Learning Healthcare Foundation Models: From Pre-training to Fine-tuning

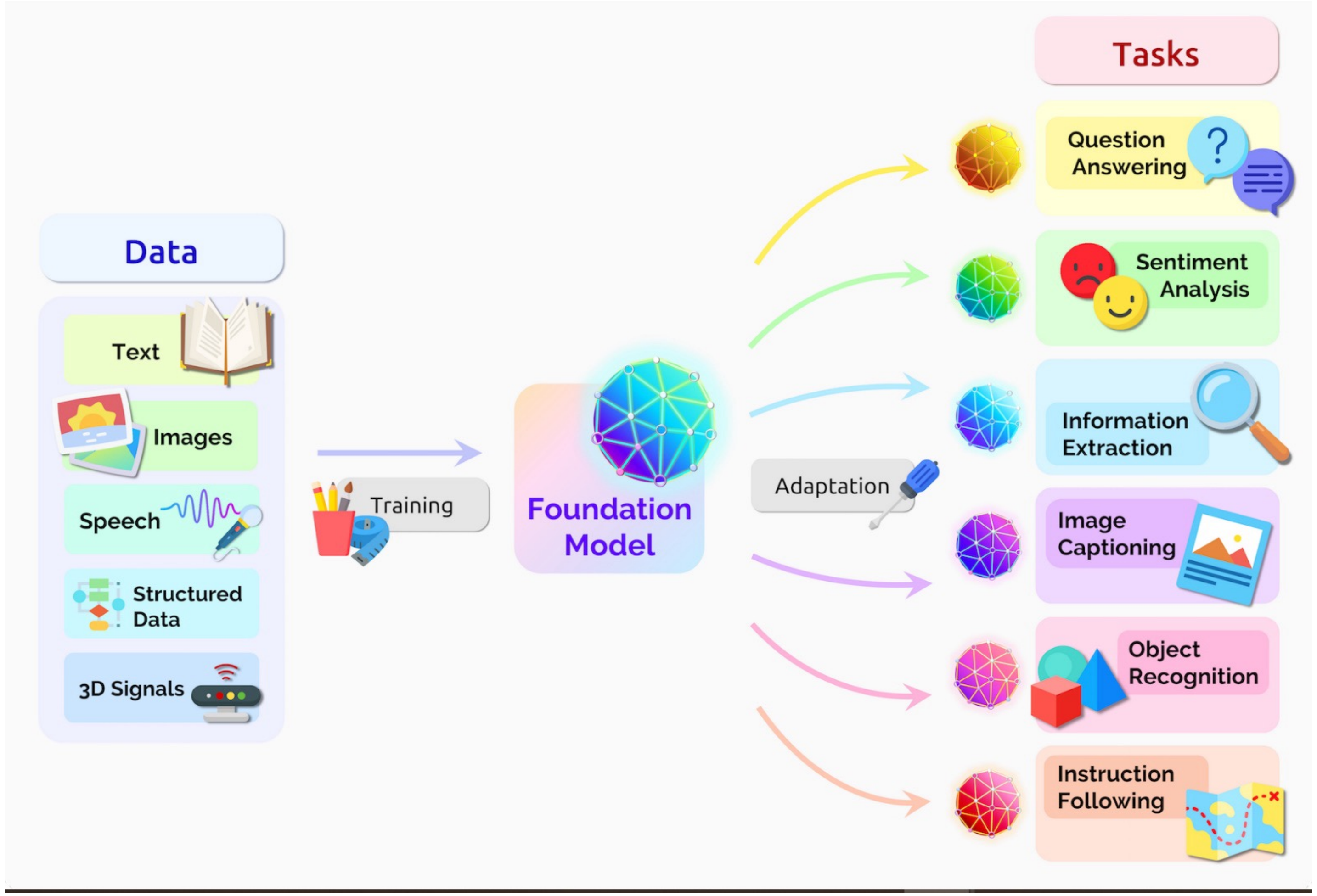
Fenglong Ma, Ph.D.

College of Information Sciences and Technology
Pennsylvania State University
fenglong@psu.edu

PennState

Personal Website: https://fenglong-ma.github.io/
Penn State Data Science Lab: https://psudslab.github.io/

Foundation Models



https://blogs.nvidia.com/blog/what-are-foundation-models/

Bard

LLaMA by Meta

Claude

BY ANTHROP\C

GPT-4

GLM-4

新一代基座大模型

a BigScience initiative

BLOOM

176B params · 59 languages · Open-access

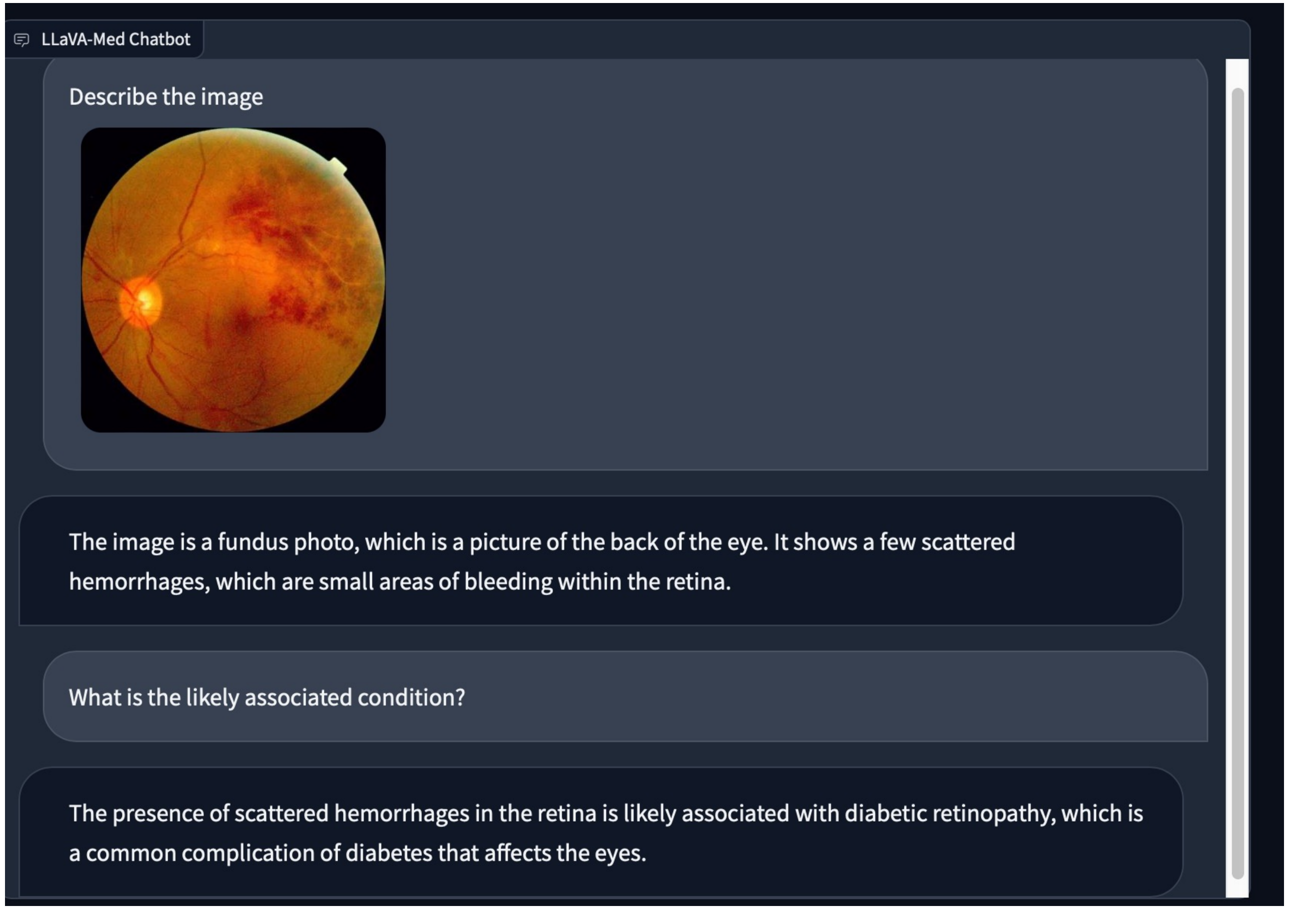
Foundation Models in General Domain

Medical Foundation Models

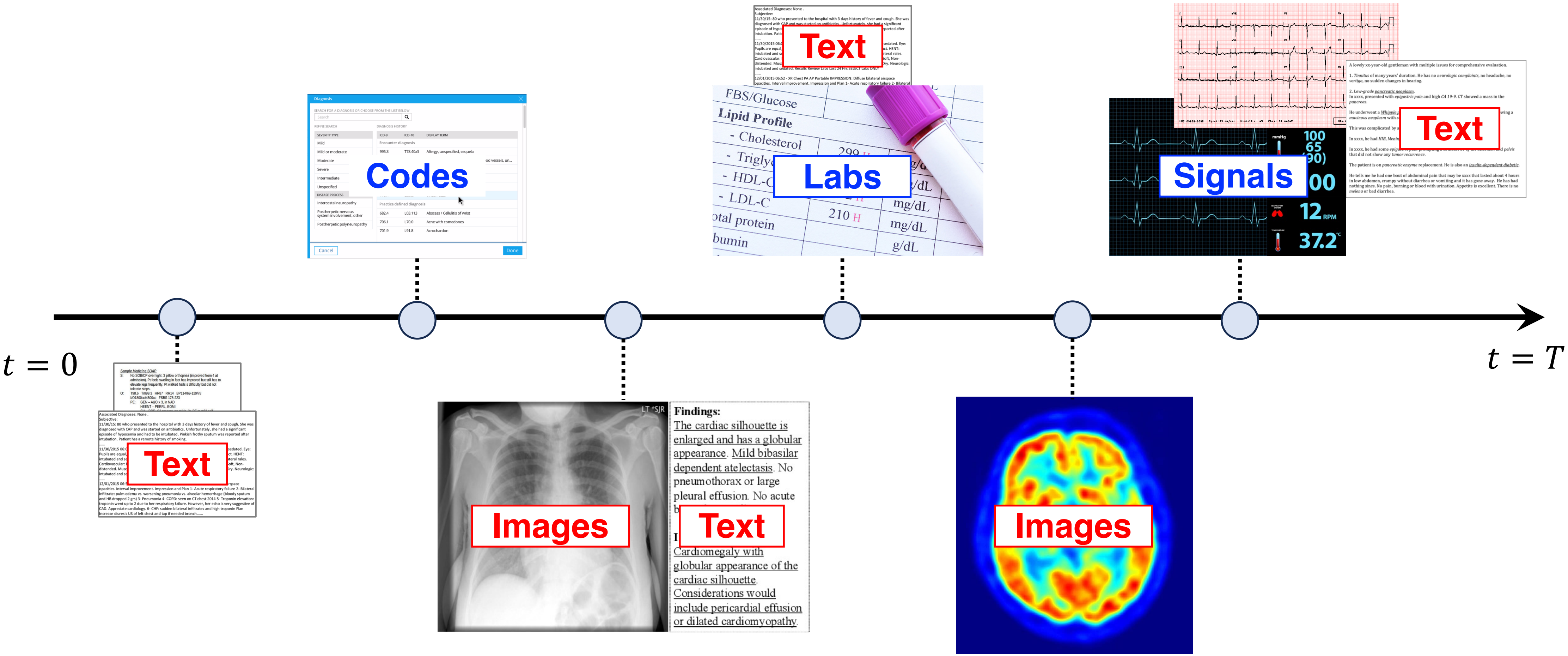
Image + Text

- LLaVA-Med

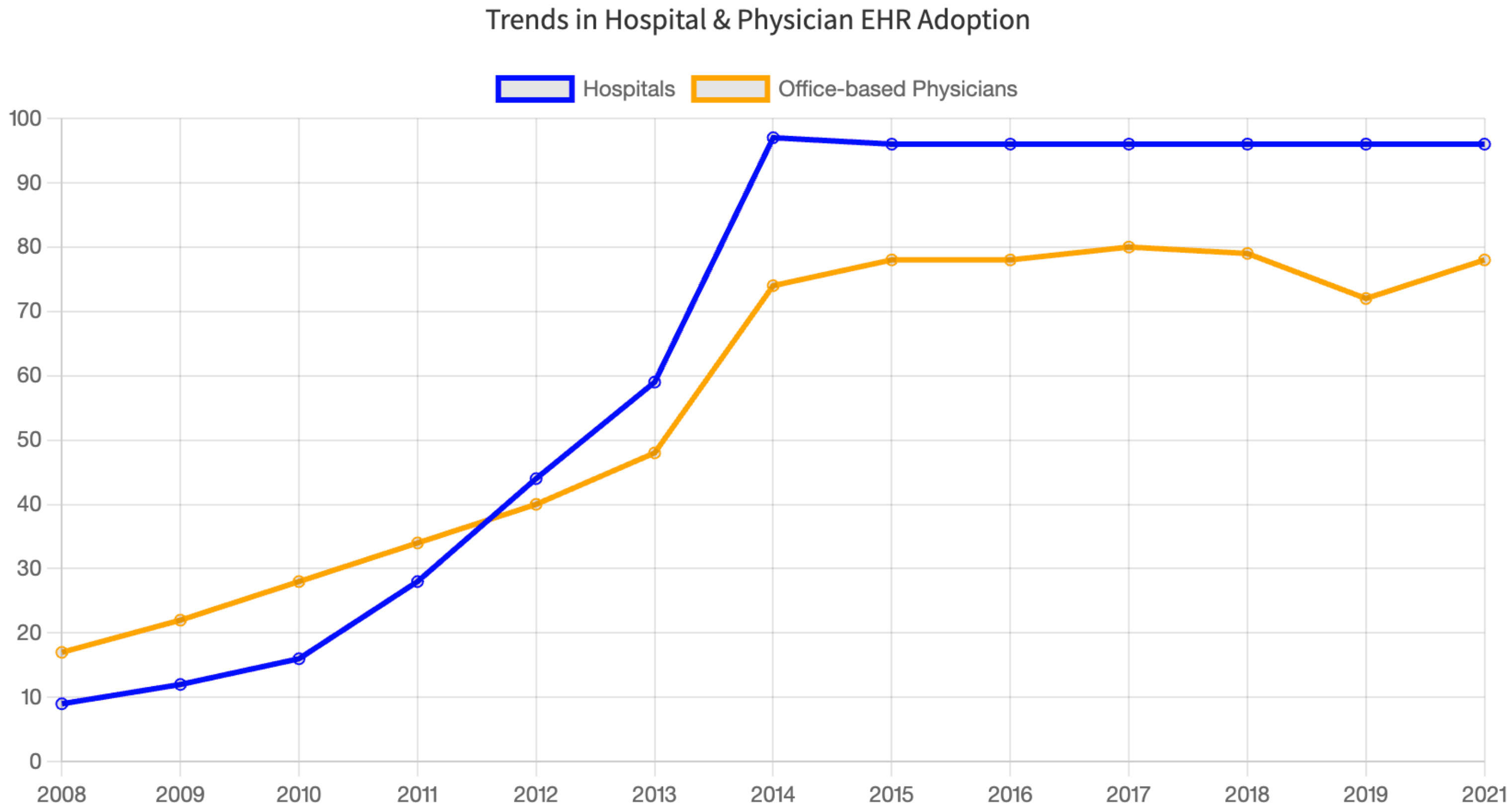
https://github.com/microsoft/LLaVA-Med



Multimodal Electronic Health Records (EHR)



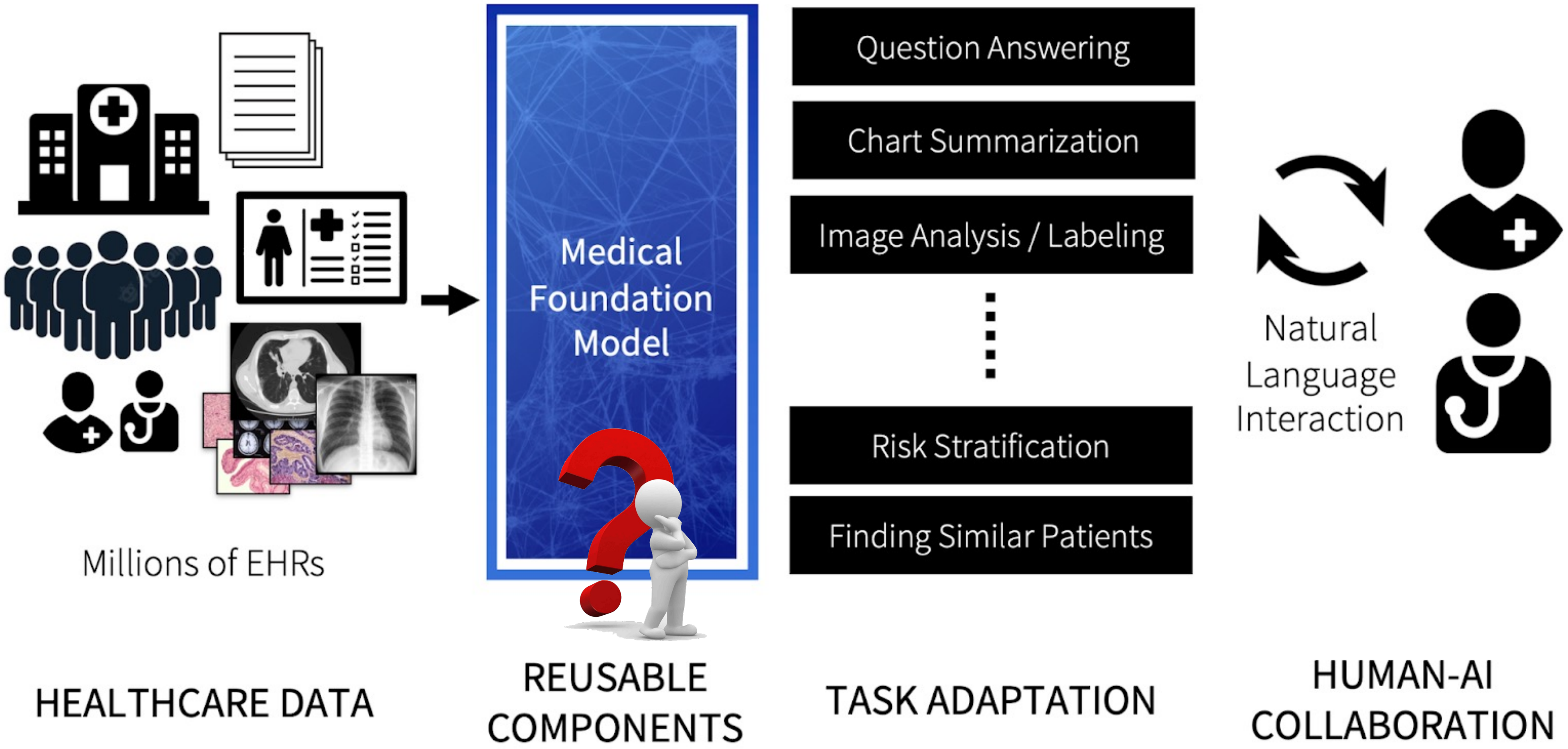
National Trends in Hospital and Physician Adoption of EHR



As of 2021, nearly 4 in 5 office-based physicians (78%) and nearly all non-federal acute care hospitals (96%) adopted a certified EHR. This marks substantial 10-year progress since 2011 when 28% of hospitals and 34% of physicians had adopted an EHR.

Office of the National Coordinator for Health Information Technology. 'National Trends in Hospital and Physician Adoption of Electronic Health Records,' Health IT Quick-Stat #61.

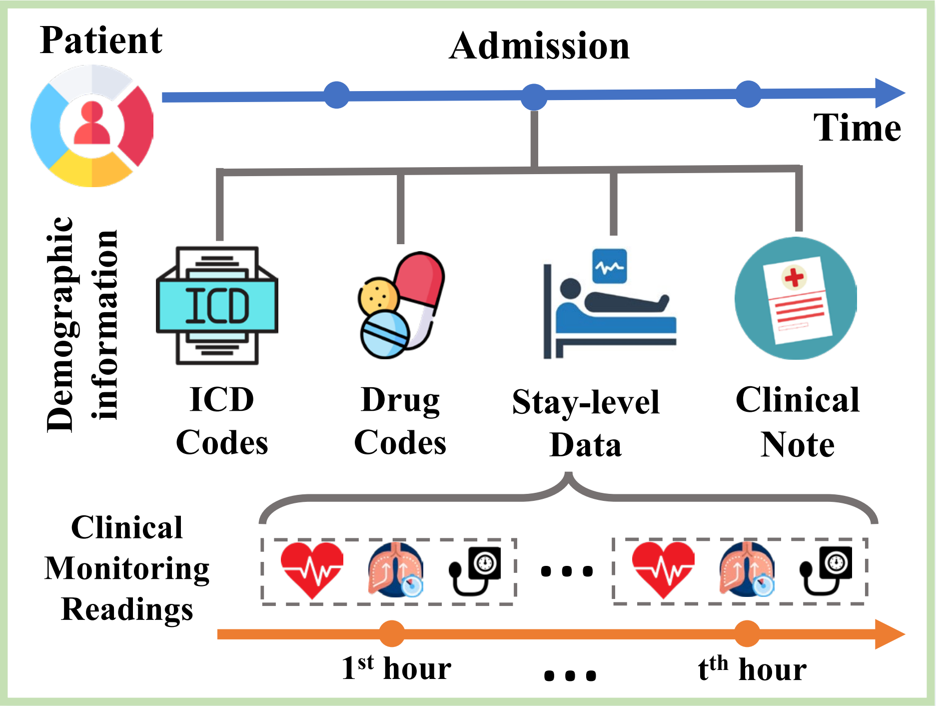
Ideal Healthcare/Medical Foundation Models



https://hai.stanford.edu/news/how-foundation-models-can-advance-ai-healthcare

Challenges of EHR Data

Hierarchical



Wang et al., *Hierarchical Pretraining on Multimodal Electronic Health Records*, EMNLP'23

Multi-sourced

Wang et al., *Unity in Diversity: Collaborative Pre-training Across Multimodal Medical Sources*, ACL'24

Wang et al., *FedMeKI: A Benchmark for Scaling Medical Foundation Models via Federated Knowledge Injection*, under review

Wang et al., *FedKIM: Adaptive Federated Knowledge Injection into Medical Foundation Models*, under review

Hierarchical Pretraining on Multimodal Electronic Health Records

Xiaochen Wang, Junyu Luo, Jiaqi Wang, Ziyi Yin, Suhan Cui, Yuan Zhong, Yaqing Wang and Fenglong Ma

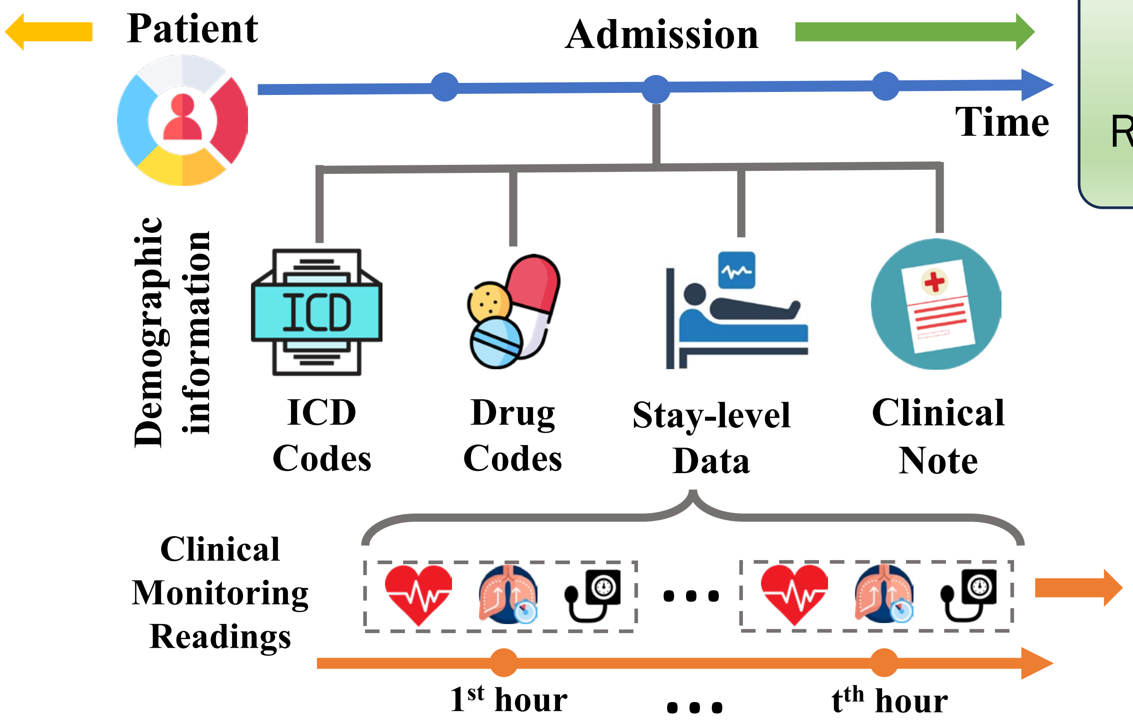
Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)

Hierarchical Pretraining on Multimodal Electronic Health Records

- Prediction tasks vary across hierarchies.

Long-term Prediction

Heath Risk Prediction after six months or one year



Shorter-term Prediction

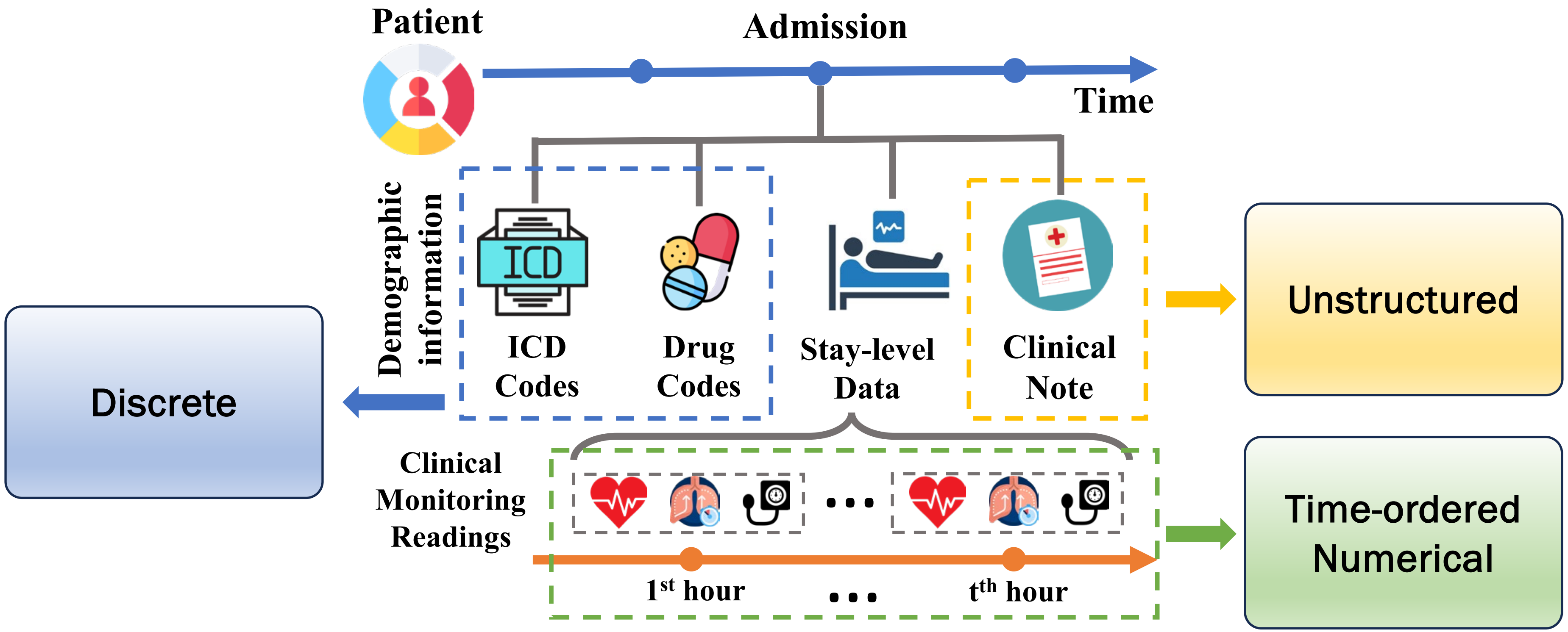
Readmission within 30 days

Hourly Prediction

Acute Respiratory Failure (ARF) within a few hours

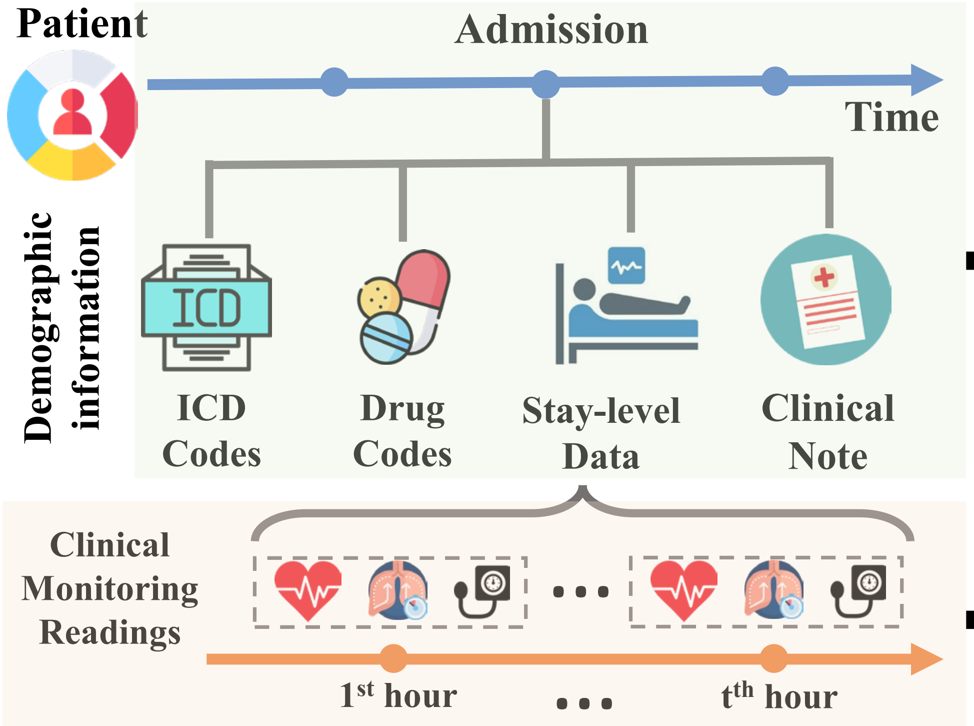
Hierarchical Pretraining on Multimodal Electronic Health Records

- Multimodal and heterogeneous EHR data



MedHMP: Hierarchical Multimodal Pretraining for Medicine

- Bottom-to-up Pretraining



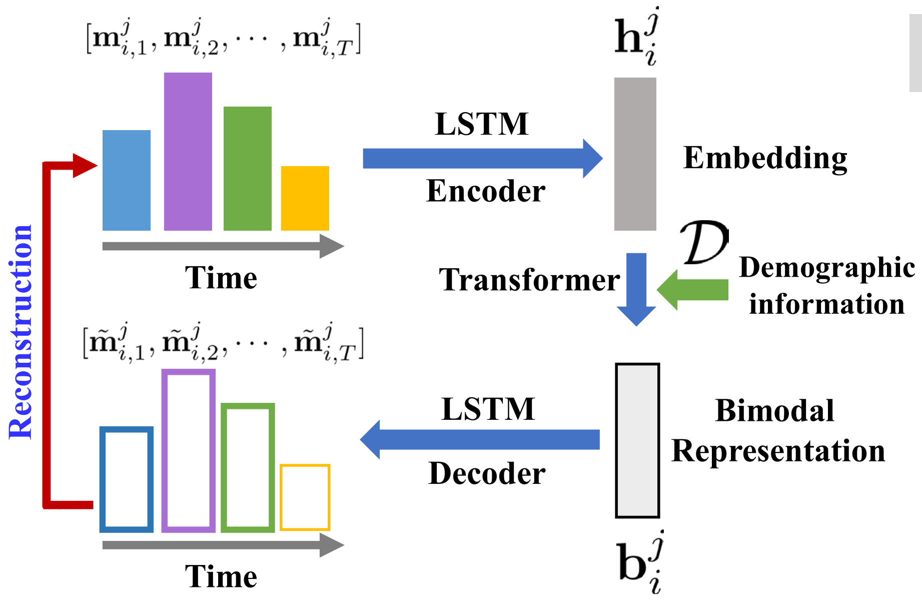
Intra-modality Pretraining
Predicting a set of masked ICD and drug codes

Inner-modality pretraining
Modality-level contrastive learning

Reconstructing the numerical time-ordered clinical features

Stay-level Self-supervised Pretraining

- Clinical Feature Reconstruction



# of admissions | # of stays | # of monitoring hours

$$\mathcal{L}_{\text{stay}} = \frac{1}{N * M * T} \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{t=1}^{T} ||\mathbf{m}^j_{i,t} - \tilde{\mathbf{m}}^j_{i,t}||_2^2.$$

$$\hat{\mathbf{b}}^j_i = \text{Softmax}\left(\frac{\mathbf{W}^Q_h \langle \mathbf{h}^j_i, \mathbf{d} \rangle \cdot \mathbf{W}^K_h \langle \mathbf{h}^j_i, \mathbf{d} \rangle}{\sqrt{d_r}}\right) \cdot \mathbf{W}^V_h \langle \mathbf{h}^j_i, \mathbf{d} \rangle,$$

$$\mathbf{b}^j_i = \text{MaxPooling}(\text{LayerNorm}(\langle \mathbf{h}^j_i, \mathbf{d} \rangle + \hat{\mathbf{b}}^j_i)),$$

Admission-level Pretraining

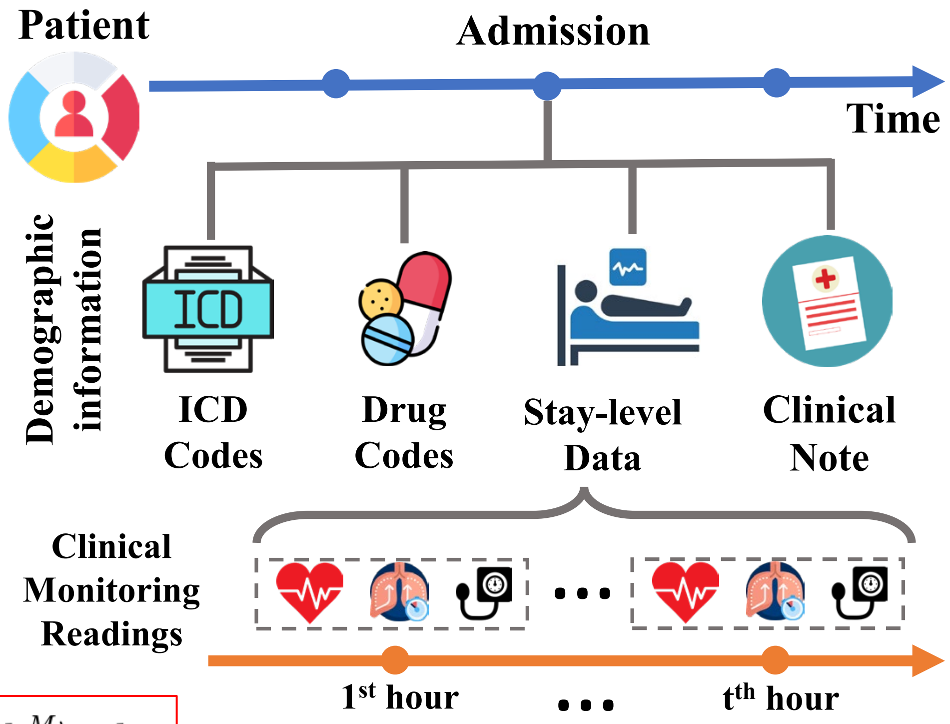
- Admission-level Feature Encoding
  - ICD Codes & Drug Codes

$$\mathbf{c}_i = \mathrm{MLP}_c(\mathcal{C}_i), \mathbf{g}_i = \mathrm{MLP}_g(\mathcal{G}_i).$$

  - Clinical Note

$$\mathrm{I}_i = \mathrm{ClinicalT5}(L_i)$$

  - Clinical Monitoring Readings

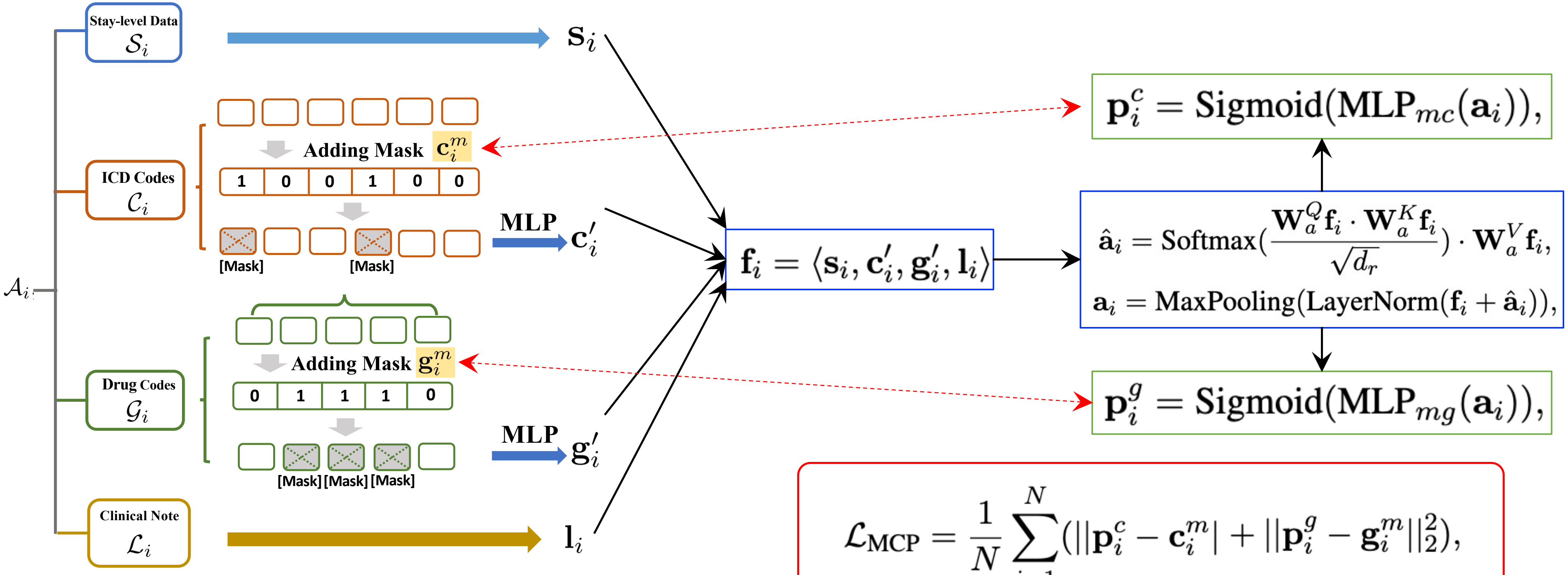


$$\hat{\mathbf{b}}_i^j = \mathrm{Softmax}\left(\frac{\mathbf{W}_h^Q \langle \mathbf{h}_i^j, \mathbf{d}\rangle \cdot \mathbf{W}_h^K \langle \mathbf{h}_i^j, \mathbf{d}\rangle}{\sqrt{d_r}}\right) \cdot \mathbf{W}_h^V \langle \mathbf{h}_i^j, \mathbf{d}\rangle,$$
$$\mathbf{b}_i^j = \mathrm{MaxPooling}(\mathrm{LayerNorm}(\langle \mathbf{h}_i^j, \mathbf{d}\rangle + \hat{\mathbf{b}}_i^j)),$$

$$\mathbf{s}_i = \mathbf{W}_s^\top \langle \mathbf{b}_i^1; \mathbf{b}_i^2; \cdots; \mathbf{b}_i^M\rangle + \mathbf{b}_s,$$

Admission-level Pretraining

- Intra-modality Mask Code Prediction



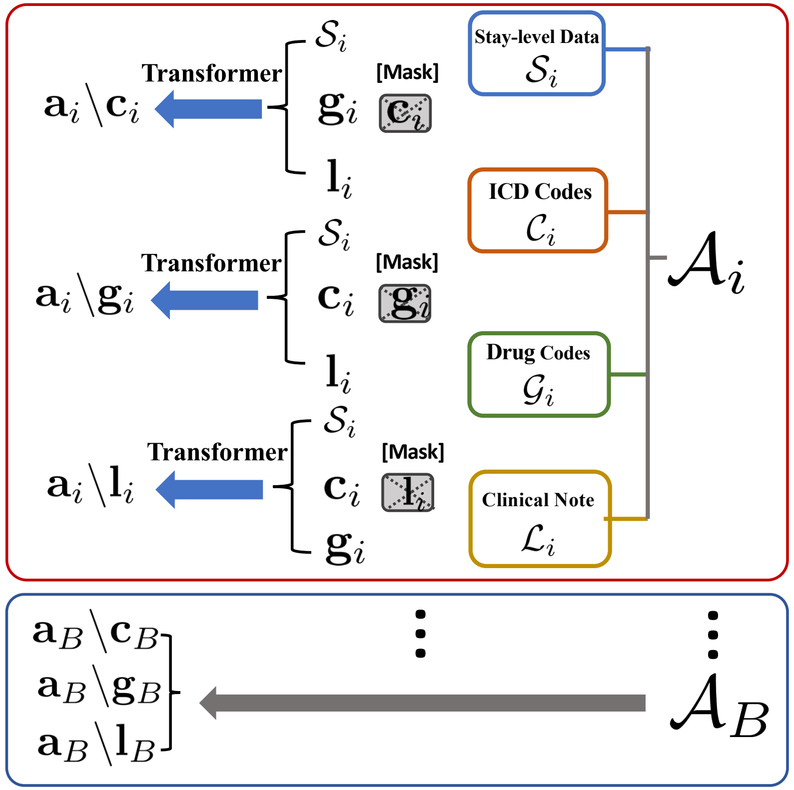
$$\mathcal{L}_{\mathrm{MCP}} = \frac{1}{N}\sum_{i=1}^{N}(||\mathbf{p}_i^c - \mathbf{c}_i^m| + ||\mathbf{p}_i^g - \mathbf{g}_i^m||_2^2),$$

Admission-level Pretraining

- Inter-modality Contrastive Learning

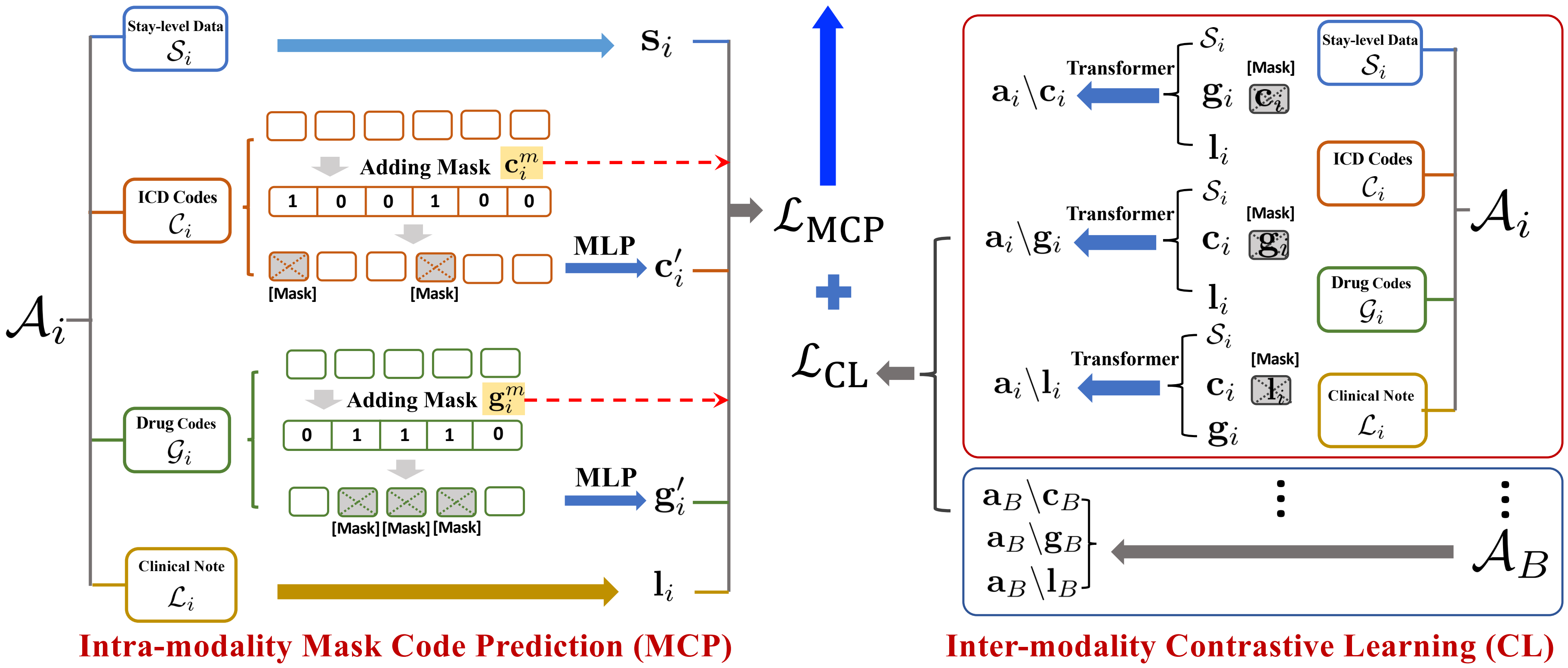
$$\mathcal{L}_{\mathrm{CL}} = \frac{1}{3N}\sum_{i=1}^{N}\sum_{\mathbf{r}_i\in\{\mathbf{c}_i,\mathbf{g}_i,\mathbf{l}_i\}} u(\mathbf{r}_i),$$

$$u(\mathbf{r}_i) = -\log\frac{e^{\mathrm{sim}(\mathbf{r}_i,\mathbf{a}_i\backslash\mathbf{r}_i)/\tau}}{\sum_{j=1,j\neq i}^{B} e^{\mathrm{sim}(\mathbf{r}_i,\mathbf{a}_j\backslash\mathbf{r}_j)/\tau}},$$

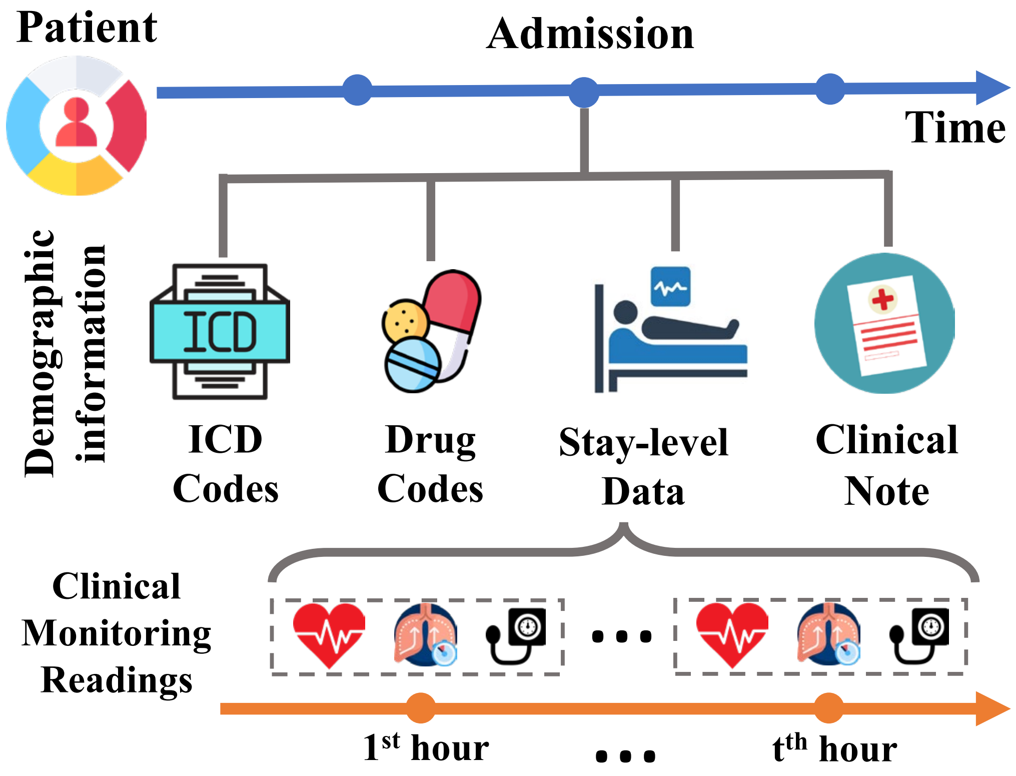


Admission-level Pretraining Loss

$$\mathcal{L}_{\text{admission}} = \mathcal{L}_{\text{MCP}} + \lambda\mathcal{L}_{\text{CL}},$$



Training MedHMP



$$\boxed{\mathcal{L}_{\text{admission}} = \mathcal{L}_{\text{MCP}} + \lambda \mathcal{L}_{\text{CL}},}$$

Convergence

MIMIC-III
MIMIC-IV

$$\boxed{\mathcal{L}_{\text{stay}} = \frac{1}{N * M * T} \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{t=1}^{T} ||\mathbf{m}_{i,t}^{j} - \tilde{\mathbf{m}}_{i,t}^{j}||_2^2.}$$

Experiments

- Stay-level Evaluation
  - Acute respiratory failure (ARF)
  - Shock
  - Mortality
  - Within 48 hours

Task	ARF		Shock		Mortality	
Metric	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
F-LSTM	69.67	10.57	70.28	23.09	81.55	**48.62**
F-CNN	69.61	10.68	69.27	23.51	80.71	42.29
RAIM	59.38	8.42	66.20	20.02	77.17	39.96
DCMN	68.98	10.07	68.68	21.72	80.05	42.93
MEDHMP	**71.66**	**14.34**	**71.04**	**24.19**	**82.17**	47.52

- Admission-level Evaluation
  - Readmission prediction
  - Within 30 days

Model	AUROC	AUPR
BertLstm	63.35	7.24
LstmBert	60.67	6.84
BertCnn	63.07	7.19
CnnBert	61.59	7.04
BertStar	61.28	6.84
StarBert	60.67	6.84
BertEncoder	61.94	6.82
EncoderBert	60.57	7.00
MEDHMP	**67.77**	**9.34**

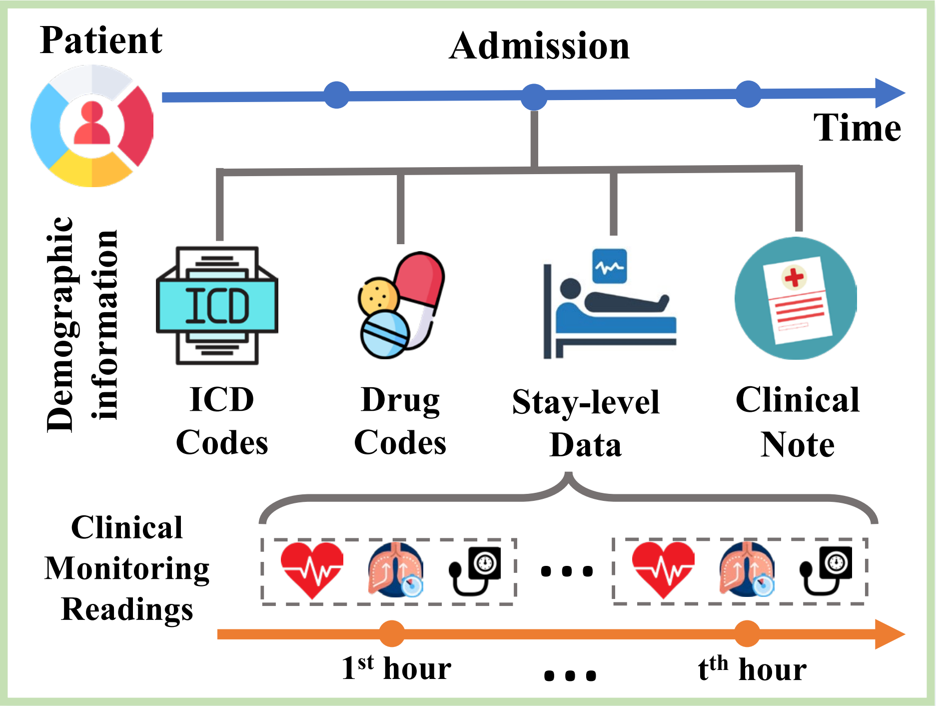
Experiments

- Patient-level Evaluation
  - Health Risk Prediction

Database	**MIMIC-III**			**TriNetX**								
Task	Heart Failure			Heart Failure			COPD			Amnesia		
Metric	AUPR	F1	KAPPA	AUPR	F1	KAPPA	AUPR	F1	KAPPA	AUPR	F1	KAPPA
LSTM_a	**57.83**	**59.40**	**35.86**	**50.16**	**46.08**	**29.26**	**50.16**	**49.34**	**34.64**	**48.68**	**49.64**	**34.46**
LSTM	**57.83**	56.70	33.03	48.20	44.44	26.64	49.52	47.76	33.44	47.92	48.80	32.98
Dipole_a	**59.71**	**60.50**	**37.68**	**47.70**	**41.86**	**25.52**	48.92	**41.06**	**28.30**	**48.74**	**45.78**	**30.78**
Dipole	59.43	58.63	36.03	47.16	40.16	24.28	**49.44**	39.48	27.86	48.36	45.63	30.40
RETAIN_a	**68.71**	**66.20**	**47.12**	**58.16**	**52.18**	**35.64**	**57.62**	**50.66**	**38.36**	**62.70**	**56.50**	**43.90**
RETAIN	67.76	65.56	45.63	57.50	50.88	34.52	57.40	49.85	37.36	62.52	56.32	43.66
AdaCare_a	58.40	**59.47**	35.77	**57.63**	**47.98**	**32.03**	54.06	**47.10**	**34.70**	**62.62**	**52.56**	**41.54**
AdaCare	**59.40**	57.58	**35.84**	55.43	45.13	31.43	**56.63**	46.60	34.53	61.62	50.54	39.22
HiTANet_a	69.42	**68.44**	**50.01**	**60.12**	**50.48**	**36.08**	**64.04**	**54.46**	**43.38**	**67.54**	**58.18**	**47.78**
HiTANet	**70.36**	66.60	46.60	54.76	47.92	32.04	60.10	52.40	39.93	63.08	54.60	43.44

Challenges of EHR Data

Hierarchical



Wang et al., *Hierarchical Pretraining on Multimodal Electronic Health Records*, EMNLP'23

Multi-sourced

Wang et al., *Unity in Diversity: Collaborative Pre-training Across Multimodal Medical Sources*, ACL'24
Wang et al., *FedMeKI: A Benchmark for Scaling Medical Foundation Models via Federated Knowledge Injection*, under review
Wang et al., *FedKIM: Adaptive Federated Knowledge Injection into Medical Foundation Models*, under review

Unity in Diversity: Collaborative Pre-training Across Multimodal Medical Sources

Xiaochen Wang, Junyu Luo, Jiaqi Wang, Yuan Zhong, Xiaokun Zhang, Yaqing Wang, Parminder Bhatia, Cao Xiao and Fenglong Ma

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)

Issues of Existing Work

- Data Scarcity

MIMIC-III

MIMIC-IV

MIM

National Institu

Clinical Center

- Limited Downstream Tasks

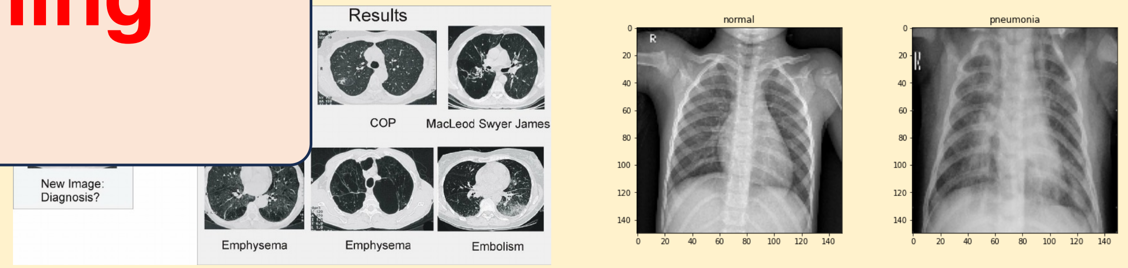
MedHMP

Stay-Level: ARF, Shock, Mortality

sion-Level: Readmission

t-Level: Health Risk

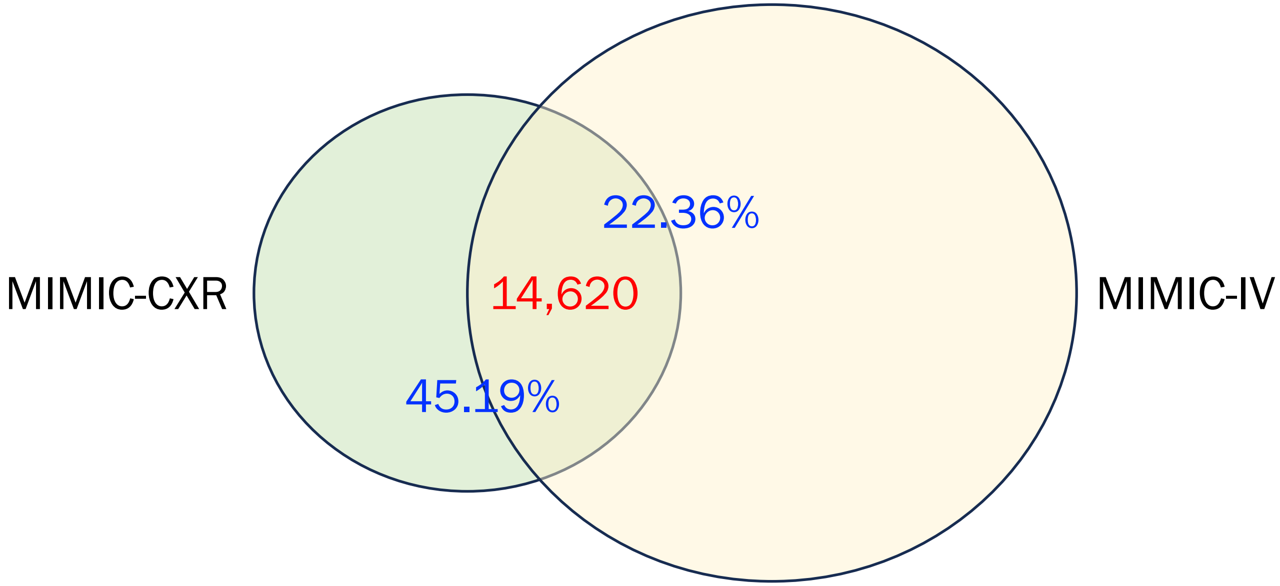
Cross-Source Pretraining



Text-Image Retrieval **Medical Image Classification**

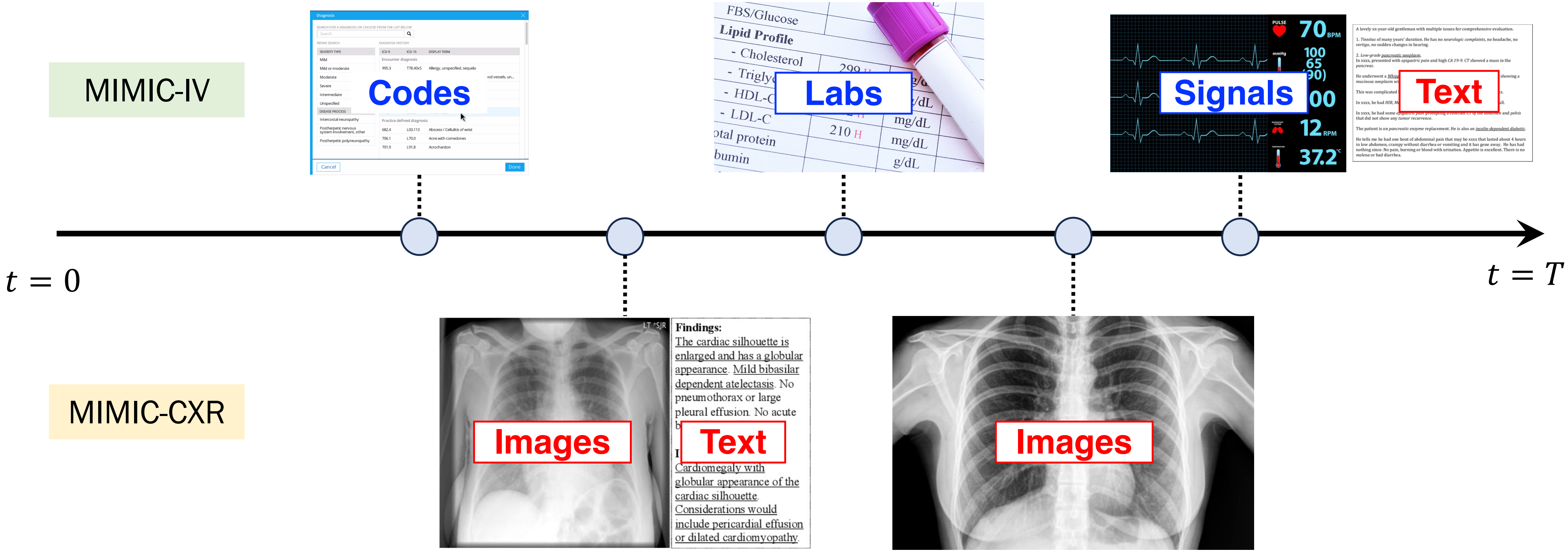
Challenges of Cross-Source Pretraining

- A small portion of overlapped patients



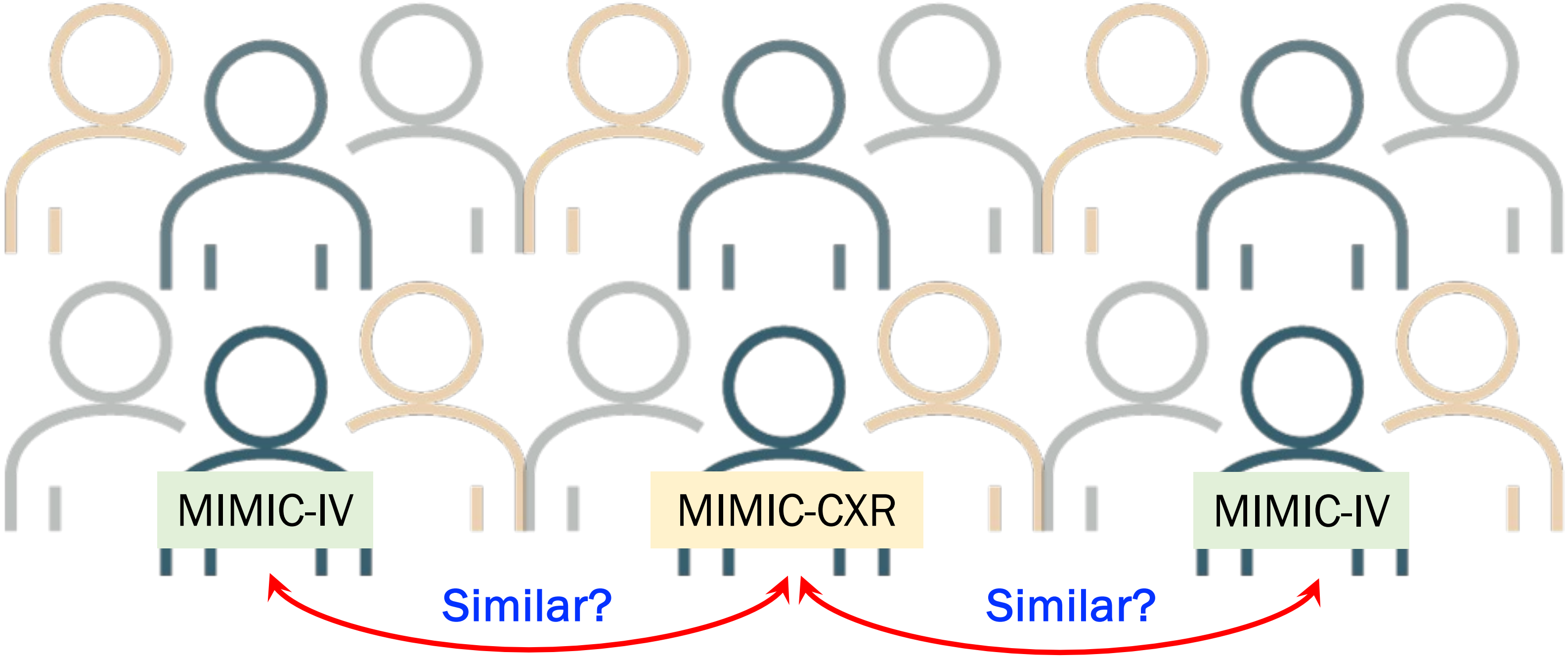
Challenges of Cross-Source Pretraining

- No perfect alignment for a patient’s data from multiple sources

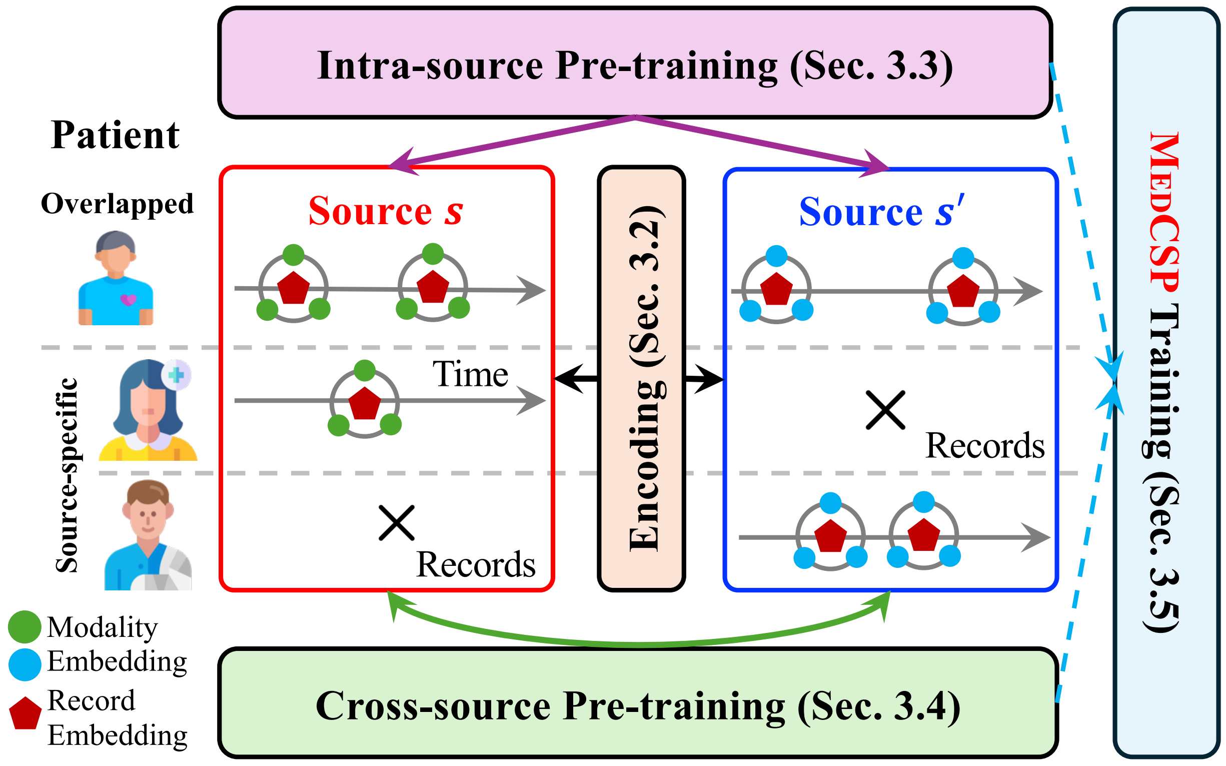


Challenges of Cross-Source Pretraining

- Hard to model implicit yet informative relationships among patients



MedCSP: Medical Cross-Source Pre-training



Intra-source Pre-training

- Modality Encoding

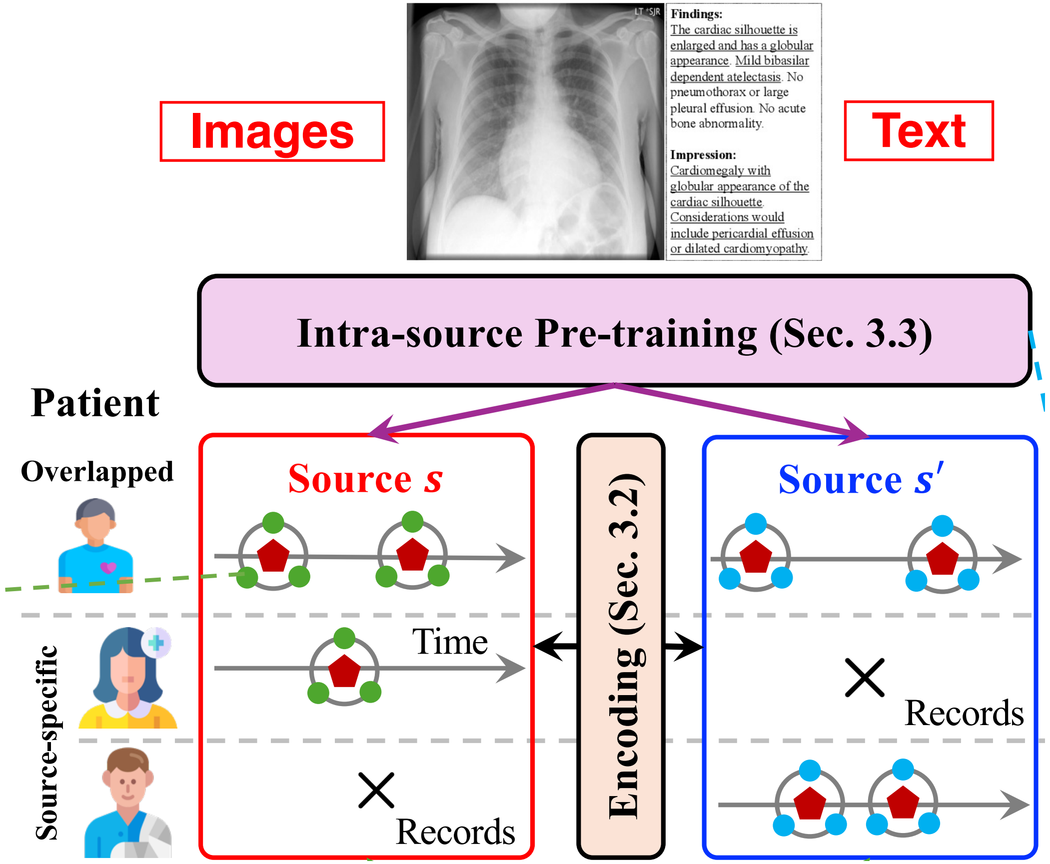
$$\mathbf{e}^{p}_{s,r,m} = \mathrm{Encoder}_{m}(\mathcal{D}^{p}_{s,r,m}),$$

Patient Index

Modality Index

Record Index

Source Index



- Ideal Solution: pair-wise modality-level contrastive learning

High Computational Complexity

Intra-source Pre-training

- Alignment-based Contrastive Learning

Patient Index

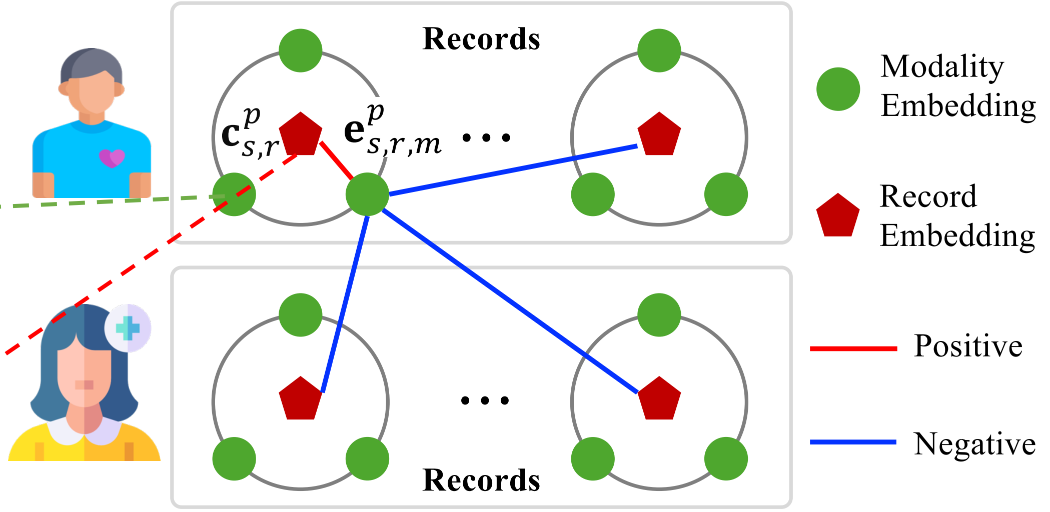
$$\mathbf{e}^{p}_{s,r,m} = \text{Encoder}_{m}(\mathcal{D}^{p}_{s,r,m}),$$

Modality Index

Record Index

Source Index

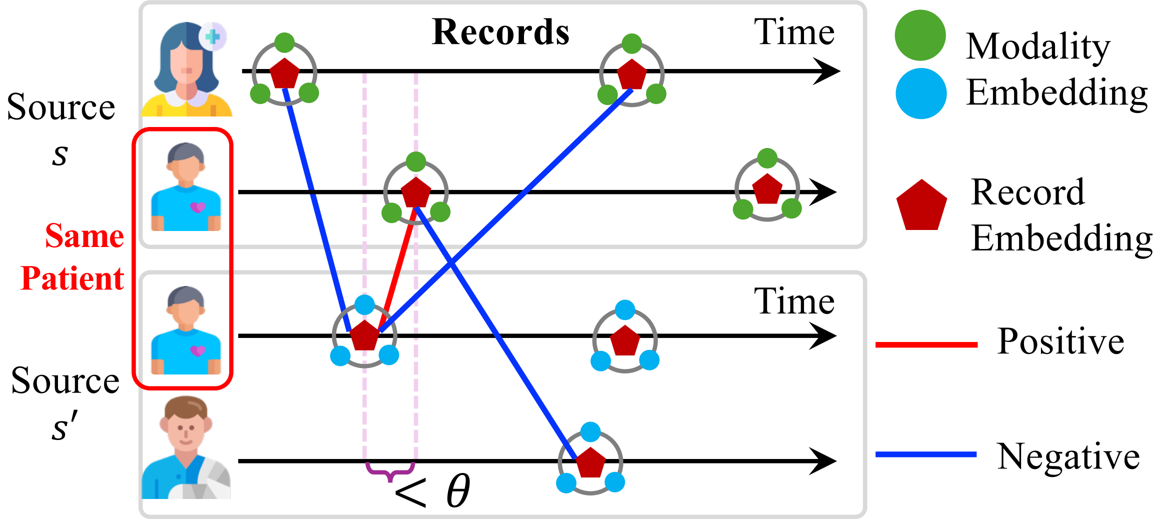
$$\mathbf{c}^{p}_{s,r} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{e}^{p}_{s,r,m},$$



$$\mathcal{L}_a = -\log \frac{\exp(\text{sim}(\mathbf{e}^{p}_{s,r,m}, \mathbf{c}^{p}_{s,r})/\tau)}{\sum_{\mathbf{c}^{p'}_{s',r'} \in \mathcal{N}_a} \exp(\text{sim}(\mathbf{e}^{p}_{s,r,m}, \mathbf{c}^{p'}_{s',r'})/\tau)},$$

Cross-source Pre-training

- NO Explicit Alignment
  - Same Patients Across Different Sources

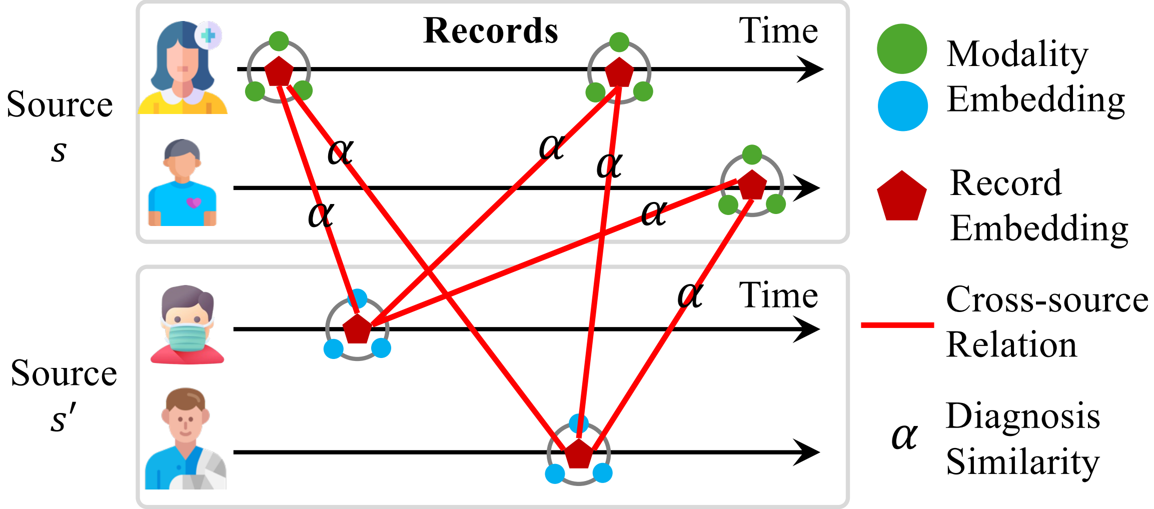


$$\mathcal{L}_p = -\log \frac{\exp(\mathrm{sim}(\mathbf{c}^p_{s,r}, \mathbf{c}^p_{\hat{s},\hat{r}})/\tau)}{\sum_{\mathbf{c}^{p'}_{s',r'} \in \mathcal{N}_p} \exp(\mathrm{sim}(\mathbf{c}^p_{s,r}, \mathbf{c}^{p'}_{s',r'})/\tau)},$$

$$s.t. \quad |T^p_{s,r} - T^p_{\hat{s},\hat{r}}| \leq \theta,$$

Cross-source Pre-training

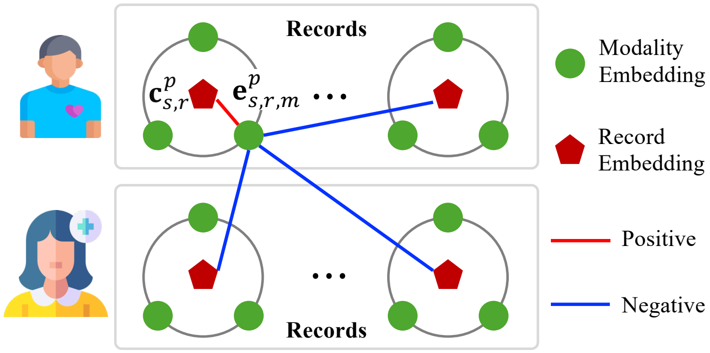
- NO Explicit Alignment
  - Patients with Similar Cohorts Across Different Sources



$$\mathcal{L}_d = -\alpha_{p,\hat{p}} \log \frac{\exp(\mathrm{sim}(\mathbf{c}^p_{s,r}, \mathbf{c}^{\hat{p}}_{\hat{s},\hat{r}})/\tau)}{\sum_{\mathbf{c}^{p'}_{s',r'} \in \mathcal{R}} \exp(\mathrm{sim}(\mathbf{c}^p_{s,r}, \mathbf{c}^{p'}_{s',r'})/\tau)},$$

Diagnosis Similarity $\alpha_{p,\hat{p}} = \mathrm{sim}(\mathbf{h}^p, \mathbf{h}^{\hat{p}}).$

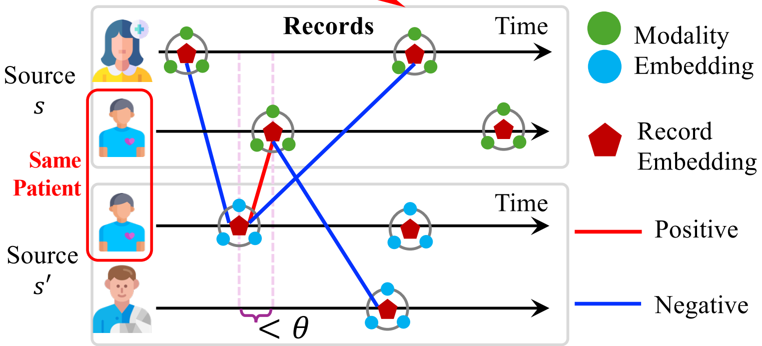
MedCSP Training

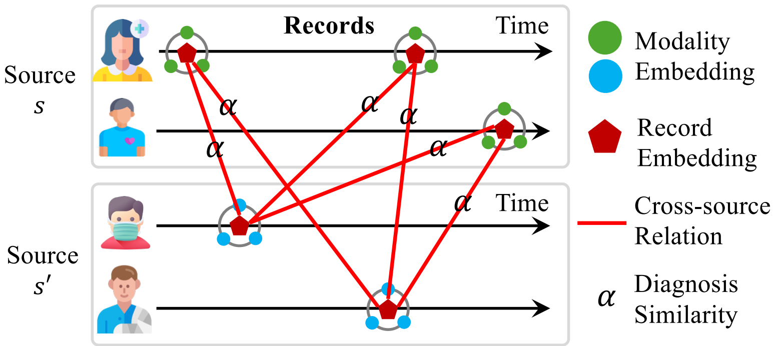


Intra-source Pre-training

$$\mathcal{L} = \mathcal{L}_A + \lambda_P \mathcal{L}_P + \lambda_D \mathcal{L}_D$$

Same Patients





Different Patients

Cross-source Pre-training

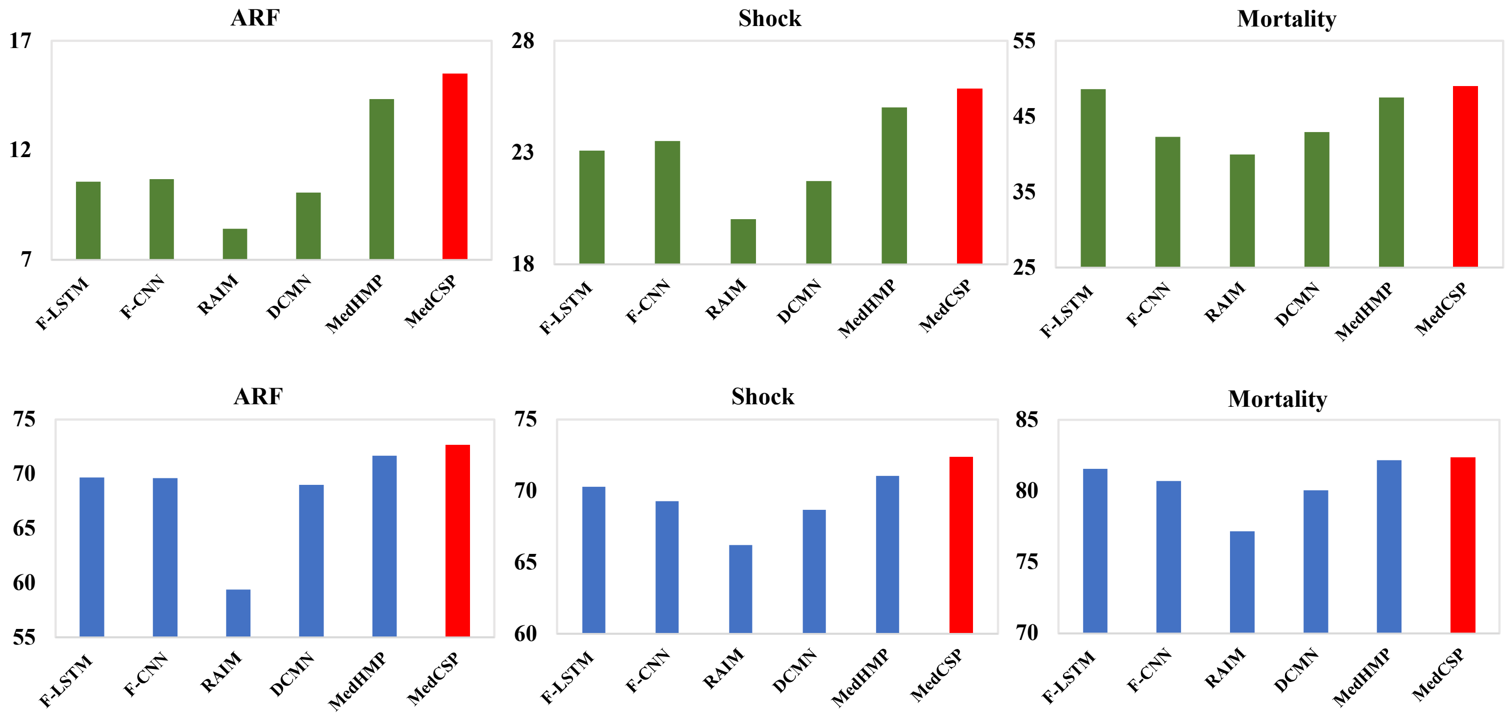
Experiments

- Pre-training and Downstream Data

Stage	Source	Dataset	# of Patients	# of Records		
Pretraining	EHR	MIMIC-IV	32,355	41,230		
	Medical Image	MIMIC-CXR	14,620	156,837		
Downstream	**Source**	**Dataset**	**Predictive Task**	**Total**	**Positive**	**Negative**
	EHR	MIMIC-III	ARF within 48 hours	5,038	402	4,636
			Shock within 48 hours	7,182	693	6,489
			Readmission within 30 days	11,695	1,581	10,114
	Medical Image	MIMIC-CXR	Image Text Retrieval	1,202	-	-
		COVID-19	Image Classification	13,808	3,616	10,192

Evaluation on EHR Source

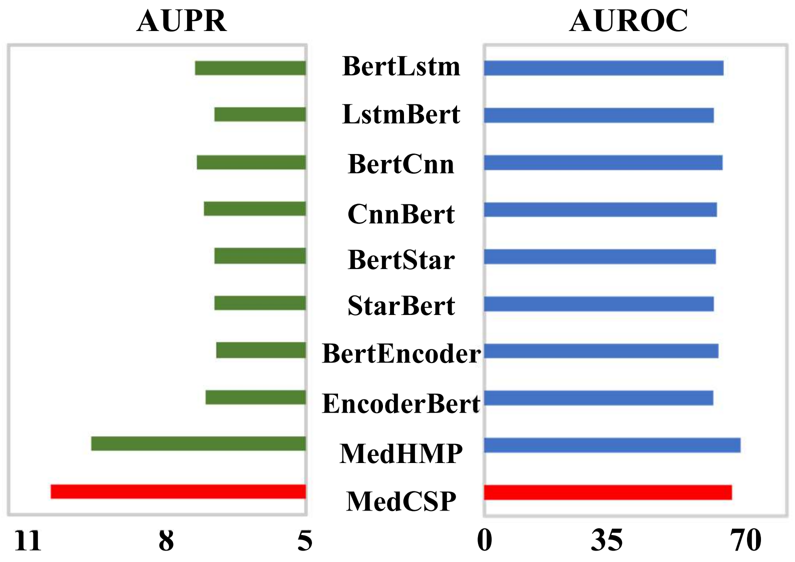
- In-ICU Criticality & Readmission Prediction



MIMIC-III & MIMIC-IV

MedHMP uses more training EHR data than MedCSP for the EHR tasks!

MIMIC-IV



Results (%) of the readmission task

Evaluation on Radiological Source

- Text-image Retrieval
  - Assess the model's ability to associate radiological images with corresponding textual descriptions correctly

Methods	Precision @K						Recall @K					
	1	5	10	20	50	100	1	5	10	20	50	100
CLIP	0.17	0.18	0.17	0.13	0.14	0.12	0.08	0.67	1.16	1.75	4.63	7.79
MedCLIP	0.08	0.10	0.08	0.09	0.08	0.08	0.04	0.23	0.44	1.03	2.07	4.21
BiomedCLIP	0.50	0.53	0.43	0.39	0.31	0.26	0.46	2.29	3.49	5.89	11.79	18.73
PubMedCLIP	0.25	0.13	0.16	0.15	0.15	0.12	0.11	0.39	0.96	1.71	4.30	7.42
CXRCLIP	0.08	0.10	0.11	0.09	0.09	0.08	0.03	0.24	0.58	0.96	2.77	4.61
LLaVAMed	0.17	0.13	0.12	0.12	0.11	0.10	0.11	0.44	0.82	1.66	3.90	7.00
MedCSP	**12.06**	**6.41**	**4.45**	**2.97**	**1.64**	**1.04**	**8.74**	**21.91**	**29.51**	**38.04**	**50.49**	**61.74**

Evaluation on Radiological Source

- Zero-shot Image Classification
  - Categorize medical images into established categories without fine-tuning

Table 3: Performance(%) comparison of the zero-shot image classification task on the COVID-19 dataset.

Methods	Precision	Recall	F1
CLIP	26.01	64.91	37.14
MedCLIP	17.80	37.28	24.10
PubMedCLIP	66.67	0.11	0.22
BiomedCLIP	97.54	21.93	35.80
CXRCLIP	30.49	96.03	47.43
LLaVAMed	26.18	100.00	41.50
MEDCSP	71.98	55.00	**62.36**

Challenges of EHR Data

Overlapped Patients

Training from Scratch

Centralized Training

Wang et al., *Unity in Diversity: Collaborative Pre-training Across Multimodal Medical Sources*, ACL'24

Multi-sourced

Fine-tuning Existing Medical Foundation Models Without Constraints?

Wang et al., *FedMeKI: A Benchmark for Scaling Medical Foundation Models via Federated Knowledge Injection*, under review

Wang et al., *FedKIM: Adaptive Federated Knowledge Injection into Medical Foundation Models*, under review

FedMeKI: A Benchmark for Scaling Medical Foundation Models via Federated Knowledge Injection

Jiaqi Wang∗, Xiaochen Wang∗, Lingjuan Lyu, Jinghui Chen and Fenglong Ma

(under review)

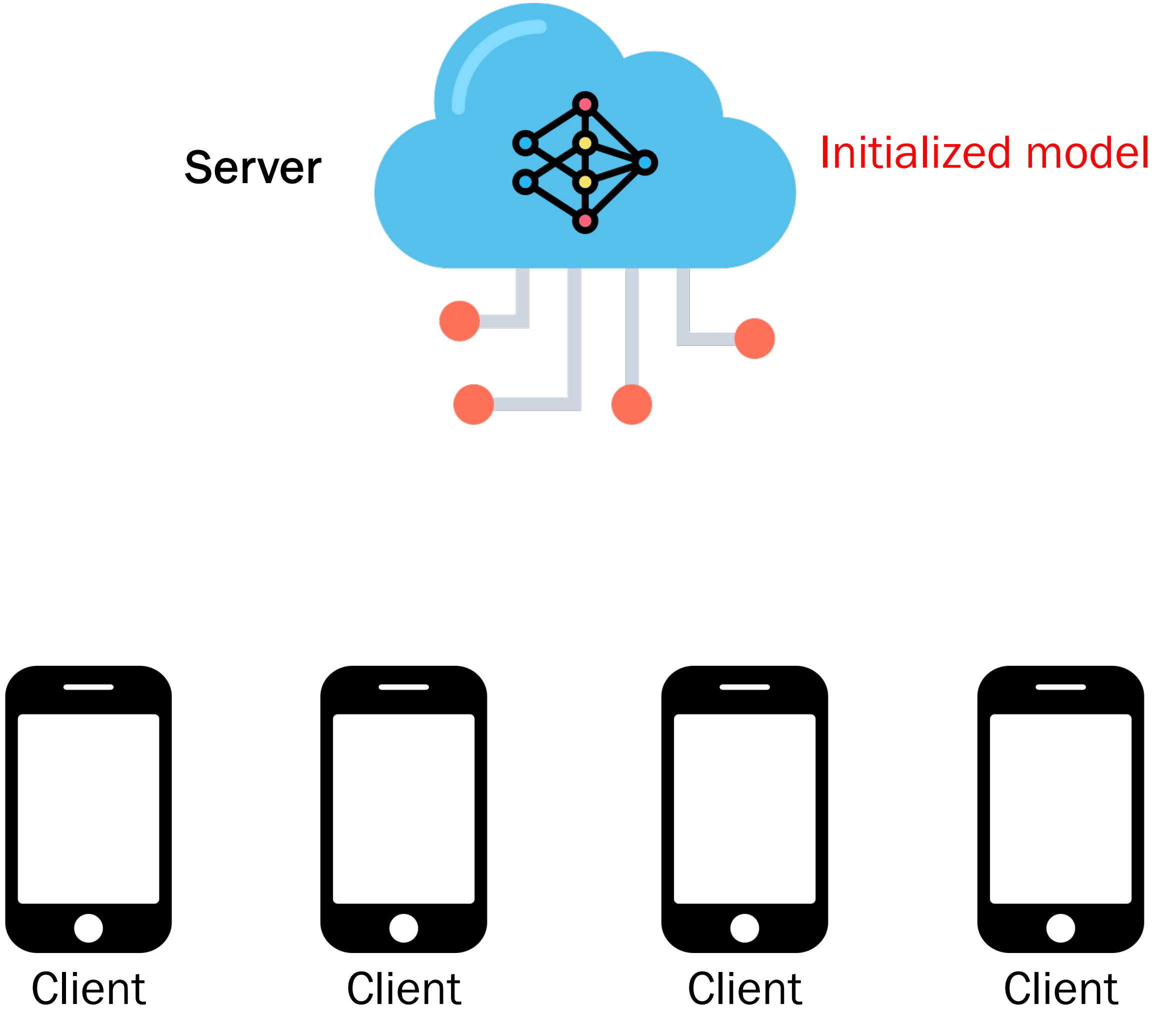
FEDKIM: Adaptive Federated Knowledge Injection into Medical Foundation Models

Xiaochen Wang∗, Jiaqi Wang∗, Houping Xiao, Jinghui Chen and Fenglong Ma

(under review)

Federated Learning

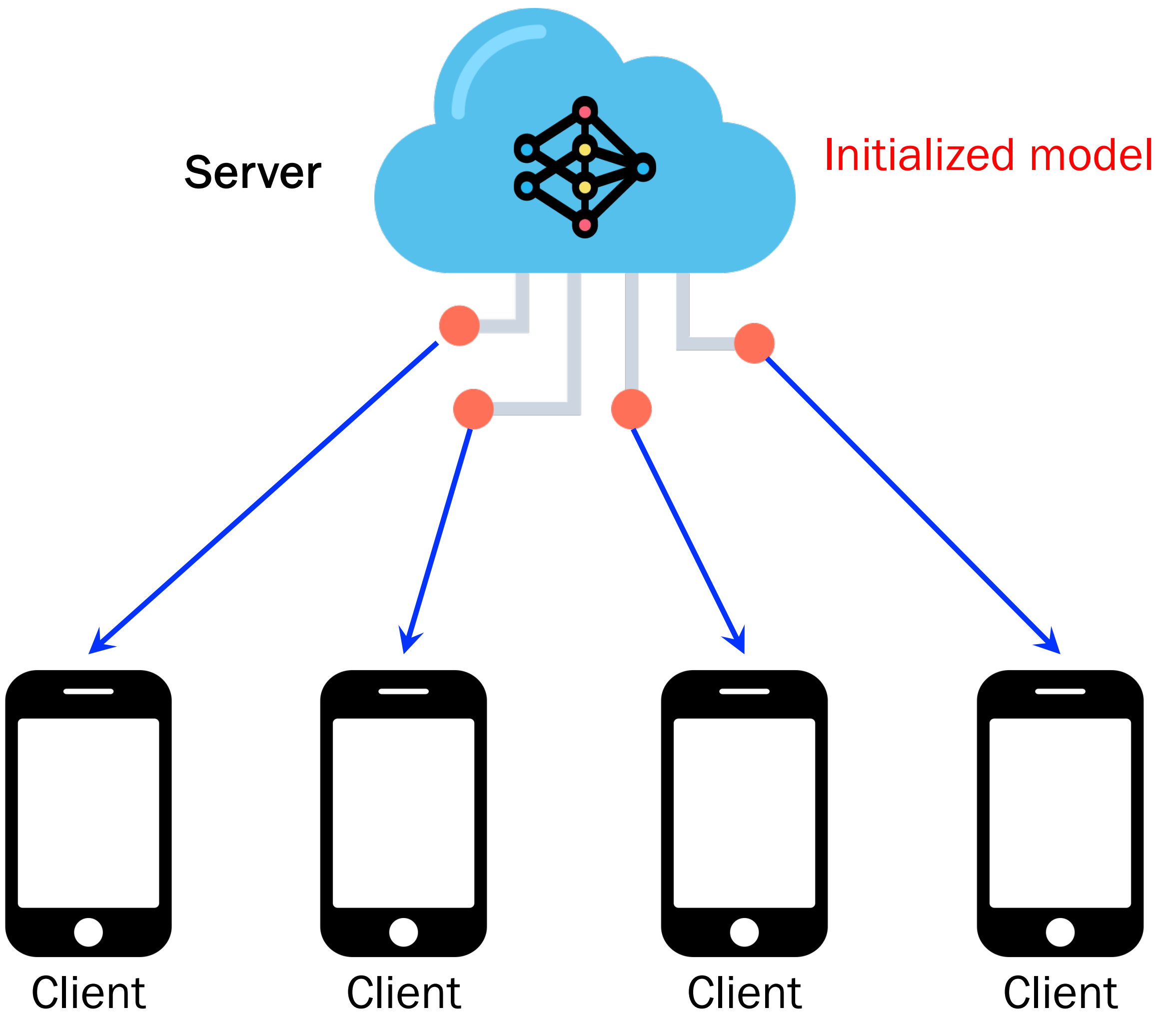
- Federated Learning (FL) aims to collaboratively train a machine learning (ML) model while keep the data decentralized.



McMahan et al. "Communication-efficient learning of deep networks from decentralized data." Artificial intelligence and statistics. PMLR, 2017.

Federated Learning

- Federated Learning (FL) aims to collaboratively train a machine learning (ML) model while keep the data decentralized.

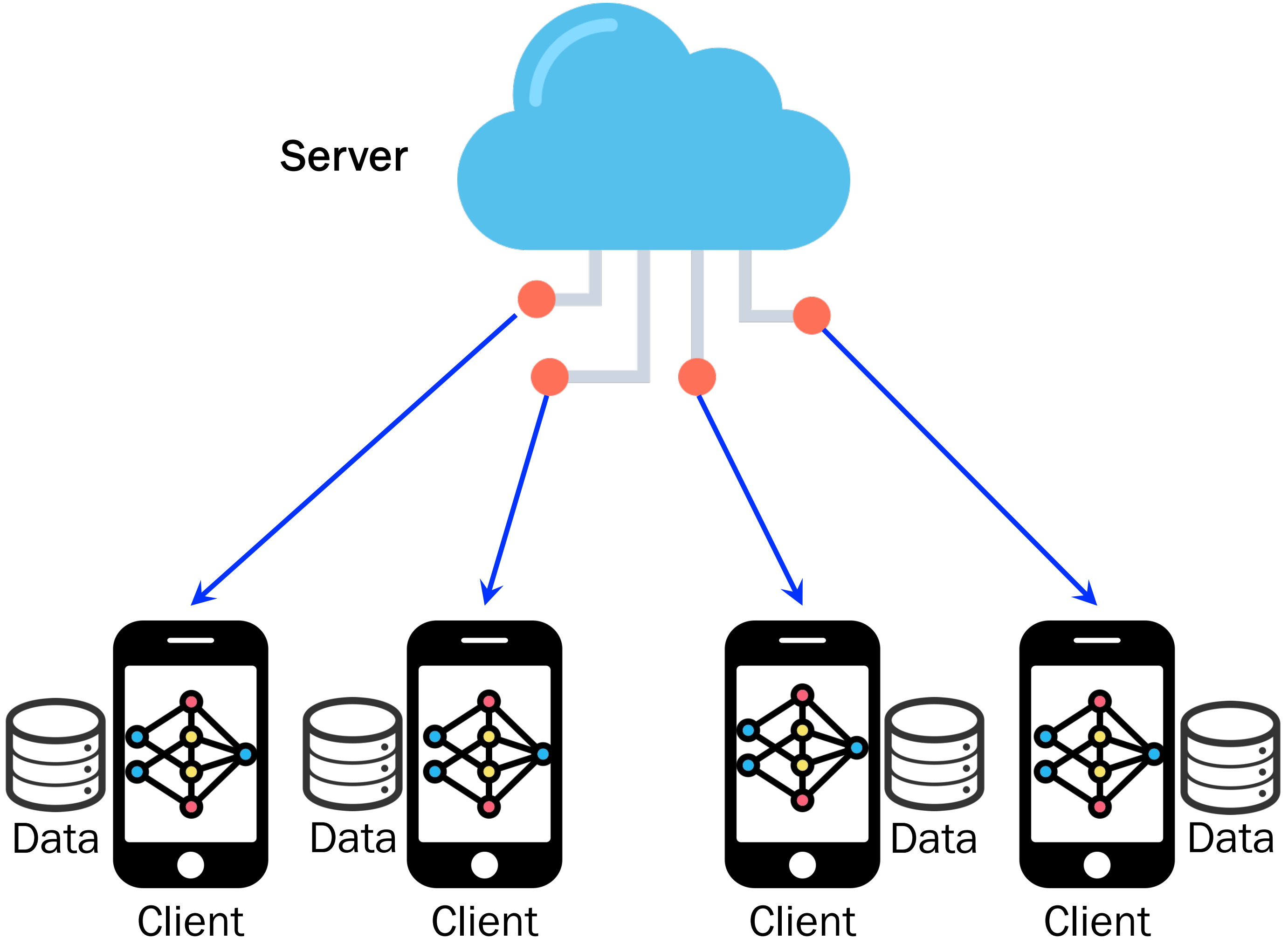


McMahan et al. "Communication-efficient learning of deep networks from decentralized data." Artificial intelligence and statistics. PMLR, 2017.

Federated Learning

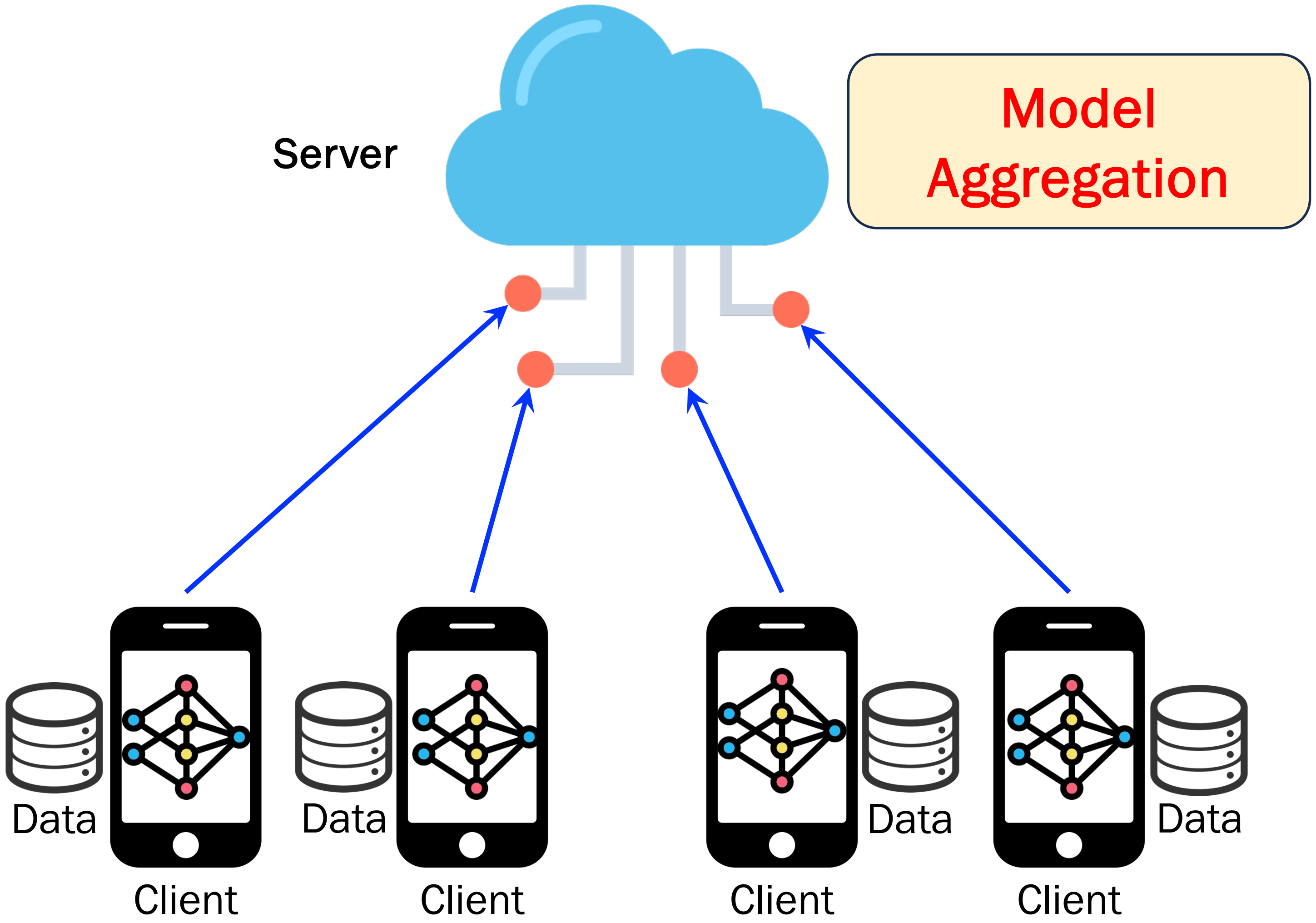
- Federated Learning (FL) aims to collaboratively train a machine learning (ML) model while keep the data decentralized.

Local Model Training



Federated Learning

- Federated Learning (FL) aims to collaboratively train a machine learning (ML) model while keep the data decentralized.

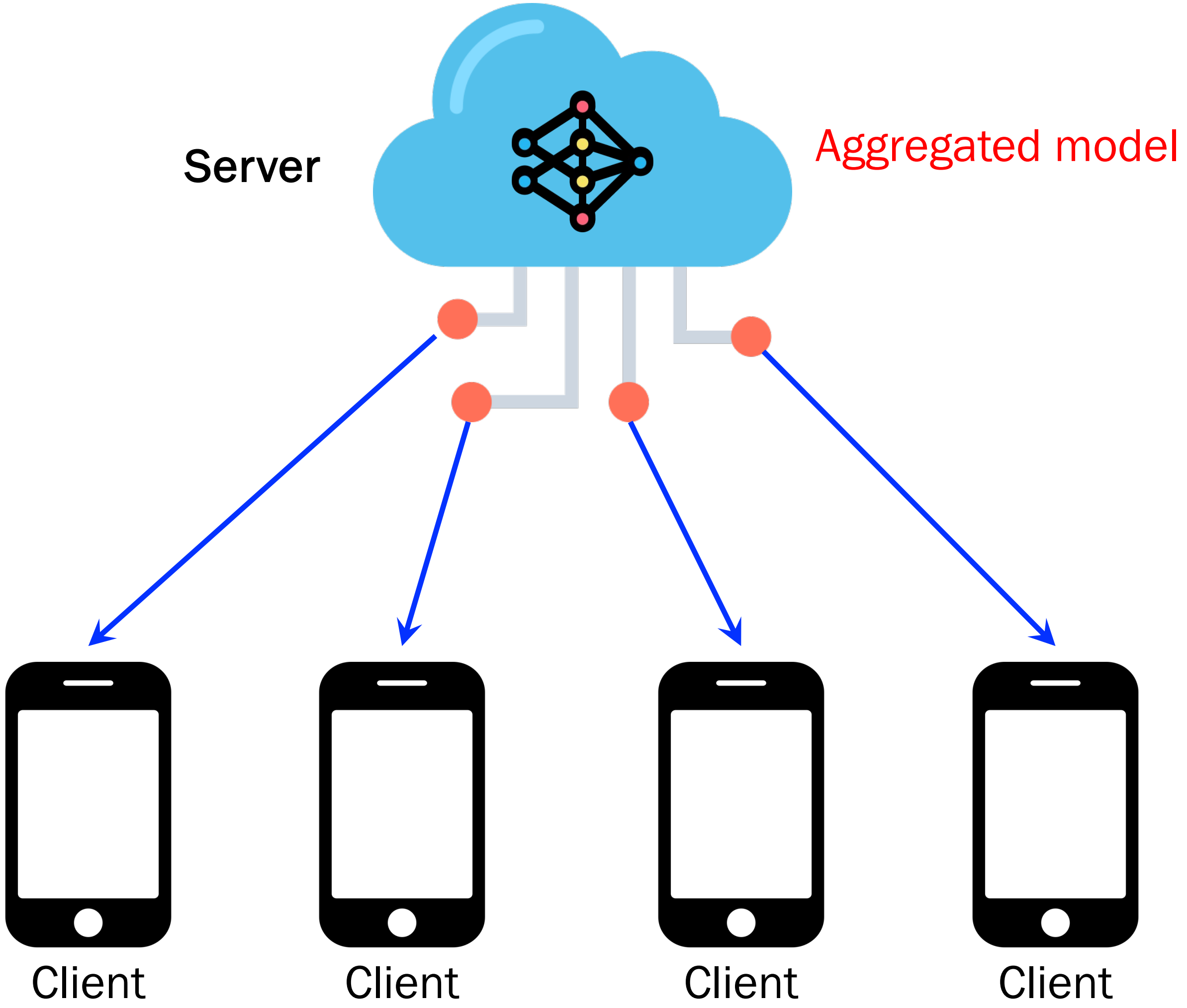


McMahan et al. "Communication-efficient learning of deep networks from decentralized data." Artificial intelligence and statistics. PMLR, 2017.

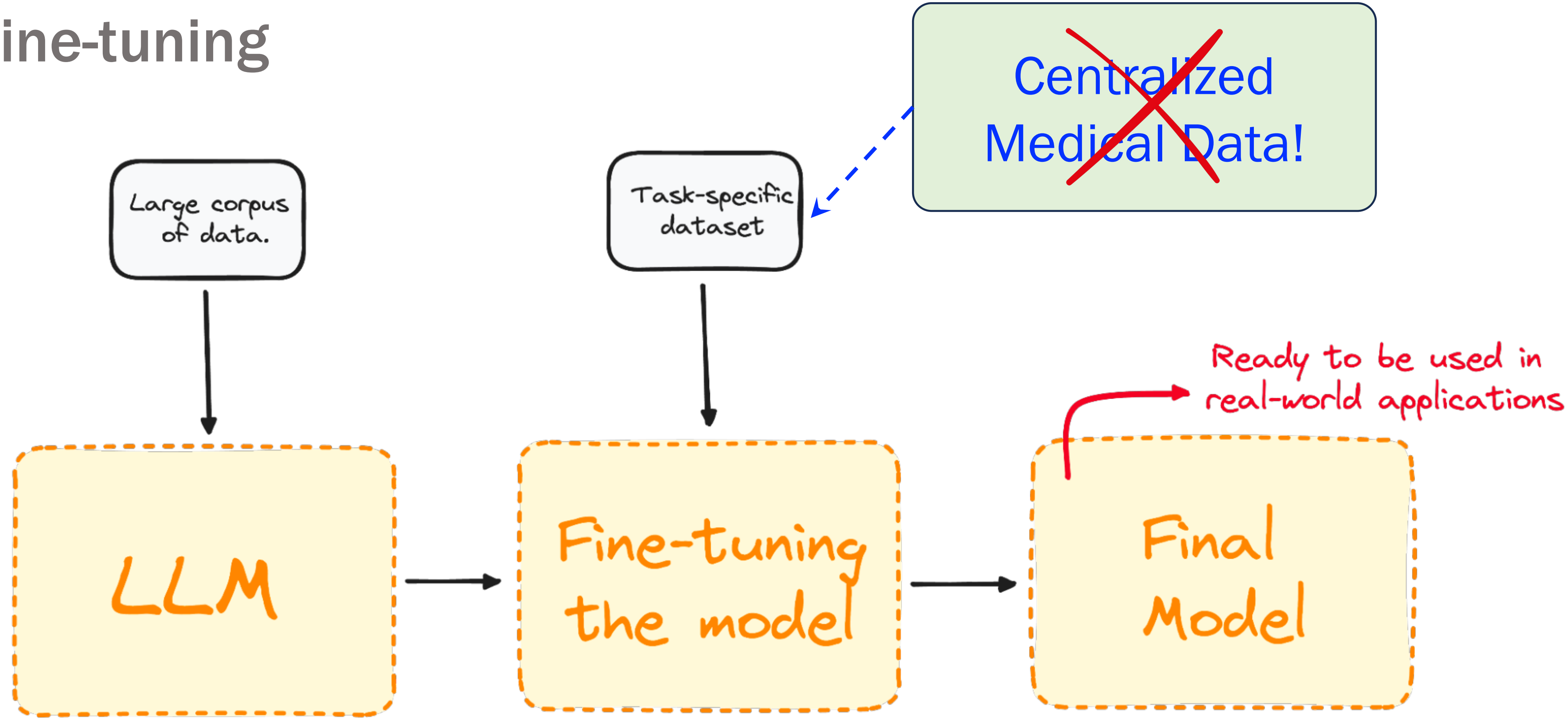
Federated Learning

- Federated Learning (FL) aims to collaboratively train a machine learning (ML) model while keep the data decentralized.

We would like the final aggregated model to be as good as the centralized solution (ideally), or at least better than what each client can learn on its own



Fine-tuning



https://www.datacamp.com/tutorial/fine-tuning-large-language-models

Existing Medical Foundation Models

Medical Foundation Model	Modalities	Tasks
MMedLM2 (Qiu et al., 2024)	Text	Question-answering
LLava-Med(Liu et al., 2023a)	Text, Image	Visual Question-answering
Med-Flamingo(Yang et al., 2023)	Text, Image	Visual Question-answering
PMC_LLAMA(Lee et al., 2023)	Text	Question-answering
BiomedGPT(Gu et al., 2021)	Text, Image	Visual Question-answering
BioMedLM(Lewis et al., 2020)	Text	Question-answering
GatorTron(Hao et al., 2020)	Text	Clinical concept extraction Medical relation extraction Semantic textual similarity Natural language inference Question-answering
Med-PaLM(Singhal et al., 2022)	Text	Question-answering
ChatDoctor(Li et al., 2023)	Text	Question-answering

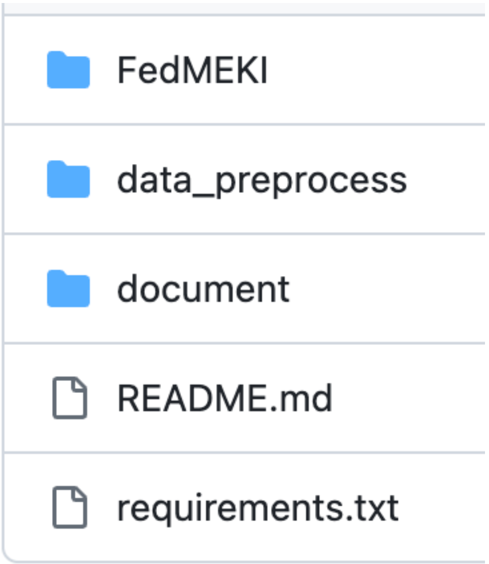
Limited modality and task adaptability

FedMeKI

- A New Task
  - Federated Medical Knowledge Injection into Medical Foundation Models

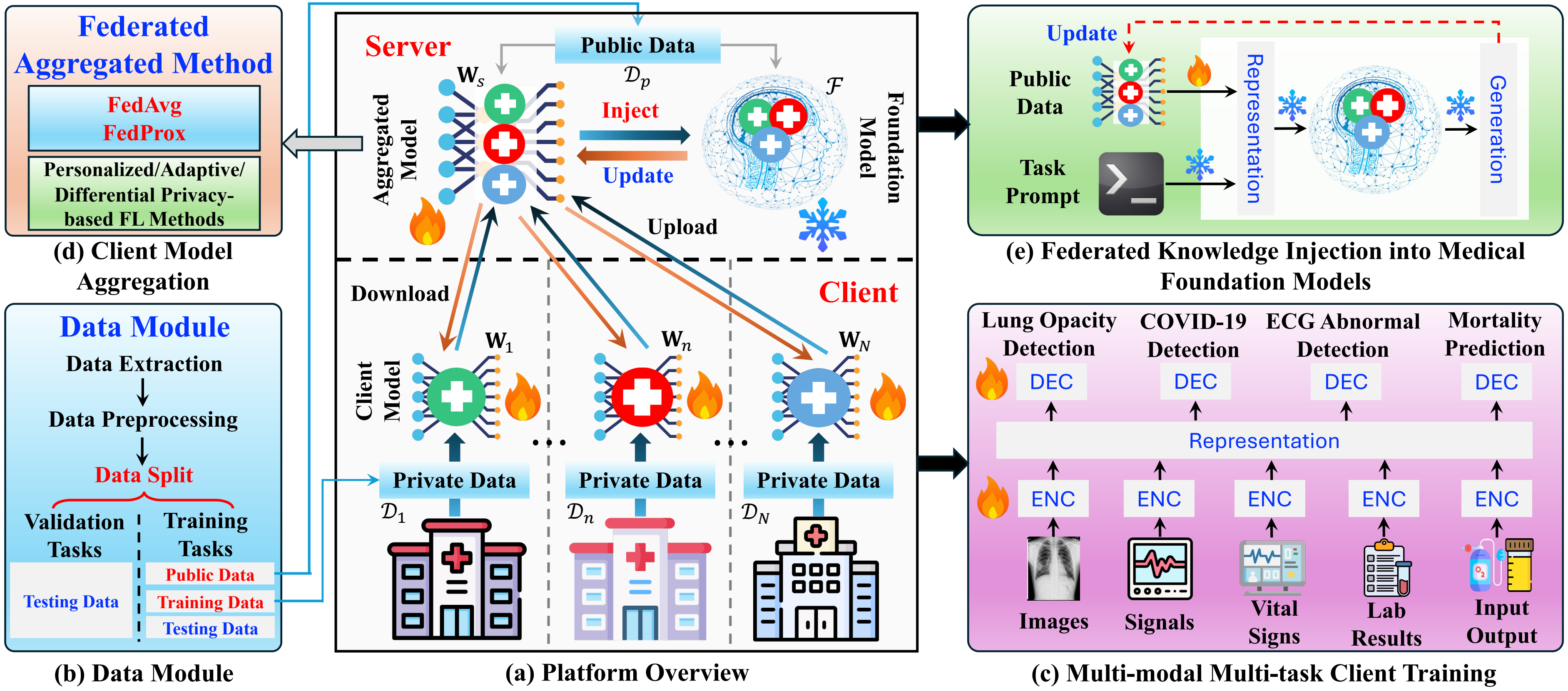
- A Comprehensive Medical Dataset
  - Eight Medical Tasks
  - Seven Medical Modalities

https://github.com/psudslab/FEDMEKI



FedMeKI

- A Novel Federated Knowledge Injection Platform

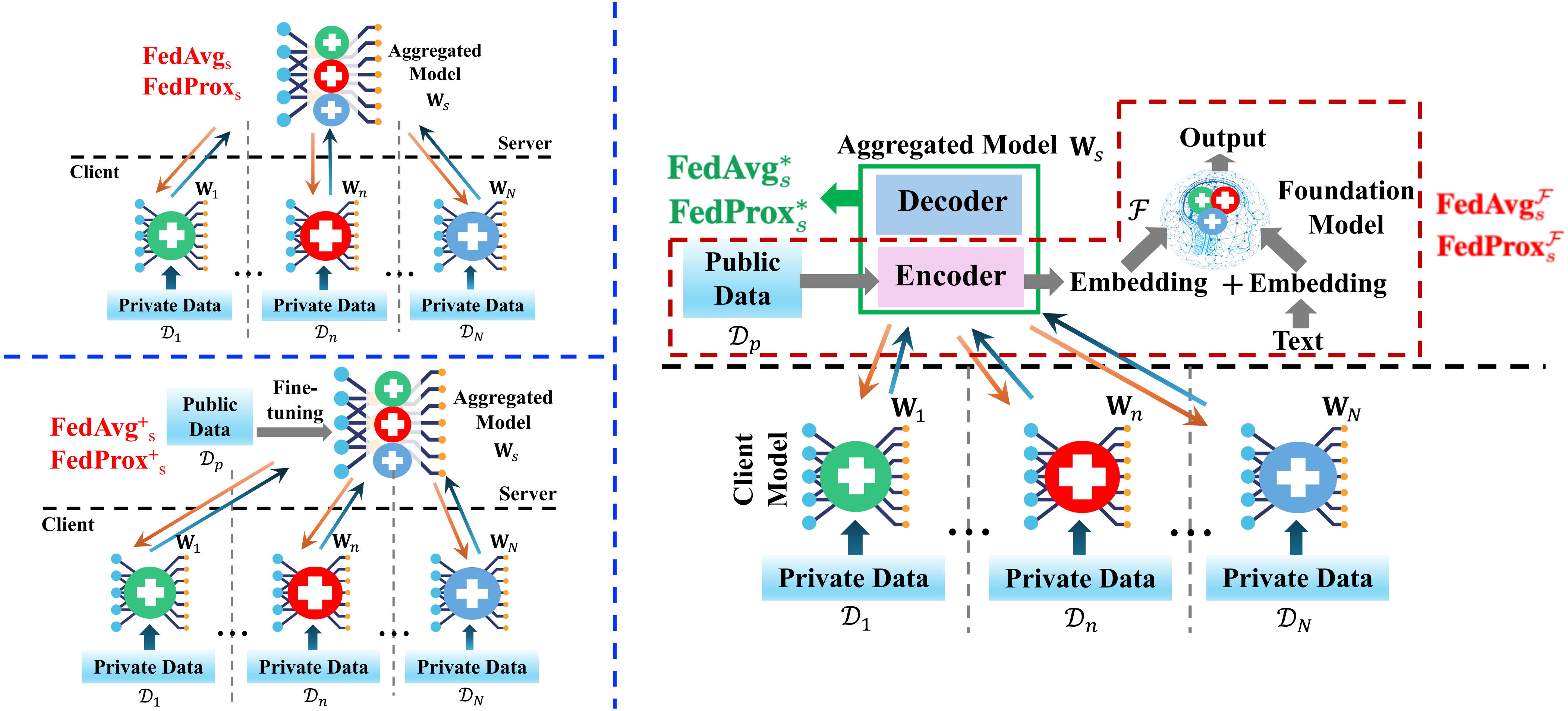


Data Preparation

Table 1: Details of data split, where we deploy 5 clients on the FEDMEKI platform.

Type	Task	Total Samples	Total Training (5 Clients)	Public Data (Server)	Development (Server)	Testing (Server)
Training Tasks	Lung Opacity Detection	18,406	12,880	1,849	1,841	1,836
	COVID-19 Detection	13,808	9,665	1,380	1,380	1,383
	ECG Abnormal Detection	21,797	15,259	2,179	2,180	2,179
	Mortality Prediction	38,129	26,690	3,812	3,812	3,813
Validation Tasks	Enlarged Cardiomediastinum Detection	234	✗	✗	✗	234
	Sepsis Prediction	1,000	✗	✗	✗	1,000
	MedVQA	1,000	✗	✗	✗	1,000
	Signal Noise Clarification	1,000	✗	✗	✗	1,000

Single-task evaluation for training tasks

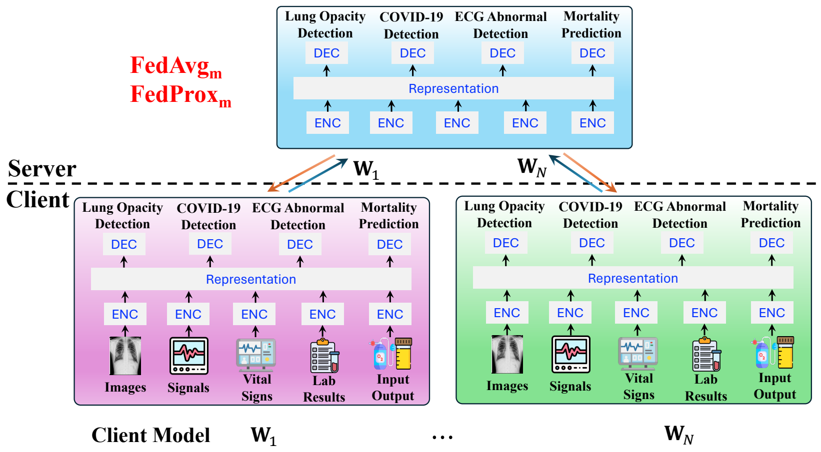


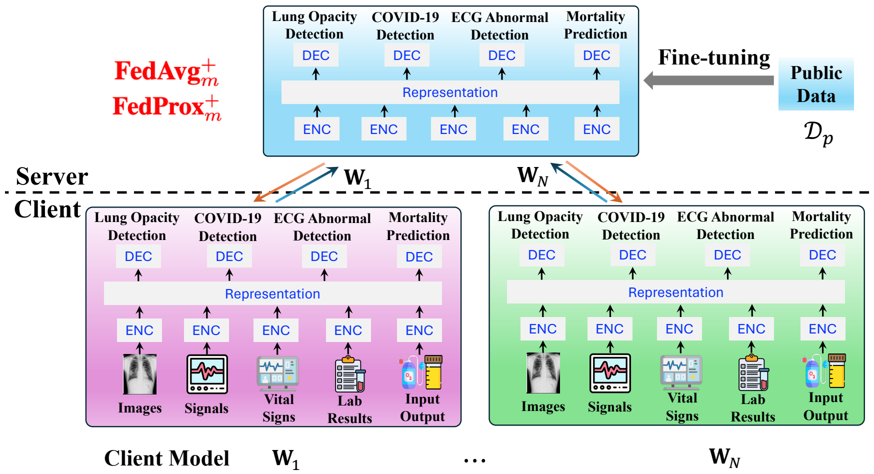
Single-task evaluation for training tasks

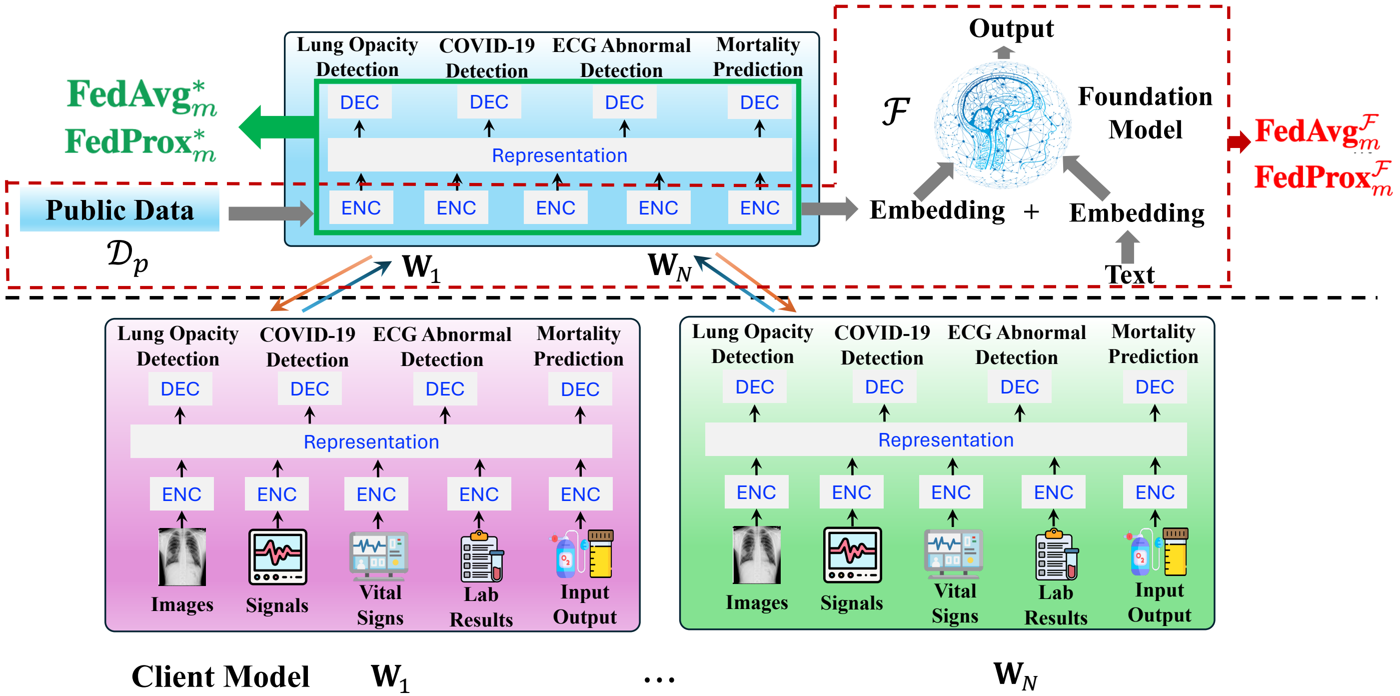
Table 2: Benchmark performance of single-task evaluation for training tasks.

Task	Metric	MMedLM-2	FedAvg				FedProx			
			FedAvg_s	FedAvg_s^{+}	FedAvg_s^{*}	$\mathrm{FedAvg}_s^{\mathcal{F}}$	$\mathrm{FedProx}_s$	$\mathrm{FedProx}_s^{+}$	$\mathrm{FedProx}_s^{*}$	$\mathrm{FedProx}_s^{\mathcal{F}}$
Lung Opacity Detection	Accuracy	✗	95.86	94.44	96.02	89.42	95.70	96.08	95.70	91.23
	Precision	✗	97.40	93.81	96.70	84.69	97.49	97.11	95.23	87.76
	Recall	✗	94.01	95.58	95.58	97.16	94.11	95.27	96.53	96.52
	F1	✗	95.31	94.69	96.14	90.50	95.77	96.18	95.87	91.93
COVID-19 Detection	Accuracy	✗	99.35	99.48	99.28	92.34	99.13	99.42	99.13	84.16
	Precision	✗	99.71	99.70	100.00	93.59	99.71	99.42	99.71	77.27
	Recall	✗	97.72	94.30	97.15	74.92	96.87	98.29	96.87	53.27
	F1	✗	98.71	96.93	98.55	79.15	98.21	98.85	98.27	63.07
ECG Abnormal Detection	Accuracy	✗	67.68	66.83	57.86	43.15	79.41	80.51	57.77	45.25
	Precision	✗	69.13	80.65	89.56	56.97	89.04	89.06	87.34	60.85
	Recall	✗	80.78	56.24	31.61	11.22	73.88	76.00	32.47	17.80
	F1	✗	74.50	66.27	46.72	18.74	80.75	82.01	47.34	27.55
Mortality Prediction	Accuracy	✗	91.98	91.66	91.61	84.11	91.98	90.12	91.61	82.41
	Precision	✗	70.00	52.86	58.33	16.35	71.05	36.45	58.33	13.87
	Recall	✗	8.70	11.42	2.17	21.43	8.39	22.98	2.17	16.64
	F1	✗	15.47	18.88	4.19	18.55	15.00	28.19	4.19	15.13

Multi-task evaluation for training tasks







Multi-task evaluation for training tasks

Table 3: Benchmark performance of multi-task evaluation for training tasks. Note that the performance of **ECG Abnormal Detection** and **Mortality Prediction** is the same as that shown in Table 2 since the modalities of these two tasks are non-overlapped with others.

Task	Metric	MMedLM-2	FedAvg				FedProx			
			$\textbf{FedAvg}_m$	$\textbf{FedAvg}_m^{+}$	$\textbf{FedAvg}_m^{*}$	$\textbf{FedAvg}_m^{\mathcal{F}}$	$\textbf{FedProx}_m$	$\textbf{FedProx}_m^{+}$	$\textbf{FedProx}_m^{*}$	$\textbf{FedProx}_m^{\mathcal{F}}$
Lung Opacity Detection	Accuracy	✗	95.42	94.23	94.88	95.48	94.77	96.24	96.41	93.13
	Precision	✗	99.66	93.51	93.68	98.22	99.54	97.22	97.63	93.74
	Recall	✗	91.48	95.48	96.64	92.95	90.33	95.48	95.37	92.95
	F1	✗	95.39	94.48	95.13	95.51	94.71	96.34	96.49	93.35
COVID-19 Detection	Accuracy	✗	99.06	98.99	99.28	98.34	99.06	99.20	98.99	86.11
	Precision	✗	99.42	98.56	99.14	96.07	99.13	98.85	98.56	65.09
	Recall	✗	96.87	97.44	98.01	97.44	97.15	98.01	97.44	97.72
	F1	✗	98.12	97.99	98.57	96.75	98.13	98.43	97.99	78.13

Zero-shot evaluation for validation tasks

Task (Modalities)	Metric	MMedLM-2	$\text{FedAvg}_m^{\mathcal{F}}$	$\text{FedProx}_m^{\mathcal{F}}$
Enlarged Cardiomediastinum Detection (medical image)	Accuracy	✗	58.54	57.26
	Precision	✗	53.33	52.57
	Recall	✗	88.07	84.40
	F1	✗	66.04	64.78
Sepsis Prediction (48 clinical features)	Accuracy	✗	39.00	39.80
	Precision	✗	2.61	3.57
	Recall	✗	55.17	75.86
	F1	✗	4.98	6.81
MedVQA (medical image + text)	BLEU	✗	1.20	1.20
	ROUGE	✗	2.43	3.42
	METEOR	✗	1.07	2.83
Signal Noise Clarification (signal + text)	BLEU	✗	0.06	0.04
	ROUGE	✗	0.29	0.23
	METEOR	✗	1.88	0.63

FedMeKI: A Benchmark for Scaling Medical Foundation Models via Federated Knowledge Injection

Jiaqi Wang*, Xiaochen Wang*, Lingjuan Lyu, Jinghui Chen and Fenglong Ma

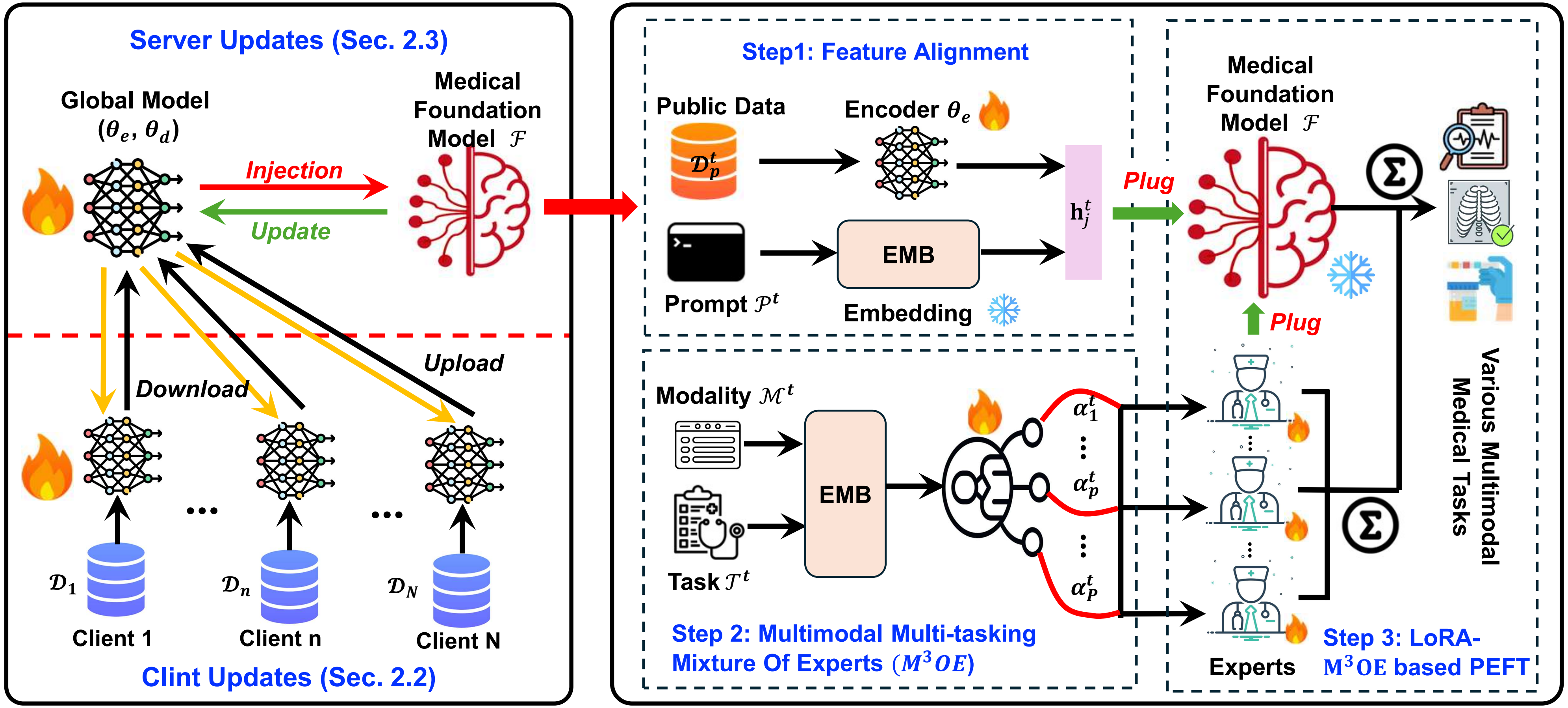
(under review)

FEDKIM: Adaptive Federated Knowledge Injection into Medical Foundation Models

Xiaochen Wang*, Jiaqi Wang*, Houping Xiao, Jinghui Chen and Fenglong Ma

(under review)

A More Advanced Model



(a) Framework Overview

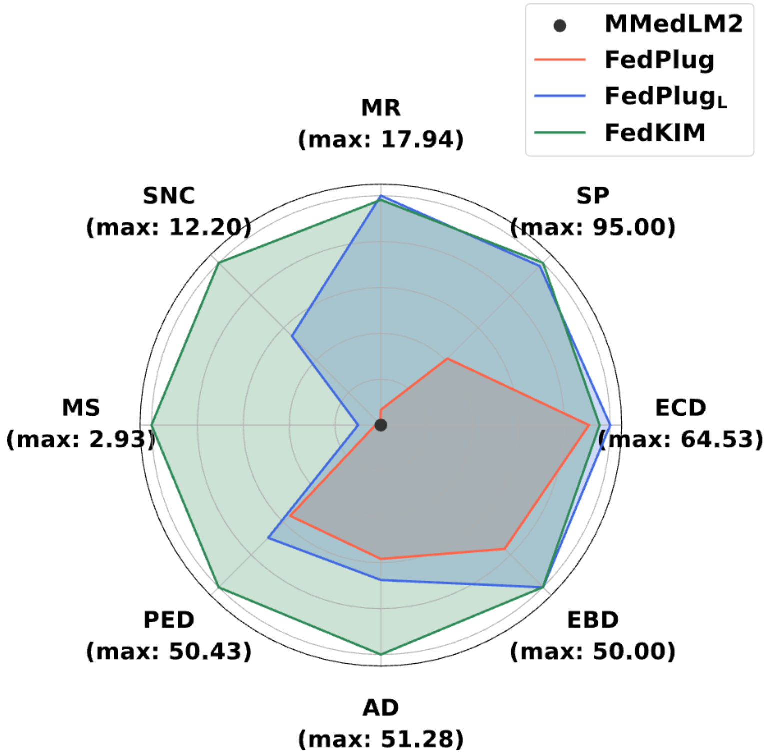
(b) Federated Knowledge Injection into Medical Foundation Model

Tasks and Modalities

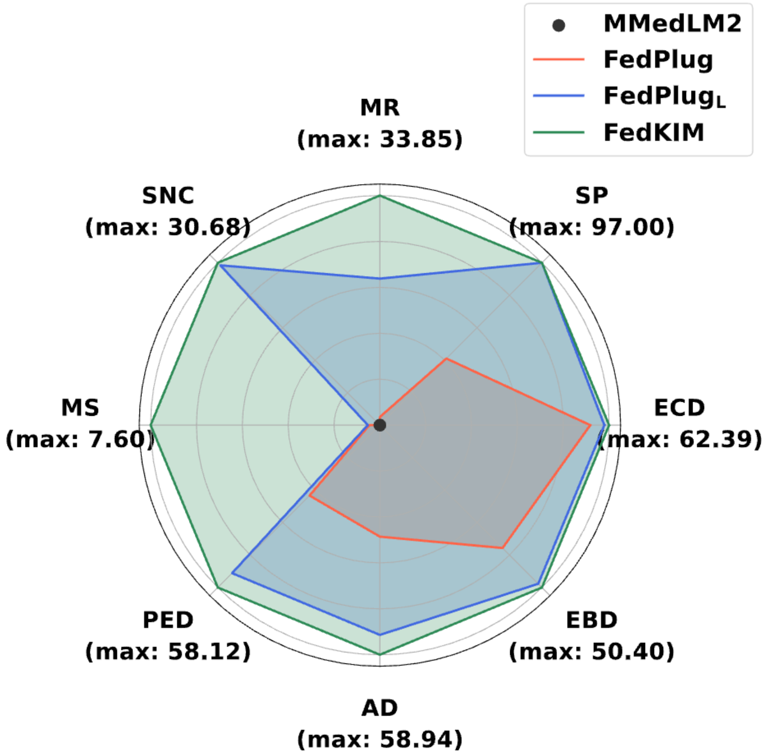
Table 2: Tasks and modalities in this study.

Task Type	Task	Modality						
		Image	**Signal**	**Vital signs**	**Lab events**	**Input**	**Output**	**Text**
Training	COVID-19 Detection (**CD**)	✓	✗	✗	✗	✗	✗	✗
	Lung Opacity Detection (**LOD**)	✓	✗	✗	✗	✗	✗	✗
	ECG Abnormal Detection (**EAD**)	✗	✓	✗	✗	✗	✗	✗
	Mortality Prediction (**MP**)	✗	✗	✓	✓	✓	✓	✗
Validation	Enlarged Cardiomediastinum Detection (**ECD**)	✓	✗	✗	✗	✗	✗	✗
	Pleural Effusion Detection (**PED**)	✓	✗	✗	✗	✗	✗	✗
	Atelectasis Detection (**AD**)	✓	✗	✗	✗	✗	✗	✗
	Ectopic Beats Detection (**EBD**)	✗	✓	✗	✗	✗	✗	✗
	Sepsis Prediction (**SP**)	✗	✗	✓	✓	✓	✓	✗
	MedVQA-RAD (**MR**)	✓	✗	✗	✗	✗	✗	✓
	MedVQA-Slake (**MS**)	✓	✗	✗	✗	✗	✗	✓
	Signal Noise Clarification (**SNC**)	✗	✓	✗	✗	✗	✗	✓

Zero-shot Evaluation



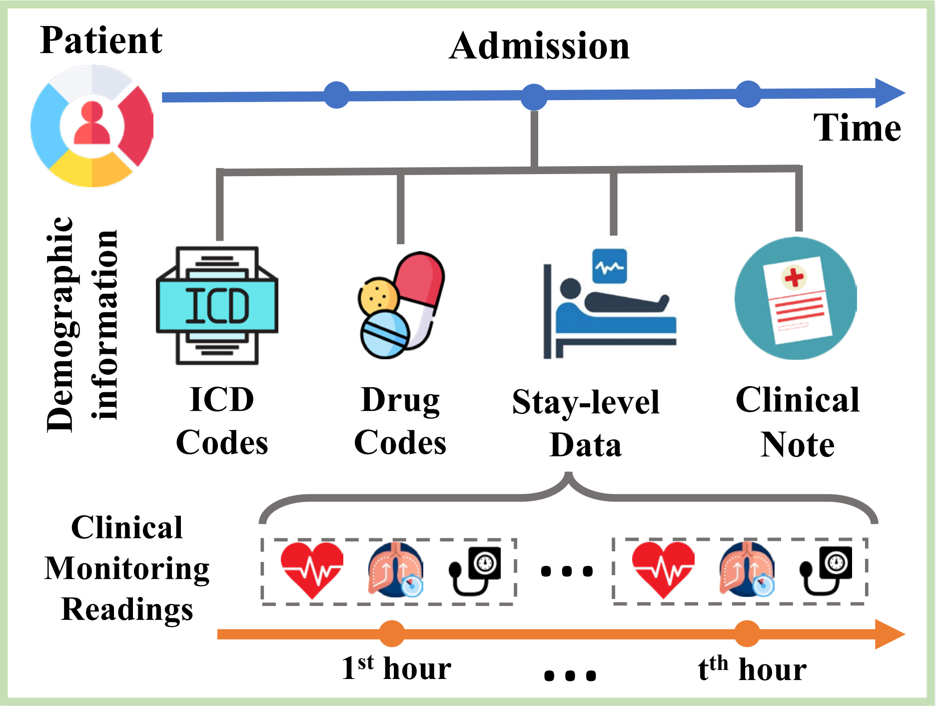
(a) FedAvg-based Knowledge Injection Performance.



(b) FedProx-based Knowledge Injection Performance.

Summary

Hierarchical



Wang et al., *Hierarchical Pretraining on Multimodal Electronic Health Records*, EMNLP'23

Multi-sourced

Wang et al., *Unity in Diversity: Collaborative Pre-training Across Multimodal Medical Sources*, ACL'24

Wang et al., *FedMeKI: A Benchmark for Scaling Medical Foundation Models via Federated Knowledge Injection*, under review

Wang et al., *FedKIM: Adaptive Federated Knowledge Injection into Medical Foundation Models*, under review

Recent Advances in Predictive Modeling with Electronic Health Records

Jiaqi Wang, Junyu Luo, Muchao Ye, Xiaochen Wang, Yuan Zhong, Aofei Chang, Guanjie Huang, Ziyi Yin, Cao Xiao, Jimeng Sun, Fenglong Ma

August 9th at 11:30 | Session: MTA: Health and medicine

Thank you!

Any questions?