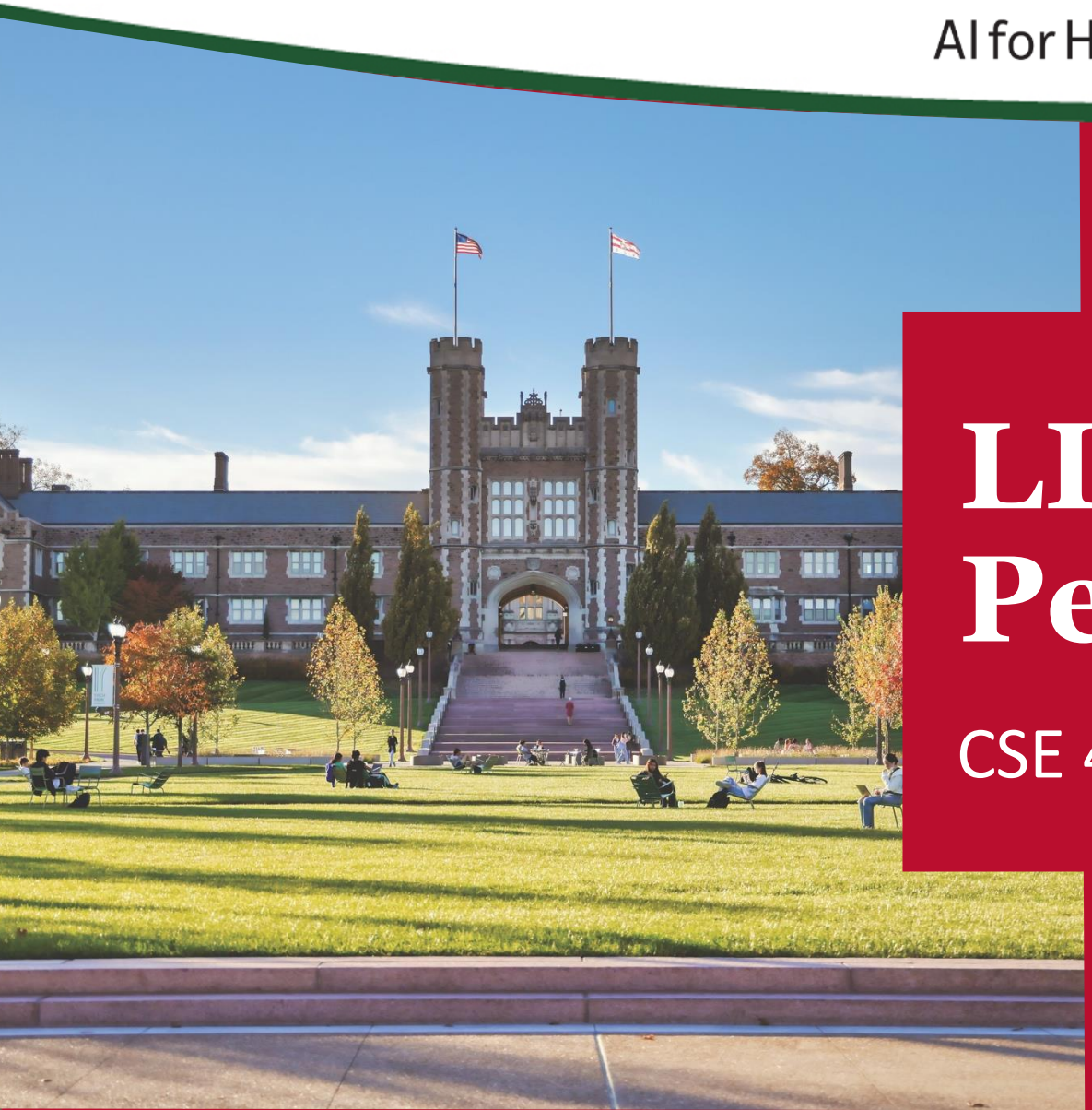
WashU McKelvey Engineering

AI for Health Institute



LLMs for Perioperative Care

CSE 4109: Introduction to AI for Health

Contents

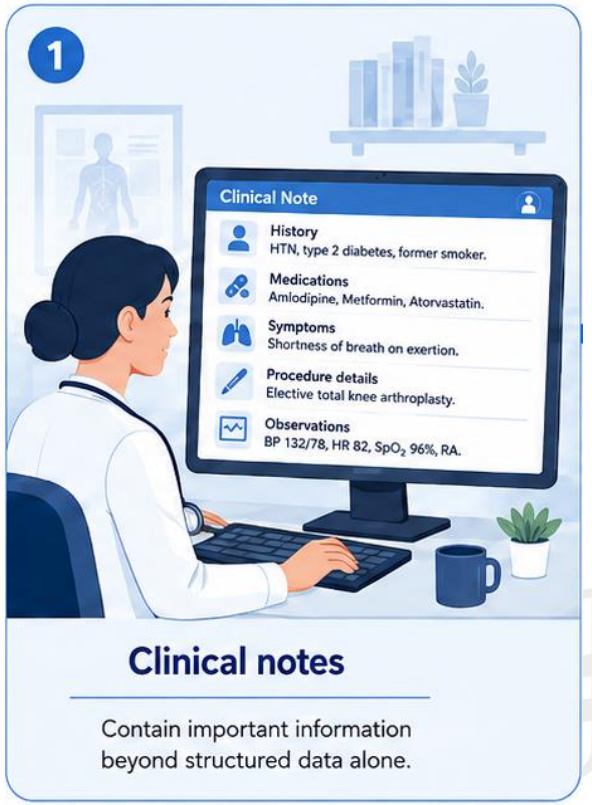
1. Why are clinical notes important?
2. Why use large language models in peri-operative care?
3. Bridging the gap: general LLMs vs clinically-oriented LLMs
4. Study: Foundational Capabilities of LLMs in medicine
5. Some resources for your final project
6. Coding session: SurgicalPLAN

Contents

1. **Why are clinical notes important?**
2. Why use large language models in peri-operative care?
3. Bridging the gap: general LLMs vs clinically-oriented LLMs
4. Study: Foundational Capabilities of LLMs in medicine
5. Some resources for your final project
6. Coding session: SurgicalPLAN

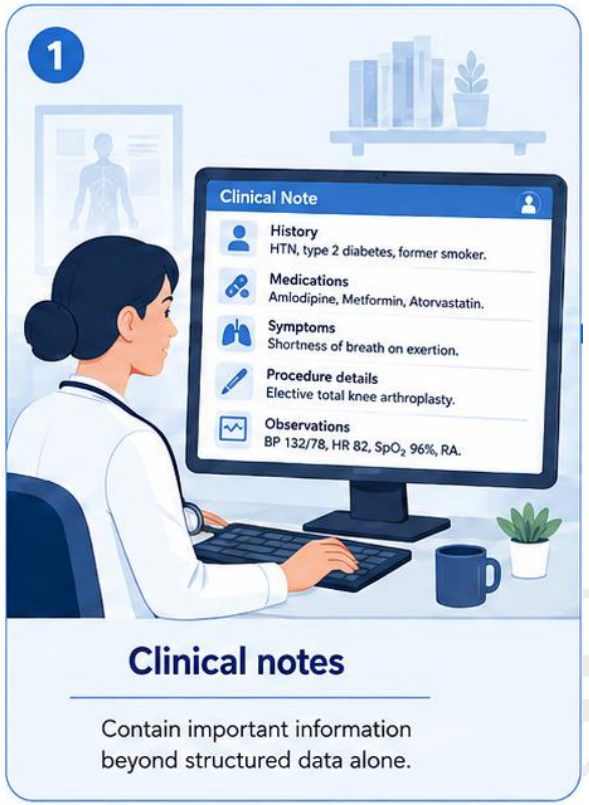
What are clinical notes? Why are they important?

- Clinicians often document:
  - Patient information (e.g., history, demographics)
  - Medications
  - Symptoms
  - Planned surgical procedures
  - Lab tests and observations
- Most information can be stored as structured Electronic Health Records (e.g., patient demographics).



What are clinical notes? Why are they important?

- But not all information could be contained as structured data.
- Clinical notes are text-based.
- Documents information that cannot be captured in a structured tabular format (e.g., planned surgical procedure details).



Contents

1. Why are clinical notes important?
2. **Why use large language models in peri-operative care?**
3. Bridging the gap: general LLMs vs clinically-oriented LLMs
4. Study: Foundational Capabilities of LLMs in medicine
5. Some resources for your final project
6. Coding session: SurgicalPLAN

LLMs in perioperative care

What is perioperative care?

- Continuum of care surrounding surgery
- Planning designed to reduce risks, support healing, and improve patient outcomes.

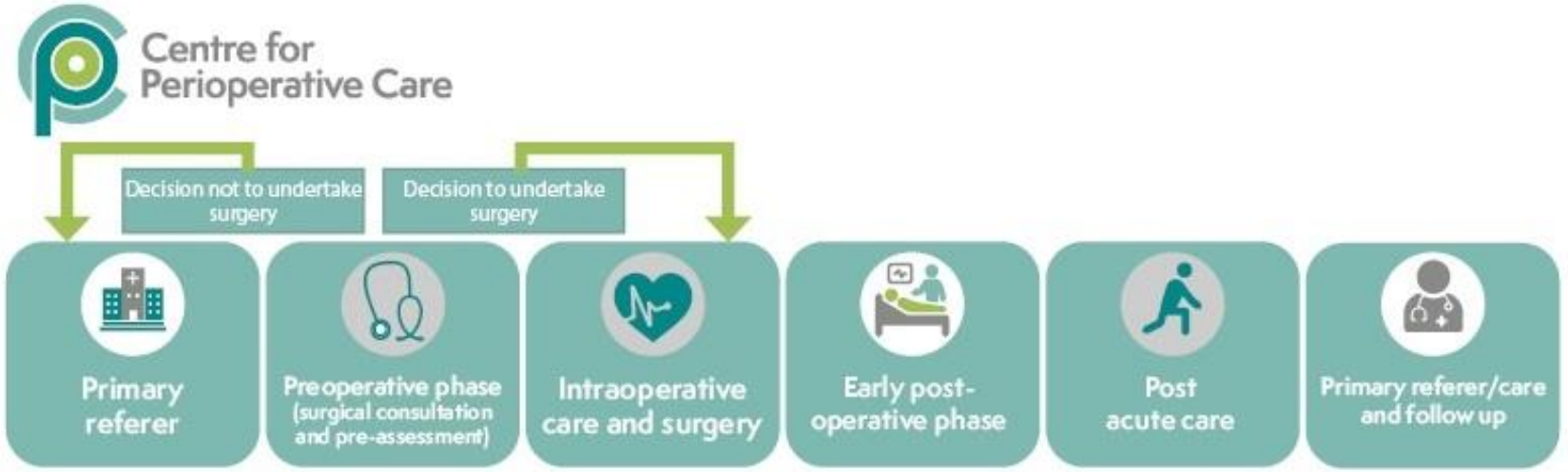


Image from Center of perioperative care

LLMs in perioperative care

Why is peri-operative care important?

- Many adverse surgical outcomes and complications are preventable.
  - Death upon discharge
  - Blood clots
  - Pneumonia
  - Delirium (confusion, disorientation)



LLMs in perioperative care

Why is peri-operative care important?

- Peri-operative care ensures comprehensive medical support provided before, during, and after surgery:
  - 🧪 Additional preoperative assessments (e.g., lab tests)
  - 🩺 Enhanced surgical preparation
  - 🏠 Discharge planning
  - 🤝 Shared decision-making
- Reduces likelihood of adverse outcomes.



LLMs in perioperative care

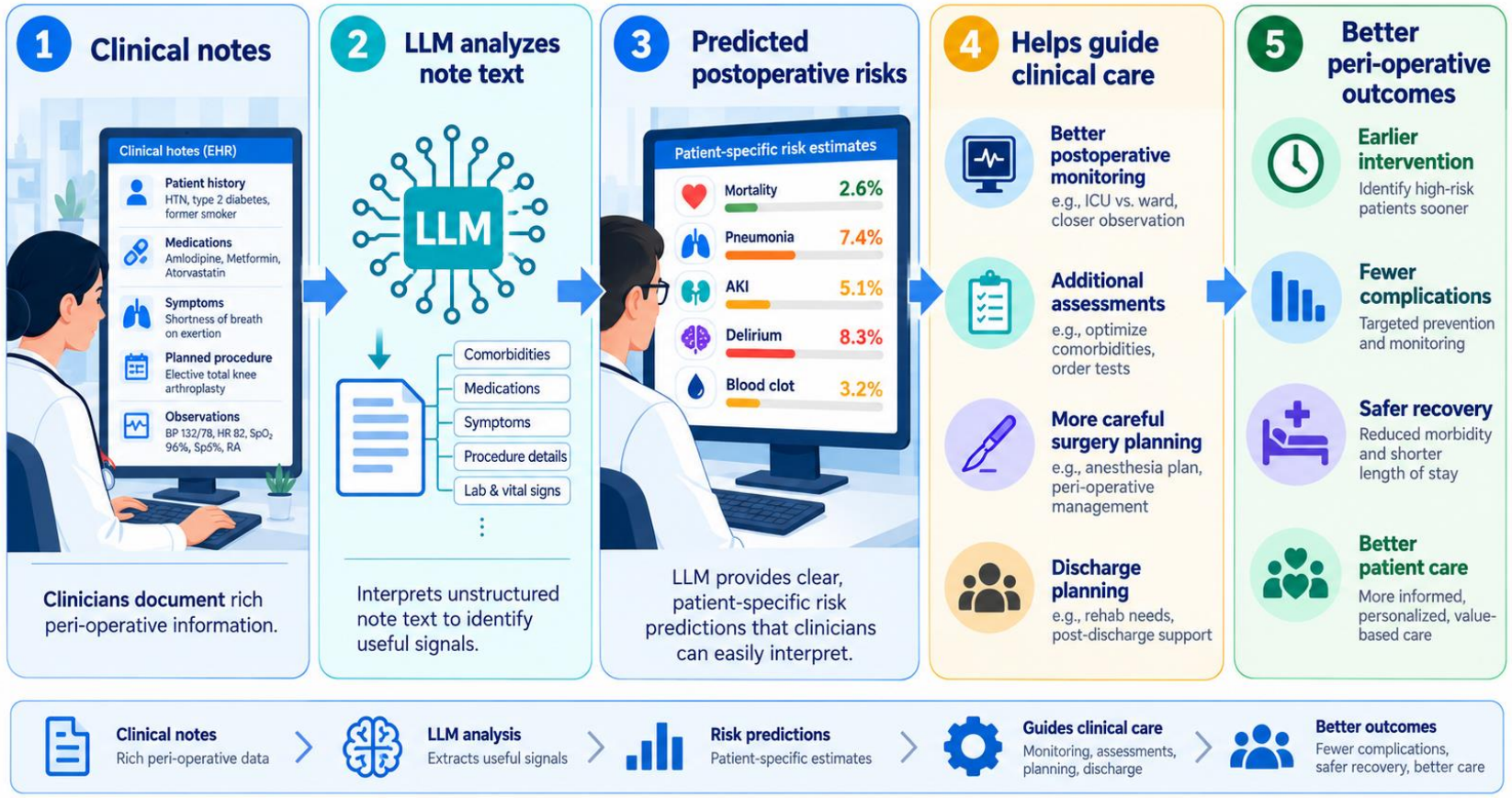
Why is peri-operative care important?

- A review by the Center of Perioperative Care, reveals that :
  - Reduce hospital stays by an average of 2 days
  - Reduce surgical complications by 30% to 80%
  - These far exceeds drug or surgical interventions



LLMs in perioperative care

What can LLMs in peri-operative care do?



LLMs in perioperative care

Why predict post-operative complications?

- Anticipatory management for perioperative contingency planning.
- Pre-emptive and early identification of risk factors.
- Reduced hospital stays and complication risks.



Image from CCI Anesthesia

Contents

1. Why are clinical notes important?
2. Why use large language models in peri-operative care?
3. **Bridging the gap: general LLMs vs clinically-oriented LLMs**
4. Study: Foundational Capabilities of LLMs in medicine
5. Some resources for your final project
6. Coding session: SurgicalPLAN

The large language model (LLM) era

What is the big deal of LLMs?

- Words, language and texts are complicated.
- Why?
  - ➢ Context is important

    *"**Charles** is giving a guest lecture today. **He** is excited."*

  - ➢ Polysemy: Same word can have different meanings

    " River **bank** ≠  Financial **bank**"

- Modeling them is complicated

The large language model (LLM) era

What is the big deal of LLMs?

- Before LLMs, most models treated words independently.
- This all changed with LLMs:
  - 🎯 Attention architecture
  - 🌐 Training models on the texts from the internet

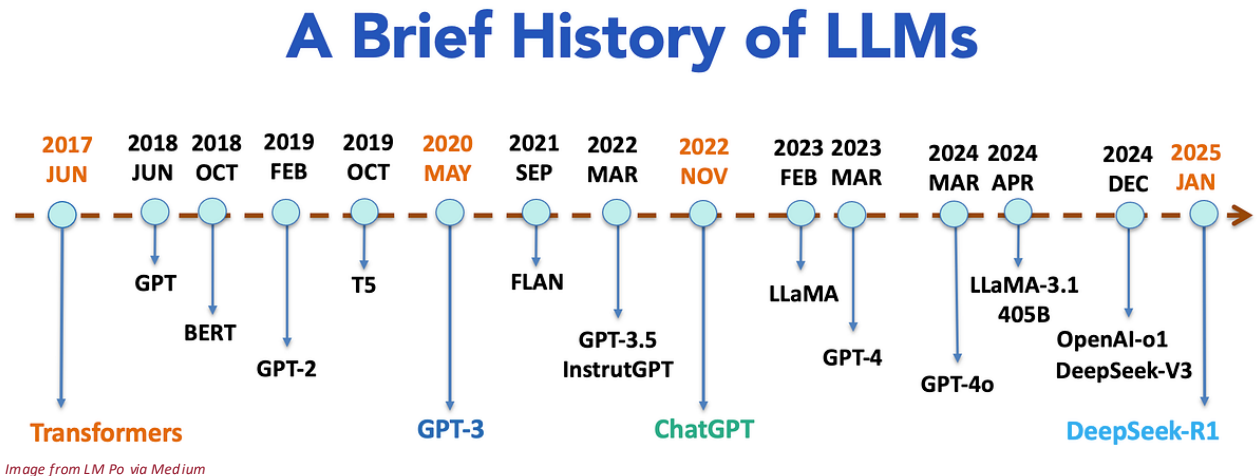


Image from LM Po via Medium

Attention in transformers

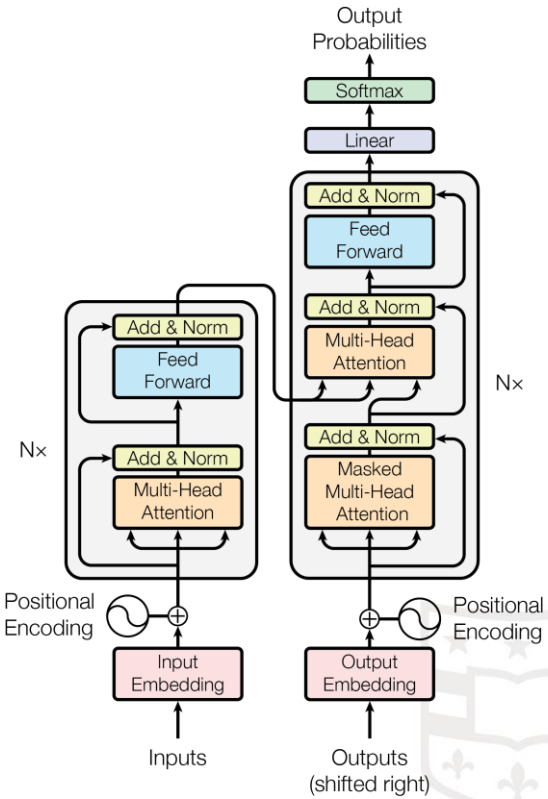


Image from LM Po via Vaswani et al.

The large language model (LLM) era

How do these models learn?

Decoder Pretraining (GPT)

- Language models like GPT and LLaMA are generative/causal language models
- No annotation needed
- Generate a sequence **x** by chain rule of **next-word prediction** from left to right

$$p(x) = \prod_{i=1}^{n} p(x_i | x_{<i})$$

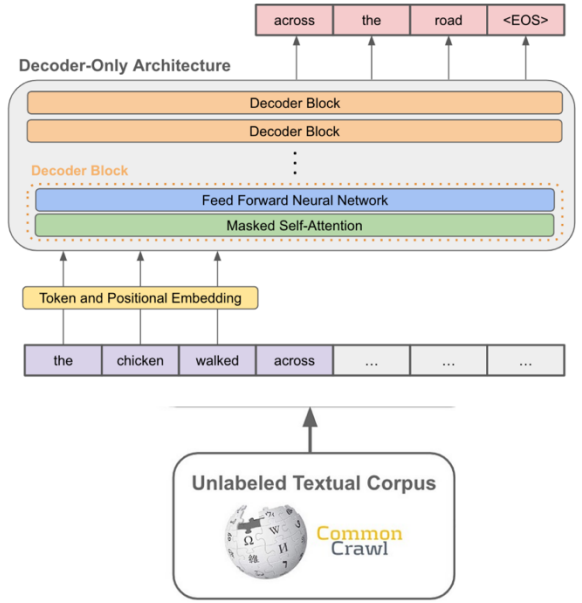
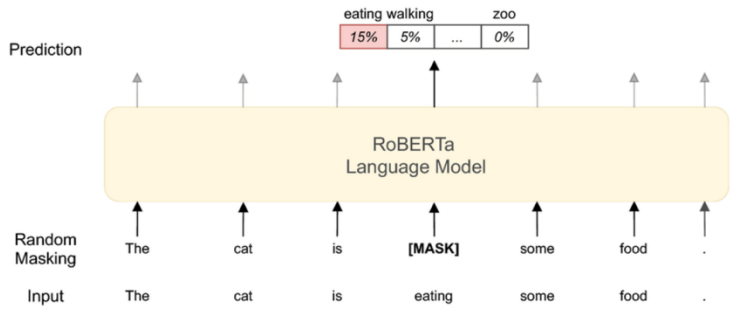


Image from Huang

Encoder Pre-training

- Pre-training objectives
  - Masked language modeling (with bidirectional attention) + Next Sentence Prediction
  - 15% of tokens are randomly corrupted (masked) for model prediction




now lets
linical
LMs
Not so
fast!

Challenges (pt 1)

The syntax of clinical notes differs from spoken language

- Language models are pre-trained on text from the internet.
- Language of the web ≠ syntax of clinical notes
- Clinical notes contains distinct jargon, medical abbreviations and clinical terminologies



Challenges (pt 1)

The syntax of clinical notes differs from spoken language

Examples:

- ‘pt’ → ‘patient’ 
  (not “Physical Therapy”  or “Point” )
- ‘PE’ → Pulmonary Embolism
  (not “Private Equity” or “Physical Education”)



Challenges (pt 2)

LLMs in clinical care is a low-resource domain

- Clinical notes are not readily open-sourced or abundant.
- Due to privacy concerns (e.g., HIPPA)
- Makes adapting these language models to clinical domains challenging



Image from IT Governance USA

Generic text can be ***readily*** trained from internet data.
This is something that ***could not be easily*** done for clinical notes.

Aren’t language models strong enough to contextualize to medical corpora?



Diving deep into language models

Case #1

Clinical Note
"***Pt** has **hx** of **PE**"*

What happens inside the language model?

Tokenization → ["pt", "has", "hx", "of", "pe"]

Mis-contextualizes 'Pt' as 'Point" or "Physical Therapist"

Same for "hx" and "pe"

Too much contextual mis-understanding
→ Mis-contextualizes entire note

Diving deep into language models

Case #2

Clinical Note
"***Hydrochlorothiazide*** *25 mg PO daily for HTN*"

What happens inside the language model?

Terms like Hydrochlorothiazide are long and unrecognized

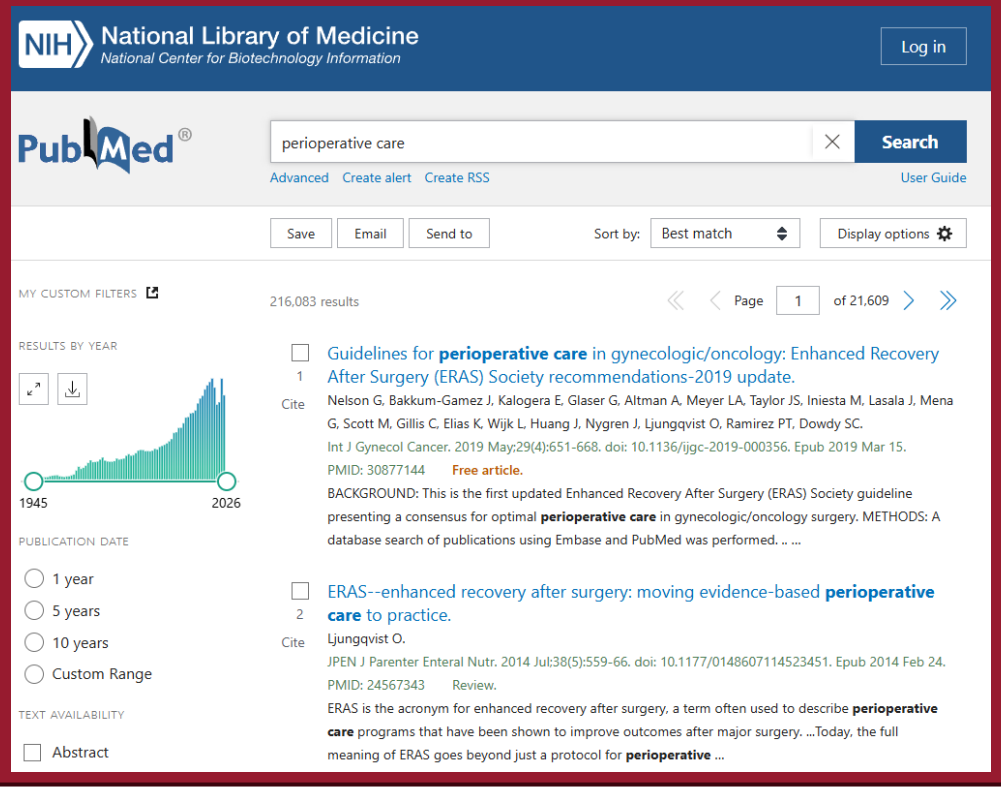
Tokenized into sub-words
→ ["*Hyd*", "*ro*", "*chlor*", "*oth*", "*iaz*", "*ide*"]

Miscontextualizes "*chlor*" associated with "*chlorine*" in chemistry

Existing solutions

Training on adjacent corpora

PubMed



Medical Information Mart for Intensive Care (MIMIC-III)

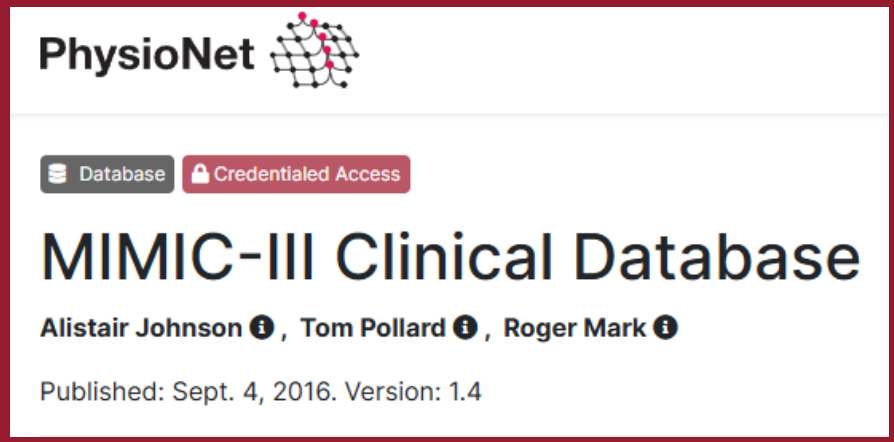




Image from Wikipedia

Existing solutions

Training on adjacent corpora

(MIMIC-III)

Exposing the model to clinical text, terminologies, and abbreviations!

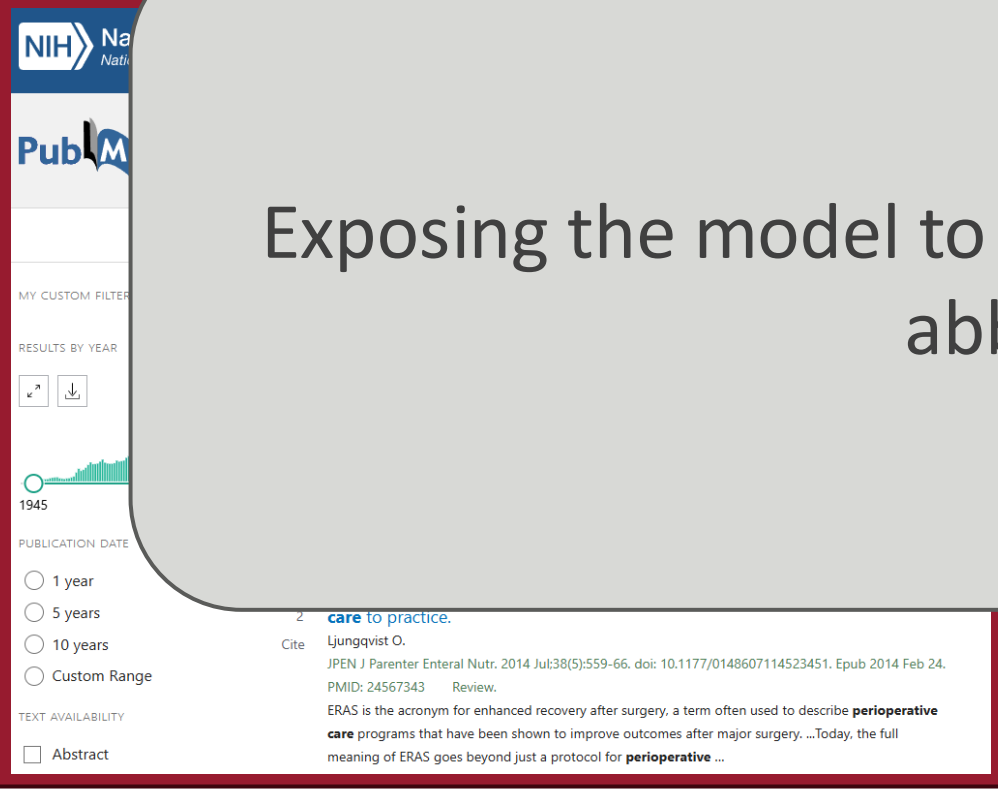
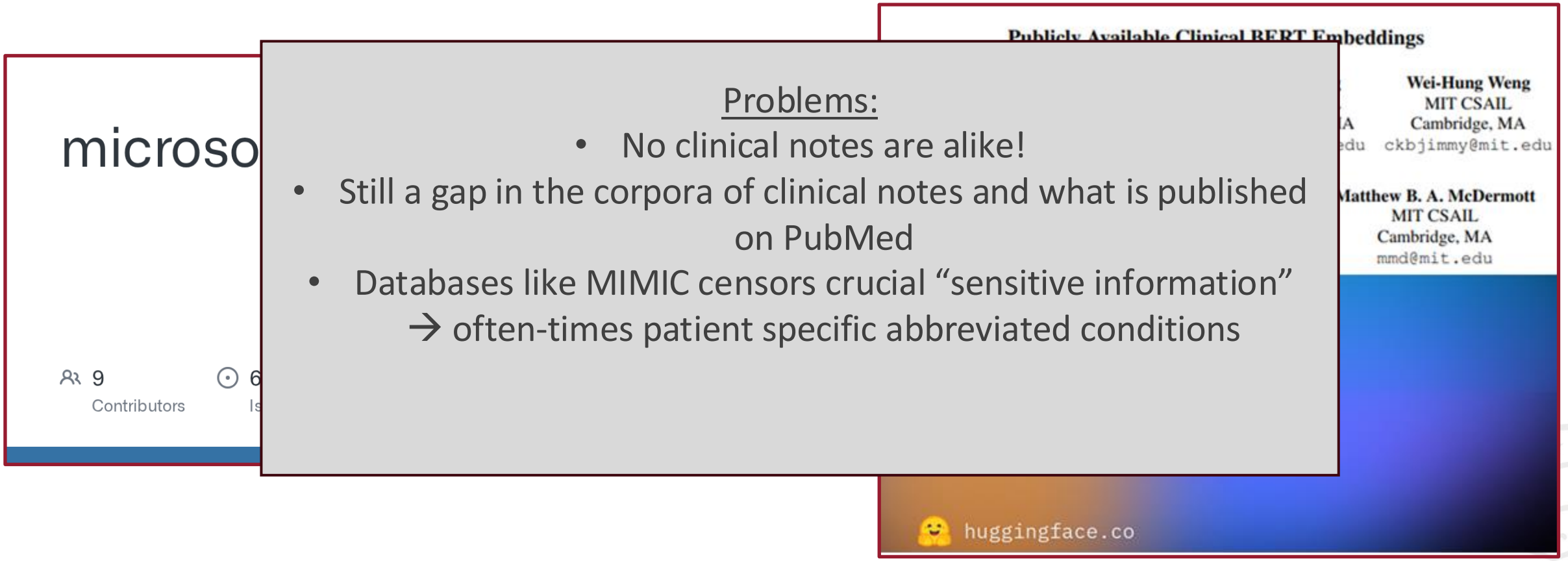




Image from Wikipedia



Problems:

- No clinical notes are alike!
- Still a gap in the corpora of clinical notes and what is published on PubMed
- Databases like MIMIC censors crucial “sensitive information”

→ often-times patient specific abbreviated conditions

Contents

1. Why are clinical notes important?
2. Why use large language models in peri-operative care?
3. Bridging the gap: general LLMs vs clinically-oriented LLMs
4. **Study: Foundational Capabilities of LLMs in medicine**
5. Some resources for your final project
6. Coding session: SurgicalPLAN

npj | digital medicine

Article | Open access | Published: 11 February 2025

The foundational capabilities of large language models in predicting postoperative risks using clinical notes

Charles Alba, Bing Xue, Joanna Abraham, Thomas Kannampallil & Chenyang Lu

npj Digital Medicine **8**, Article number: 95 (2025) | Cite this article

9430 Accesses | **8** Citations | **62** Altmetric | Metrics

Abstract

Clinical notes recorded during a patient's perioperative journey holds immense informational value. Advances in large language models (LLMs) offer opportunities for bridging this gap. Using 84,875 preoperative notes and its associated surgical cases from 2018 to 2021, we examine the performance of LLMs in predicting six postoperative risks using various fine-tuning strategies. Pretrained LLMs outperformed traditional word embeddings by an absolute AUROC of 38.3% and AUPRC of 33.2%. Self-supervised fine-tuning further improved

Download PDF

Sections | Figures | References

Abstract

Introduction

Results

Discussion

Methods

Data availability

Code availability

References

Acknowledgements

Author information

Work with Drs. Bing Xue, Thomas Kannampallil (Medicine) , Joanna Abraham (Medicine), Chengyang Lu

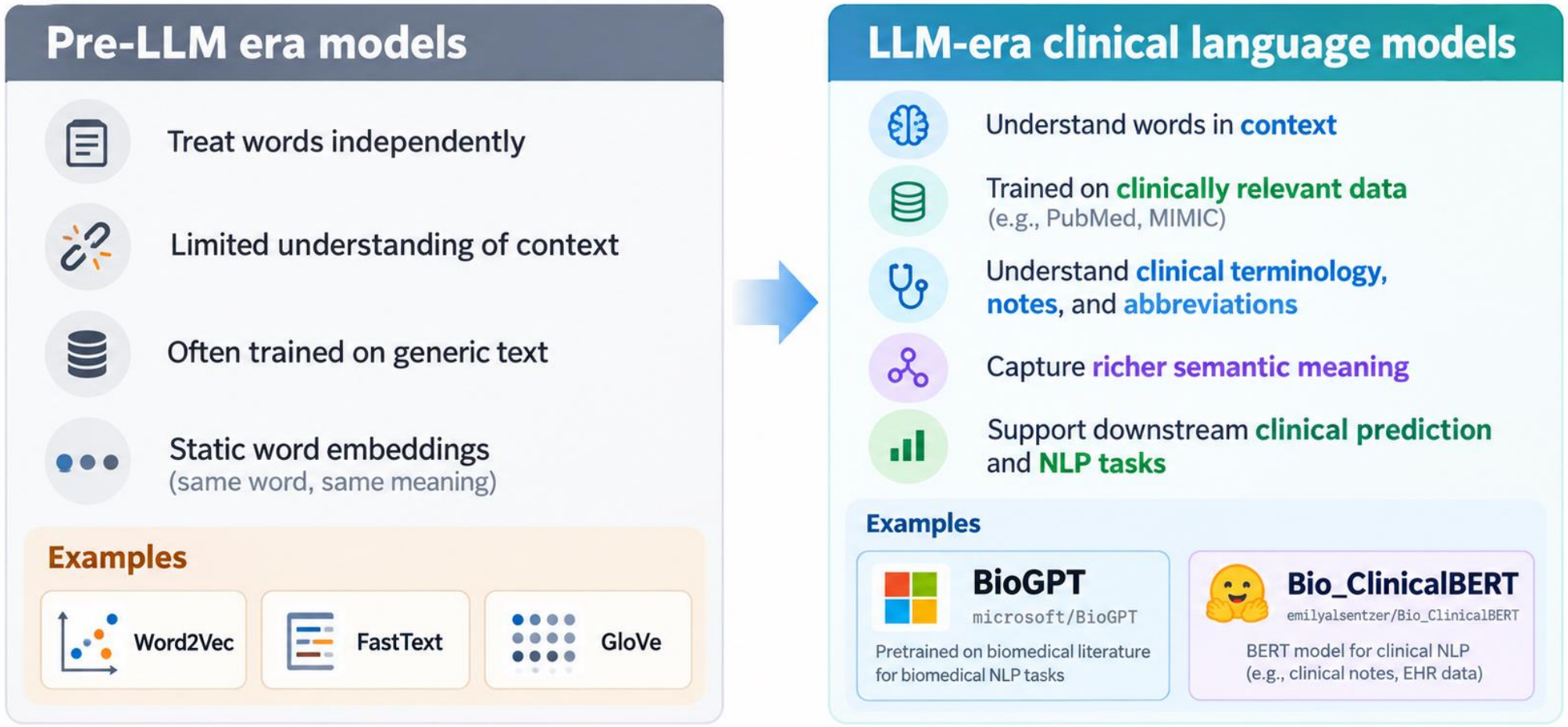
Settings and design

- 84,875 historical clinical notes and outcomes from Barnes Jewish Hospital
- From the pre-operative clinical notes, can it predict the following adverse post-surgical outcomes:
  - Death 30-days after getting discharge.
  - Pneumonia
  - Delirium
  - Acute Kidney Injury
  - Deep Vein Thrombosis
  - Pulmonary Embolism



Image from Barnes Jewish

Do Clinical LLMs really help?

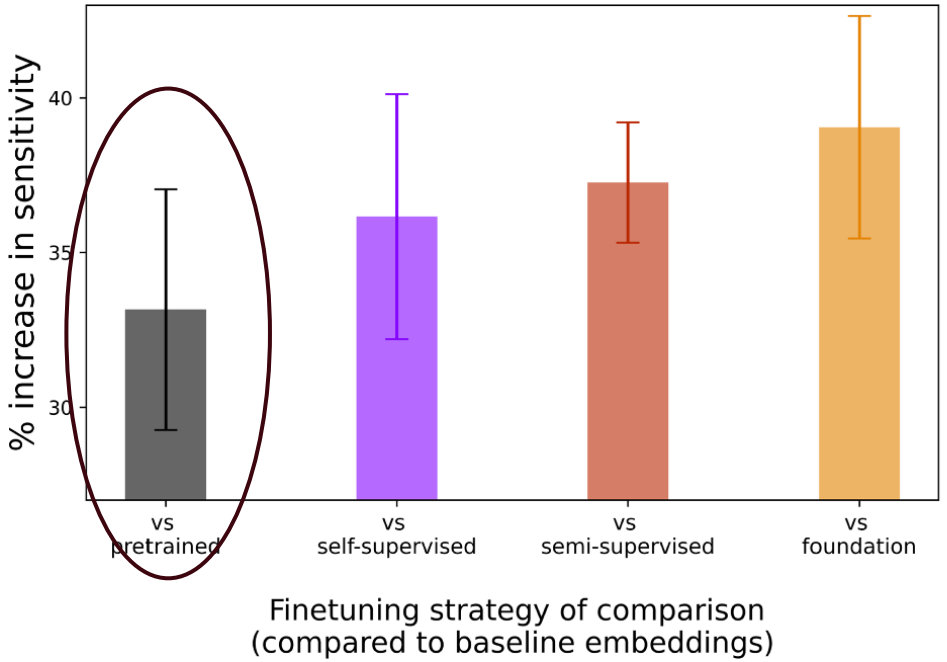


Model	Death in 30 days		PE		Pneumonia		DVT		AKI		Delirium	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
cbow	Mean: 0.528 CI: (0.409, 0.648)	Mean: 0.023 CI: (0.015, 0.031)	Mean: 0.506 CI: (0.418, 0.593)	Mean: 0.004 CI: (0.002, 0.006)	Mean: 0.526 CI: (0.384, 0.668)	Mean: 0.009 CI: (0.001, 0.016)	Mean: 0.524 CI: (0.457, 0.59)	Mean: 0.006 CI: (0.005, 0.007)	Mean: 0.56 CI: (0.488, 0.632)	Mean: 0.156 CI: (0.125, 0.187)	Mean: 0.501 CI: (0.464, 0.539)	Mean: 0.474 CI: (0.441, 0.507)
Doc2vec	Mean: 0.479 CI: (0.348, 0.611)	Mean: 0.021 CI: (0.012, 0.03)	Mean: 0.517 CI: (0.466, 0.567)	Mean: 0.004 CI: (0.004, 0.004)	Mean: 0.495 CI: (0.347, 0.643)	Mean: 0.006 CI: (0.003, 0.01)	Mean: 0.531 CI: (0.443, 0.619)	Mean: 0.007 CI: (0.004, 0.01)	Mean: 0.523 CI: (0.421, 0.624)	Mean: 0.146 CI: (0.092, 0.199)	Mean: 0.484 CI: (0.439, 0.53)	Mean: 0.466 CI: (0.417, 0.514)
fastText	Mean: 0.725 CI: (0.67, 0.781)	Mean: 0.05 CI: (0.04, 0.06)	Mean: 0.652 CI: (0.602, 0.701)	Mean: 0.007 CI: (0.005, 0.01)	Mean: 0.696 CI: (0.643, 0.749)	Mean: 0.016 CI: (0.008, 0.024)	Mean: 0.694 CI: (0.642, 0.746)	Mean: 0.014 CI: (0.011, 0.017)	Mean: 0.726 CI: (0.702, 0.75)	Mean: 0.273 CI: (0.239, 0.307)	Mean: 0.565 CI: (0.541, 0.589)	Mean: 0.533 CI: (0.513, 0.554)
GloVe	Mean: 0.818 CI: (0.807, 0.83)	Mean: 0.128 CI: (0.118, 0.139)	Mean: 0.664 CI: (0.628, 0.701)	Mean: 0.01 CI: (0.007, 0.013)	Mean: 0.765 CI: (0.732, 0.799)	Mean: 0.04 CI:(0.017, 0.063)	Mean: 0.723 CI: (0.7, 0.745)	Mean: 0.019 CI: (0.013, 0.024)	Mean: 0.81 CI: (0.805, 0.815)	Mean: 0.441 CI: (0.43, 0.451)	Mean: 0.666 CI: (0.652, 0.681)	Mean: 0.636 CI: (0.613, 0.66)
bioClinicalBERT	Mean: 0.85 CI: (0.84, 0.861)	Mean: 0.156 CI: (0.138, 0.173)	Mean: 0.683 CI: (0.621, 0.745)	Mean: 0.008 CI: (0.006, 0.011)	Mean: 0.809 CI: (0.785, 0.833)	Mean: 0.043 CI: (0.027, 0.059)	Mean: 0.76 CI: (0.723, 0.796)	Mean: 0.02 CI: (014, 0.027)	Mean: 0.83 CI: (0.828, 0.831)	Mean: 0.469 CI: (0.457, 0.48)	Mean: 0.68 CI: (0.663, 0.697)	Mean: 0.653 CI: (0.626, 0.68)
bioGPT	Mean: 0.862 CI: (0.851, 0.872)	Mean: 0.161 CI: (0.141, 0.182)	Mean: 0.711 CI: (0.679, 0.743)	Mean: 0.011 CI: (0.005, 0.017)	Mean: 0.818 CI: (0.8, 0.837)	Mean: 0.047 CI: (0.037, 0.058)	Mean: 0.773 CI: (0.734, 0.813)	Mean: 0.024 CI: (0.016, 0.032)	Mean: 0.835 CI: (0.833, 0.838)	Mean: 0.478 CI: (0.465, 0.492)	Mean: 0.691 CI: (0.672, 0.71)	Mean: 0.664 CI: (0.638, 0.69)
ClinicalBERT	Mean: 0.855 CI: (0.842, 0.867)	Mean: 0.155 CI: (0.137, 0.173)	Mean: 0.717 CI: (0.691, 0.743)	Mean: 0.013 CI: (0.009, 0.017)	Mean: 0.806 CI: (0.784, 0.827)	Mean: 0.04 CI: (0.024, 0.056)	Mean: 0.764 CI: (0.73, 0.799)	Mean: 0.022 CI: (0.015, 0.03)	Mean: 0.83 CI: (0.827, 0.833)	Mean: 0.469 CI: (0.458, 0.48)	Mean: 0.686 CI: (0.671, 0.702)	Mean: 0.66 CI: (0.634, 0.686)

Results

Pretrained LLMs hold promise in predicting postoperative risks

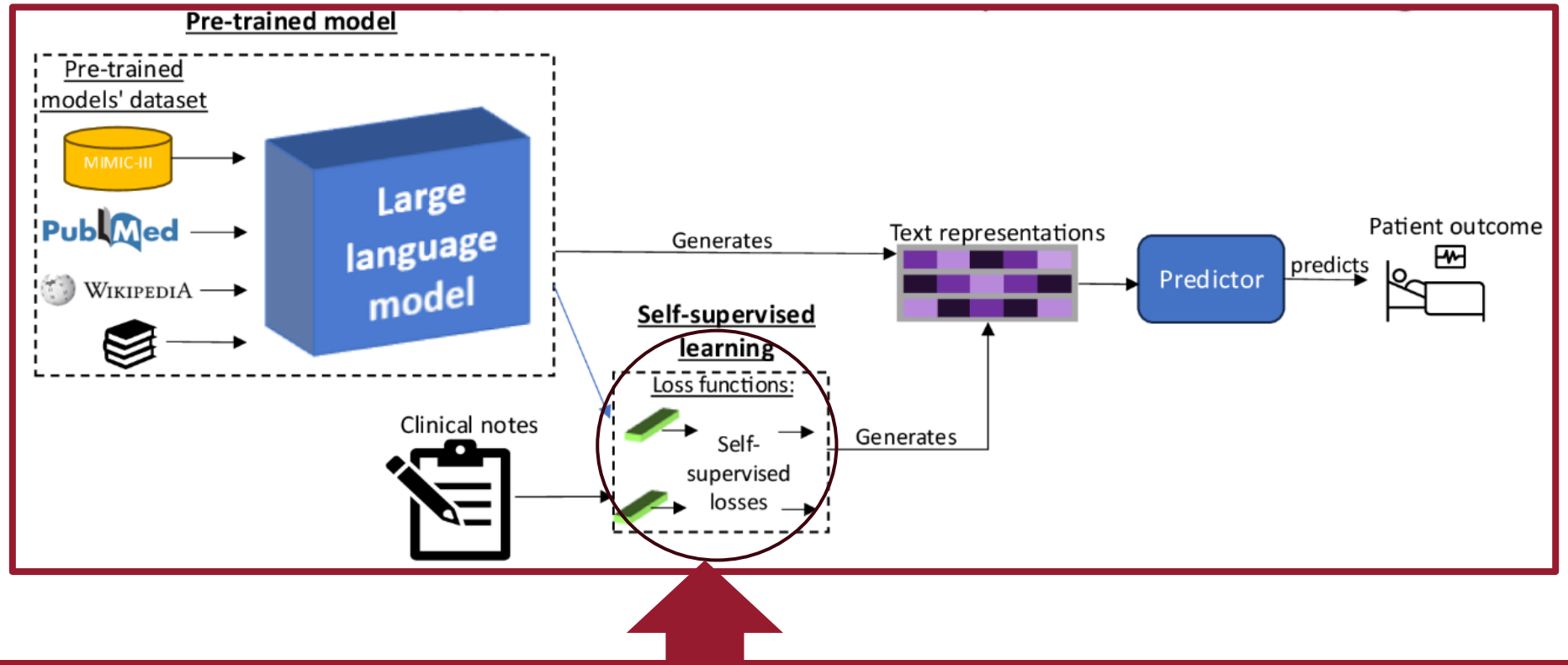
→ **33%** increase in sensitivity over traditional NLP methods



The number of high-risk patients correctly identified per 100 cases that were missed by baseline models.

Methods

Using pre-trained model vs self-supervised training



We are essentially taking these pretrained Clinical LLMs and “adapting” them to our very own notes.

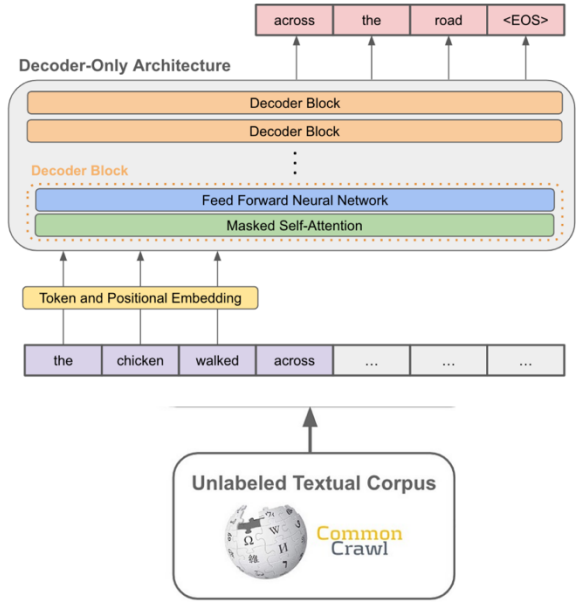
Self-supervised loses

Recap from a few slides above

Decoder Pretraining (GPT)

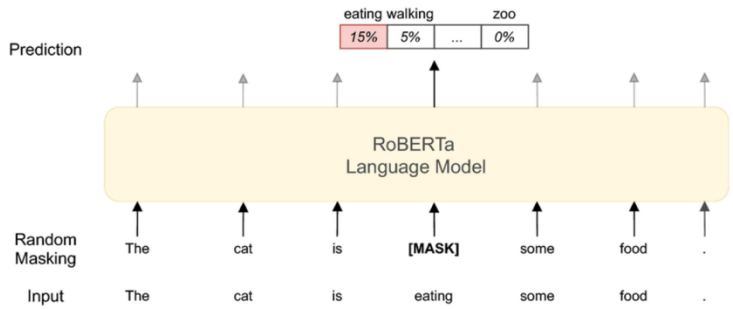
- Language models like GPT and LLaMA are generative/causal language models
- No annotation needed
- Generate a sequence **x** by chain rule of **next-word prediction** from left to right

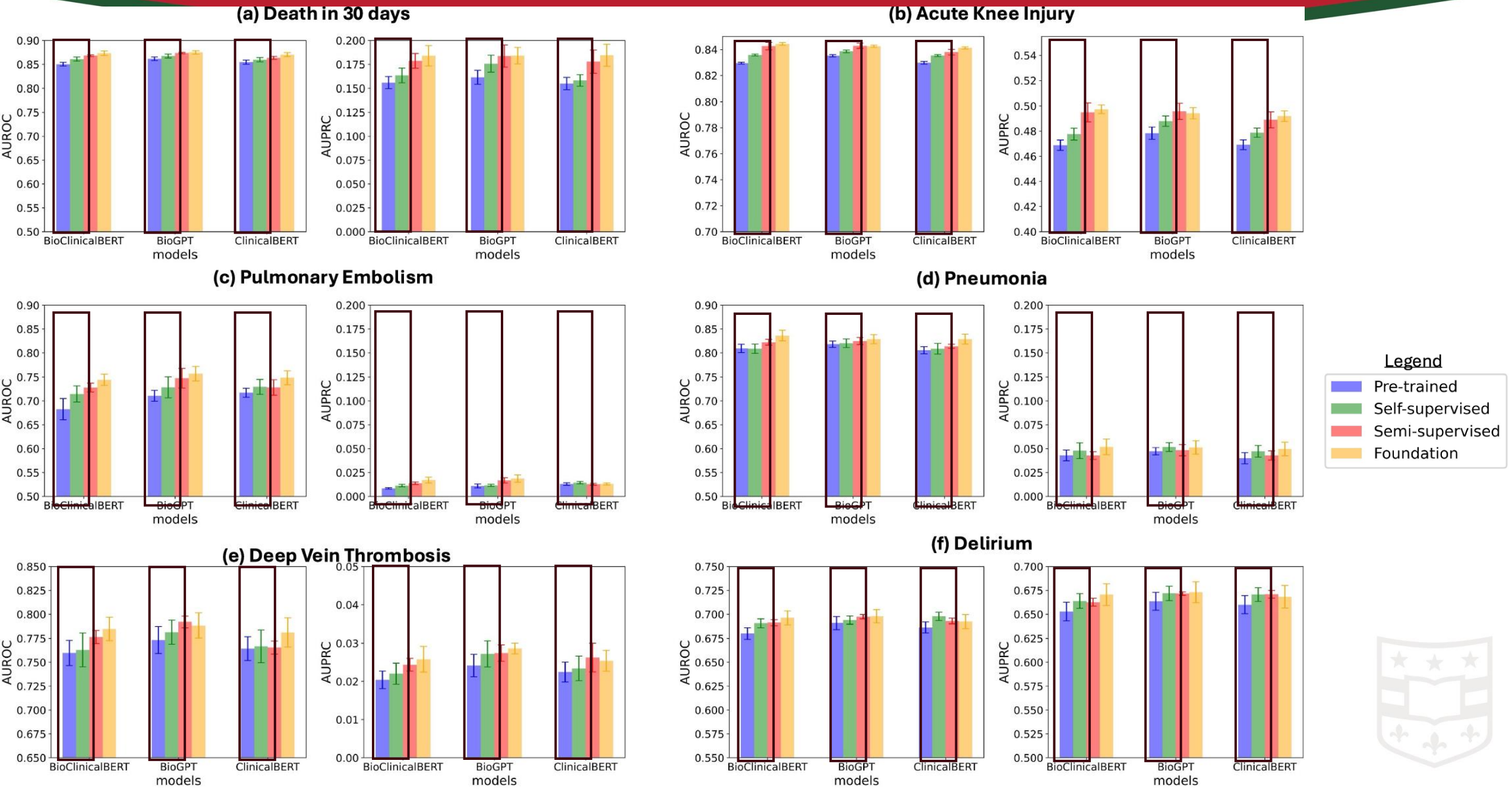
$$p(x) = \prod_{i=1}^{n} p(x_i|x_{<i})$$



Encoder Pre-training

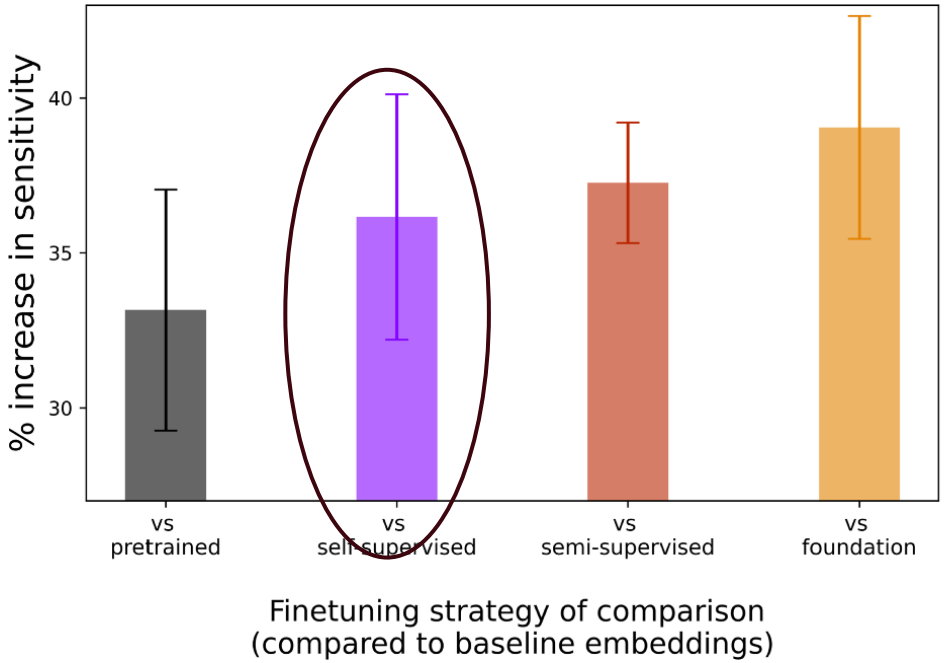
- Pre-training objectives
  - Masked language modeling (with bidirectional attention) + Next Sentence Prediction
  - 15% of tokens are randomly corrupted (masked) for model prediction




(a) Death in 30 days
(b) Acute Knee Injury
(c) Pulmonary Embolism
(d) Pneumonia
(e) Deep Vein Thrombosis
(f) Delirium
AUROC
AUPRC
BioClinicalBERT
BioGPT
ClinicalBERT
models
Legend
Pre-trained
Self-supervised
Semi-supervised
Foundation

Results

Increase in self-supervised training highlights gaps in corpora of pre-trained models and clinical notes.



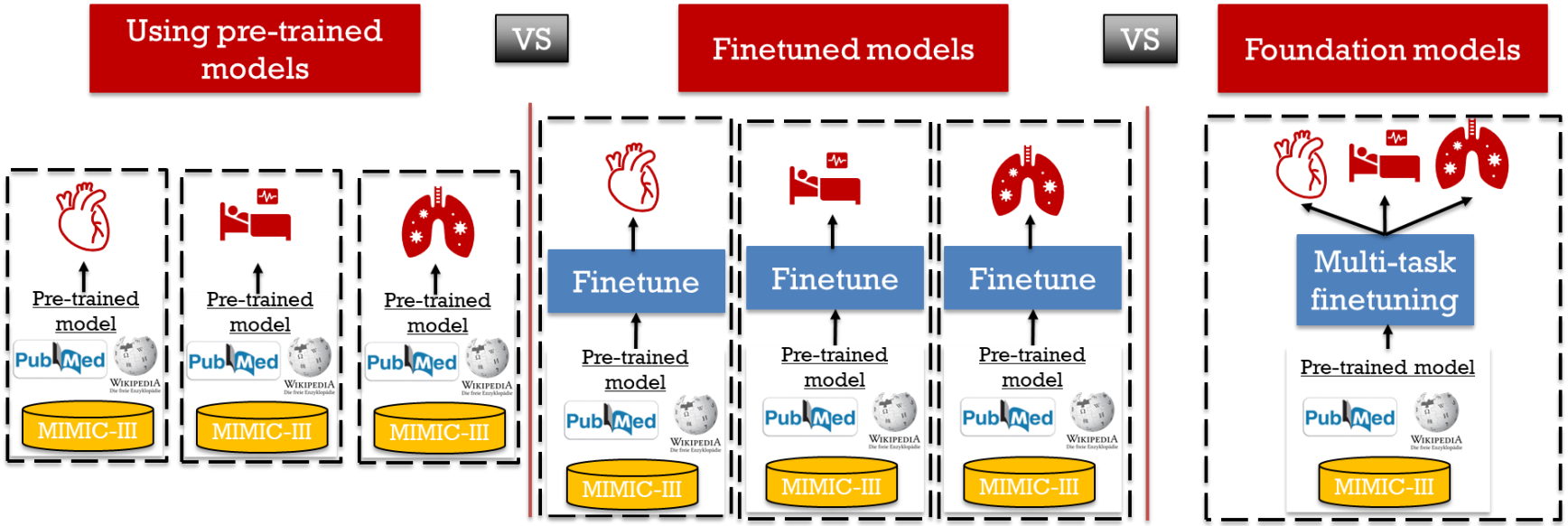
The number of high-risk patients correctly identified per 100 cases that were missed by baseline models.

Why?

- *Clinical LMs may have been exposed to clinical texts, abbreviations and terminologies.*
  - *But it may not be precisely identical to what is seen in our own clinical notes.*
- *We adapted these Clinical LLMs, 'familiarizing' them with the text of our clinical notes*

Methods

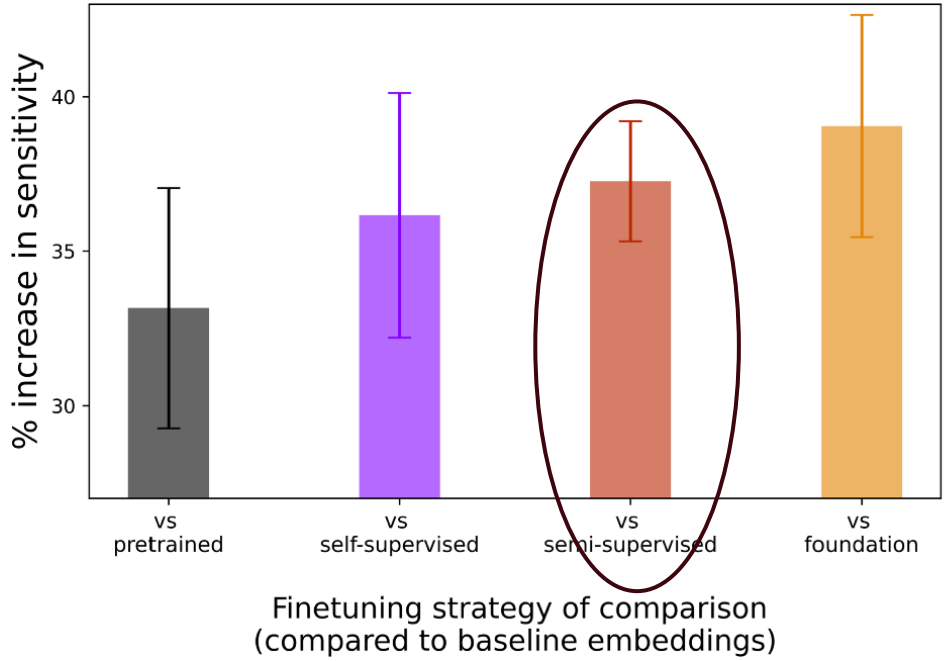
Foundation models



- ✓Foundation models via multi-task learning
- ✓Knowledge sharing across all outcomes
- ✓Unified model can predict multiple outcomes

Results

Finetuning with labels incorporated can further boost performances



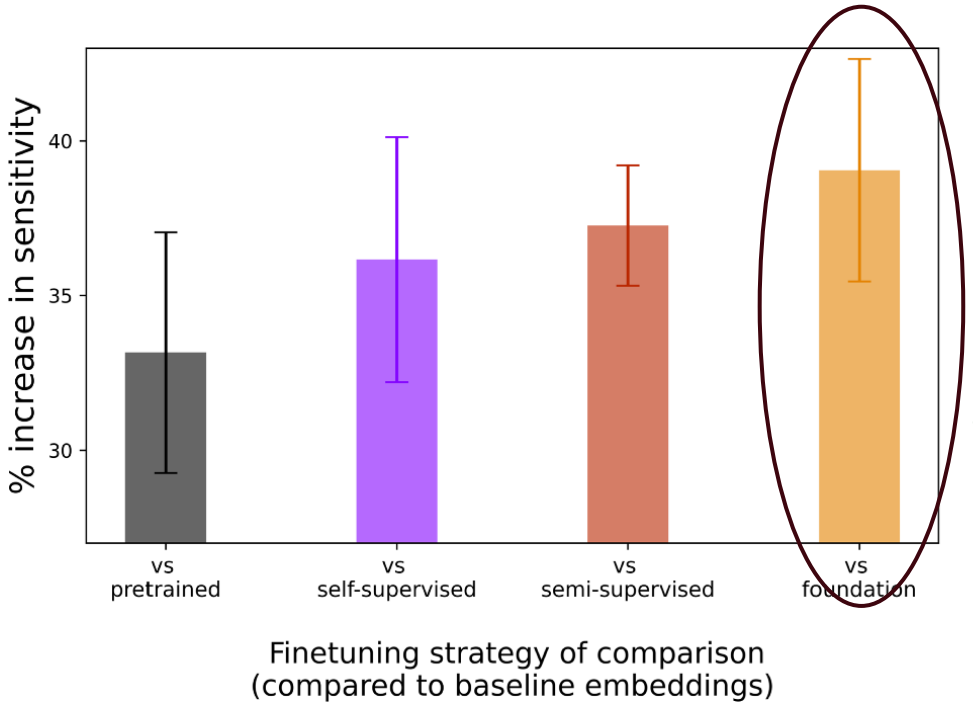
The number of high-risk patients correctly identified per 100 cases that were missed by baseline models.

Why?

Tokens linked to specific outcomes are effectively represented

Results

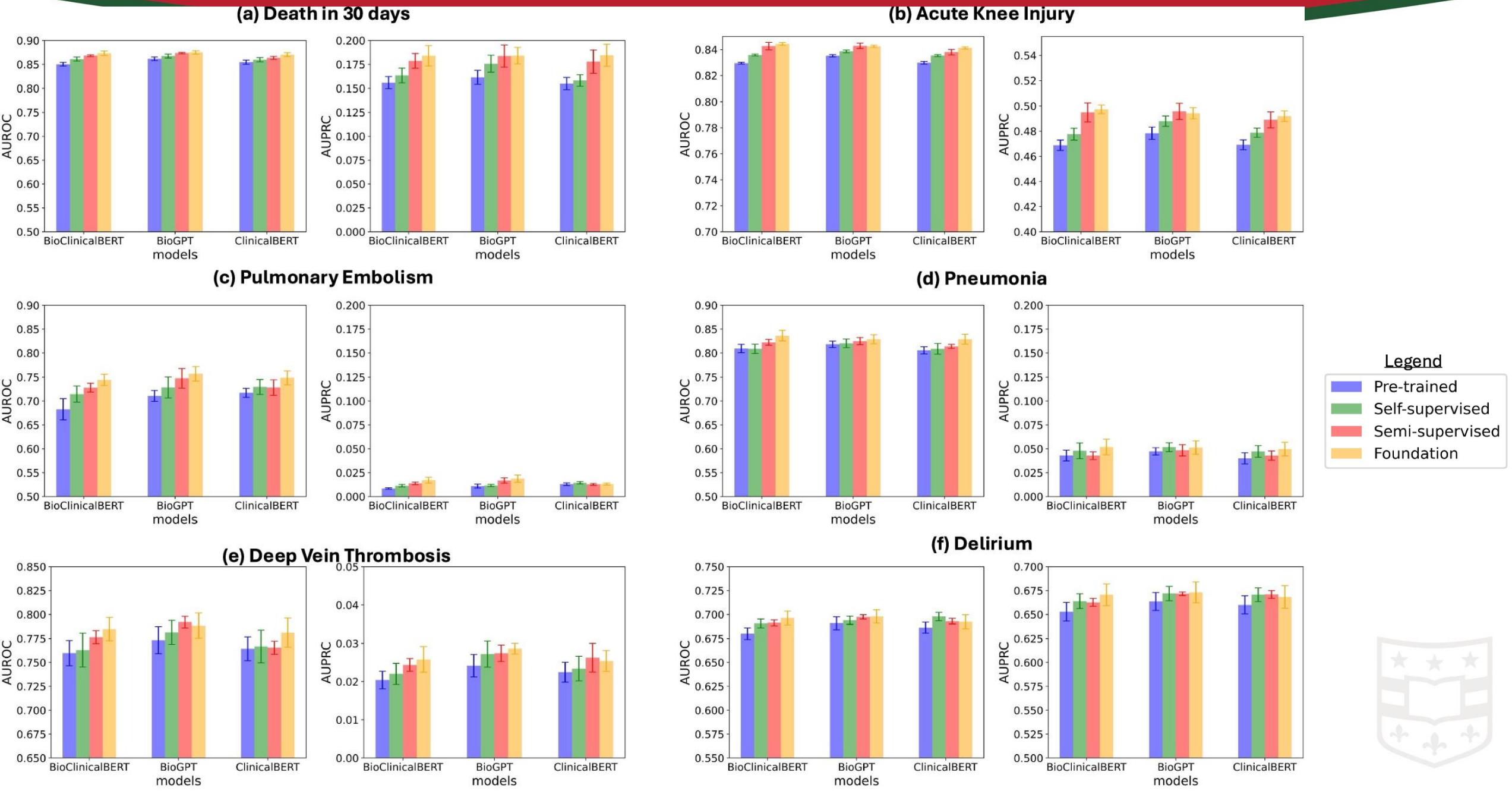
Unified foundation model showcase superior performances.
→ **39%** increase in sensitivity compared to baseline



The number of high-risk patients correctly identified per 100 cases that were missed by baseline models.

Why?

Knowledge sharing across outcomes via multi-task learning

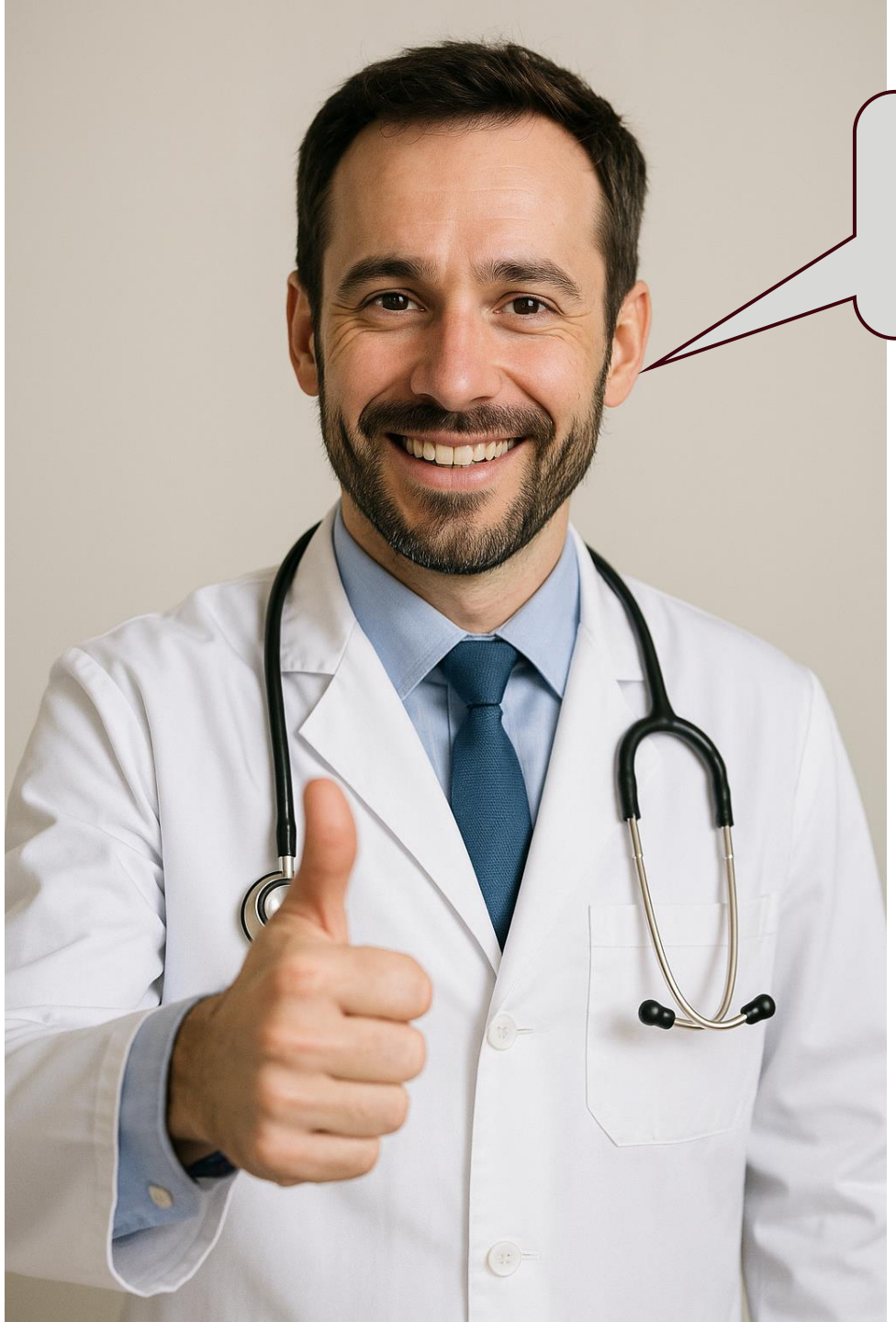

(a) Death in 30 days
(b) Acute Knee Injury
(c) Pulmonary Embolism
(d) Pneumonia
(e) Deep Vein Thrombosis
(f) Delirium
AUROC
AUPRC
BioClinicalBERT
BioGPT
ClinicalBERT
models
Legend
Pre-trained
Self-supervised
Semi-supervised
Foundation

Why foundation models?

Beyond performances, what are the other practical benefits if we were to ever embed this into clinical workflows?

- ✓Superior performances: generalize across outcomes
- ✓Versatility: A single model can predict multiple outcomes simultaneously.
- ✓Efficiency: Tuning one model for 6 outcomes is more efficient than tuning six models for each outcome




Predicting potential post-operative risks with better performances are great!

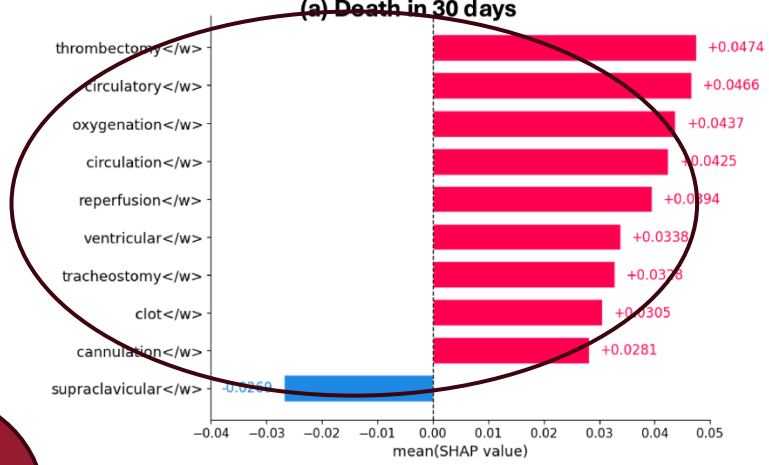

But how do I really know if your model has truly adapted to my clinical notes?

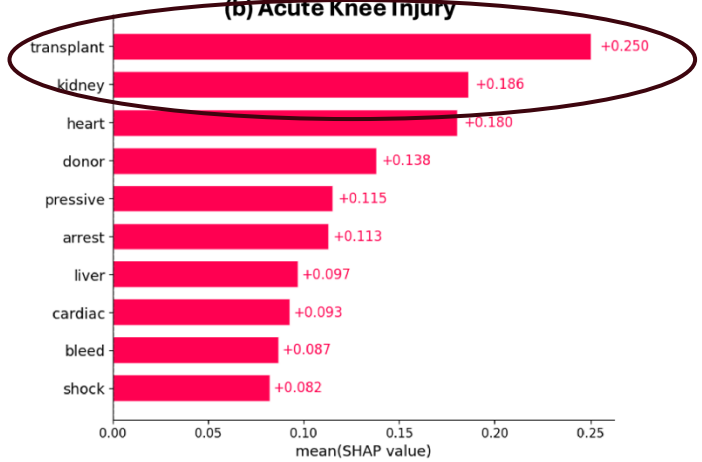
Model variant	Prompt	Base model	Our model
BioClinicalBERT	"[MASK] underwent surgery to remove tumor."	*She* underwent surgery to remove tumor.	Patient underwent surgery to remove tumor.
	"Patient complains about pain. Patient is [MASK]."	Patient complains about pain. Patient is *comfortable*.	Patient complains about pain. Patient is anxious.
	"Patient suffers from excessive [MASK]. Patient should be sent to ICU."	Patient suffers from excessive anxiety. Patient should be sent to ICU.	Patient suffers from excessive burn. Patient should be sent to ICU.
	"Patient's family has history of diabetes. Surgeon should watch out for [MASK]."	Patient's family has history of diabetes. Surgeon should watch out for diabetes.	Patient's family has history of diabetes. Surgeon should watch out for procedures.
BioGPT	"Patient suffers from excessive bleeding. Patient should be ... ..."	Patient suffers from excessive bleeding. Patient should be sent to a hospital for further investigation.	Patient suffers from excessive bleeding. Patient should be sent to a trauma room or a vascular access room
	"Patient complains of extreme pain. Patient is probably a ... ..."	Patient complains of extreme pain. Patient is probably a *good candidate for atrial of conservative treatment*	Patient complains of extreme pain. Patient is probably a case of lumbar
	"Patient's family has history of high blood pressure. Avoid ... ..."	Patient's family has history of high blood pressure. Avoid use of antihypertensive drugs.	Patient's family has history of high blood pressure. Avoid invasive diagnostic procedure if possible

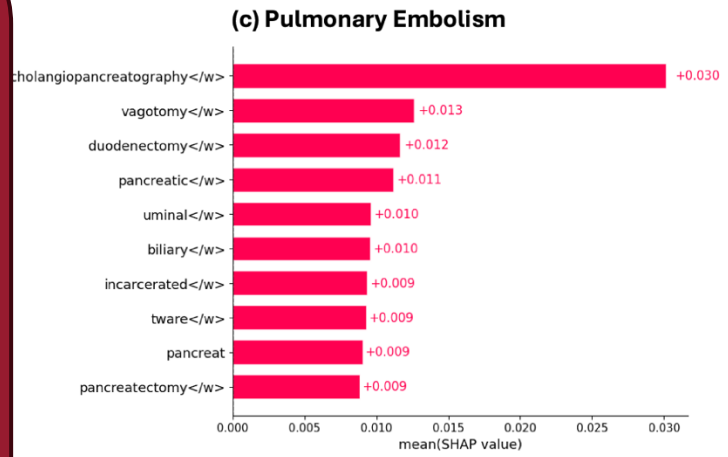
The prompts demonstrate the safety and adaptive nature of our model in perioperative care use. Outputs of the prompts are underlined and results deemed potentially harmful or biased are *italicized*.

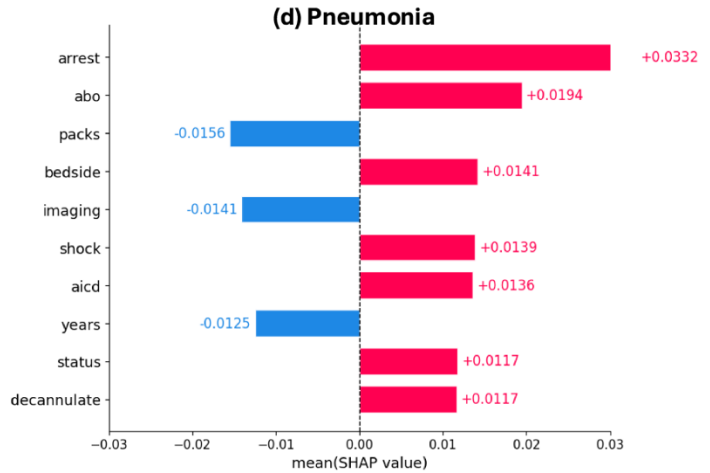

Okay, anticipating these risks are great, but how do I know which terms elicit certain predictions?

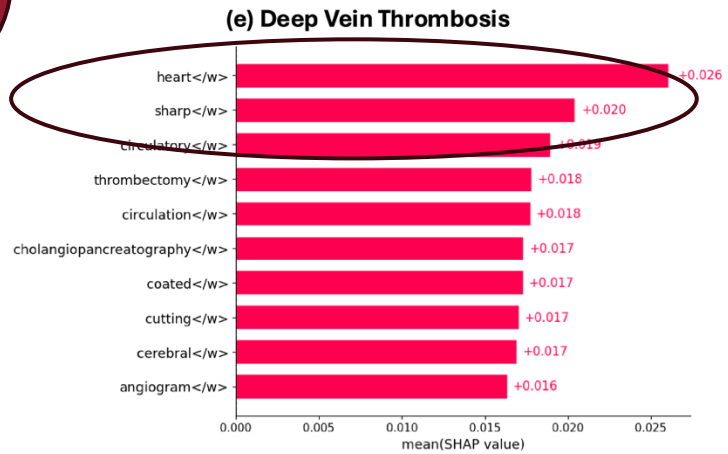
Keywords of procedures / organs commonly linked with associated risks!

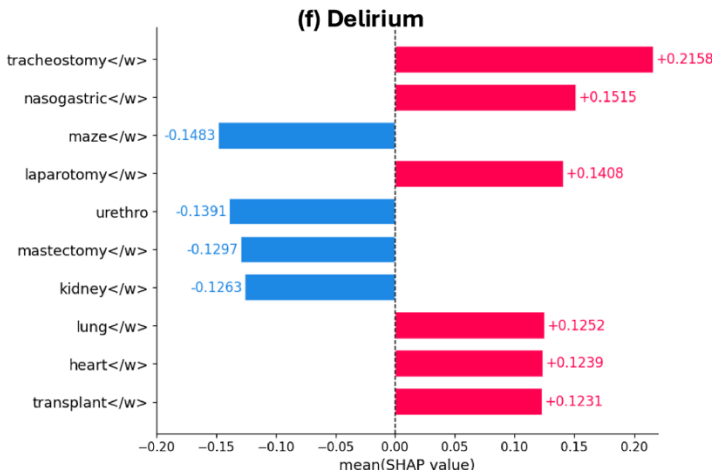












Limitations

Clinical limi

- Will physicians trade these found
  simpler tools (e.g., NSQIP ACS Ri
  they are accustomed to?

Share your thoughts, ideas and suggestions in the critique!

ological limitations

e with larger models (e.g., llama)?
performance improvement vs
itional resources

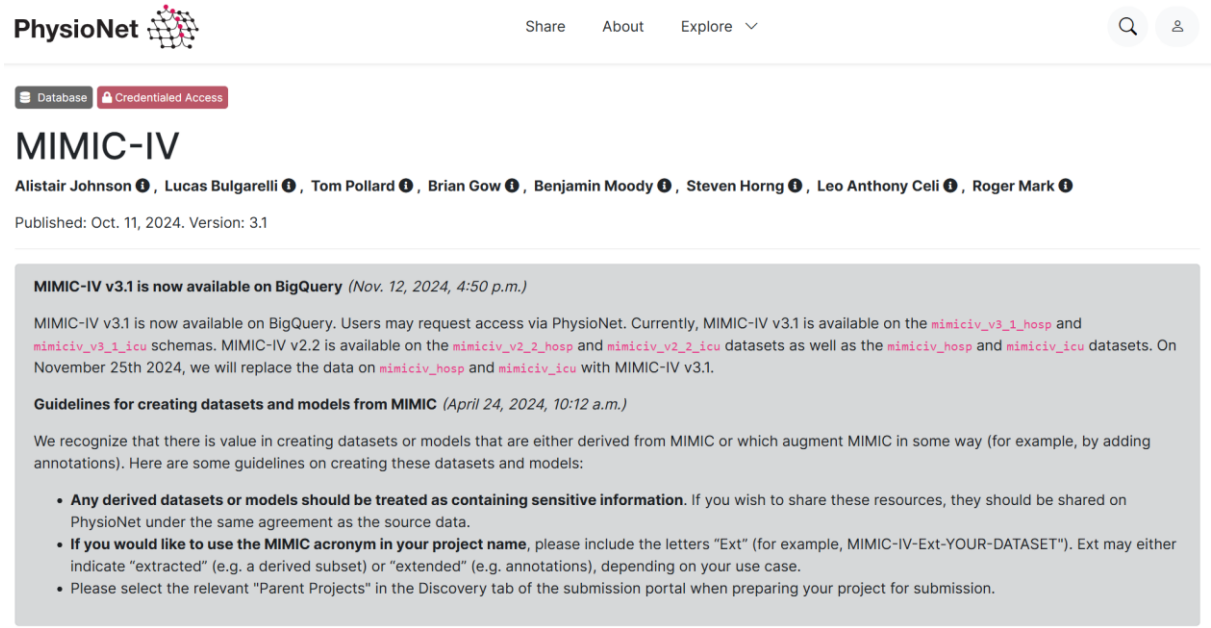
ubsets of certain populations make
it cha ging to measure fairness

Contents

1. Why are clinical notes important?
2. Why use large language models in peri-operative care?
3. Bridging the gap: general LLMs vs clinically-oriented LLMs
4. Study: Foundational Capabilities of LLMs in medicine
5. **Some resources for your final project**
6. Coding session: SurgicalPLAN

Datasets

- MIMIC-III: https://physionet.org/content/mimiciii/1.4/
- MIMIC-IV: https://physionet.org/content/mimiciv/3.1/



PhysioNet Share About Explore

Database Credentialed Access

MIMIC-IV

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, Roger Mark

Published: Oct. 11, 2024. Version: 3.1

MIMIC-IV v3.1 is now available on BigQuery *(Nov. 12, 2024, 4:50 p.m.)*

MIMIC-IV v3.1 is now available on BigQuery. Users may request access via PhysioNet. Currently, MIMIC-IV v3.1 is available on the `mimiciv_v3_1_hosp` and `mimiciv_v3_1_icu` schemas. MIMIC-IV v2.2 is available on the `mimiciv_v2_2_hosp` and `mimiciv_v2_2_icu` datasets as well as the `mimiciv_hosp` and `mimiciv_icu` datasets. On November 25th 2024, we will replace the data on `mimiciv_hosp` and `mimiciv_icu` with MIMIC-IV v3.1.

Guidelines for creating datasets and models from MIMIC *(April 24, 2024, 10:12 a.m.)*

We recognize that there is value in creating datasets or models that are either derived from MIMIC or which augment MIMIC in some way (for example, by adding annotations). Here are some guidelines on creating these datasets and models:

- **Any derived datasets or models should be treated as containing sensitive information**. If you wish to share these resources, they should be shared on PhysioNet under the same agreement as the source data.
- **If you would like to use the MIMIC acronym in your project name**, please include the letters "Ext" (for example, MIMIC-IV-Ext-YOUR-DATASET"). Ext may either indicate "extracted" (e.g. a derived subset) or "extended" (e.g. annotations), depending on your use case.
- Please select the relevant "Parent Projects" in the Discovery tab of the submission portal when preparing your project for submission.

Packages

These are must-have packages for any LLM project!

- Pytorch (*must have!*)
- Transformers (*must have!*)
- Datasets (*must have!*)
- Tokenizers
- huggingface_hub



Image from Pytorch.com

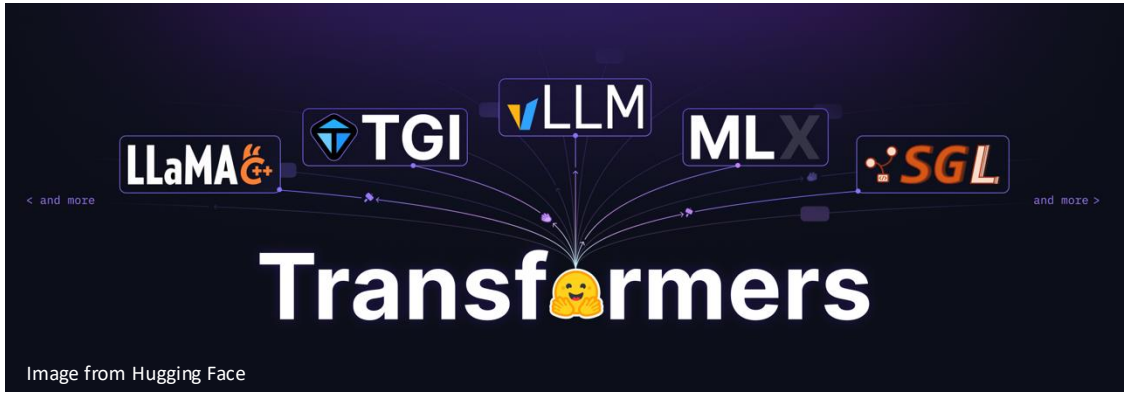


Image from Hugging Face

Sites

Where can I find open-sourced models?

- Huggingface: https://huggingface.co/
- Ollama: https://ollama.com/

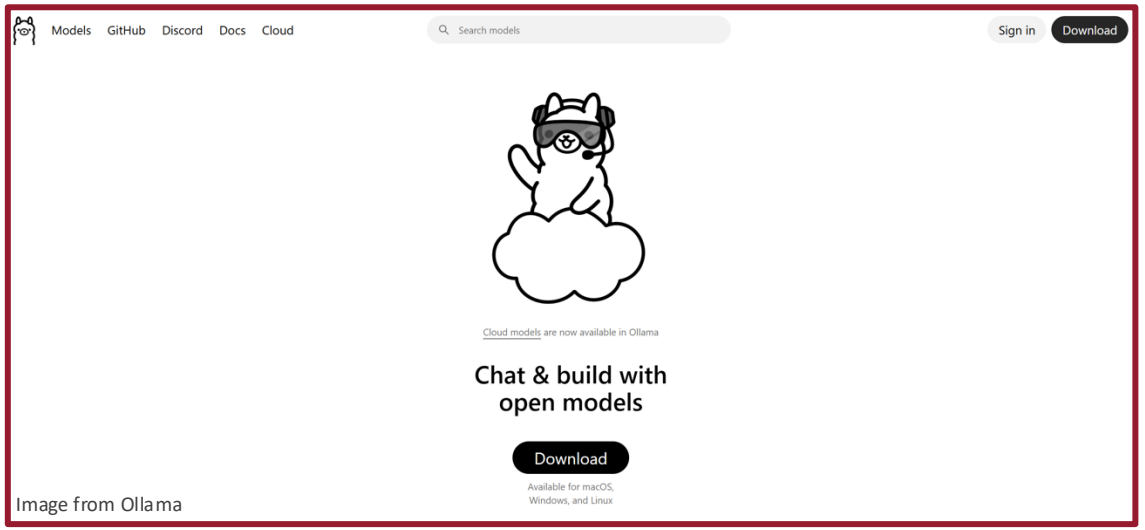


Image from Ollama

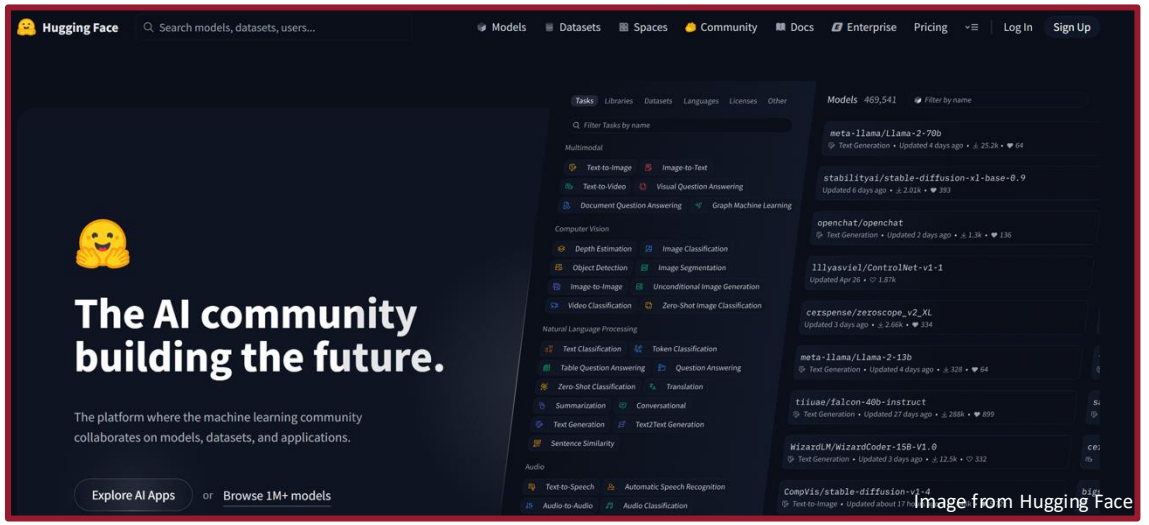
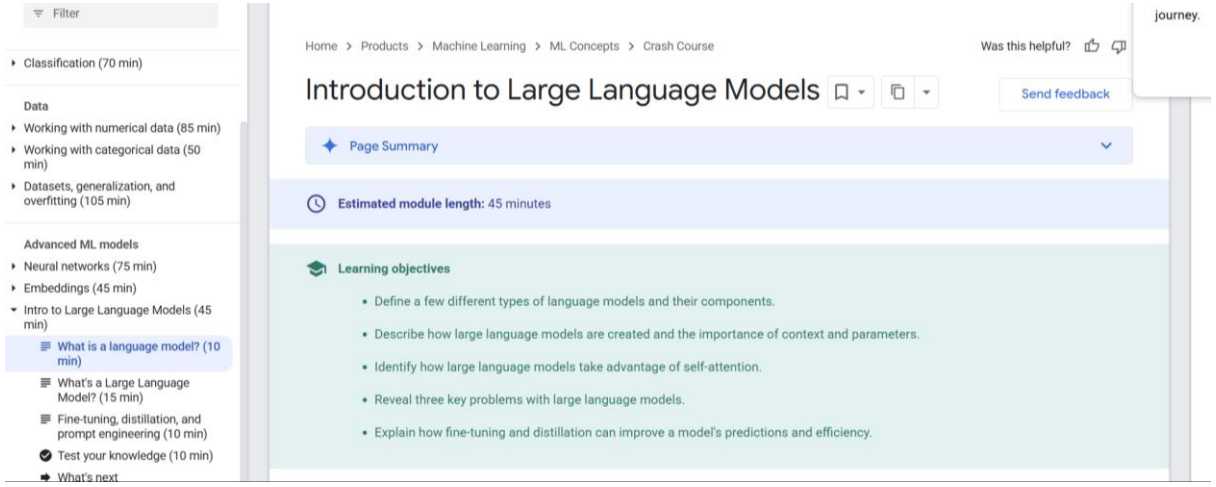


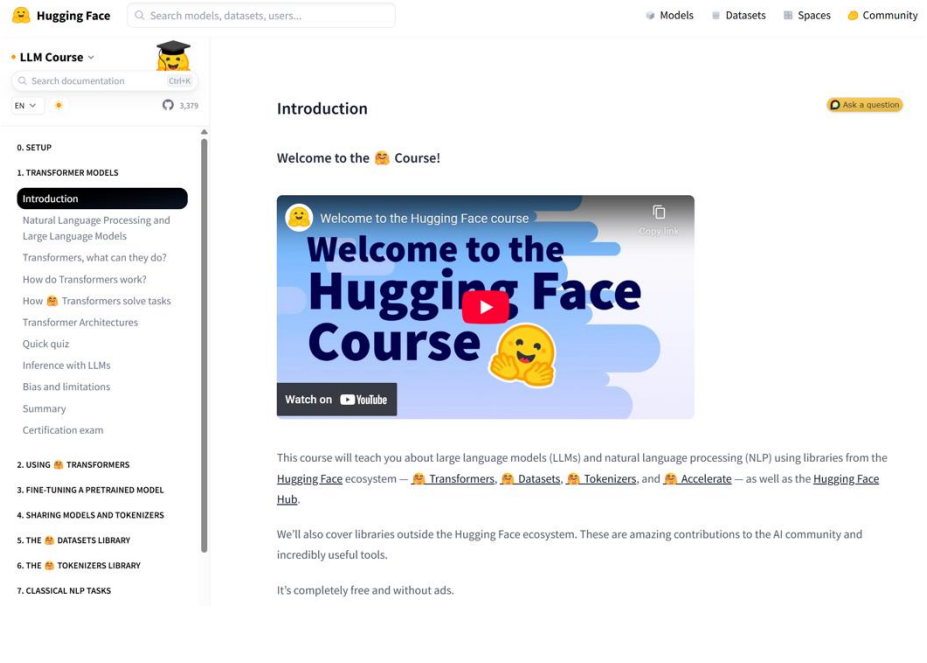
Image from Hugging Face

Tutorials

I am new to LLMs. Where can I learn about LLMs?

- HF “Intro to LLM course” 🤗 : https://huggingface.co/learn/llm-course/
- Google's ML Crash Course 💻: https://developers.google.com/machine-learning/crash-course/llm



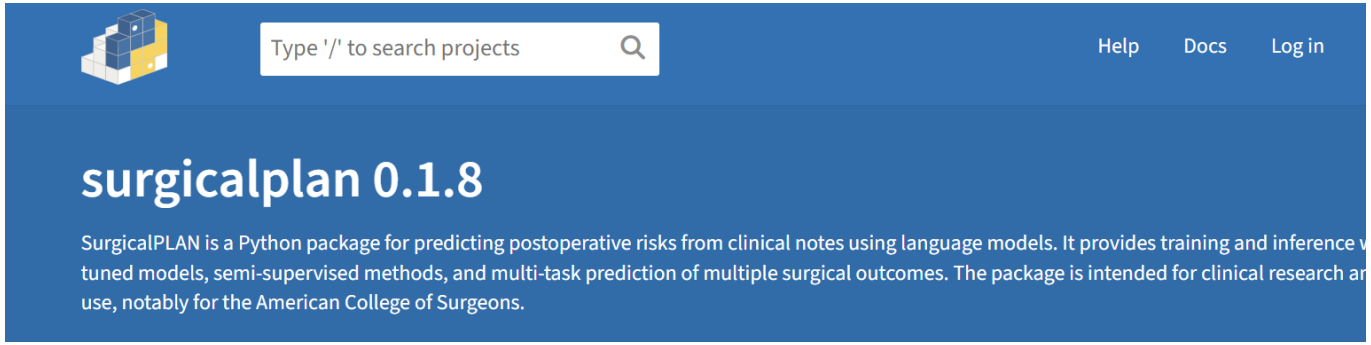


Tutorials

Want use the same methods presented here in your project?

- *SurgicalPLAN* package:
  - Website & documentation: https://websites.washu.edu/surgicalplan/
  - PyPI package: https://pypi.org/project/surgicalplan/





Contents

1. Why are clinical notes important?
2. Why use large language models in peri-operative care?
3. Bridging the gap: general LLMs vs clinically-oriented LLMs
4. Study: Foundational Capabilities of LLMs in medicine
5. Are there a more efficient ways to train LLMs?
6. Some resources for your final project
7. **Coding session: SurgicalPLAN**

Go to this link: **https://shorturl.at/oDhC7**

ACS_demo.ipynb

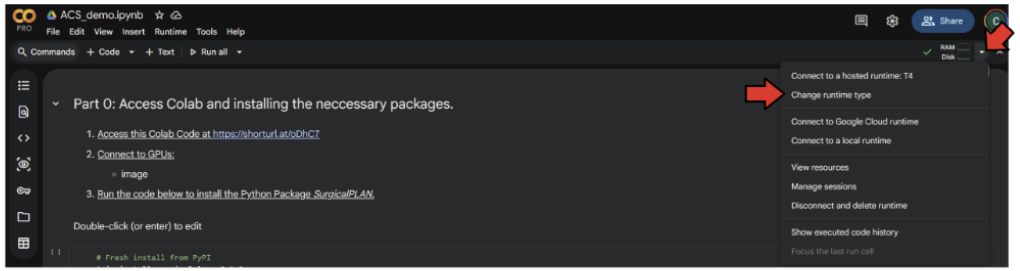
File Edit View Insert Runtime Tools Help

Commands + Code + Text Run all Connect A100

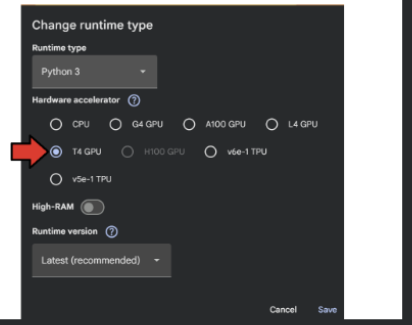
Part 0: Access Colab and installing the neccessary packages.

1. Access this Colab Code at https://shorturl.at/oDhC7
2. Connect to GPUs:

Step 1: Click on "▼" at the upper-right and select "Change runtime type"



Step 2: Select "T4 GPU" and click "Save"



3. Run the code below to install the Python Package *SurgicalPLAN.*

Thank you!



Questions?

