Large Language Models for Clinical Information Extraction and Beyond

Healthcare

NLP

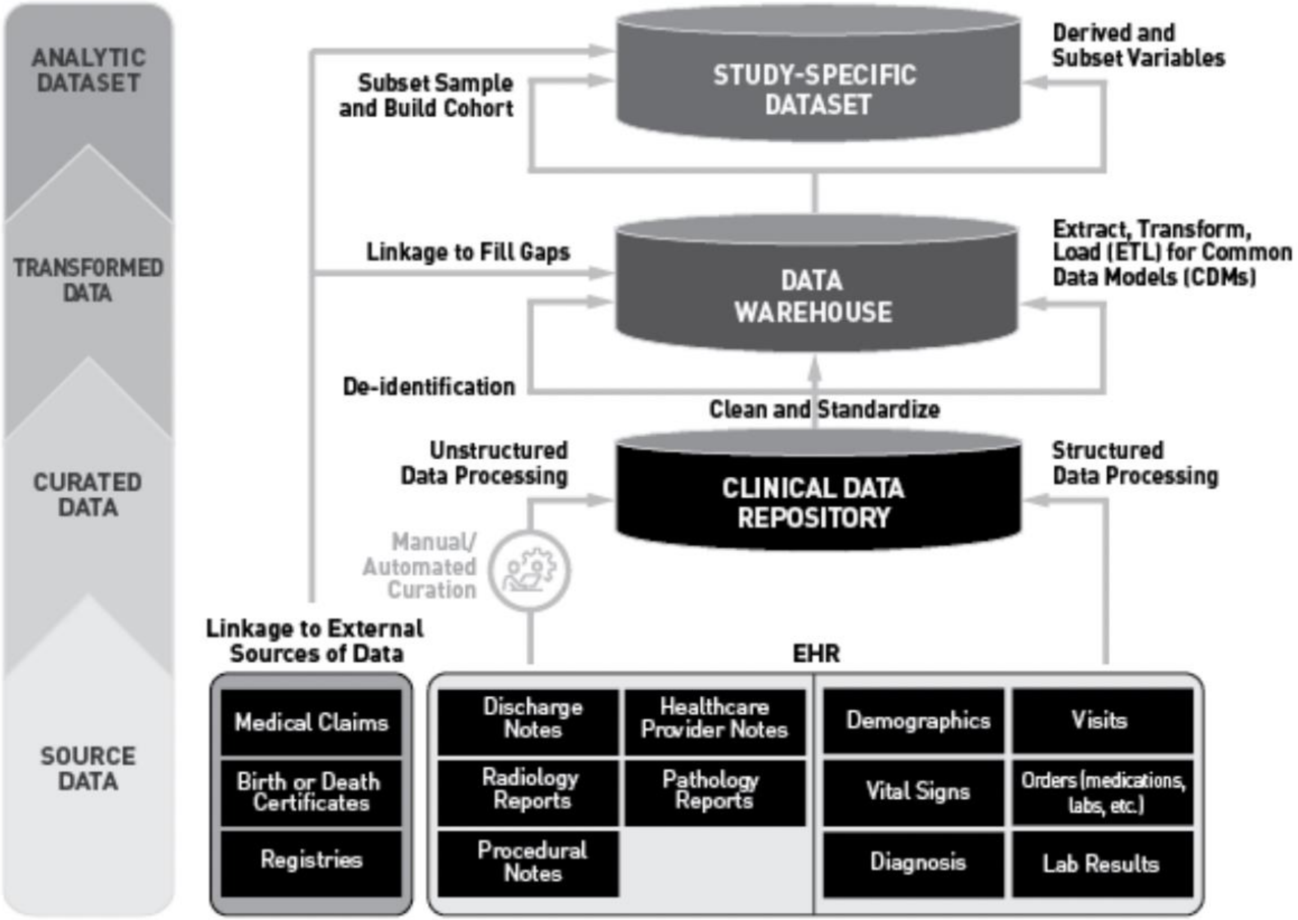
Hua Xu PhD

December 6th, 2024

Electronic Health Records (EHRs) for Clinical Research

- EHRs (and linked data) becomes an enabling resource for clinical and translational research

RWD Guidance2024 *



* Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision-Making for Drug and Biological Products Guidance for Industry. FDA July 2024

Textual Documents in EHRs

Admit 10/23
Medical History: 71 yo woman h/o DM, HTN, Dilated CM/CHF, Afib s/p embolic event, chronic diarrhea, admitted with SOB. CXR pulm edema. Rx’d Lasix.
Social History: PT isolates to self in her apartment.
All: none
Meds Lasix 40mg IVP bid, ASA, Coumadin 5, Prinivil 10, glucophage 850 bid, glipizide 10 bid, immodium prn

Medical History

Social History

Treatment Response

More details …

Information Extraction (IE) from Clinical Notes

Named Entity Recognition - NER

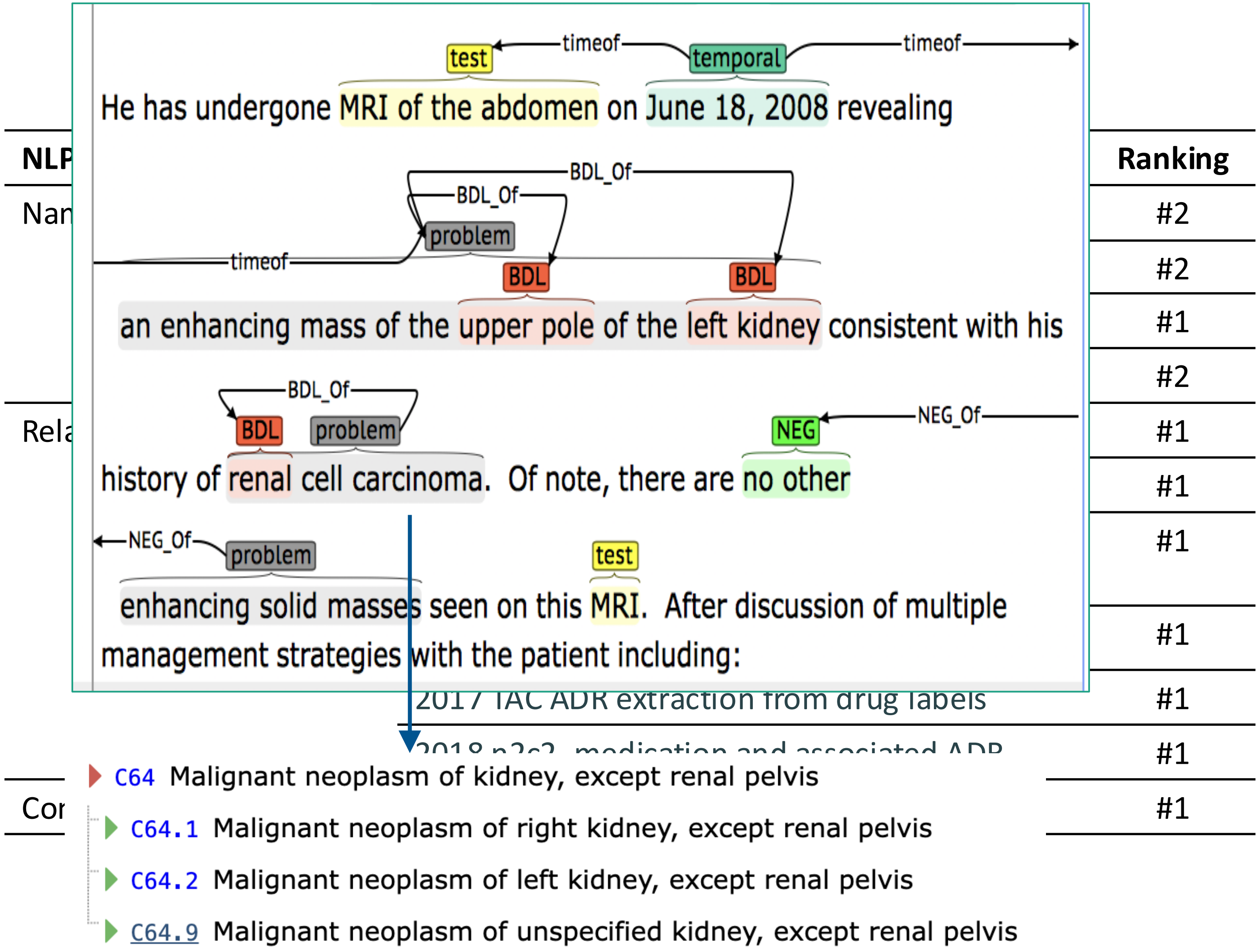
Recognize boundary and type of an entity mention in the text

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



OHDSI NLP Working Group

- A multi-stakeholder, interdisciplinary collaborative to bring out the value of health data through large-scale analytics

OHDSI Collaborators:
- >2,770 researchers in academia, industry and government
- >21 countries

OHDSI Data Network:
- >133 databases from 18 countries
- 1.9 billion patient records (duplicates)
- ~369 million non-US patients

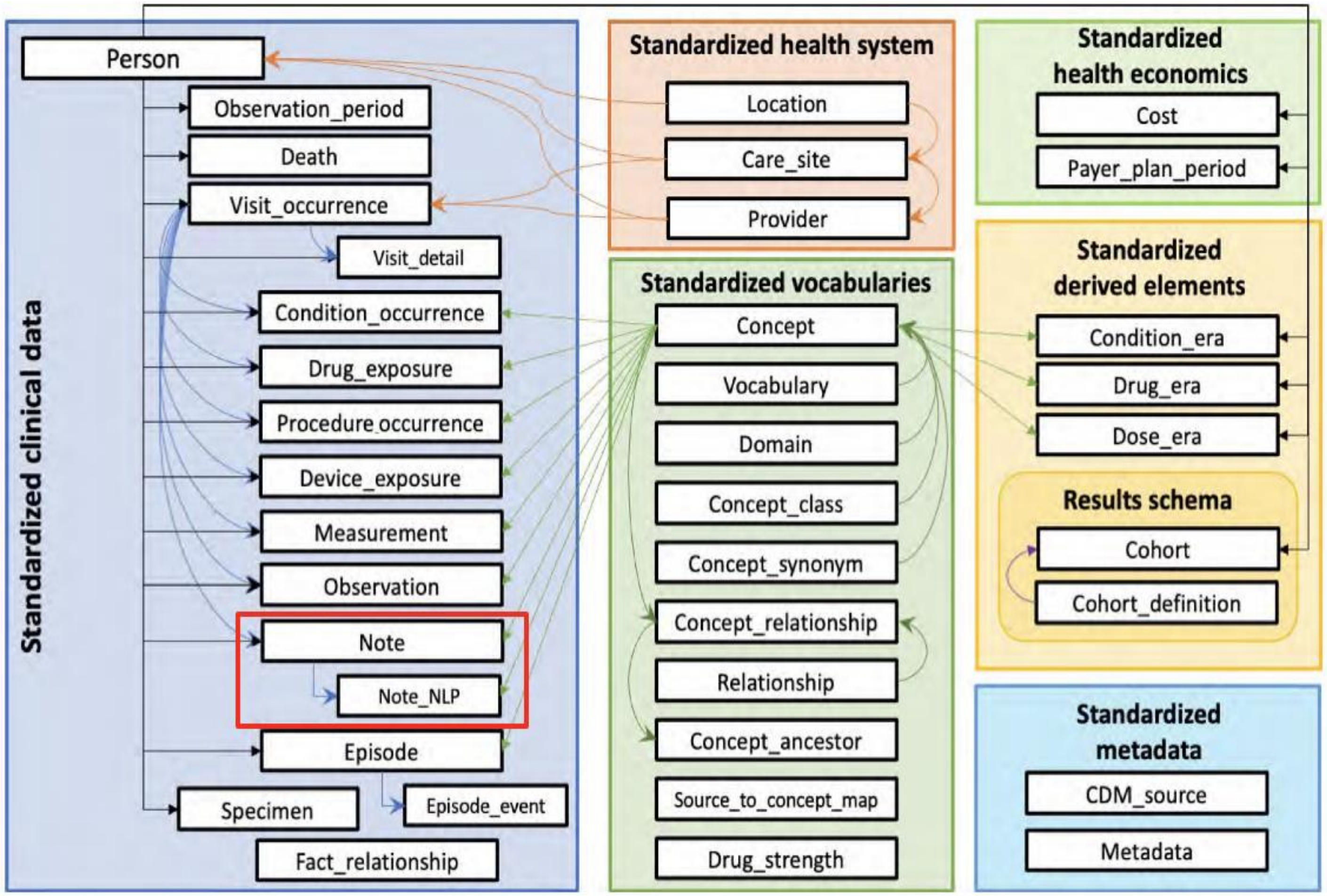
- OHDSI NLP Workgroup - stablished in 2015, with the goal to promote the use of textual data in EHRs for real world studies

  - 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

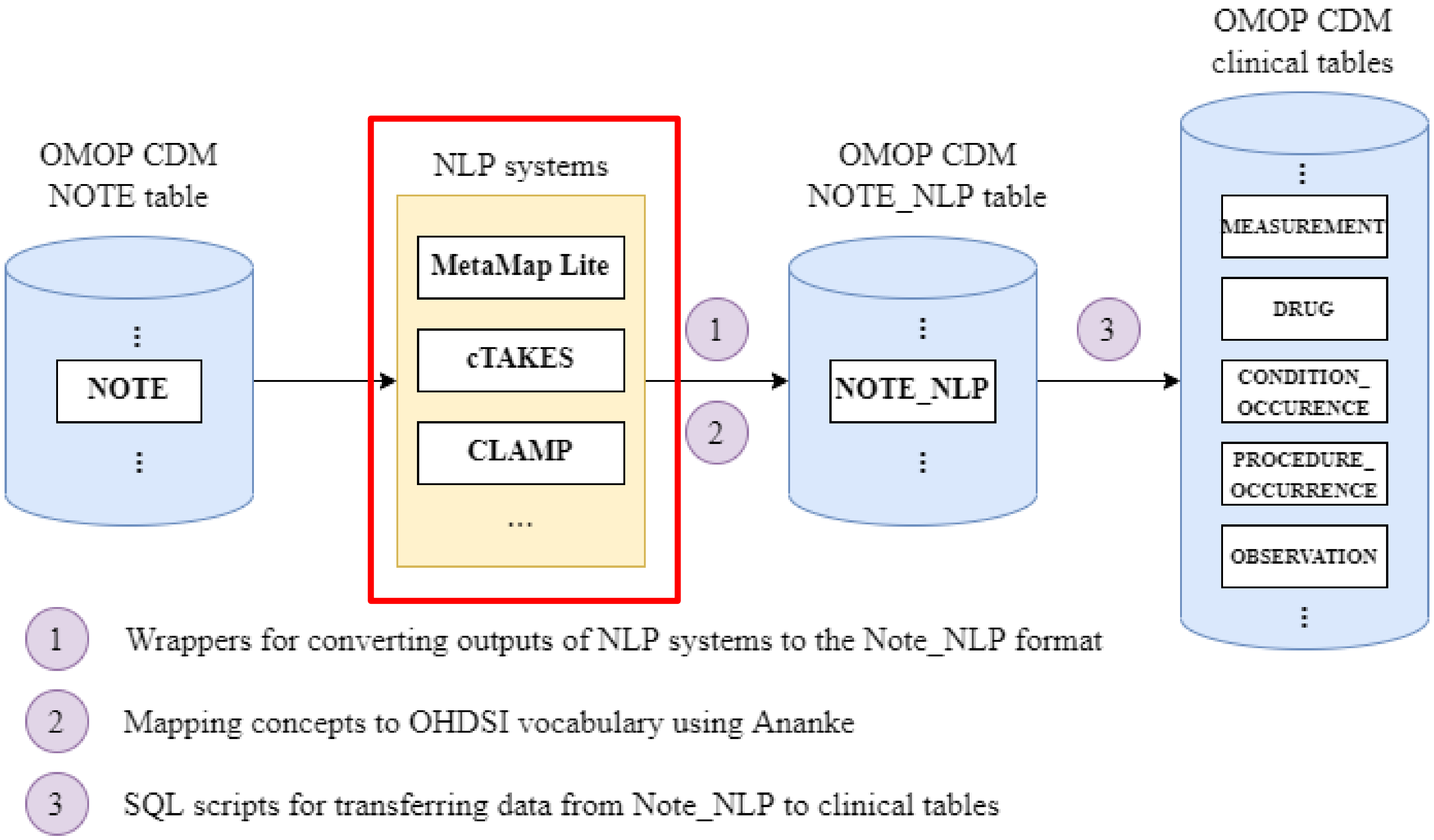
Representing Clinical Texts and NLP Outputs in OMOP CDM

- 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)



NLP Workflow for Textual Data in CDM

- Run 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.

Clinical IE – Machine Learning and Deep Learning-based Approaches

- Task: 2010 i2b2 challenge – entity recognition for problem, treatment, and test in discharge summaries

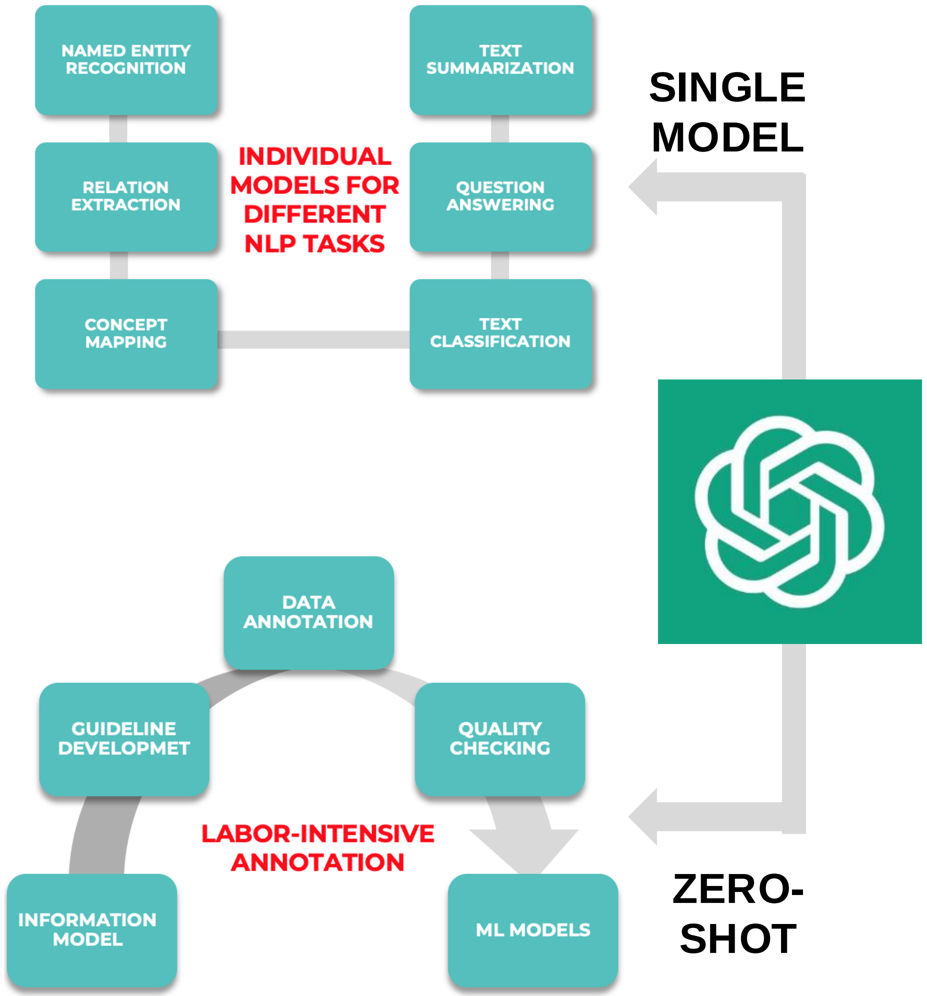
Algorithms	Feature	F1
CRFs (Jiang et al., 2010) (#2 in challenge)	Bag of words	77.33
	Optimized features	**83.60**
Semi-Markov (deBruijn B, et al., 2010) (#1 in challenge)	Optimized features + Brown clustering	**85.23**
SSVMs (Tang et al., 2014)	Optimized features + Brown clustering + Random indexing	85.82
CNN (Wu et al., 2015)	Word embedding	82.77
Bi-LSTM-CRF (Wu et al., 2017)	Word embedding	85.91
BERT (Si et al., 2020)	Pre-trained language model - BERT, fine tuned on clinical text	**90.25**

Clinical IE using GPTs

HU

Please extract dosage and frequency for each drug mentioned in the following text: "Meds: Lasix 40mg IVP bid, ASA, Coumadin 5, Prinivil 10, glucophage 850 bid, glipizide 10 bid, immodium prn"

Drug	Dosage	Frequency
Lasix	40mg	IVP bid
ASA	Not specified	Not specified
Coumadin	5	Not specified
Prinivil	10	Not specified
Glucophage	850	bid
Glipizide	10	bid
Imodium	Not specified	prn

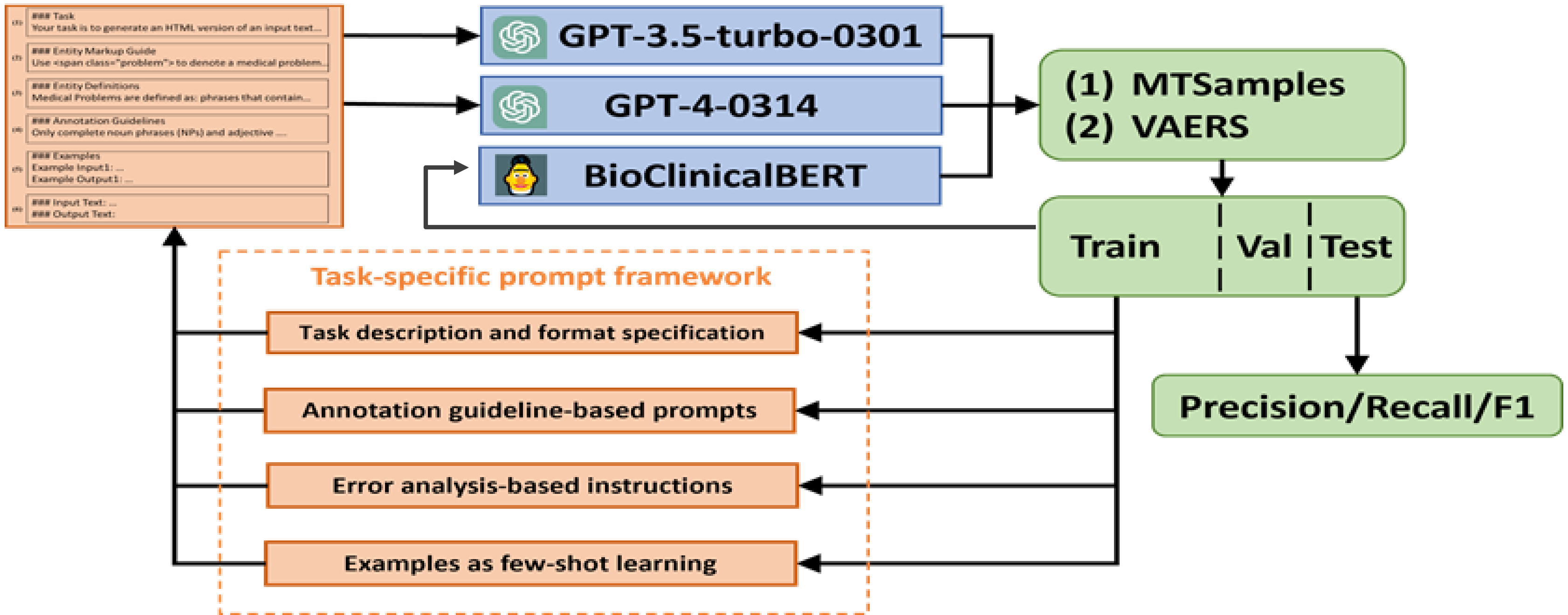


Clinical IE #1 – Prompt Engineering

- 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)
  - Vaccine adverse event reporting system – VAERS (91 safety reports)

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



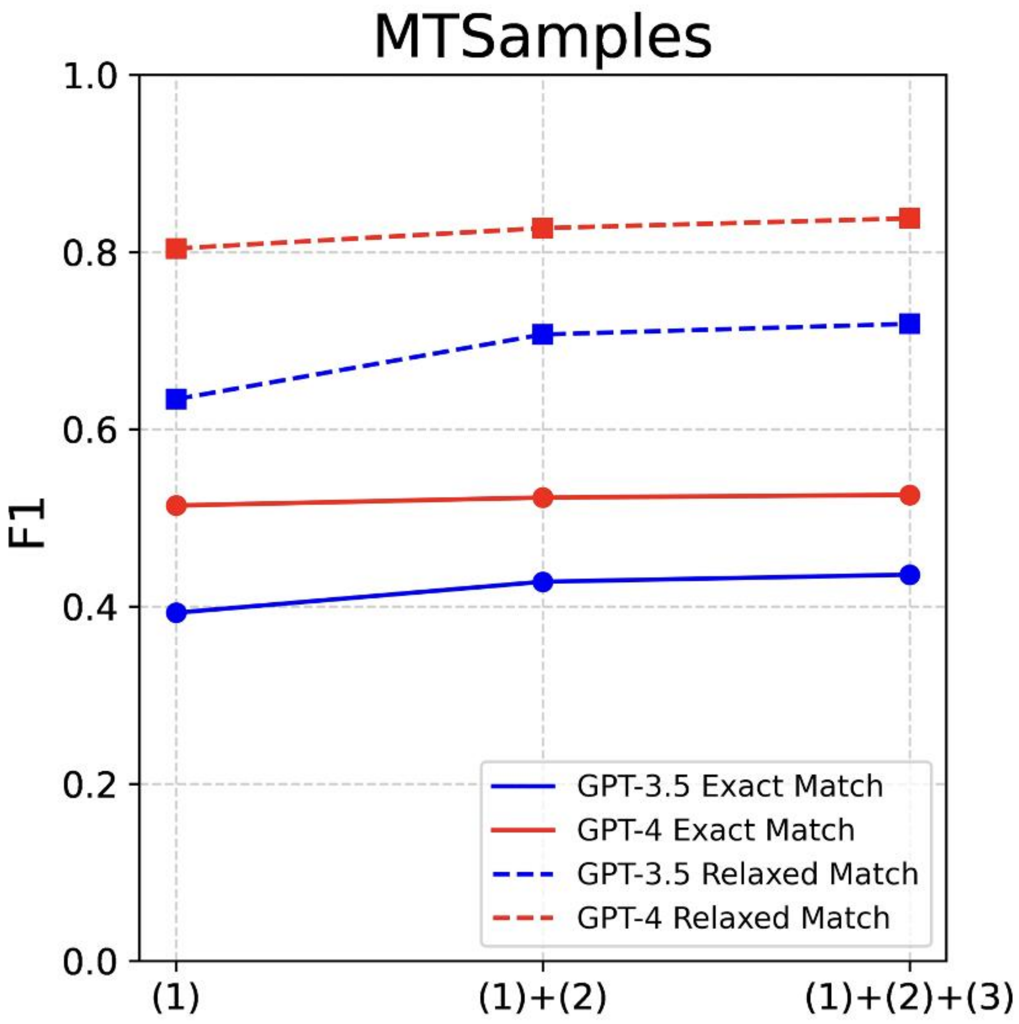
https://arxiv.org/abs/2303.16416

Prompt Details

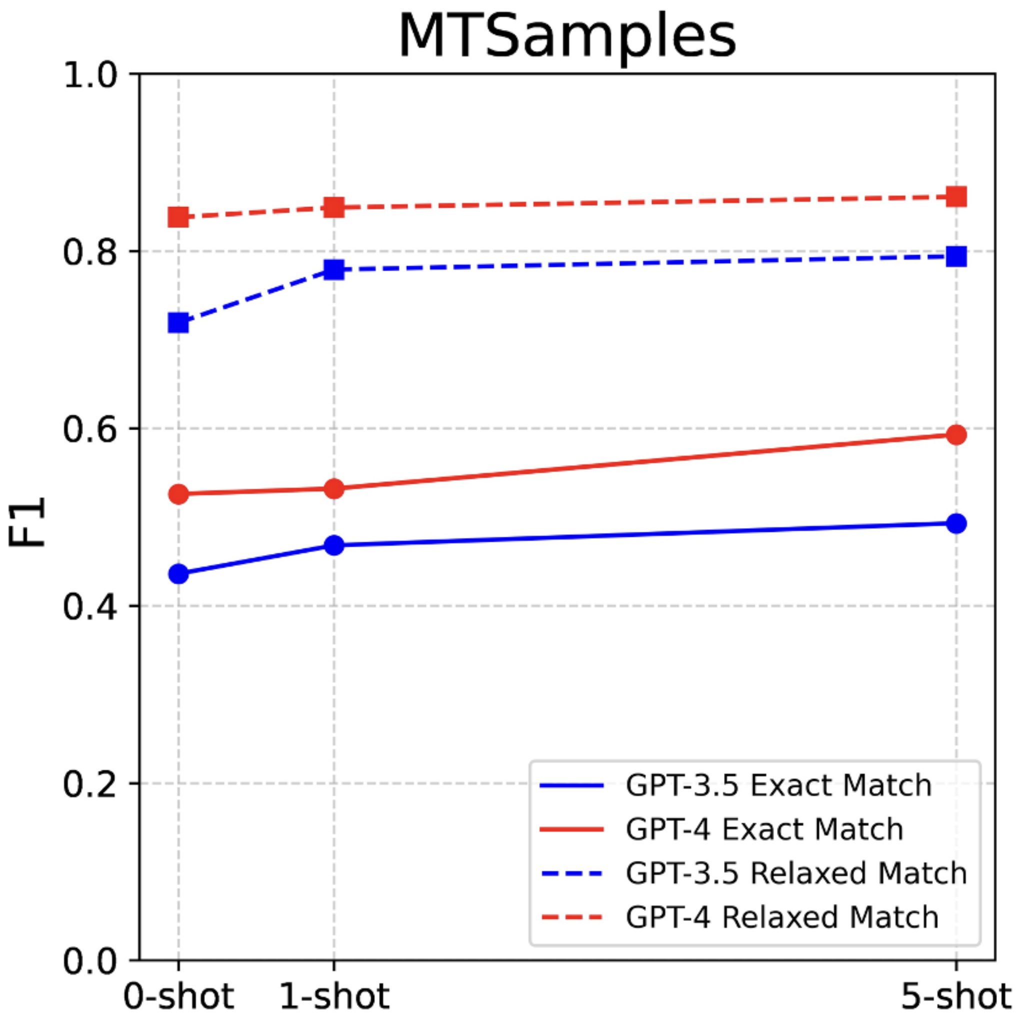
###Task:
Your task is to generate an HTML version of an input text, marking up specific entities. The entities to be identified are: 'medical problems', 'treatments', and 'tests'.
###Entity markup guide:
Use HTML tags to highlight these entities. Each should have a class attribute indicating the type 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

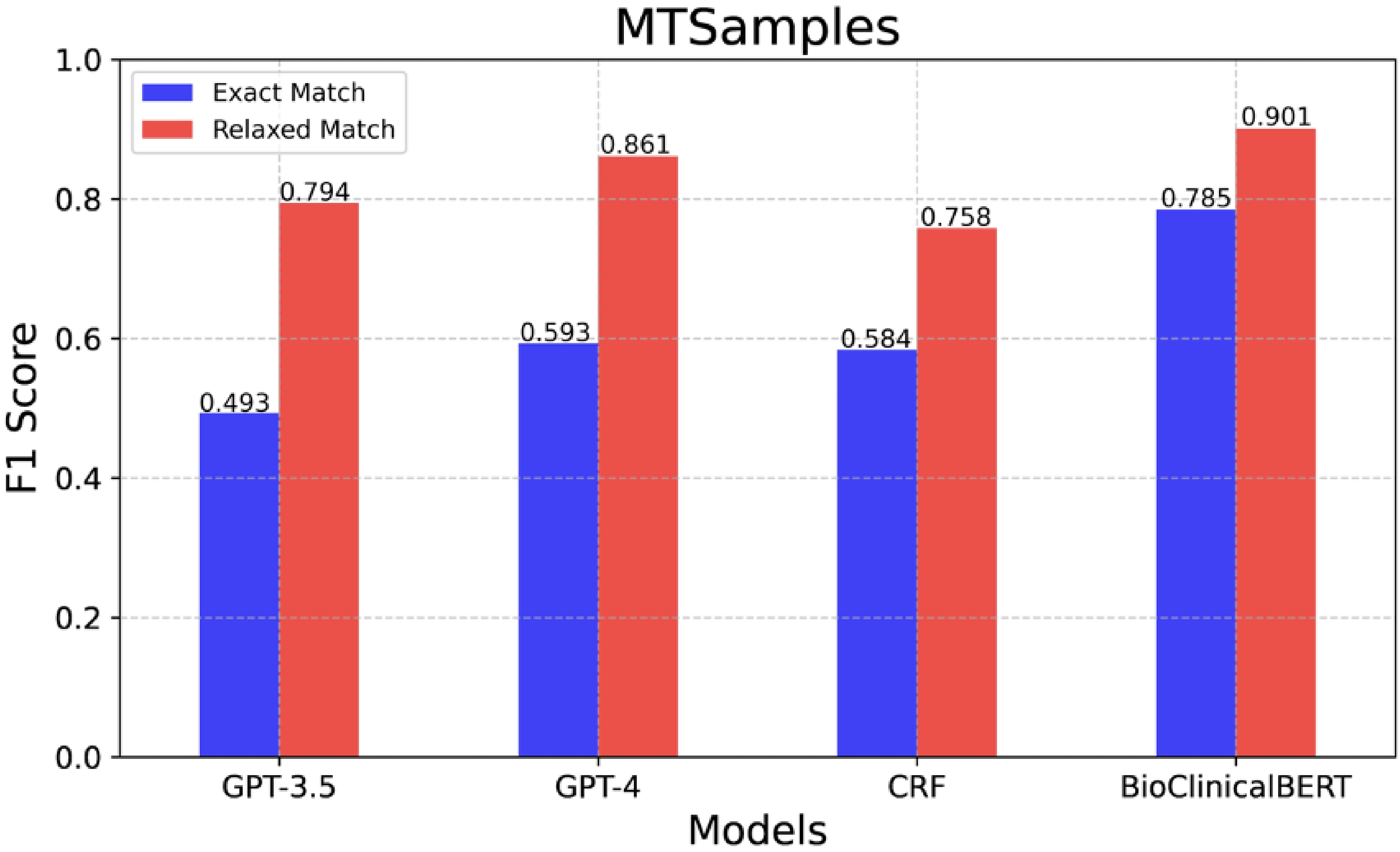


Prompt Strategies



Few-shot Learning

Results – Comparing ML, DL, and LLMs



Evaluations of GPTs on Different Biomedical NLP Tasks

- **Objective:** Establish the baseline performance of GPT 3.5 and GPT 4 on *12* biomedical datasets across *6* NLP tasks

- **NLP tasks and datasets:**
  - Named entity recognition
  - Relation extraction
  - Document classification
  - Question answering
  - Text summarization
  - Text simplification

- **Models:**
  - GPT-3.5-turbo-0301
  - GPT-4-0314
  - LLaMA 2, PMC LLaMA
  - BERT and BART

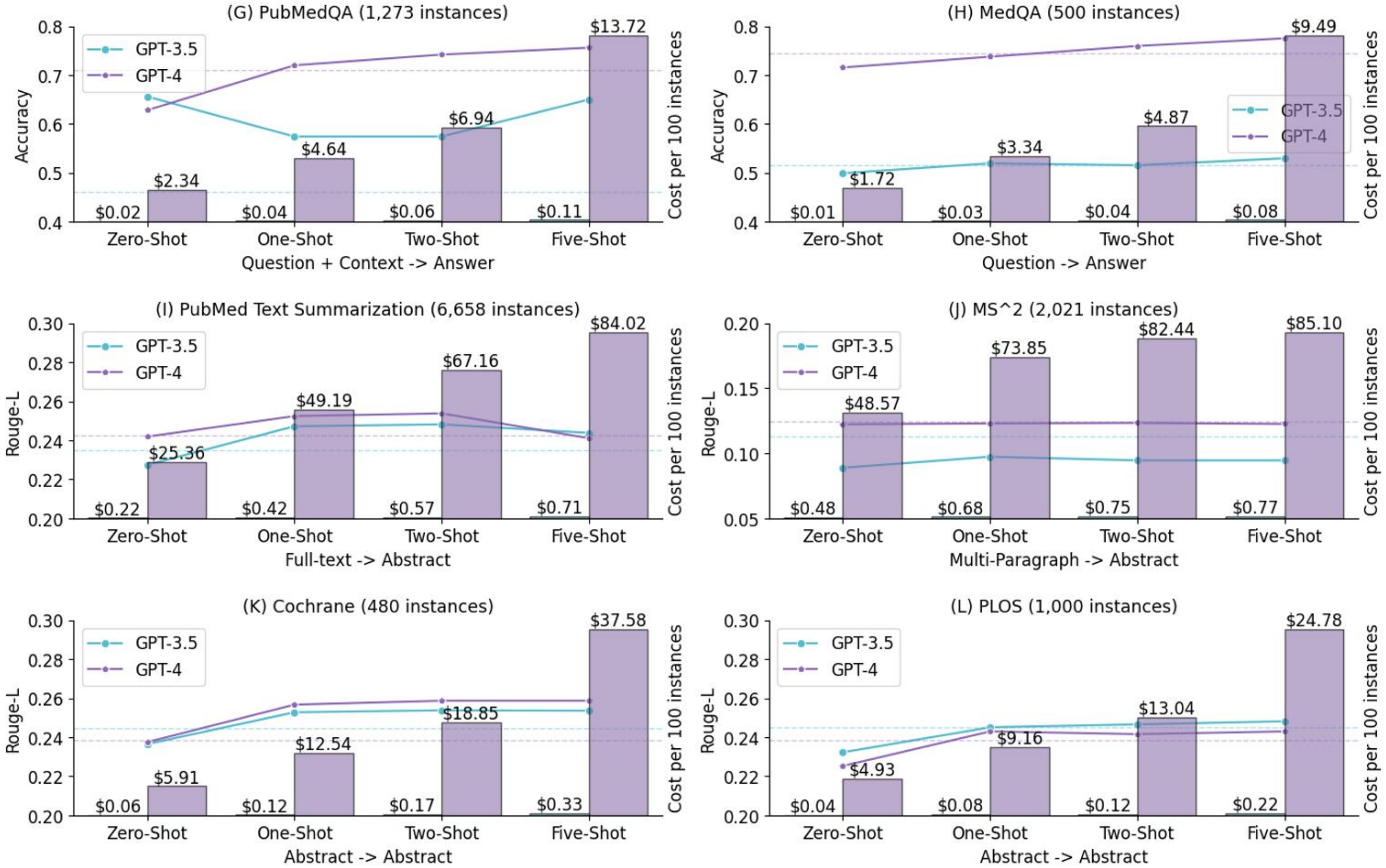
	Training	Validation	Testing	Primary metrics
Named entity recognition				
BC5CDR-chemical [43]	4,560	4,581	4,797	Entity-level F1 [43, 44]
NCBI-disease [45]	5,424	923	940	Entity-level F1 [16, 45]
Relation extraction				
ChemProt [46]	19,460	11,820	16,943	Macro F1 [47]
DDI2013 [48]	18,779	7,244	5,761	Macro F1 [48, 49]
Multi-label document classification				
HoC [50]	1,108	157	315	Macro F1 [50, 51]
LitCovid [52]	24,960	6,239	2,500	Macro F1 [52]
Question answering				
MedQA 5-option [53]	10,178	1,272	1,273	Accuracy [53]
PubMedQA [55]	190,142	21,127	500	Accuracy [55]
Text summarization				
PubMed Text Summarization[1][56]	117,108	6,631	6,658	Rouge-L [56]
MS^2[2][59]	14,188	2,021	-	Rouge-L [59]
Text simplification				
Cochrane PLS [61]	3,568	411	480	Rouge-L [61]
PLOS Text Simplification [64]	26,124	1,000	1,000	Rouge-L [64]

Chen Q et al. Benchmarking large language models for biomedical natural language processing applications and recommendations. Nature Communications (in press). 2024

Results - Performance

		SOTA results before the LLMs (Foundation model)	Zero/Few-shot									Fine-tuned	
			Zero-shot			One-shot			Five-shot				
			GPT-3.5	GPT-4	LLaMA 2 13B	GPT-3.5	GPT-4	LLaMA 2 13B	GPT-3.5	GPT-4	LLaMA 2 13B[2]	LLaMA 2 13B	PMC LLaMA 13B
Named entity recognition													
BC5CDR-chemical	Entity F1	0.9500 [80] (PubMedBERT)	0.6274	0.7993	0.3944	0.7133	0.8327*	0.6276	0.7228	0.7979	0.5530	0.9149	0.9063
NCBI Disease	Entity F1	0.9090 [80] (PubMedBERT)	0.4060	0.5827	0.2211	0.4817	0.5988	0.3811	0.4309	0.6389*	0.4847	0.8682*	0.8353
Relation extraction													
ChemProt	Macro F1	0.7344 [81] (BioBERT)	0.1345	0.3250	0.1392	0.1280	0.3391	0.0718	0.1758	0.3756	0.0967	0.4612*	0.3111
DDI2013	Macro F1	0.7919 [49] (BioBERT)	0.2004	0.2968	0.1305	0.2126	0.3312	0.1779	0.1706	0.3276	0.1663	0.6218	0.5700
Multi-label document classification													
HoC	Macro F1	0.8882 [51] (BioBERT)	0.6722	0.7109	0.1285	0.6671	0.7093	0.3072	0.6994	0.7099	0.1797	0.6957*	0.4221
LitCovid	Macro F1	0.8921 [51] (BioBERT)	0.5967	0.5883	0.3825	0.6009	0.5901	0.4808	0.6179	0.6077	0.3305	0.5725*	0.4273
Question answering													
MedQA (5-Option)	Accuracy	0.4195[1] [82] (BioLinkBERT)	0.4988	0.7156	0.2522	0.5161	0.7439	0.2899	0.5208	0.7651*	0.3504	0.4462*	0.3975
PubMedQA	Accuracy	0.7340 [82] (BioLinkBERT)	0.6560	0.6280	0.5520	0.4600	0.7100	0.2660	0.6920	0.7580*	0.6000	0.8040*	0.7680
Text summarization													
PubMed	Rouge-L	0.4316 [83] (BART)	0.2274	0.2419	0.1190	0.2351	0.2427	0.0989	0.2423	0.2444	0.1629	0.1857*	0.1684
MS^2	Rouge-L	0.2080 [59] (BART)	0.0889	0.1224	0.0948	0.1132	0.1248	0.0320	0.1013	0.1218	0.1205	0.0934*	0.0059
Text simplification													
Cochrane PLS	Rouge-L	0.4476 [84] (BART)	0.2365	0.2375	0.2081	0.2447	0.2385	0.2207	0.2470	0.2469	0.2283	0.2355	0.2370
PLOS	Rouge-L	0.4368 [64] (BART)	0.2323	0.2253	0.2121	0.2449*	0.2386	0.1836	0.2416	0.2409	0.1656	0.2583	0.2577
Macro-average		0.6536	0.3814	0.4561	0.2362	0.3848	0.4750	0.2614	0.4052	0.4862	0.2866	0.5131	0.4422

Results - Cost



Results - Recommendations

	Zero/few-shot	Fine-tuning
Highly recommend **Question answering** Reasoning-related	**Top-choice:** GPT-4 **Good-choice:** Closed-source LLMs only (e.g., starting with GPT-3.5) + Advanced Prompt Engineering	Open-source LLMs
Recommend **Summarization** **Simplification** Generation-related	**Good-choice:** Closed-source LLMs only (e.g., starting with GPT-3.5)	**Strong baseline to try first:** fine-tuned BART models **Open-source LLMs:** if input context length fits
Good to try **Document-level classification** Semantic understanding-related	**Good-choice:** Closed-source LLMs only (e.g., starting with GPT-3.5) + Advanced Prompt Engineering	**Strong baseline to try first:** fine-tuned BERT models **Open-source LLMs:** if input context length fits
Less recommend **Extraction** Extractive tasks	Less recommended	**Top-choice:** fine-tuned BERT models **Open-source LLMs:** if input context length fits

General Recommendations

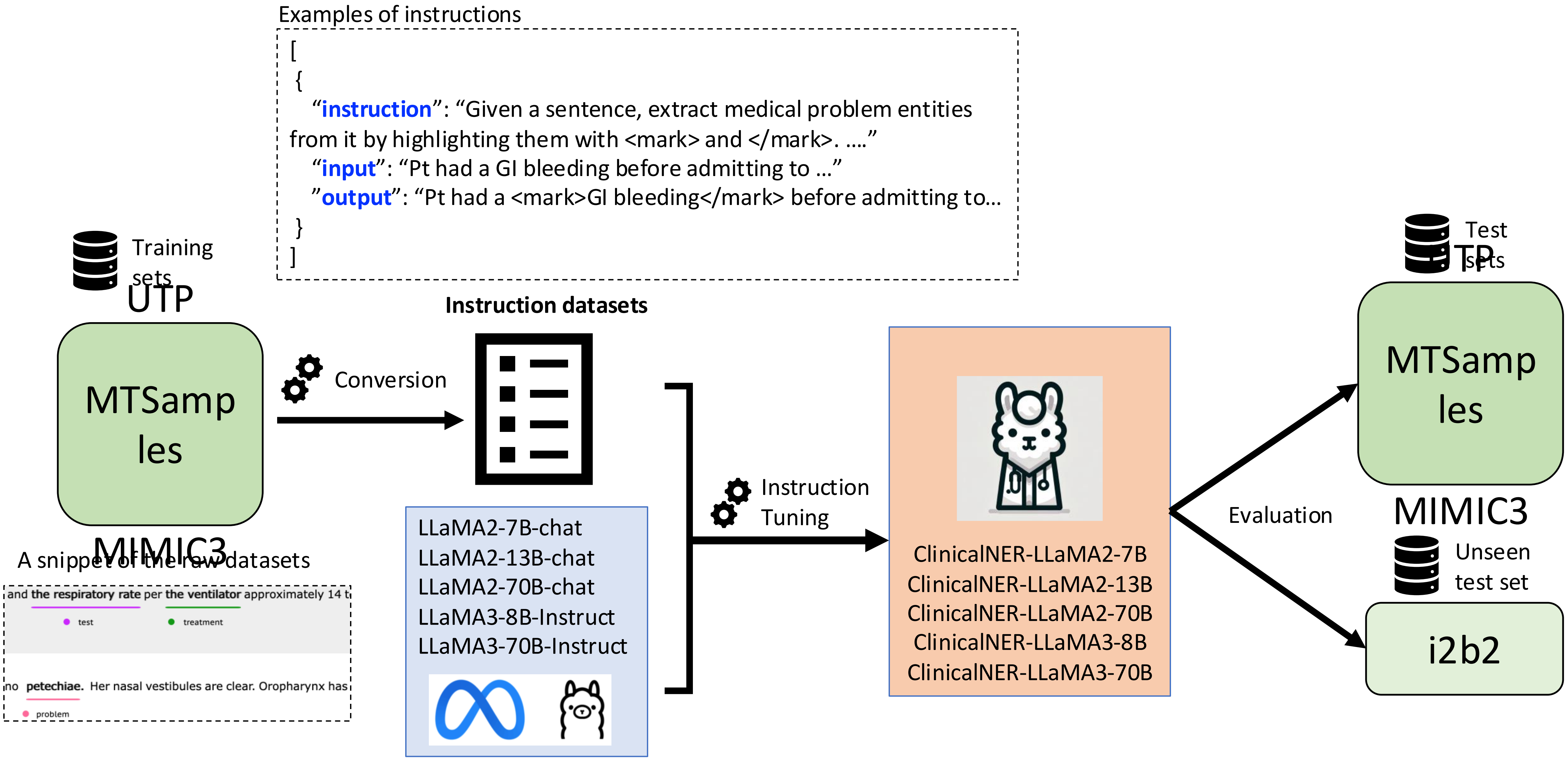
1. Stay aware of inconsistent, missing, and hallucinated responses; providing even one example could reduce such cases; manual review is essential
2. GPT-3.5 is a reliable baseline option given its performance and cost-effectiveness; apply GPT-4 especially for tasks requiring advanced reasoning abilities
3. Apply advanced prompt engineering especially for tasks requiring reasoning and semantic understanding

Clinical IE #2: Instruction Tuning of LLaMA

- Motivation:
  - Supervised fine tuning of LLaMA for clinical NER tasks
- Clinical NER Task:
  - Extract problems, drugs, labs, and other treatments from clinical notes.
- Clinical NER datasets:
  - UT Physicians
  - MTSample
  - MIMIC-III
  - I2b2
- Models:
  - LLaMA2-7B, 13B, and 70B
  - LLaMA3-8B and 70B
  - BioMedBERT

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…
}
]
Training sets
UTP
MTSamples
MIMIC3
A snippet of the raw datasets
and the respiratory rate per the ventilator approximately 14 t
test
treatment
no petechiae. Her nasal vestibules are clear. Oropharynx has
problem
Conversion
Instruction datasets
LLaMA2-7B-chat
LLaMA2-13B-chat
LLaMA2-70B-chat
LLaMA3-8B-Instruct
LLaMA3-70B-Instruct
Instruction Tuning
ClinicalNER-LLaMA2-7B
ClinicalNER-LLaMA2-13B
ClinicalNER-LLaMA2-70B
ClinicalNER-LLaMA3-8B
ClinicalNER-LLaMA3-70B
Evaluation
Test sets
UTP
MTSamples
MIMIC3
Unseen test set
i2b2

Results

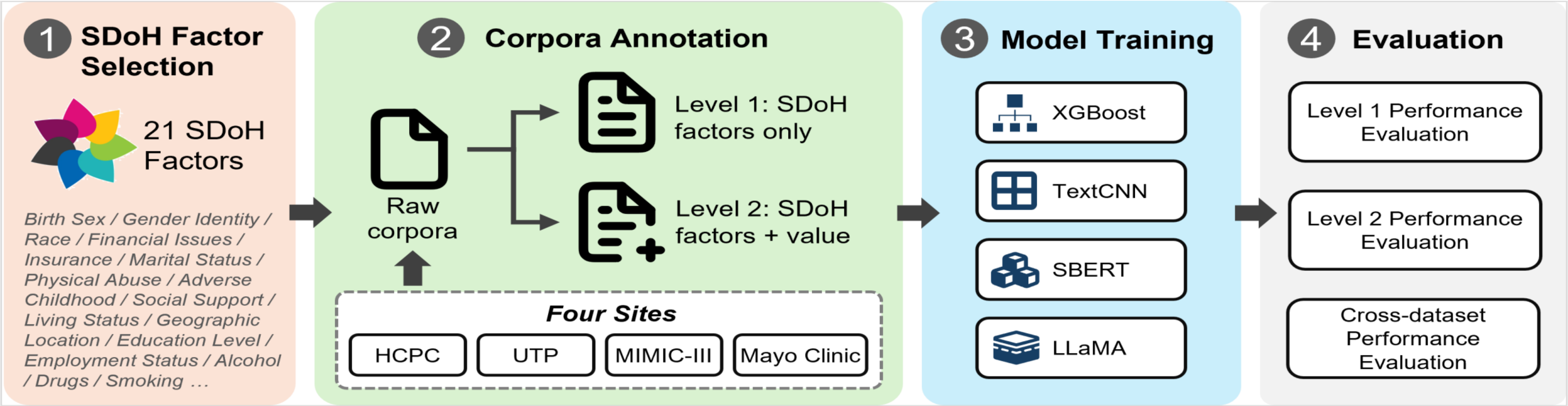
- Performance

Datasets	LLAMA2-7b		LLAMA2-13b		LLAMA2-70b		LLAMA3-8b		LLAMA3-70b		BioMedBERT	
	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 (seconds/note, UTP)

Speed	LLAMA2-7b	LLAMA2-13b	LLAMA2-70b	LLAMA3-8b	LLAMA3-70b	BioMedBERT
Train	42.3	72.6	304.2	39.2	273.9	18.9
Test	6.2	8.2	45.3	4.1	**39.1**	**0.2**

LLMs for Extracting Social Determinants of Health



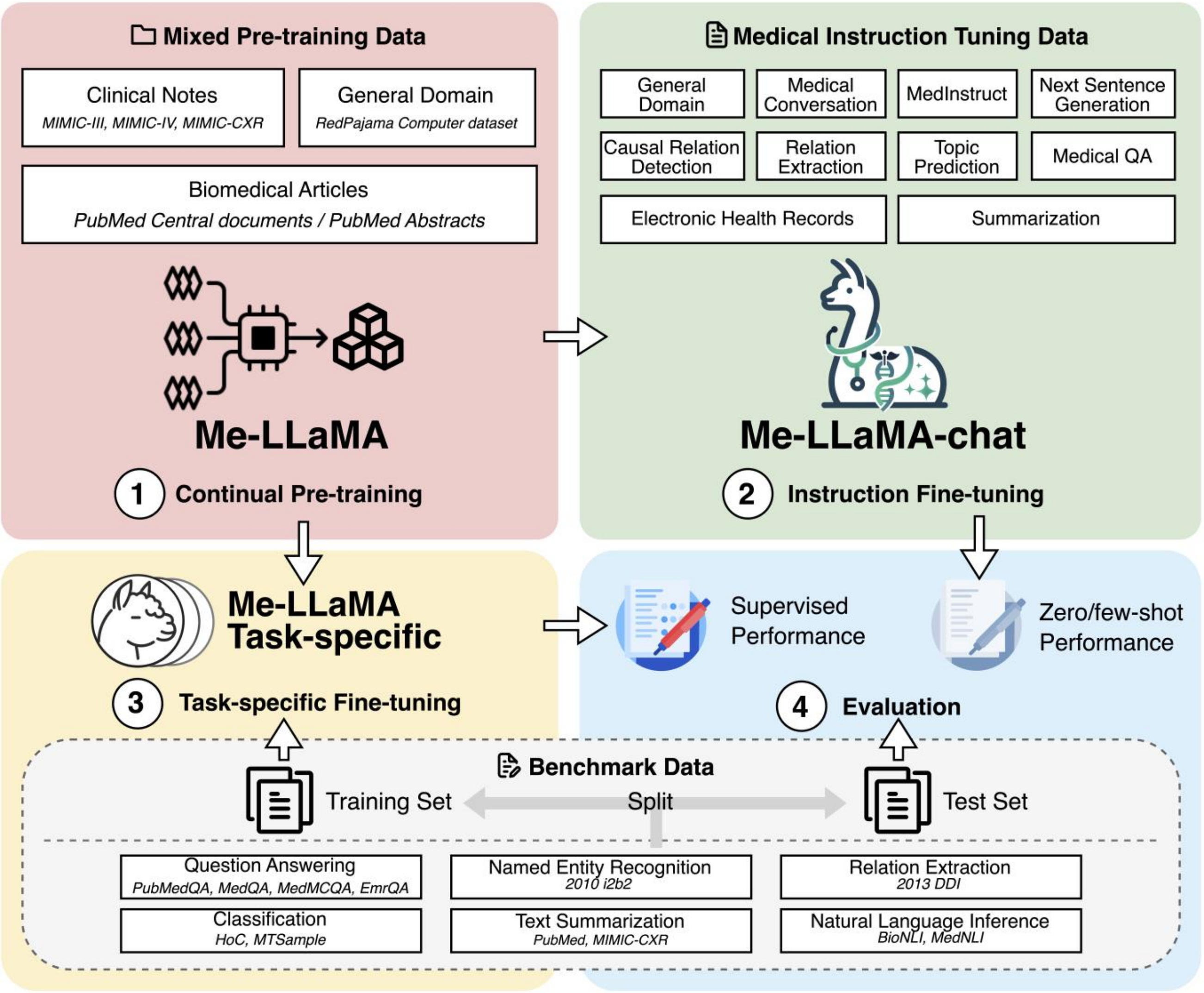
SDoH	Example
Geographic location	Pt born and raised in Rio Grande, Mexico. Location-Raised Location-Born
Sex, Race	The patient is an 80-year-old WF Race-Caucasian Sex-Female
Employment status	Pt has been unemployed for past year … EmploymentUnemployed-Present
Social support	… with her peers and has a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

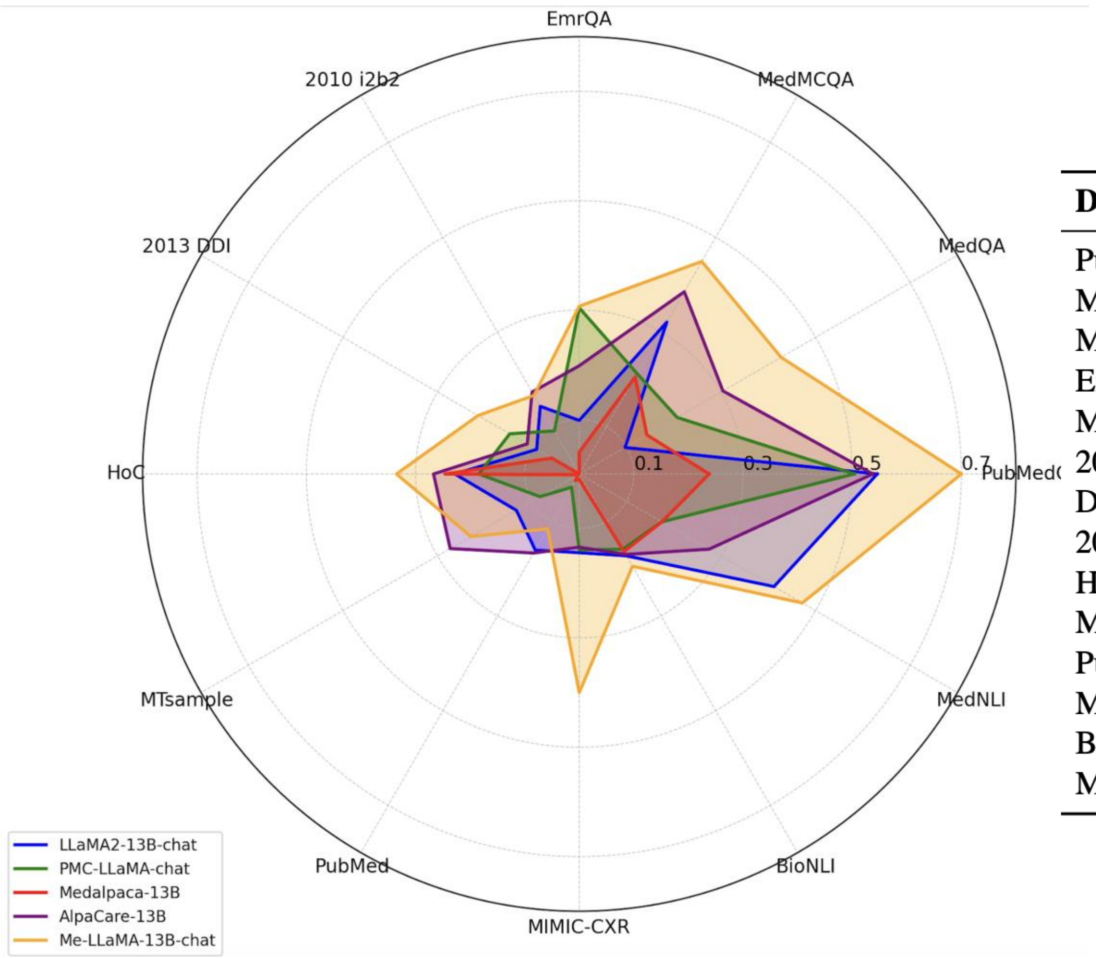
Clinical IE #3 (and Beyond): Continual Pre-training LLaMA

- Continual pre-training: Trained on 129B tokens of biomedical data, with 100,000+ GPU hours
- Instruction fine-tuning: Trained on 200K+ medical QA pairs, with 1,000+ GPU hours
- Task-specific fine-tuning: Trained and evaluated on 6 tasks, 12 datasets
- Available at 13B and 70B models

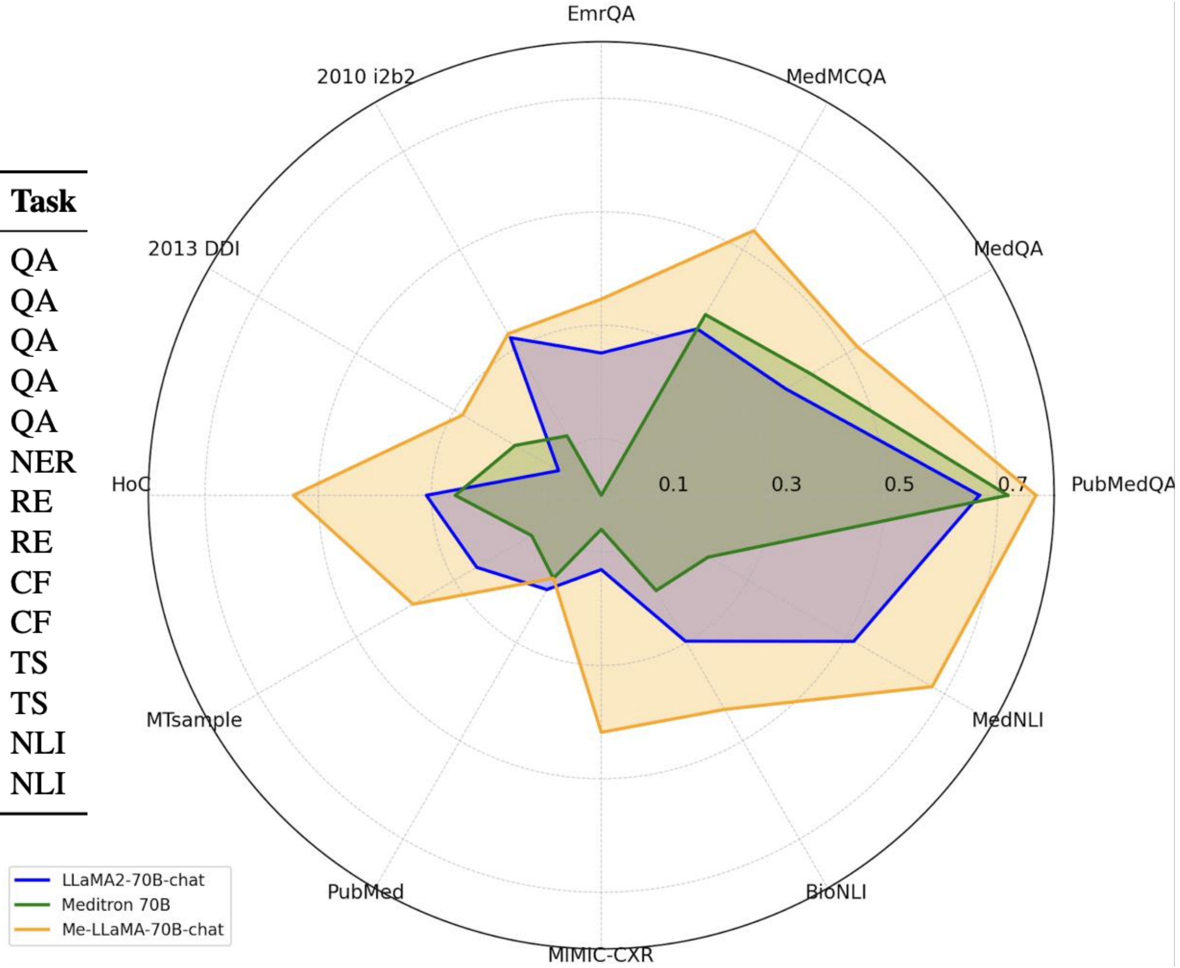
Xie Q et al. Me LLaMA: Foundation Large Language Models for Medical Applications arxiv, 2024



Me LLAMA: Outperform Existing Open Medical LLMs



Data	Task
PubMedQA	QA
MedQA	QA
MedMCQA	QA
EmrQA	QA
MMLU	QA
2012 i2b2	NER
DDI2013	RE
2018 n2c2	RE
HoC	CF
MTSample	CF
PubMedSum	TS
MIMIC-CXR	TS
BioNLI	NLI
MedNLI	NLI



Best on 9 out of 12 datasets on 13B

Best on 11 out of 12 datasets on 70B

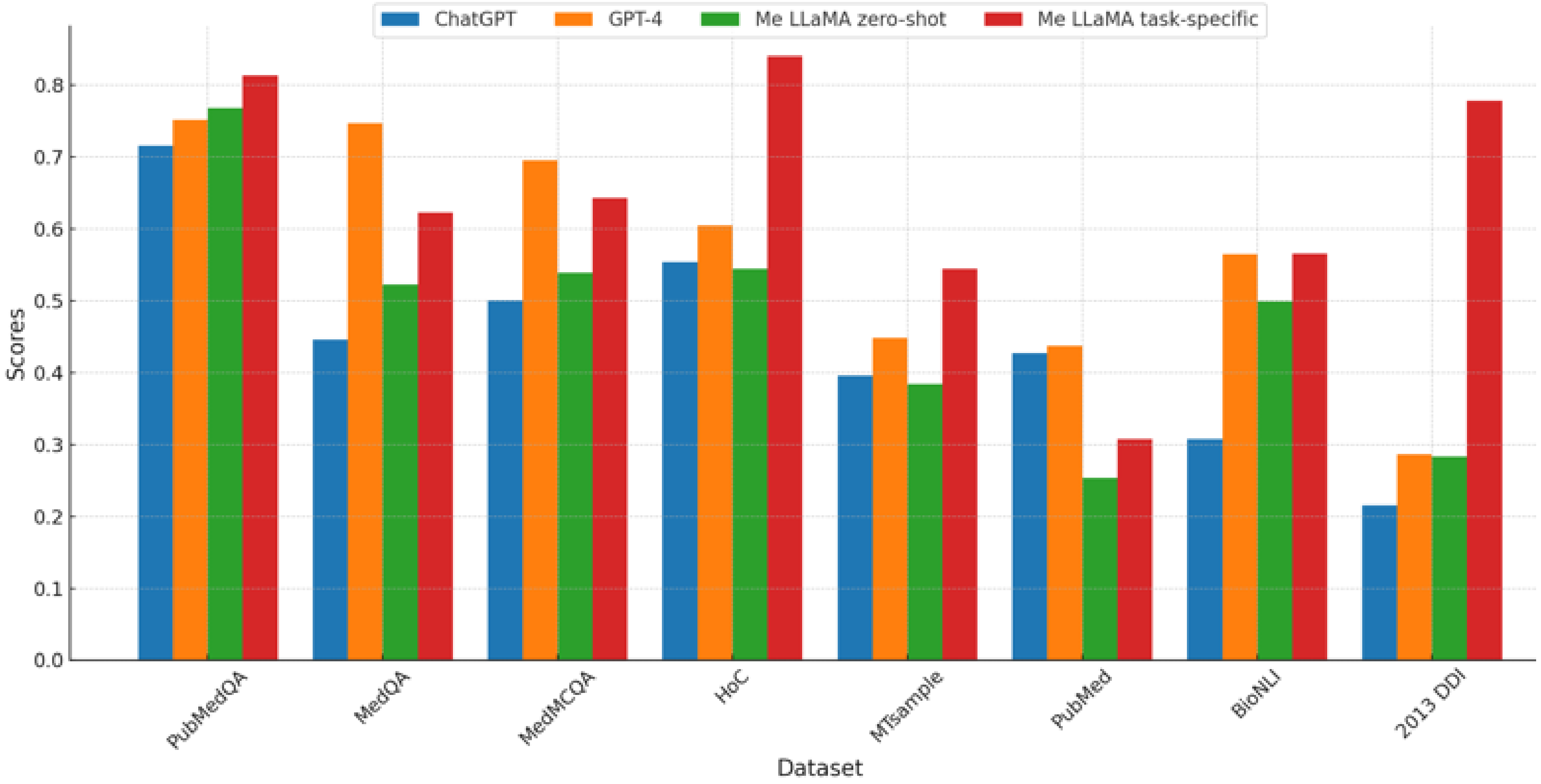
Me LLaMA vs. ChatGPT and GPT-4

Zero-shot learning

- Outperform ChatGPT on 5/8
- Underperform GPT-4 on 7/8

Supervised learning

- Outperform GPT-4 on 5/8
- Outperform ChatGPT on 7/8



Summary of LLMs for Clinical Information Extraction

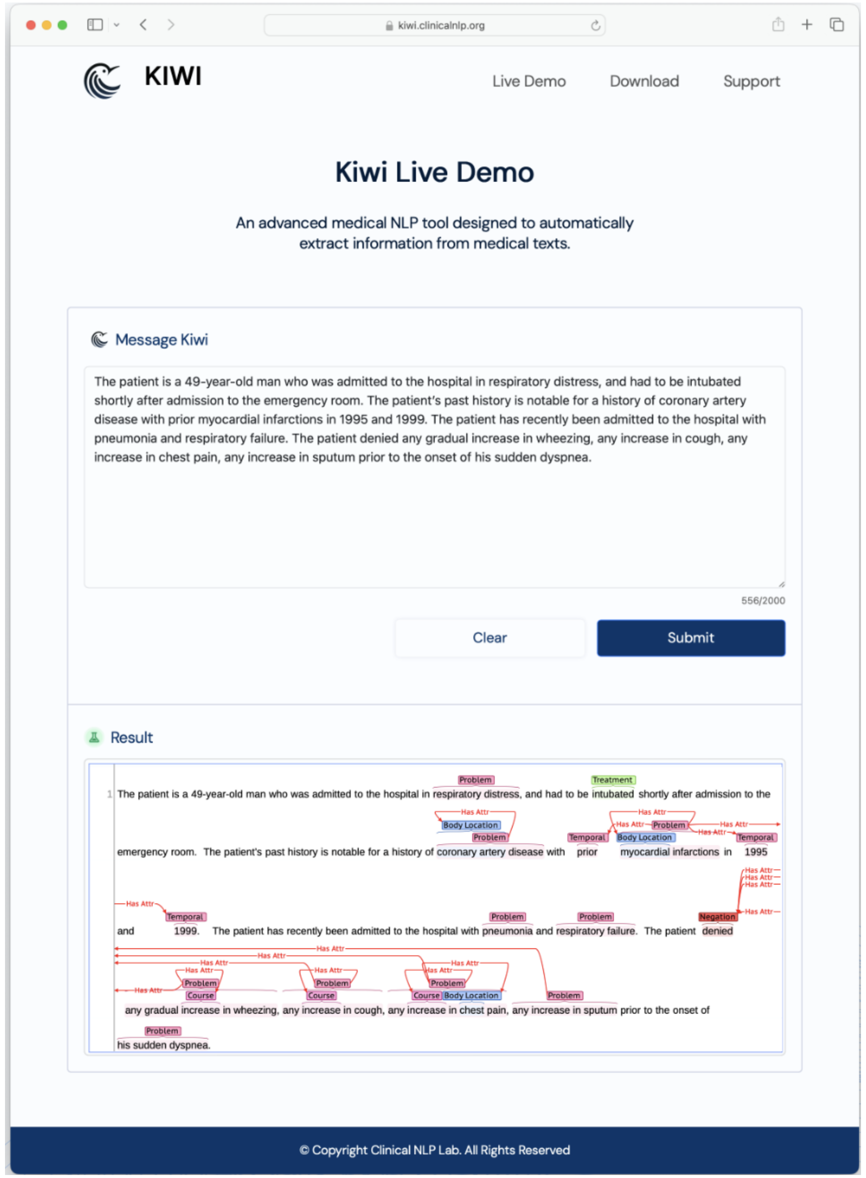
- LLMs vs BERT
  - LLMs with few-shot learning showed reasonable but lower performance than fine-tuned BERT models for clinical IE tasks
  - LLMs with instruction tuning showed better performance and generalizability than fine-tuned BERT models for clinical IE tasks, especially for general domain entities
- GPT vs. LLaMA
  - Zero-shot performance, fine-tuning, data privacy, costs, expertise, integration
- Ready for switching from BERT to LLMs for Clinical IE tasks
  - Performance, costs, infrastructure, speed, data availability

The LLM field is highly dynamic, with rapid advancements and continual changes!

KIWI - A LLM-based Clinical Information Extraction System

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

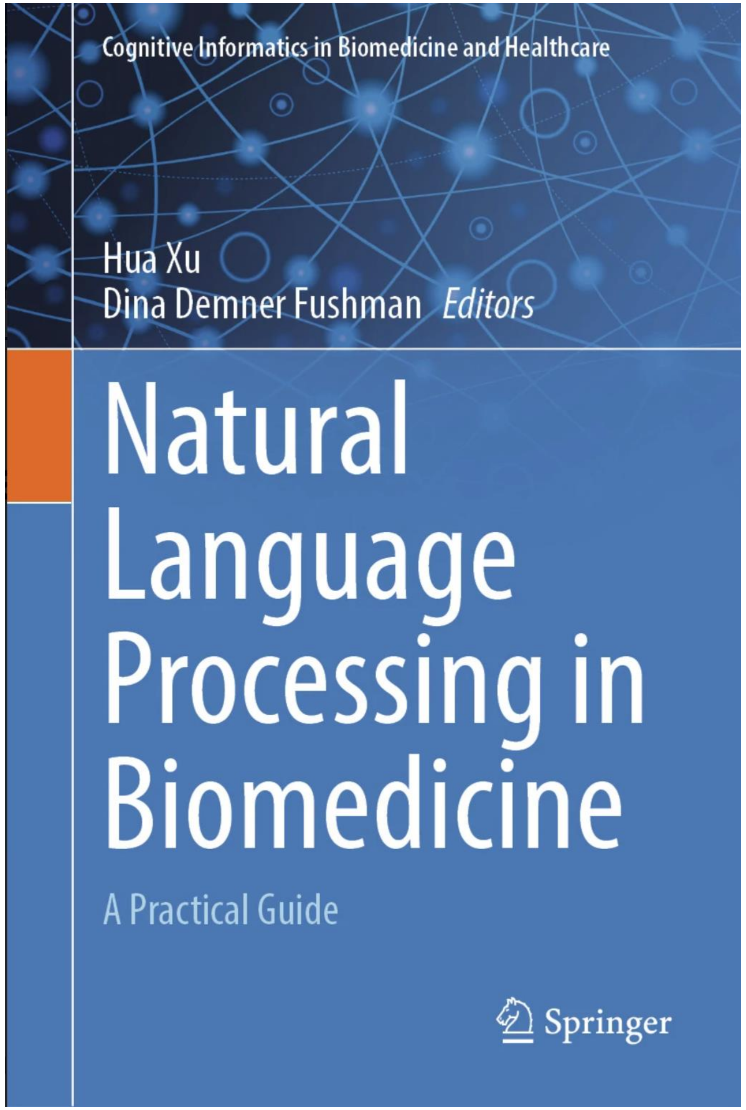
https://kiwi.clinicalnlp.org/



Book – Natural Language Processing in Biomedicine

- A textbook covers broad topics within the application of NLP in biomedicine.

- Three sections:
  - Basics of NLP including linguistic information, ML, DL, LLM algorithms
  - Common biomedical NLP tasks such as NER, RE, IR, QA etc.
  - How to build NLP solutions for different biomedical texts: clinical notes, literature, social media etc.



https://link.springer.com/book/10.1007/978-3-031-55865-8

MedViz System Demo

Acknowledgement

- **Contributors**
  - Chen, Qingyu
  - Xie, Qianqian
  - He, Huan
  - Hu, Yan
  - Kuttichi Keloth, Vipina
  - Hong, Na
  - Gilman, Chris
  - Lin, Fongci
  - Peng, Xueqing
  - Raja, Kalpana
  - KJ Richmond
  - Cardoso, Joao
  - Ng, Chi Wing
  - Ondov, Brian
  - Zhang, Jeffrey
  - Qiang, Lingfei
  - Zhang, Vincent
  - Qiaozhu Mei
  - Yutong Xie
  - Qijia Liu
  - Dennis Shung
  - Aokun Chen
  - Cheng Peng
  - Yonghui Wu
  - Jiang Bian
  - Zuo, Xu
  - Dennis Shuang
  - Andrew Taylor
  - Past lab members
  - Collaborators

- **Grants**
  - 1RF1AG072799
  - U24LM013755
  - RM1HG011558
  - R01LM013519
  - U54HG012510
  - OT2OD035404
  - R01AG073435
  - R01AG080429
  - R01AG078154
  - Past grants

Thank you!

Questions?

hua.xu@yale.edu