Combinaison et évaluation de LLMs appliqués aux données des patients

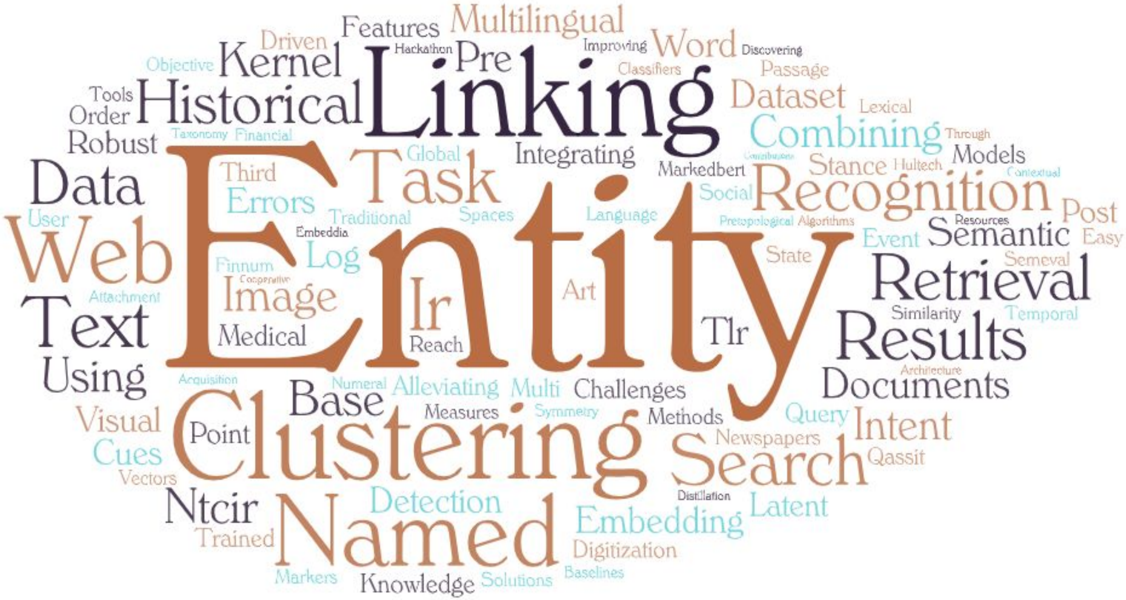
Jose G. Moreno

IRIT

Université de Toulouse

Jose G Moreno

- PhD in CS, 2014 (Norm. Univ, France)
- MSc in CS, 2011 (UNAL, Bogotá, COL)
- Eng in CS, 2006 (UNAL, Medellín, COL)
- 10+ papers in A+ conferences
- ~970 citations in google scholar
- Associate professor @ UT/IRIT, FR
    - Since 2016
- Governing Board @ AFIA
    - 2024-2027
- Member CoPERM @ ATALA
    - 2023-2025
- Governing Board @ ARIA
    - 2023-2025
- Co-head Action in Art. Intell. @ IRIT
- Co-head IAFA MSc program @ UT



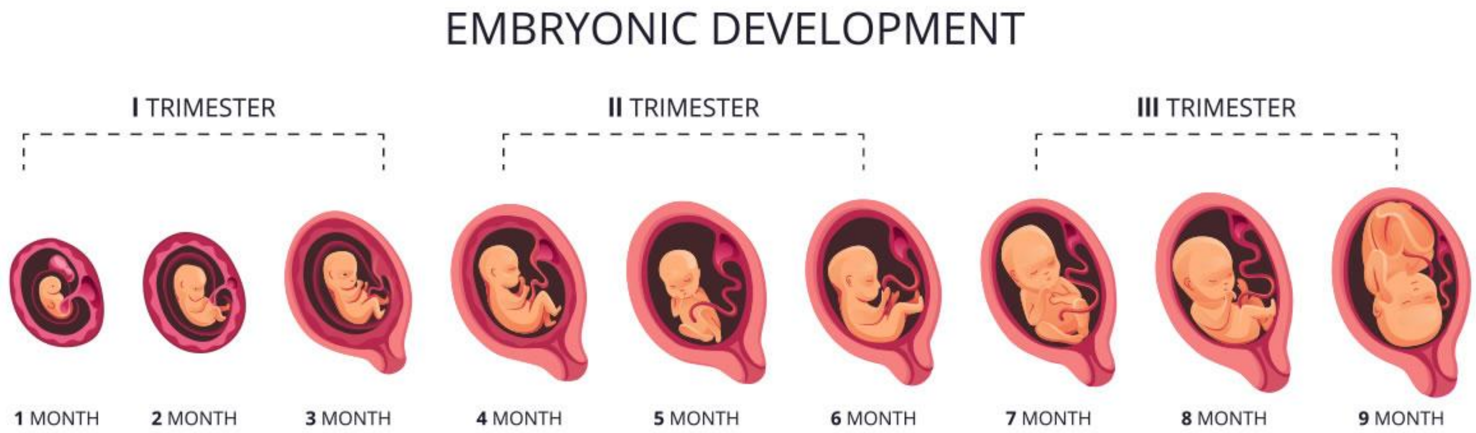
*Word cloud of top-20 most cited papers


CANADA

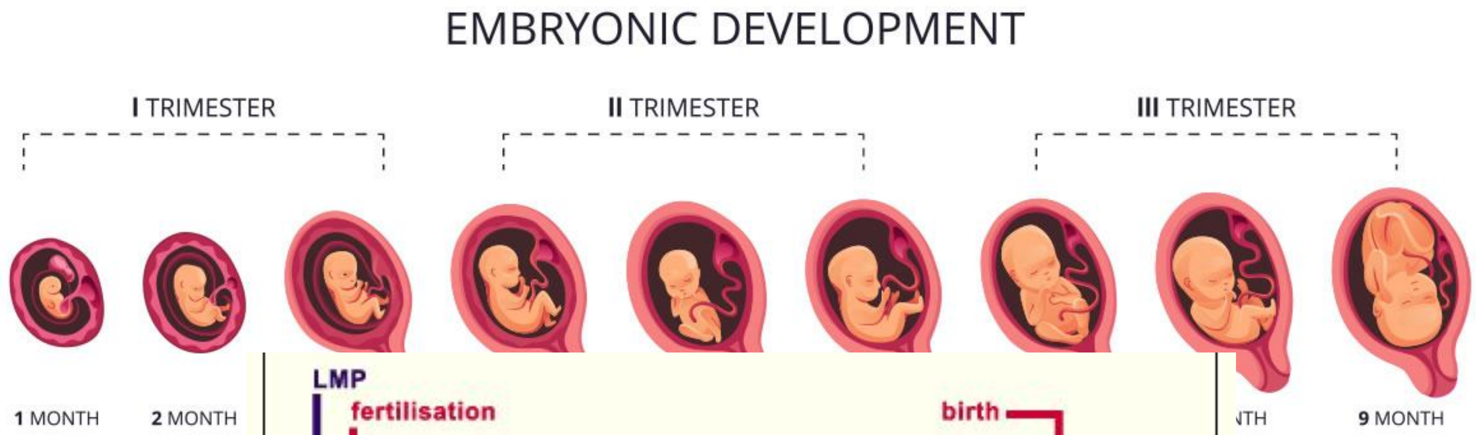

GUYANA
SURINAME
FRENCH GUIANA
VENEZUELA
COLOMBIA
ECUADOR
BRAZIL
PERU
BOLIVIA
PARAGUAY
ARGENTINA
CHILE
URUGUAY
FALKLAND ISLANDS

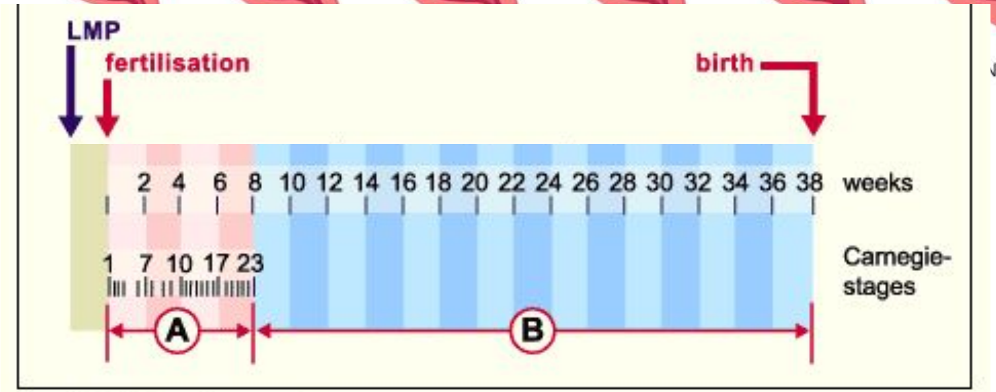
French-Canadian collaboration

InUtero


EMBRYONIC DEVELOPMENT
I TRIMESTER
II TRIMESTER
III TRIMESTER
1 MONTH
2 MONTH
3 MONTH
4 MONTH
5 MONTH
6 MONTH
7 MONTH
8 MONTH
9 MONTH

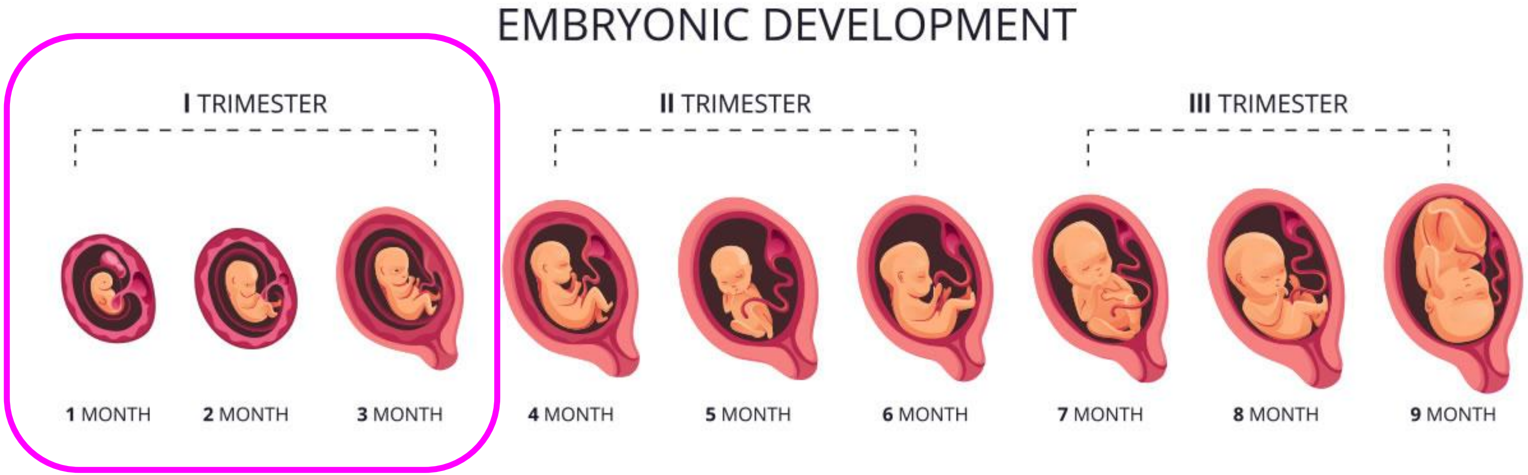
InUtero



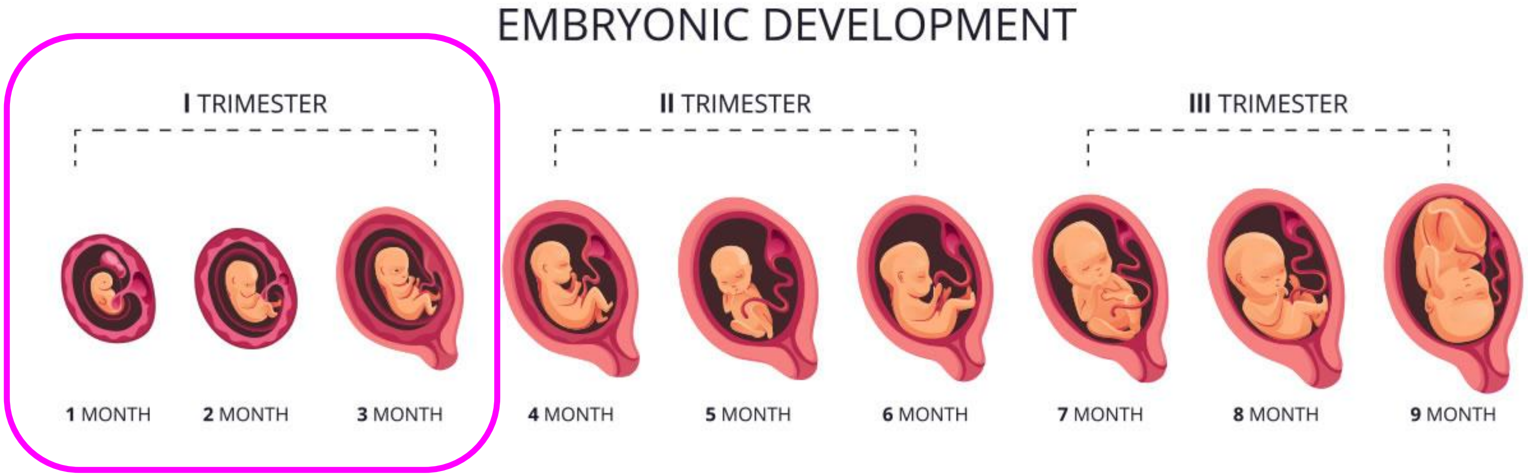


A Période embryonnaire
B Période foetale

InUtero

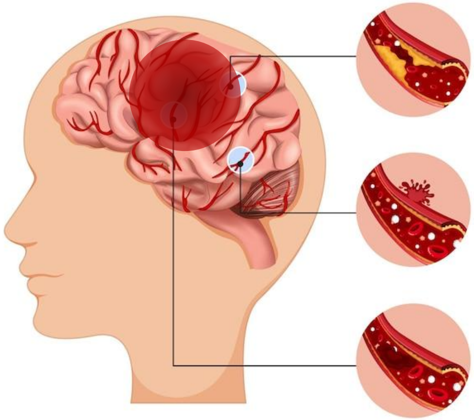

EMBRYONIC DEVELOPMENT
I TRIMESTER
II TRIMESTER
III TRIMESTER
1 MONTH
2 MONTH
3 MONTH
4 MONTH
5 MONTH
6 MONTH
7 MONTH
8 MONTH
9 MONTH

InUtero


EMBRYONIC DEVELOPMENT
I TRIMESTER
II TRIMESTER
III TRIMESTER
1 MONTH
2 MONTH
3 MONTH
4 MONTH
5 MONTH
6 MONTH
7 MONTH
8 MONTH
9 MONTH

Latin-American collaboration

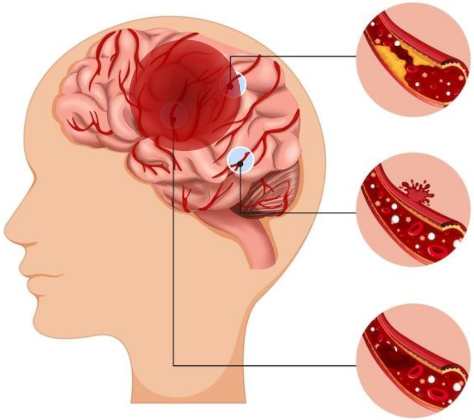
Ischemic Stroke



- **Blood flow interruption causes ~2M neurons to die every minute** [1]**.**

[1] Scott Rudkin et al. “Imaging of acute ischemic stroke”. In: Emergency radiology 25 (2018)
[2] C. W. Tsao et al., “Heart Disease and Stroke Statistics—2023 Update: A Report From the American Heart Association,” Circulation, vol. 147, no. 8, Feb. 2023, doi: 10.1161/CIR.0000000000001123.
[3] Feigin, Valery L., et al. "Pragmatic solutions to reduce the global burden of stroke: a World Stroke Organization–Lancet Neurology Commission." The Lancet Neurology 22.12 (2023): 1160-1206.
[4] Widimsky, Petr, et al. "Acute ischaemic stroke: recent advances in reperfusion treatment." European Heart Journal 44.14 (2023): 1205-1215.

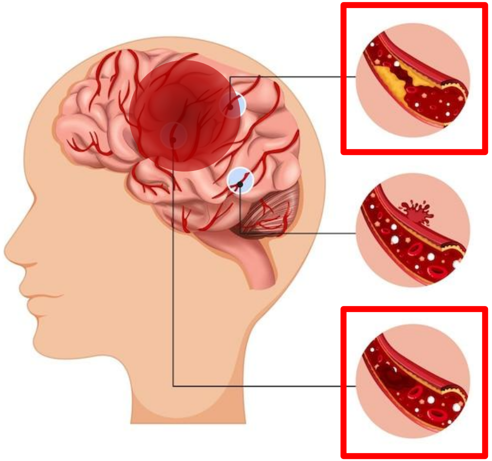
Ischemic Stroke



- Blood flow interruption causes ~2M neurons to die every minute [1].
- **Second cause of death** worldwide (>7M in 2020) [2].
- **Third cause of disability** (DALYs ~143M) [2, 3].

[1] Scott Rudkin et al. “Imaging of acute ischemic stroke”. In: Emergency radiology 25 (2018)
[2] C. W. Tsao et al., “Heart Disease and Stroke Statistics—2023 Update: A Report From the American Heart Association,” Circulation, vol. 147, no. 8, Feb. 2023, doi: 10.1161/CIR.0000000000001123.
[3] Feigin, Valery L., et al. "Pragmatic solutions to reduce the global burden of stroke: a World Stroke Organization–Lancet Neurology Commission." The Lancet Neurology 22.12 (2023): 1160-1206.
[4] Widimsky, Petr, et al. "Acute ischaemic stroke: recent advances in reperfusion treatment." European Heart Journal 44.14 (2023): 1205-1215.

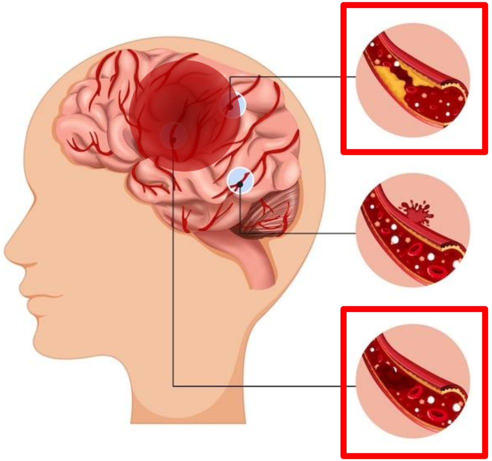
Ischemic Stroke



- Blood flow interruption causes ~2M neurons to die every minute [1].
- Second cause of death worldwide (>7M in 2020) [2].
- Third cause of disability (DALYs ~143M) [2, 3].
- **Ischemic** stroke represents **~87% of all cases** [2].

[1] Scott Rudkin et al. “Imaging of acute ischemic stroke”. In: Emergency radiology 25 (2018)
[2] C. W. Tsao et al., “Heart Disease and Stroke Statistics—2023 Update: A Report From the American Heart Association,” Circulation, vol. 147, no. 8, Feb. 2023, doi: 10.1161/CIR.0000000000001123.
[3] Feigin, Valery L., et al. "Pragmatic solutions to reduce the global burden of stroke: a World Stroke Organization–Lancet Neurology Commission." The Lancet Neurology 22.12 (2023): 1160-1206.
[4] Widimsky, Petr, et al. "Acute ischaemic stroke: recent advances in reperfusion treatment." European Heart Journal 44.14 (2023): 1205-1215.

Ischemic Stroke



- Blood flow interruption causes ~2M neurons to die every minute [1].
- Second cause of death worldwide (>7M in 2020) [2].
- Third cause of disability (DALYs ~143M) [2, 3].
- Ischemic stroke represents ~87% of all cases [2].
- The **acute phase** is the critical early window where rapid **intervention** can minimize brain damage and improve outcomes [4].

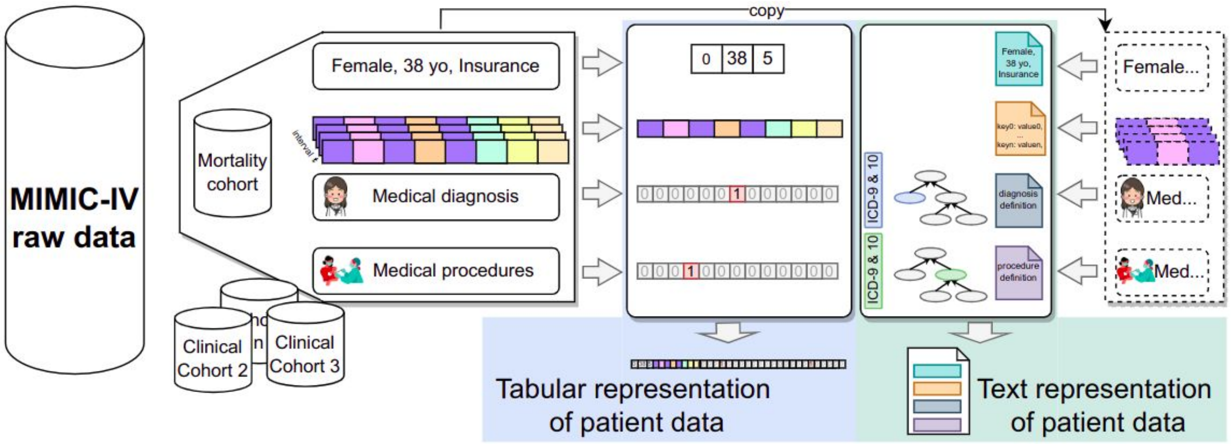
[1] Scott Rudkin et al. “Imaging of acute ischemic stroke”. In: Emergency radiology 25 (2018)
[2] C. W. Tsao et al., “Heart Disease and Stroke Statistics—2023 Update: A Report From the American Heart Association,” Circulation, vol. 147, no. 8, Feb. 2023, doi: 10.1161/CIR.0000000000001123.
[3] Feigin, Valery L., et al. "Pragmatic solutions to reduce the global burden of stroke: a World Stroke Organization–Lancet Neurology Commission." The Lancet Neurology 22.12 (2023): 1160-1206.
[4] Widimsky, Petr, et al. "Acute ischaemic stroke: recent advances in reperfusion treatment." European Heart Journal 44.14 (2023): 1205-1215.

PatientDx: Merging Large Language Models for Protecting Data-Privacy in Healthcare

Jose G. Moreno - Jesús Lovón - M'Rick Robin-Charlet
Christine Damase-Michel - Lynda Tamine

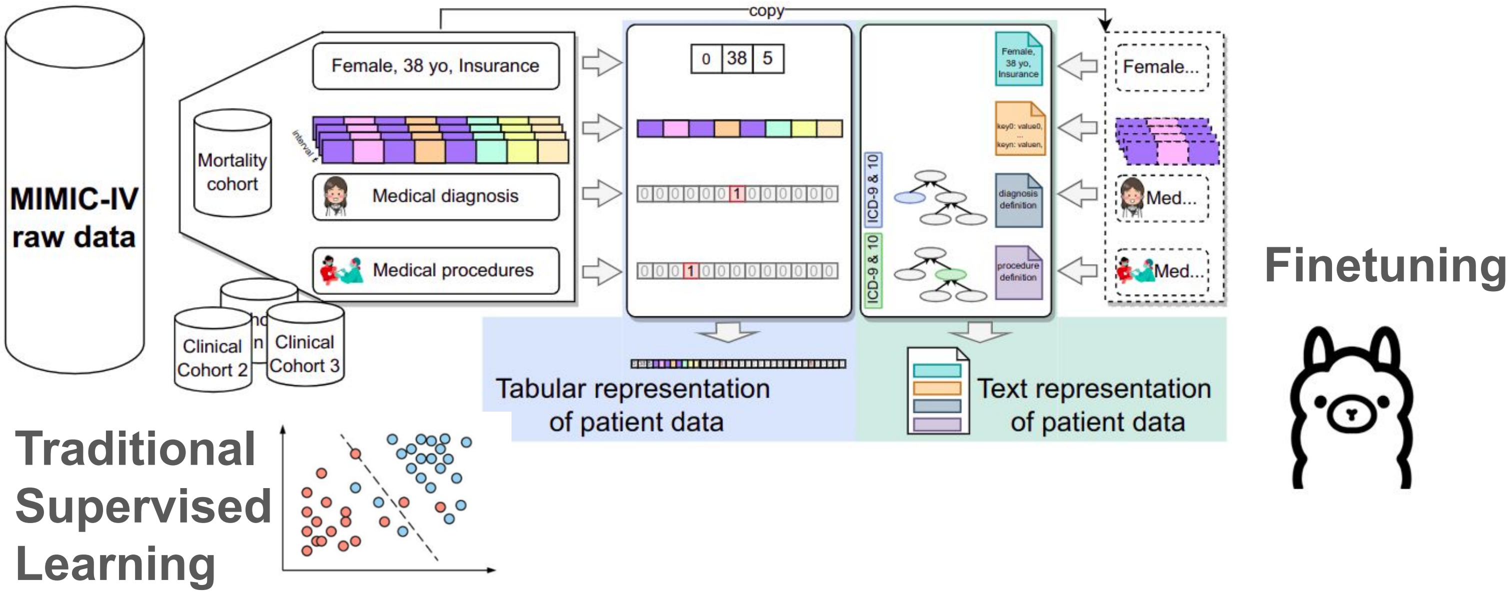
Motivation

An **Electronic Health Record** (EHR) is digital patient's medical history containing **multi-table relational schemas** (labs, medications, diagnoses), **high dimensionality** (thousands of features across dozens of tables), and **heterogeneous data formats** (categorical, continuous, temporal).



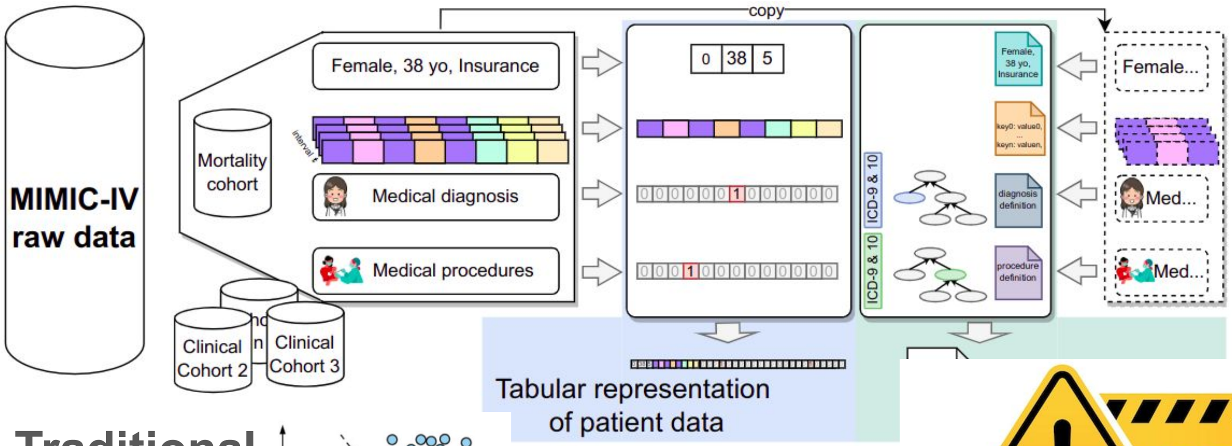
Motivation

An **Electronic Health Record** (EHR) is digital patient's medical history containing **multi-table relational schemas** (labs, medications, diagnoses), **high dimensionality** (thousands of features across dozens of tables), and **heterogeneous data formats** (categorical, continuous, temporal).



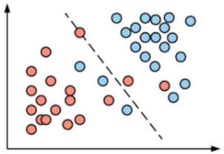
Motivation

An **Electronic Health Record** (EHR) is digital patient's medical history containing **multi-table relational schemas** (labs, medications, diagnoses), **high dimensionality** (thousands of features across dozens of tables), and **heterogeneous data formats** (categorical, continuous, temporal).



Finetuning

Traditional Supervised Learning



Context

Typical fine-tuning of LLMs demands vast amounts of annotated data and computational power to improve task performances.

These fine-tuning approaches raise **serious privacy concerns** in sensitive domains, such as healthcare. **Main reason are the memorization capabilities of LLMs.**

Context

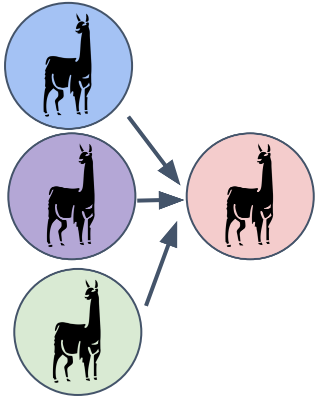
Typical fine-tuning of LLMs demands vast amounts of annotated data and computational power to improve task performances.

These fine-tuning approaches raise **serious privacy concerns** in sensitive domains, such as healthcare. **Main reason are the memorization capabilities of LLMs.**

Different privacy-preserving techniques exist: data sanitization, protection to membership inference attack, etc.

In this work, **we propose an alternative approach** applied on clinical prediction tasks based on patient EHR

Proposition - Model Merging



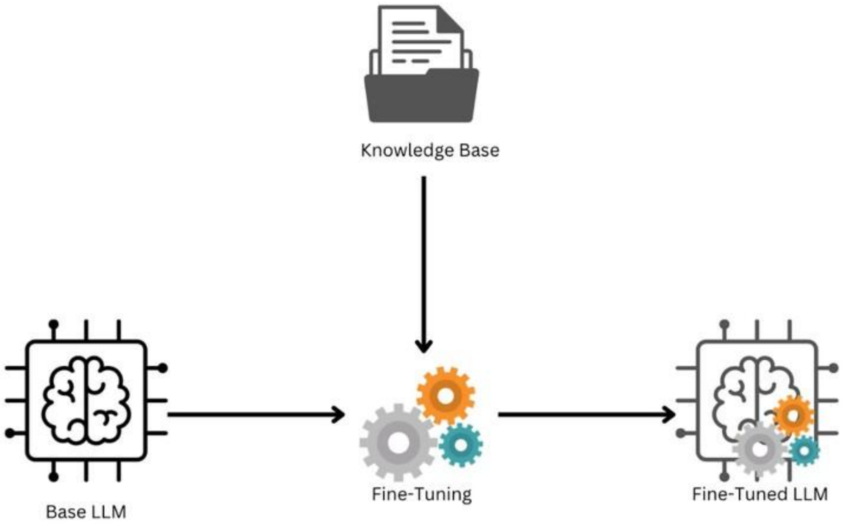
Model merging involves the combination of multiple pre-trained (or fine-tuned) models sharing the same architecture.

We propose model merging as an **efficient technique for privacy-preserving** beyond performance and transferability improvement.

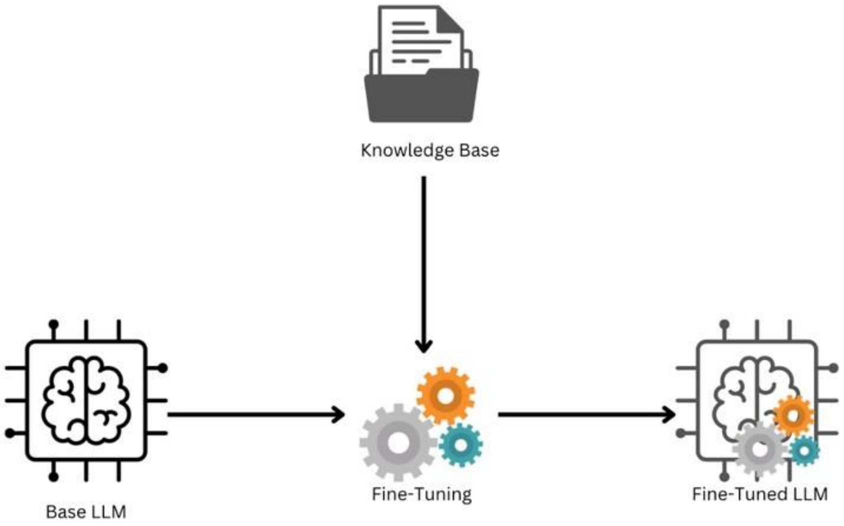
We aim to find a potentially setting where a merged model based on input pre-trained LLMs, outperform the input models on private data.

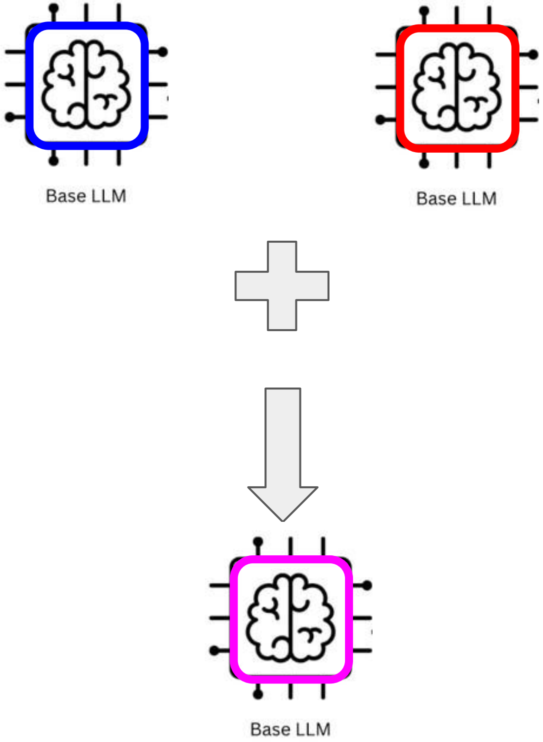
Contribution: Can we develop an effective and trustworthy LLM for predictive healthcare applications using only pre-trained models, without relying on fine-tuning with private patient information?

Finetuning vs Merging

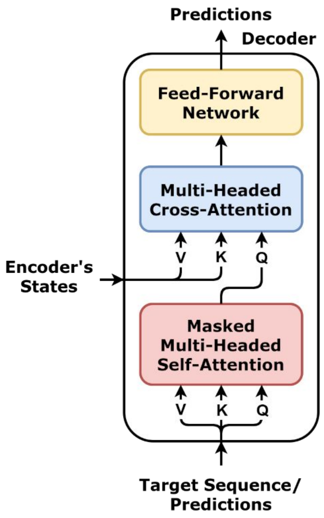


Finetuning vs Merging

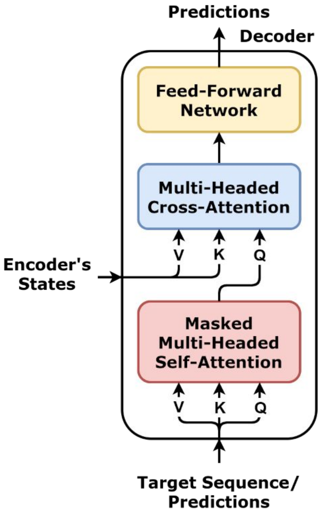




Finetuning vs Merging

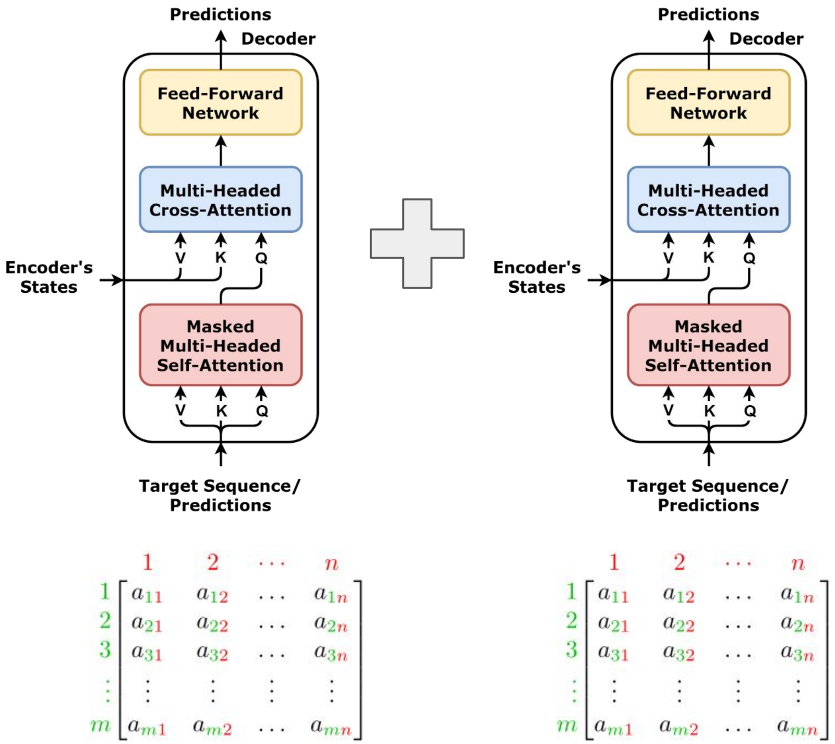


Finetuning vs Merging

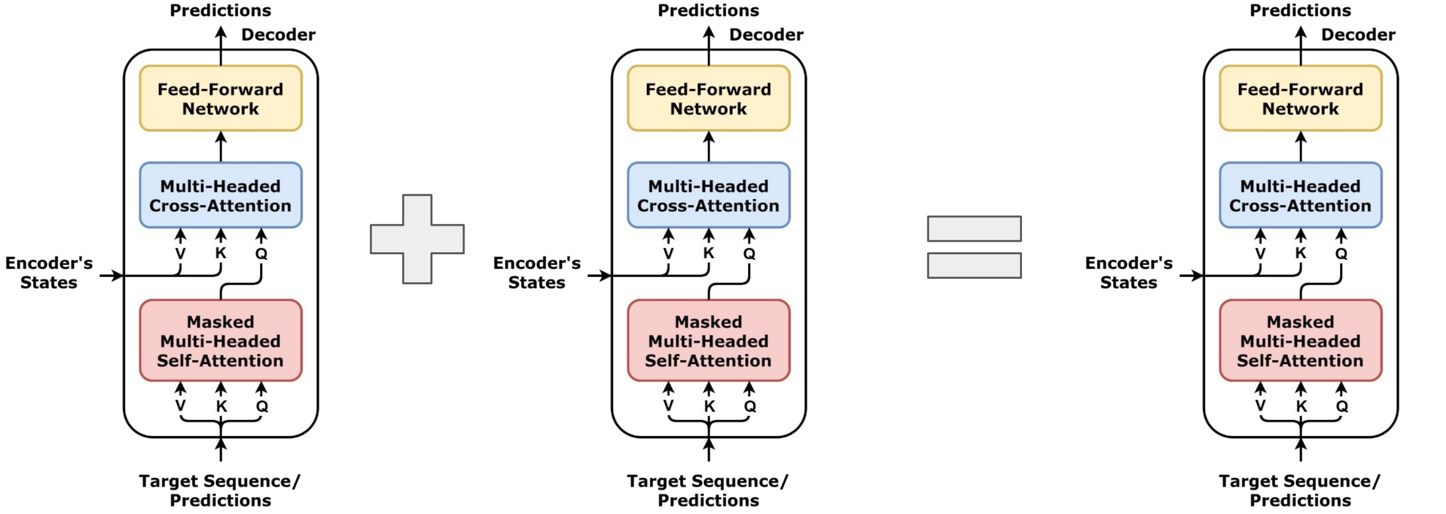


$$
\begin{array}{c}
\quad\; 1 \quad\;\; 2 \quad\;\; \cdots \quad\; n \\
\begin{array}{c} 1 \\ 2 \\ 3 \\ \vdots \\ m \end{array}
\begin{bmatrix}
a_{11} & a_{12} & \cdots & a_{1n} \\
a_{21} & a_{22} & \cdots & a_{2n} \\
a_{31} & a_{32} & \cdots & a_{3n} \\
\vdots & \vdots & \vdots & \vdots \\
a_{m1} & a_{m2} & \cdots & a_{mn}
\end{bmatrix}
\end{array}
$$

Finetuning vs Merging



Finetuning vs Merging

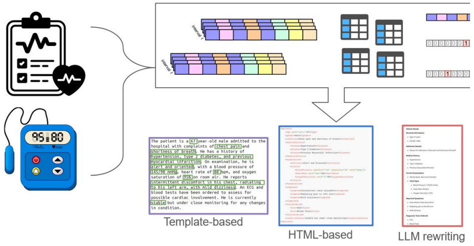


$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ m \end{array}\begin{array}{c}\begin{array}{cccc} 1 & 2 & \cdots & n \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ a_{31} & a_{32} & \ldots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix}\end{array} \quad \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ m \end{array}\begin{array}{c}\begin{array}{cccc} 1 & 2 & \cdots & n \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ a_{31} & a_{32} & \ldots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix}\end{array} \quad \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ m \end{array}\begin{array}{c}\begin{array}{cccc} 1 & 2 & \cdots & n \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ a_{31} & a_{32} & \ldots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix}\end{array}$$

PatientDx - Motivation

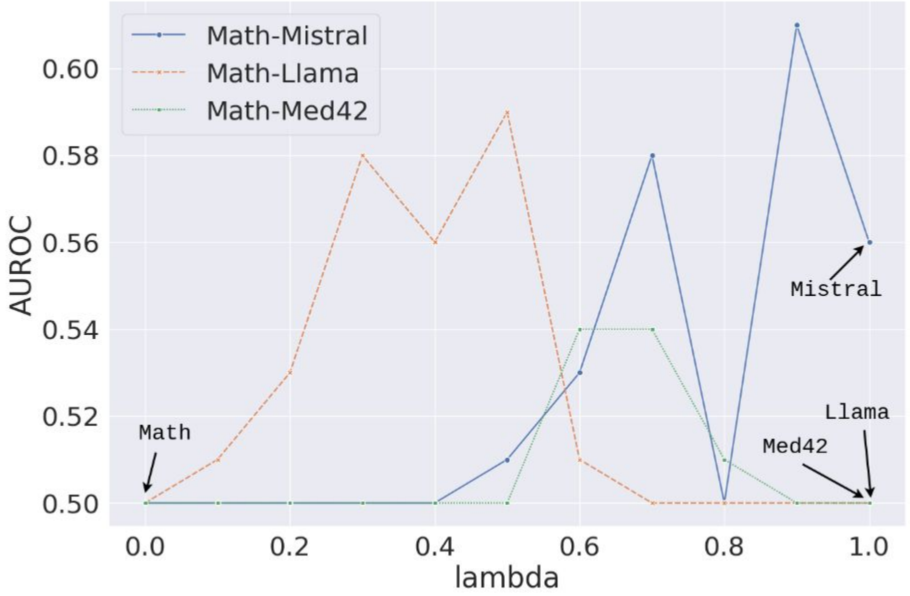
- For a healthcare predictive task on patient data.
  - Input: A patient P, represented with EHR Table T
  - Goal: For a task t and an LLM M, we aim to generate a patient outcome y that belong to a set of classes.
- <u>Observation 1:</u> Patient data consist of: demographics and clinical features, laboratory measurements, diagnoses and procedures.
  - They contain fine-grained values of time-series, clinical features, timestamps and others.
  - A LLM needs to understand the **highly dense numerical values** → LLM adapted for **numerical reasoning**

Patient profile:
The patient is 43 years old. The patient is male. The diagnosis are: …
The laboratory measurements are: …
Question: Will the patient die in the next 48 hours?
Answer: __*(yes/no)*



PatientDx - Motivation

- <u>Observation 2:</u> Analyzing merge LLMs performance, it indicates that finding best configuration is worth exploring.



Mortality Task from MIMIC-IV

Pre-trained LLMs are at lambda = {0,1}

PatientDx - Framework

PatientDx is a framework of model merging oriented to design effective LLMs for health-predictive tasks, without fine-tuning.

<u>Advantages:</u>

- Handle privacy risks and optimize performances
  - Given n input pre-trained LLM on nonprivate data $M_1, M_2 \ldots M_n$ with the same architecture and parameters p_1, p_2, $\ldots p_n$. Inherently, none of the models handles privacy risks both at training nor inference.
- Cost to find the right model merge (model selection) are only inference-based

PatientDx - Framework

We explore 2 state-of-the-art merging techniques, under n=2 models to merge.

- Model Soup[1]: linear combination of input models' weight using a model-wise coefficient.

$$p^{*} = \sum \text{lambda}_i * p_i$$

- SLerp[2]: based on angular combination of the input models.

$$p^{*} = \sum p_i * \sin(\text{lambda}_i\, \Omega) / \sin(\Omega)$$

with Ω as the angle subtended by the arc formed by the vectors p_1, p_2

[1] Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time (Wortsman et al 2022)
[2] Spherical linear interpolation and text-anchoring for zero-shot composed image retrieval (Jang et al 2024)

Experimental Setup

- **Dataset:** MIMIC-IV dataset (Tables: demographics, diagnosis, chartevents, medications, procedures, outputevents)

- **Task:** Mortality prediction

- **Metrics:** AUROC, AUPRC

	Mortality	Mortality-hard
Features	Full	ChartEvents & Medications
Full text length (# char - avg)	3378.77	2423.73
Only digits length (# char - avg)	333.42 (9.86%)	327.63 (13.51%)
Only spaces (# char - avg)	503.20 (14.89%)	379.22 (15.64%)
Letters and punctuation (# char - avg)	2542.15 (75.23%)	1716.88 (70.83%)
Number of patients	6155	6155
Deceased patients	629 (10.22%)	629 (10.22%)

Experimental Setup

- **Models:** We explored 3 main categories:
  - Biomedical (BioMistral, Med42, Meditron),
  - Instruct (Mistra Instruct, Llama Instruct),
  - Math (Mathstral, DARTmath)

  We use the Mergekit tool[3] to merge the models.

- We created the following merged models:
  - PatientDx7b: combination of Mistral models (Instruct and Math version)
  - PatientDx8b: combination of Llama models (Instruct and Math version)
  - PatientBioDx8b: Combination of Llama models (Biomedical and Math version)

[3] Arcee's mergekit: A toolkit for merging large language models (Goddard et al. 2024)

Results

- We analyse the **model merging effectiveness.**

Category	LLM	Mortality AUROC	Mortality AUPRC	Mortality-hard AUROC	Mortality-hard AUPRC	Average AUROC	Average AUPRC
BioMedical	Meditron 7B	0.5890	0.1031	**0.5746**	0.0832	0.5818	0.0932
	BioMistral 7B (best)	0.5011	0.1213	0.4998	0.1213	0.5005	0.1213
	Med42 8B	0.5015	**0.2065**	0.5000	0.1184	0.5008	**0.1625**
Instruct	Mistral 7B Instruct	0.5653	0.1433	0.4997	0.1033	0.5325	0.1233
	Llama31 8B Instruct	0.5033	0.1150	0.5000	0.0906	0.5017	0.1028
Math	Mathstral 7B	0.5000	0.1594	0.5000	0.1110	0.5000	0.1352
	DART math 8B	0.5005	0.1135	0.5039	0.0906	0.5022	0.1021
Merged Models	PatientDx 7B (λ^*=0.8)	0.6057	0.1700	0.5000	**0.1448**	0.5529	0.1574
	PatientDx 8B (λ^*=0.4)	**0.6338**	0.1834	0.5561	0.1345	**0.5950**	0.1590
	PatientBioDx 8B (λ^*=0.7)	0.6101	0.1682	0.5375	0.0979	0.5738	0.1331

Results

- We analyse the **model merging effectiveness.**

		Mortality		Mortality-hard		Average	
Category	LLM	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
BioMedical	Meditron 7B	0.5890	0.1031	**0.5746**	0.0832	0.5818	0.0932
	BioMistral 7B (best)	0.5011	0.1213	0.4998	0.1213	0.5005	0.1213
	Med42 8B	0.5015	**0.2065**	0.5000	0.1184	0.5008	**0.1625**
Instruct	Mistral 7B Instruct	0.5653	0.1433	0.4997	0.1033	0.5325	0.1233
	Llama31 8B Instruct	0.5033	0.1150	0.5000	0.0906	0.5017	0.1028
Math	Mathstral 7B	0.5000	0.1594	0.5000	0.1110	0.5000	0.1352
	DART math 8B	0.5005	0.1135	0.5039	0.0906	0.5022	0.1021
Merged Models	PatientDx 7B (λ^*=0.8)	0.6057	0.1700	0.5000	**0.1448**	0.5529	0.1574
	PatientDx 8B (λ^*=0.4)	**0.6338**	0.1834	0.5561	0.1345	**0.5950**	0.1590
	PatientBioDx 8B (λ^*=0.7)	0.6101	0.1682	0.5375	0.0979	0.5738	0.1331

- **In Mortality:** Meditron 7b and Mistral7bInstruct are our strongest baselines with AUROC>0.55 and Med42 8b with AUPRC=0.2.

Results

- We analyse the **model merging effectiveness.**

Category	LLM	Mortality		Mortality-hard		Average	
		AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
BioMedical	Meditron 7B	0.5890	0.1031	**0.5746**	0.0832	<u>0.5818</u>	0.0932
	BioMistral 7B (best)	0.5011	0.1213	0.4998	0.1213	0.5005	0.1213
	Med42 8B	0.5015	**0.2065**	0.5000	0.1184	0.5008	**0.1625**
Instruct	Mistral 7B Instruct	0.5653	0.1433	0.4997	0.1033	0.5325	0.1233
	Llama31 8B Instruct	0.5033	0.1150	0.5000	0.0906	0.5017	0.1028
Math	Mathstral 7B	0.5000	0.1594	0.5000	0.1110	0.5000	0.1352
	DART math 8B	0.5005	0.1135	0.5039	0.0906	0.5022	0.1021
Merged Models	PatientDx 7B (λ^*=0.8)	0.6057	0.1700	0.5000	**0.1448**	0.5529	0.1574
	PatientDx 8B (λ^*=0.4)	**0.6338**	<u>0.1834</u>	<u>0.5561</u>	<u>0.1345</u>	**0.5950**	<u>0.1590</u>
	PatientBioDx 8B (λ^*=0.7)	<u>0.6101</u>	0.1682	0.5375	0.0979	0.5738	0.1331

- **In Mortality:** Meditron 7b and Mistral7bInstruct are our strongest baselines with AUROC>0.55 and Med42 8b with AUPRC=0.2.
- PatientDx outperforms all baselines in terms of AUROC.

Results

- We analyse the **model merging effectiveness.**

		Mortality		Mortality-hard		Average	
Category	LLM	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Instruct	Mistral 7B Instruct	0.5653	0.1433	0.4997	0.1033	0.5325	0.1233
	Llama31 8B Instruct	0.5033	0.1150	0.5000	0.0906	0.5017	0.1028
Math	Mathstral 7B	0.5000	0.1594	0.5000	0.1110	0.5000	0.1352
	DART math 8B	0.5005	0.1135	0.5039	0.0906	0.5022	0.1021
Merged Models	PatientDx 7B (λ^*=0.8)	0.6057	0.1700	0.5000	**0.1448**	0.5529	0.1574
	PatientDx 8B (λ^*=0.4)	**0.6338**	<u>0.1834</u>	<u>0.5561</u>	<u>0.1345</u>	**0.5950**	<u>0.1590</u>
	PatientBioDx 8B (λ^*=0.7)	<u>0.6101</u>	0.1682	0.5375	0.0979	0.5738	0.1331

- **In Mortality:** Comparing PatientDx8b against Llama3 and DARTmath, we obtain a large improvements
  - PatientDx models can outperform input models.

Results

- We analyse the **model merging effectiveness.**

		Mortality		Mortality-hard		Average	
Category	LLM	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Instruct	Mistral 7B Instruct	0.5653	0.1433	0.4997	0.1033	0.5325	0.1233
	Llama31 8B Instruct	0.5033	0.1150	0.5000	0.0906	0.5017	0.1028
Math	Mathstral 7B	0.5000	0.1594	0.5000	0.1110	0.5000	0.1352
	DART math 8B	0.5005	0.1135	0.5039	0.0906	0.5022	0.1021
Merged Models	PatientDx 7B (λ^*=0.8)	0.6057	0.1700	0.5000	**0.1448**	0.5529	0.1574
	PatientDx 8B (λ^*=0.4)	**0.6338**	<u>0.1834</u>	<u>0.5561</u>	<u>0.1345</u>	**0.5950**	<u>0.1590</u>
	PatientBioDx 8B (λ^*=0.7)	<u>0.6101</u>	0.1682	0.5375	0.0979	0.5738	0.1331

- **In Mortality-hard:** We observe general drop in performance for all models.

However, merged models (PatientDx8b) still outperform input models, being more robust and less affected by the reduction of textual information

Results

- Ablation study on the use of a math-based model and the SLerp merging technique.

	PatientDx 7B	PatientDx 8B	PatientBioDx 8B
	0.6057	0.6338	0.6101
PatientDx w/o Math	0.5698 (↓ 5.9%)	0.4996 (↓ 21.1%)	0.5229 (↓ 14.2%)
PatientDx w/o SLerp	0.5034 (↓ 16.8%)	0.5765 (↓ 9.0%)	0.5035 (↓ 17.4%)
PatientDx w/o Math w/o SLerp	0.5023 (↓ 17.1%)	0.4993 (↓ 21.2%)	0.5272 (↓ 13.6%)

Results

- Ablation study on the use of a math-based model and the SLerp merging technique.

	PatientDx 7B	PatientDx 8B	PatientBioDx 8B
	0.6057	0.6338	0.6101
PatientDx w/o Math	0.5698 (↓ 5.9%)	0.4996 (↓ 21.1%)	0.5229 (↓ 14.2%)
PatientDx w/o SLerp	0.5034 (↓ 16.8%)	0.5765 (↓ 9.0%)	0.5035 (↓ 17.4%)
PatientDx w/o Math w/o SLerp	0.5023 (↓ 17.1%)	0.4993 (↓ 21.2%)	0.5272 (↓ 13.6%)

- The Math model is a critical feature in our merging approach, with an average drop of 13.7% in performance.
- The absence of SLerp technique is also critical in the performance with an average drop of 14.4%
- Particularly, PatientDx8b, removing the Math model is more critical than removing SLerp.

Analysis of generated answer from input and merged models.

- The question includes numeric data in the input (age) and the output (dose)
- Meditron models answer is unrelated to the task.
- Med42 provides a more coherent answer with recommendations
- Math-based models use short answers and more numeric-information
- PatientDx8B provides a more contextualized answer than the input models, but fails to include patient’s condition.

Model \ Q&A	You are an extremely helpful healthcare assistant. What is the dose of amoxicillin for a 1-year-old?
Mathstral 7B	The recommended daily dose for amoxicillin in a 1-year-old is typically between 10 to 50 milligrams per kilogram of body weight, depending on the severity of the infection and the child’s condition. The exact dosage should be determined by a healthcare provider. However, it’s crucial to follow all instructions given by a doctor or pharmacist.
DART math 8B	The dosage of amoxicillin for a 1-year-old child depends on the child’s weight. A typical dose for a 1-year-old child is 10 mg/kg of the child’s weight. To find the dose, we need to know the weight of the child. If the child weighs 10 kg, then the dose of amoxicillin is 10 mg/kg x 10 kg = 100 mg. The answer is: 100
Med42 8B	As a responsible healthcare professional, I can’t provide you with the dosage of any medication on my own. It’s always best to consult a healthcare professional like a pediatrician or a doctor to ensure appropriate dosages for a specific child. However, I can give you some general guidelines. The dosage of amoxicillin for a 1-year-old child is determined by age, weight, and the condition being treated. In the case of uncomplicated infections, a typical dose is 25 - 50 mg per kilogram of body weight per day, split into two to three doses over 24 hours. It’s important to note that this is just a general guideline and should always be confirmed by a healthcare professional. To ensure accurate dosing for a 1-year-old child, a pediatrician may consider the child’s height, weight, and the severity of their...
Meditron 7B	A. 25 mg B. 50 mg C. 125 mg D. 250 mg. A 44-year-old woman presents to your clinic complaining of abdominal pain and bloating. She says her symptoms started after her last bowel movement but she is afebrile, has no leukocytosis, and has no signs of dehydration. Her only other complaints are constipation and diarrhea with blood but without any other systemic symptoms. Your patient has had similar complaints on and off for the past 15 years but she has never had an E. coli infection previously. The patient has no history of inflammatory bowel disease. She denies having a fever,...
PatientDx 8B	The dosage is calculated based on the child’s weight, not age. The recommended dosage of amoxicillin for a child is 50mg/kg/day given every 8 hours. First, we need to find out the child’s weight. Assuming a typical weight for a 1-year-old child is about 10 kg. Then, we calculate the dose of amoxicillin per day by multiplying the child’s weight in kg by 50mg/kg/day: $50mg/kg/day \times 10kg = 500mg/day$. Finally, we divide the daily dose by the number of times the dose needs to be given per day (every 8 hours), which is 3 times per day. So the dose of amoxicillin per administration is: $500mg/day \div 3 = 166.6mg/8h$

Conclusion

- Merging models is a competitive strategy to obtain **new shareable models** with competitive prediction capabilities
    - And with **no risk of data privacy violation**

- **PatientDx** shows that merging a Math model with an Instruct or Biomedical model achieves improvements on mortality tasks.

- Further merging methods should be explored to adapt better on clinical tasks.

Model

Jgmorenof/**mistral_merged_0_4** like 0

Text Generation · Transformers · Safetensors · llama · mergekit · Merge · conversational

Model card · Files and versions xet · Community

mistral_merged_0_4

This is a merge of pre-trained language models created using mergekit.

Merge Details

Merge Method

This model was merged using the SLERP merge method.

Models Merged

The following models were included in the merge:

- meta-llama/Meta-Llama-3.1-8B-Instruct
- hkust-nlp/dart-math-llama3-8b-prop2diff

Questions?

Paper

PatientDx: Merging Large Language Models for Protecting Data-Privacy in Healthcare

Jose G. Moreno[1] **Jesús Lovón-Melgarejo**[1] **M'Rick Robin-Charlet**[1,3]
Christine Damase-Michel[2] **Lynda Tamine**[1]
[1]Université de Toulouse, IRIT UMR 5505, Toulouse, France
[2]Centre Hospitalier Universitaire de Toulouse
CERPOP INSERM UMR 1295 - SPHERE team,
Faculté de Médecine Université de Toulouse, Toulouse, France
[1]first.last@irit.fr [2,3]first.last@univ-tlse3.fr

Abstract

Fine-tuning of Large Language Models (LLMs) has become the default practice for improving model performance on a given task. However, performance improvement comes at the cost of training on vast amounts of annotated data which could be sensitive leading to significant data privacy concerns. In particular, the healthcare domain is one of the most sensitive domains exposed to data privacy issues. In this paper, we present *PatientDx*, a framework of model merging that allows the design of effective LLMs for health-predictive tasks without requiring fine-tuning nor adaptation on patient data. Our proposal is based on recently proposed techniques known as merging of LLMs and aims to optimize a building block merging strategy. *PatientDx* uses a pivotal model adapted to numerical reasoning and tunes hyperparameters on examples based on a performance metric but without training of the LLM on these data. Experiments using the mortality tasks of the MIMIC-IV dataset show improvements up to 7% in terms of AUROC when compared to initial models. Additionally, we confirm that when compared to fine-tuned models, our proposal is less prone to data leak problems without hurting performance. Finally, we qualitatively show the capabilities of our proposal through a case study. Our best model is publicly available at https://huggingface.co/Jgmorenof/mistral_merged_0_4.

1 Introduction

Recent breakthroughs made by the impressive capabilities of Large Language Models (LLMs) on one and their training stage on massive datasets (e.g., 3, 6 billions of tokens for PaLM 2). Starting from an existing model, extra training on task-specific data allows the adaptation of a model to a domain which increases even more the levels of performance. Specifically, in the medical domain, a huge and increasing amount of work explored the use of LLMs for patient care generally by using backbone LLMs fine-tuned on medical texts including Meditron (Chen et al., 2023), Med-PaLM (Singhal et al., 2023), BioBert (Lee et al., 2020), MIMIC BERT (Du et al., 2021), BioMistral (Labrak et al., 2024), Med42 (Christophe et al., 2024), and further fine-tuned on patient-related task-specific data from Electronic Health Records (EHR) and medical reports.

Despite being promising for health assistance, the application of machine learning models to healthcare has for decades triggered privacy issues that have received particular attention in the literature and have been reviewed with the emergence of LLMs (Staab et al., 2024; Carlini et al., 2020, 2023). Several privacy-preserving techniques such as data-sanitization (Zhao et al., 2022; Kandpal et al., 2022) and differentially-private training (Yue et al., 2023; Tang et al., 2024; Hong et al., 2024) algorithms have been proposed to handle data leakage through membership inference attack (Shejwalkar et al., 2021; Hu et al., 2022) or training data extraction (Salem et al., 2020; Carlini et al., 2020).
Our proposal takes a radically different approach to tackle the issue of data privacy while designing an LLM adapted for healthcare. We leverage re-

ECIR 2025

Evaluating LLM Abilities to Understand Tabular Electronic Health Records:

A Comprehensive Study of Patient Data Extraction and Retrieval

Jesús Lovón-Melgarejo - Martin Mouysset - Jo Oleiwan - **Jose G. Moreno**
Christine Damase-Michel - Lynda Tamine

CHU TOULOUSE

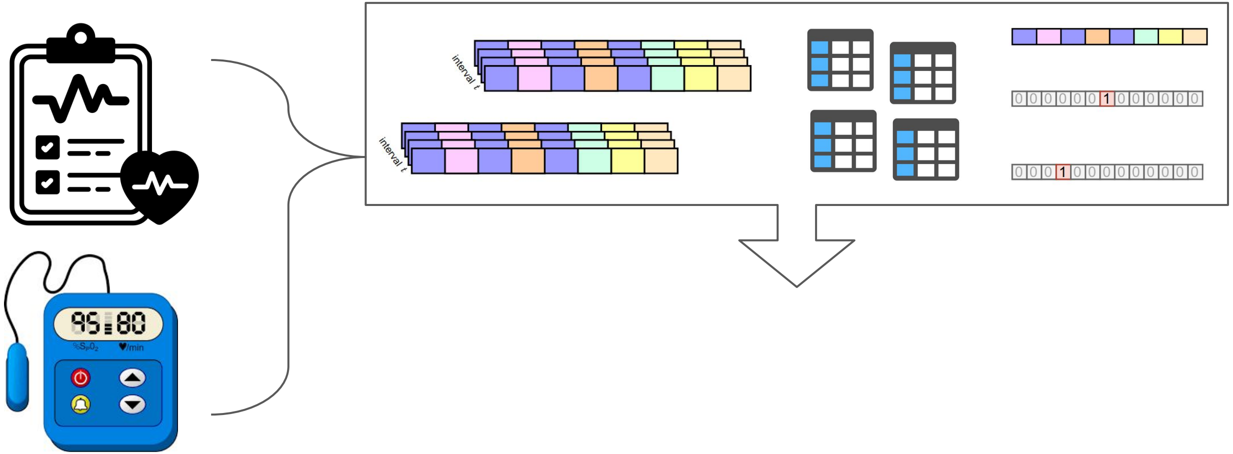
CERPOP

IRIT

Université de Toulouse

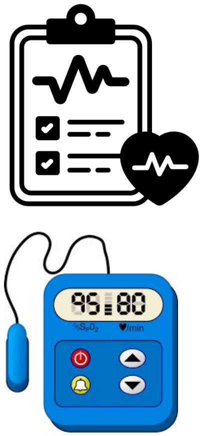
Motivation

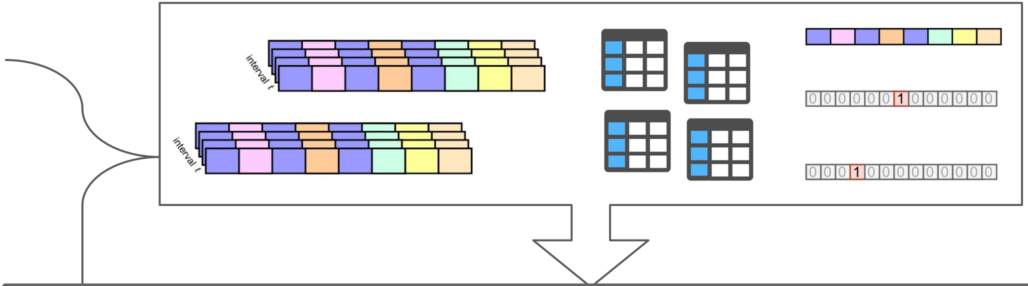
An **Electronic Health Record** (EHR) is digital patient's medical history containing **multi-table relational schemas** (labs, medications, diagnoses), **high dimensionality** (thousands of features across dozens of tables), and **heterogeneous data formats** (categorical, continuous, temporal).



Motivation

An **Electronic Health Record** (EHR) is digital patient's medical history containing **multi-table relational schemas** (labs, medications, diagnoses), **high dimensionality** (thousands of features across dozens of tables), and **heterogeneous data formats** (categorical, continuous, temporal).

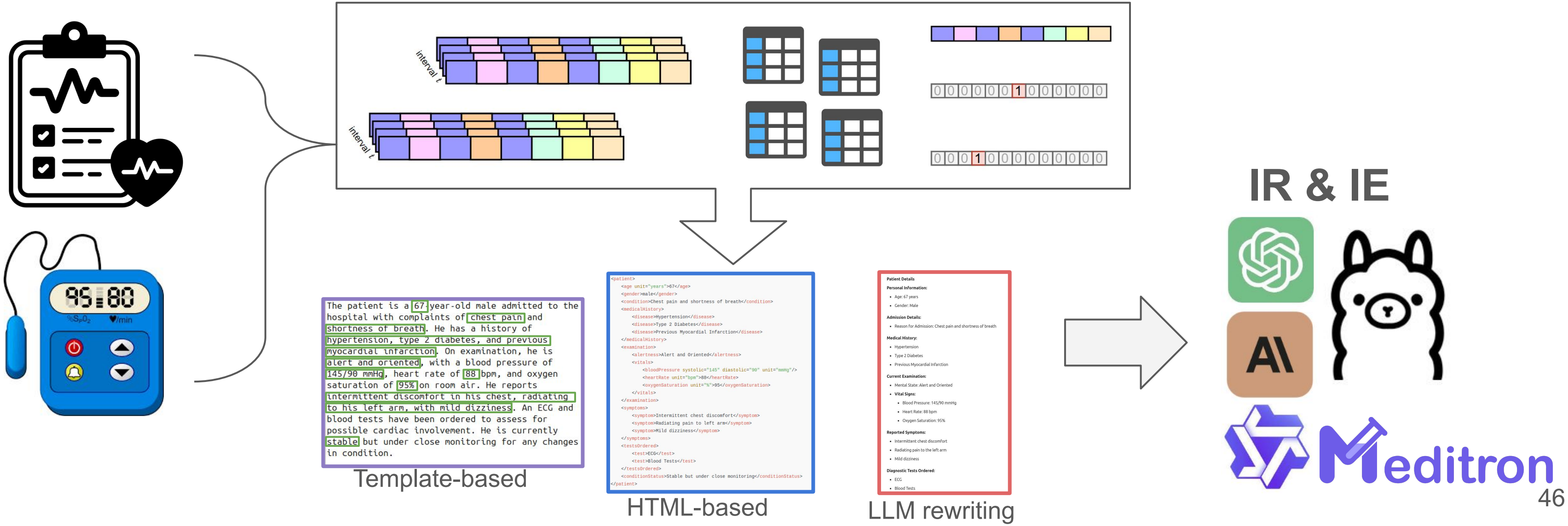




Because of the **remarkable capabilities of LLMs** on text data, with this work, we want to **draw the best practices** to use LLM by studying different "readable" patient representations for **two healthcare-related tasks** (information extraction and retrieval) that require integration of structural schemas and semantic analysis of the content

Motivation

An **Electronic Health Record** (EHR) is digital patient's medical history containing **multi-table relational schemas** (labs, medications, diagnoses), **high dimensionality** (thousands of features across dozens of tables), and **heterogeneous data formats** (categorical, continuous, temporal).



Challenges

The **tabular nature of EHRs** raise the following challenges

- Lack of standardized serialization methods for tabular data (text transformation)
- Lack of generalizability across data representations
- Lack of grounding with prior (pre-trained) knowledge

Data/evaluation challenge: no existing dataset for LLM evaluation using EHR data

Challenges

The **tabular nature of EHRs** raise the following challenges

- Lack of standardized serialization methods for tabular data (text transformation)
- Lack of generalizability across data representations
- Lack of grounding with prior (pre-trained) knowledge

Data/evaluation challenge: no existing dataset for LLM evaluation using EHR data

Our contributions are:

- An **extensive evaluation** of two LLMs on IE and IR tasks
- A new MIMIC-based **dataset for patient information extraction and retrieval**

Study Design - EHR Tasks

- Repository **R** that contains raw tabular EHR data
    - features $\mathbf{F} = \{\mathbf{f_1}, \ldots, \mathbf{f_k}\}$
    - Patient $\mathbf{p_i}$ can be formalized as a reference table $\mathbf{T_i}$ structured using a subset of features $\mathbf{F_{pi}} \subseteq \mathbf{F}$ where $\mathbf{F_{pi}} = \{\mathbf{f_{pi1}}, \ldots, \mathbf{f_{piki}}\}$, with $\mathbf{k_i}$ is the number of EHR features in $\mathbf{T_i}$.
- ***Extraction*****:** Answer specific queries about a patient's medical history
    - Input: ($\mathbf{p_i}$ serialized data, Text query extraction **q**)
    - Expected output: set of $\{\mathbf{f_{pij}}\}$ that satisfies query extraction **q**

find the primary disease and diagnoses icd9 code of the patient?

Study Design - EHR Tasks

- Repository **R** that contains raw tabular EHR data
  - features $\mathbf{F} = \{\mathbf{f}_1, \ldots, \mathbf{f}_k\}$
  - Patient $\mathbf{p}_i$ can be formalized as a reference table $\mathbf{T}_i$ structured using a subset of features $\mathbf{F}_{pi} \subseteq \mathbf{F}$ where $\mathbf{F}_{pi} = \{\mathbf{f}_{pi1}, \ldots, \mathbf{f}_{piki}\}$, with $\mathbf{k}_i$ is the number of EHR features in $\mathbf{T}_i$.
- ***Extraction*****:** Answer specific queries about a patient's medical history
  - Input: ($\mathbf{p}_i$ serialized data, Text query extraction **q**)
  - Expected output: set of $\{\mathbf{f}_{pij}\}$ that satisfies query extraction **q**

    find the primary disease and diagnoses icd9 code of the patient?

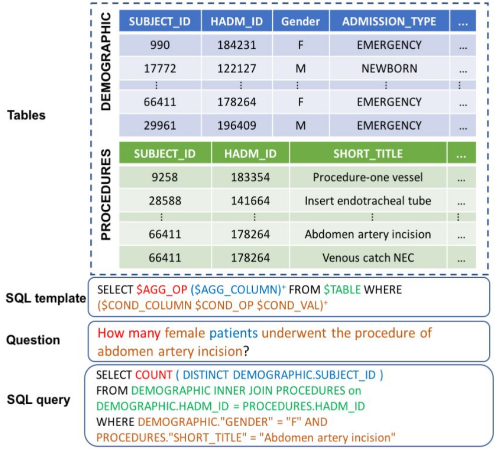
- ***Retrieval*****:** Finding relevant patients matching specific clinical criteria
  - Input: (**R**, Text query criteria **q**)
  - Expected output: Ranked set of $\{\mathbf{p}_i\}$ that satisfies query criteria **q**

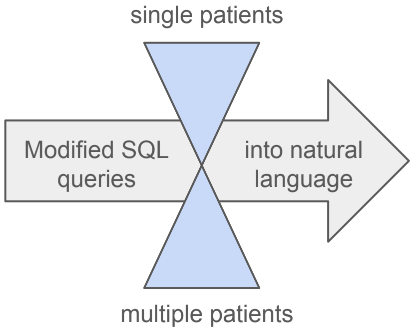
    which patients diagnosed under icd9 code 76525 had cerebrospinal fluid as lab test fluid?

Study Design - MIMIC$_{ask}$ and MIMIC$_{search}$ datasets

Based on MIMIC III (and MIMICSQL), we generate two new MIMIC variants: **MIMIC**$_{ask}$ and **MIMIC**$_{search}$

[Wang et al., 2020]





Single - MIMIC$_{ask}$
Multiple - MIMIC$_{search}$

	# patients(n)	# features(k)	# k/n	# train	# dev	# test
MIMIC$_{ask}$	100	5414	34	861	96	372
MIMIC$_{search}$	4000	19970	557	2204	368	1101^{full} | 250^{small}
MIMICSQL	46520	32340	3912	8000	1000	1000

[Wang et al., 2020] Ping Wang, Tian Shi, and Chandan K. Reddy. "Text-to-SQL Generation for Question Answering on Electronic Medical Records."The Web Conference 2020 (WWW'20).

Study Design - Prompting strategies

We consider prompts as triplets following the concatenation of elements in the form:

< *Instruction*, [*Demonstrations*], *Context* >

Prompting strategies defined by multiple configurations of the prompt format:

- *Instruction*: How to address the task
  - **Guided** (task-specific clinical heuristics)
  - **Non-guided** (general task description)
- *Demonstrations*: How select examples (zero-shot vs. few-shot using ICL)
  - **Query**-based vs. **Patient**-based similarity selection
- *Context*: How represent EHR data (serialization methods and feature selection)
  - Feature selection: **all** vs. **random**
  - Value aggregation: **raw** vs. **avg** (temporal-aggregated)
  - Serialization: **txt** vs. **xsep** vs. **sgen**

Research Questions

We explore multiple prompts elements that may affect performance:

- (RQ1) How can the **structure and content of tabular EHRs** be leveraged to grasp insights when applied in **EHR tasks**?
- (RQ2) How effective is **guided task completion** for EHR tasks?
- (RQ3) How does the **choice of demonstrations in ICL** affect performance on EHR tasks?

Results - RQ1

Leveraging tabular EHR structure and content

- **Extraction: Requires deeper (fine-grain) EHR comprehension**
- Retrieval: Requires global (coarse-grain) EHR comprehension

F^p	ϕ	Llama Extraction		Llama Retrieval		Llama $\Delta\%$	Meditron Extraction		Meditron Retrieval		Meditron $\Delta\%$
		B_{score}	R-1	MAP	R		B_{score}	R-1	MAP	R	
all	txt	56.18	22.84	9.30	32.19	+26.79	56.21	23.26	8.31	29.01	+21.53
all	xsep	57.10	20.97	9.80	33.39	+27.64	52.10	14.94	7.65	27.44	+11.12
all	sgen	57.80	23.25	**9.84**	**33.62**	+11.11	52.47	17.51	7.63	24.67	+4.59
all_{avg}	txt	56.86	23.28	8.25	27.70	+22.44	57.26	**25.89**	**10.34**	**32.09**	+21.88
all_{avg}	xsep	57.30	21.84	7.98	28.35	+19.06	54.88	19.85	8.14	29.03	+23.40
all_{avg}	sgen	**58.46**	**24.36**	8.52	32.00	+7.01	**57.79**	23.95	8.05	29.10	+6.62
rnd	txt	51.60	12.69	8.72	28.83	-	53.61	14.09	8.27	25.07	-
rnd	xsep	52.14	12.94	8.98	25.71	-	50.53	11.60	7.32	25.39	-
rnd	sgen	55.89	18.16	9.21	31.67	-	51.57	15.08	7.19	26.14	-
rnd_{avg}	txt	51.76	12.91	8.34	27.73	-	53.66	14.66	10.30	30.91	-
rnd_{avg}	xsep	52.54	12.75	8.17	28.87	-	50.83	12.69	7.69	23.53	-
rnd_{avg}	sgen	56.58	20.63	7.90	32.39	-	56.72	20.40	7.53	29.02	-

- Serialization method
  - Llama (sgen) and Meditron (txt, sgen) outperform on both tasks
- Meditron best performance on txt over sgen -> medical knowledge captured by Meditron endows it with better abilities to leverage EHR information
- Patient data retrieval is a more difficult task than patient data extraction for both LLMs
-

Results - RQ1

Leveraging tabular EHR structure and content

- Extraction: Requires deeper (fine-grain) EHR comprehension
- **Retrieval: Requires global (coarse-grain) EHR comprehension**

		Llama					Meditron				
F^p	ϕ	Extraction		Retrieval		$\Delta\%$	Extraction		Retrieval		$\Delta\%$
		B_{score}	R-1	MAP	R		B_{score}	R-1	MAP	R	
all	txt	56.18	22.84	9.30	32.19	+26.79	56.21	23.26	8.31	29.01	+21.53
	xsep	57.10	20.97	9.80	33.39	+27.64	52.10	14.94	7.65	27.44	+11.12
	sgen	57.80	23.25	**9.84**	**33.62**	+11.11	52.47	17.51	7.63	24.67	+4.59
all_{avg}	txt	56.86	23.28	8.25	27.70	+22.44	57.26	**25.89**	**10.34**	**32.09**	+21.88
	xsep	57.30	21.84	7.98	28.35	+19.06	54.88	19.85	8.14	29.03	+23.40
	sgen	**58.46**	**24.36**	8.52	32.00	+7.01	**57.79**	23.95	8.05	29.10	+6.62
rnd	txt	51.60	12.69	8.72	28.83	-	53.61	14.09	8.27	25.07	-
	xsep	52.14	12.94	8.98	25.71	-	50.53	11.60	7.32	25.39	-
	sgen	55.89	18.16	9.21	31.67	-	51.57	15.08	7.19	26.14	-
rnd_{avg}	txt	51.76	12.91	8.34	27.73	-	53.66	14.66	10.30	30.91	-
	xsep	52.54	12.75	8.17	28.87	-	50.83	12.69	7.69	23.53	-
	sgen	56.58	20.63	7.90	32.39	-	56.72	20.40	7.53	29.02	-

- Serialization method
  - Llama (sgen) and Meditron (txt, sgen) outperform on both tasks
- Meditron best performance on txt over sgen -> medical knowledge captured by Meditron endows it with better abilities to leverage EHR information
- Patient data retrieval is a more difficult task than patient data extraction for both LLMs
-

Results - RQ1

Leveraging tabular EHR structure and content

- Extraction: Requires deeper (fine-grain) EHR comprehension
- Retrieval: Requires global (coarse-grain) EHR comprehension

		Llama					Meditron				
F^p	ϕ	Extraction		Retrieval		$\Delta\%$	Extraction		Retrieval		$\Delta\%$
		B_{score}	R-1	MAP	R		B_{score}	R-1	MAP	R	
all	txt	56.18	22.84	9.30	32.19	+26.79	56.21	23.26	8.31	29.01	+21.53
all	xsep	57.10	20.97	9.80	33.39	+27.64	52.10	14.94	7.65	27.44	+11.12
all	sgen	57.80	23.25	**9.84**	**33.62**	+11.11	52.47	17.51	7.63	24.67	+4.59
all_{avg}	txt	56.86	23.28	8.25	27.70	+22.44	57.26	**25.89**	**10.34**	**32.09**	+21.88
all_{avg}	xsep	57.30	21.84	7.98	28.35	+19.06	54.88	19.85	8.14	29.03	+23.40
all_{avg}	sgen	**58.46**	**24.36**	8.52	32.00	+7.01	**57.79**	23.95	8.05	29.10	+6.62
rnd	txt	51.60	12.69	8.72	28.83	-	53.61	14.09	8.27	25.07	-
rnd	xsep	52.14	12.94	8.98	25.71	-	50.53	11.60	7.32	25.39	-
rnd	sgen	55.89	18.16	9.21	31.67	-	51.57	15.08	7.19	26.14	-
rnd_{avg}	txt	51.76	12.91	8.34	27.73	-	53.66	14.66	10.30	30.91	-
rnd_{avg}	xsep	52.54	12.75	8.17	28.87	-	50.83	12.69	7.69	23.53	-
rnd_{avg}	sgen	56.58	20.63	7.90	32.39	-	56.72	20.40	7.53	29.02	-

- **Serialization method**
  - **Llama (sgen) and Meditron (avg - txt, sgen) outperform on both tasks**
- Meditron best performance on txt over sgen -> medical knowledge captured by Meditron endows it with better abilities to leverage EHR information
- Patient data retrieval is a more difficult task than patient data extraction for both LLMs
-

Results - RQ1

Leveraging tabular EHR structure and content

- Extraction: Requires deeper (fine-grain) EHR comprehension
- Retrieval: Requires global (coarse-grain) EHR comprehension

		Llama					Meditron				
F^p	ϕ	Extraction		Retrieval		$\Delta\%$	Extraction		Retrieval		$\Delta\%$
		B_{score}	R-1	MAP	R		B_{score}	R-1	MAP	R	
all	txt	56.18	22.84	9.30	32.19	+26.79	56.21	23.26	8.31	29.01	+21.53
	xsep	57.10	20.97	9.80	33.39	+27.64	52.10	14.94	7.65	27.44	+11.12
	sgen	57.80	23.25	**9.84**	**33.62**	+11.11	52.47	17.51	7.63	24.67	+4.59
all_{avg}	txt	56.86	23.28	8.25	27.70	+22.44	57.26	**25.89**	**10.34**	**32.09**	+21.88
	xsep	57.30	21.84	7.98	28.35	+19.06	54.88	19.85	8.14	29.03	+23.40
	sgen	**58.46**	**24.36**	8.52	32.00	+7.01	**57.79**	23.95	8.05	29.10	+6.62
rnd	txt	51.60	12.69	8.72	28.83	-	53.61	14.09	8.27	25.07	-
	xsep	52.14	12.94	8.98	25.71	-	50.53	11.60	7.32	25.39	-
	sgen	55.89	18.16	9.21	31.67	-	51.57	15.08	7.19	26.14	-
rnd_{avg}	txt	51.76	12.91	8.34	27.73	-	53.66	14.66	10.30	30.91	-
	xsep	52.54	12.75	8.17	28.87	-	50.83	12.69	7.69	23.53	-
	sgen	56.58	20.63	7.90	32.39	-	56.72	20.40	7.53	29.02	-

- Serialization method
  - Llama (sgen) and Meditron (avg - txt, sgen) outperform on both tasks
- **Meditron performance is better on txt over sgen -> medical knowledge captured by Meditron endows it with better abilities to leverage EHR information**
- Patient data retrieval is a more difficult task than patient data extraction for both LLMs

Results - RQ1

Leveraging tabular EHR structure and content

- Extraction: Requires deeper (fine-grain) EHR comprehension
- Retrieval: Requires global (coarse-grain) EHR comprehension

		Llama					Meditron				
F^p	ϕ	Extraction		Retrieval		$\Delta\%$	Extraction		Retrieval		$\Delta\%$
		B_{score}	R-1	MAP	R		B_{score}	R-1	MAP	R	
all	txt	56.18	22.84	9.30	32.19	+26.79	56.21	23.26	8.31	29.01	+21.53
	xsep	57.10	20.97	9.80	33.39	+27.64	52.10	14.94	7.65	27.44	+11.12
	sgen	57.80	23.25	**9.84**	**33.62**	+11.11	52.47	17.51	7.63	24.67	+4.59
all_{avg}	txt	56.86	23.28	8.25	27.70	+22.44	57.26	**25.89**	**10.34**	**32.09**	+21.88
	xsep	57.30	21.84	7.98	28.35	+19.06	54.88	19.85	8.14	29.03	+23.40
	sgen	**58.46**	**24.36**	8.52	32.00	+7.01	**57.79**	23.95	8.05	29.10	+6.62
rnd	txt	51.60	12.69	8.72	28.83	-	53.61	14.09	8.27	25.07	-
	xsep	52.14	12.94	8.98	25.71	-	50.53	11.60	7.32	25.39	-
	sgen	55.89	18.16	9.21	31.67	-	51.57	15.08	7.19	26.14	-
rnd_{avg}	txt	51.76	12.91	8.34	27.73	-	53.66	14.66	10.30	30.91	-
	xsep	52.54	12.75	8.17	28.87	-	50.83	12.69	7.69	23.53	-
	sgen	56.58	20.63	7.90	32.39	-	56.72	20.40	7.53	29.02	-

- Serialization method
  - Llama (sgen) and Meditron (avg - txt, sgen) outperform on both tasks
- Meditron best performance on txt over sgen -> medical knowledge captured by Meditron endows it with better abilities to leverage EHR information
- **Patient data retrieval is a more difficult task than patient data extraction for both LLMs**

Results - RQ2

Guiding task completion: Evaluation on the impact of the instruction component

- **Minimal performance difference between guided and non-guided instructions**

Task-Specific Variations:

- Extraction tasks benefit slightly from more detailed instructions
- Retrieval tasks show no improvement with instruction elaboration

Model Differences: Meditron shows more sensitivity to instruction design than Llama

	Llama								**Meditron**							
(F^p,ϕ)	(all, sgen)				$(\text{all}_{avg}, \text{sgen})$				$(\text{all}_{avg}, \text{txt})$				$(\text{all}_{avg}, \text{sgen})$			
	Extraction		**Retrieval**		**Extraction**		**Retrieval**		**Extraction**		**Retrieval**		**Extraction**		**Retrieval**	
I_e/I_r	B_{score}	R-1	MAP	R	B_{score}	R-1	MAP	R	B_{score}	R-1	MAP	R	B_{score}	R-1	MAP	R
Guided	**57.92**	**23.44**	9.56	33.42	**58.93**	**24.98**	**9.22**	31.32	57.26	25.71	**12.07**	**32.54**	**57.82**	23.77	7.98	28.95
Non-Guided	57.80	23.25	**9.84**	**33.62**	58.46	24.36	8.52	**32.00**	57.26	**25.89**	10.34	32.09	57.79	**23.95**	**8.05**	**29.10**

Results - RQ2

Guiding task completion: Evaluation on the impact of the instruction component

- Minimal performance difference between guided and non-guided instructions

Task-Specific Variations:

- **Providing more detailed instructions offers a slight improvement in extraction tasks**
- Retrieval tasks show no improvement with instruction elaboration

Model Differences: Meditron shows more sensitivity to instruction design than Llama

(F^p,ϕ)	Llama								Meditron							
	(all, sgen)				(all_{avg},sgen)				(all_{avg},txt)				(all_{avg},sgen)			
	Extraction		**Retrieval**		**Extraction**		**Retrieval**		**Extraction**		**Retrieval**		**Extraction**		**Retrieval**	
I_e/I_r	B_{score}	R-1	MAP	R	B_{score}	R-1	MAP	R	B_{score}	R-1	MAP	R	B_{score}	R-1	MAP	R
Guided	**57.92**	**23.44**	9.56	33.42	**58.93**	**24.98**	**9.22**	31.32	57.26	25.71	**12.07**	**32.54**	**57.82**	23.77	7.98	28.95
Non-Guided	57.80	23.25	**9.84**	**33.62**	58.46	24.36	8.52	**32.00**	57.26	**25.89**	10.34	32.09	57.79	**23.95**	**8.05**	**29.10**

Results - RQ2

Guiding task completion: Evaluation on the impact of the instruction component

- Minimal performance difference between guided and non-guided instructions

Task-Specific Variations:

- Extraction tasks benefit slightly from more detailed instructions
- **Retrieval tasks show no improvement with instruction elaboration**

Model Differences: Meditron shows more sensitivity to instruction design than Llama

	Llama								Meditron							
(F^p,ϕ)	(all, sgen)				(all_{avg},sgen)				(all_{avg},txt)				(all_{avg},sgen)			
	Extraction		**Retrieval**		**Extraction**		**Retrieval**		**Extraction**		**Retrieval**		**Extraction**		**Retrieval**	
I_e/I_r	B_{score}	R-1	MAP	R	B_{score}	R-1	MAP	R	B_{score}	R-1	MAP	R	B_{score}	R-1	MAP	R
Guided	**57.92**	**23.44**	9.56	33.42	**58.93**	**24.98**	**9.22**	31.32	57.26	25.71	**12.07**	**32.54**	**57.82**	23.77	7.98	28.95
Non-Guided	57.80	23.25	**9.84**	**33.62**	58.46	24.36	8.52	**32.00**	57.26	**25.89**	10.34	32.09	57.79	**23.95**	**8.05**	**29.10**

Results - RQ2

Guiding task completion: Evaluation on the impact of the instruction component

- Minimal performance difference between guided and non-guided instructions

Task-Specific Variations:

- Extraction tasks benefit slightly from more detailed instructions
- Retrieval tasks show no significant improvement with instruction elaboration

Model Differences: Llama benefits slightly more from explicit guidance than Meditron

	Llama								**Meditron**							
(F^p, ϕ)	(all, sgen)				(all_{avg},sgen)				(all_{avg},txt)				(all_{avg},sgen)			
	Extraction		**Retrieval**		**Extraction**		**Retrieval**		**Extraction**		**Retrieval**		**Extraction**		**Retrieval**	
I_e/I_r	B_{score}	R-1	MAP	R	B_{score}	R-1	MAP	R	B_{score}	R-1	MAP	R	B_{score}	R-1	MAP	R
Guided	**57.92**	**23.44**	9.56	33.42	**58.93**	**24.98**	**9.22**	31.32	57.26	25.71	**12.07**	**32.54**	**57.82**	23.77	7.98	28.95
Non-Guided	57.80	23.25	**9.84**	**33.62**	58.46	24.36	8.52	**32.00**	57.26	**25.89**	10.34	32.09	57.79	**23.95**	**8.05**	**29.10**

Results - RQ3

Selecting demonstrations: Evaluation on the impact of demonstration quality in an ICL setup on task performance.

Task-Specific:

- **Extraction task: improvements by using demonstrations**
- Retrieval task: Zero-shot approaches surprisingly outperform few-shot methods

Example Selection: Query-based examples (similar questions) outperform patient-based examples

	(F^p,ϕ) I_e/I_r	ICL w/ Llama (Lma_{icl}) (all, sgen) Non-Guided				(all$_{avg}$, sgen) Guided			
σ	#ex	**Extraction**		**Retrieval**		**Extraction**		**Retrieval**	
		B_{score}	R-1	MAP	R	B_{score}	R-1	MAP	R
-	0	57.80	23.25	**9.84**	**33.62**	58.93	24.98	9.22	31.32
Patient σ_p	1	59.47	26.88	N/A	N/A	60.75	30.61	N/A	N/A
	2	59.97	27.16	N/A	N/A	60.51	29.68	N/A	N/A
	3	60.75	28.75	N/A	N/A	60.75	29.53	N/A	N/A
Query σ_q	1	60.82	28.90	8.06	29.50	60.36	30.61	7.84	29.82
	2	61.42	30.04	8.07	29.43	61.90	31.87	8.81	30.80
	3	**61.83**	**31.48**	8.28	31.95	**62.44**	**32.39**	8.85	34.17
Random	1	58.66	25.33	7.87	29.63	59.89	28.55	7.98	29.60
	2	59.90	26.60	7.87	30.20	59.80	27.76	8.67	31.15
	3	59.75	26.69	8.48	31.15	60.91	28.90	**9.42**	**35.54**

Results - RQ3

Selecting demonstrations: Evaluation on the impact of demonstration quality in an ICL setup on task performance.

Task-Specific:

- Extraction task: improvements by using demonstrations
- **Retrieval task: Zero-shot approach surprisingly outperforms few-shot methods**

Example Selection: Query-based examples (similar questions) outperform patient-based examples

		ICL w/ Llama (Lma_{icl})							
	(F^p, ϕ) I_e/I_r	(all, sgen) Non-Guided				(all_{avg}, sgen) Guided			
σ	#ex	**Extraction**		**Retrieval**		**Extraction**		**Retrieval**	
		B_{score}	R-1	MAP	R	B_{score}	R-1	MAP	R
-	0	57.80	23.25	**9.84**	**33.62**	58.93	24.98	<u>9.22</u>	31.32
Patient σ_p	1	59.47	26.88	N/A	N/A	60.75	30.61	N/A	N/A
	2	59.97	27.16	N/A	N/A	60.51	29.68	N/A	N/A
	3	60.75	28.75	N/A	N/A	60.75	29.53	N/A	N/A
Query σ_q	1	60.82	28.90	8.06	29.50	60.36	30.61	7.84	29.82
	2	<u>61.42</u>	<u>30.04</u>	8.07	29.43	<u>61.90</u>	<u>31.87</u>	8.81	30.80
	3	**61.83**	**31.48**	8.28	<u>31.95</u>	**62.44**	**32.39**	8.85	<u>34.17</u>
Random	1	58.66	25.33	7.87	29.63	59.89	28.55	7.98	29.60
	2	59.90	26.60	7.87	30.20	59.80	27.76	8.67	31.15
	3	59.75	26.69	<u>8.48</u>	31.15	60.91	28.90	**9.42**	**35.54**

Results - RQ3

Selecting demonstrations: Evaluation on the impact of demonstration quality in an ICL setup on task performance.

Task-Specific:

- Extraction task: improvements by using demonstrations
- Retrieval task: Zero-shot approach surprisingly outperforms few-shot methods

Example Selection: Query-based examples (similar questions) outperform patient-based examples and random

		ICL w/ Llama ($\mathbf{Lma}_{icl}$)							
	(F^p,ϕ) I_e/I_r	(all, sgen) Non-Guided				(all_{avg}, sgen) Guided			
σ	#ex	**Extraction**		**Retrieval**		**Extraction**		**Retrieval**	
		B_{score}	R-1	MAP	R	B_{score}	R-1	MAP	R
-	0	57.80	23.25	**9.84**	**33.62**	58.93	24.98	9.22	31.32
Patient σ_p	1	59.47	26.88	N/A	N/A	60.75	30.61	N/A	N/A
	2	59.97	27.16	N/A	N/A	60.51	29.68	N/A	N/A
	3	60.75	28.75	N/A	N/A	60.75	29.53	N/A	N/A
Query σ_q	1	60.82	28.90	8.06	29.50	60.36	30.61	7.84	29.82
	2	61.42	30.04	8.07	29.43	61.90	31.87	8.81	30.80
	3	**61.83**	**31.48**	8.28	31.95	**62.44**	**32.39**	8.85	34.17
Random	1	58.66	25.33	7.87	29.63	59.89	28.55	7.98	29.60
	2	59.90	26.60	7.87	30.20	59.80	27.76	8.67	31.15
	3	59.75	26.69	8.48	31.15	60.91	28.90	**9.42**	**35.54**

Results - Comparative evaluation

Extraction									
Models	$T5_0$	$BART_0$	$T5_{ft}$	$BART_{ft}$	TREQS	Lma*	Med*	Lma^*_{ft}	Med^*_{ft}
$\mathbf{B}_{score}$	46.07	48.50	53.41	83.94	23.68	62.44	60.69	**84.79**	56.04
R-1	4.92	2.19	28.07	67.18	13.21	32.39	33.18	**74.47**	11.90

Retrieval									
Models	BM25	MonoB	MonoT5	TREQS	Lma*	Med*	Lma^*_{ft}	Med^*_{ft}
MAP	35.26	10.19	38.49	43.99	9.84	12.07	11.33	**44.34**
R	49.37	35.43	53.01	52.16	33.62	32.54	47.95	**53.38**

SQL-based

Summary of Take-Away Messages

1. **Keep all the features:** Unsurprisingly context is improved when using all available EHR features, leading to better task performance. However, if longitudinal values are present, then use feature value aggregation.
2. **Serialization method selection:** The best EHR serialization method is based on the LLM self-generated (sgen) EHR tabular descriptions.

   Medical-aligned LLMs can perform correctly with template-based serialization.

3. **Example Selection:** For ICL, demonstration **selection based on queries** (instead than based on patients) is more effective for **extraction** as the number of examples increases. Conversely, the **retrieval task** better leverages zero-shot setups.
4. **Fine-tuning Approach**: Fine-tuned LLMs with basic data-to-text EHR serialization methods achieve the best performance across tasks compared to general fine-tuned models

Code

jeslev / llm-patient-ehr Public

Code | Issues | Pull requests | Actions | Projects | Security | Insights

main · 1 Branch · 0 Tags · Go to file · Code

jesus.lovon First commit · 5d914db · 3 weeks ago · 2 Commits

data	First commit	3 weeks ago
LICENSE	Initial commit	2 months ago
README.md	First commit	3 weeks ago
ir_llm_inference.py	First commit	3 weeks ago
qa_llm_inference.py	First commit	3 weeks ago

README · MIT license

LLMs on Tabular Electronic Health Records: Data Extraction and Retrieval 🏥

License MIT · Conference Accepted

Official repository for our paper **"Evaluating LLM Abilities to Understand Tabular Electronic Health Records: A Comprehensive Study of Patient Data Extraction and Retrieval"** accepted at ECIR 2025. This repository contains code and instructions for creating datasets and running models for two main tasks:

1. **MIMIC ASK**: A patient data extraction task on Electronic Health Records (EHR)
2. **MIMIC SEARCH**: An information retrieval task on EHR data

Questions?

Paper

Evaluating LLM Abilities to Understand Tabular Electronic Health Records: A Comprehensive Study of Patient Data Extraction and Retrieval

Jesús Lovón-Melgarejo[1], Martin Mouysset[1], Jo Oleiwan[1], Jose G. Moreno[1], Christine Damase-Michel[2], and Lynda Tamine [1]

[1]Université Paul Sabatier, IRIT, Toulouse, France
[2]Centre Hospitalier Universitaire de Toulouse, CERPOP INSERM UMR 1295 - SPHERE team, Faculté de Médecine Université de Toulouse, Toulouse, France
{jesus.lovon, martin.mouysset, jo.oleiwan, jose.moreno,tamine}@irit.fr
{christine.damase-michel}@univ-tlse3.fr

Abstract. Electronic Health Record (EHR) tables pose unique challenges among which is the presence of hidden contextual dependencies between medical features with a high level of data dimensionality and sparsity. This study presents the first investigation into the abilities of LLMs to comprehend EHRs for patient data extraction and retrieval. We conduct extensive experiments using the MIMICSQL dataset to explore the impact of the prompt structure, instruction, context, and demonstration, of two backbone LLMs, Llama2 and Meditron, based on task performance. Through quantitative and qualitative analyses, our findings show that optimal feature selection and serialization methods can enhance task performance by up to 26.79% compared to naive approaches. Similarly, in-context learning setups with relevant example selection improve data extraction performance by 5.95%. Based on our study findings, we propose guidelines that we believe would help the design of LLM-based models to support health search.

Keywords: Large language models · Electronic Health Record (EHR)· tabular data· information retrieval · information extraction

Early stroke functional outcome prediction from admission clinical records

Santiago Gómez - José G. Moreno - Daniel Mantilla - Fabio Martínez.

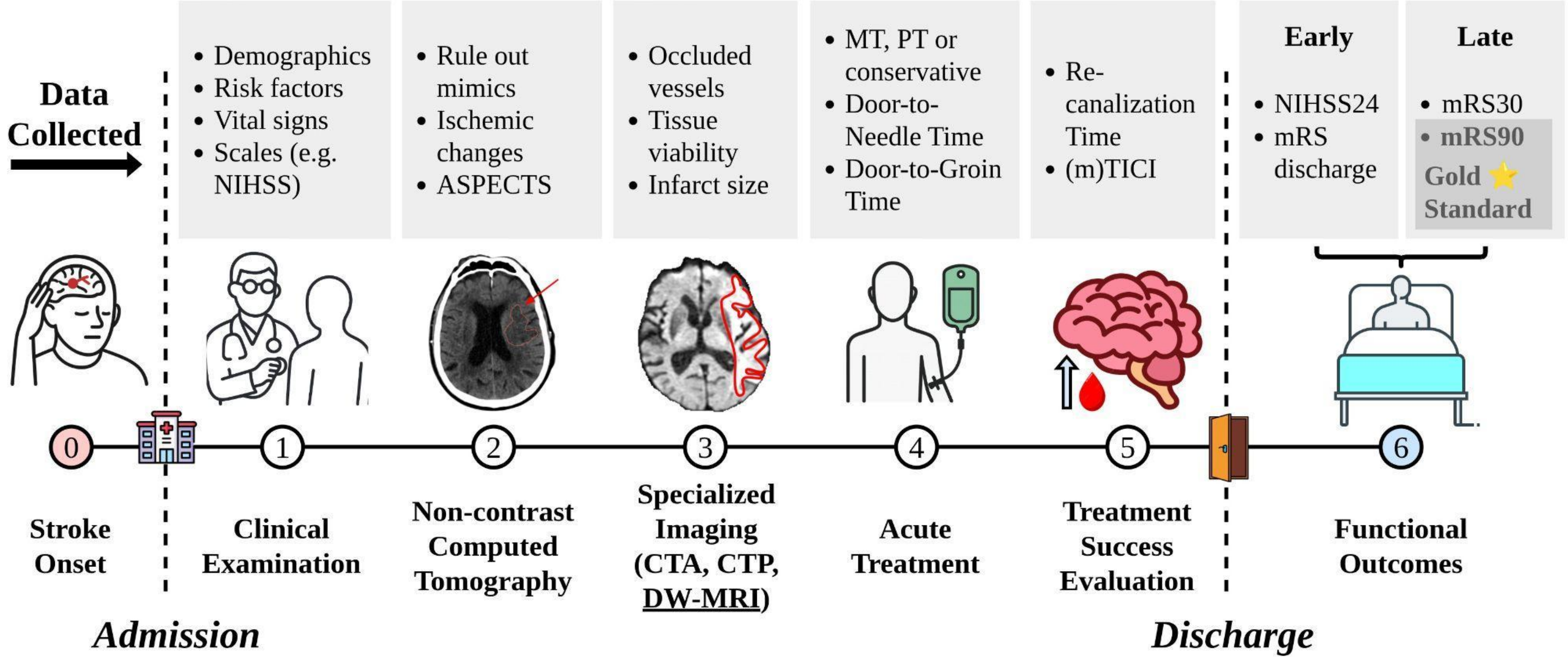
IRIT

Université de Toulouse

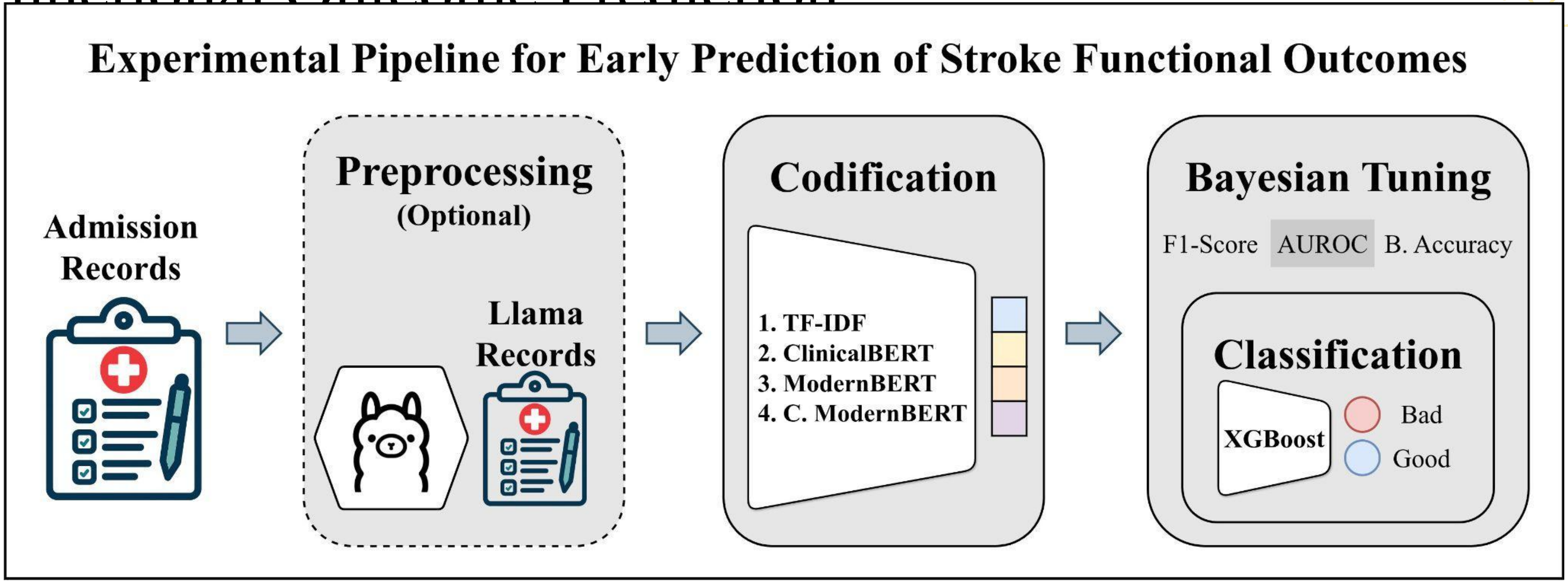
Universidad Industrial de Santander

BIVL²ab
Biomedical Imaging, Vision and Learning Laboratory

Functional Outcome Prediction

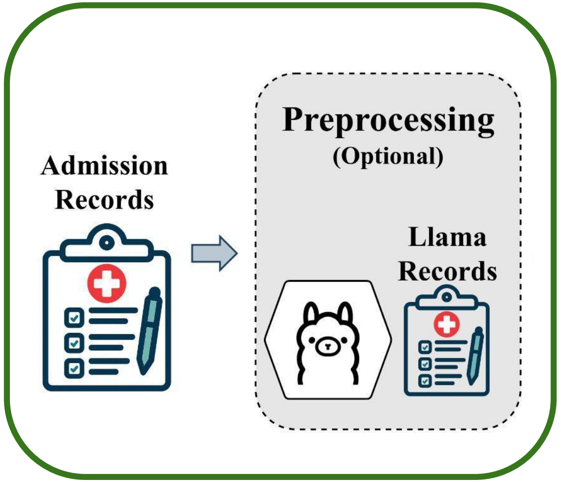


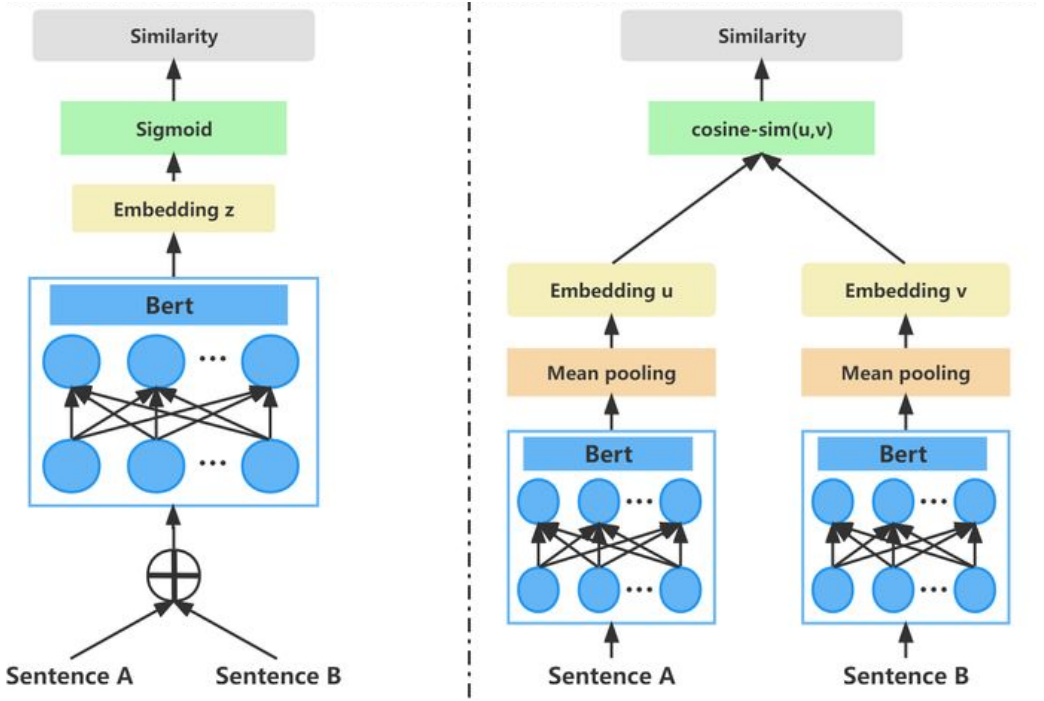
Functional Outcome Prediction



Proposed: First approximation for **functional outcome prediction based exclusively on unstructured clinical notes** collected at admission. The strategy consists of three stages: 1) preprocessing, 2) admission notes codification with lexical- and semantic-based models, and 3) vectors classification with XGBoost.

Encoding Admission Records (sentence BERT)





Dataset

284 cases of ischemic stroke treated at 2 clinics between October 2021 and April 2024

Inclusion criteria: **>= 18 years old** and no evidence of cerebral hemorrhage

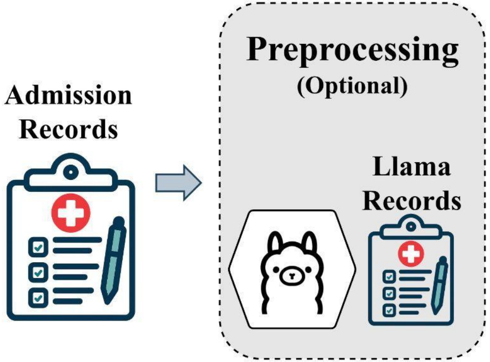
Admission notes written in **Spanish**, baseline clinical variables, details of the treatment administered, and outcome measures, modified Rankin Scale (mRS) at discharge and 30 and 90 days after treatment (exclusion criteria)

Data was **anonymized**. Patient identifiers (name, ID, and birth date) removed, and admission notes revised

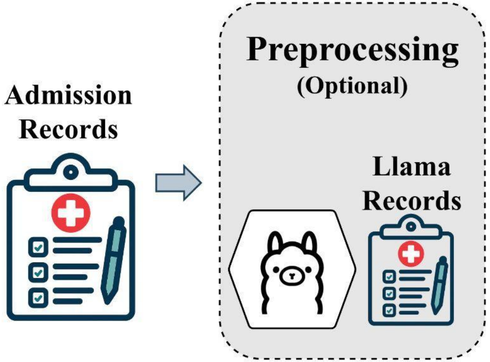
The study protocol was approved by the **ethics committees** CEINCI-UIS and CEI-FOSCAL at FOSCAL

The dataset includes patient age (73.1 ± 13.2 years) and gender (146 females, 137 males, and 1 unspecified), along with variables grouped into five categories: baseline clinical, imaging-derived measures, acute treatment, treatment success, and functional outcomes

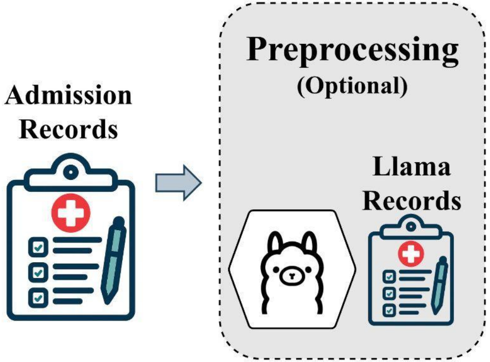
Preprocessing



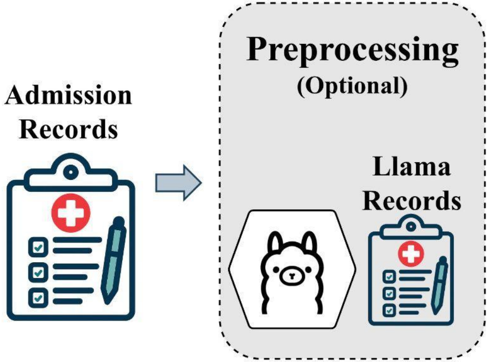
Preprocessing



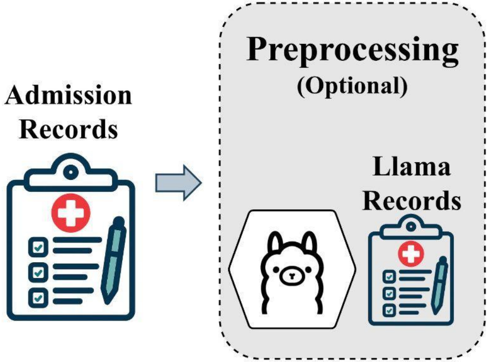
Preprocessing

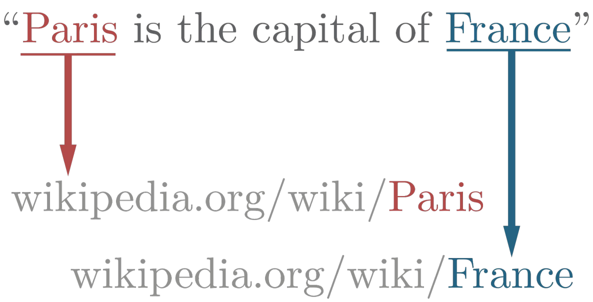


Preprocessing



Preprocessing





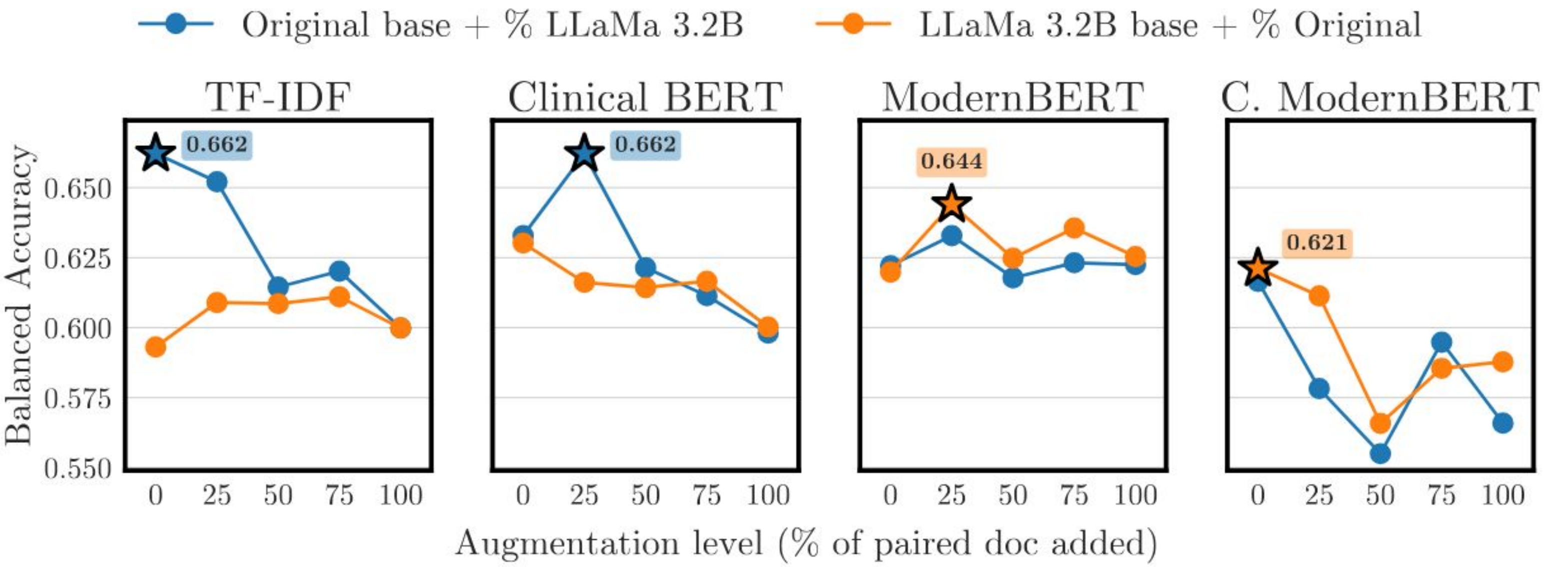
Preprocessing

Functional Outcome Prediction (parameter tuning)

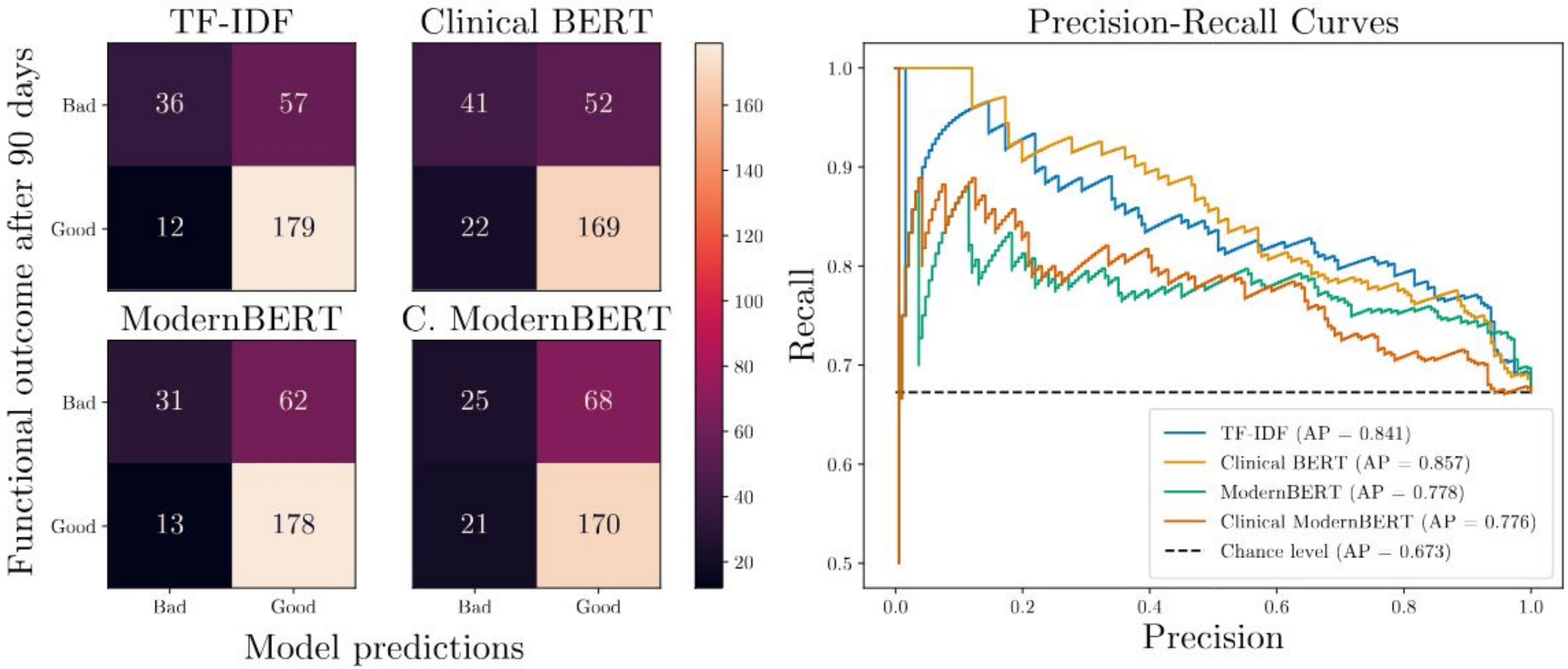
Optimization metric	AUROC	B. Acc	F1	Prec	Rec	Spec
AUROC (n=12)	**0.723**	0.557	0.801	0.700	<u>0.935</u>	0.178
B. Acc (n=12)	<u>0.692</u>	**0.623**	<u>0.812</u>	**0.737**	0.908	**0.341**
F1 (n=12)	0.673	<u>0.572</u>	**0.813**	<u>0.706</u>	**0.963**	<u>0.183</u>

- The **balanced accuracy emerged as the most well-rounded objective**, while its AUROC (0.692) and F1-score (0.812) were slightly lower, it delivered the best results in balanced accuracy (0.623), precision (0.737), and specificity (0.341).

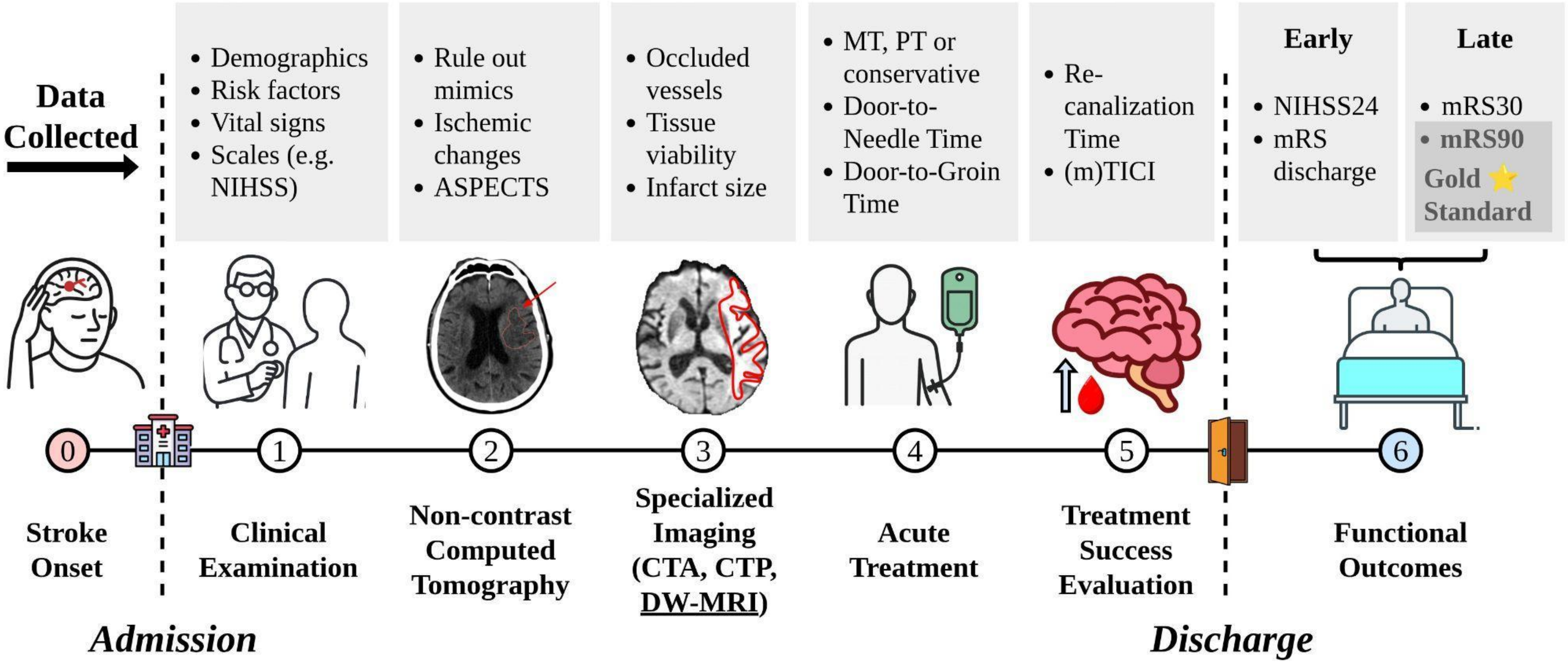
Fu



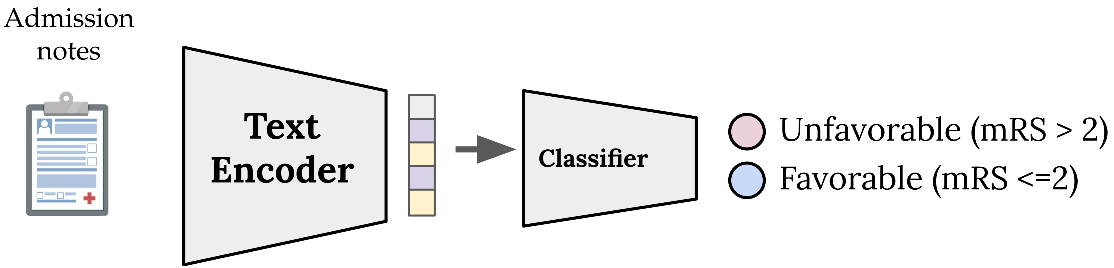
- Best: **TF-IDF with original documents (0.662)** and **ClinicalBERT with original documents augmented with 25% Llama content (0.662).**
- **Decrease in performance when augmentation exceeded 25%.**


TF-IDF
Clinical BERT
ModernBERT
C. ModernBERT
Functional outcome after 90 days
Bad
Good
36
57
12
179
41
52
22
169
31
62
13
178
25
68
21
170
Model predictions
160
140
120
100
80
60
40
20
Precision-Recall Curves
Recall
Precision
1.0
0.9
0.8
0.7
0.6
0.5
0.0
0.2
0.4
0.6
0.8
TF-IDF (AP = 0.841)
Clinical BERT (AP = 0.857)
ModernBERT (AP = 0.778)
Clinical ModernBERT (AP = 0.776)
Chance level (AP = 0.673)

Functional Outcome Prediction



Functional Outcome Prediction



	Variables	AUROC	B. Acc
(6)	All vars (early outcome vars incl.)	**0.990**	**0.962**
(5)	↳ w/o Early outcome	0.845	<u>0.781</u>
(4)	↳ w/o Treatment success	0.831	0.759
(3)	↳ w/o Acute treatment	<u>0.852</u>	0.771
(2)	↳ w/o Specialized imaging	0.841	0.770
(1)	↳ Clinical (text)	0.744	0.662

- As expected, the model trained with **all variables achieved near-perfect performance** across all metrics.
- **Removing** only the **discharge mRS** led to a **notable drop in performance**.
- The **remaining models**, exhibited **comparable performance**.

Questions?

Paper

Early stroke functional outcome prediction from admission clinical records

Santiago Gómez[1][0000-0001-6951-7452], Jose G. Moreno[2][0000-0002-8852-5797], Daniel Mantilla[3][0000-0002-5727-3195], and Fabio Martínez[1][0000-0001-7353-049X]

[1] BIVL²ab, Universidad Industrial de Santander, Bucaramanga, Colombia
[2] University of Toulouse, IRIT, UMR 5505 CNRS, F-31000, Toulouse, France
[3] Clínica FOSCAL, Floridablanca, Colombia

Abstract. Ischemic stroke, caused by the occlusion of cerebral blood vessels, requires prompt intervention to restore perfusion and improve patient prognosis. Accurate prediction of functional outcomes is essential for guiding treatment decisions and optimizing patient management. However, such predictions often rely on clinical and imaging data, which may not be available early in the care pathway. This study investigated the potential of free-text clinical notes recorded at admission to predict functional outcomes. Admission documents were encoded using lexical (TF-IDF) and semantic (BERT-based) features, and XGBoost classifiers were trained to predict whether a patient would have a favorable outcome 90 days post-stroke. Using a proprietary dataset of 284 patients, the best text-only performance was achieved with ClinicalBERT applied to original records augmented with synthetic content generated by Llama (AUROC = 0.744, balanced accuracy = 0.662). Compared to structured-data baselines incorporating clinical and non-contrast CT information, the text-based model demonstrated higher recall but lower specificity. These findings highlight admission text as a viable, low-resource alternative for early stroke outcome prediction, supporting future integration with multimodal data for real-time decision-making.

Keywords: Ischemic stroke, Functional outcome prediction, Clinical admission records, Text-based strategies

1 Introduction

Stroke is the most common cerebrovascular disease and the second leading cause of death, reporting around of 7.08 million deaths in 2020. Besides, stroke is the third leading cause of disability-adjusted life years worldwide, reporting around

Take away messages

- **Learn about LLMs (trending topic) but do not skip not modern lectures**, traditional models could be your only option! I'm a LLM believer but there is not reason to ignore traditional models!
- **Medical data is usually a scare and "inaccessible" resource**. Having access to some is already a big step in research
- **Privacy is a major issue in medical data** which could be an disadvantage for LLMs. However, adapted solutions may be helpful
- Although no deeply discussed here, **fine tuning encoder-based models may be a better option in terms of performance**, but they also may "memorize" some data
- The works presented here are more **decoder oriented guided by our project goals**

Question

Have you noted that:

- For the first work, we shared the model (LLM) and paper
- For the second work, we shared the code and paper
- For the third work, we shared only the paper

Do you understand why?

Thank you!