



AI and Large Language Models at NYU Langone

John Speakman, AVP Research and Education
Information Technology

HPC User Forum
April 10, 2024

Digital  Health



Confession: I am neither of these people

**Yindalon
Aphinyanaphongs, MD,
PhD**

Assistant Professor, Department of Population Health
Assistant Professor, Department of Medicine



NYU Langone Provider

Eric K. Oermann, MD

Specialties: Neurosurgery, Spine Surgery

Treats: Adults

Language: English

Phone: 212-263-2480

[Schedule Appointment](#)



Agenda

1. Who are we?
2. Generative AI in Medicine
3. AI at NYU Langone Health
4. NYUTron: our generative AI LLM
5. HPC Support



NYU Langone Health's Mission

Our trifold mission to **serve**, **teach**, and **discover** is achieved daily through an integrated academic culture devoted to excellence in patient care, education, and research.



NYU Langone by the Numbers

105,973
Hospital
Discharges

2,073
Beds

4
Emergency
Departments

6
Inpatient
Locations

191
Operating
Rooms

8.37M
Patients in Our
Electronic Health Record

14,393
Births

\$12B
In Hospital
Revenue

\$4.9B
In Philanthropy
Efforts

13M SF+
Space
Portfolio

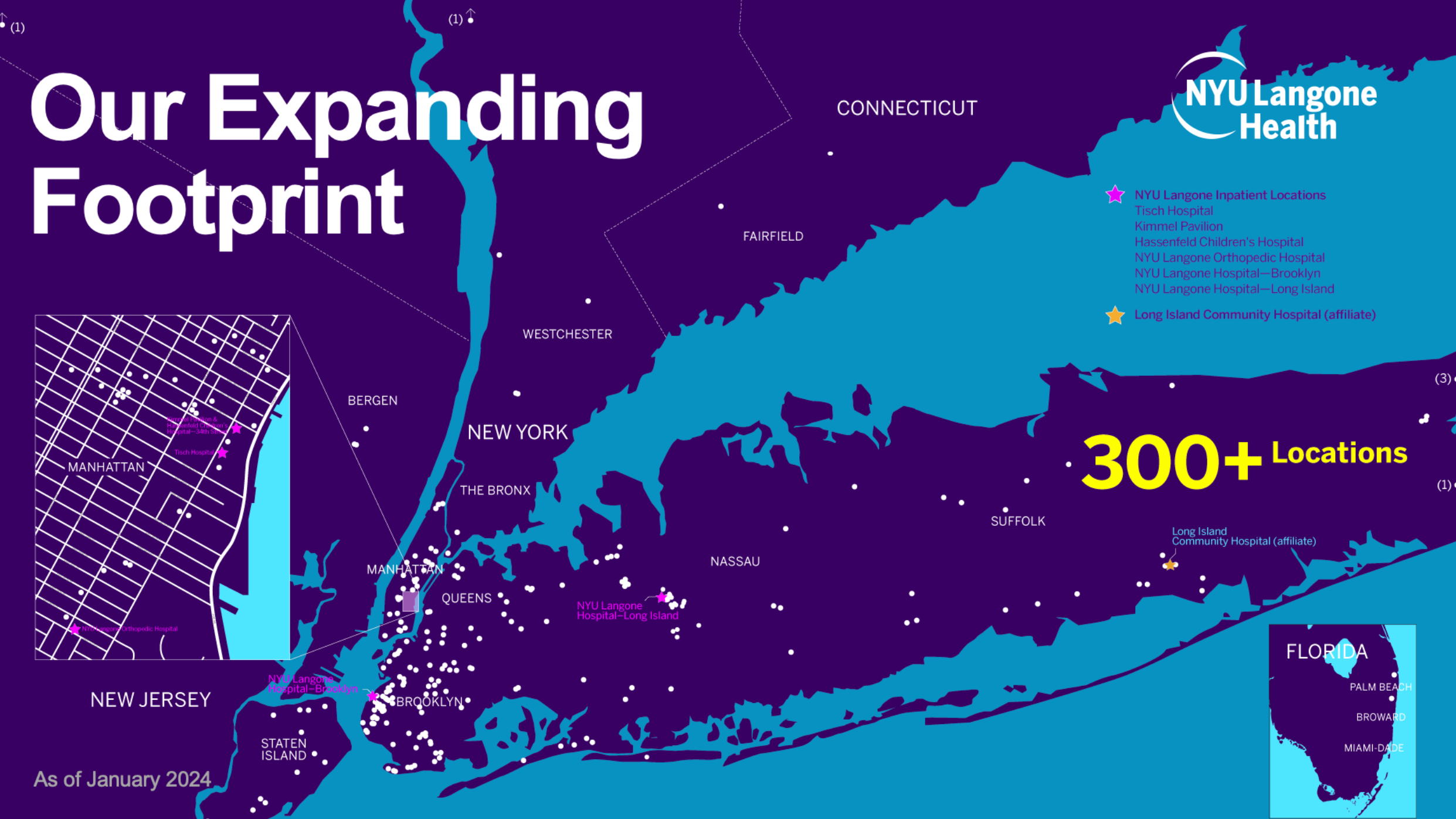
286,464
Emergency
Department Visits

10.1M
Outpatient Visits
per Year

Our Expanding Footprint



- ★ NYU Langone Inpatient Locations
 - Tisch Hospital
 - Kimmel Pavilion
 - Hassentfeld Children's Hospital
 - NYU Langone Orthopedic Hospital
 - NYU Langone Hospital—Brooklyn
 - NYU Langone Hospital—Long Island
- ★ Long Island Community Hospital (affiliate)



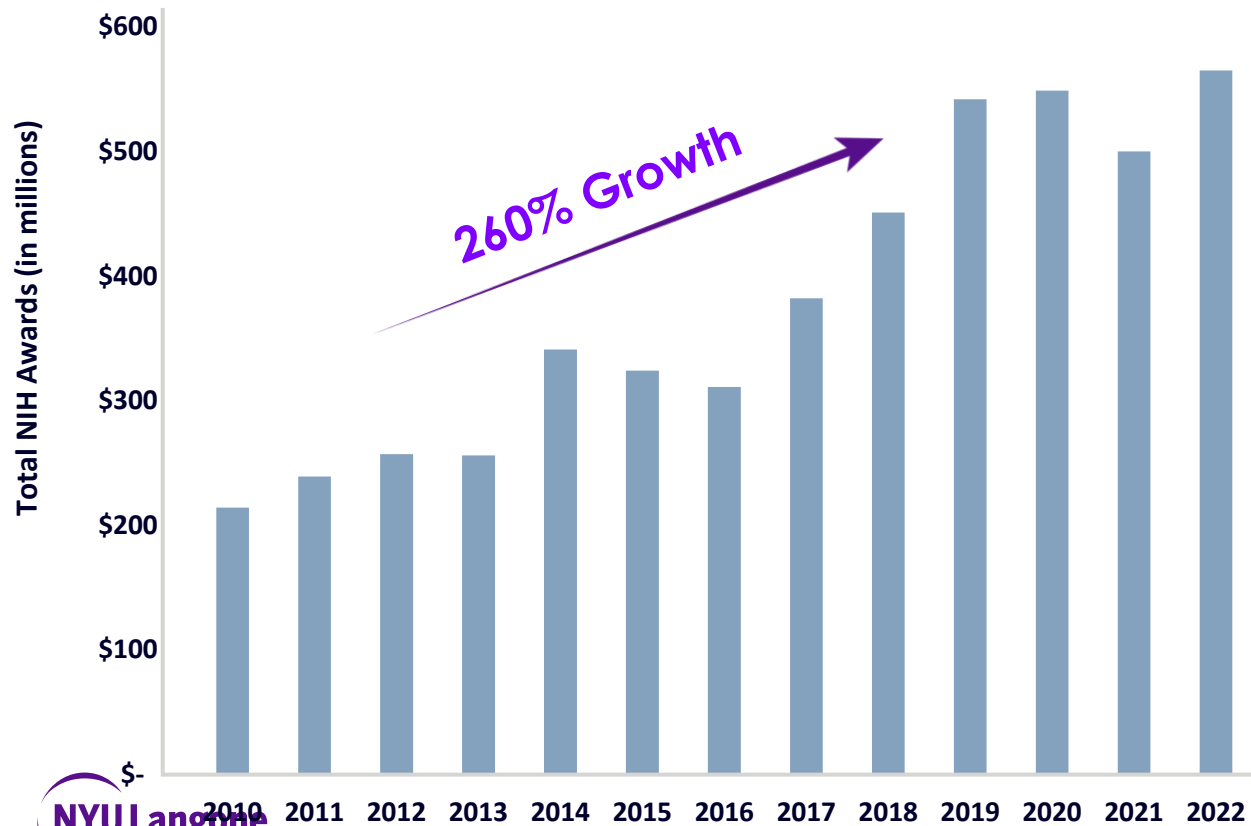
300+ Locations



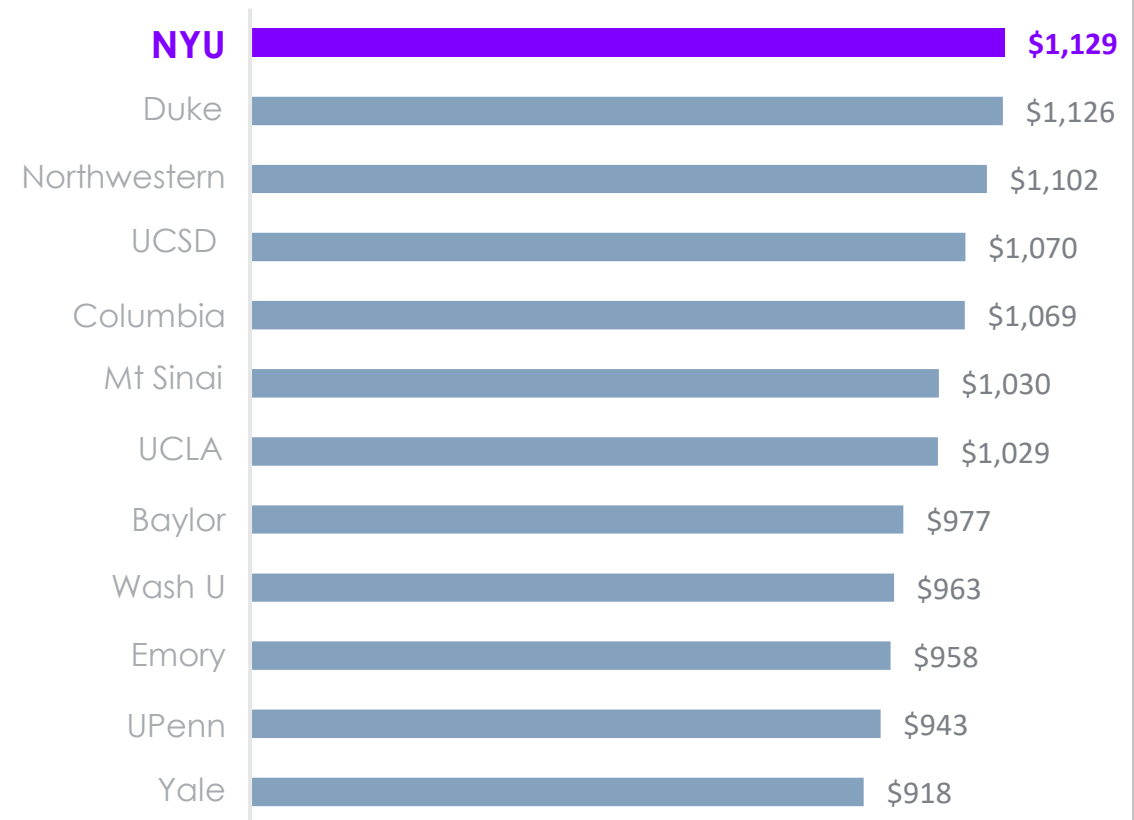
As of January 2024

Research Metrics

NIH Awards to NYU Grossman School of Medicine (2010-2022):
\$252m to \$549m



NIH Funding Per Research Faculty (FY22)



Harnessing Big Data For Biomedical Research



Epic

DATA GATHERING

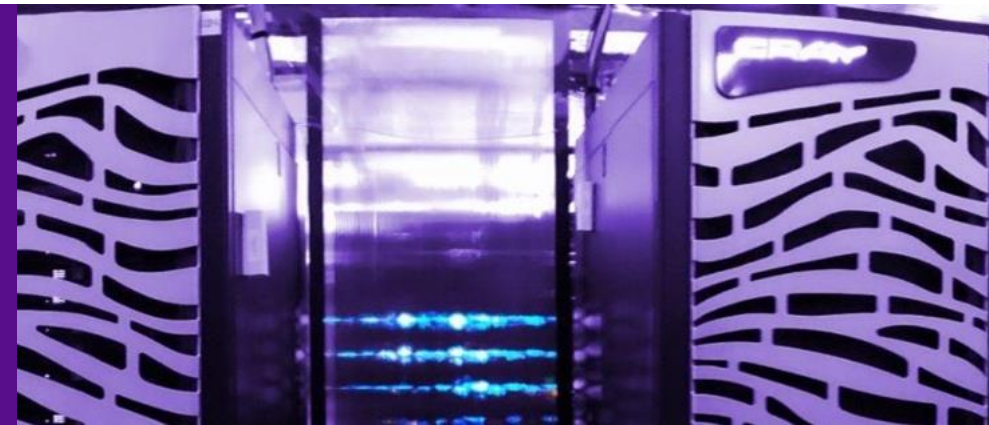
Electronic health records

- One of earliest adopter: **12+ years of clinical data**
- Over **8.9 million** patient records
- **300,000,000+** vital sign measurements
- **350,000,000+** documentation notes
- **650,000,000+** test results



Ultraviolet HPC Cluster

- #362 in Top500 (without the new SuperPOD): awaiting June top500
- #16 in US academic
- Only system in the Top 500 dedicated to healthcare and the life sciences



Generative AI in Medicine



Healthcare is an optimal place for AI transformation

- AI can be a new member of the healthcare team - part scribe, part co-pilot
- This will redefine some roles
 - Demand for healthcare far outstrips supply
 - Patients want care that is both more cutting-edge AND more human
 - AI can tackle the data overload facing clinicians, enabling higher-quality, safer, more efficient, and more precise care
- Ultimately, AI will free up clinicians for more patient interaction and higher order activities

Google Bard: “What medications are in this image and what are they used for?”

The medications in the image are:

- **Dextrose 50% in Water** - A solution of glucose that is used to treat low blood sugar.
- **Normal Saline** - A solution of salt water that is used to keep the body hydrated.
- **Vancomycin** - An antibiotic that is used to treat infections caused by bacteria.
- **Cefazolin** - An antibiotic that is used to treat infections caused by bacteria.
- **Fentanyl** - A pain medication that is used to treat severe pain.
- **Midazolam** - A sedative that is used to calm patients before surgery or other procedures.
- **Atropine** - A medication that is used to treat slow heart rate, low blood pressure, and other medical problems.
- **Diazepam** - A sedative that is used to treat anxiety and seizures.
- **Lidocaine** - A local anesthetic that is used to numb pain.
- **Epinephrine** - A medication that is used to treat allergic reactions and other medical emergencies.





“Tell me about the potential side effects of this anti-bacterial that my wife was prescribed for an infected finger”

The medication in the image you've shared is Cephalexin 500 mg capsules, which is indeed an antibiotic. Cephalexin is a cephalosporin antibiotic and is used to treat various bacterial infections. Common side effects of cephalexin include:

1. *Diarrhea*
2. *Nausea*
3. *Vomiting*
4. *Stomach pain*
5. *Dizziness*
6. *Fatigue*
7. *Joint pain*

Considerations

- **Hallucinations:** Fabricated facts, values
- **Lack of evidence:** Few randomized clinical trials; most studies have been on structured cases and not the real-world vagaries of healthcare and documentation
- **Automation bias:** the AI sounds like an expert, speaks confidently, and generates realistic text that people tend to trust without question
- **Non-generalizable models:** What works for one health system/specialty may not perform as well in others

AI at NYU Langone Health



MCIT Department of Health Informatics: Division of Applied Artificial Intelligence

Mission Statement:

Develop, translate, and apply artificial intelligence, machine learning, and predictive analytic models in support of the highest quality of care and operational advancement

- Empower the organization to harness the power of AI/ML
- Treat AI/ML as a core capability
- Align with marketplace
- Emphasize practical applications

Goals:

#1: To infuse NYU Langone with AI at all levels of the organization

#2: Be an Institutional Resource for Research and Operations

Over 30 Live AI models, e.g.: Reconciliation of Documentation and Orders

Goals: Reconcile intent to order a critical medication and a corresponding order is placed for that medication.

Status: Live today with GPT for

- Anti-coagulation
- Immunosuppressants

Results: Statistically significant increase in proportion of anti-coagulation orders within 6 hours of chart documentation


FENGI:

-Diet: Regular

-DVT ppx: Lovenox

#CODE STATUS: No orders of the

⚠ Possible Missing Immunosuppression



Clinical note indicates that patient is receiving tacrolimus but there is no active order.

Note Author: Attending P Nyu Ipo Inpatient, MD
Note Type: Progress Notes
Note Content: Assessment and Plan: The patient was ordered Envarsus XR as an immunosuppressant to prevent further episodes of
Date/Time: 2023-07-13T16:10:40Z

⚠ Acknowledge Reason

Medication(s) not indicated; will ensure... Note did not indicate plan for said medi...

Another Example: Triaging Epic In-Basket Messages

Goal: Prioritizing High Acuity Epic Inbasket Messages.

Status:

- **Validation:** Running silently for all incoming Epic inbasket messages.
- **Current Workflow:** prioritizes messages in nursing pools.

*Subject: Fussiness
Patient age: 3 months*

Hi Dr xxx., xxx has been fussy for the past two days. She's been crying almost every hour. She has no fever though. We tried giving her gripe water, which usually works but this time her crying is non-stop. We also noticed her cry has changed with episodes like she's gasping for air. This is the first time we've noticed this. Should we take her to the ER at this point?

*Model score: 78.98
RN called – Advise to go to ER
Patient had a follow up visit at NYU. Dx: Constipation and Breath-holding spell*

NYU Langone Health OpenAI Initiatives



Epic Generative AI

1. *LIVE* - Draft **response to MyChart** medical advice and medication refill requests
2. *Testing* - **Natural language** search embedded in **analytics** tools (SlicerDicer & Cosmos)
3. *Proposed* - **Quick summary** of patients for clinicians and end-of-shift hand-off report for Nursing



Ambient Documentation Powered by GPT

1. Live – proof of concept DAX for Providers – ambient note creation (Primary Care -24 NYULH physicians)
- **Current:** Primary Care
 - **Future:** Specialty Practices



NYU Generative AI (Private instance)

HIPAA compliant, secure for internal use

NYUTron



NYUTron (pronounced: new-tron): nature

Explore content ▾ About the journal ▾ Publish with us ▾

[nature](#) > [articles](#) > article

Article | [Open Access](#) | [Published: 07 June 2023](#)

Health system-scale language models are all-purpose prediction engines

[Lavender Yao Jiang](#), [Xujin Chris Liu](#), [Nima Pour Nejatian](#), [Mustafa Nasir-Moin](#), [Duo Wang](#), [Anas Abidin](#), [Kevin Eaton](#), [Howard Antony Riina](#), [Ilya Laufer](#), [Paawan Punjabi](#), [Madeline Miceli](#), [Nora C. Kim](#), [Cordelia Orillac](#), [Zane Schnurman](#), [Christopher Livia](#), [Hannah Weiss](#), [David Kurland](#), [Sean Neifert](#), [Yosef Dastagirzada](#), [Douglas Kondziolka](#), [Alexander T. M. Cheung](#), [Grace Yang](#), [Ming Cao](#), [Mona Flores](#), ... [Eric Karl Oermann](#)  [+ Show authors](#)

[Nature](#) (2023) | [Cite this article](#)

43k Accesses | **1291** Altmetric | [Metrics](#)

<https://doi.org/10.1038/s41586-023-06160-y>

- **7.25 million** notes
- **4.1 billion** words
- **9.5 years** of data
- **109 million** parameter BERT-like model
- **24** A100s for 3 weeks

a

Clinical
task



Physician

In-hospital mortality prediction

How likely is the patient to die in the hospital before discharge?

Binned comorbidity index imputation

Without structured ICDS, how sick/chronically ill is the patient?

30-day all-cause readmission prediction

How likely is the patient to come back within 30 days of discharge?

Operational
task



Admin

Binned LOS prediction

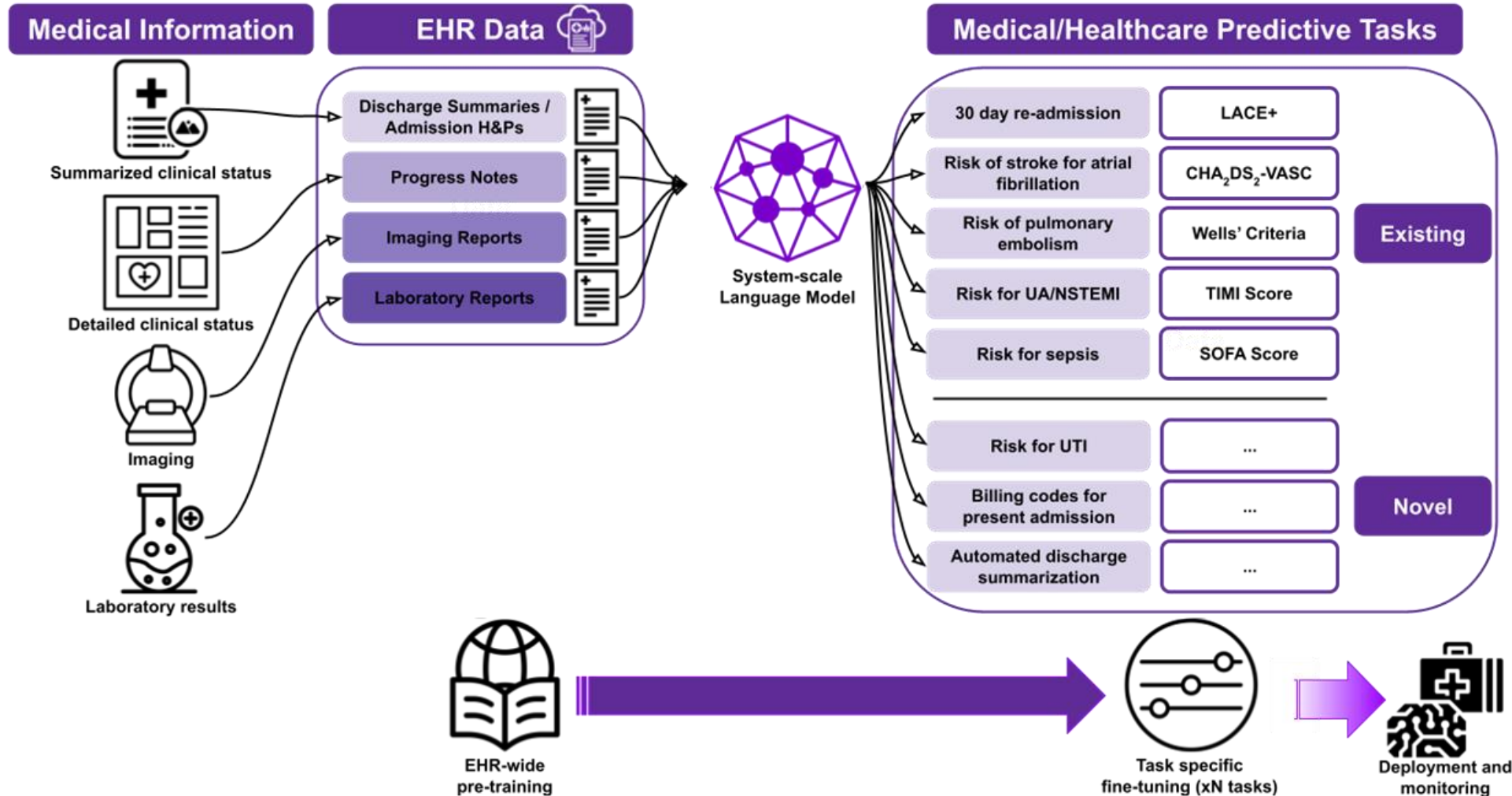
How long will the patient stay in the hospital?

Insurance denial prediction

How likely is the patient's insurance claim to be denied?

This article is in the 99th percentile (ranked 180th) of the 235,626 tracked articles of a similar age in all journals and the 97th percentile (ranked 22nd) of the 729 tracked articles of a similar age in *Nature*

Hypothesis: If we train a LLM at the scale of our health system then we'll have a flexible tool for predictive analytics



Epic Electronic Health Record

The EHR records everything relevant to the care of each patient: everything measured, recorded and done

We used EHR records from Jan 2011 to May 2020.

Patients (N)

- 336,000 patients for pretraining
- 413,845 patients for finetuning



A unified approach to solve the last-mile problems

NARRATIVE & IMPRESSION

Pre-treatment WBS I 123 dated 6/5/19.

Findings:

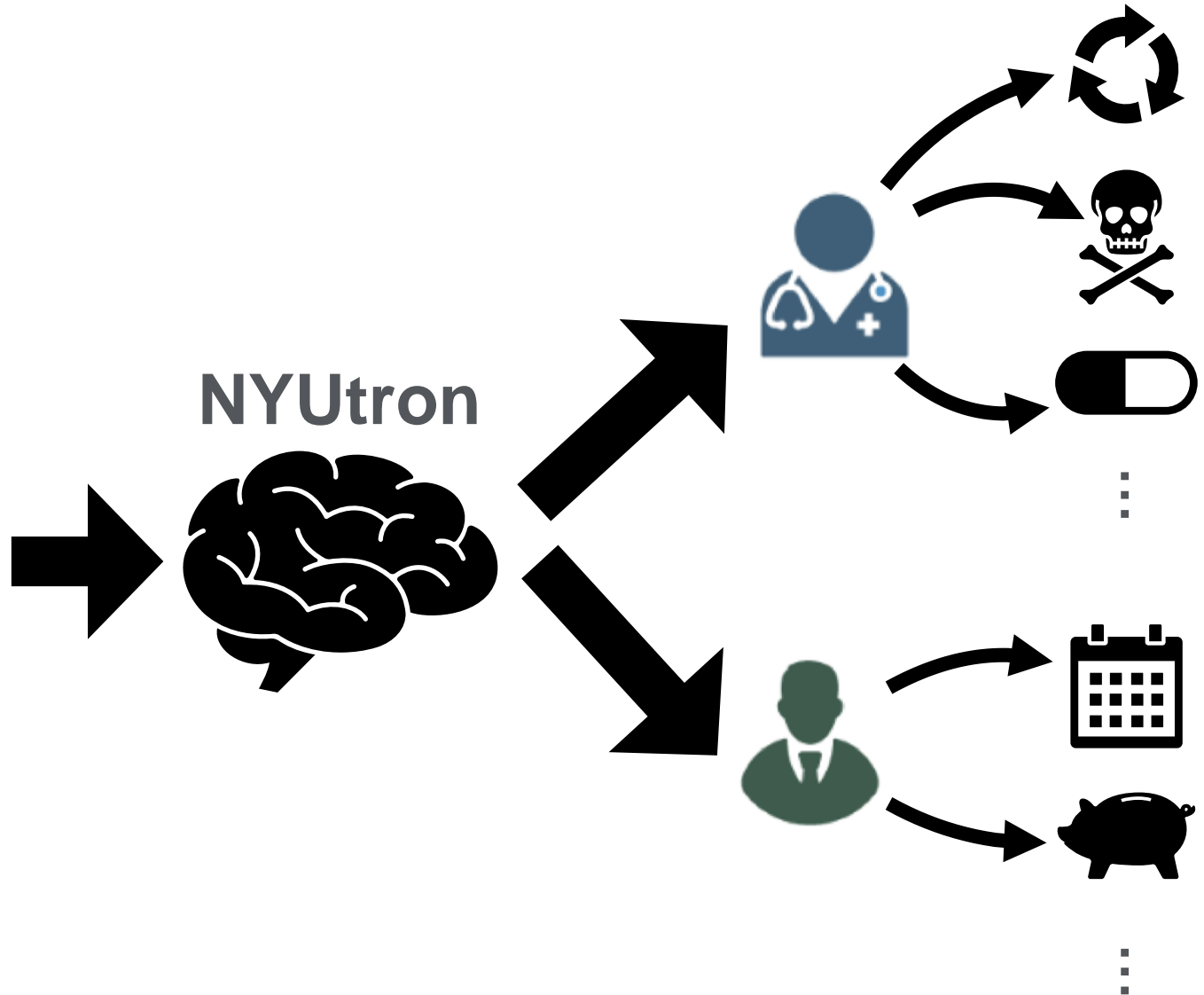
Redemonstration of two foci of increased tracer uptake in the mid-line in the neck, corresponding to same sites of the pre-therapy I 123 scan. SPECT/CT localizes these foci of uptake to the thyroglossal duct and the thyroid remnant in the left thyroid bed.

A focus of tracer uptake is seen in the right posterior maxilla, most likely dental in etiology. There is faint tracer uptake in the right neck, in the region of the right parotid gland, likely representing retained activity in the right parotid which is somewhat asymmetric. A focus of trace uptake in the right lower abdomen corresponds to physiological tracer activity in the bowel on SPECT/CT.

No new suspicious focus of tracer uptake is seen.

On CT, right maxillary sinus mucosal thickening. Thyroidectomy. No cervical lymphadenopathy. Visualized upper lung fields are normal.

Electronic Signature: I personally reviewed the images and



Benchmark clinical and operational tasks

Clinical Task



Physician



In-hospital mortality prediction

How likely will the patient die in the hospital before discharge?



Binned comorbidity index imputation

Without structured icds, how sick / chronically ill is the patient?



30-day all-cause readmission prediction

How likely will the patient come back within 30 days of discharge?

Operational Task



Admin



Binned LOS (length-of-stay) prediction

How long will the patient stay in the hospital?



Insurance denial prediction

How likely will the patient's insurance claim get denied?

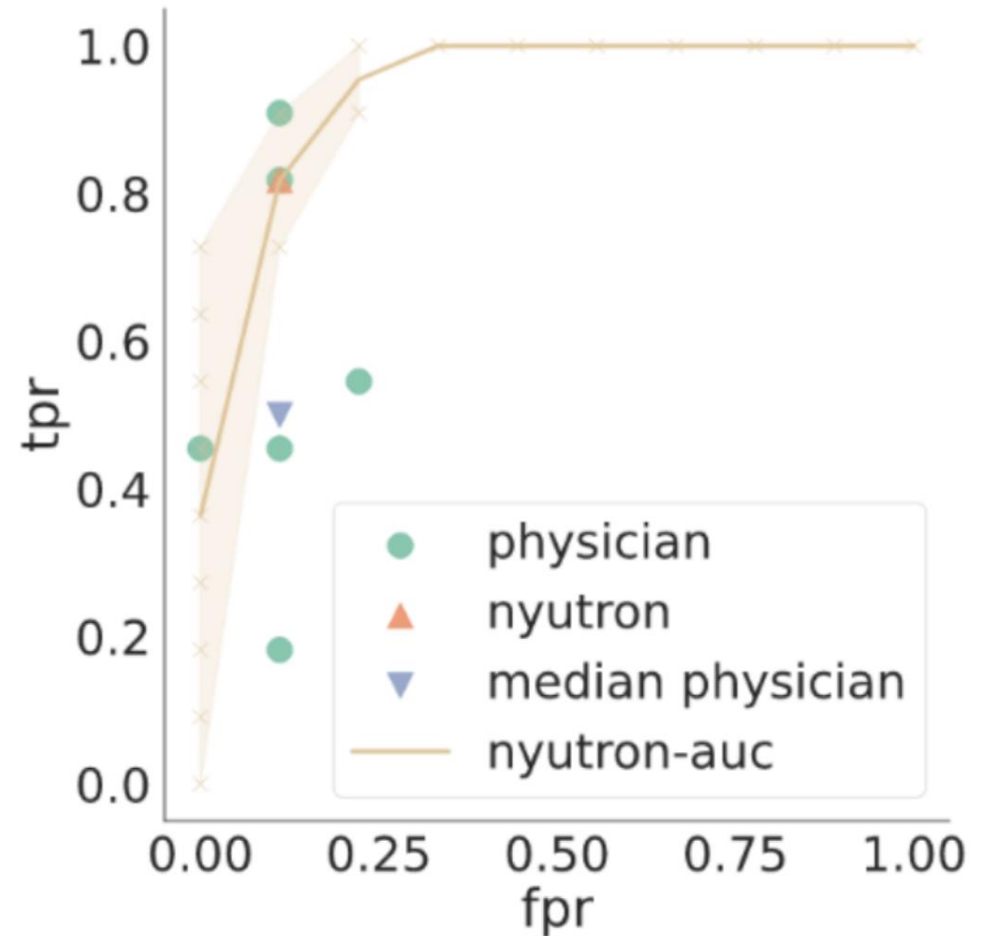
NYUTron: results



NYUtron is competitive with physicians

Readmission prediction

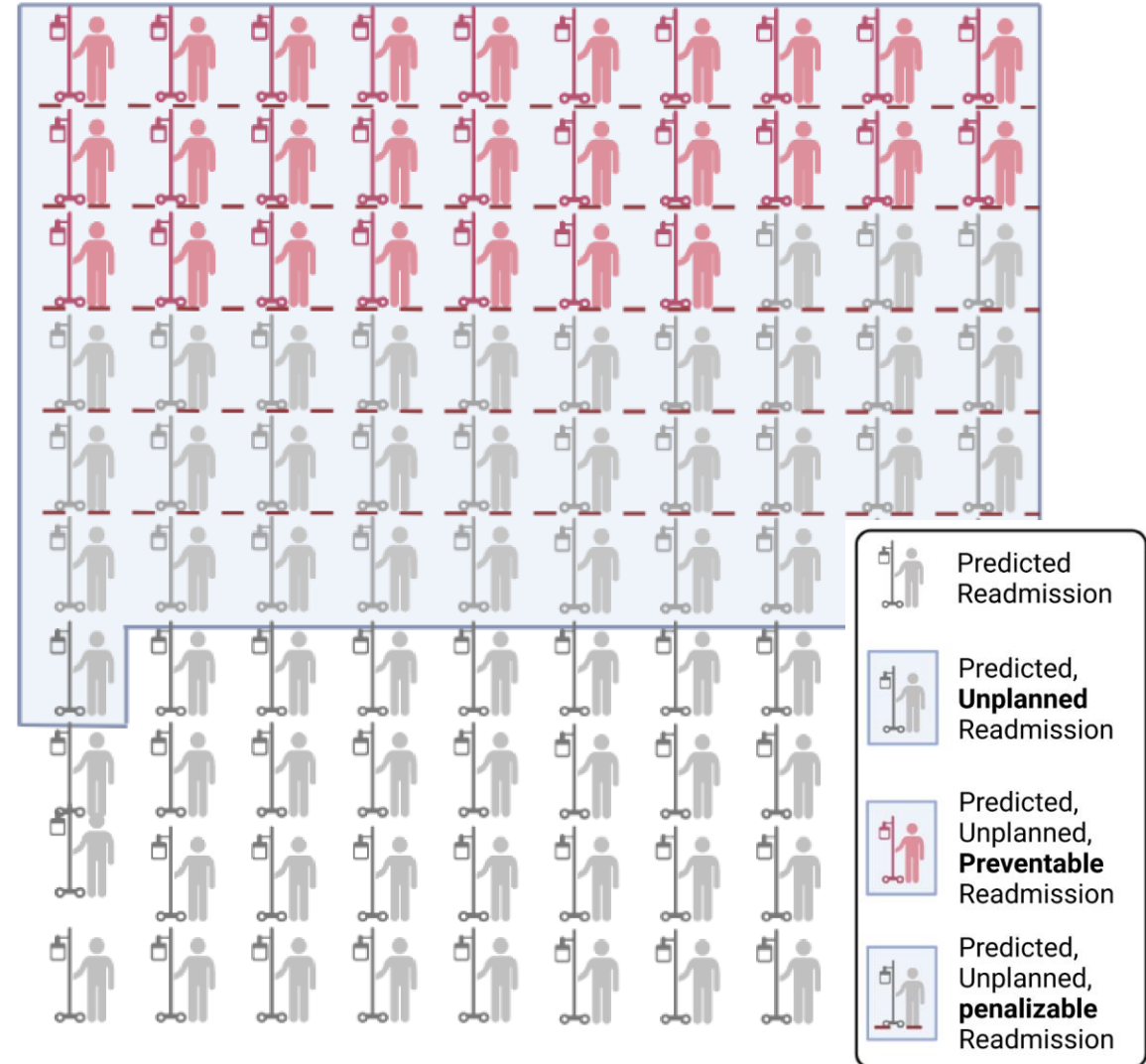
- We'll focus on readmission prediction...
- 6 physicians at different levels of seniority from NYU Langone
- NYUtron is better than the median physician, but the best physician performs better than NYUtron.



Prospective, observational study

A prospective, single-arm, non-interventional trial

- Deployed Spring/Summer 2022
- Detailed analysis by physicians
 - A panel of six physicians
 - 100 randomly sampled readmitted cases captured by NYUTron
- A significant portion correspond to preventable (50%) readmission



Takeaways

- **NYUTron models are general purpose clinical predictive engines**
 - We can predict 30-day all-cause readmission risk at the time the clinician signs their notes
 - We also have tasks for length of stay and in-hospital mortality prediction
- **Seamless integration with clinical workflows is paramount**
- **The big bottlenecks are *hardware, engineering expertise, and datasets***
 - **Hardware:** Most academic medical institutions don't invest in significant HPC capacity
 - **Engineering:** Consequently, most academics don't pretrain large LLMs on optimized in-house software stacks
 - **Datasets:** The lack of massive un-labelled and large, well labelled datasets is the single biggest barrier to research

Supporting AI with HPC



Resources: Ultraviolet

- 197 nodes, 646 GPUs, 13,736 CPUs, Active Storage 15PB

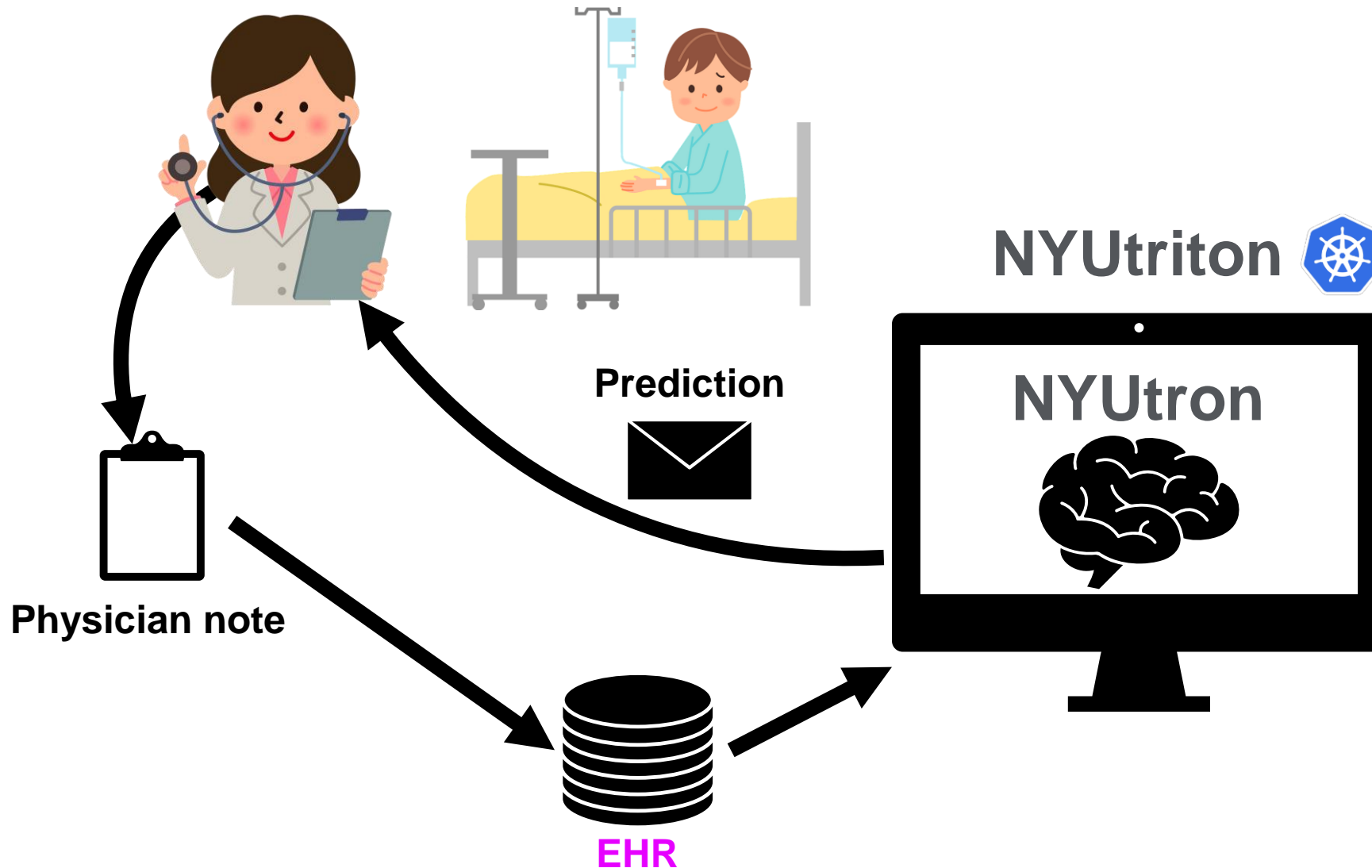
Condo	# of Nodes	CPU (sockets/core/HT/model)	# GPU/Model/NV Link	RAM (GB)
SuperPOD	16	2/56/2/Intel(R) Xeon(R) Platinum 8480CL	8/H100/NV18X	2048
Radiology	10	2/40/2/Intel(R) Xeon(R) Platinum 8380	4/A100/NV12	1024
A100-4	48	2/24/1/Intel(R) Xeon(R) Gold 6342	4/a100/NV12	512
A100-8	3	2/48/1/AMD EPYC 7643	8/A100/NV12X	1024
A100nv	7	2/32/1/AMD EPYC 7543	4/A100/NV12X	512
NL Condo (10 GPU)	2	2/64/1/AMD EPYC 7763	10/L40/PXB	512
NL Condo (6 GPU)	8	2/64/1/AMD EPYC 7763	6/L40/PXB	512
UV/gn	25	2/20/1/Intel(R) Xeon(R) Gold 6148	4/V100/NV2	376
UV/gpu	8	2/20/1/Intel(R) Xeon(R) Gold 6148	8/V100/NV1	754

Reservation paradigm

- *“I need to reserve at least 64 A100 GPUs for two weeks to make an LLM/an ensemble of LLMs for...”*
- Shorter training time is important to stay competitive
- With ~2,000 users, fair share resource scheduling is not sufficient to support the quantity of resources needed to support LLMs
- Challenges with high demand and limited resources: checkpointing, education, optimization

Deployment

Modified NVIDIA Triton Inference Server: NYUTriton



More boxes





Thank You!