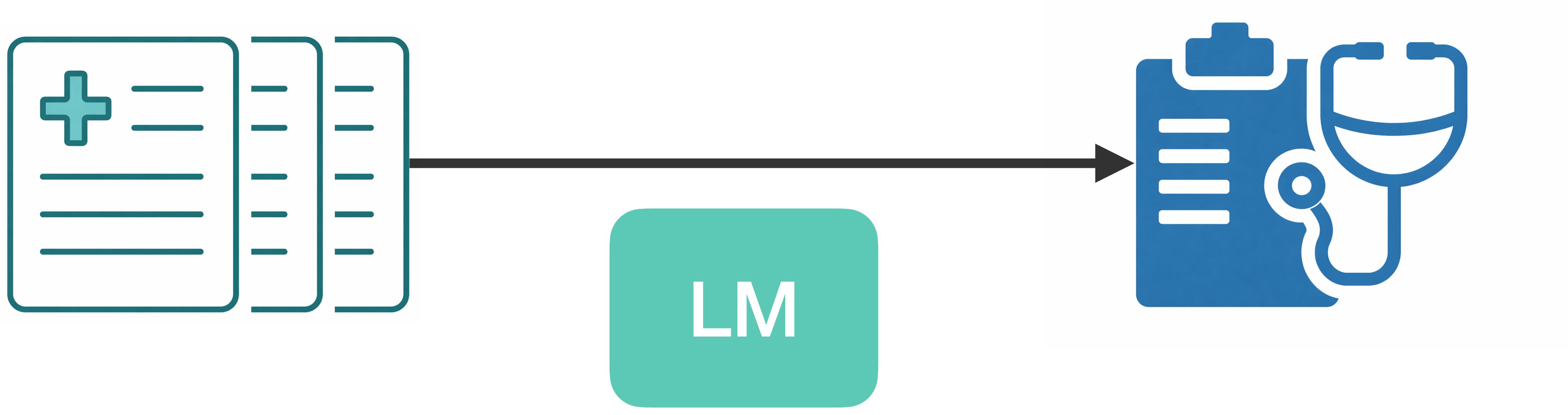
Clinical Information Extraction using Language Models, and its Role in Disease Prediction



Sitong Zhou

Advisors: Mari Ostendorf, Meliha Yetisgen

UW, Electrical and Computer Engineering

Why we need unstructured EHRs

E*lectronic*
H*ealth*
R*ecords*

complementary

Unstructured EHRs

free-texts

rich with nuance

e.g clinical reports,
radiology reports

Structured EHRs

tabular data (convenient)

incomplete & missing details

e.g ICD diagnosis codes,
demographic forms, labs

Why we need unstructured EHRs

Structured EHRs

incomplete & missing details

structured: ICD codes

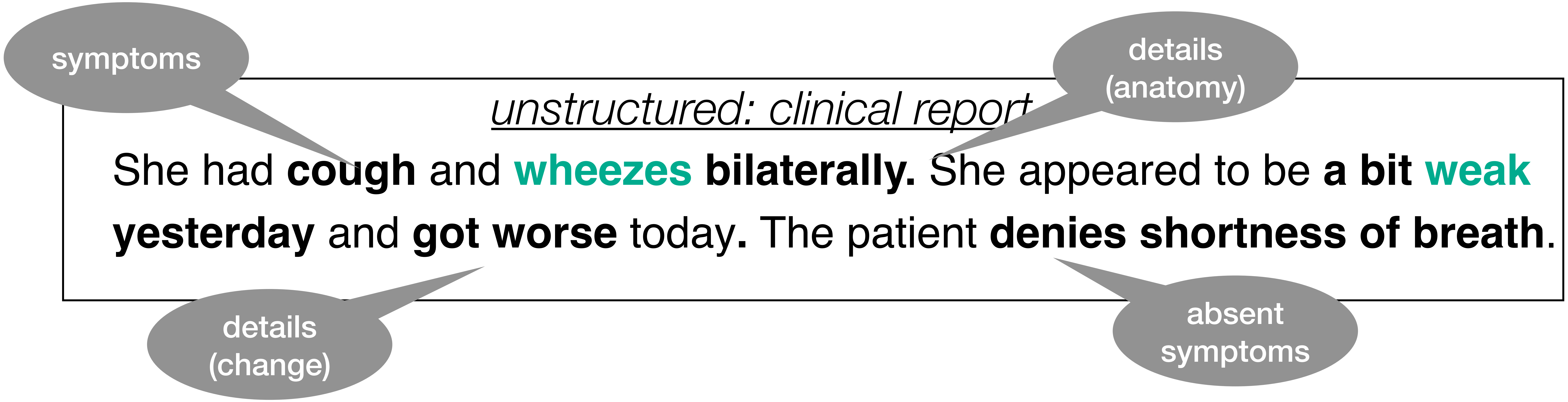
She had **cough** and **wheezes bilaterally.** She appeared to be **a bit weak yesterday** and **got wo** R06.2 day. The patient **denies shortness of b** R53.1 .

Why we need unstructured EHRs

complementary

Unstructured EHRs → ***Structured EHRs***

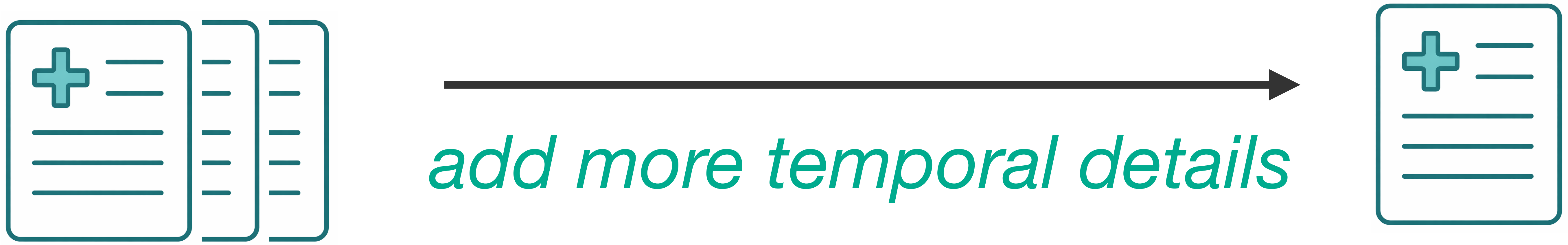
rich with nuance



<u>*unstructured: clinical report*</u>

She had **cough** and **wheezes** **bilaterally.** She appeared to be **a bit** **weak** **yesterday** and **got worse** today**.** The patient **denies shortness of breath**.

Why we need longitudinal unstructured EHRs



add more temporal details

Longitudinal reports

temporal change details:
How long has it been stable, how fast it grows, details in the past …

Single report

short phrases for changes:
stable, size increased, resolving

Why we need longitudinal unstructured EHRs

Single report

short phrases for changes:

Chest CT
stable subsolid pulmonary nodule …

Why we need longitudinal unstructured EHRs

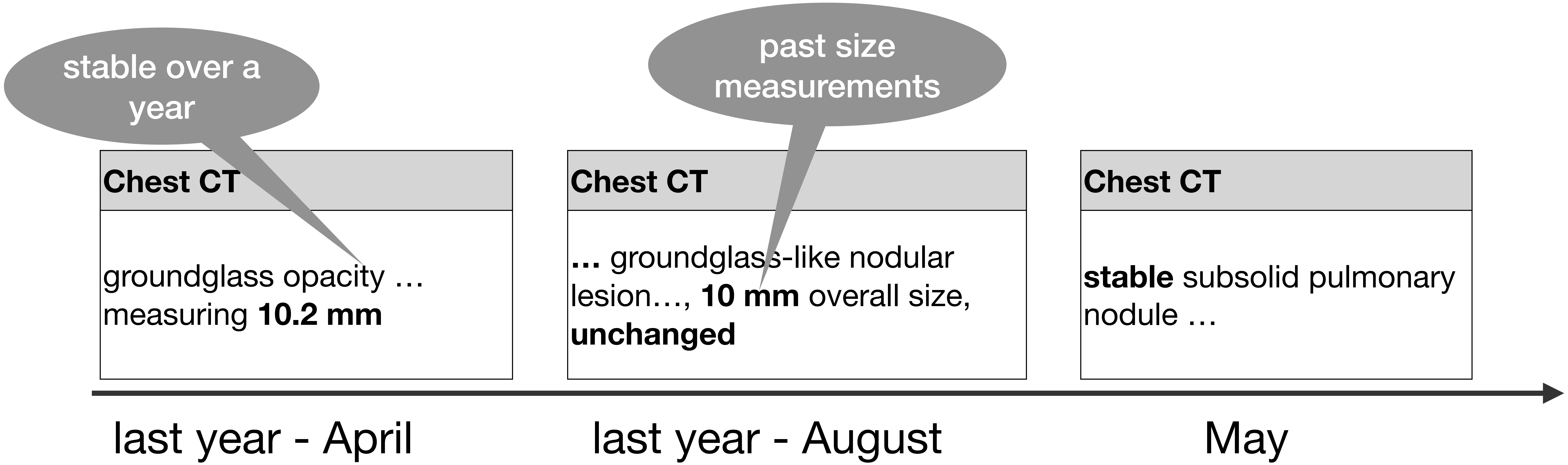
add temporal change details

Longitudinal reports

temporal change details

Single report

short phrases for changes



stable over a year

past size measurements

Chest CT

groundglass opacity … measuring **10.2 mm**

last year - April

Chest CT

… groundglass-like nodular lesion…, **10 mm** overall size, **unchanged**

last year - August

Chest CT

stable subsolid pulmonary nodule …

May

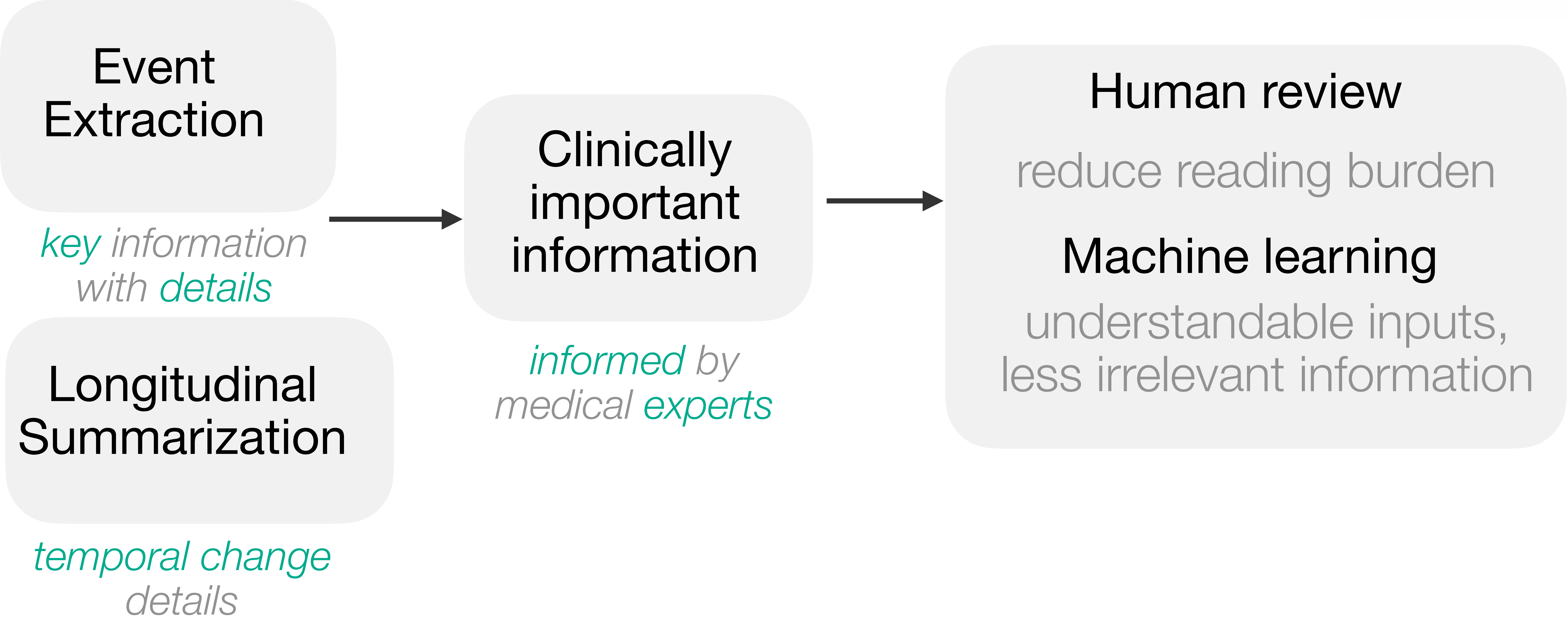
Why IE from unstructured EHRs

inconvenient

***I**nformation*
***E**xtraction*

IE *(two tasks)*

Decision support



Event Extraction

key information with details

Longitudinal Summarization

temporal change details

Clinically important information

informed by medical experts

Human review

reduce reading burden

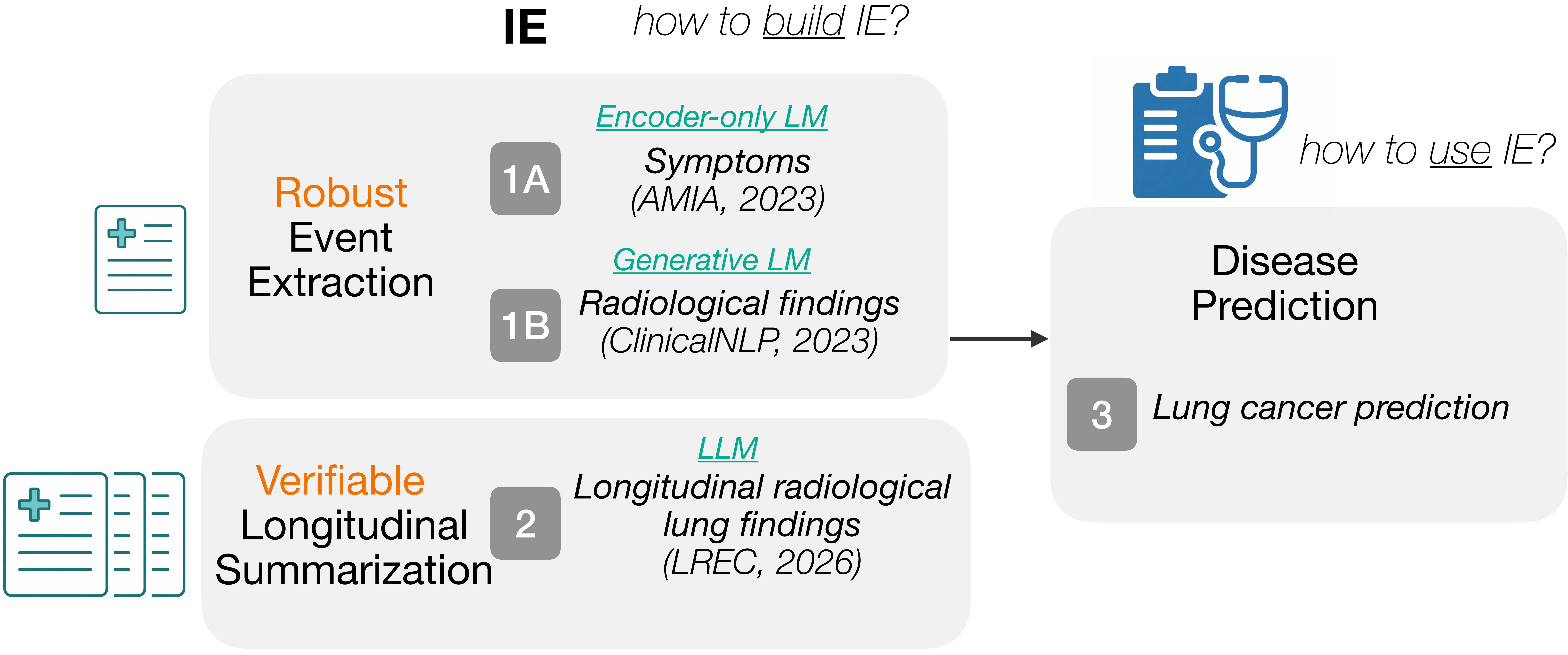
Machine learning

understandable inputs, less irrelevant information

Challenges of building clinical IE using LMs

IE	LM	Challenge	Clinical Usage
Robust Event Extraction	**encoder-only LM** BERT (2018) BioClinical BERT (2019) **generative LM** T5 (2019) ClinicalT5 (2022)	Domain generalization ***supervised learning*** *on limited expert annotations*	Fairness
Verifiable Longitudinal Summarization	**LLM** Llama 3 (2024) GPT 4o (2024)	Long context understanding, factuality	Trust
Efficient generative methods		Costs / latency	Scalability

Overview: clinical IE using LMs for Disease Prediction



IE *how to build IE?*

Robust
Event
Extraction

Encoder-only LM

1A *Symptoms*
(AMIA, 2023)

Generative LM

1B *Radiological findings*
(ClinicalNLP, 2023)

Verifiable
Longitudinal
Summarization

LLM

2 *Longitudinal radiological lung findings*
(LREC, 2026)

how to use IE?

Disease
Prediction

3 *Lung cancer prediction*

Efficient generative methods

1A

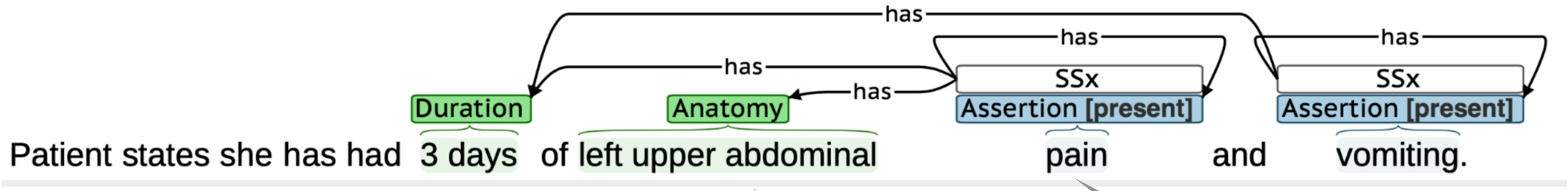
Robust event extraction on symptoms with encoder-only LM

Generalizing through Forgetting -- Domain Generalization for Symptom Event Extraction in Clinical Notes

Zhou, Lybarger, Yetisgen, Ostendorf, AMIA 2023

Symptom event extraction

Symptom event schema *(Lybarger et al, 2021)*



Patient states she has had 3 days of left upper abdominal pain and vomiting.

Arguments (details)

Trigger (key)

What is an event?

highlight in raw texts

What is a symptom event?

Arguments:

Assertion

present, absent,,

Anatomy,

left upper abdominal

Duration,

3 days

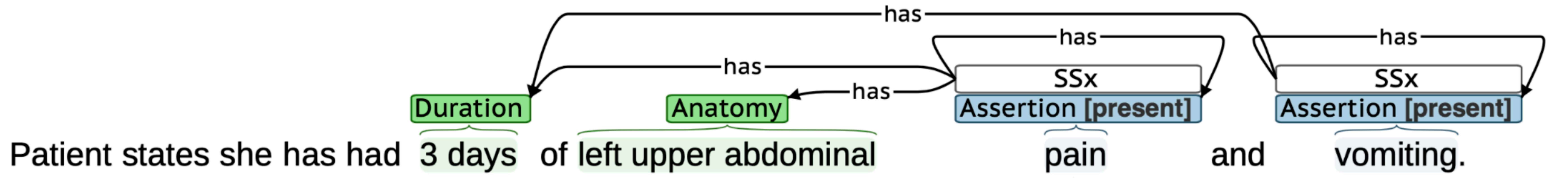
…

Trigger (SSx):

Symptom (e.g., pain, vomiting)

Symptom event extraction

Symptom event schema *(Lybarger et al, 2021)*



What is an event?

What is a symptom event?

Arguments (details)

Trigger (key)

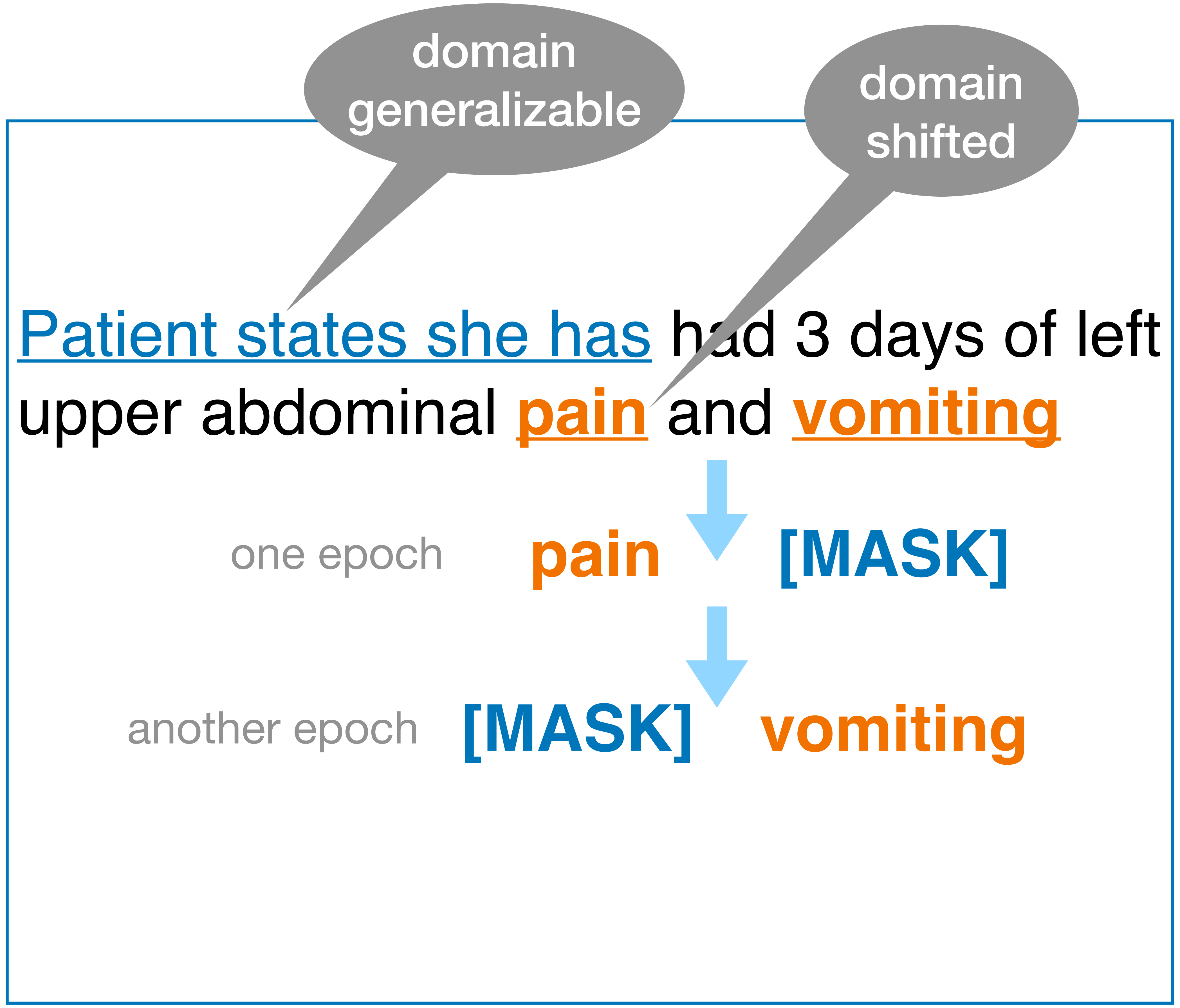
Shift with domains

Arguments:
Assertion
present, absent,,
Anatomy,
left upper abdominal
Duration,
3 days
…

Trigger (SSx):
Symptom (e.g., pain, vomiting)

hospital, specialty, time, etc

Dynamically masking frequent trigger phrases



domain generalizable

domain shifted

Patient states she has had 3 days of left upper abdominal **pain** and **vomiting**

one epoch **pain** **[MASK]**

another epoch **[MASK]** **vomiting**

during training

Problem (Memorization):
Learn shortcuts about frequent symptoms

Solution (Forgetting):

Dynamic masking on frequent symptoms

A data manipulation strategy

no extra data, only removing
unnecessary information in training data

clinic *patients*
(during pandemic) → ***cancer*** *patients*

Dynamic masking helps

additional benefits beyond pretraining

Entity (F1)	***Bio+Clinical BERT***	+ Mask	**Further Adaptive Pretraininig**	+ Mask
Trigger	77.2	**78.5****	78.6	**79.9***
Assertion	72.2	**73.1***	74	**74.9**
Anatomy	55.3	**56.7**	57.2	**59.4**
Characteristics	22.5	**22.8**	**26.8**	25.8
Duration	**46.5**	44.1	46.3	**48.4**

**(- -) and *(-) indicate significant gain(loss) p-value < 0.01 and <0.05

Architecture: *SpERT*

Adaptive pretraining:
on UWM notes

Robust
Event
Extraction

1A
Encoder-only LM
Symptoms
(AMIA, 2023)

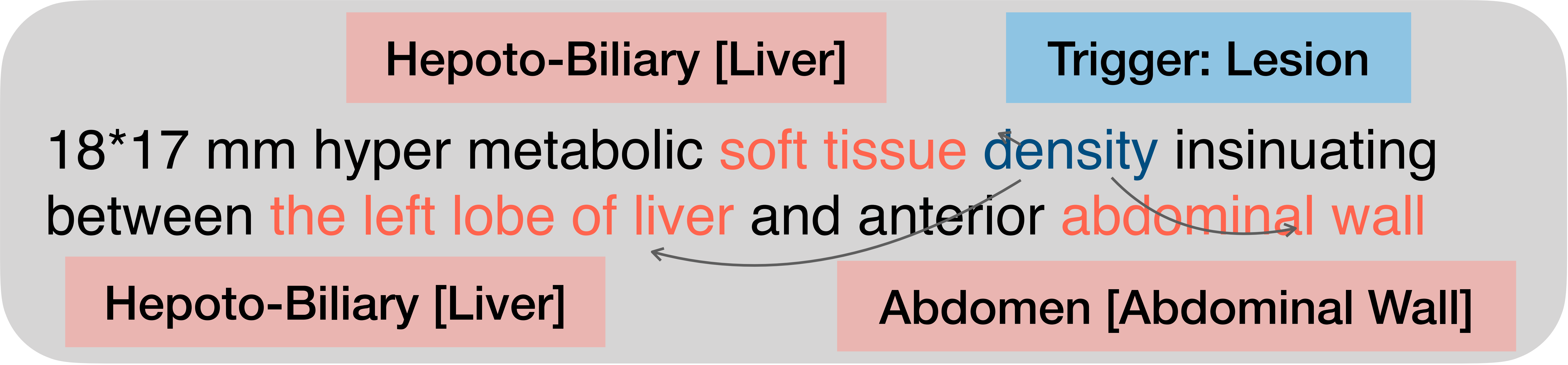
1B
Generative LM
Radiological findings
(ClinicalNLP, 2023)

1B Robust event extraction on radiological findings with generative LM

Building blocks for complex tasks: Robust generative event extraction for radiology reports under domain shifts

Zhou, Yetisgen, Ostendorf, ClinicalNLP 2023

Radiological event schema



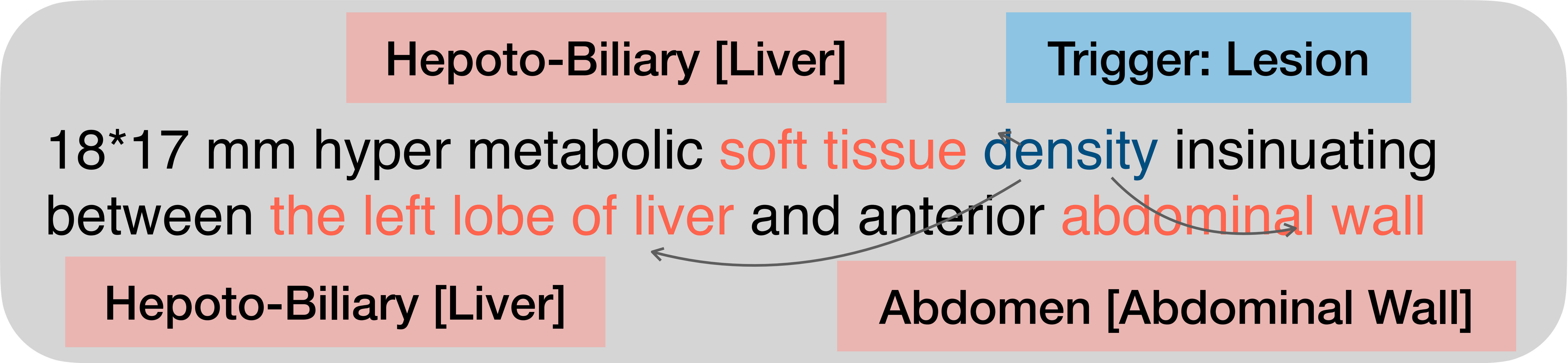
18*17 mm hyper metabolic soft tissue density insinuating between the left lobe of liver and anterior abdominal wall

Subtasks

trigger & anatomy
(span and type)

(help filter in later use)

Radiological event schema



Hepoto-Biliary [Liver]

Trigger: Lesion

18*17 mm hyper metabolic soft tissue density insinuating between the left lobe of liver and anterior abdominal wall

Hepoto-Biliary [Liver]

Abdomen [Abdominal Wall]

Subtasks

trigger span
trigger type
anatomy span
anatomy type

1-step inference

– SFT with generative models

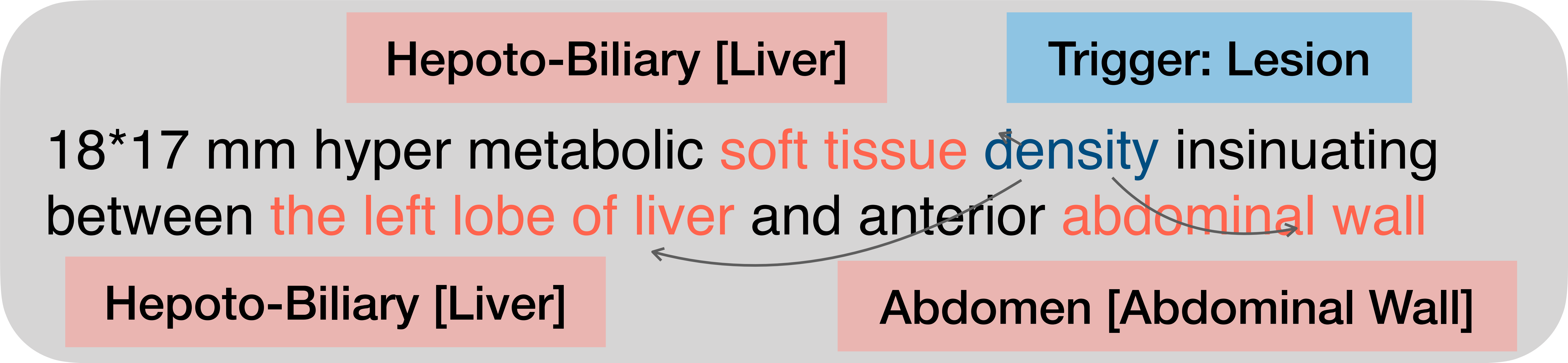
Q: What are medical findings in this sentence? What anatomy they occur in? which anatomy category they belong to among listed options?

A: density [Lesion]
soft tissue [Hepato-Bilary | Liver]
left lobe of liver [Hepato-Bilary | Liver]
abdominal wall ~~[Hepato-Bilary | Liver]~~

if training data tend to have all anatomy types the same

looks like hallucination in LLM

Radiological event schema



Hepoto-Biliary [Liver]

Trigger: Lesion

18*17 mm hyper metabolic soft tissue density insinuating between the left lobe of liver and anterior abdominal wall

Hepoto-Biliary [Liver]

Abdomen [Abdominal Wall]

Subtasks

trigger span
trigger type
anatomy span
anatomy type

1-step inference

– SFT with generative models

Q: What are medical findings in this sentence? What anatomy they occur in? which anatomy category they belong to among listed options?

A: density [Lesion]
soft tissue [Hepato-Bilary | Liver]
left lobe of liver [Hepato-Bilary | Liver]
abdominal wall ~~[Hepato-Bilary | Liver]~~

if training data tend to have all anatomy types the same

higher latency

Better off with separate runs
(compositionality gap)

Reasoning helps in LLM 1-step *(Press et al, 2022)*

Building blocks to help reasoning

SFT with T5 models (much smaller than LLM)

1-step inference

Q: What are medical findings in this sentence? What anatomy they occur in? which anatomy category they belong to among listed options?

A: density [Lesion]
soft tissue [Hepato-Bilary | Liver]
left lobe of liver [Hepato-Bilary | Liver]
abdominal wall [Abdomen | Abdominal Wall]

with building blocks

train on complex & simpler tasks

(stack of blocks) *(single block)*

domain shifted

domain generalizable

. . .

A: state: anatomy span
answer: soft tissue, left lobe of liver, abdominal wall

A: state: anatomy type
answer: soft tissue [Hepato-Bilary | Liver]

A: state: anatomy type
answer: left lobe of liver [Hepato-Bilary | Liver]

A: state: anatomy type
answer: abdominal wall [Abdomen | Abdominal Wall]

one subtask
one block
(encapsulate)

Reducing the 1-step compositionality gap

MRI -> PET (***local*** *exam ->* ***global*** *exam)*, **T5-base models**

	anatomy types	multi-step	1-step	1-step w/ building blocks
e.g: Respiratory	- Parent level	48.6	44.9	**48.3**
e.g: Lung	- Child level	44.6	41.2	**44.8**

One-step inference suffers from compositionality gap

- 8%

+8%

Multitask learning with blocks recovers the gap

Learning on simpler tasks enable small LMs to reason for complex tasks

Conclusion: robust event extraction

Robust Event Extraction

1A *Encoder-only LM*
Symptoms
(AMIA, 2023)

1B *Generative LM*
Radiological findings
(ClinicalNLP, 2023)

Realistic domain shift has multiple reasons:

Symptom distribution, subtask dependencies

Removing unnecessary information (shifted across domains) from training data helps generalization

symptom phrases, other subtasks

Shorter training data helps generalization when generating longer outputs

IE

Robust
Event
Extraction

1A
Encoder-only LM
Symptoms
(AMIA, 2023)

1B
Generative LM
Radiological findings
(ClinicalNLP, 2023)

Verifiable
Longitudinal
Summarization

2
LLM
Longitudinal radiological lung findings
(under review, 2025)

2

Verifiable longitudinal summarization on radiological lung findings with LLMs

RadTimeline: Timeline Summarization for Longitudinal Radiological Lung Findings

Zhou, Yetisgen, Ostendorf, LREC 2026

Why we need longitudinal radiology reports

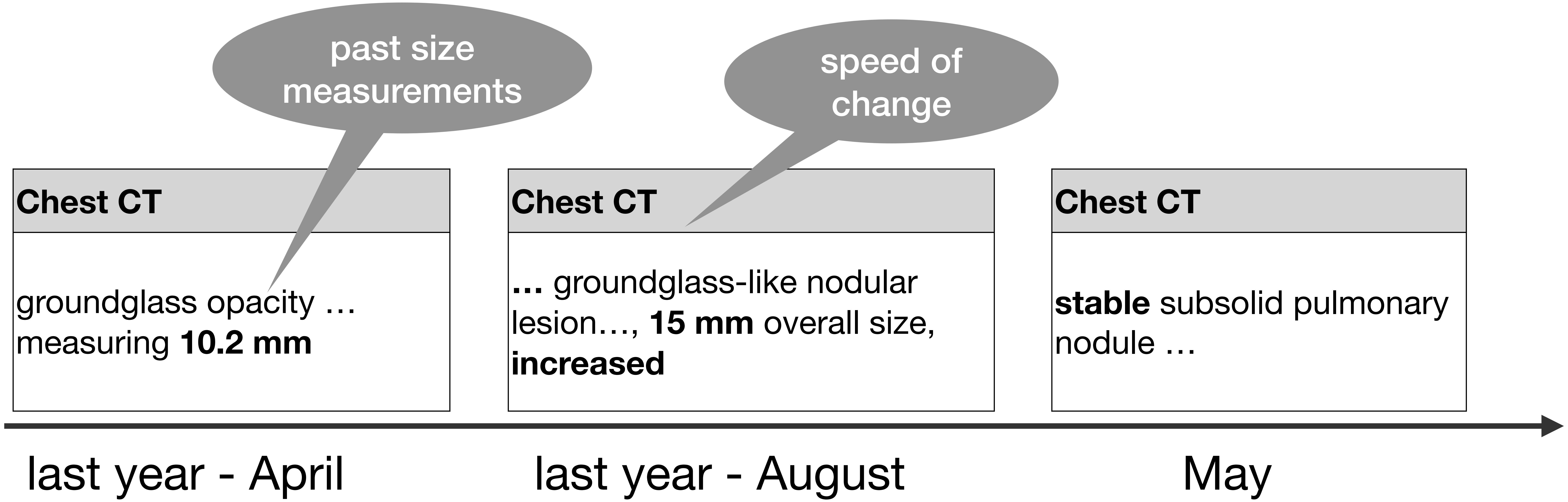
Longitudinal reports

temporal change details

Single report

short phrases for changes

lung finding change details inform lung cancer risk.



past size measurements

speed of change

Chest CT

groundglass opacity … measuring **10.2 mm**

Chest CT

… groundglass-like nodular lesion…, **15 mm** overall size, **increased**

Chest CT

stable subsolid pulmonary nodule …

last year - April

last year - August

May

Reading Challenges: time-consuming & cognitive overload

limited time for clinicians

synthesizing distributed facts

...

Chest CT

… predominantly **groundglass opacity** in the posterior segment of the **right lower lobe** … measuring as much as **2.0 cm**; alveolar **groundglass opacity** in the posterior segment of the **right upper lobe** .. measuring **10.2 mm** …

last year - April

Chest CT

… **Right upper lobe**, indistinct mostly **groundglass-like nodular lesion**…, **15 mm** overall size, increased**; Right lower lobe**, posterior basal segment, stellate-like **groundglass opacity**, **2 cm**, unchanged **…**

last year - August

Chest CT

… **stable subsolid pulmonary nodule** in the **right upper lobe; stable subsolid pulmonary nodule** in the **right lower lobe; new 12 mm peripheral subsolid opacity** in the **right upper lobe** anterior laterally, …

May

...

Automatic Longitudinal Summarization: what format?

Direct summarization
as free-texts

(Chien et al, 2024)

personalized, but complicates fact-checking

Verifiable structured summarization
group timestamped facts into fixed topics

(Verma et al, 2025)

Verifiable facts, but fixed topics lack granularity.

Radiological finding groups change with patients

RadTimeline (ours): timeline summarization
group timestamped facts with personalized group names

(Zhou et al, 2025)

Verifiable facts & personalized related finding groups

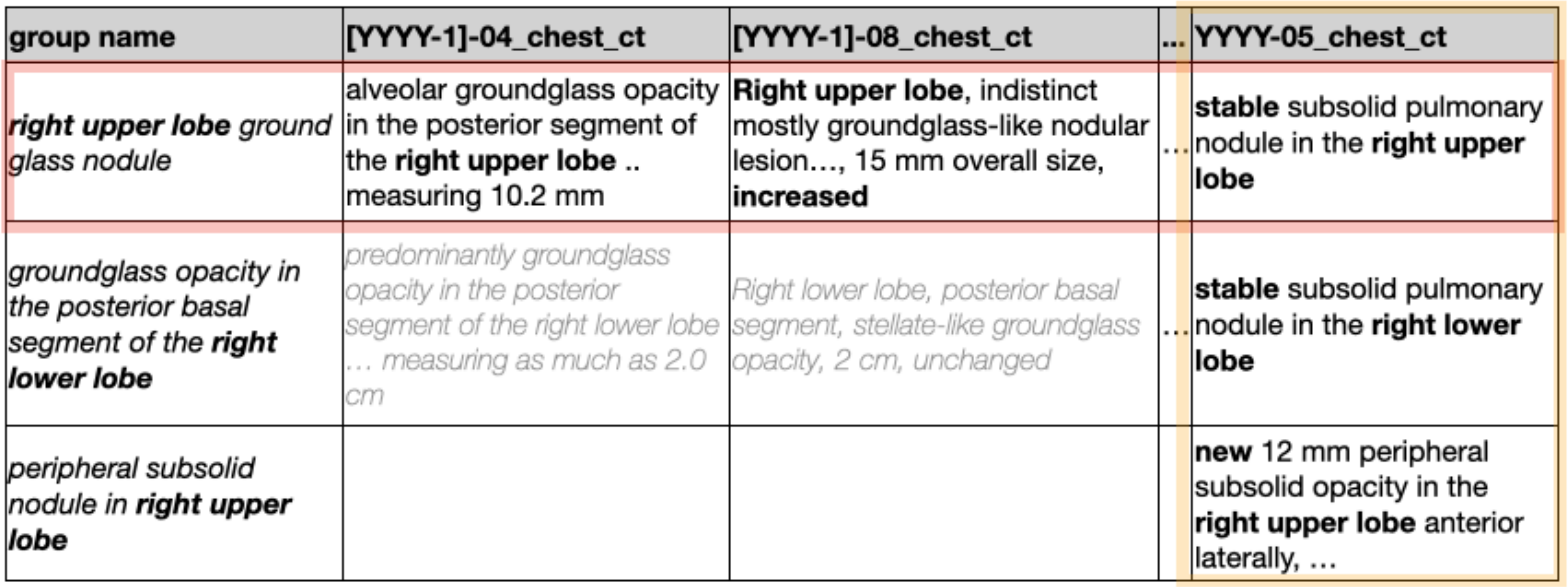
RadTimeline: contributions

- Structured summarization format
- RadTimeline evaluation dataset
- 3-step generation approach (no training required)

RadTimeline: structured summarization format

Columns: verifiable **findings** from each report — *header : report title*

Rows: personalized groups for related findings — *header : **group name***



group name	**[YYYY-1]-04_chest_ct**	**[YYYY-1]-08_chest_ct**	**...**	**YYYY-05_chest_ct**
right upper lobe *ground glass nodule*	alveolar groundglass opacity in the posterior segment of the **right upper lobe** .. measuring 10.2 mm	**Right upper lobe**, indistinct mostly groundglass-like nodular lesion…, 15 mm overall size, **increased**	…	**stable** subsolid pulmonary nodule in the **right upper lobe**
*groundglass opacity in the posterior basal segment of the **right lower lobe***	*predominantly groundglass opacity in the posterior segment of the right lower lobe … measuring as much as 2.0 cm*	*Right lower lobe, posterior basal segment, stellate-like groundglass opacity, 2 cm, unchanged*	…	**stable** subsolid pulmonary nodule in the **right lower lobe**
*peripheral subsolid nodule in **right upper lobe***				**new** 12 mm peripheral subsolid opacity in the **right upper lobe** anterior laterally, …

RadTimeline: structured summarization format

Columns: verifiable **findings** from each report — *header : report title*

each cell is a piece of fact containing critical details (size, trend, ..)

group name	[YYYY-1]-04_chest_ct	[YYYY-1]-08_chest_ct	…	**YYYY-05_chest_ct**
right upper lobe ground glass nodule	alveolar groundglass opacity in the posterior segment of the right upper lobe .. measuring 10.2 mm	Right upper lobe, indistinct mostly groundglass-like nodular lesion…, 15 mm overall size, increased	…	**stable** subsolid pulmonary nodule in the **right upper lobe**
groundglass opacity in the posterior basal segment of the right lower lobe	*predominantly groundglass opacity in the posterior segment of the right lower lobe … measuring as much as 2.0 cm*	*Right lower lobe posterior basal segment, stellate-like groundglass opacity, 2 cm, unchanged*	…	**stable** subsolid pulmonary nodule in the **right lower lobe**
peripheral subsolid nodule in right upper lobe				**new** 12 mm peripheral subsolid opacity in the **right upper lobe** anterior laterally, …

RadTimeline: structured summarization format

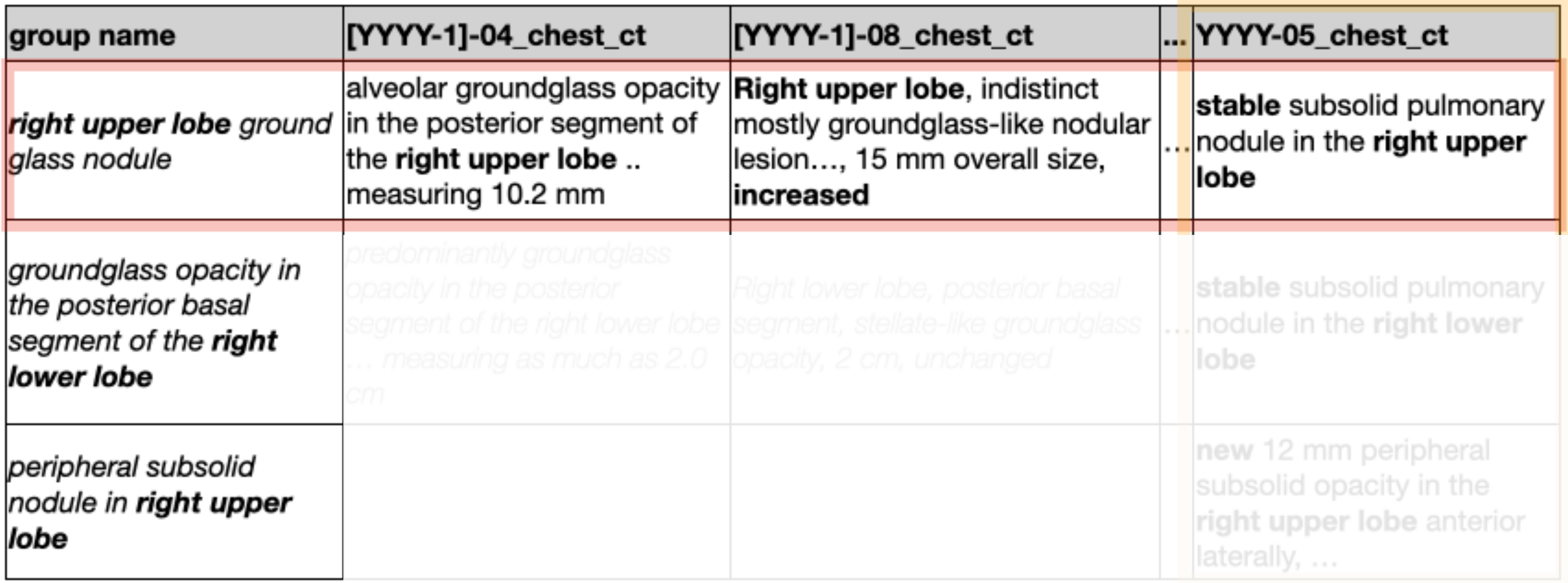
Rows: personalized groups for related findings

— *header : **group name***

*Each finding has a **group assignment***

each group name is distinguishable from other groups

avoid too generic / specific



group name	[YYYY-1]-04_chest_ct	[YYYY-1]-08_chest_ct	...	YYYY-05_chest_ct
***right upper lobe** ground glass nodule*	alveolar groundglass opacity in the posterior segment of the **right upper lobe** .. measuring 10.2 mm	**Right upper lobe**, indistinct mostly groundglass-like nodular lesion…, 15 mm overall size, **increased**	…	**stable** subsolid pulmonary nodule in the **right upper lobe**
*groundglass opacity in the posterior basal segment of the **right lower lobe***	*predominantly groundglass opacity in the posterior segment of the right lower lobe … measuring as much as 2.0 cm*	*Right lower lobe, posterior basal segment, stellate-like groundglass opacity, 2 cm, unchanged*	…	**stable** subsolid pulmonary nodule in the **right lower lobe**
*peripheral subsolid nodule in **right upper lobe***				**new** 12 mm peripheral subsolid opacity in the **right upper lobe** anterior laterally, …

RadTimeline: gold standard timelines for evaluation

EHR raw data: radiology reports from a lung cancer prediction study

10 patients (6 with future lung cancer) with 65 reports in a 5-year history window

Focus on lung findings from chest-related CT/XRay reports

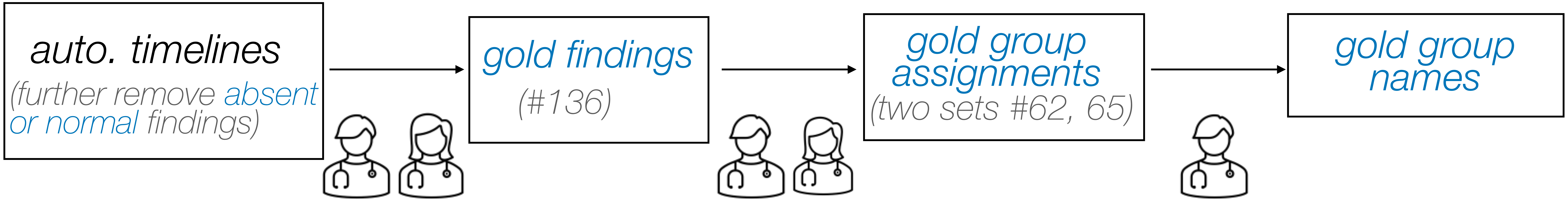
RadTimeline: gold standard timelines for evaluation

EHR raw data: radiology reports from a lung cancer prediction study

10 patients (6 with future lung cancer) with 65 reports in a 5-year history window

Focus on lung findings from chest-related CT/XRay reports

Timeline annotations: *Manually corrected autogenerated timelines using experts' feedback*



auto. timelines
(further remove absent or normal findings)

gold findings
(#136)

gold group assignments
(two sets #62, 65)

gold group names

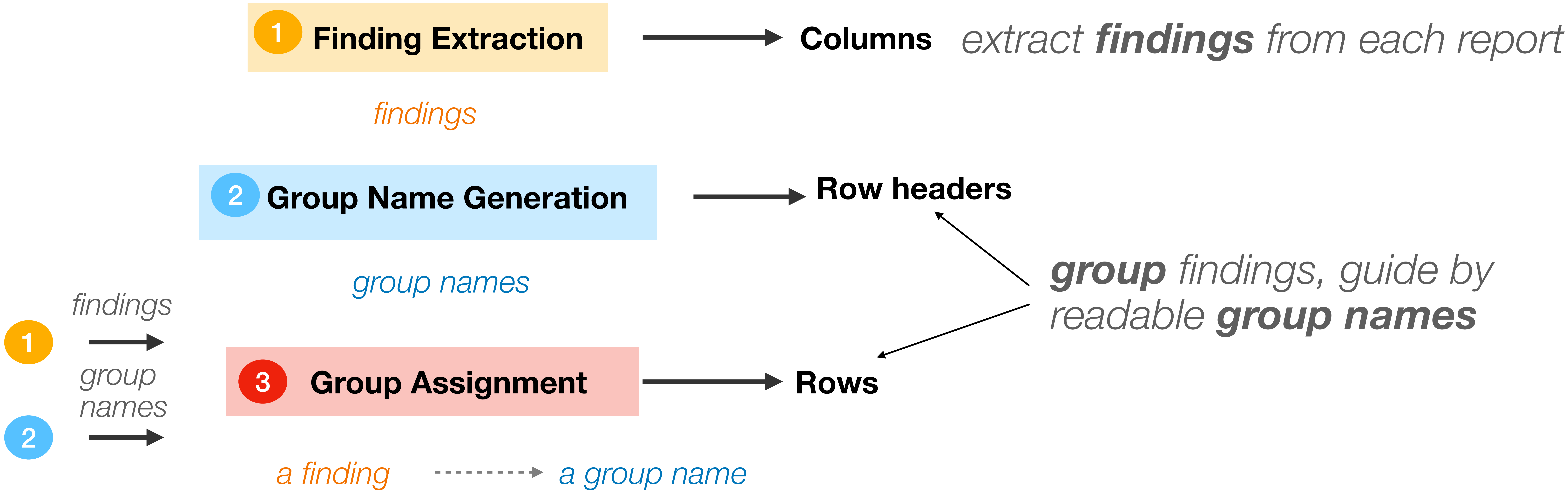
Identify missing and irrelevant, incorrect findings

Correct grouping and group names

Revises group names for each set

Findings are corrected based on their feedback

Three-step timeline generation



1 **Finding Extraction** → **Columns** *extract **findings** from each report*

findings

2 **Group Name Generation** → **Row headers**

group names

findings

group names

3 **Group Assignment** → **Rows**

a finding → *a group name*

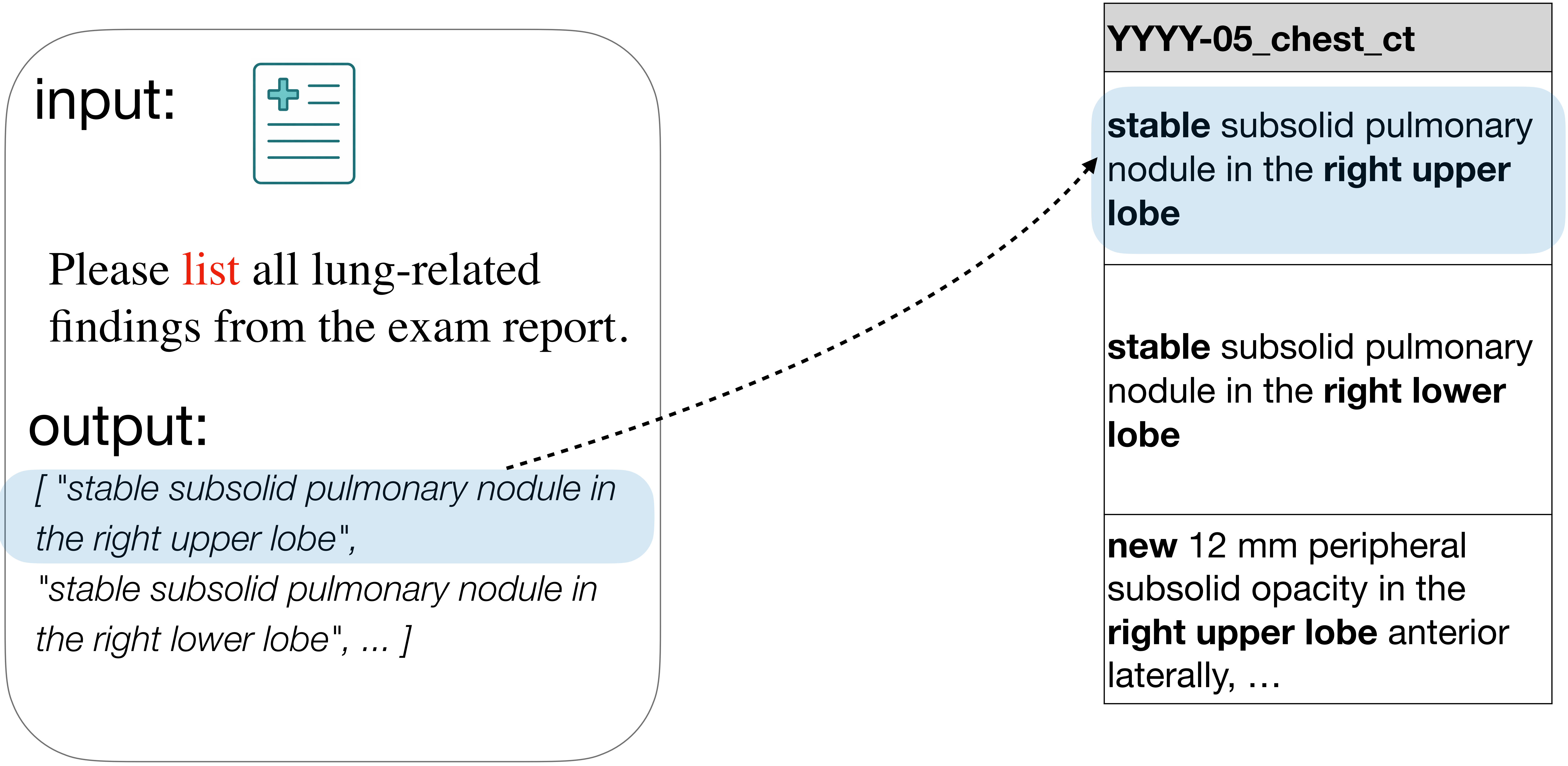
*group findings, guide by readable **group names***

Step 1: finding extraction

① **Finding Extraction** → **Columns**

LLM (single report)

*extract **findings** from each report*



input:

Please list all lung-related findings from the exam report.

output:

["stable subsolid pulmonary nodule in the right upper lobe",
"stable subsolid pulmonary nodule in the right lower lobe", ...]

YYYY-05_chest_ct
stable subsolid pulmonary nodule in the **right upper lobe**
stable subsolid pulmonary nodule in the **right lower lobe**
new 12 mm peripheral subsolid opacity in the **right upper lobe** anterior laterally, …

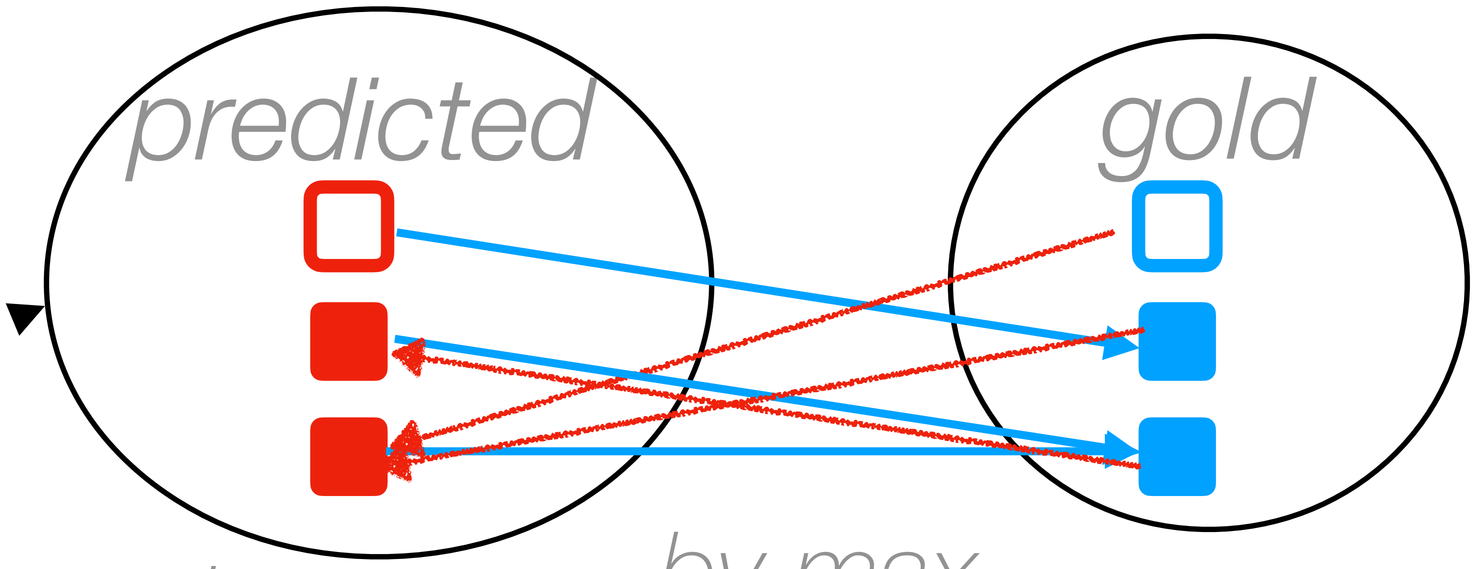
Finding extraction results

Good completeness given single-report inputs for both LLMs

High unnecessary information for Llama 3.1

includes wrong, irrelevant, redundant, unfiltered absent/normal

version	ROUGE-L ↑	% Missing ↓	% Unnecessary ↓
Llama 3.1 8B Instruct	70	**3**	32
GPT-4o	**75**	6	**15**



filter absent/normal predictions automatically

by max ROUGE-L

🟥☐ ***ROUGE-L:*** *avg. over predictions*

☐ ***% Missing:*** *unmatched gold*

☐ ***% Unnec. :*** *unmatched predicted*

Step 2: group name generation

② **Group Name Generation** → **Row headers**

LLM

input: *full contexts / finding lists*

Please group the lung-related findings recorded over time and list the group names.

output:

List of group names:
Right upper lobe ground glass nodule,
Right lower lobe ground glass opacity with central nodular component, ...

group name
***right upper lobe** ground glass nodule*
*groundglass opacity in the posterior basal segment of the **right lower lobe***
*peripheral subsolid nodule in **right upper lobe***

Group name generation results

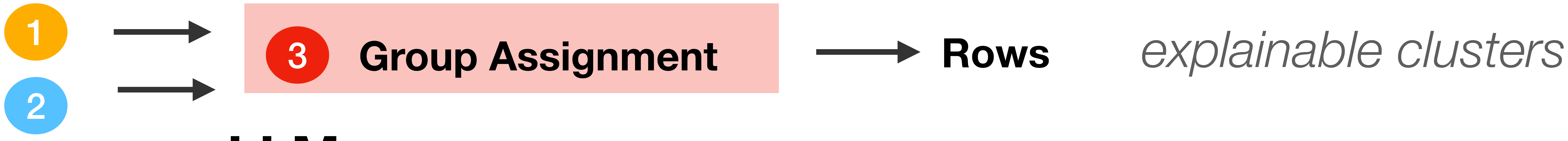
Scoring similar to findings, but average the scores against the two groupings.

GPT-4o significantly improves over Llama 3.1 in ROUGE-L and unnecessary names, but behind human agreement

version	ROUGE-L ↑	% Missing ↓	% Unnec. ↓
Llama 3.1 (Full)	38	**25**	60
GPT-4o (Full)	**48**	30	**37**
GPT-4o (Fdg.)	52	13	**46**
Human	**87**	9	**9**

Using the full context vs. just finding list hurts name quality but reduces unnecessary names

Step 3: group assignment



LLM

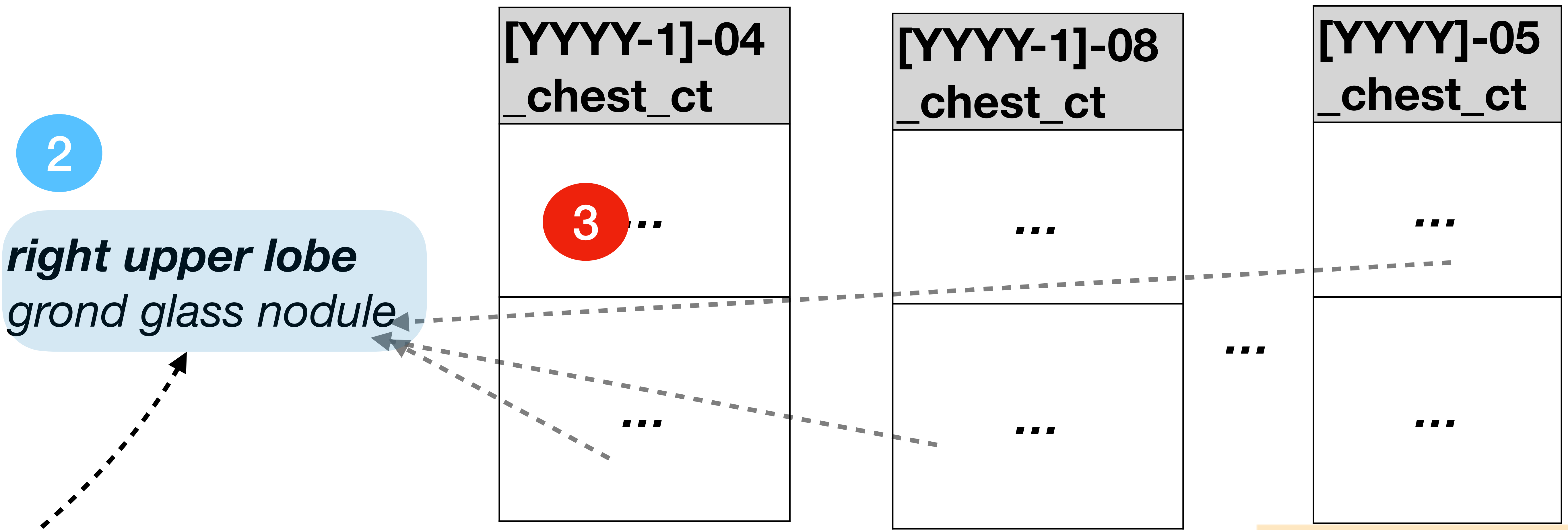
input: *single finding*
& group names

Please find the group best describing this finding given group labels.

List of group names: Right upper lobe ground glass nodule, ….

output:

right upper lobe
ground glass nodule



group name	[YYYY-1]-04_chest_ct	[YYYY-1]-08_chest_ct	...	YYYY-05_chest_ct
right upper lobe *ground glass nodule*	alveolar groundglass opacity in the posterior segment of the **right upper lobe** .. measuring 10.2 mm	**Right upper lobe**, indistinct mostly groundglass-like nodular lesion…, 15 mm overall size, **increased**	...	**stable** subsolid pulmonary nodule in the **right upper lobe**

Group assignment results

report CoNLL F1 for clustering on **gold findings**

Few shot only helps the weaker model

group names	ZeroShot (Llama 3.1)	FewShot (Llama 3.1)	ZeroShot (GPT-4o)	FewShot (GPT-4o)	Embedder
Llama-3.1	66	74	77	73	70
GPT-4o	**75**	**79**	**82**	**80**	69
Gold					**83**

Group name quality matters

GPT-4o is close to human (82)

With gold group names, embedder can be good

much cheaper

Embedder: E5-Mistral 7B-instruct

Conclusion: verifiable & personalized longitudinal summarization

LLM can generate lung findings

good completeness, high irrelevant information

LLM can group gold findings

near human performance
cheaper models may be effective with better group names

Extensible framework for tracking problems

symptom trajectory, treatment trajectory

Conclusion: verifiable longitudinal summarization

LLM can generate lung findings

good completeness, concerning irrelevant information

LLM can group gold findings

near human performance
smaller models can be improved with better intermediate group names

Extensible framework for tracking problems

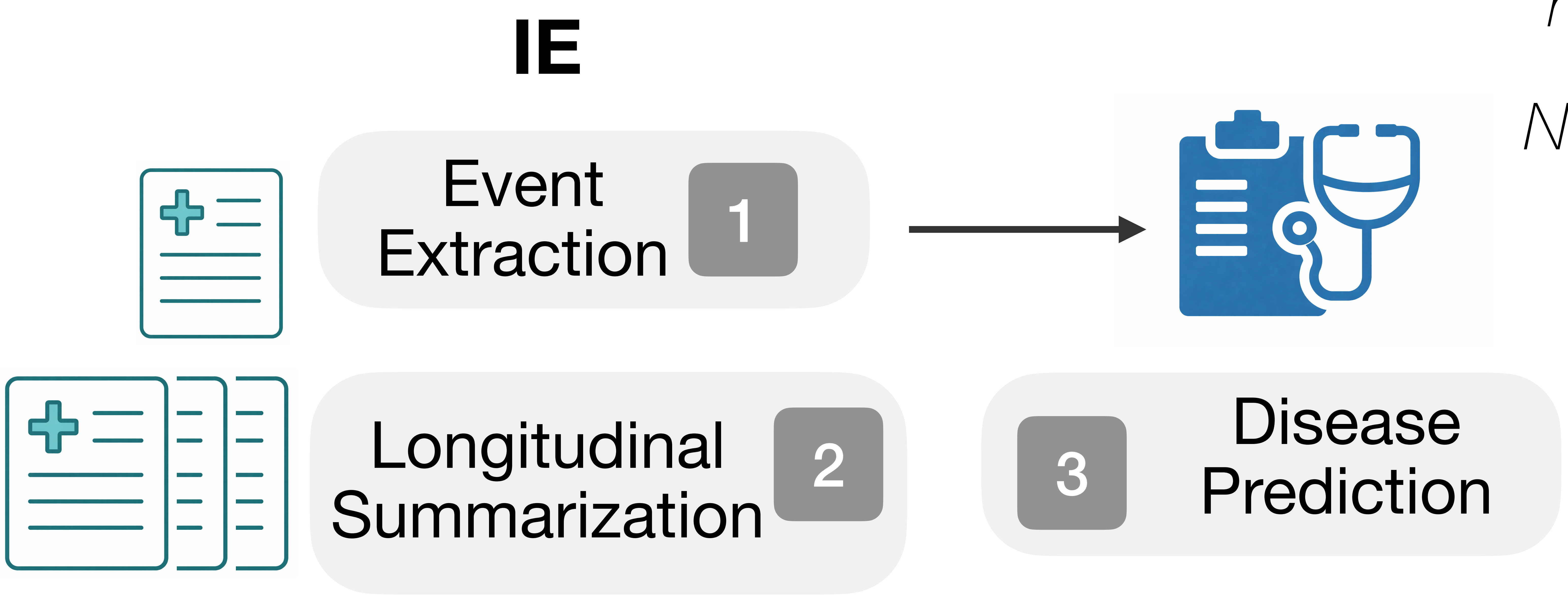
symptom trajectory, treatment trajectory, etc

Verifiable
Longitudinal
Summarization

2

LLM
Longitudinal radiological lung findings
(LREC, 2026)

3 Lung cancer prediction with IE



IE

Event Extraction 1

Longitudinal Summarization 2

3 Disease Prediction

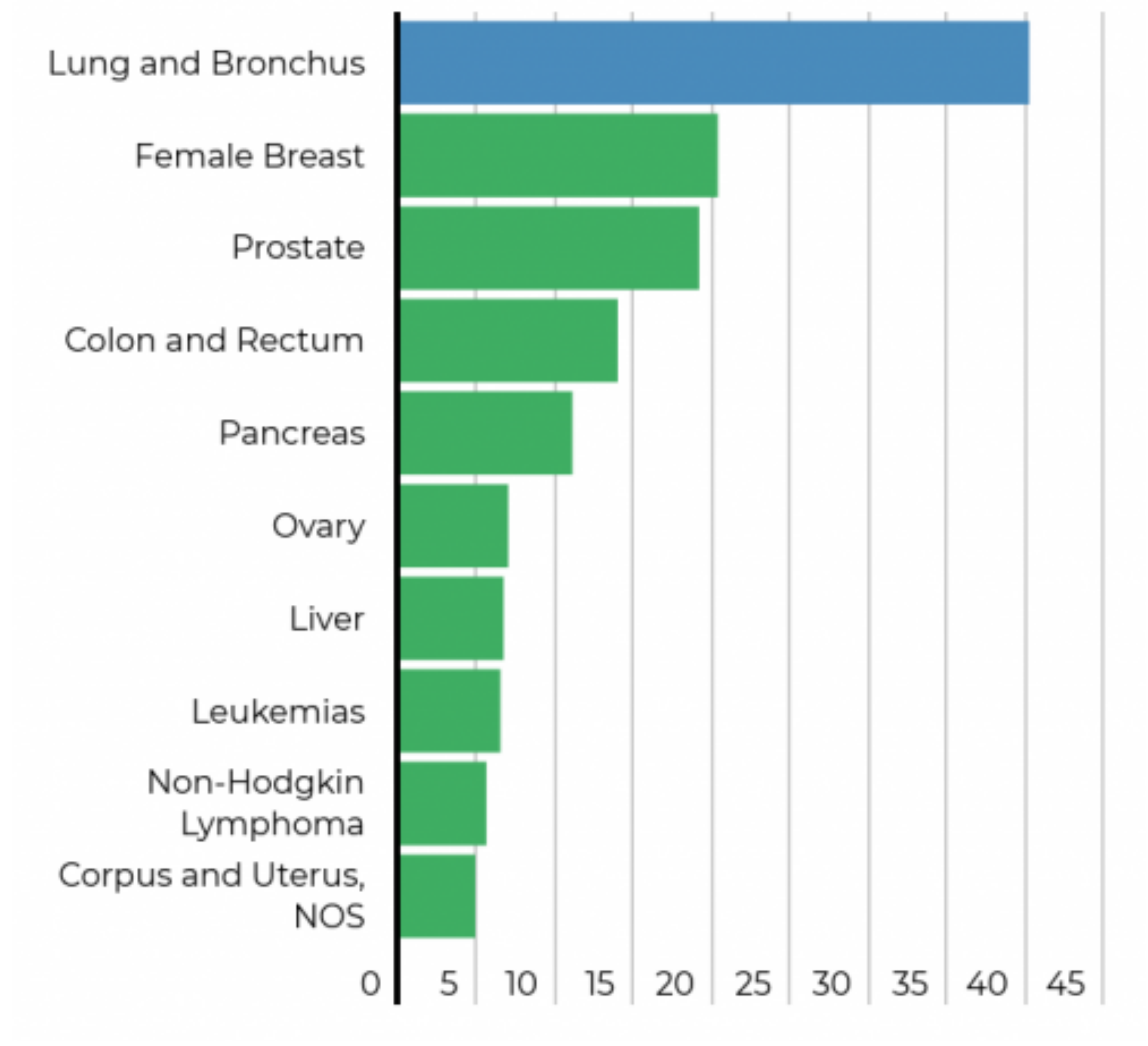
how to use IE?

NLP efforts. -> clinical impact?

Early detection of lung cancer reduces mortality

The leading cause of cancer deaths

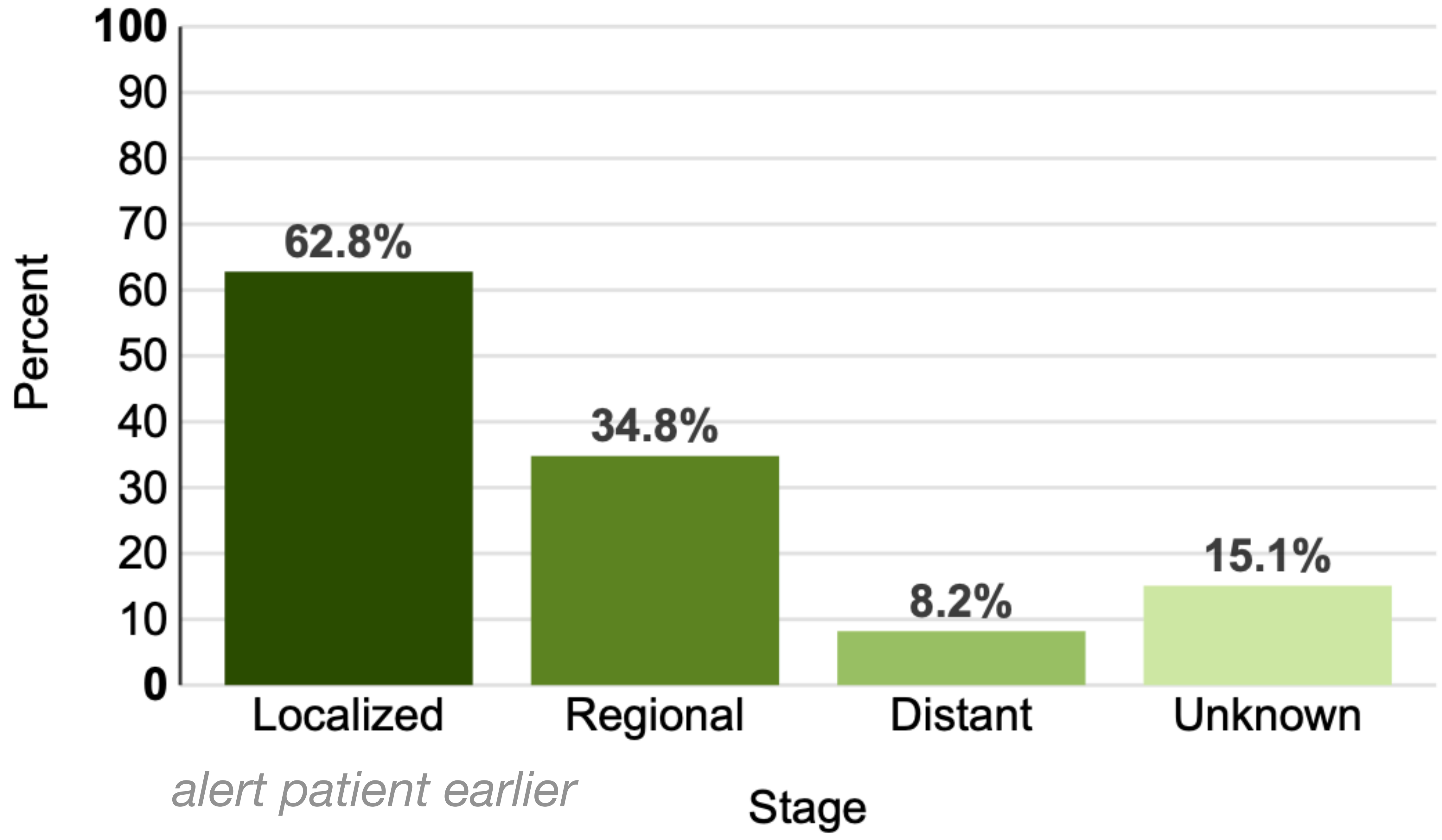
Lung Cancer Rates in the United States (per 100K)



Source: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute

Early detection saves life

5-Year Relative Survival



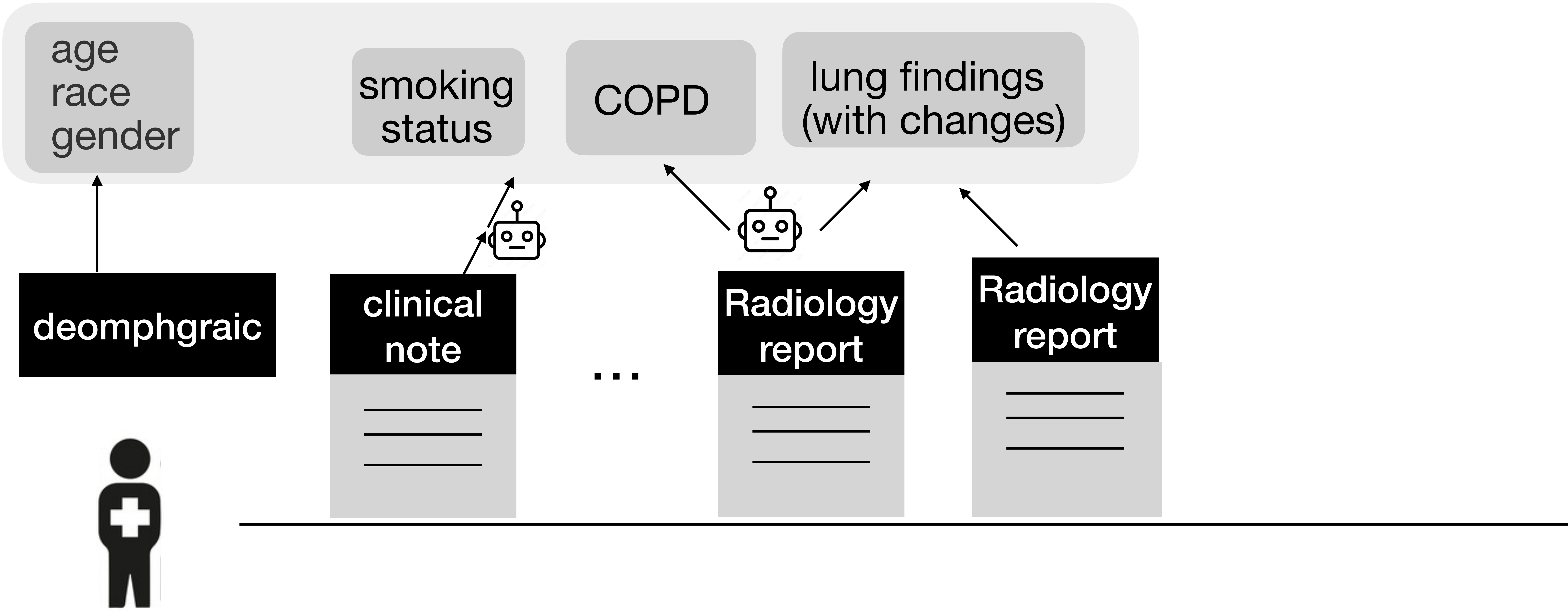
alert patient earlier with EHR signals

NIH NATIONAL CANCER INSTITUTE
Surveillance, Epidemiology, and End Results Program

Lung cancer risk factors in unstructured EHRs

rich literature

- Age
- Smoking
- COPD *Chronic Obstructive Pulmonary Disease*
- Abnormal lung findings
  - type, size, temporal change, etc



age
race
gender

smoking status

COPD

lung findings (with changes)

deomphgraic

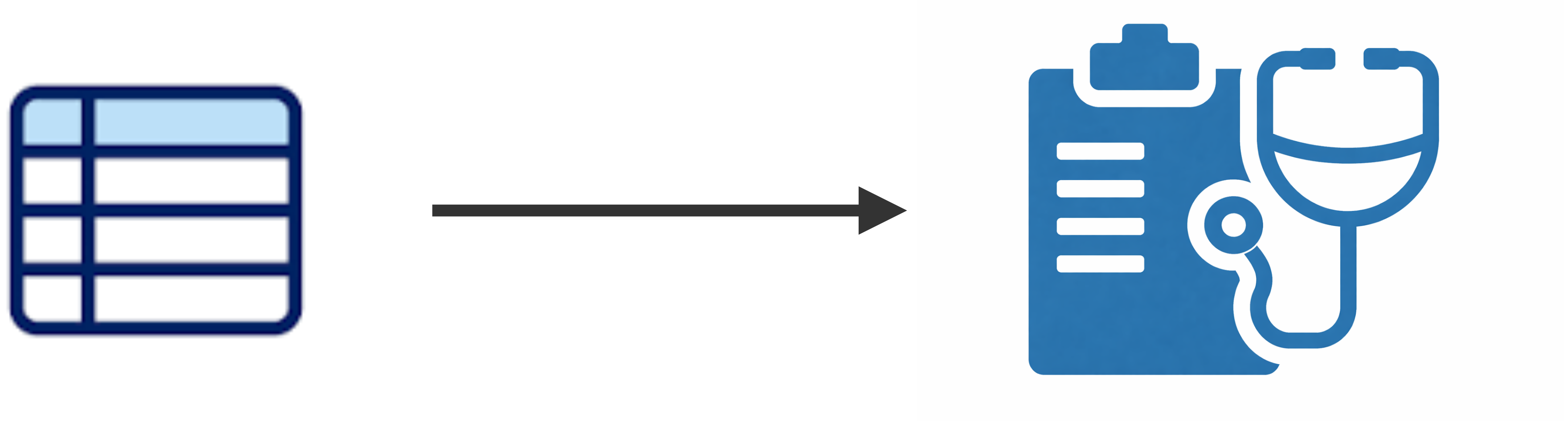
clinical note

...

Radiology report

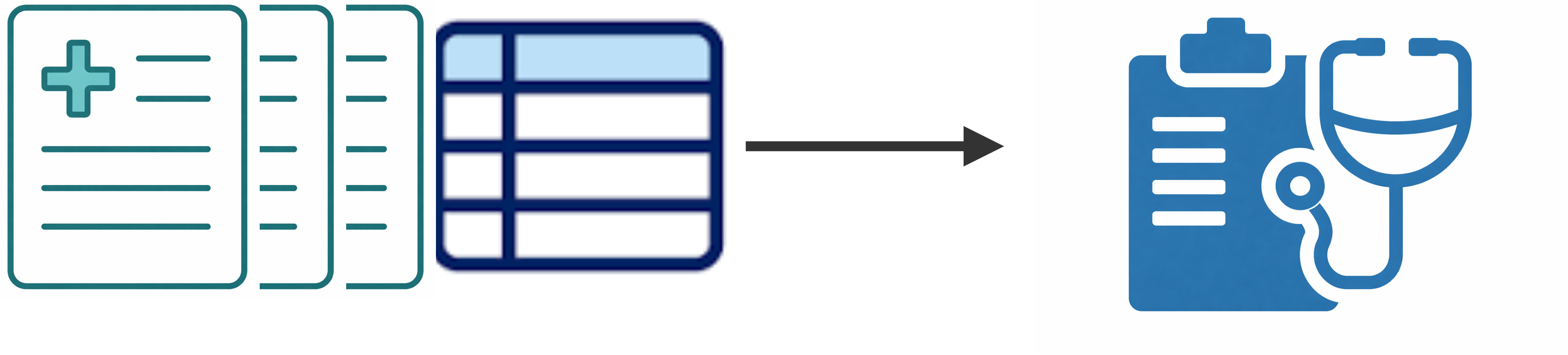
Radiology report

Prior work for lung cancer prediction with EHRs



Use structured data

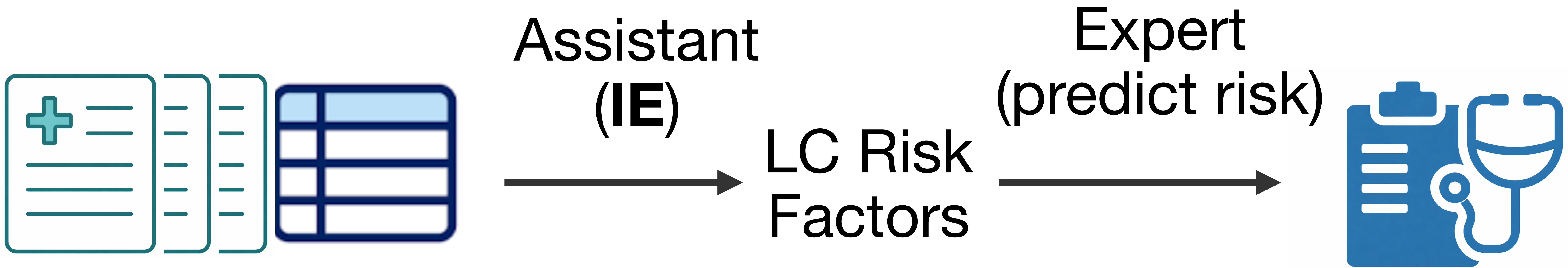
miss information from unstructured data



Use unstructured data

summarization, concept extraction

lack of understanding



Assistant (**IE**)

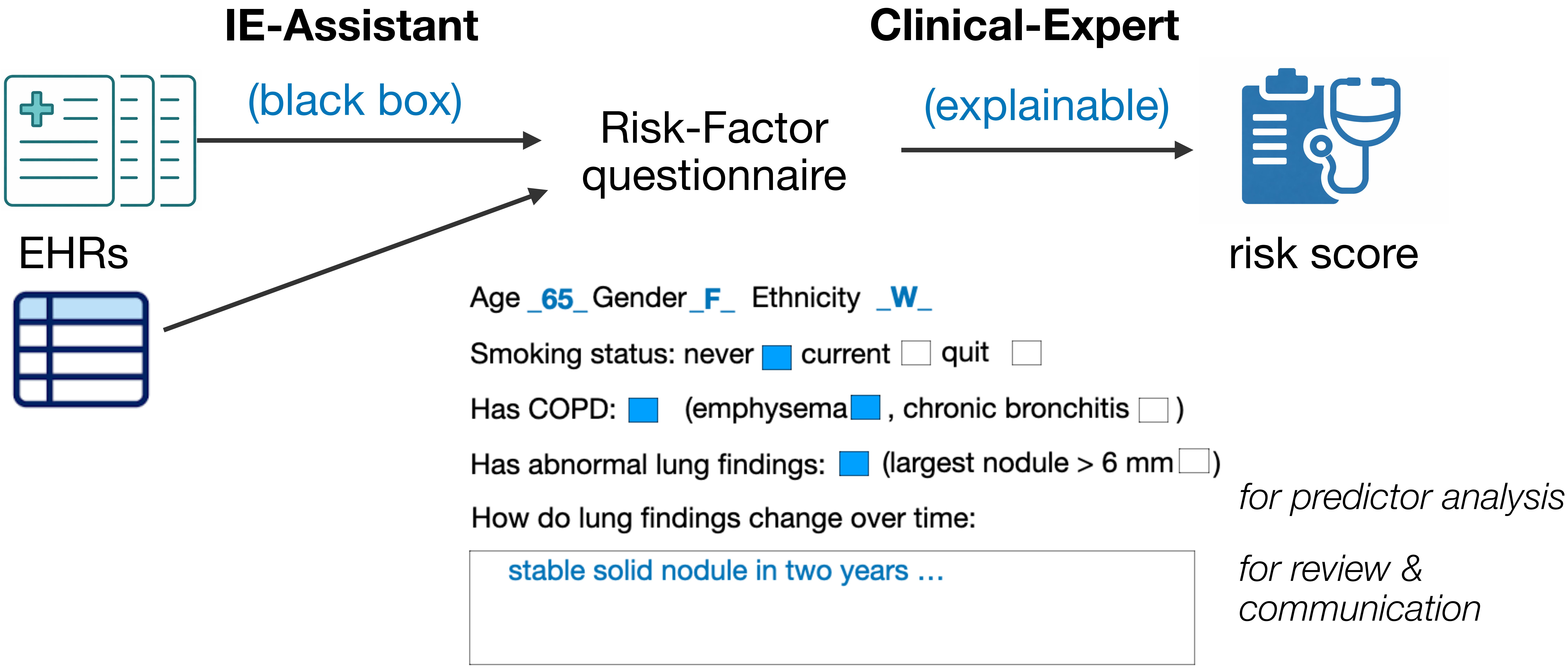
LC Risk Factors

Expert (predict risk)

black box

explainable

IE-Assistant & Clinical-Expert Risk Prediction



IE-Assistant

(black box)

Clinical-Expert

(explainable)

EHRs

Risk-Factor questionnaire

risk score

Age _65_ Gender_F_ Ethnicity _W_

Smoking status: never ☑ current ☐ quit ☐

Has COPD: ☑ (emphysema ☑, chronic bronchitis ☐)

Has abnormal lung findings: ☑ (largest nodule > 6 mm ☐)

How do lung findings change over time:

stable solid nodule in two years …

for predictor analysis

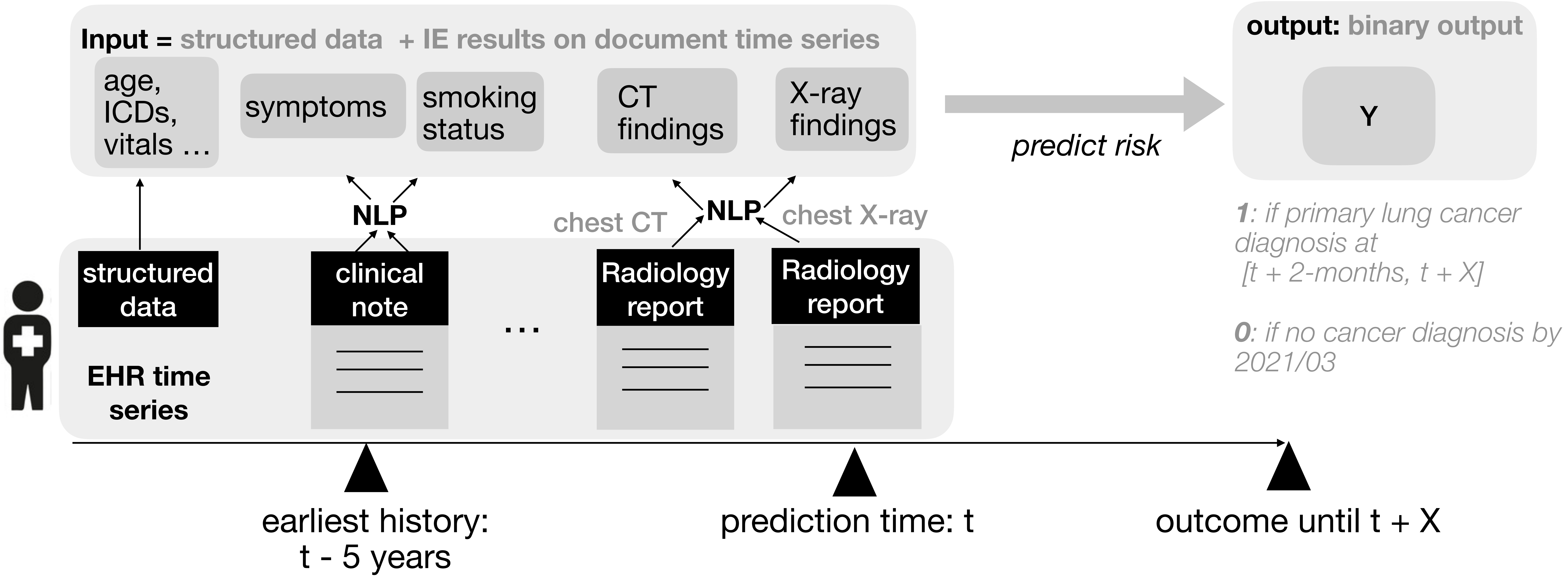
for review & communication

Assistant - Expert Lung Cancer Prediction (ours)

- Task design
- Data: a case-control cohort
- IE approach (Assistant)
- Disease prediction approach (Expert)

Task: Predict lung cancer risk

At time (t) of **chest radiology exam** *(CT or X-ray)*,
predict primary lung cancer diagnosis outcome within risk period ***X*** *(3 years)*.



Input = structured data + IE results on document time series

age, ICDs, vitals …

symptoms

smoking status

CT findings

X-ray findings

predict risk

output: binary output

Y

1*: if primary lung cancer diagnosis at [t + 2-months, t + X]*

0*: if no cancer diagnosis by 2021/03*

NLP

chest CT

NLP

chest X-ray

structured data

clinical note

…

Radiology report

Radiology report

EHR time series

earliest history: t - 5 years

prediction time: t

outcome until t + X

Data: A case-control cohort

scale up IE

1:10 case-control cohort. (3964 cases, 39640 controls)

case: have a valid lung cancer diagnosis in the cancer registry

control: no records of any cancer in the cancer registry

each case / control has a prediction time radiology note prior to diagnosis related to chest (Chest CT, Chest XRay, Abdomen CT)

patient's age is above 40 at the prediction time.

IE-Assistant: fill a risk-factor questionnaire

numerical
categorical

Age _**65**_ Gender _**F**_ Ethnicity _**W**_

Smoking status: never ☑ current ☐ quit ☐

Has COPD: ☑ (emphysema ☑, chronic bronchitis ☐)

Has abnormal lung findings: ☑ (largest nodule > 6 mm ☐)

free-texts → How do lung findings change over time:

stable solid nodule in two years …

IE-Assistant: fill a risk-factor questionnaire

numerical
categorical

Smoking status: never ☑ current ☐ quit ☐

Has COPD: ☑ (emphysema ☑, chronic bronchitis ☐)

Has abnormal lung findings: ☑ (largest nodule > 6 mm ☐)

Merge structured data with LLM QA results *on longitudinal inputs*

for higher data completeness

What is the smoking status of the patient?

Does this patient has COPD, emphysema, and chronic bronchitis?

What is the largest lung nodule size in mm?

IE-Assistant: fill a risk-factor questionnaire

free-texts → How do lung findings change over time:

stable solid nodule in two years …

- Event Extraction finding list (in time order) *from recent to old*

  filter for present & lung findings

1B *Event Extraction on Radiological findings Park et al, 2024*

- LLM direct summarization

2 *Longitudinal Summarization radiological lung findings*

- LLM findings (RadTimeline)

  (in time order) *column by column*

  (in group order) *row by row*

contextualized on related findings

Model risk prediction - predict binary outcome

Xgboost

only accept numerical/categorical features

DistillBERT

accept all features after verbalization

vignette *style*

a case study style for medical education

This patient is a six-five year old White Female current smoker, has COPD emphysema, has lung abnormality. ….

Impact of individual risk factor completeness in risk prediction

XGBoost results		data completeness boost % with LLM		AUC (demographic baseline = 0.664		AP (demographic baseline = 0.148	
Category	Variable	*Case*	*Control*	Structured Data	Merge with LLM	Structured Data	Merge with LLM
Smoking	current	*22*	*57*	70.8	**71.4**	19.4	**20.6**
	never	*13*	*33*	72.6	**74.2**	19.2	**20.6**
	quit	*26*	*40*	66.5	**66.9**	15.3	**15.9**
COPD	copd	*48*	*96*	**70.8**	70.4	**18.2**	18.0
	emphysema	*116*	*225*	69.0	**71.3**	18.2	**20.4**
Abnormal lung finding	has-lung-abnormal	*43*	*33*	68.8	**71.4**	18.7	**21.8**

higher completeness than structured data

IE amplifies effective individual predictors with higher completeness

Longitudinal free-text features

small model VS LLM

time order VS group order

free-text features with DistilBERT	AUC	AP
Event extraction (time order)	**83.7**	**36.5**
LLM direct summarization	81.6	35.2
LLM RadTimeline (time order)	82.5	36.6
LLM RadTimeline (group order)	81.6	32.5

Small event extraction model is the best, followed by LLM findings in time order

Using group order hurt LLM finding features

LLM: Llama 3.1-8b-instruct

Longitudinal free-text features - event extraction

DistilBERT	5-year history (default)		1-month history	
	AUC	AP	AUC	AP
findings (only lung)	**83.7**	**36.5**	**83.3**	**37.2**
findings (all body parts)	82.0	35.9	81.9	35.7

Irrelevant information hurts (information not related to lung)

even with a large cohort

Classifying anatomy types in event extraction (small LMs) can filter irrelevant information and helps disease prediction

Conclusion: lung cancer risk prediction

IE helps lung cancer prediction

boost data completeness

improve the predictive power of individual risk factors

Irrelevant information hurts

similar to domain generalization experiments

LM behaves different when the input sequence is reordered

Potential reason: positional bias

Verifiable
Longitudinal
Summarization

2

LLM
Longitudinal radiological lung findings
(under review, 2025)

Conclusion

IE helps disease prediction

so many more risk factors to extract
to help so many more diseases

Real Data

For domain generalization

For long context / longitudinal understanding

Domain generalization lessons

Reduce unnecessary inputs (include shortening)

shorter data helps pretraining long context LLM

we still do supervised learning for clinical models

prompt compression?

Sequence order matters

in extracting longer sequence

in understanding longitudinal inputs

LM positional bias?