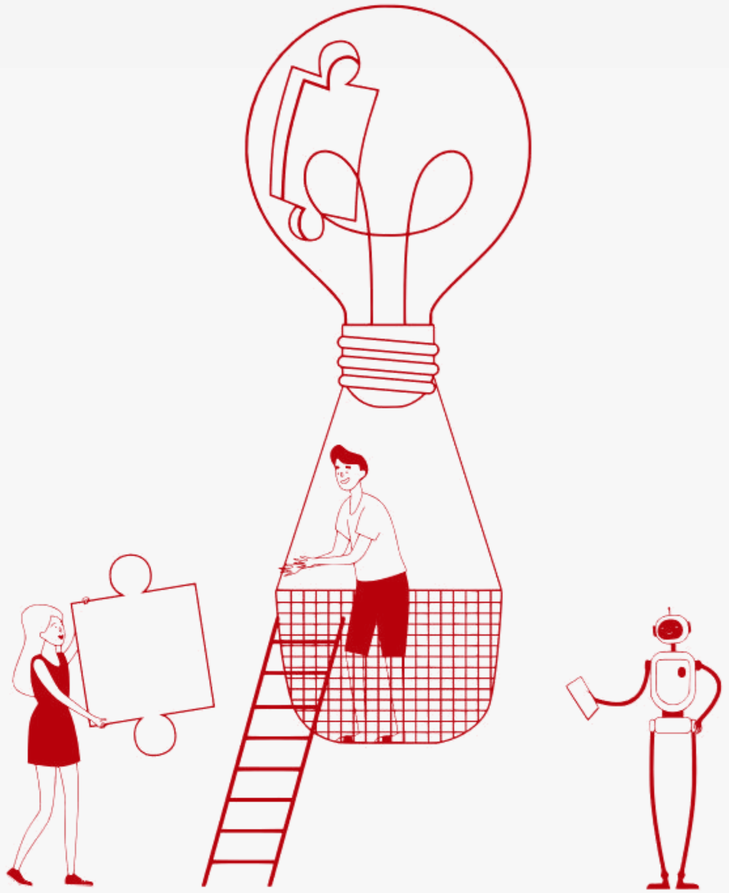
https://thegradient.pub/human-centered-explainable-ai/

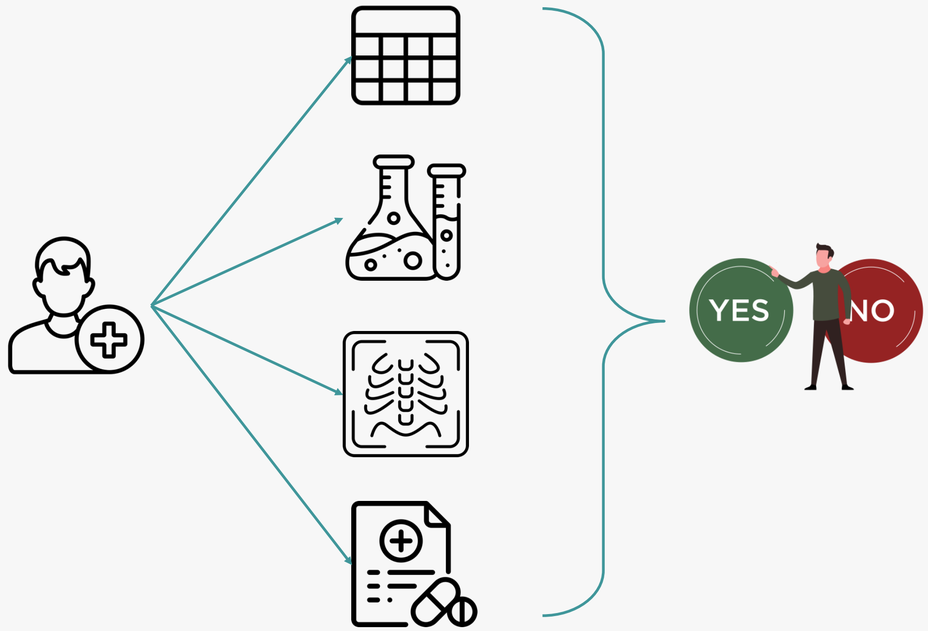
xHAIM: Improved Performance and Explainability



Periklis Petridis, Georgios Margaritis, Vasiliki Stoumpou, Dimitris Bertsimas

The Problem

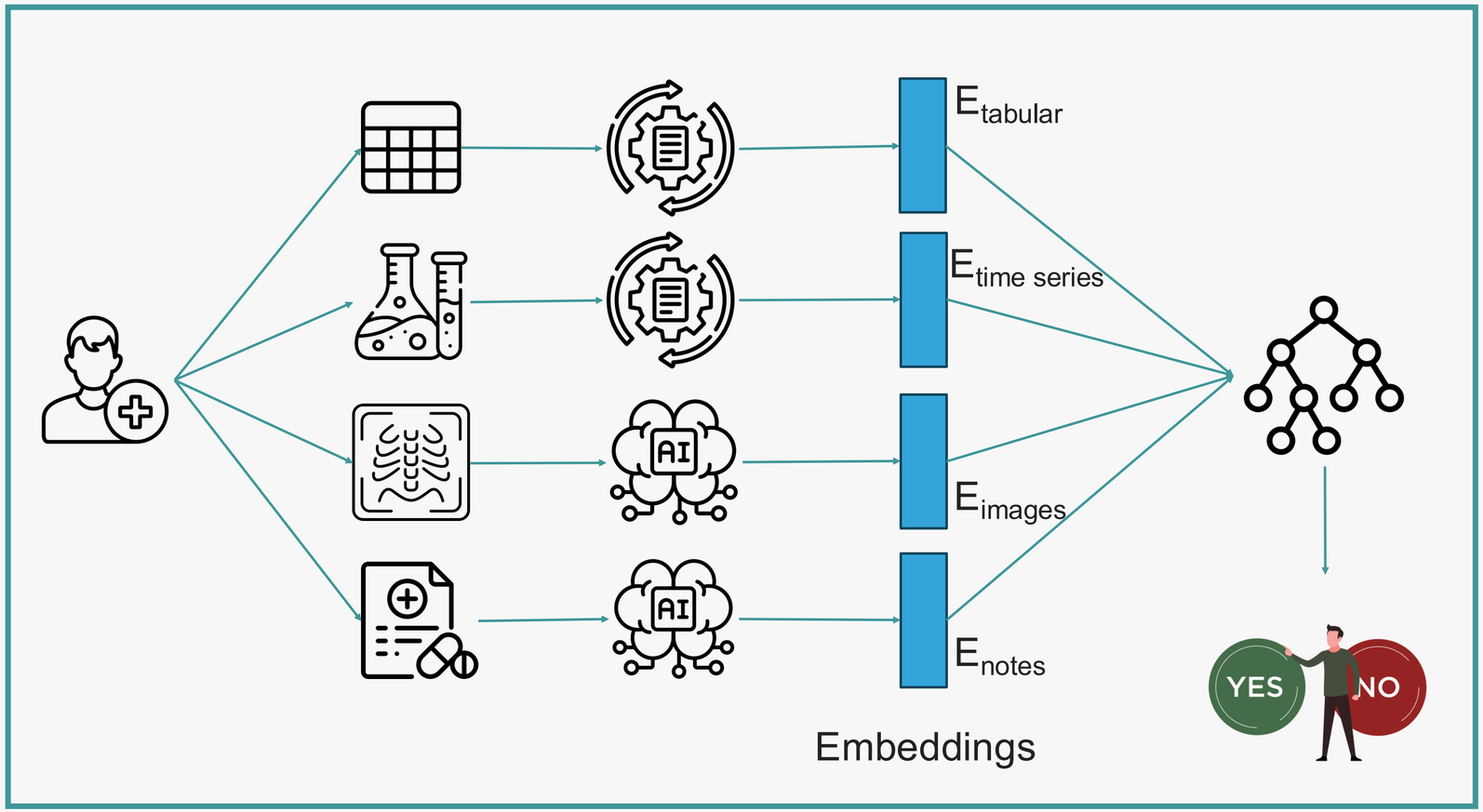
- Patients with multimodal data:
  - Tabular (e.g. demographics, medications)
  - Time series (e.g. lab values, vitals)
  - Images (e.g. X-Rays)
  - Notes (e.g. Radiology Reports, Past Medical History)
- Can we leverage these data to predict a diagnosis/outcome?



Tasks

Pathology	Outcomes
Pneumonia	Mortality
Diabetes	Length of stay

Existing approach: HAIM[1]



- The HAIM framework takes as input the multimodal data
- Transforms them to embeddings using modality-specific models
- Concatenates them (fusion)
- Uses them as input to a final classification model
- This classification model can be trained to perform any downstream task (e.g. predict if patients have pneumonia or not)

1. Soenksen, L.R., Ma, Y., Zeng, C. *et al.* Integrated multimodal artificial intelligence framework for healthcare applications. *npj Digit. Med.* **5**, 149 (2022). https://doi.org/10.1038/s41746-022-00689-4

What are potential issues?

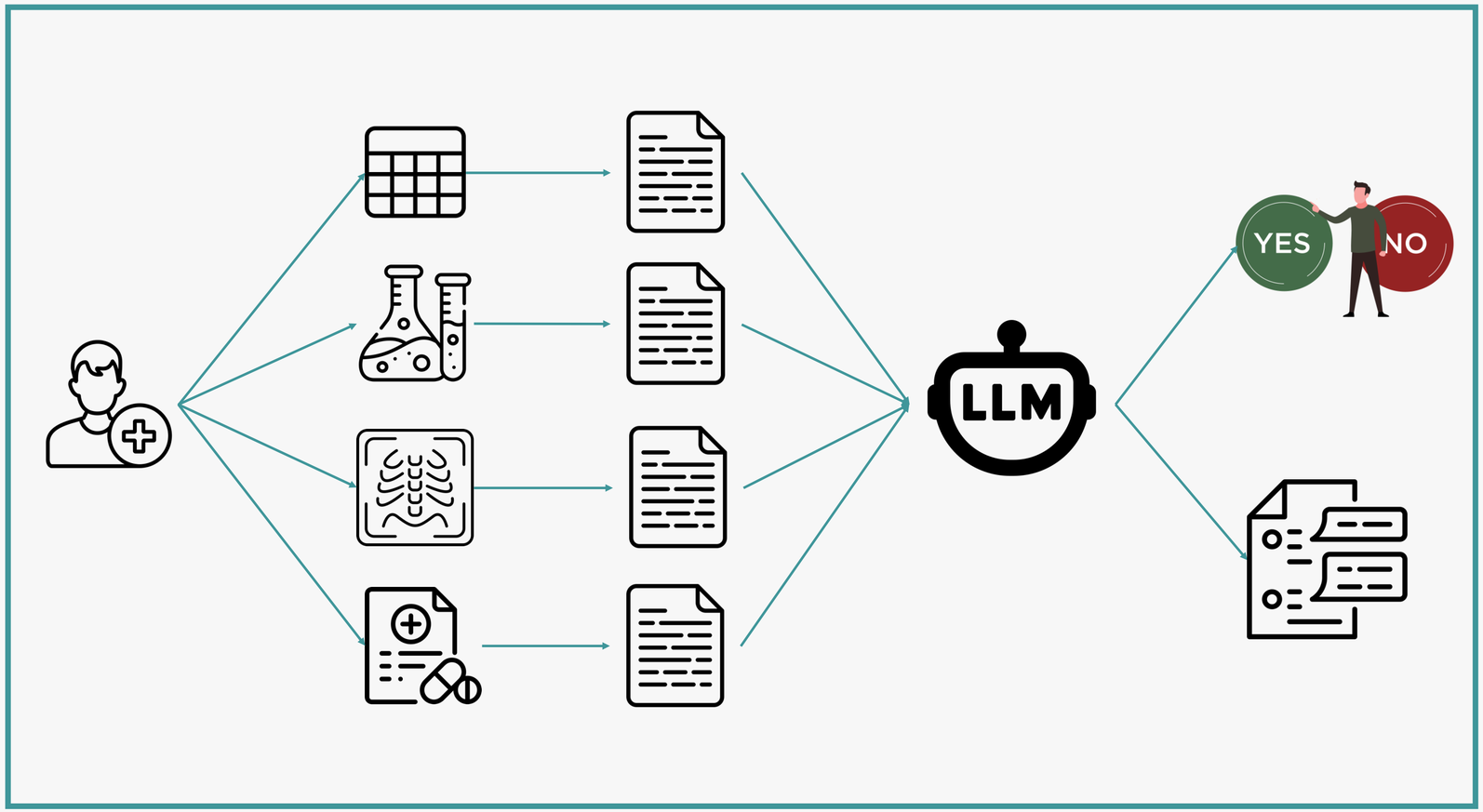
Too much input data, that can often be irrelevant to the final downstream task.

Input to the final classification model can be **noisy**, since a single embedding needs to capture all these details.

No explainability (black box architecture)

Makes adoption and clinical deployment harder.

Existing approach: Zero-shot LLM



- Another approach is to convert every available data source to text
- Provide all the raw information to an LLM
- Ask the LLM to provide a prediction and an explanation

What are potential issues?

Too much input data, hard to fit to an LLM.

Unsupervised predictions, not optimized for prediction accuracy, poorly calibrated.

Not trained/finetuned on our data. Finetuning proprietary LLMs on sensitive patient data is not trivial.

Our approach: xHAIM

Improve Input

Step 0: Define task & keywords (e.g. “hypertension”)

Step 1: Find the most **relevant patient data** for our task

Step 2: Produce a **single summary** encompassing all patient information

Solves: **Too much data**

Improve Predictions

Step 3: Feed summaries into **HAIM** for **better predictive performance**

Explainability

Step 4: Combine the following for **GenAI explanations**:

i) ML predictions,

ii) summaries and raw patient data,

iii) medical knowledge

Solves: **Transparency**

Converting multimodal data to text

- Tabular and time series data can be converted to natural language*:
  - Tabular*: Patient is a 80 year old male. Patient is prescribed 100 mg of Insulin.*
  - Time series: *Patient's average blood pressure is 60/120 mmHg, with its last measurement being 75/130 mmHg.*
- Images, along with their corresponding reports, are passed as a prompt to a Multimodal LLM (e.g. Qwen) and a final text description is generated.
- Notes can be used as they are.

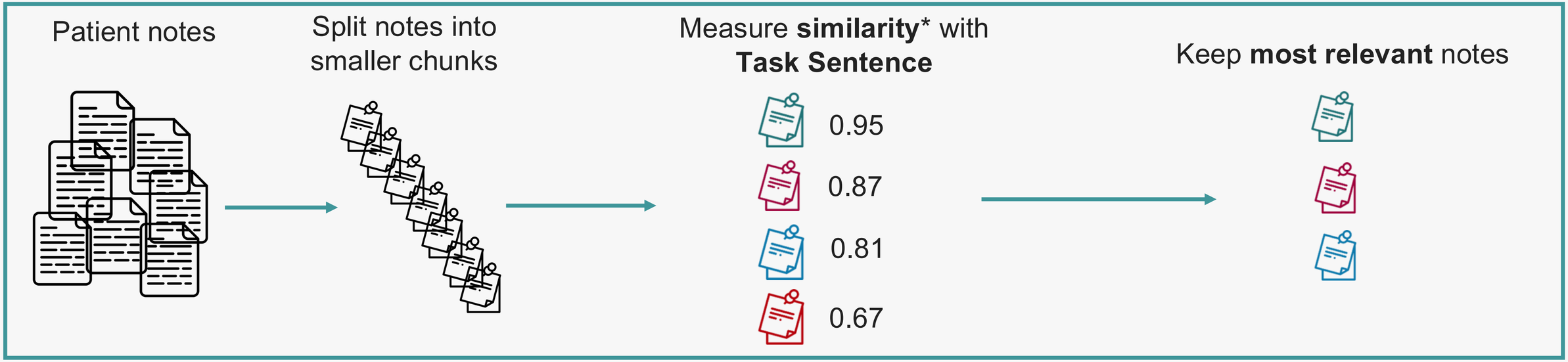
* Tabular and time series data can also just be processed as in regular HAIM, as applying xHAIM to them is less beneficial compared to the other modalities.

Part 1: Finding relevant data

We define a task sentence, that is related to the final task.

Example task sentence

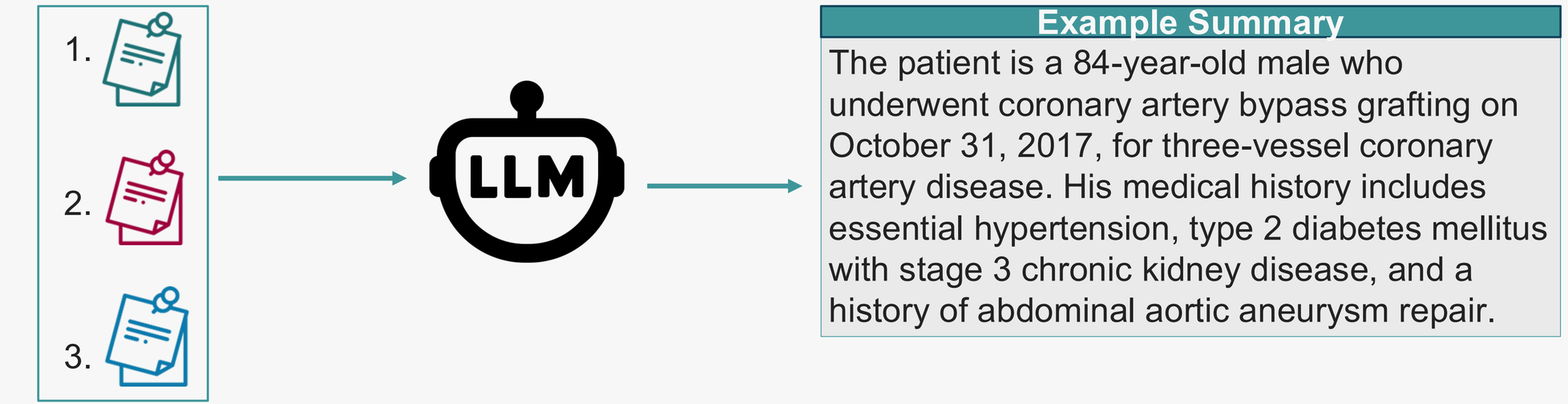
"essential hypertension, hypertensive, high blood pressure, HTN, systolic, diastolic"



*We measure similarity with a combination of:

- Embedding similarity: We encode both the target sentence t as well as the N resulting chunks as embeddings (S_t, and S_1, S_2, …, S_N) respectively. We find the top-k (e.g. 10) chunks (S_i) that are more similar to our target sentence (S_t)
- BM25 (keyword/frequency based)

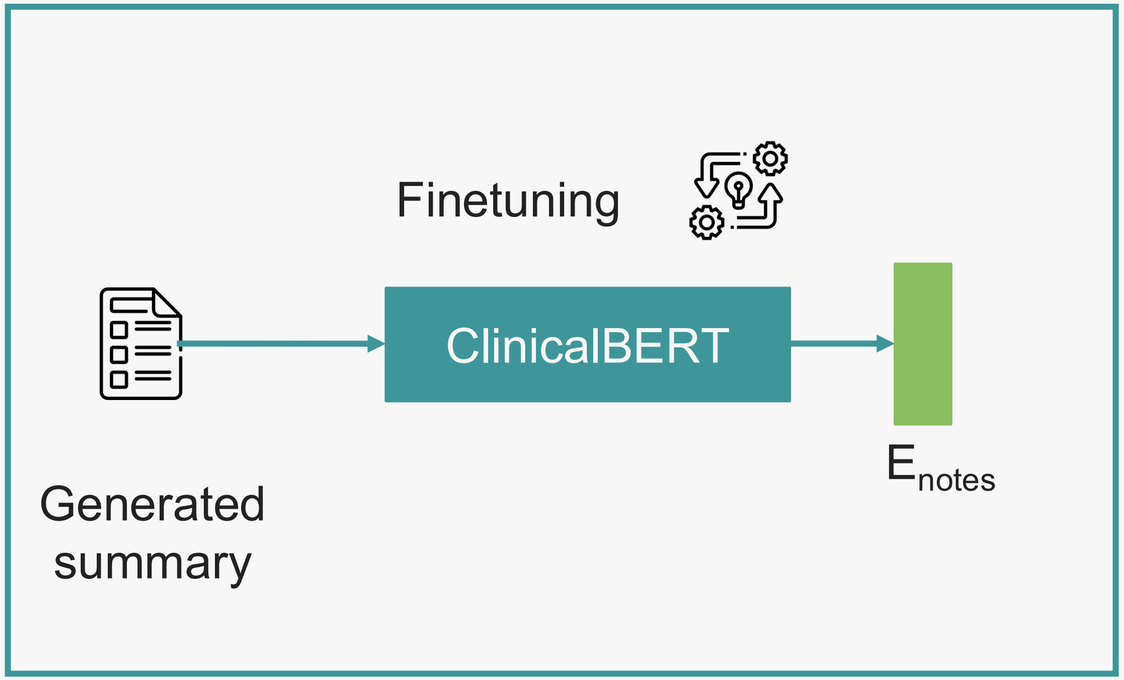
Part 2: Generating summary for each modality



We want to create a uniform, comprehensive paragraph that summarizes all the information from the patient's most relevant chunks for the specific modality.

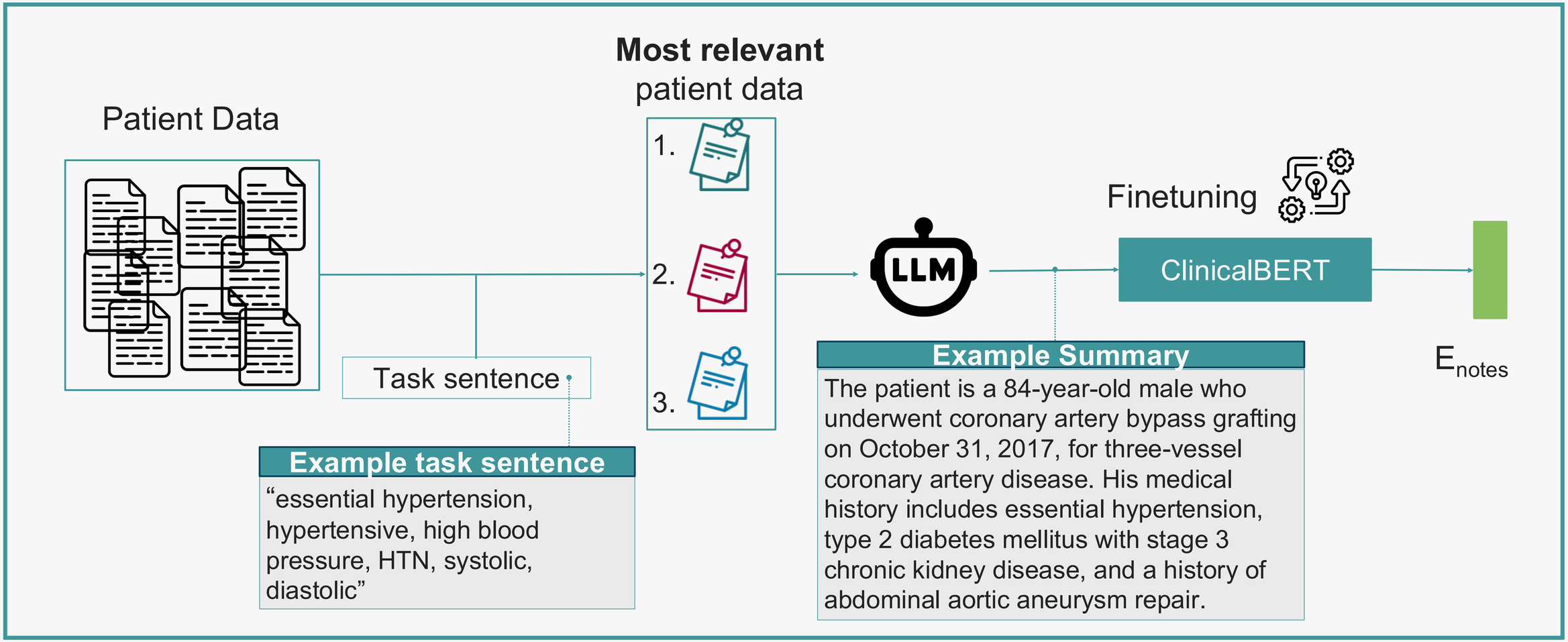
This summary is much more compact than the previous input, which can be thousands of words.

Part 3: Finetuning ClinicalBERT for each modality

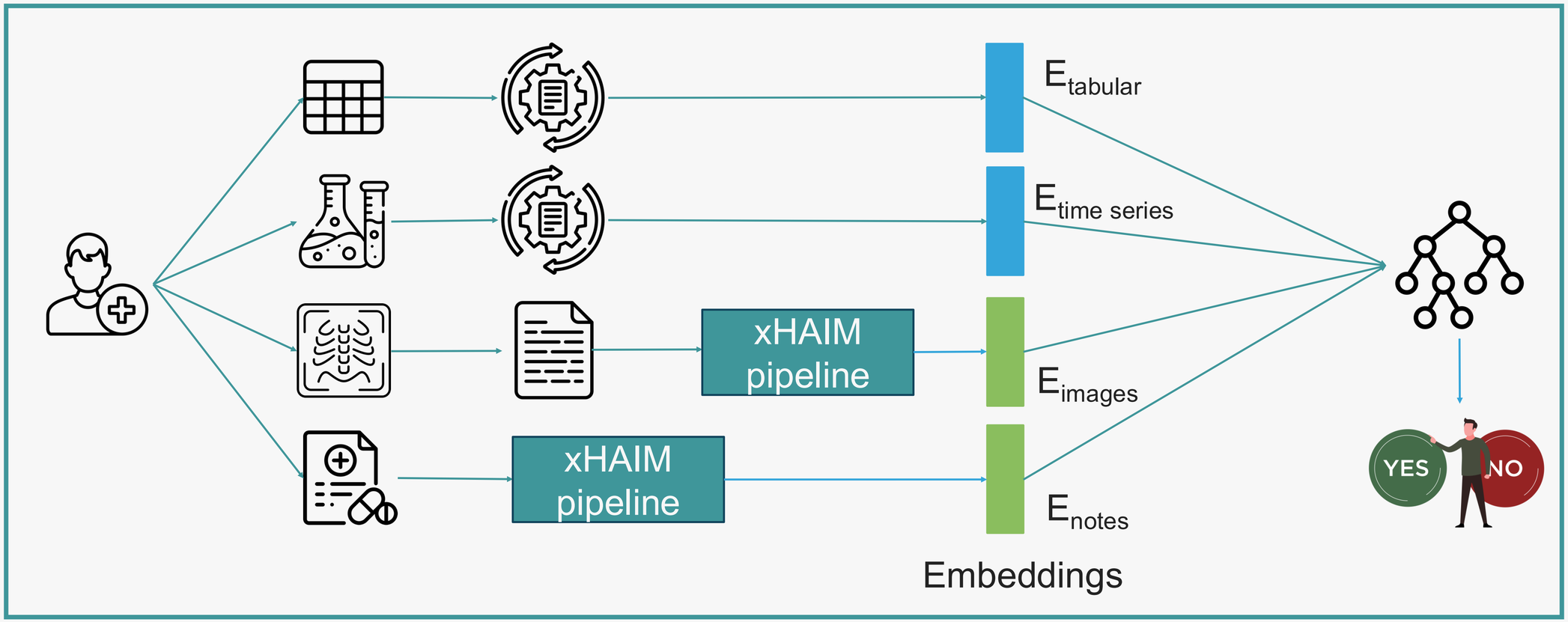


- In the original HAIM pipeline, all modalities were converted to representations, called embeddings, that come from pretrained models (e.g. DenseNet for images, ClinicalBERT for notes)
- These embeddings come from models that have been trained on large amounts of data, that are different from ours
- The size of the generated summaries and the fact that all modalities are encoded as text, enables **effective finetuning** of pretrained models (e.g. ClinicalBERT) to get representations that are more tailored to our specific dataset

Bringing it all together (for e.g. the notes modality)

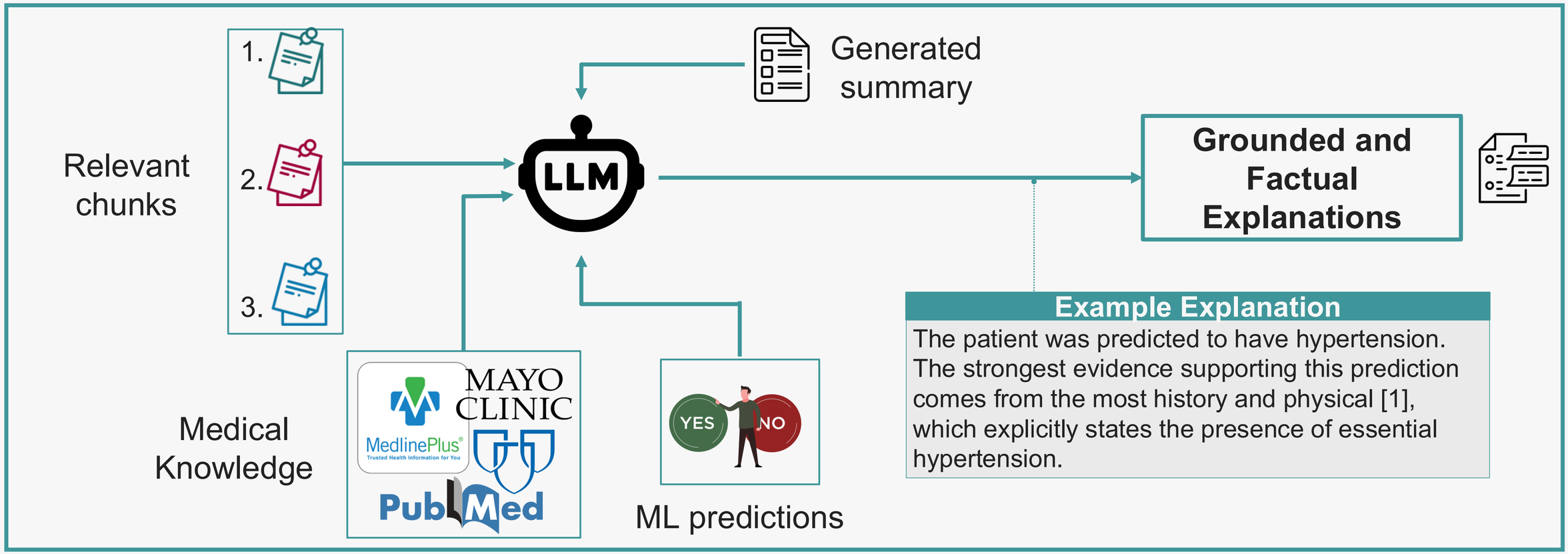


Part 4: Training the final classification model



- The embeddings coming from the finetuned models are concatenated and used as input to the final classification model.
- This results in improved performance across modalities, since the data quality used to train the model is significantly increased compared to the baseline (HAIM).

Part 5: Explainability



We pass the relevant chunks, the generated summary, the ML predictions and the related medical knowledge through a Generative AI model and we generate an explanation of the model's prediction, with specific citations in the original raw text.

Experiments

- We utilized MIMIC-IV data from ICU patients (demographics, lab values, chest X-Rays and radiology and process reports)
- In our experiments, we worked with:

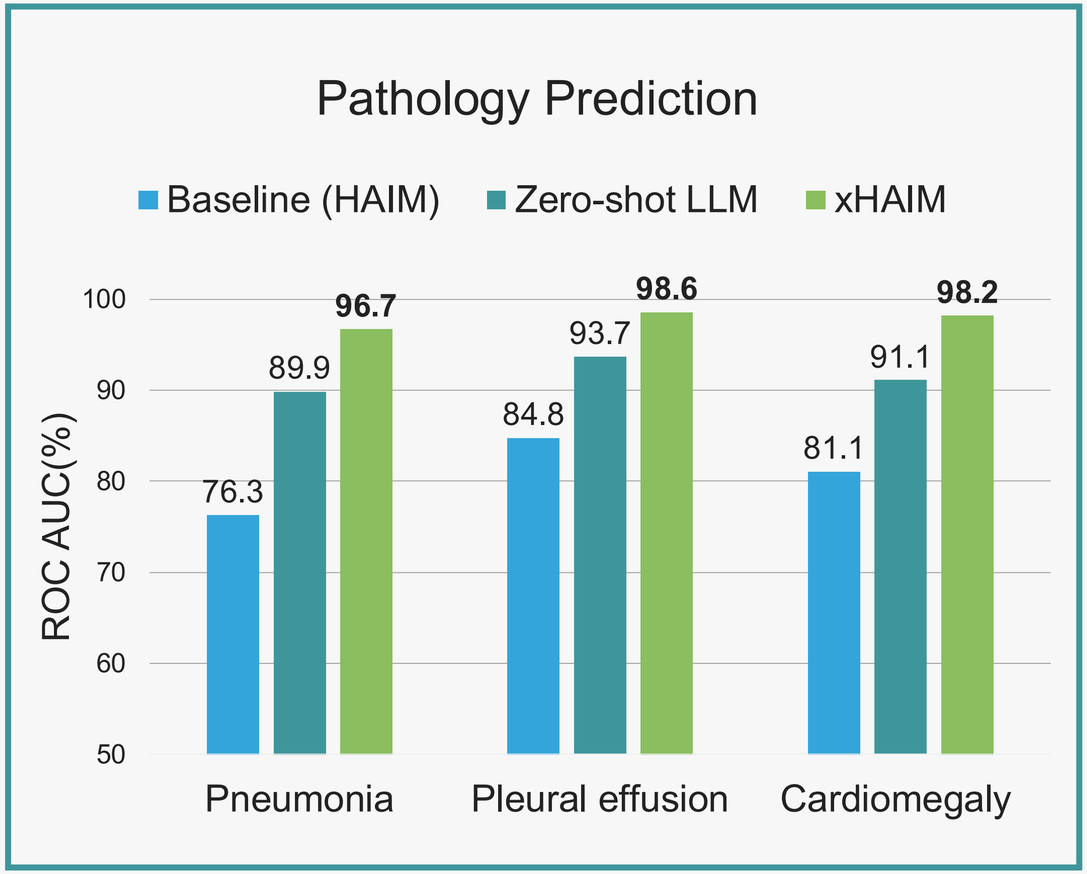
1. Tasks for which the information can be easily inferred from the input data (**information extraction-like tasks**):

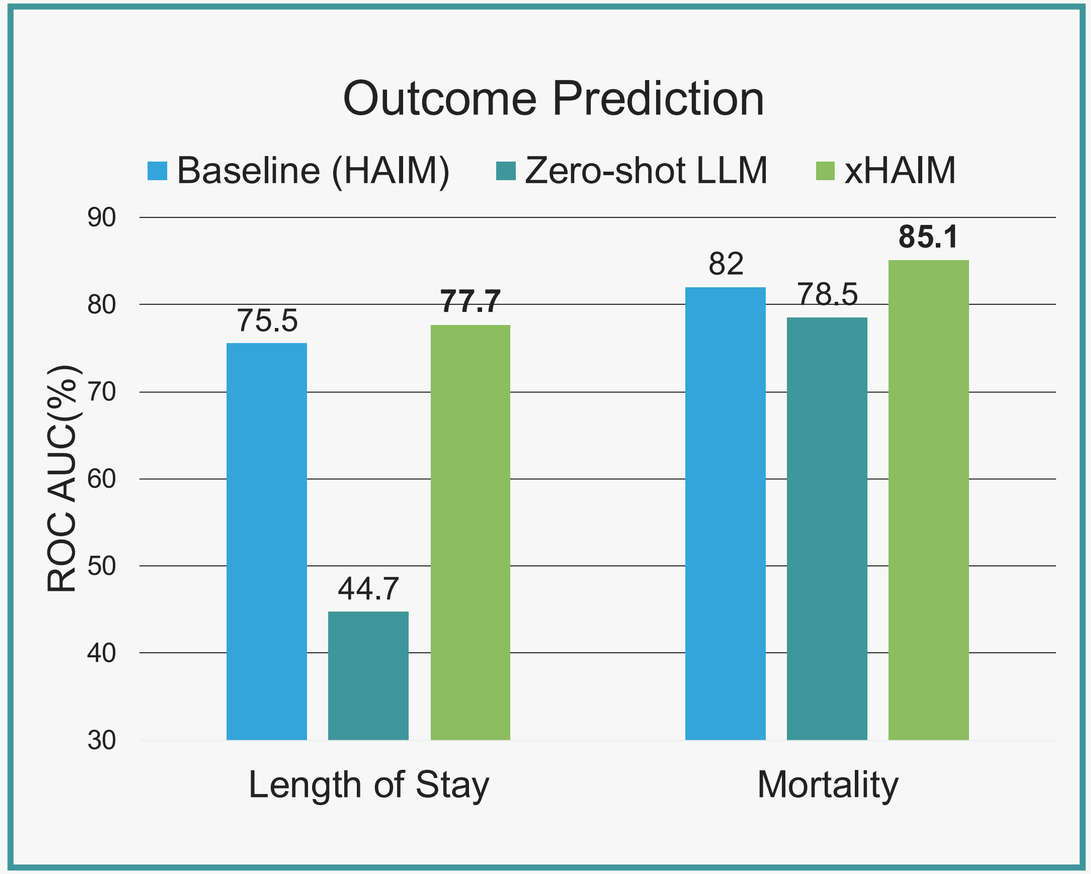
   - Pneumonia
   - Pleural Effusion
   - Cardiomegaly

2. **Prospective outcomes**

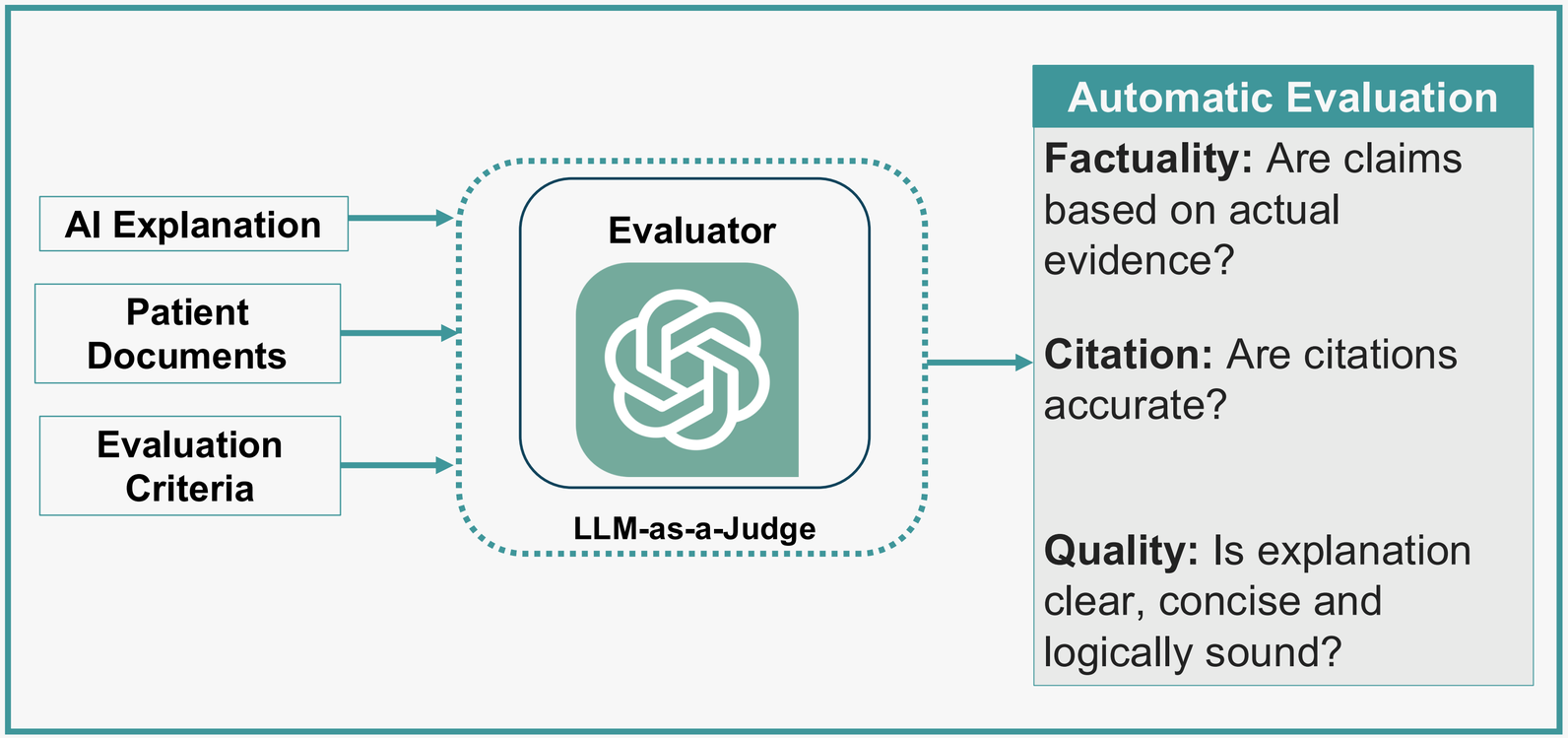
   - Mortality within 48 hours,
   - Length of stay > 7 days in the ICU)

Results (Evaluating Prediction Performance)



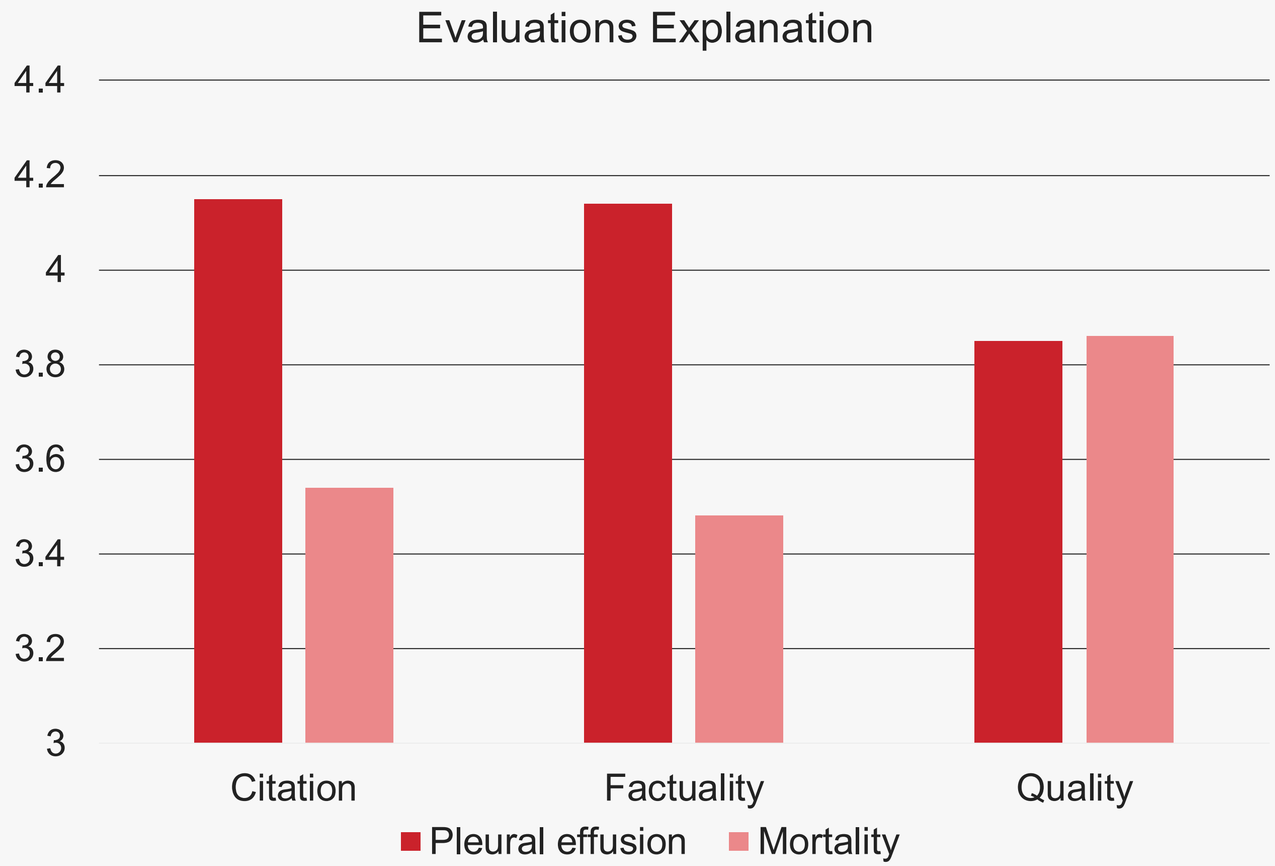


Results (Evaluating Explanation Quality)



- We created a scoring system (3 main categories: Citation quality, Factual correctness and Overall Quality (verbosity etc.) and a thorough rubric)
- Human annotators used this system and annotated 50 summaries for pleural effusion and 50 summaries for mortality
- **LLM-as-a-Judge framework:** one generates critiques, and the other incorporates both the critique and the explanation to assign final scores to the same summaries.
- After establishing adequate overlap between the score distributions of human annotators and LLM-as-a-Judge, we can trust the LLM-as-a-Judge and we use this system to generate evaluations for 1,000 explanations

Results (Evaluating Explanation Quality over 1,000 cases)



"Mortality within 48 hours" is harder to generate explanations for, because it is difficult even for humans to find evidence (e.g. citations) for such an outcome in the relevant documents

Explanation Example: citation of raw patient data

Generated Explanation

The patient has been predicted to have pleural effusion (prediction 1) with a model output probability of 1.00, based on consistent documentation of pleural effusion across multiple radiology reports. The strongest evidence supporting this prediction includes the most recent chest radiology note **[1]**, which describes a "small left pleural effusion is likely," and the second-most recent report **[2]**, which explicitly mentions "small bilateral pleural effusions." No documents contradict the presence of pleural effusion, and the model's maximum confidence likely reflects this consistent, multi-source documentation.

[1] Radiology Note – 05/25/2021

ap chest, man with a right apical pneumothorax. impression. small left pleural effusion is likely….

[2] Radiology Note - 05/22/2021

single frontal view of the chest. there are persistent low lung volumes, bibasilar atelectasis larger on the left side and small bilateral pleural effusions. moderate interstitial pulmonary edema has increased...

Conclusions

- **Improved performance:**
  - Find relevant data to the task at hand
  - Downstream models use dense summaries instead of raw data, avoiding noise
  - AUC improvements of up to 2-20% (additive)

- **Improved explainability:**
  - Explanations grounded to actual raw patient data