need

To test a model for this ED, you would need:

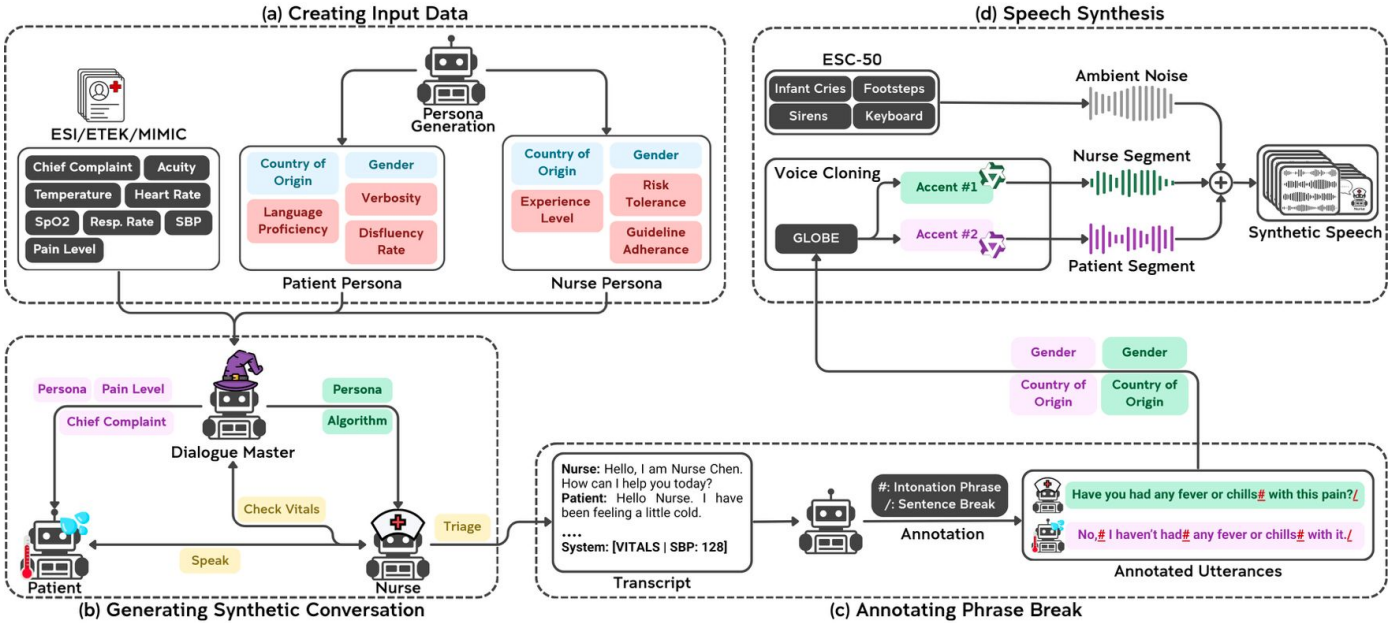
- **conversations, not finished records**
- **their voice and their accent, said aloud**

What the ED has:



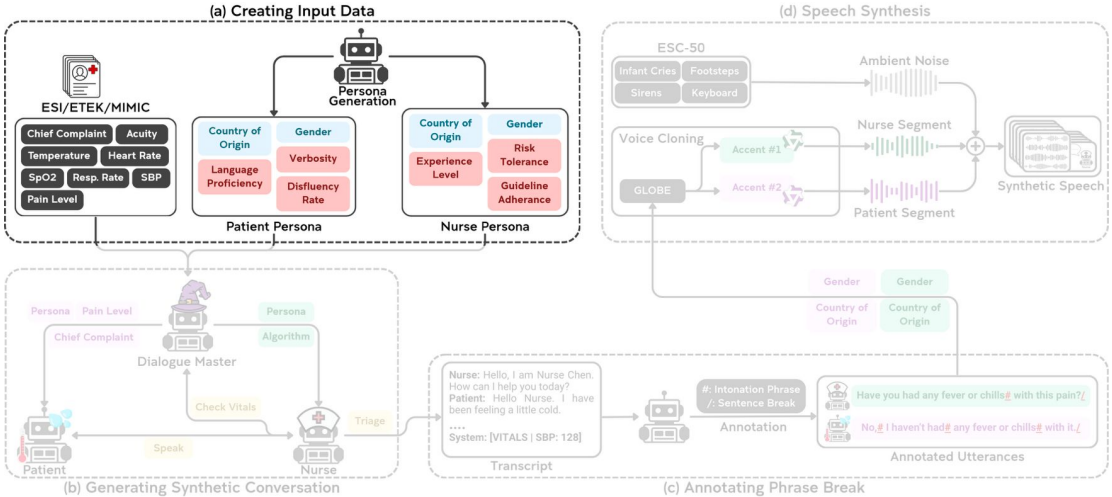
TriageSim: A Conversational Emergency Triage Simulation Framework from Structured Electronic Health Records (**Srirag et al., 2026**)

A Simulation Framework for Triage Conversations



TriageSim: A Conversational Emergency Triage Simulation Framework from Structured Electronic Health Records (**Srirag et al., 2026**)

Creating Input Data



MIMIC-IV-ED* [Johnson et al. 2023a]
MIMIC-IV-Note* [Johnson et al. 2023c]
Discharge Me* [Xu et al. 2024a]

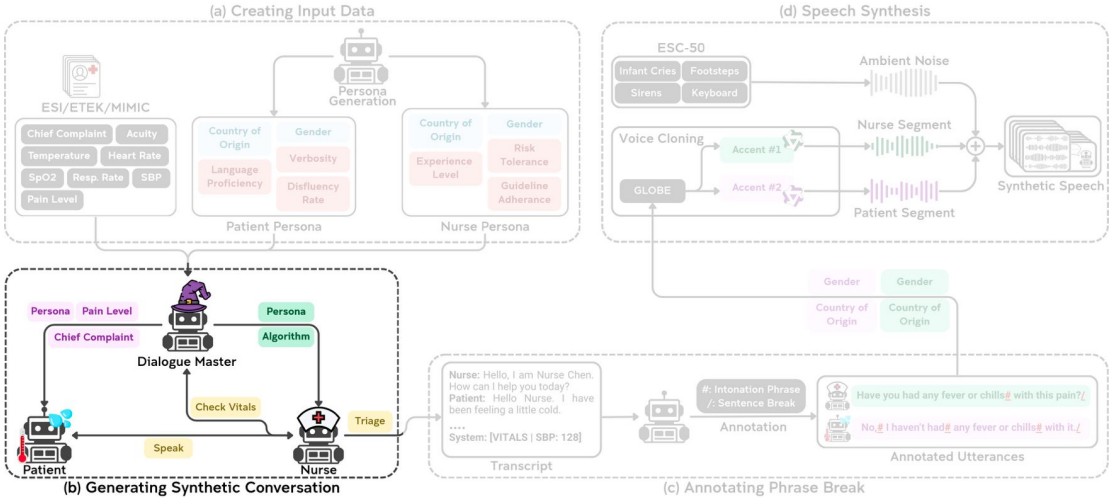
ESI Handbook [Gilboy et al. 2005]
ETEK Handbook [ETEK 2024]
MTSSamples [mts 2025]

+ Triage

- **Seed data** from ED records, the ESI handbook and the ETEK manual
- **Personas** generated with Gemini-3-Pro and GPT-5.2-Pro
- **Patient**: demographics, communication traits, cognitive traits
- **Nurse**: demographics, experience, risk tolerance, guideline adherence

TriageSim: A Conversational Emergency Triage Simulation Framework from Structured Electronic Health Records (**Srirag et al., 2026**)

Generating Synthetic Conversation

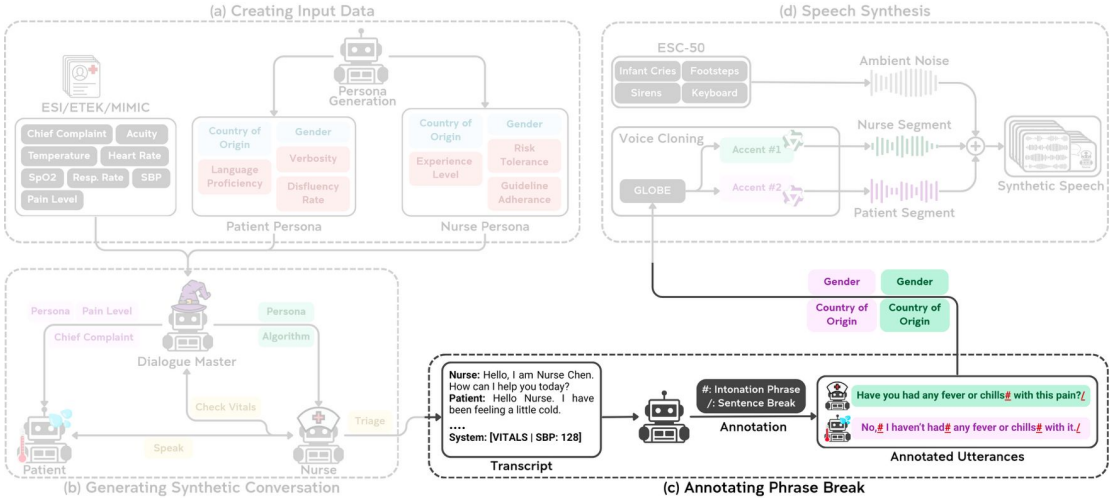


+ Dialogue

- **Dialogue Master**: sees the full ground truth and sets up the other agents
- **Nurse agent**: persona and triage algorithm; speaks, checks vitals, triages
- **Patient agent**: persona, chief complaint and pain level

TriageSim: A Conversational Emergency Triage Simulation Framework from Structured Electronic Health Records (**Srirag et al., 2026**)

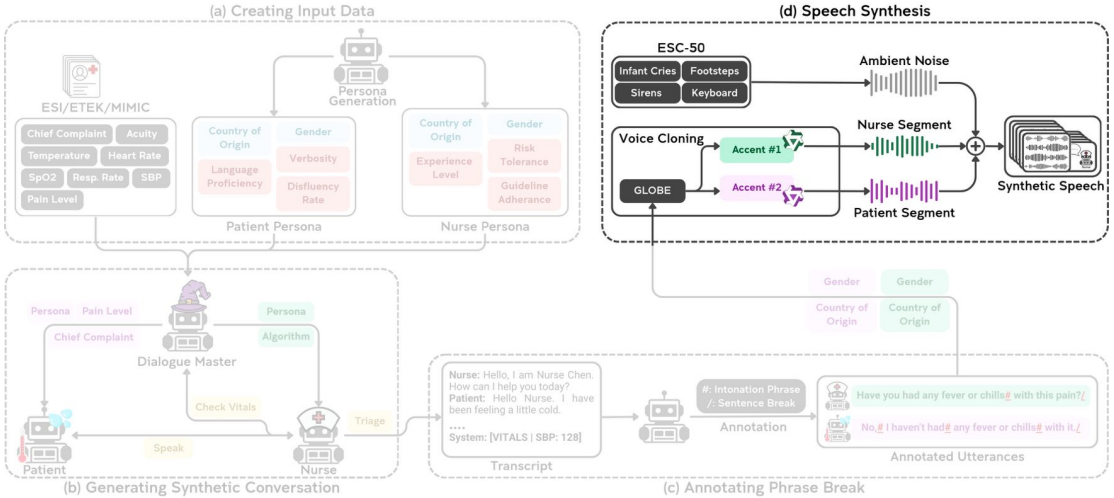
Generating Phrase Break



- **Phrase boundaries** annotated, for explicit control of prosody in synthesis
- **#** intonation phrase, **/** sentence break
- Uses the framework of ***Lee et al. (2025)***

TriageSim: A Conversational Emergency Triage Simulation Framework from Structured Electronic Health Records (**Srirag et al., 2026**)

Speech Synthesis



- **Voice cloning**: a voice from GLOBE, sampled by the speaker's demographics, cloned with Qwen-3-TTS
- **Ambient noise**: infant cries, footsteps, sirens, from ESC-50

+ Audio

TriageSim: A Conversational Emergency Triage Simulation Framework from Structured Electronic Health Records (**Srirag et al., 2026**)

Dataset Statistics

Acuity	Di.	Sp.	Ut.	Tokens	Hours
1	73	69	9.95	21K	2.02
2	231	138	11.46	78K	7.30
3	309	147	12.21	109K	10.37
4	112	108	11.95	42K	3.68
5	89	83	12.07	30K	2.77
Total	814	150	11.52	280K	26.14

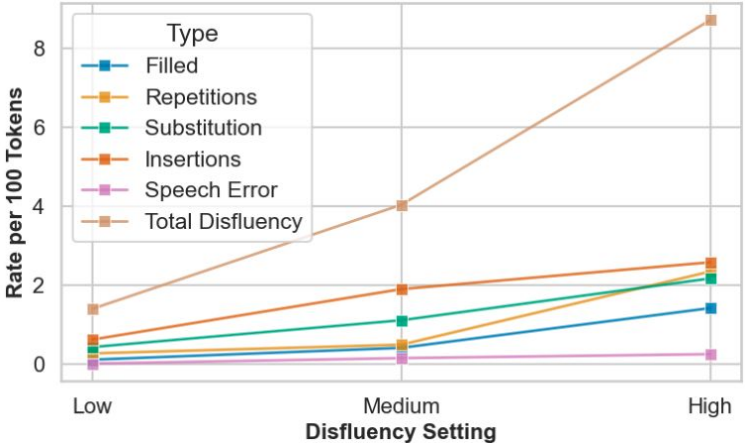
Corpus statistics for each acuity level

Di.: #conversations
Sp.: #speakers
Ut.: #utterances

TriageSim: A Conversational Emergency Triage Simulation Framework from Structured Electronic Health Records (**Srirag et al., 2026**)

Robustness

Linguistic and Behavioural Fidelity

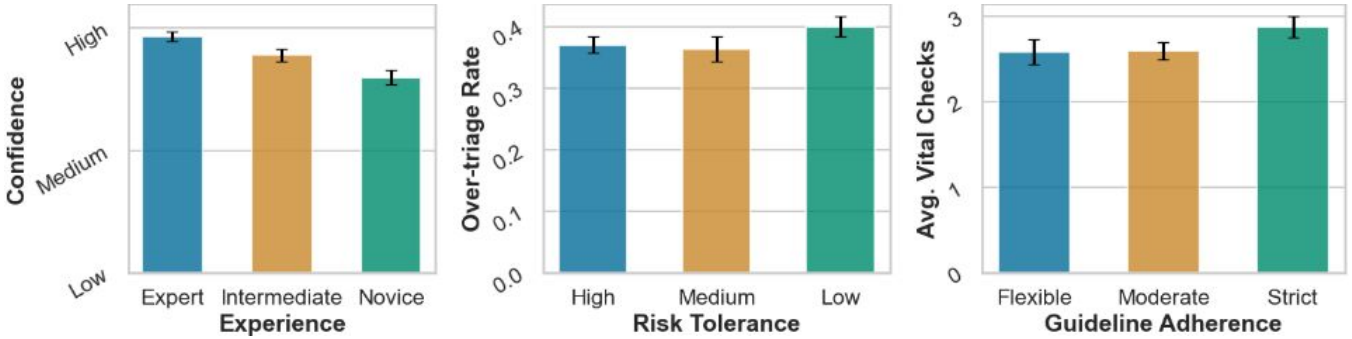


Type	Example
Repetition	**It's... it's** maybe six.
Filled pause	I was **uh** feeling dizzy.
Insertion	Yes, **actually,** I've been going less frequently.
Substitution	I have a kidney-**surgery, no transplant.**
Speech error	**O-okay...** please help her.

- **Patient disfluency** follows its setting: ρ = 0.57
- Setting vs. disfluency probability: ρ = 0.45

TriageSim: A Conversational Emergency Triage Simulation Framework from Structured Electronic Health Records (**Srirag et al., 2026**)

Linguistic and Behavioural Fidelity



- **Confidence** in the final level increases with **experience**
- Slightly more **over-triage** with lower **risk tolerance**
- **Strict** nurse agents ask for more **vital checks**

TriageSim: A Conversational Emergency Triage Simulation Framework from Structured Electronic Health Records (**Srirag et al., 2026**)

Acoustic Fidelity

Following *Wang et al., (2025)*:

- **Intelligibility**: Whisper word error rate, **10.8** overall

	Int. (↓)	Cons. (↑)
Overall	10.8	99.98
Speaker Role		
Nurse	5.7	99.97
Patient	16.0	99.97
Patient Accent Group		
Australian	14.2	100.0
Chinese	13.2	100.0
Indian	16.7	99.88
Middle Eastern	19.4	100.0

Acoustic Fidelity

Following SpeechDialogueFactory (Wang et al., 2025):

- **Intelligibility**: Whisper word error rate, **10.8** overall
- **Consistency**: **99.98**, each speaker keeps one voice
- **Quality**: UTMOS **3.42**, moderate-to-high naturalness

	Int. (↓)	Cons. (↑)
Overall	10.8	99.98
Speaker Role		
Nurse	5.7	99.97
Patient	16.0	99.97
Patient Accent Group		
Australian	14.2	100.0
Chinese	13.2	100.0
Indian	16.7	99.88
Middle Eastern	19.4	100.0

TriageSim: A Conversational Emergency Triage Simulation Framework from Structured Electronic Health Records (**Srirag et al., 2026**)

Robustness

Medical Fidelity

A medical expert read 50 generated transcripts:

- Drafted the **chief complaint**: similarity **0.83** with the record's

Metric	Score
Chief Complaint Sim.	0.83
Red Flag Precision	0.94
Red Flag Recall	0.96
Red Flag F1	0.94

TriageSim: A Conversational Emergency Triage Simulation Framework from Structured Electronic Health Records (**Srirag et al., 2026**)

Robustness

Medical Fidelity

A medical expert read 50 generated transcripts:

- Drafted the **chief complaint**: similarity **0.83** with the record's
- Checked the **red flags** the nurse agent named: precision **0.94**, recall **0.96**

Metric	Score
Chief Complaint Sim.	0.83
Red Flag Precision	0.94
Red Flag Recall	0.96
Red Flag F1	0.94

TriageSim: A Conversational Emergency Triage Simulation Framework from Structured Electronic Health Records (**Srirag et al., 2026**)

What You Would Need

To test a model for this ED, you would need:

- **conversations, not finished records** → 814 conversations, generated from structured records

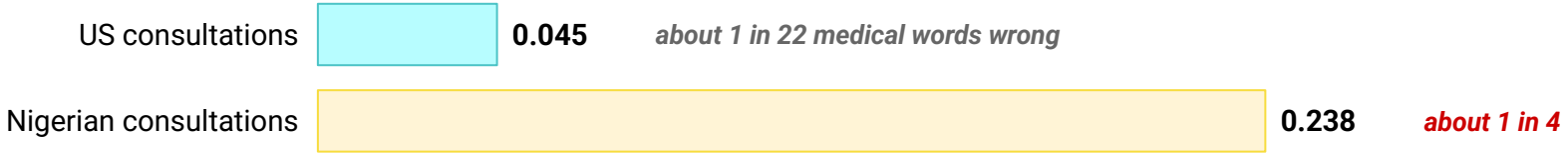
- **their voice and their accent, said aloud** → 150 speakers: Australian, Chinese, Indian, Middle Eastern; ED noise

PhysioNet

TriageSim: A Conversational Emergency Triage Simulation Framework from Structured Electronic Health Records (**Srirag et al., 2026**)

Another Accent

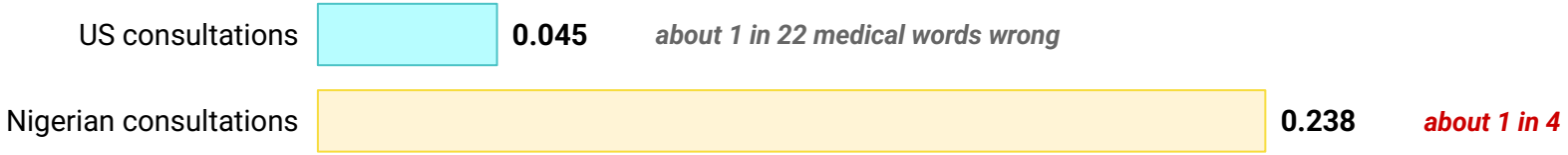
Medical-term word error rate: the same kind of consultation, two accents



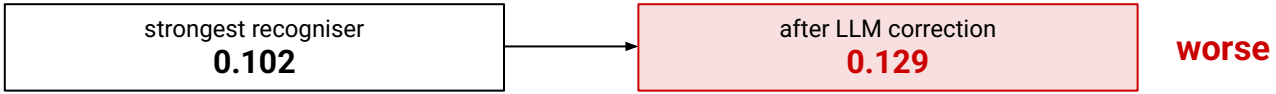
The Multicultural Medical Assistant: Can LLMs Improve Medical ASR Errors Across Borders? (**Adedeji et al., 2025**)

Another Accent

Medical-term word error rate: the same kind of consultation, two accents



Ask an LLM to correct the transcript? (overall WER, Nigerian consultations)

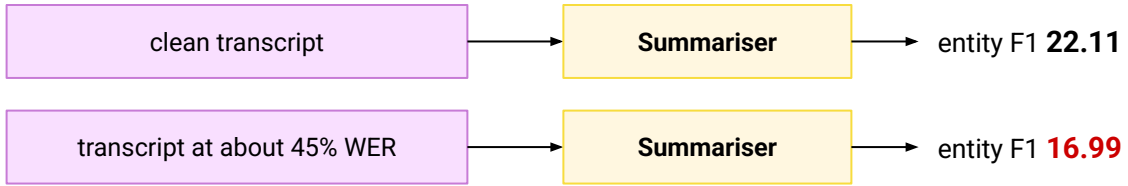


The Multicultural Medical Assistant: Can LLMs Improve Medical ASR Errors Across Borders? (**Adedeji et al., 2025**)

As Noise-Aware Training

x is:

- Dialogue
- Speech

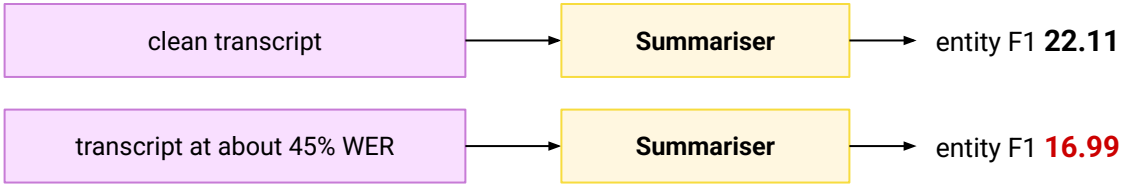


MEDSAGE: Enhancing Robustness of Medical Dialogue Summarization to ASR Errors with LLM-generated Synthetic Dialogues (**Binici et al., 2025**)

As Noise-Aware Training

x is:

- Dialogue
- Speech



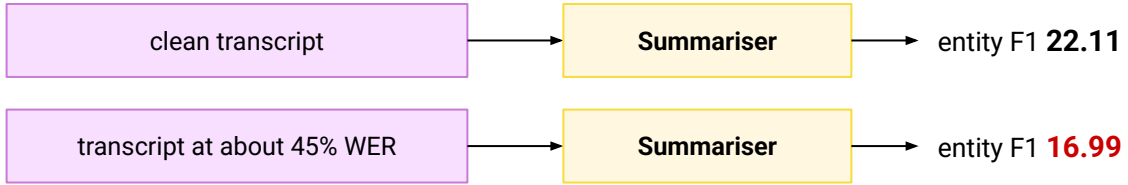
Clean it with an LLM first? Worse.

MEDSAGE: Enhancing Robustness of Medical Dialogue Summarization to ASR Errors with LLM-generated Synthetic Dialogues (**Binici et al., 2025**)

As Noise-Aware Training

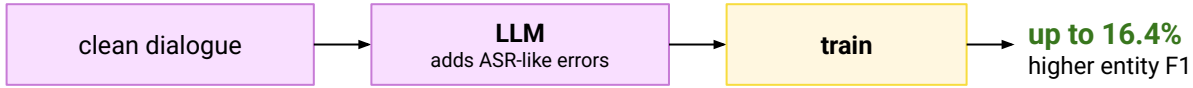
x is:

- Dialogue
- Speech
- **Simulated ASR errors**



Clean it with an LLM first? Worse.

Train on the noise instead:



MEDSAGE: Enhancing Robustness of Medical Dialogue Summarization to ASR Errors with LLM-generated Synthetic Dialogues (**Binici et al., 2025**)

Temporal Validity

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid x)$$

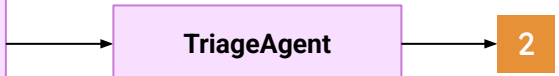
the finished record

Was it tested only on what was known at the time of the decision?

Temporal Validity

A Finished Vignette

A 24-year-old man with sharp chest pain and shortness of breath. HR 118, RR 24, SpO_2 93%. Scaphoid pinned two weeks ago. Unsure about leg swelling.



a vignette: triage case note

TriageAgent: Towards Better Multi-Agents Collaborations for Large Language Model-Based Clinical Triage (**Lu et al., 2024**)

Temporal Validity

A Finished Vignette

A 24-year-old man with sharp chest pain and shortness of breath. HR 118, RR 24, SpO_2 93%. Scaphoid pinned two weeks ago. Unsure about leg swelling.

→ **TriageAgent** → 2

a vignette: triage case note

At the front door, it arrives a turn at a time:



I've got this sharp pain in my chest, and I can't catch my breath.



Any recent surgery or injuries?



They put pins in my scaphoid two weeks ago.

?

?

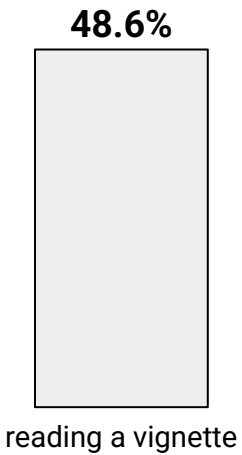
What level, after each line?

TriageAgent: Towards Better Multi-Agents Collaborations for Large Language Model-Based Clinical Triage (**Lu et al., 2024**)

Temporal Validity

Vignette to Conversation

GPT-4, same cases, diagnostic accuracy

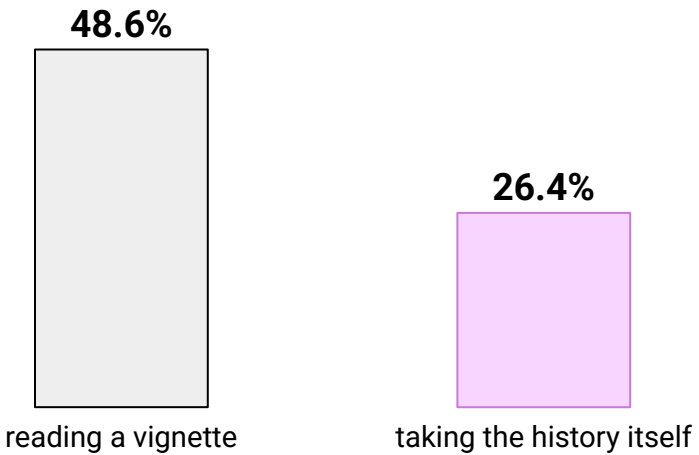


An evaluation framework for clinical use of large language models in patient interaction tasks (**Johri et al., 2025**)

Temporal Validity

Vignette to Conversation

GPT-4, same cases, diagnostic accuracy

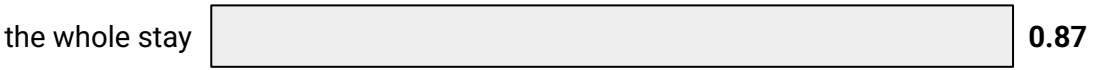


- Same model, same cases
- Only **who gathers the information** changed

An evaluation framework for clinical use of large language models in patient interaction tasks (**Johri et al., 2025**)

Scored at Decision Time

A sepsis model, scored on data from…

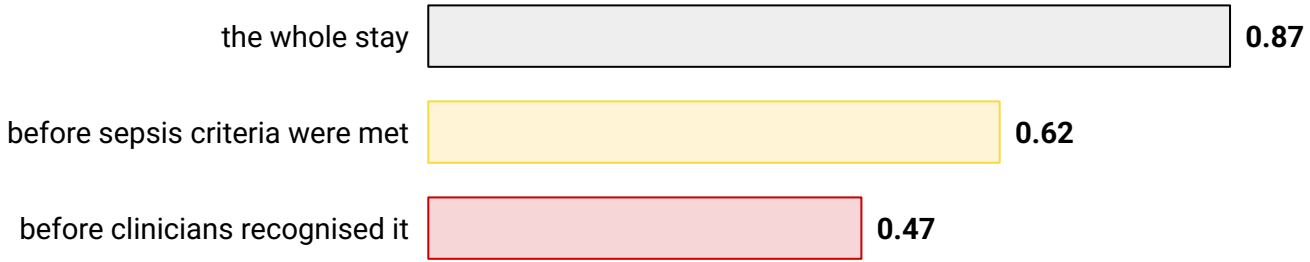


Evaluation of Sepsis Prediction Models before Onset of Treatment (**Kamran et al., 2024**)

Temporal Validity

Scored at Decision Time

A sepsis model, scored on data from…



- Below 0.5 is **worse than a coin flip**
- Part of the score came from **what the clinicians had already done**

Evaluation of Sepsis Prediction Models before Onset of Treatment (**Kamran et al., 2024**)

Temporal Validity

As Sequential Triage

x is:

- Triage record

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid x)$$

the finished record

chief complaint: chest pain, short of breath | HR 118 | RR 24 | SpO_2 93% | BP 128/78 | age 24 | male → **LLM** → ? *one level, from everything at once*

- How triage LLMs are usually tested: on the **record the nurse completed after the conversation**

LLMs Anchor on Chief Complaint and Fail to Integrate Evidence in Sequential Clinical Triage (**Srirag et al., 2026**)

Temporal Validity

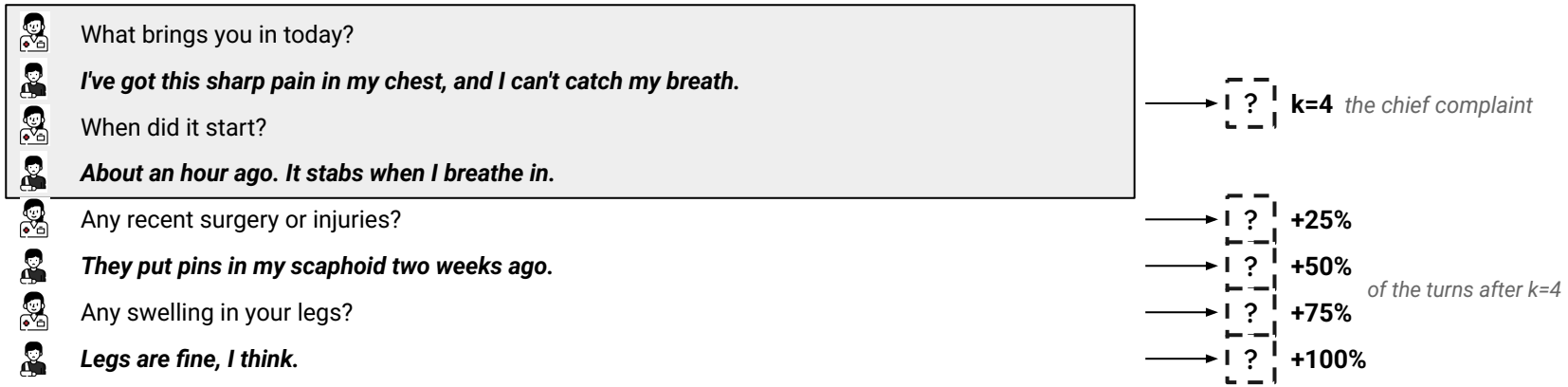
As Sequential Triage

x is:

- **Conversation so far**

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid x)$$

the conversation so far, at checkpoint k



LLMs Anchor on Chief Complaint and Fail to Integrate Evidence in Sequential Clinical Triage (**Srirag et al., 2026**)

Temporal Validity

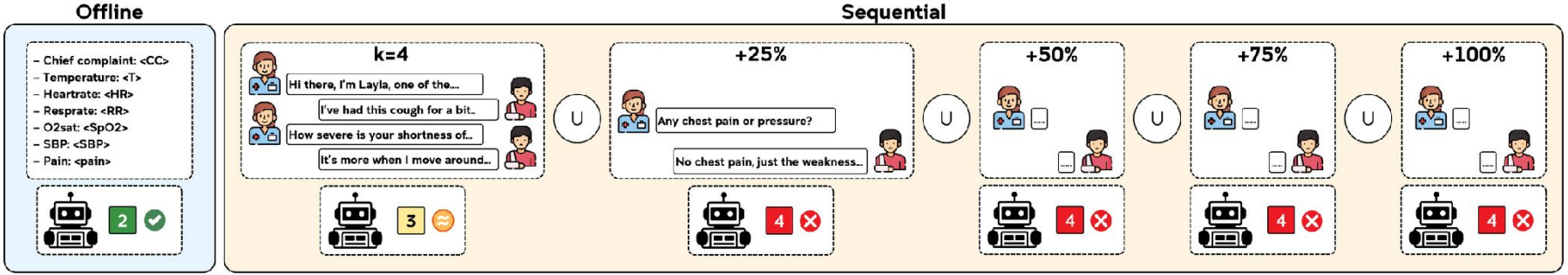
As Sequential Triage

x is:

- **Conversation so far**

$$\hat{y} = \arg\max_{y \in Y} p_\theta(y \mid x)$$

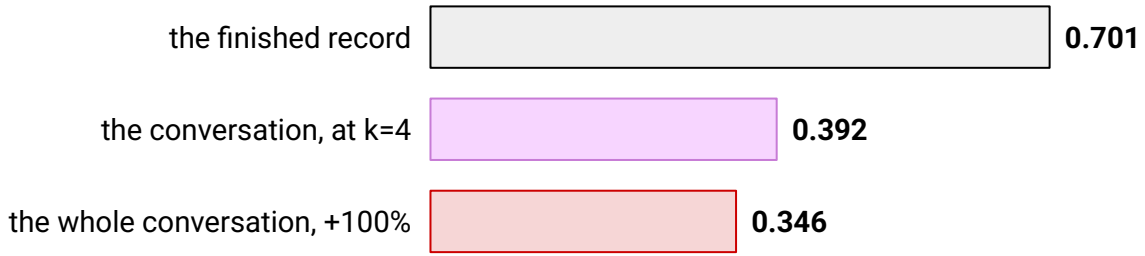
the conversation so far, at checkpoint k



LLMs Anchor on Chief Complaint and Fail to Integrate Evidence in Sequential Clinical Triage (**Srirag et al., 2026**)

Told Turn by Turn

Claude Opus 4.6: agreement with the true ESI level (QWK)

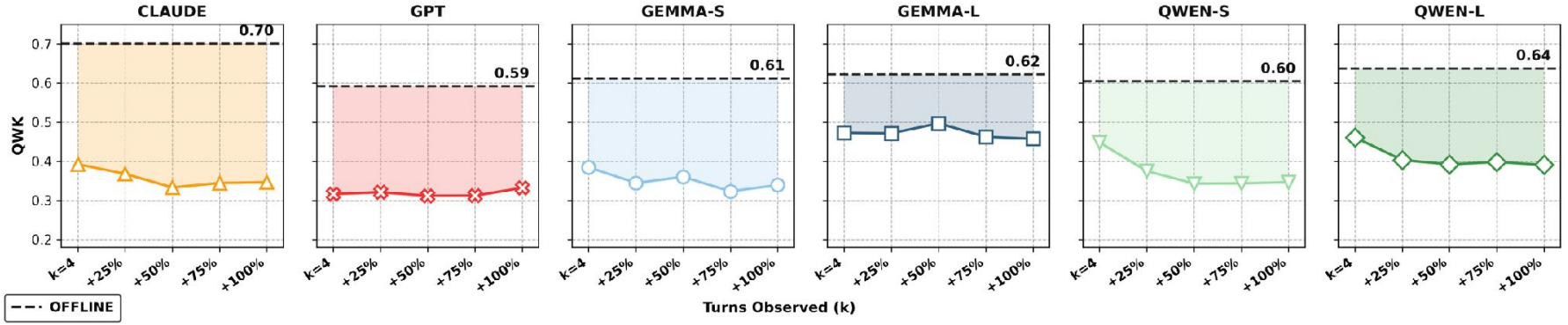


- The same cases, told as a conversation: **about half the agreement**
- **More turns did not help**

LLMs Anchor on Chief Complaint and Fail to Integrate Evidence in Sequential Clinical Triage (**Srirag et al., 2026**)

Temporal Validity

Told Turn by Turn



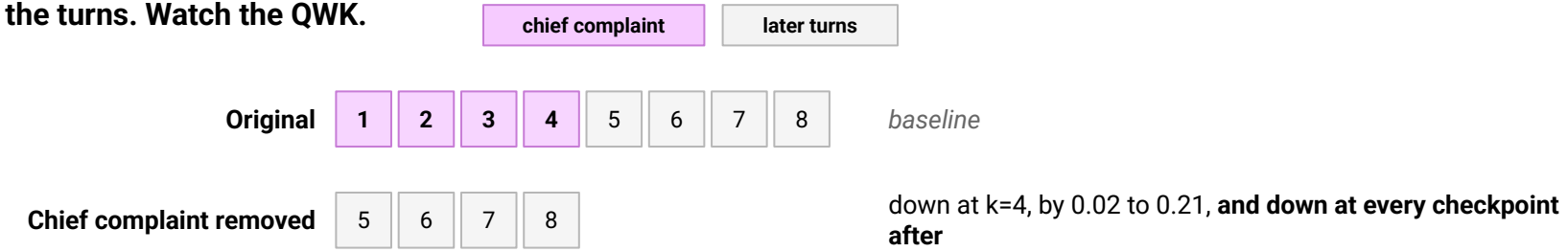
- Every model, at every checkpoint, **below its finished-record score** (dashed)
- **Flat from k=4**
- Reasoning mode, plan-and-solve and self-consistency **did not lift it**

LLMs Anchor on Chief Complaint and Fail to Integrate Evidence in Sequential Clinical Triage (**Srirag et al., 2026**)

Temporal Validity

Anchored on the Chief Complaint

Rearrange the turns. Watch the QWK.

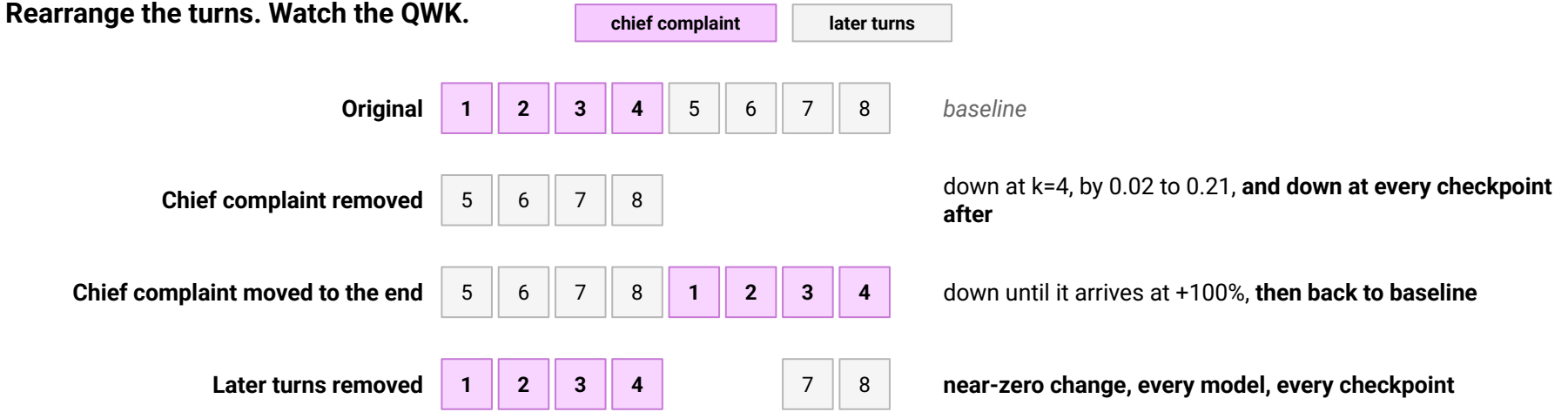


LLMs Anchor on Chief Complaint and Fail to Integrate Evidence in Sequential Clinical Triage (**Srirag et al., 2026**)

Temporal Validity

Anchored on the Chief Complaint

Rearrange the turns. Watch the QWK.

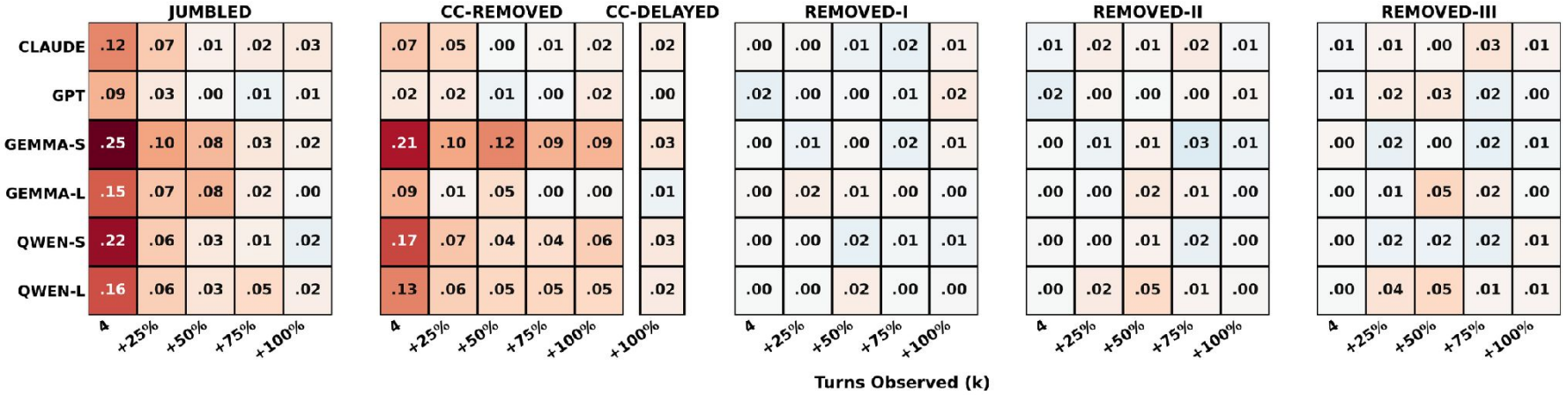


- The level is set by **the first four turns**

LLMs Anchor on Chief Complaint and Fail to Integrate Evidence in Sequential Clinical Triage (**Srirag et al., 2026**)

Temporal Validity

Anchored on the Chief Complaint

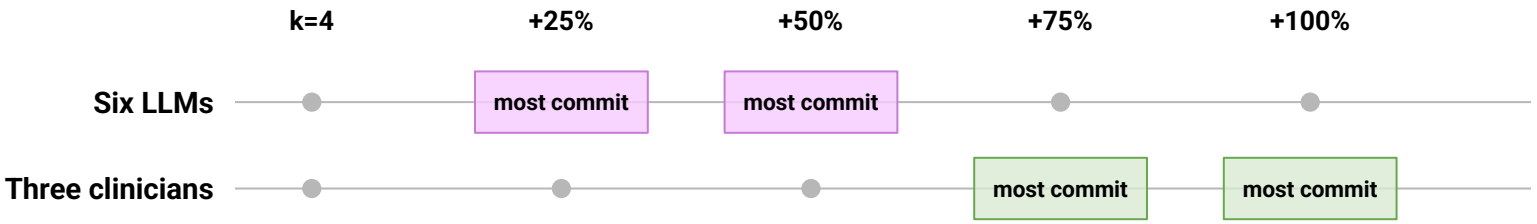


- QWK moves only when **the first four turns** are touched

LLMs Anchor on Chief Complaint and Fail to Integrate Evidence in Sequential Clinical Triage (**Srirag et al., 2026**)

Clinicians Wait, Models Commit

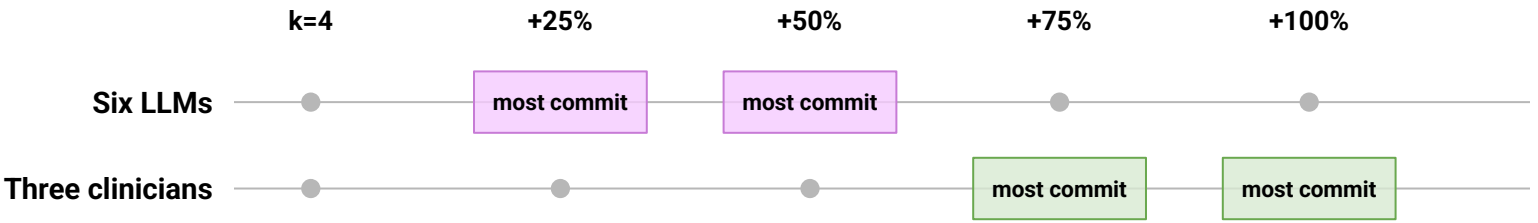
At each checkpoint: commit to a level, or wait for more turns.



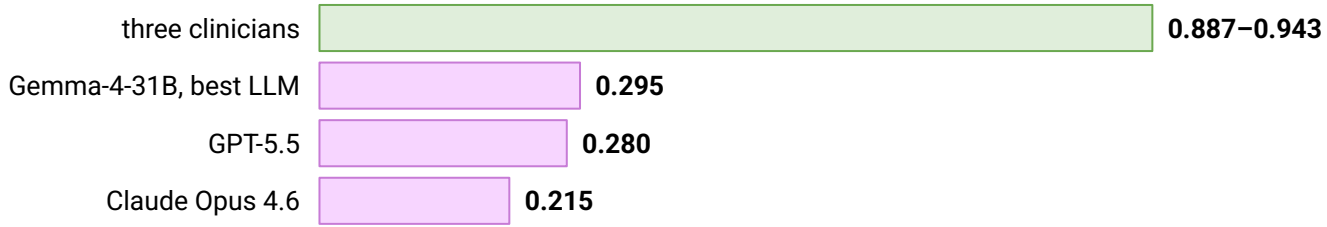
LLMs Anchor on Chief Complaint and Fail to Integrate Evidence in Sequential Clinical Triage (**Srirag et al., 2026**)

Clinicians Wait, Models Commit

At each checkpoint: commit to a level, or wait for more turns.



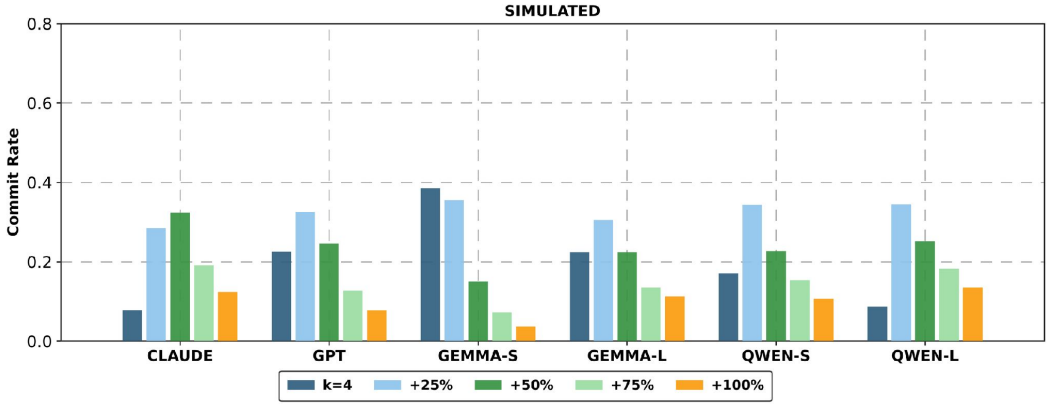
Agreement with the true level, at the moment of commit (same 50 conversations)

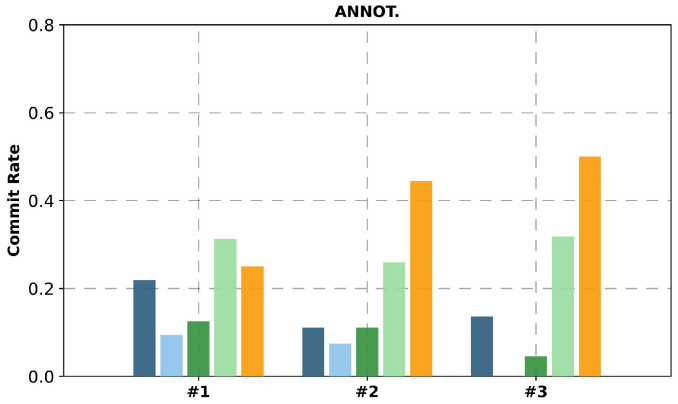


LLMs Anchor on Chief Complaint and Fail to Integrate Evidence in Sequential Clinical Triage (**Srirag et al., 2026**)

Temporal Validity

Clinicians Wait, Models Commit





	GEMMA-S	GEMMA-L	QWEN-S	QWEN-L	CLAUDE	GPT	ANNOT #1	ANNOT #2	ANNOT #3
QWK	0.009	0.295	0.016	0.195	0.215	0.280	0.929	0.943	0.887

LLMs Anchor on Chief Complaint and Fail to Integrate Evidence in Sequential Clinical Triage (**Srirag et al., 2026**)

Extracted, Not Integrated

1 **Extract.** Find what the later turns say.

Chief complaint, rewritten from the later turns alone: **0.57–0.77** cosine with the real one

Red flags both Claude and GPT find: **3.8** per ESI 1 conversation, **1.8** per ESI 5

- **It extracts**

2 **Integrate.** Let it move the level.

LLMs Anchor on Chief Complaint and Fail to Integrate Evidence in Sequential Clinical Triage (**Srirag et al., 2026**)

Temporal Validity

Extracted, Not Integrated

1 **Extract.** Find what the later turns say.

Chief complaint, rewritten from the later turns alone: **0.57–0.77** cosine with the real one

Red flags both Claude and GPT find: **3.8** per ESI 1 conversation, **1.8** per ESI 5

- **It extracts**

2 **Integrate.** Let it move the level.

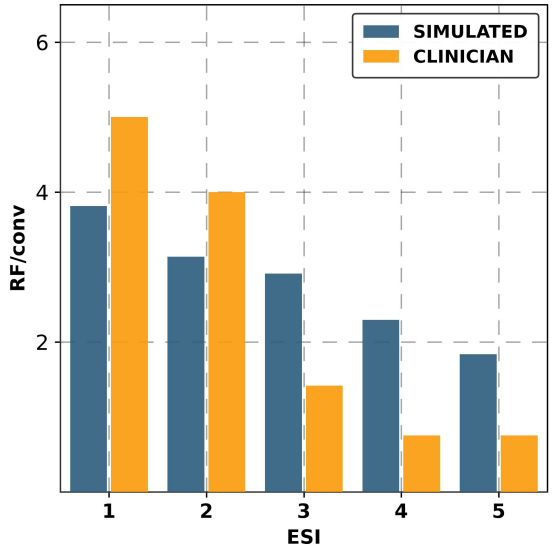
NLL, k=4 to +100%: **up 0.74 to 1.82 nats**, every open-weight model

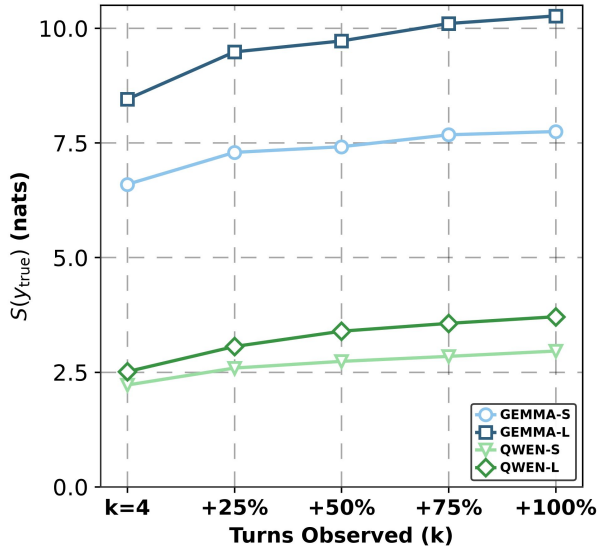
- More evidence, and the true level becomes **less likely**
- **It does not integrate**

LLMs Anchor on Chief Complaint and Fail to Integrate Evidence in Sequential Clinical Triage (**Srirag et al., 2026**)

Temporal Validity

Extracted, Not Integrated





- Test triage models **on the conversation, checkpoint by checkpoint**

LLMs Anchor on Chief Complaint and Fail to Integrate Evidence in Sequential Clinical Triage (**Srirag et al., 2026**)

Evaluation and Clinical Validation

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid x)$$

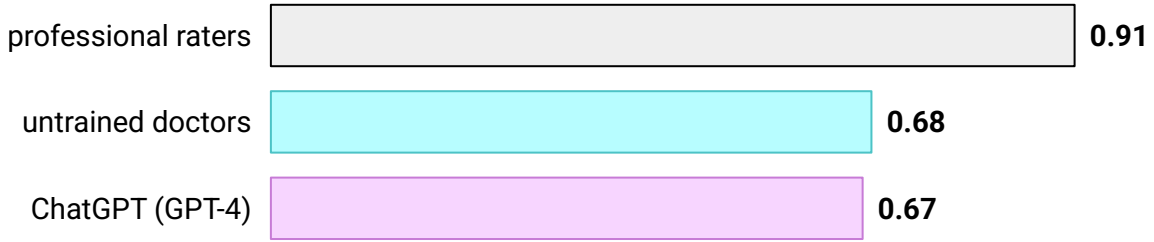
scored right or wrong

Do its scores measure what matters to the patient?

Evaluation

Agreement Hides Direction

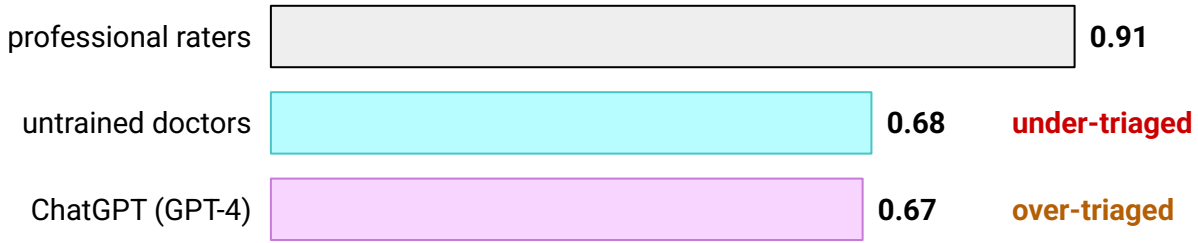
Agreement with the reference triage level (weighted κ)



Triage Performance Across Large Language Models, ChatGPT, and Untrained Doctors in Emergency Medicine: Comparative Study (**Masanneck et al., 2024**)

Agreement Hides Direction

Agreement with the reference triage level (weighted κ)



- Almost the **same κ**
- Errors in **opposite directions**

Under-triage

- A level too low: the patient waits when he should not.

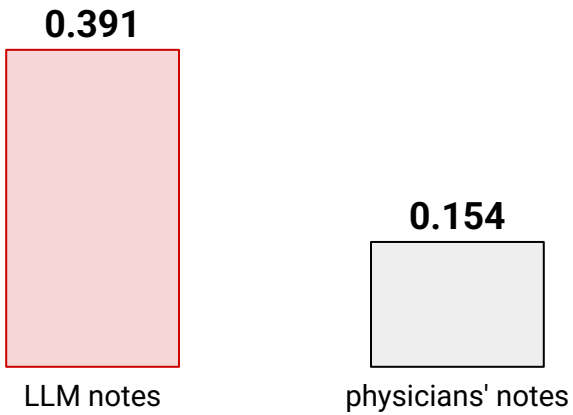
Over-triage

- He is seen sooner than needed.

Triage Performance Across Large Language Models, ChatGPT, and Untrained Doctors in Emergency Medicine: Comparative Study (**Masanneck et al., 2024**)

The Metric and the Clinicians

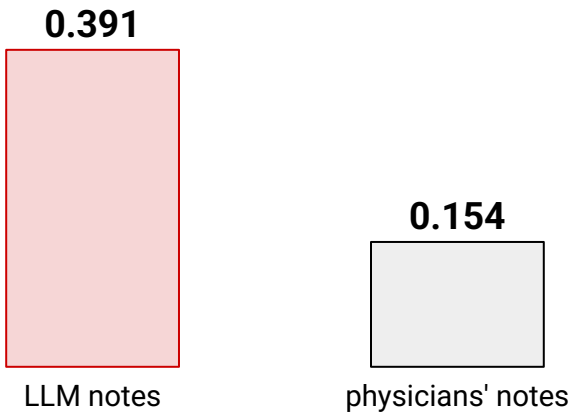
What ROUGE-L says



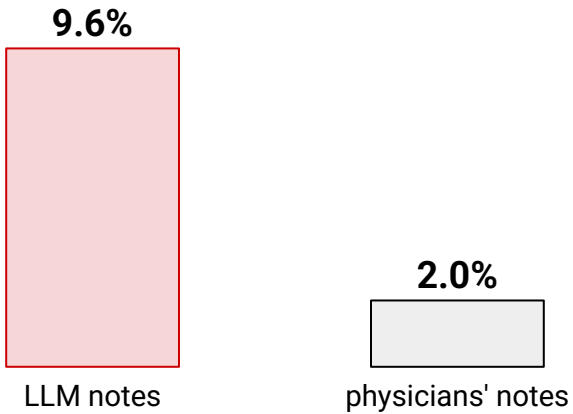
Developing and Evaluating Large Language Model–Generated Emergency Medicine Handoff Notes (**Hartman et al., 2024**)

The Metric and the Clinicians

What ROUGE-L says



Notes with an error, by clinicians

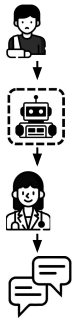


Developing and Evaluating Large Language Model–Generated Emergency Medicine Handoff Notes (**Hartman et al., 2024**)

As a Prospective Study

x is:

- **Real patients**



98 real patients

chat with AMIE, in text

a physician watches every chat, and can stop it

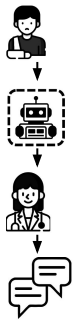
checked later, against the final diagnosis

Conversational diagnostic artificial intelligence in ambulatory primary care: a prospective feasibility study (**Brodeur et al., 2026**)

As a Prospective Study

x is:

- **Real patients**



98 real patients

chat with AMIE, in text

a physician watches every chat, and can stop it

checked later, against the final diagnosis

The final diagnosis was…



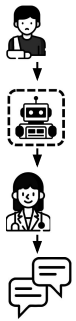
Conversational diagnostic artificial intelligence in ambulatory primary care: a prospective feasibility study (**Brodeur et al., 2026**)

Evaluation

As a Prospective Study

x is:

- **Real patients**



98 real patients

chat with AMIE, in text

a physician watches every chat, and can stop it

checked later, against the final diagnosis

The final diagnosis was…



Not yet shown

- the ED: this was primary care
- speech: this was text
- a second site

Conversational diagnostic artificial intelligence in ambulatory primary care: a prospective feasibility study (**Brodeur et al., 2026**)

Deployment and Workflow Integration

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid x)$$

Beyond a research paper

Would the ED adopt it: is it safe to run, and does it fit the work?

Rarely Tested in Care

Where the evidence for clinical NLP comes from

Our survey of ED NLP

- Papers that report a deployment inside an ED: none

519 studies of LLMs in health care (Bedi et al.)

- Used real patient care data: 5%
- Measured deployment considerations: 4.6%

From Triage to Discharge: A Survey of NLP Tasks, Methods, and Open Challenges in the Emergency Department (**Srirag et al., 2026**)

Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review (**Bedi et al., 2025**)

Rarely Tested in Care

Where the evidence for clinical NLP comes from

Our survey of ED NLP

- Papers that report a deployment inside an ED: **none**

519 studies of LLMs in health care (Bedi et al.)

- Used real patient care data: **5%**
- Measured deployment considerations: **4.6%**

- NLP is judged on **benchmarks**; adoption is decided **in care**

From Triage to Discharge: A Survey of NLP Tasks, Methods, and Open Challenges in the Emergency Department (**Srirag et al., 2026**)

Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review (**Bedi et al., 2025**)

Deployment

Generated Text Enters the Record

Ambient AI notes from real visits: 356 rated by the physicians who signed them

- Omissions in 18% of notes
- Hallucinations in 11.5%
- An error posing serious or imminent risk in 5.3%
- Physicians changed a median 9.0% of the words (1.9% to 69.3% across physicians)

Quality of Clinical Notes Created by Ambient Listening Generative AI: Pragmatic Prospective Pilot Study (**Taylor et al., 2026**)

Deployment

Generated Text Enters the Record

Ambient AI notes from real visits: 356 rated by the physicians who signed them

- Omissions in **18%** of notes
- Hallucinations in **11.5%**
- An error posing **serious or imminent risk** in **5.3%**
- Physicians changed a median **9.0%** of the words (**1.9% to 69.3%** across physicians)

- The text goes **into the record**, so someone **checks and signs every note**

Quality of Clinical Notes Created by Ambient Listening Generative AI: Pragmatic Prospective Pilot Study (**Taylor et al., 2026**)

The Model Changes Under You

The same hosted model, GPT-4, three months apart

- Prime number questions: 84.0% in March 2023, 51.1% in June
- USMLE questions: 86.6% in March, 82.4% in June

How Is ChatGPT's Behavior Changing Over Time? (**Chen et al., 2024**)

The Model Changes Under You

The same hosted model, GPT-4, three months apart

- Prime number questions: 84.0% in March 2023, **51.1%** in June
- USMLE questions: 86.6% in March, **82.4%** in June

- A hosted LLM can **change under a deployed tool**
- A tool validated in March is **not the tool running in June**

How Is ChatGPT's Behavior Changing Over Time? (**Chen et al., 2024**)

AJCAI 2026 - Post Survey



Break

See you in 10 mins

Readiness

Put on your practitioner’s hat

Integration Readiness Levels

	Where it runs	The most it may do alone	
IRL2	Noisy, incomplete or adversarial simulation	Manual review, plus a safety gateway	simulation
IRL1	Static or scripted sandbox	Manual review before any commit	

LLM-Based Multi-Agent Systems for Clinical Workflows: A Survey of AI Hospitals (**Yao & Yu, 2026**)

Integration Readiness Levels

	Where it runs	The most it may do alone	
IRL4	Limited human-in-the-loop pilot	Suggests bounded steps; approval for high stakes	replay, then a pilot
IRL3	Shadow replay on real data, no live impact	It proposes; humans decide	
IRL2	Noisy, incomplete or adversarial simulation	Manual review, plus a safety gateway	simulation
IRL1	Static or scripted sandbox	Manual review before any commit	

LLM-Based Multi-Agent Systems for Clinical Workflows: A Survey of AI Hospitals (**Yao & Yu, 2026**)

Integration Readiness Levels

	Where it runs	The most it may do alone	
IRL6	Scaled across sites or languages	Supervised autonomy, periodic audits	rollout
IRL5	Limited rollout, monitored end to end	Acts by default; exceptions reviewed	
IRL4	Limited human-in-the-loop pilot	Suggests bounded steps; approval for high stakes	replay, then a pilot
IRL3	Shadow replay on real data, no live impact	It proposes; humans decide	
IRL2	Noisy, incomplete or adversarial simulation	Manual review, plus a safety gateway	simulation
IRL1	Static or scripted sandbox	Manual review before any commit	

LLM-Based Multi-Agent Systems for Clinical Workflows: A Survey of AI Hospitals (**Yao & Yu, 2026**)

Each Frontier Gates a Rung

	Where it runs	
IRL6	Scaled across sites or languages	→ **Generalisability**
IRL5	Limited rollout, monitored end to end	→ **Deployment**
IRL4	Limited human-in-the-loop pilot	→ **Evaluation**
IRL3	Shadow replay on real data, no live impact	→ **Temporal Validity**
IRL2	Noisy, incomplete or adversarial simulation	→ **Robustness**
IRL1	Static or scripted sandbox	*it works, on a fixed test set*

- To climb a rung, answer its **frontier's question**, with evidence

LLM-Based Multi-Agent Systems for Clinical Workflows: A Survey of AI Hospitals (**Yao & Yu, 2026**)

Our Patient's Triage Model

TriageAgent, from Module 1, rung by rung

IRL4	
IRL3	
IRL2	
IRL1	**Today:** handbook vignettes; it reasons from the protocol, no local data

TriageAgent: Towards Better Multi-Agents Collaborations for Large Language Model-Based Clinical Triage (**Lu et al., 2024**)

Our Patient's Triage Model

TriageAgent, from Module 1, rung by rung

IRL4	**Then, through Evaluation:** a supervised pilot; a nurse watches; under-triage counted
IRL3	**Then, through Temporal Validity:** a silent run on real triage, scored at every turn
IRL2	**Next, through Robustness:** TriageSim conversations, in four accents, as recognised speech
IRL1	**Today:** handbook vignettes; it reasons from the protocol, no local data

TriageAgent: Towards Better Multi-Agents Collaborations for Large Language Model-Based Clinical Triage (**Lu et al., 2024**)

Activity 1

Rate it

Rate it

1. **Read the card.** Where did it run, and what did it show?
2. **Climb from IRL1.** Stop at the first rung it has not earned: that is its level.
3. **Read its claim.** What would it do alone? That rung is its claim.
4. **Name the blocker.** Which frontier gates the next rung?

TriageAgent is at IRL1, claims IRL4, and is blocked by Robustness because it has only seen clean, finished vignettes.

Rate it

Card	Evidenced IRL	Claimed IRL	Blocked by	Evidence row
A TriageGPT-ESI				
B Waiting-Room Companion				
C ShiftScribe				
D AdmitAhead				
E LinguaTriage				
F ClearGoHome				

Activity 2

Promote it

Promote it

1. **Next rung.** Which rung, and which frontier gates it?
2. **The study.** Where, on which patients, who watches, compared with what?
3. **One number.** About harm to patients, with its threshold.
4. **The way back down.** What sends it back a rung, and who is told?

Card _ goes to IRL _ through _: we run _, and it moves up only if _.

Promote it

1. **Next rung.** Which rung, and which frontier gates it?
2. **The study.** Where, on which patients, who watches, compared with what?
3. **One number.** About harm to patients, with its threshold.
4. **The way back down.** What sends it back a rung, and who is told?

TriageAgent goes to IRL2 through Robustness: we run TriageSim in four accents, and it moves up only if under-triage of ESI 1–2 stays at 5% or less in every accent.

Promote it

Card	Next rung	Through	The study	Number, threshold	Way back down
A TriageGPT-ESI					
B Waiting-Room Companion					
C ShiftScribe					
D AdmitAhead					
E LinguaTriage					
F ClearGoHome					

Tutorial Plan

Module 1: NLP for Emergency Department

Module 2: Typical Tasks and Methods

Module 3: Frontiers & Readiness



Module 4: CSE Researchers in Clinical AI

TRIBOT

Medicine & Health

School of Population Health

Home About us Study areas Study Research Partner with us

TRIBOT - Multilingual AI for emergency care

Transforming emergency care with AI-powered multilingual triage

Contact us

Trial TRIBOT





Dr Padmanesan Narasimhan (Padma)

CIA-A · Founder & Lead Investigator

Senior Lecturer & Lead, Digital Health & AI — UNSW Medicine & Health, School of Population Health; ED Clinician, SWSLHD. WHO-recognised expert in AI and Digital Health with a strong track record in epidemiology, health systems, and emergency medicine.



Roberto Forero

Senior Research Fellow and Post Graduate Coordinator SWS Clinical Campuses, School of Clinical Medicine

View Profile →



Aditya Joshi

Senior Lecturer

View Profile →



Sandra Hale

Professor of Interpreting and Translation

View Profile →



Jitendra Jonnagaddala

View Profile →



Usman Iqbal

Adjunct Associate Professor

View Profile →

Mohammed Mohsin

View Profile →

CSE Researchers in Clinical AI

Are you a CS PhD aspirant/CS researcher interested in research in NLP for ED (and broadly Clinical AI)?

The following slides are **reflective** and **subjective**.

Publicly Available Datasets

Data Portal

PhysioNet

Some datasets have credentialled access

May require: (a) Login, (b) Training, (c) Data Usage Agreement

Shared Tasks

MEDIQA-RADEVAL @ NAACL-ClinicalNLP 2027

Overview of the MEDIQA-Chat 2023 Shared Tasks on the Summarization & Generation of Doctor-Patient Conversations

Overview of the First Shared Task on Clinical Text Generation: RRG24 and "Discharge Me!"

Creating/collecting own datasets

Annotation:

- Multiple annotators
- Define annotator expertise; report inter-annotator agreement

Ethics Approval:

- Please consult your institution’s ethics review committee (or similar name).
- Publicly available social media data may require ethics approval too.

Methodology

Prompting may be tempting: Does your data usage permission allow for commercial API usage?

Have you measured for variance/bias in your results?

Methodology informed by practitioner needs/wisdom is valuable.

Interdisciplinary Researchers

Clarify terminology

“Coding”

Appropriately navigate power relationships

Calibrate expectations with respect to AI

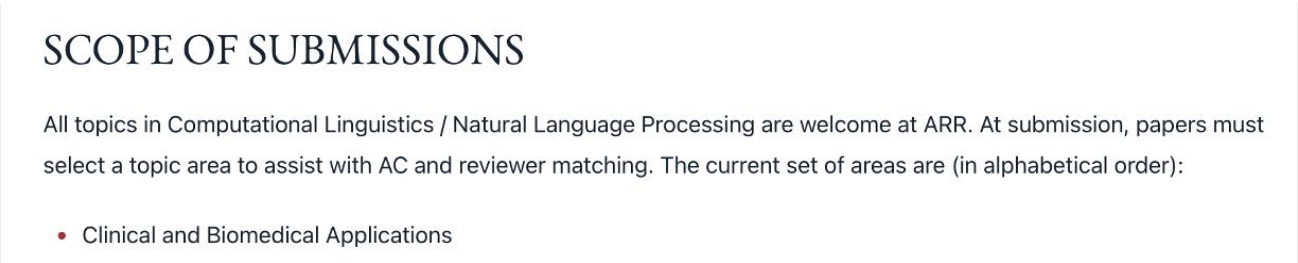
NLP Research Venues

Workshops:

- Workshop on Biomedical Language Processing (BioNLP) (25th in 2026!)
- Clinical Natural Language Processing Workshop (5th in 2027; upcoming)

Special Interest Group (SIG) SIGCLINICAL for “engagement among members of the ACL, clinicians, and clinical researchers ” https://sigclinical.org/

ACL Rolling Review



SCOPE OF SUBMISSIONS

All topics in Computational Linguistics / Natural Language Processing are welcome at ARR. At submission, papers must select a topic area to assist with AC and reviewer matching. The current set of areas are (in alphabetical order):

- Clinical and Biomedical Applications

Tutorial Plan

Module 1: NLP for Emergency Department

Module 2: Typical Tasks and Methods

Module 3: Frontiers & Readiness

Module 4: CSE Researchers in Clinical AI



Learning Outcomes

LO1: Describe the landscape of ED NLP: tasks, datasets, modelling paradigms, evaluation in triage, diagnosis and disposition. (Modules 1 and 2)

LO2: Analyse the deployment constraints specific to the ED and the methods that currently address each. (Module 3)

LO3: Examine the desiderata an NLP system must satisfy to be credible in an ED, and apply them to critique a proposed design. (Modules 3 and 4)

LO4: Design a research project for ED in NLP by identifying a relevant research question, resources (dataset, simulation), methodology. (Modules 1 - 4)

Summary

Module 1: NLP for Emergency Department

Module 2: Typical Tasks and Methods

Module 3: Frontiers & Readiness

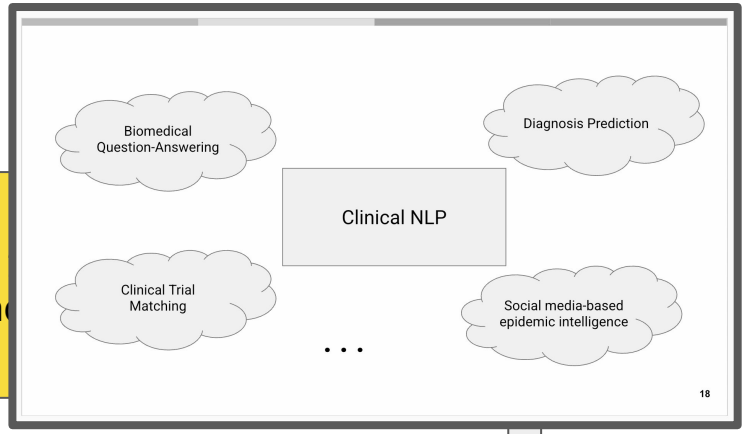
Module 4: CSE Researchers in Clinical AI



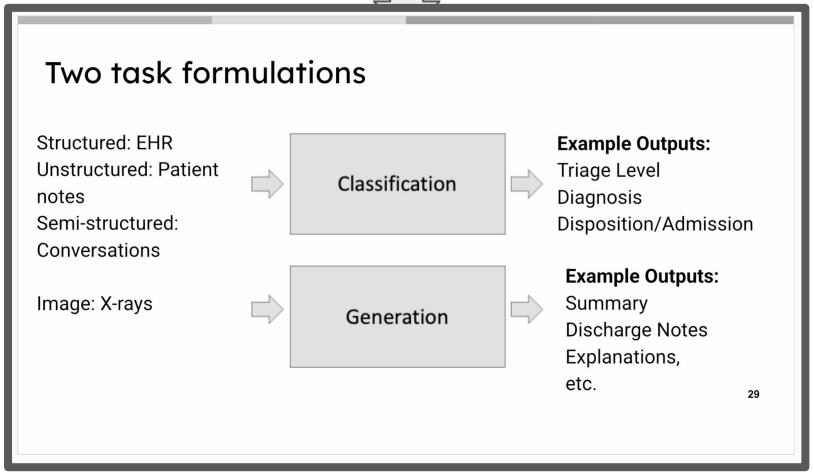
Summary

Module 1: NLP for Emergency Department

Module Tasks an



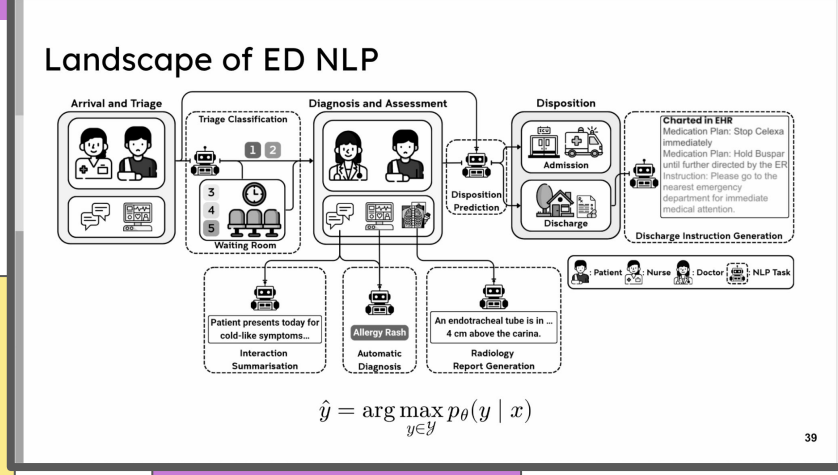
AJCAI 2026 - Post Survey



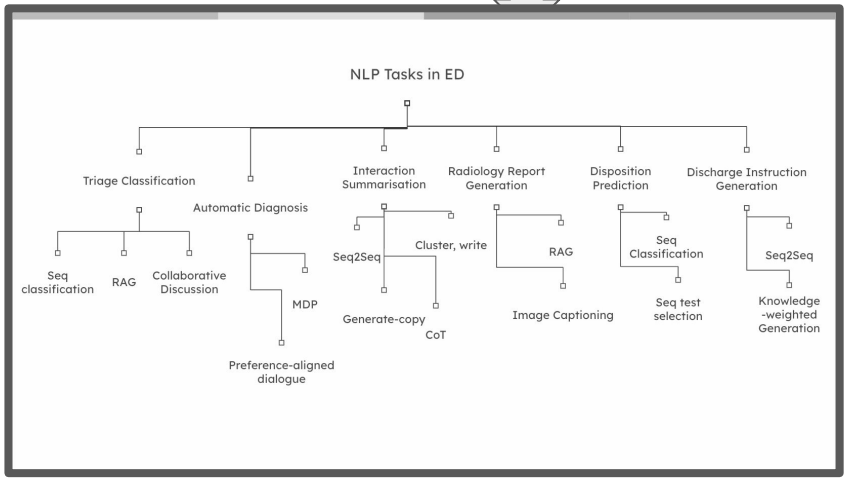
Summary

Module 1: NLP for Emergency Department

Module 2: Typical Tasks and Methods



AJCAI 2026 - Post Survey



Summary

Frontiers

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid x)$$

Generalisability	**params** was learned at one hospital, on one scale
Robustness	**input** was typed, clean and complete
Temporal Validity	**input** was the finished record
Evaluation	**output** was scored right or wrong
Deployment	**output** is where the paper stops

From Triage to Discharge: A Survey of NLP Tasks, Methods, and Open Challenges in the Emergency Department (**Srirag et al., 2026**)

154

Module 2: Typical asks and Methods

Module 3: Frontiers & Readiness

AJCAI 2026 - Post Survey

Readiness

Integration Readiness Levels

	Where it runs	The most it may do alone	
IRL6	Scaled across sites or languages	Supervised autonomy, periodic audits	rollout
IRL5	Limited rollout, monitored end to end	Acts by default; exceptions reviewed	
IRL4	Limited human-in-the-loop pilot	Suggests bounded steps; approval for high stakes	replay, then a pilot
IRL3	Shadow replay on real data, no live impact	It proposes; humans decide	
IRL2	Noisy, incomplete or adversarial simulation	Manual review, plus a safety gateway	simulation
IRL1	Static or scripted sandbox	Manual review before any commit	

LLM-Based Multi-Agent Systems for Clinical Workflows: A Survey of AI Hospitals (**Yao & Yu, 2026**)

226

Summary

NLP Research Venues

Workshops:

Workshop on Biomedical Language Processing (BioNLP) (25th in 2026!)

Clinical Natural Language Processing Workshop (5th in 2027; upcoming)

Special Interest Group (SIG) SIGCLINICAL for "engagement among members of the ACL, clinicians, and clinical researchers " https://sigclinical.org/

ACL Rolling Review

SCOPE OF SUBMISSIONS

All topics in Computational Linguistics / Natural Language Processing are welcome at ARR. At submission, papers must select a topic area to assist with AC and reviewer matching. The current set of areas are (in alphabetical order):

- Clinical and Biomedical Applications

238

Module
Tasks an

CSE Researchers in Clinical AI

Are you a CS PhD aspirant/CS researcher interested in research in NLP for ED (and broadly Clinical AI)?

The following slides are **reflective** and **subjective**.

233

AJCAI 2026 - Post Survey

Module 4: CSE Researchers in Clinical AI

Emergency Department

A Real-World Edge Case for Clinical NLP

Dipankar Srirag, Aditya Joshi
UNSW Sydney
d.srirag@unsw.edu.au | aditya.joshi@unsw.edu.au

AJCAI 2026 - Post Survey

UNSW-NLP