Yale SCHOOL OF MEDICINE
*Biomedical Informatics and Data Science*

# OHDSI Working Group Spotlights – NLP WG

Hua Xu, PhD and Vipina Kuttichi Keloth, PhD, Yale University

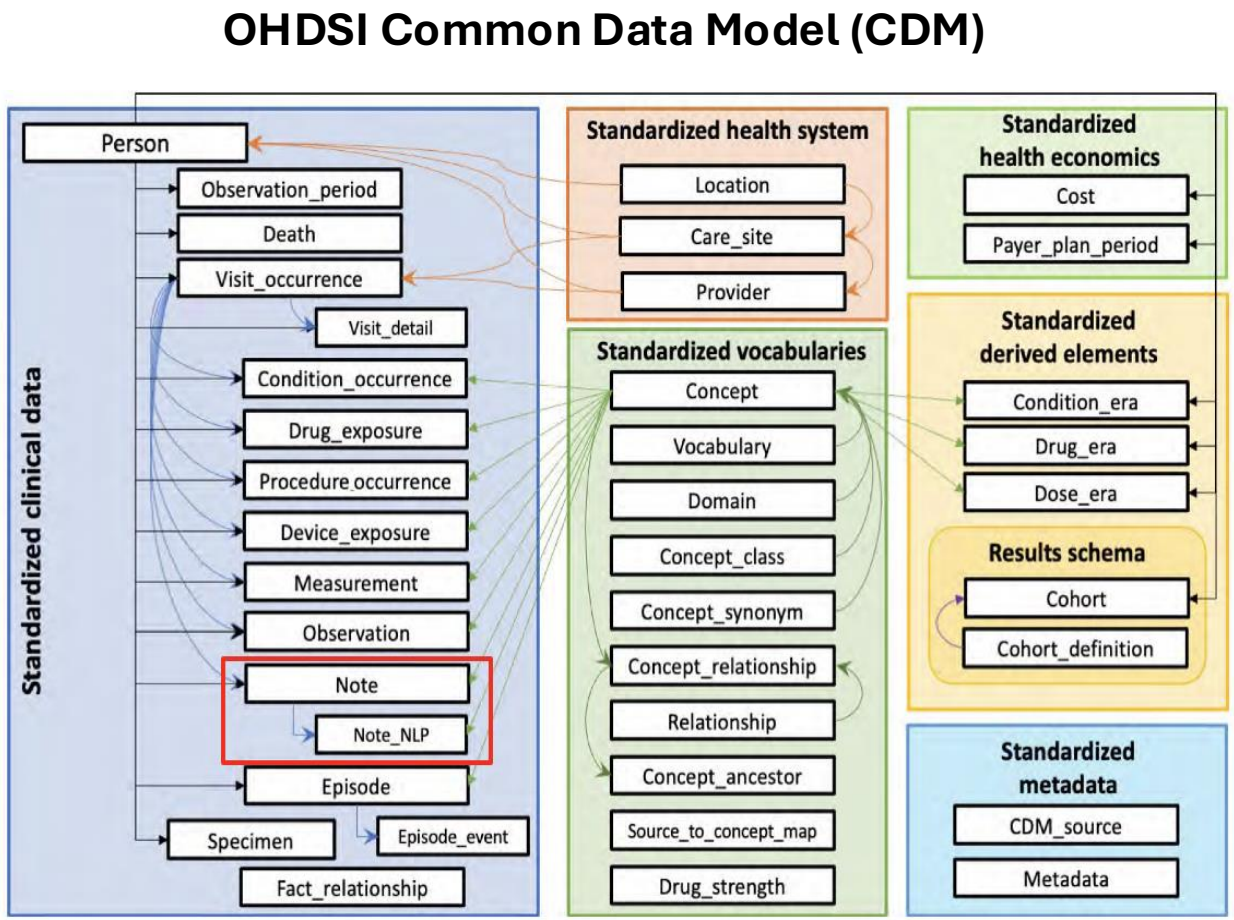
July 14, 2026

# OHDSI NLP Working Group

- Established in 2015, with the goal to promote the use of textual data in electronic health records (EHRs) for observational studies under the OHDSI umbrella
- Three objectives:
  - Develop standard representations for clinical text and NLP output data
  - Build methods and tools to facilitate textual data processing
  - Conduct cross-institutional studies and disseminate best practice of using textual data for real world evidence generation
- Available at https://www.ohdsi.org/web/wiki/doku.php?id=projects:workgroups:nlp-wg

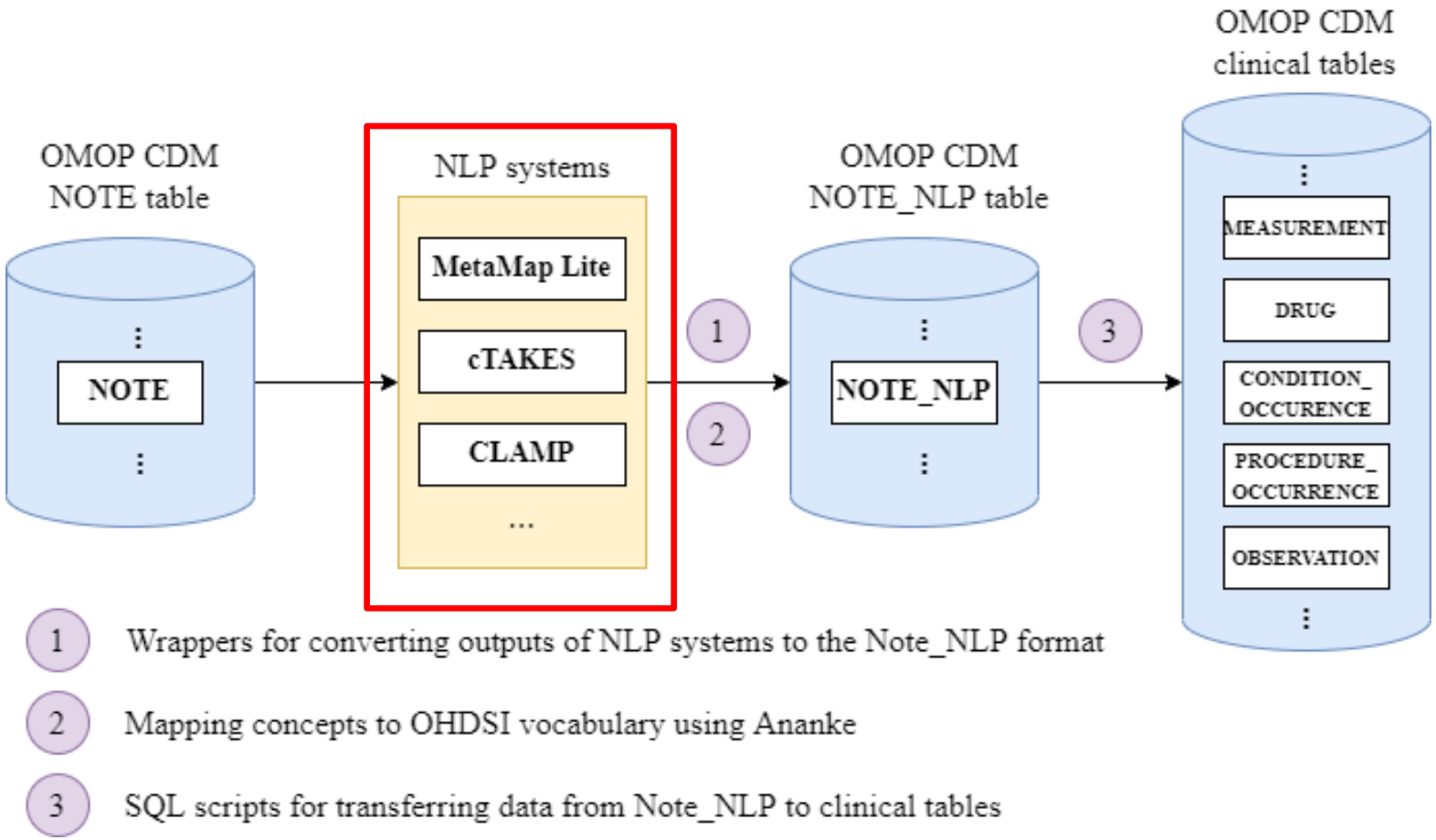
# Representing Clinical Texts and NLP Outputs in OMOP CDM

- To enable the storing of clinical text and the information extracted by the NLP tools from the text into the OMOP CDM
  - Note table - includes the unstructured clinical documentation of patients in EHRs, along with additional meta information (e.g., dates the notes were recorded, types of notes)
  - Note_NLP table - store select NLP outputs from clinical notes (e.g., name and concept id, modifiers)



# Workflow for Textual Data in CDM

- Run natural language processing (NLP) systems to process textual notes in NOTE table
- Convert NLP system output into NOTE_NLP table
- Transfer concepts from NOTE_NLP to clinical tables in CDM




Keloth VK et al. Representing and utilizing clinical textual data for real world studies: An OHDSI approach. J Biomed Inform. 2023 Jun;142:104343. doi: 10.1016/j.jbi.2023.104343.

# Information Extraction from Clinical Notes

## Named Entity Recognition - NER

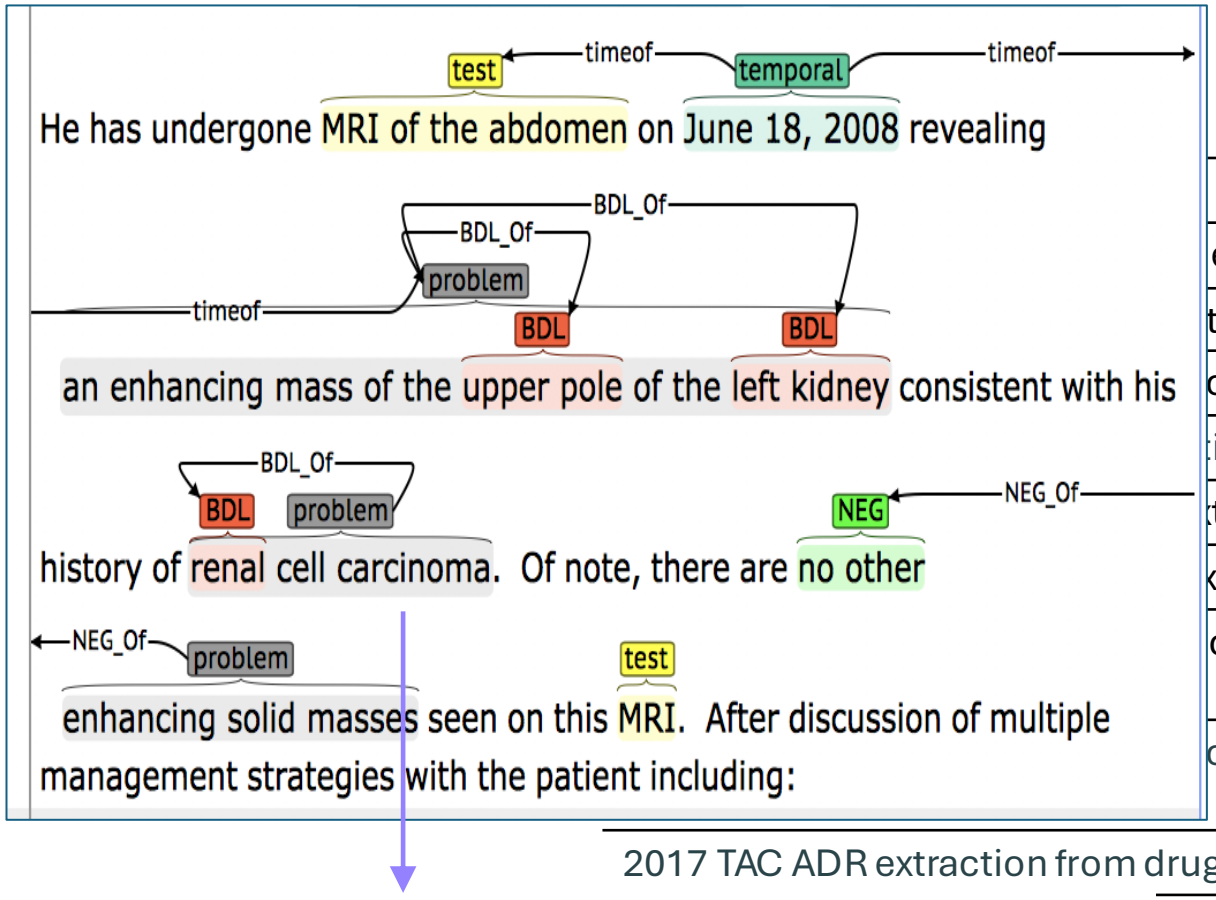
Recognize boundary and type of an entity mention in the text

## Relation Extraction - RE

Extract modifiers of main entities, such as negation, subject, conditional, certainty, temporal etc.

## Concept Normalization - CN

Link an entity to a concept in an ontology, also called entity linking




2017 TAC ADR extraction from drug labels

- C64 Malignant neoplasm of kidney, except renal pelvis
  - C64.1 Malignant neoplasm of right kidney, except renal pelvis
  - C64.2 Malignant neoplasm of left kidney, except renal pelvis
  - C64.9 Malignant neoplasm of unspecified kidney, except renal pelvis

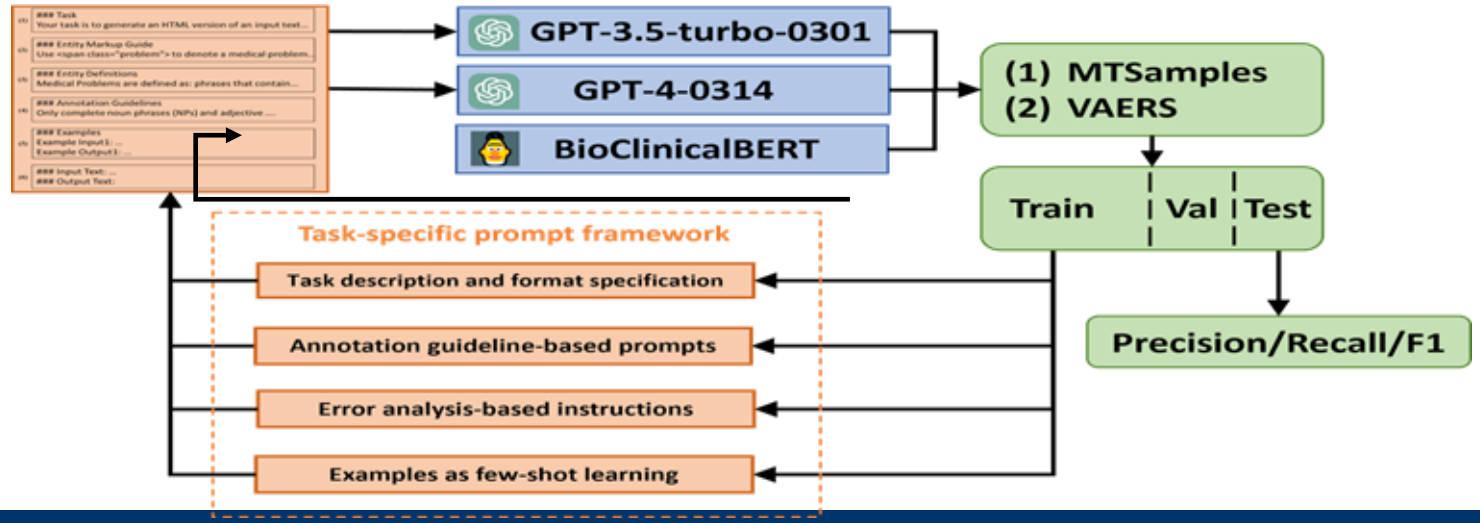
extraction
t extraction
cognition
ion
traction
xtraction
ced
on
ociated ADR
ng

# A Study of GPT Models for Clinical NER

- Objective: Investigate the potential of GPT-3.5 and GPT-4 models for clinical NER tasks and compare the performance with existing models (e.g., BioClinicalBERT)

- Datasets:
  - MTSamples (163 discharge summaries): Medical Problems, Treatments, and Tests

- Models:
  - GPT-3.5-turbo-0301
  - GPT-4-0314
  - BioClinicalBERT



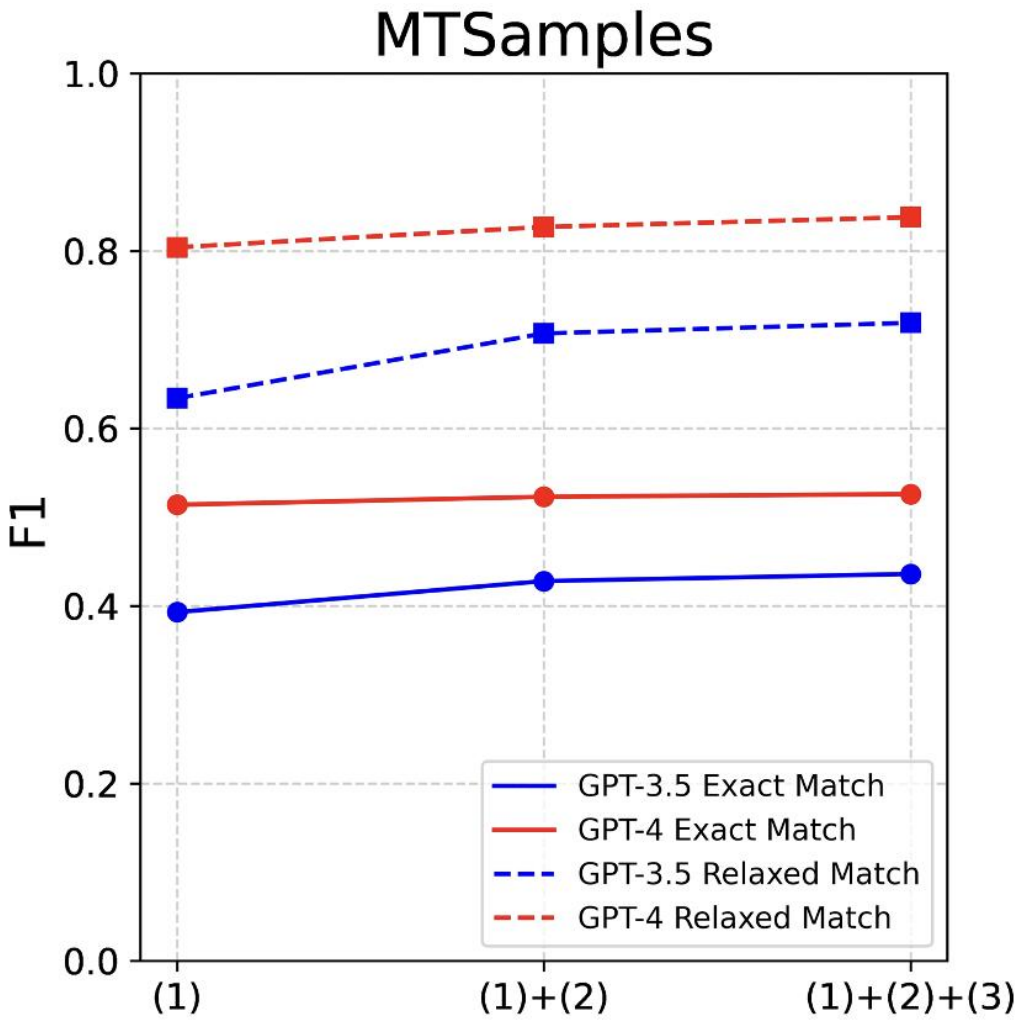

Hu Y et al. Improving Large Language Models for Clinical Named Entity Recognition via Prompt Engineering. JAMIA 2024

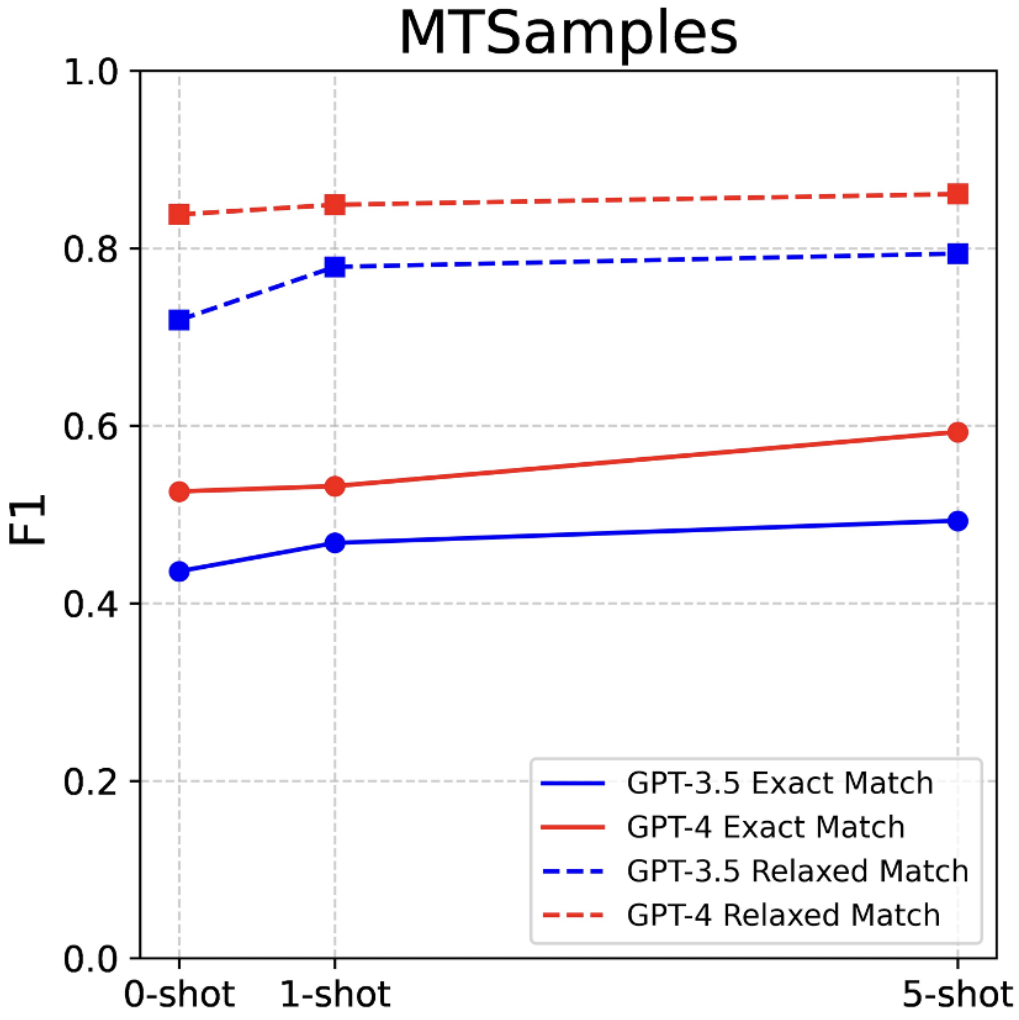
# Prompt Details

###Task:
Your task is to generate an HTML version of an input text, marking up specific entities. The entities to be identified are: 'medical problems', 'treatments', and 'tests'.
###Entity markup guide:
Use HTML  tags to highlight these entities. Each  should have a class attribute indicating the type of the entity. Use  to denote a medical problem, <span
…
###Entity definitions:
Medical Problems are defined as phrases that contain observations … Treatments are defined as …
###Annotation guidelines:
Only complete noun phrases (NPs) and adjective phrases (APs) should be marked. Terms that fit …
###Examples:
Example input1: At the time of admission , he denied fever , diaphoresis , …
Example output1: At the time of admission , he denied fever , diaphoresis …
###Input text: <add input sentence here>

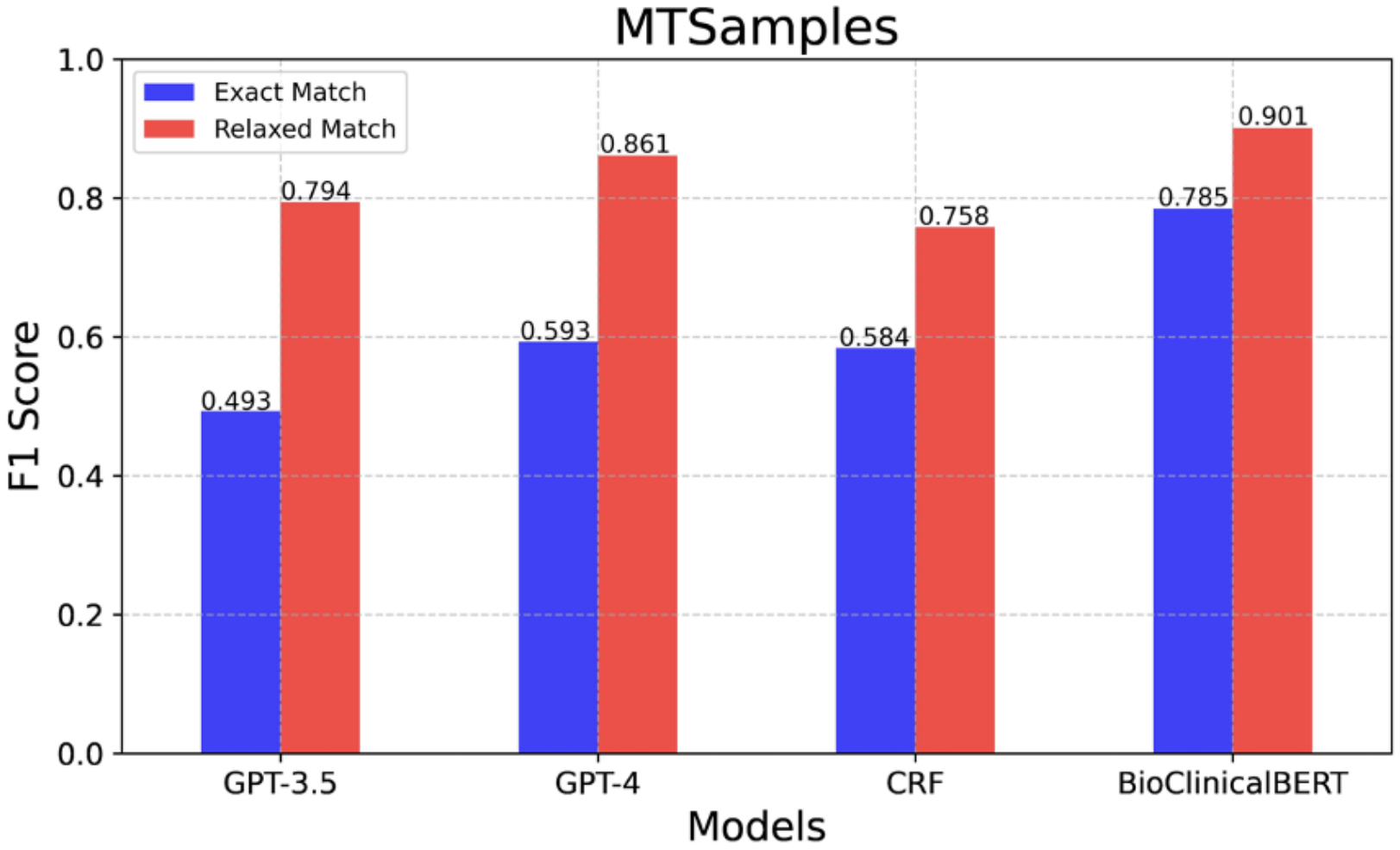
…

# Results – Prompt Strategies and Few-shot Learning



# Results – Comparing ML, DL, and LLMs




Hu Y et al. Improving Large Language Models for Clinical Named Entity Recognition via Prompt Engineering. JAMIA 2024

# Instruction Tuning LLaMA for Clinical NER

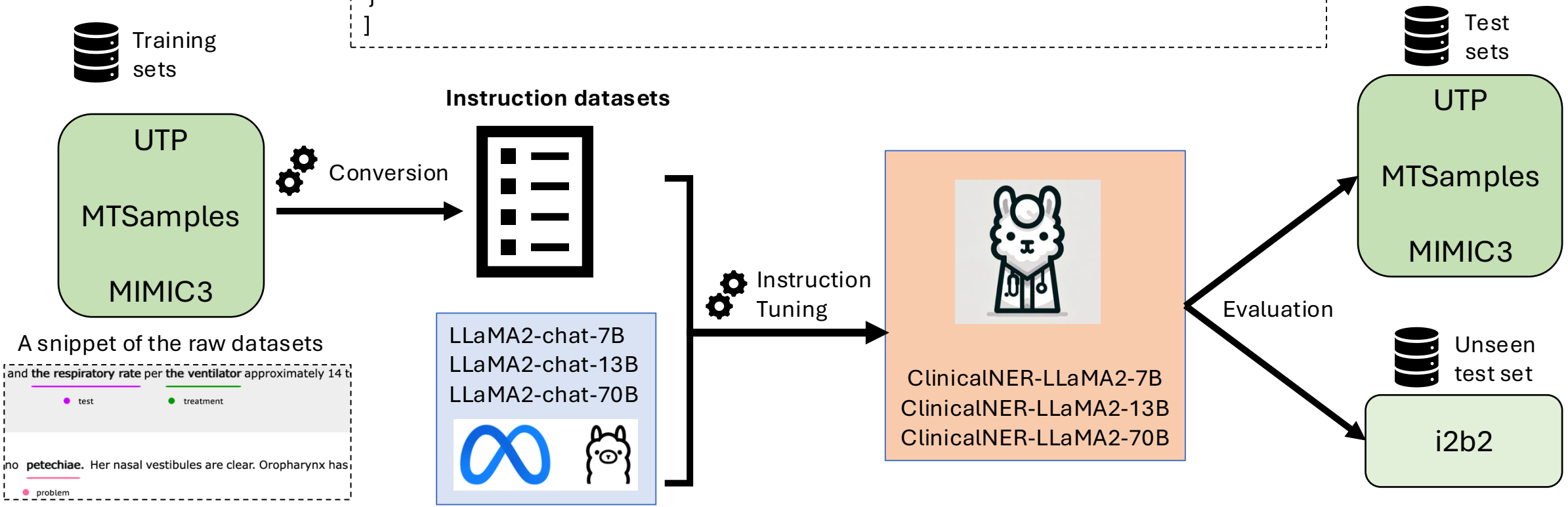
- Motivation:
  - NER is typically conceptualized as a sequence labeling task, whereas LLMs are optimized for text generation and reasoning tasks.
  - We propose an instruction-based learning paradigm that transforms biomedical NER from a sequence labeling task into a generation task
- Clinical NER Task: Extract problems, drugs, labs, and other treatments from clinical notes.
- Clinical NER datasets:
  - UT Physicians
  - MTSample
  - MIMIC-III
  - I2b2
- Models:
  - LLaMA2/3-7B, 13B, and 70B
  - BioClinicalBERT

| Source | Split | Number of documents |
|---|---|---|
| UTP | Train & Test | 1172 for train, 50 for test |
| MTSamples | Train & Test | 92 for train, 50 for test |
| MIMIC3 | Train & Test | 23 for train, 25 for test |
| i2b2 | Test | 50 for test |

# Methods

Examples of instructions

```
[
 {
   "instruction": "Given a sentence, extract medical problem entities from it by highlighting them with <mark> and </mark>. ...."
   "input": "Pt had a GI bleeding before admitting to ..."
   "output": "Pt had a <mark>GI bleeding</mark> before admitting to...
 }
]
```



# Results

- Performance

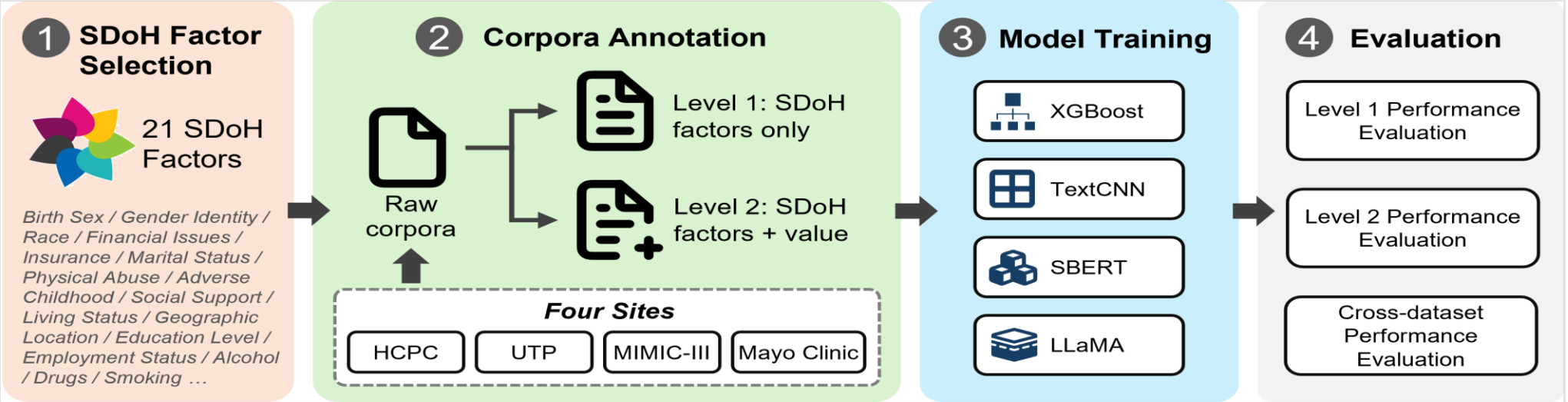
Hu Y, Zuo X, Zhou Y, Peng X, Huang J, Keloth VK, Zhang VJ, Weng RL, Shyr C, Chen Q, Jiang X, Roberts KE, Xu H. Information extraction from clinical notes: are we ready to switch to large language models? JAMIA 2026

| Datasets | LLAMA2-7b | | LLAMA2-13b | | LLAMA2-70b | | LLAMA3-8b | | LLAMA3-70b | | BERT | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | F1 (Exact) | F1 (Relax) | F1 (Exact) | F1 (Relax) | F1 (Exact) | F1 (Relax) | F1 (Exact) | F1 (Relax) | F1 (Exact) | F1 (Relax) | F1 (Exact) | F1 (Relax) |
| UTP | 0.929 | 0.963 | 0.932 | 0.964 | 0.931 | 0.964 | 0.929 | 0.965 | **0.932** | 0.964 | **0.921** | 0.957 |
| MTSample | 0.860 | 0.923 | 0.868 | 0.928 | 0.871 | 0.928 | 0.869 | 0.931 | **0.876** | 0.934 | **0.833** | 0.910 |
| MIMIC-III | 0.838 | 0.926 | 0.847 | 0.933 | 0.847 | 0.933 | 0.843 | 0.930 | **0.855** | 0.939 | **0.810** | 0.911 |
| i2b2 | 0.846 | 0.921 | 0.853 | 0.925 | 0.860 | 0.926 | 0.852 | 0.926 | **0.872** | 0.932 | **0.798** | 0.896 |

- Speed (i2b2)

| Speed | LLAMA2-7b | LLAMA2-13b | LLAMA2-70b | LLAMA3-8b | LLAMA3-70b | BERT |
|---|---|---|---|---|---|---|
| Sec/note | 8.0 | 13.9 | 35.1 | 6.7 | **35.7** | **1.6** |

# LLMs for Extracting Social Determinants of Health




| SDoH | Example |
|---|---|
| Geographic location | Pt born and raised in Rio Grande, Mexico.<br>Location-Raised<br>Location-Born |
| Sex, Race | The patient is an 80-year-old WF<br>Race-Caucasian Sex-Female |
| Employment status | Pt has been unemployed for past year …<br>EmploymentUnemployed-Present |
| Social support | … with her peers and has a good social support network<br>SocialSupport-Strong |
| Isolation | He has been very isolative and refuses to …<br>Isolation-Yes |
| Food insecurity | … said he couldn't afford to eat balanced meals.<br>FoodInsecure-Yes |

| Dataset | XGBoost | TextCNN | Sent. BERT | LLaMA |
|---|---|---|---|---|
| SDoH factors only | | | | |
| HCPC | 0.907/0.803/0.851 | 0.895/0.781/0.834 | 0.880/0.858/0.869 | 0.941/0.913/**0.927** |
| UTP | 0.982/0.935/0.958 | 0.980/0.927/0.952 | 0.979/0.948/0.963 | 0.990/0.979/**0.984** |
| MIMIC-III | 0.887/0.780/0.830 | 0.841/0.732/0.782 | 0.890/0.821/0.854 | 0.934/0.840/**0.883** |
| Mayo | 0.852/0.799/0.825 | 0.823/0.734/0.775 | 0.892/0.672/0.766 | 0.953/0.938/**0.945** |
| SDoH factors + values | | | | |
| HCPC | 0.821/0.690/0.750 | 0.824/0.569/0.673 | 0.826/0.751/0.786 | 0.903/0.869/**0.886** |
| UTP | 0.946/0.880/0.912 | 0.889/0.815/0.850 | 0.957/0.882/0.918 | 0.982/0.932/**0.956** |
| MIMIC-III | 0.802/0.649/0.717 | 0.737/0.430/0.543 | 0.805/0.674/0.734 | 0.877/0.801/**0.837** |
| Mayo | 0.795/0.711/0.750 | 0.770/0.572/0.656 | 0.878/0.629/0.732 | 0.935/0.901/**0.918** |

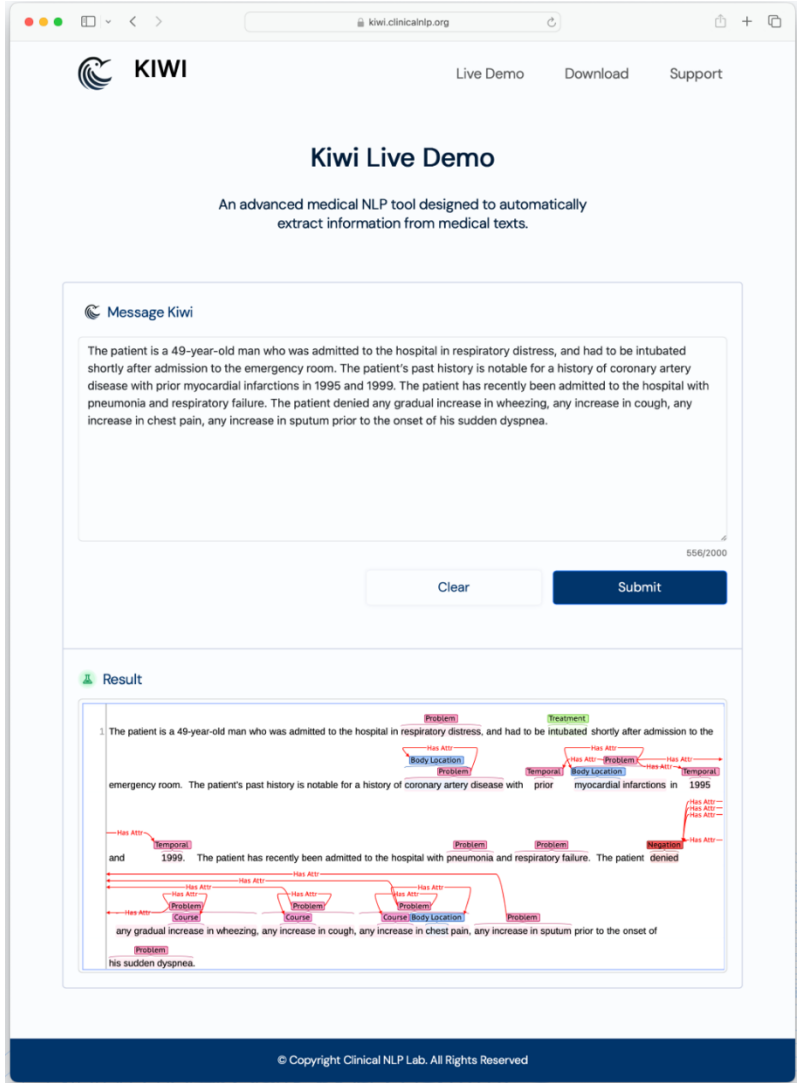
Keloth, V.K., Selek, S., Chen, Q. et al. Social determinants of health extraction from clinical notes across institutions using large language models. npj Digit. Med. 8, 287 (2025)

# KIWI - An LLM-based Clinical Information Extraction System

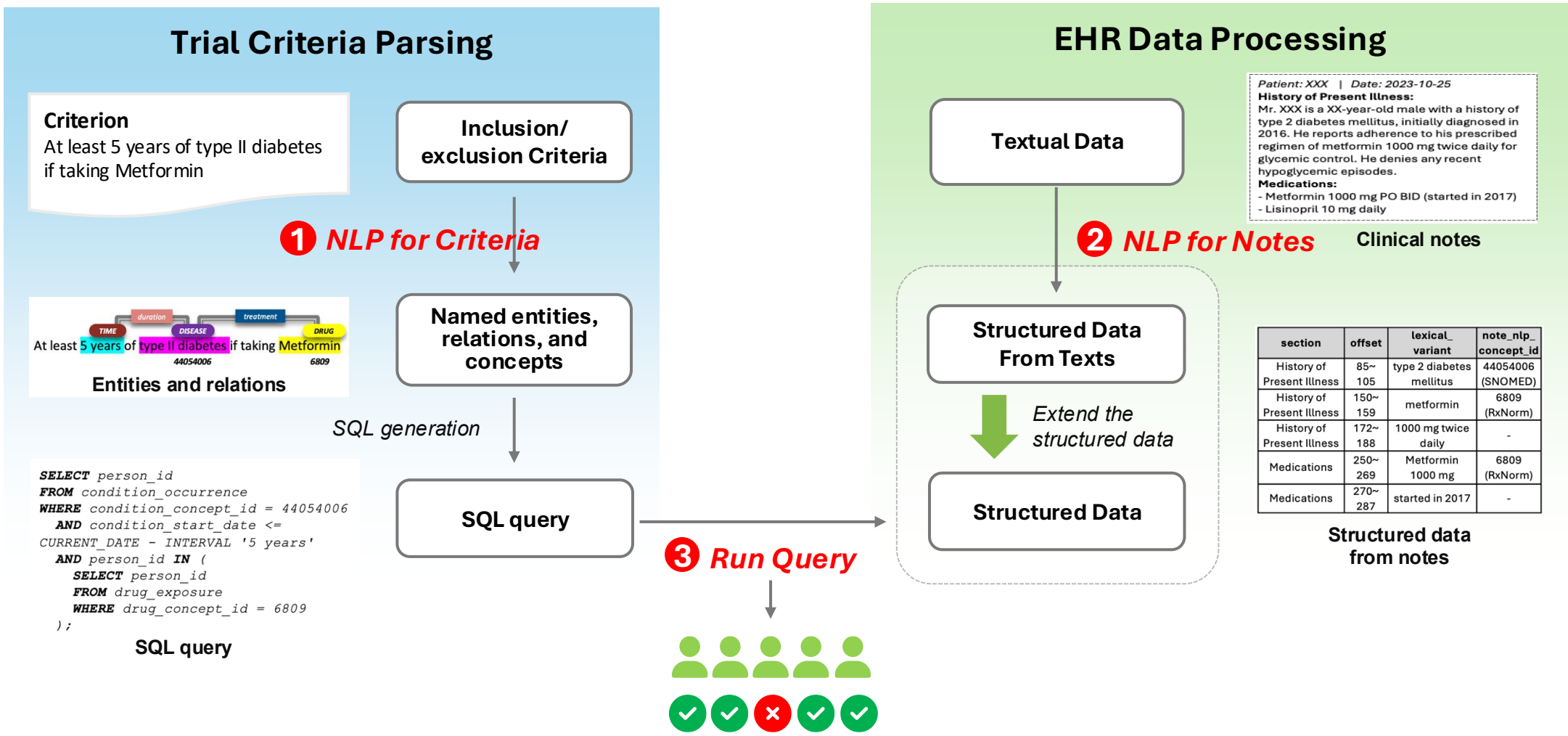
KIWI

- Provide both LLaMA and BERT models for clinical information extraction
- Offer general and disease specific pipelines
- Available as a docker image for easy installation

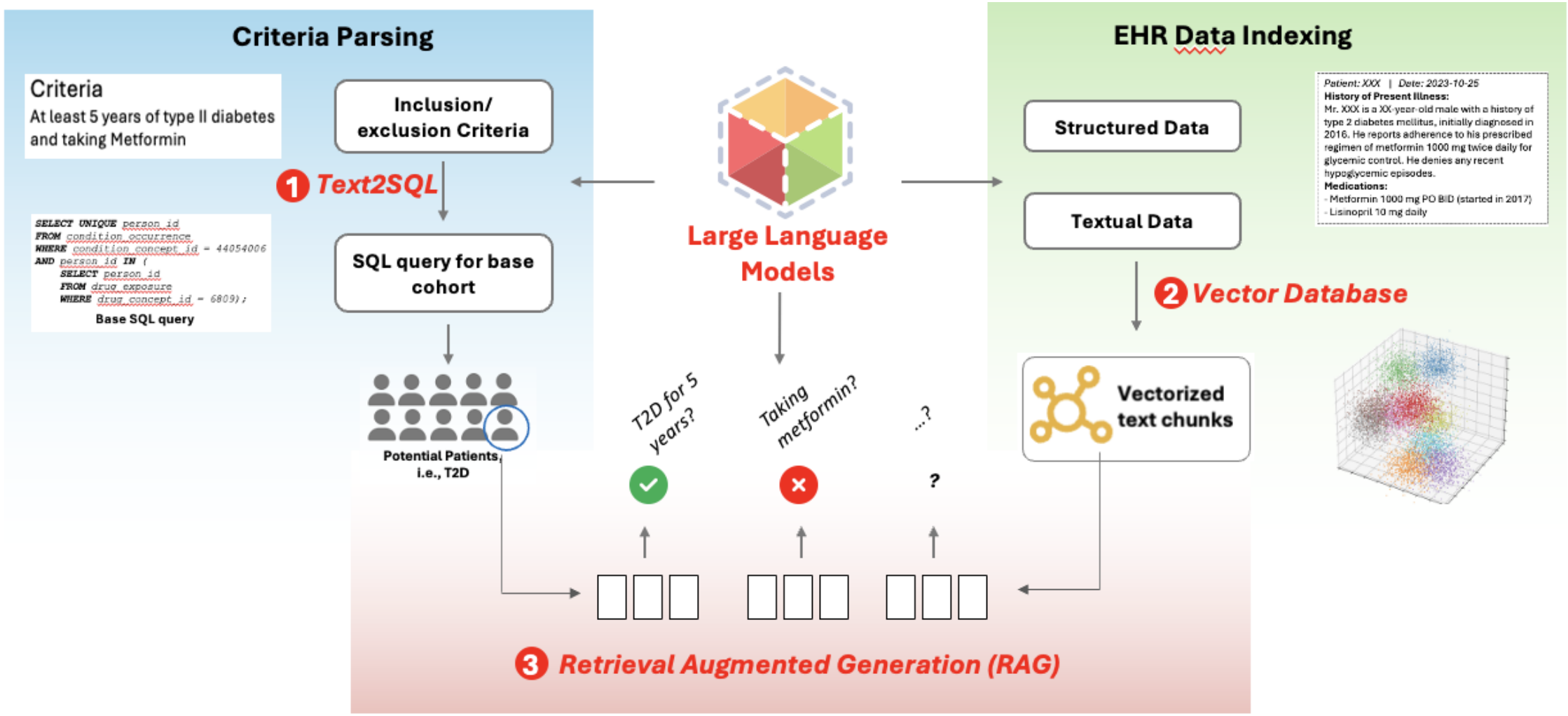
https://kiwi.clinicalnlp.org/



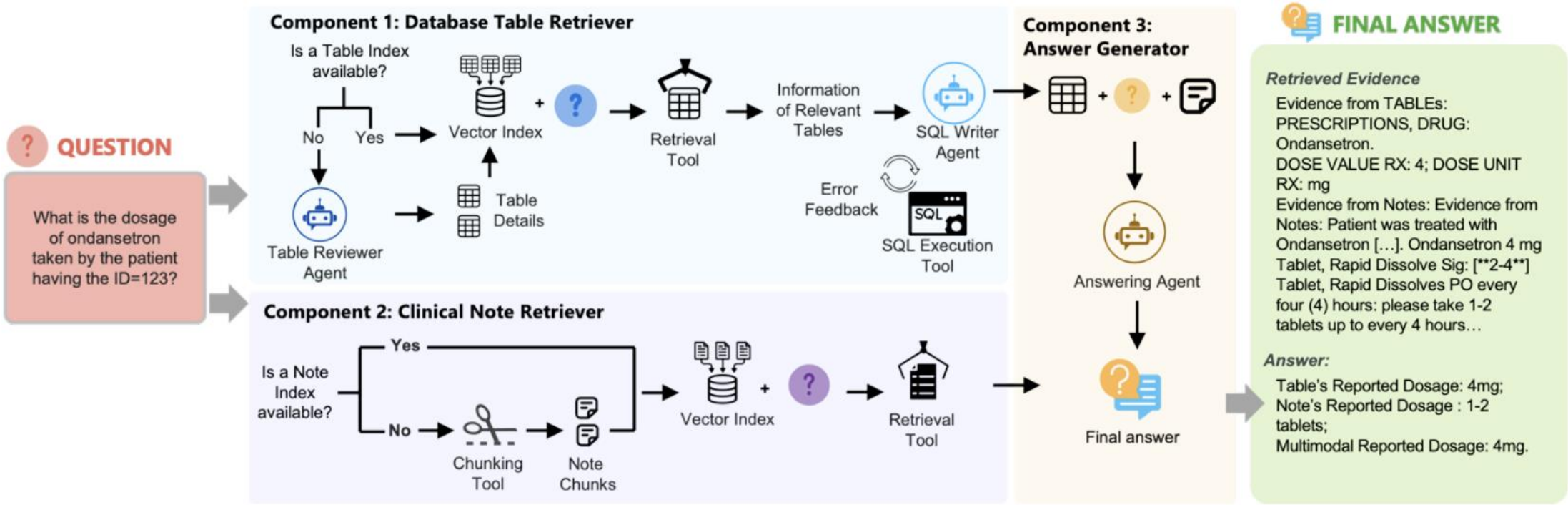
# Current NLP-based Phenotyping Approaches



# A New Workflow for RAG-based EHR Phenotyping



# RAG-based Phenotyping Architecture




## Framework

(1) Structured data retrieval – text2sql
(2) Unstructured notes retrieval – vector database
(3) Answer generation by LLM

# Results

| Setting | Backbone Model | Accuracy |
|---|---|---|
| **MultimodalQA** | **ClinicalBERT** | **0.6108** |
| **LLM only** | **GPT-4o** | **0.3993** |
| | **Llama3-70B-Instruct** | **0.1021** |
| | **Llama3.1-70B-Instruct** | **0.3686** |
| **RAG-EHR (w/o training the answer generator)** | **GPT-4o** | **0.5104** |
| | **Llama3-70B-Instruct** | **0.3378** |
| | **Llama3.1-70B-Instruct** | **0.4702** |
| **RAG-EHR (fine-tuned answer generator)** | **Llama3-70B-Instruct** | **0.7978** |
| | **Llama3.1-70B-Instruct** | **0.8132** |

| Dataset | Structured QA | Unstructured QA | Multimodal QA |
|---|---|---|---|
| DrugEHRQA* | 5,559 | 5,693 | 27,795 |

*Bardhan, Jayetri, et al. "Drugehrqa: A question answering dataset on structured and unstructured electronic health records for medicine related queries." *arXiv preprint arXiv:2205.01290* (2022).

Qian L et al. EHRNavigator: A Multi-Agent System for Patient-Level Clinical Question Answering over Heterogeneous Electronic Health Records. https://arxiv.org/abs/2601.10020

# Disease Prediction Models Based on Large Real-World Data

nature

Article

## Learning the natural history of human disease with generative transformers

https://doi.org/10.1038/s41586-025-09529-3

Received: 18 May 2024

Accepted: 13 August 2025

Published online: 17 September 2025

Open access

Check for updates

Artem Shmatko[1,2,3,13], Alexander Wolfgang Jung[2,4,5,6,13], Kumar Gaurav[2,13], Søren Brunak[4,7], Laust Hvas Mortensen[5,7,8], Ewan Birney[2] ✉, Tom Fitzgerald[2] ✉ & Moritz Gerstung[1,2,9,10,11,12] ✉

Decision-making in healthcare relies on understanding patients' past and current health states to predict and, ultimately, change their future course[1–3]. Artificial intelligence (AI) methods promise to aid this task by learning patterns of disease progression from large corpora of health records[4,5]. However, their potential has not been fully investigated at scale. Here we modify the GPT[6] (generative pretrained transformer) architecture to model the progression and competing nature of human diseases. We train this model, Delphi-2M, on data from 0.4 million UK Biobank participants and validate it using external data from 1.9 million Danish individuals with no change in parameters. Delphi-2M predicts the rates of more than 1,000 diseases, conditional on each individual's past disease history, with accuracy comparable to that of existing single-disease models. Delphi-2M's generative nature also enables sampling of synthetic future health trajectories, providing meaningful estimates of potential disease burden for up to 20 years, and enabling the training of AI models that have never seen actual data. Explainable AI methods[7] provide insights into Delphi-2M's predictions, revealing clusters of co-morbidities within and across disease chapters and their time-dependent consequences on future health, but also highlight biases learnt from training data. In summary, transformer-based models appear to be well suited for predictive and generative health-related tasks, are applicable to population-scale datasets and provide insights into temporal dependencies between disease events, potentially improving the understanding of personalized health risks and informing precision medicine approaches.

## Generative Medical Event Models Improve with Scale

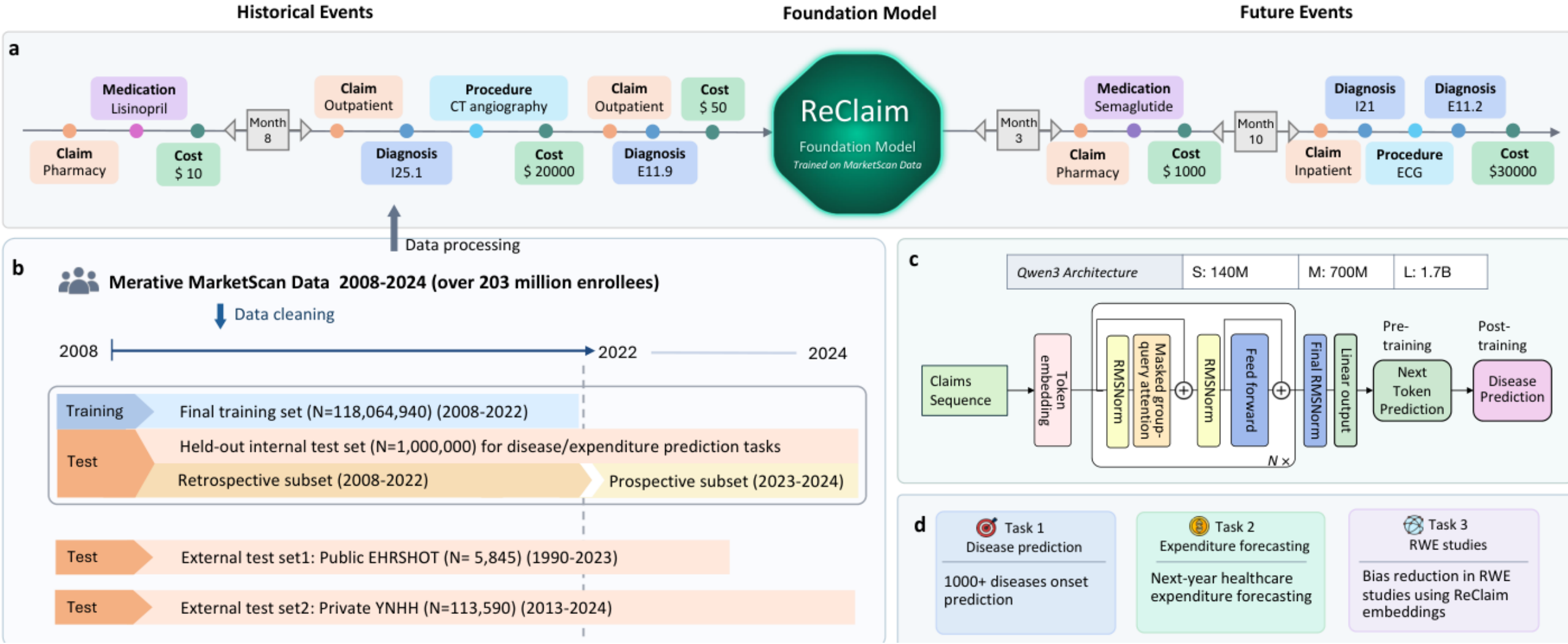
Shane Waxler*[1], Paul Blazek*[1], Davis White*[1], Daniel Sneider*[1], Kevin Chung[1], Mani Nagarathnam[1], Patrick Williams[1], Hank Voeller[1], Karen Wong[1], Matthew Swanhorst[1], Sheng Zhang[2], Naoto Usuyama[2], Cliff Wong[2], Tristan Naumann[2], Hoifung Poon[2], Andrew Loza[3], Daniella Meeker[3, 4], Seth Hain[1], and Rahul Shah†[1]

[1]Epic Systems
[2]Microsoft Research
[3]Yale School of Medicine
[4]Cosmos Governing Council

**Abstract**

Realizing personalized medicine at scale calls for methods that distill insights from longitudinal patient journeys, which can be viewed as a sequence of medical events. Foundation models pretrained on large-scale medical event data represent a promising direction for scaling real-world evidence generation and generalizing to diverse downstream tasks. Using Epic Cosmos, a dataset with medical events from de-identified longitudinal health records for 16.3 billion encounters over 300 million unique patient records from 310 health systems, we introduce the Curiosity models, a family of decoder-only transformer models pretrained on 118 million patients representing 115 billion discrete medical events (151 billion tokens). We present the largest scaling-law study of medical event data, establishing a methodology for pretraining and revealing power-law scaling relationships for compute, tokens, and model size. Consequently, we pretrained a series of compute-optimal models with up to 1 billion parameters. Conditioned on a patient's real-world history, Curiosity autoregressively predicts the next medical event to simulate patient health timelines. We studied 78 real-world tasks, including diagnosis prediction, disease prognosis, and healthcare operations. Remarkably for a foundation model with generic pretraining and simulation-based inference, Curiosity generally outperformed or matched task-specific supervised models on these tasks, without requiring task-specific fine-tuning or few-shot examples. Curiosity's predictive power consistently improves as the model and pretraining scale. Our results show that Curiosity, a generative medical event foundation model, can effectively capture complex clinical dynamics, providing an extensible and generalizable framework to support clinical decision-making, streamline healthcare operations, and improve patient outcomes.

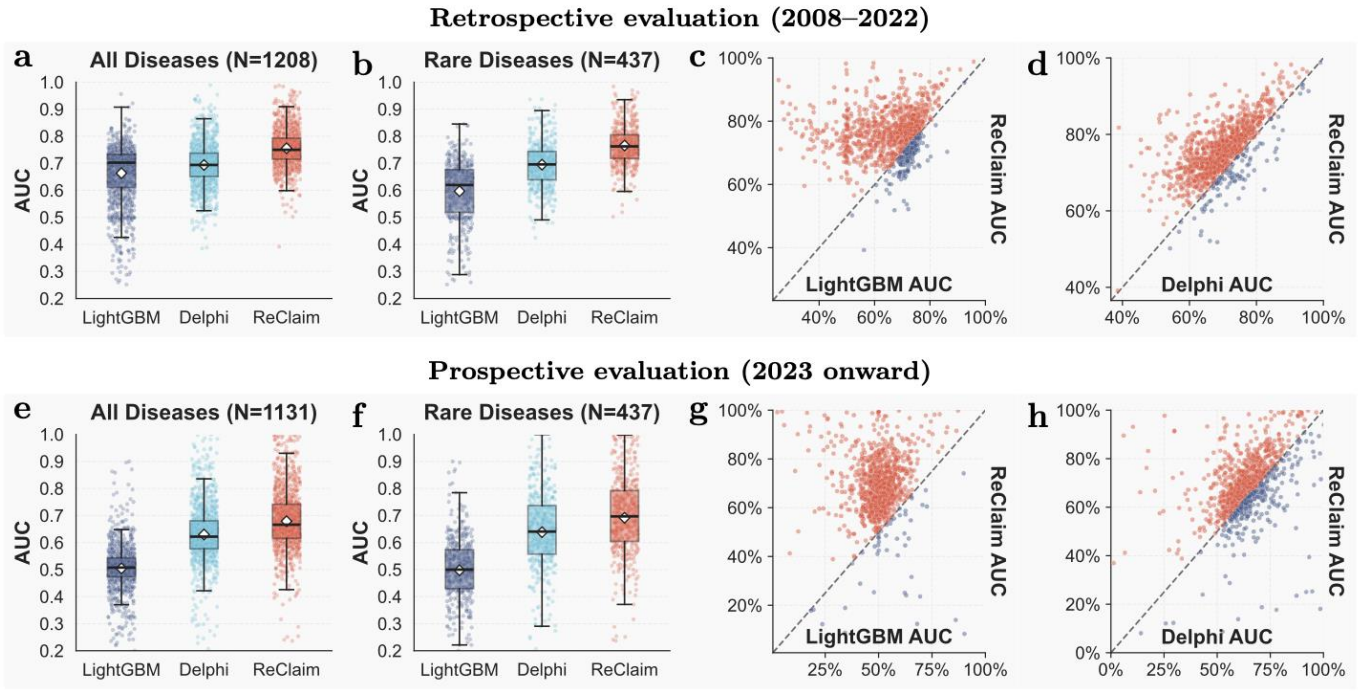
# ReClaim Overview




Ma F et al. Foundation Models to Unlock Real-World Evidence from Nationwide Medical Claims. https://arxiv.org/abs/2605.02740v2

# Main result: stronger disease onset prediction

Across both retrospective and prospective evaluations, ReClaim outperforms LightGBM and Delphi on most disease endpoints

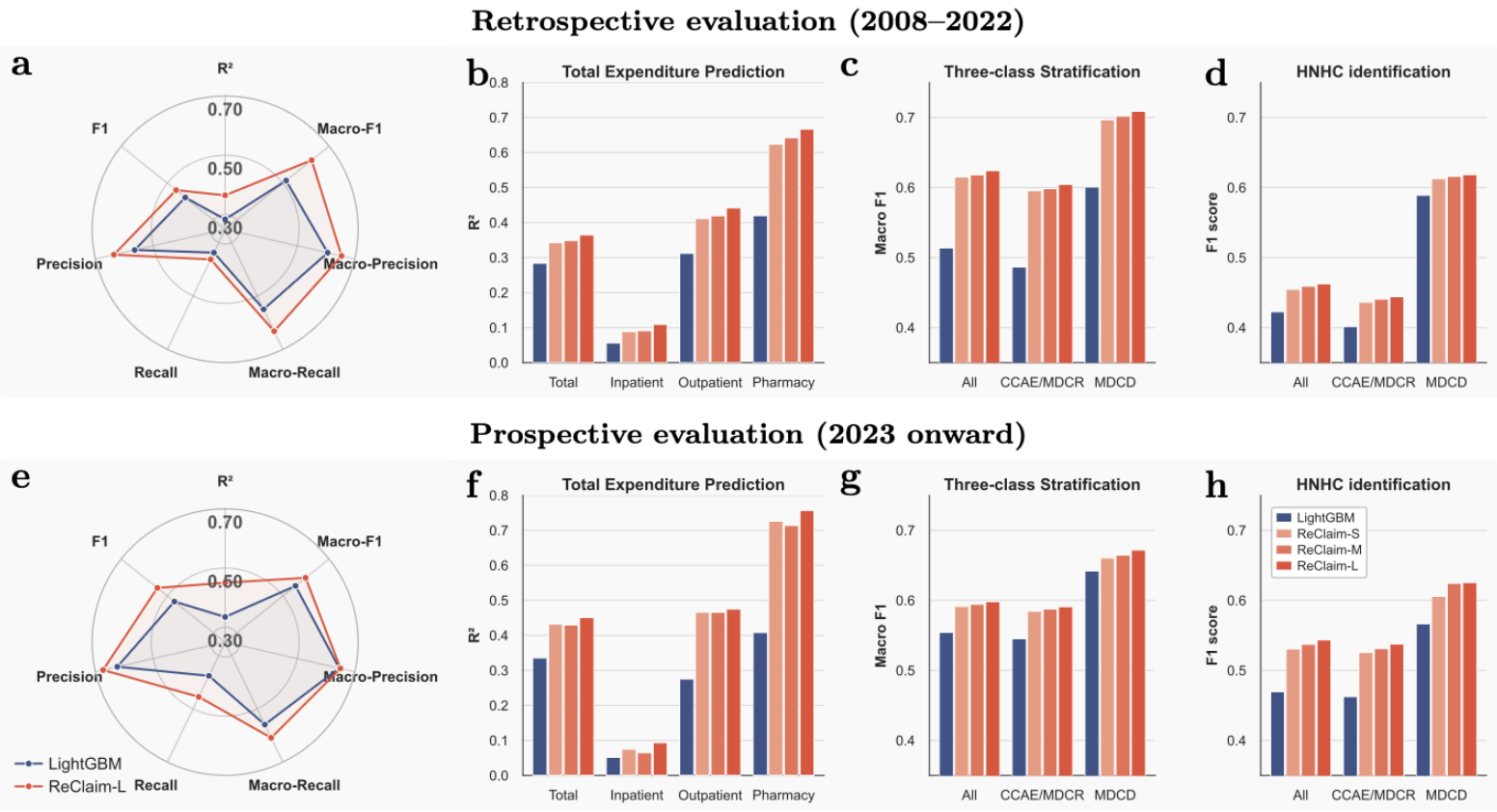



## Key findings

- Retrospective mean AUC: ReClaim 75.57%, Delphi 69.36%, LightGBM 66.34%
- Prospective mean AUC: ReClaim 67.89%, Delphi 62.97%, LightGBM 50.44%
- Largest gains are seen for rare diseases
- ReClaim beats LightGBM on 79.9% of retrospective endpoints and 95.8% of prospective endpoints
- ReClaim beats Delphi on 92.0% of retrospective endpoints and 77.0% of prospective endpoints

# Beyond diagnosis: ReClaim also improves expenditure forecasting

The model is not only predictive for diseases; it also captures utilization patterns linked to future spending

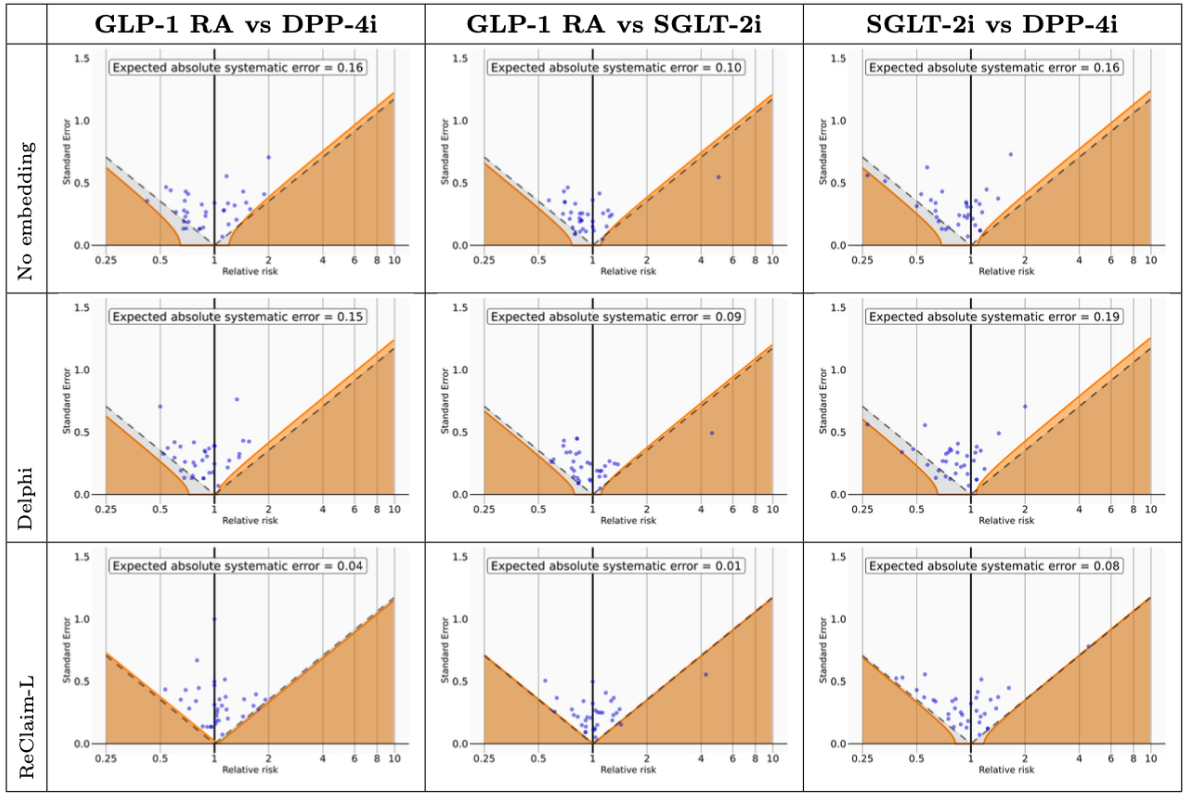



- Retrospective total expenditure $R^2$ improves from 0.2835 to 0.365 over LightGBM
- Prospective total expenditure $R^2$ improves from 0.335 to 0.451
- ReClaim also improves 3-class expenditure stratification and high-need/high-cost identification
- The paper links clinical risk with financial burden in one shared representation

The model's utility goes beyond diagnosis and into healthcare cost estimation.

# RWE generation: learned embeddings reduce residual bias

ReClaim-derived representations improve confounding adjustment in an observational RWE setting using an embedding-augmented target trial emulation comparing three antidiabetic drug classes for psychiatric outcomes.




## Result pattern

- Study compares GLP-1 RAs, SGLT-2is, and DPP-4is in 7,246 eligible individuals
- For Negative Control Outcomes (NCO), points closer to the null indicate lower residual systematic bias after matching
- ReClaim embeddings achieve the lowest EASE in all three pairwise treatment comparisons
- Examples: 0.16 → 0.04, 0.10 → 0.01, and 0.16 → 0.08 versus no embeddings

Learned representations by ReClaim are useful for real-world evidence generation.

# Ongoing studies

## Predictors of diagnostic transition from major depressive disorder to bipolar disorder: a retrospective observational network study (Psychiatry)

Lead: Ming Huang (ming.huang@austin.utexas.edu); Christophe Lambert (cglambert@salud.unm.edu)

Meeting: 3rd Friday of the month – 4 to 5 PM ET

Interested in joining: Marty Alvarez (Marta.Alvarez@tuftsmedicine.org)

## Characterizing the Anticancer Treatment Trajectory and Pattern in Patients Receiving Chemotherapy for Cancer (Oncology)

Lead: Liwei Wang (Liwei.Wang@austin.utexas.edu); Georgina Kennedy (georgina.kennedy@unsw.edu.au)

Georgina building an Oncology NLP Endpoint catalogue and NLP resource library (Reusable prompts, extractors, schemas, and tools for clinical oncology NLP)

# Other ongoing studies

## Social Determinants of Health and Treatment Outcomes in Type 2 Diabetes: A Multi-Site Analysis of LEGEND-T2DM study

Select a Target-Comparator-Outcome

GLP1-RA – Sulfonylureas – 3-point MACE

Repeat the analysis (multisite) stratified by the 3 SDoH factors (Social Support, Housing instability, Substance use)

Structured data (e.g., ICD-10 SDoH codes)

Structured + semi-structured (e.g., flowsheets)

Structured + semi-structured + unstructured (e.g., NLP-derived from clinical notes)

# OHDSI NLP WG meetings

## Monthly meeting

Second Wednesday of the month – 2 PM – 3 PM ET

NLP Workgroup @ OHDSI Annual Symposium

Thursday Oct 22 – 1:30 – 3:30 PM

Agenda: Research updates, Panel

# Thank you!
# Questions?

hua.xu@yale.edu