LLM Rubric Representation Learning

LLMs can construct powerful representations and streamline sample-efficient supervised learning

Ilker Demirel, Lawrence Shi, Zeshan Hussain, David Sontag
https://lrrlpaper.github.io/

Representational challenges in supervised learning

Nursing Note Example

XX/XX/20XX	X:XXpm

The patient is a 64 year old female whose chief complaint is an exacerbation of chronic migraines. Patient reports an 8/10 pain level that decreases to 7/10 on the current prescription of 600mg of ibuprofen once a day. Patient states she feels nauseous, which is a common symptom of her migraine exacerbations. She describes the quality of migraine as a "stabbing, throbbing feeling." Patient states she struggles with bright lighting and loud sounds because they worsen her migraines. In these instances, she says her pain score rises to 9/10.

Patient is AAOx3. HR=95 bpm, RR=20 bpm, T=98.6, BP=130/90, PO2=98%. Bowel sounds present and bowel movements remain regular. No complaints of impaired vision nor memory. Patient states there are no life stressors adding to her current migraine exacerbation. Patient states her sleep schedule is normal.

Mayowa Okumba, APRN, ordered a basic metabolic panel (BMP) and an MRI scheduled for tomorrow. Patient is aware and consents to these procedures. Will continue to monitor vitals and for any other signs of discomfort.

Cherise McDonald, RN

Image Source

IntelyCare

Clinical notes, free-text

EXAMINATION: CHEST (PA AND LAT)

INDICATION: ___ year old woman with ?pleural effusion // ?pleural effusion

TECHNIQUE: Chest PA and lateral

COMPARISON: ___

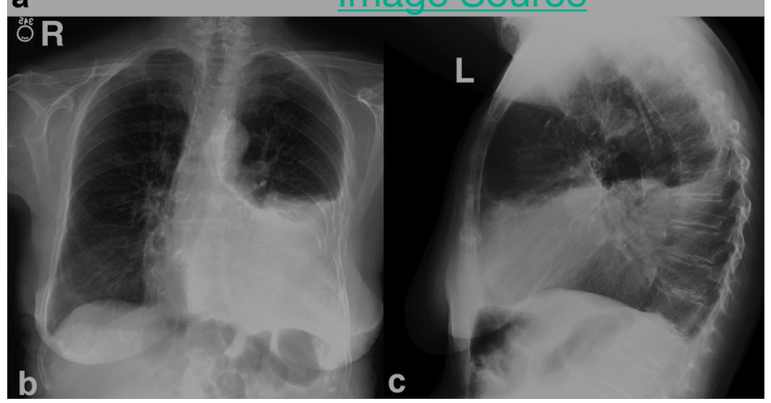
FINDINGS:

Cardiac size cannot be evaluated. Large left pleural effusion is new. Small right effusion is new. The upper lungs are clear. Right lower lobe opacities are better seen in prior CT. There is no pneumothorax. There are mild degenerative changes in the thoracic spine

IMPRESSION:

Large left pleural effusion

Image Source



Medical images

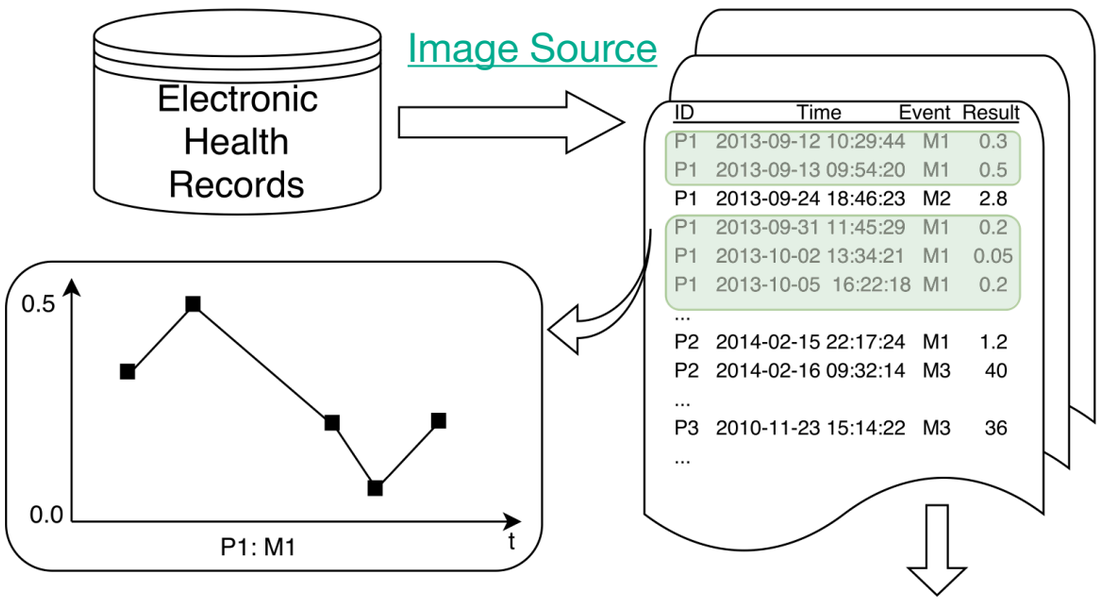
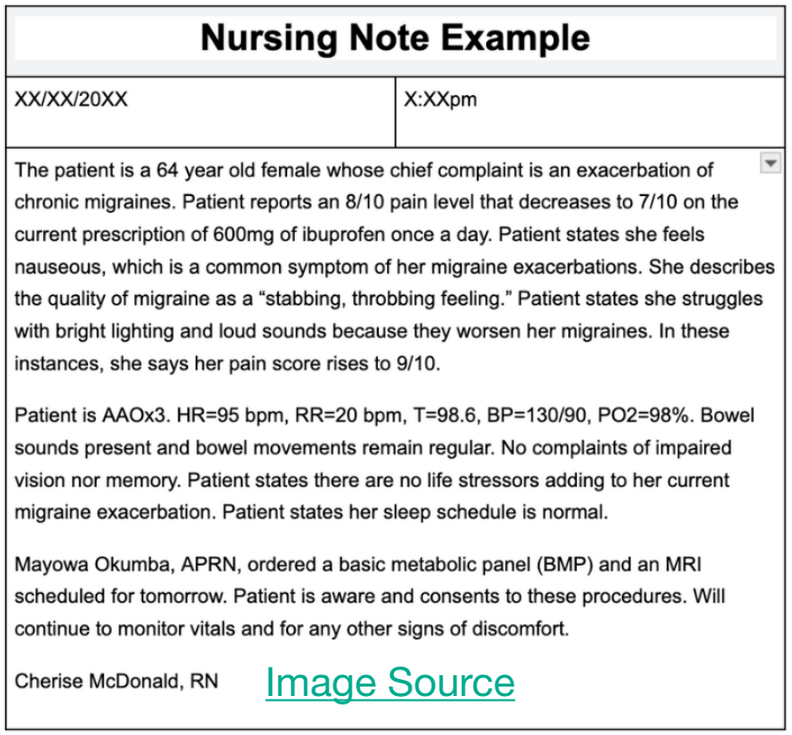


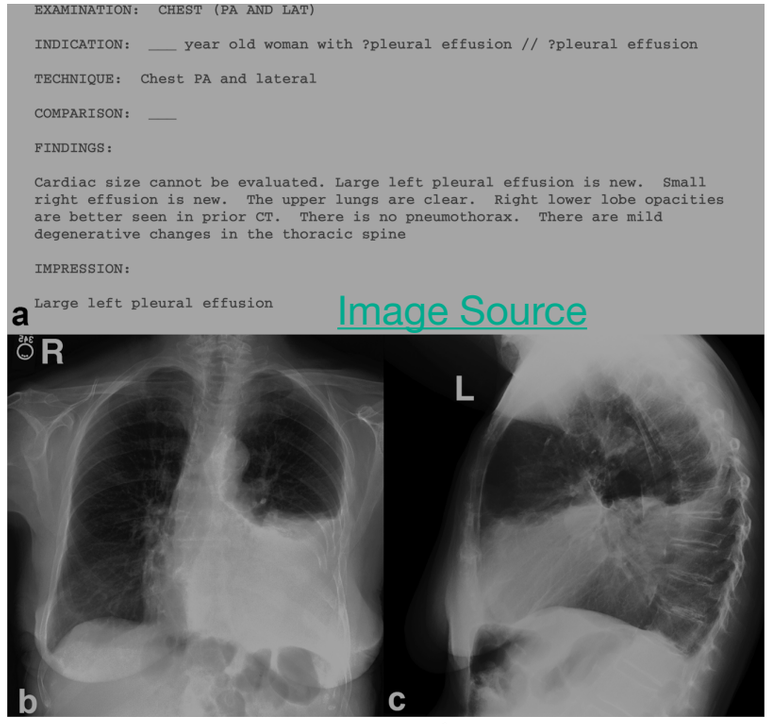
Image Source

Structured tables / Time-series

Representational challenges in supervised learning

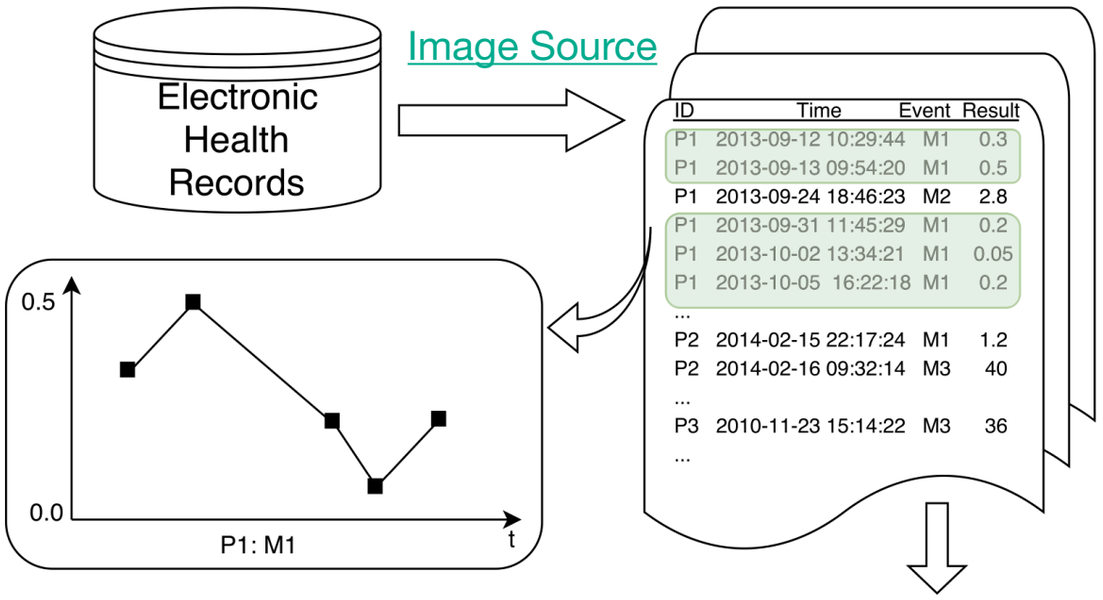


Clinical notes, free-text



Medical images

Supervised learning (e.g. binary clf.)



Structured tables / Time-series

- What to include in the input?
- How to represent the input ($\mathbf{X}$)?
- What ML models to use?

Representational challenges in supervised learning

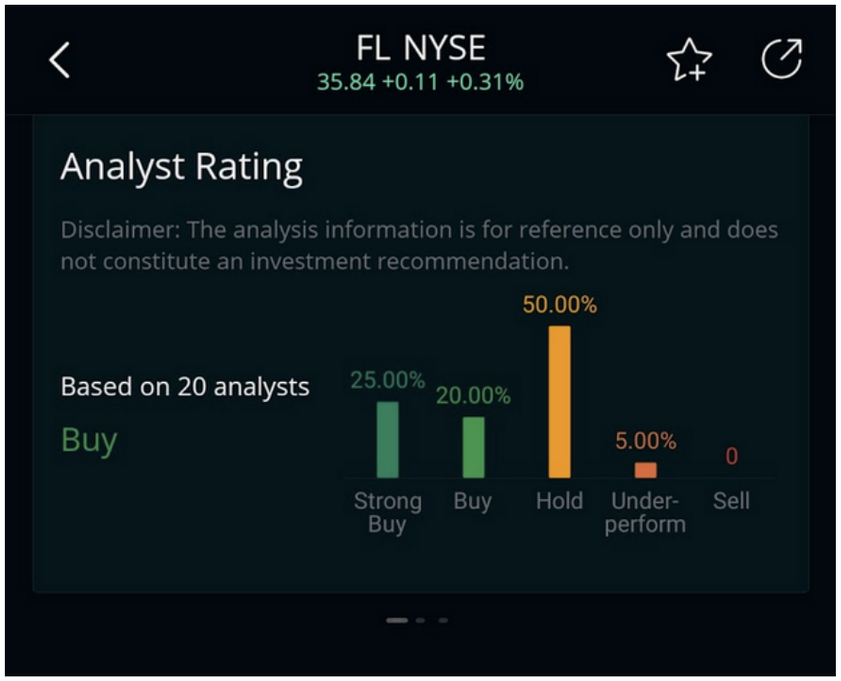
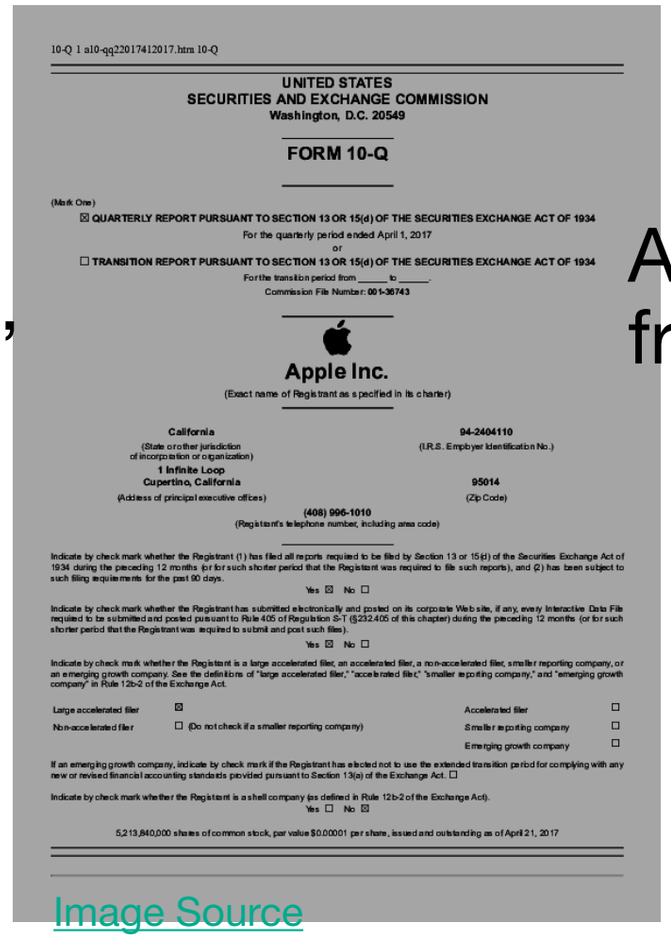


Image Source

Analyst ratings, tabular



10-Q 1 a10-qq22017412017.htm 10-Q

UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549

FORM 10-Q

(Mark One)

☒ **QUARTERLY REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934**
For the quarterly period ended April 1, 2017
or
☐ **TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934**
For the transition period from _____ to _____.
Commission File Number: **001-36743**

Apple Inc.
(Exact name of Registrant as specified in its charter)

California	**94-2404110**
(State or other jurisdiction of incorporation or organization)	(I.R.S. Employer Identification No.)
1 Infinite Loop, Cupertino, California	**95014**
(Address of principal executive offices)	(Zip Code)

(408) 996-1010
(Registrant's telephone number, including area code)

Indicate by check mark whether the Registrant (1) has filed all reports required to be filed by Section 13 or 15(d) of the Securities Exchange Act of 1934 during the preceding 12 months (or for such shorter period that the Registrant was required to file such reports), and (2) has been subject to such filing requirements for the past 90 days.
Yes ☒ No ☐

Indicate by check mark whether the Registrant has submitted electronically and posted on its corporate Web site, if any, every Interactive Data File required to be submitted and posted pursuant to Rule 405 of Regulation S-T (§232.405 of this chapter) during the preceding 12 months (or for such shorter period that the Registrant was required to submit and post such files).
Yes ☒ No ☐

Indicate by check mark whether the Registrant is a large accelerated filer, an accelerated filer, a non-accelerated filer, smaller reporting company, or an emerging growth company. See the definitions of "large accelerated filer," "accelerated filer," "smaller reporting company," and "emerging growth company" in Rule 12b-2 of the Exchange Act.

Large accelerated filer	☒	Accelerated filer	☐
Non-accelerated filer	☐ (Do not check if a smaller reporting company)	Smaller reporting company	☐
		Emerging growth company	☐

If an emerging growth company, indicate by check mark if the Registrant has elected not to use the extended transition period for complying with any new or revised financial accounting standards provided pursuant to Section 13(a) of the Exchange Act. ☐

Indicate by check mark whether the Registrant is a shell company (as defined in Rule 12b-2 of the Exchange Act).
Yes ☐ No ☒

5,213,840,000 shares of common stock, par value $0.00001 per share, issued and outstanding as of April 21, 2017

Image Source

Annual reports, free-text

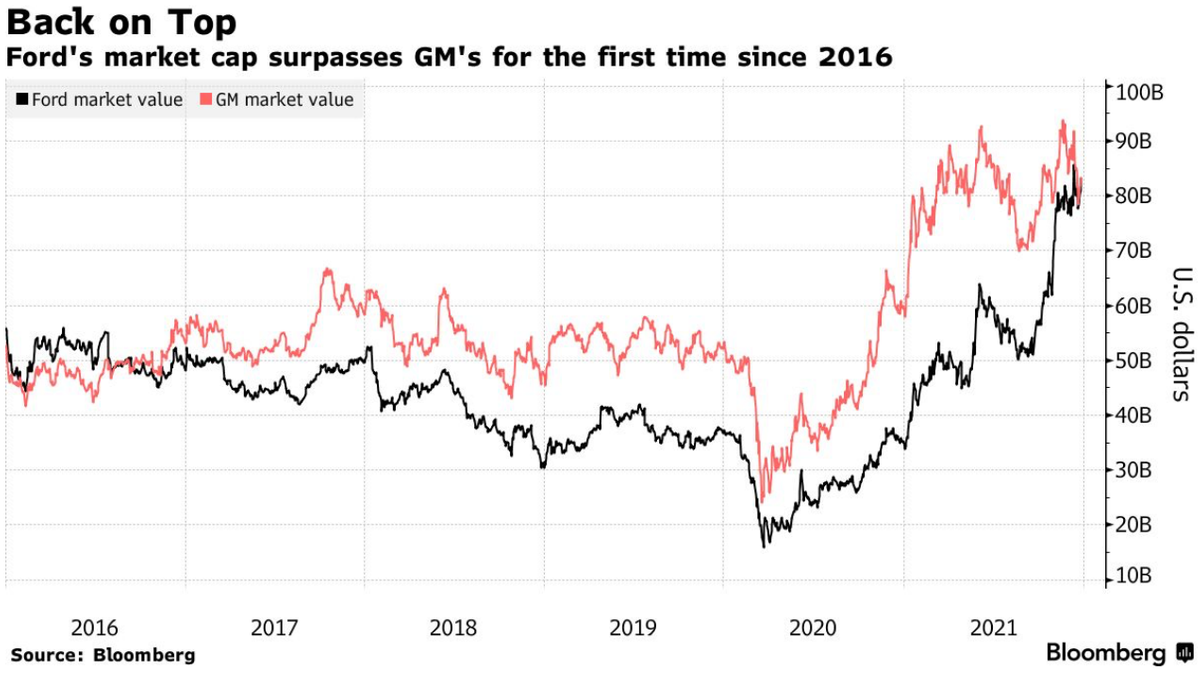


Image Source

Market cap, time-series

LLMs as an interface to complex / heterogeneous datasets

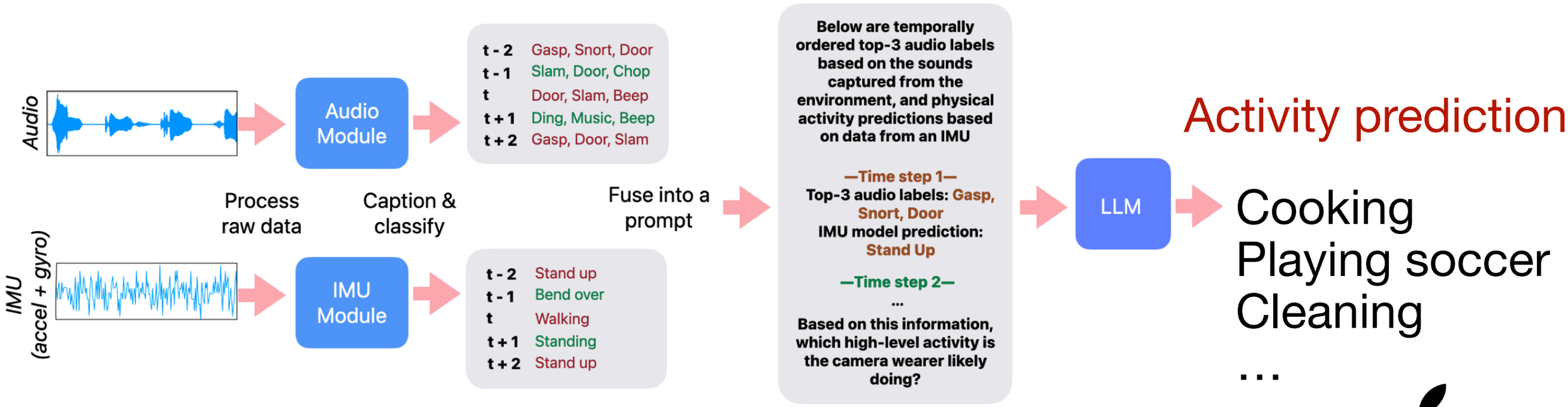
```
cell: cell_a
2024/06/02 17:00:00 PDT, Day:Fri Week:21
search_space:
  {'JOB/data_pipeline/PRODUCTION_WORKLOAD':
    ['machineE', 'machineA', 'none_selected']}
assignments: {"JOB/data_pipeline/PRODUCTION_WORKLOAD": "machineA"}
distributions:
- platform: {machineA}
  num_machines: 1.239e+03
  low_level_zones: 5.200e+01
  mid_level_zones: 5.200e+01
  high_level_zones: 4.300e+01
  resources: 5.481e+05
job_profiles:
- job: {user: data_pipeline, group_name: data_pipeline_workers}
  platform_profiles:
    machineD: {mean_mips_per_resource_usage: 8.165e+02}
    machineA: {mean_mips_per_resource_usage: 9.590e+02}
    machineF: {mean_mips_per_resource_usage: 8.321e+02}
    machineC: {mean_mips_per_resource_usage: 7.098e+02}
  limits:
    job_requested_resource_limit: 1.217e+04
    job_requested_num_vms: 1087
```

Text-to-text regression to predict system performance from configuration [Paper]

LLMs as an interface to complex / heterogeneous datasets

```
cell: cell_a
2024/06/02 17:00:00 PDT, Day:Fri Week:21
search_space:
  {'JOB/data_pipeline/PRODUCTION_WORKLOAD':
    ['machineE', 'machineA', 'none_selected']}
assignments: {"JOB/data_pipeline/PRODUCTION_WORKLOAD": "machineA"}
distributions:
- platform: {machineA}
  num_machines: 1.239e+03
  low_level_zones: 5.200e+01
  mid_level_zones: 5.200e+01
  high_level_zones: 4.300e+01
  resources: 5.481e+05
job_profiles:
- job: {user: data_pipeline, group_name: data_pipeline_workers}
  platform_profiles:
    machineD: {mean_mips_per_resource_usage: 8.165e+02}
    machineA: {mean_mips_per_resource_usage: 9.590e+02}
    machineF: {mean_mips_per_resource_usage: 8.321e+02}
    machineC: {mean_mips_per_resource_usage: 7.098e+02}
  limits:
    job_requested_resource_limit: 1.217e+04
    job_requested_num_vms: 1087
```

Text-to-text regression to predict system performance from configuration [Paper]



Activity prediction from text-transformed multimodal data [Paper]

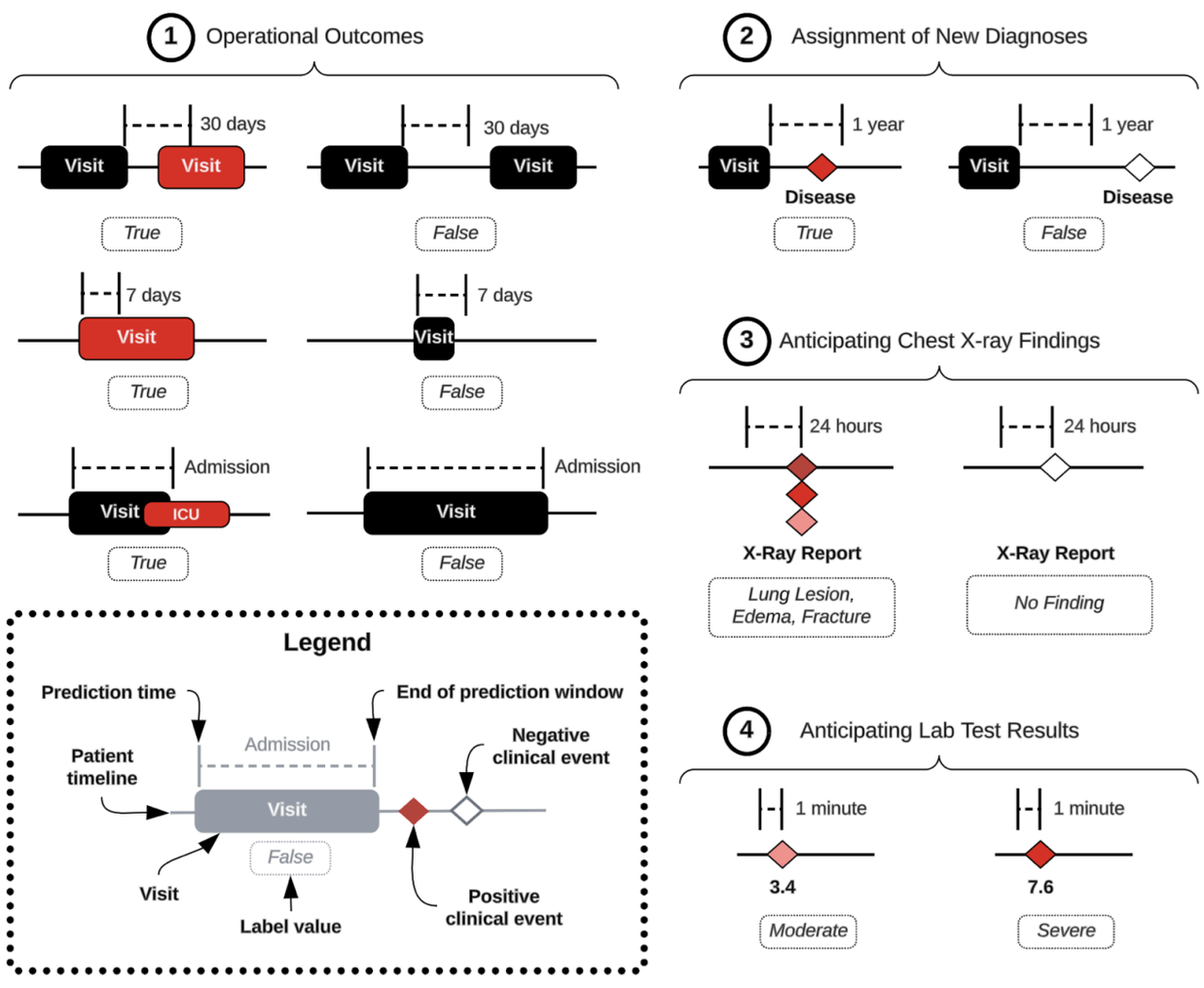
EHRSHOT benchmark

Wornow et al, 2023. ShahLab @ Stanford

Full longitudinal EHR of 6,739 patients

- Temporally ordered visits
- Coded events such as
- Labs, diagnoses, medications…

Category	Task	Train	Val	Test
Operational Outcomes (3)	ICU transfer	2402 (113)	100 (50)	2037 (85)
	Length of stay >7 days	2569 (681)	100 (50)	2195 (552)
	30-day readmission	2608 (370)	100 (50)	2189 (260)
Assignment of New Diagnosis (6)	Hypertension	1259 (182)	100 (50)	1258 (159)
	Hyperlipidemia	1684 (205)	100 (50)	1317 (172)
	Pancreatic cancer	2576 (155)	100 (50)	2220 (56)
	Celiac disease	2623 (62)	22 (11)	2222 (21)
	Lupus	2570 (104)	66 (33)	2243 (20)
	Acute MI	2534 (175)	100 (50)	2127 (144)
Anticipating Lab Results (5)	Thrombocytopenia	2000 (1000)	100 (50)	2000 (1000)
	Hyperkalemia	2000 (1000)	100 (50)	1896 (948)
	Hypoglycemia	2000 (1000)	100 (50)	1566 (783)
	Hyponatremia	2000 (1000)	100 (50)	2000 (1000)
	Anemia	2000 (1000)	100 (50)	2000 (1000)
Chest X-ray Findings (1)	Chest X-ray abnormality	2000 (1000)	100 (50)	2000 (1000)



EHRSHOT Paper

EHRSHOT benchmark

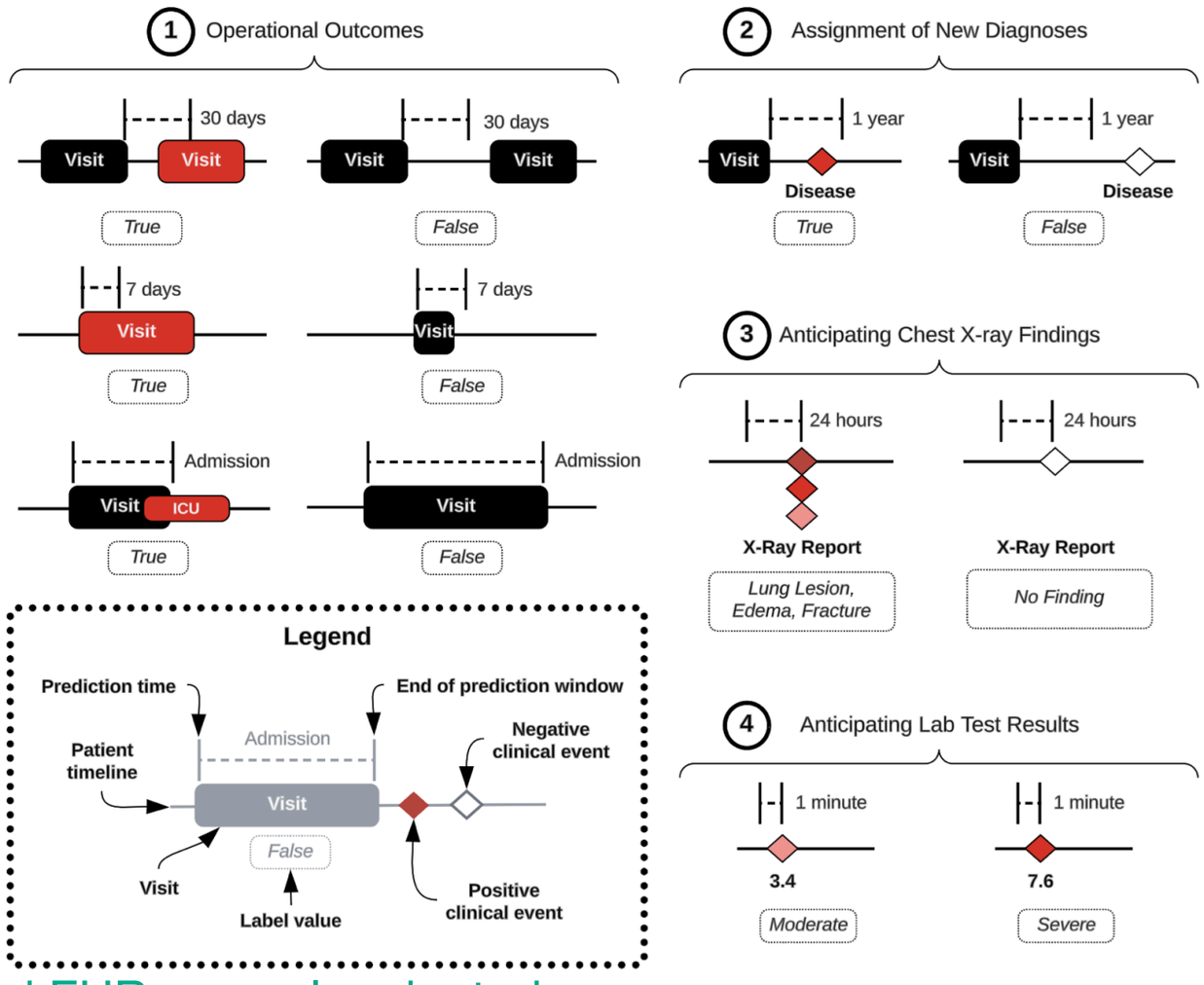
Wornow et al, 2023. ShahLab @ Stanford

Full longitudinal EHR of 6,739 patients

- Temporally ordered visits
- Coded events such as
- Labs, diagnoses, medications…

```
# Naive Text Serialization
## Patient Demographics
- Patient age:  78, FEMALE [...]
## Detailed Past Medical Visits
### Inpatient Visit (14 days to pred.  time, current visit)
#### Conditions
- Acute posthemorrhagic anemia
- pH measurement, venous:  7.25, 7.31, 7.31 [...]
#### Medications
- furosemide 20 MG Oral Tablet
- pantoprazole 20 MG Delayed Release Oral Tablet [...]
#### Procedures
- Chest x-ray
- Electrocardiogram report [...]
### Emergency Room Visit (87 days before prediction time)
#### Conditions
- Benign essential hypertension
- Chest pain [...]
#### Medications
- 2ML ondansetron 2MG/ML inject.
- nitroglycerin 0.4 MG [...]
#### Procedures
- Ct angiography
- Comprehen metabolic panel[...]
```

Category	Task	Train	Val	Test
Operational Outcomes (3)	ICU transfer	2402 (113)	100 (50)	2037 (85)
	Length of stay >7 days	2569 (681)	100 (50)	2195 (552)
	30-day readmission	2608 (370)	100 (50)	2189 (260)
Assignment of New Diagnosis (6)	Hypertension	1259 (182)	100 (50)	1258 (159)
	Hyperlipidemia	1684 (205)	100 (50)	1317 (172)
	Pancreatic cancer	2576 (155)	100 (50)	2220 (56)
	Celiac disease	2623 (62)	22 (11)	2222 (21)
	Lupus	2570 (104)	66 (33)	2243 (20)
	Acute MI	2534 (175)	100 (50)	2127 (144)
Anticipating Lab Results (5)	Thrombocytopenia	2000 (1000)	100 (50)	2000 (1000)
	Hyperkalemia	2000 (1000)	100 (50)	1896 (948)
	Hypoglycemia	2000 (1000)	100 (50)	1566 (783)
	Hyponatremia	2000 (1000)	100 (50)	2000 (1000)
	Anemia	2000 (1000)	100 (50)	2000 (1000)
Chest X-ray Findings (1)	Chest X-ray abnormality	2000 (1000)	100 (50)	2000 (1000)



Text-serialized EHR example adapted from Hegselmann et al (2025)

EHRSHOT Paper

EHR text / rubric representations

Rubric: A recipe for transforming *naive* text serialization of input into a more structured / standardized representation

 # Naive Text Serialization

Patient Demographics
- Patient age: 78, FEMALE [...]

Detailed Past Medical Visits

Inpatient Visit (14 days to pred. time, current visit)

Conditions
- Acute posthemorrhagic anemia
- pH measurement, venous: 7.25, 7.31, 7.31 [...]

Medications
- furosemide 20 MG Oral Tablet
- pantoprazole 20 MG Delayed Release Oral Tablet [...]

Procedures
- Chest x-ray
- Electrocardiogram report [...]

Emergency Room Visit (87 days before prediction time)

Conditions
- Benign essential hypertension
- Chest pain [...]

Medications
- 2ML ondansetron 2MG/ML inject.
- nitroglycerin 0.4 MG [...]

Procedures
- Ct angiography
- Comprehen metabolic panel[...]

NaiveText

 # Task-conditioned local rubric generation prompt

GOAL: Read the patient's text-serialized EHR and write a compact reasoning trace that characterizes the patient's risk profile for the following clinical outcome prediction task: {task_query}

--- START OF EHR DATA ---
{NaiveText_Serialization (x^{text})}
--- END OF EHR DATA ---

Your output MUST follow this exact structure:

1. Patient Snapshot
2. Main Risk Factors
3. Protective Factors
4. What's Unknown / Could Swing the Risk
5. Weighing and Aggregating the Evidence
6. Overall Risk Impression

Example task query:
Will the patient develop hypertension in the next year?

EHR text / rubric representations

Rubric: A recipe for transforming *naive* text serialization of input into a more structured / standardized representation

```
# Naive Text
## Patient D
- Patient ag
## Detailed
### Inpatien
pred. time,
#### Conditi
- Acute post
- pH measure
7.31, 7.31 |
#### Medicat
- furosemide
- pantoprazo
Release Oral
#### Procedu
- Chest x-ra
- Electrocar
### Emergenc
days before
#### Conditi
- Benign ess
- Chest pain
#### Medicat
- 2ML ondans
- nitroglyce
#### Procedu
- Ct angiogr
- Comprehen
```

NaiveText

```
# Local Rubric Representation

1. Patient Snapshot
27 yo hispanic male. Recurrent
cardiology visits for
congenital anomaly of coronary
artery [...]

2. Main Risk Factors
- Congenital coronary artery
anomaly (established structural
predisposition to myocardial
ischemia/infarction).
- Tobacco exposure (smokeless
tobacco reported) [...]

3. Protective Factors
- Young age (27) | lower
baseline atherosclerotic burden
relative to older adults.
- Normal BMI (21-22).
- No documented diabetes
(glucose in normal range) or
chronic renal impairment [...]

6. Overall Risk Impression
Elevated risk of another
acute myocardial infarction
[...]. Rationale: although
the patient is young and has
favorable metabolic parameters,
the combination of a congenital
coronary anomaly [...]
```

Local-Rubric

(High inference cost)

```
# Task-conditioned local rubric generation prompt

GOAL: Read the patient's text-serialized EHR and
write a compact reasoning trace that characterizes
the patient's risk profile for the following
clinical outcome prediction task: {task_query}

--- START OF EHR DATA ---
{NaiveText_Serialization ($x^{text}$)}
--- END OF EHR DATA ---

Your output MUST follow this exact structure:

1. Patient Snapshot
2. Main Risk Factors
3. Protective Factors
4. What's Unknown / Could Swing the Risk
5. Weighing and Aggregating the Evidence
6. Overall Risk Impression
```

Example task query:
Will the patient develop hypertension in the next year?

Our rubric-transformed text-serializations (Paper)

EHR text / rubric representations

Rubric: A recipe for transforming *naive* text serialization of input into a more structured / standardized representation

Downstream learning: linear heads over text-embeddings (we use Qwen3-Emb-8B)

 # Naive Text Serialization

Patient Demographics
- Patient age: 78, FEMALE [...]

Detailed Past Medical Visits

Inpatient Visit (14 days to pred. time, current visit)

Conditions
- Acute posthemorrhagic anemia
- pH measurement, venous: 7.25, 7.31, 7.31 [...]

Medications
- furosemide 20 MG Oral Tablet
- pantoprazole 20 MG Delayed Release Oral Tablet [...]

Procedures
- Chest x-ray
- Electrocardiogram report [...]

Emergency Room Visit (87 days before prediction time)

Conditions
- Benign essential hypertension
- Chest pain [...]

Medications
- 2ML ondansetron 2MG/ML inject.
- nitroglycerin 0.4 MG [...]

Procedures
- Ct angiography
- Comprehen metabolic panel[...]

NaiveText

 # Local Rubric Representation

1. Patient Snapshot
27 yo hispanic male. Recurrent cardiology visits for congenital anomaly of coronary artery [...]

2. Main Risk Factors
- Congenital coronary artery anomaly (established structural predisposition to myocardial ischemia/infarction).
- Tobacco exposure (smokeless tobacco reported) [...]

3. Protective Factors
- Young age (27) | lower baseline atherosclerotic burden relative to older adults.
- Normal BMI (21–22).
- No documented diabetes (glucose in normal range) or chronic renal impairment [...]

6. Overall Risk Impression
Elevated risk of another acute myocardial infarction [...]. Rationale: although the patient is young and has favorable metabolic parameters, the combination of a congenital coronary anomaly [...]

Local-Rubric

(High inference cost)

 # Global Rubric Representation

3. Demographics
- 55 | FEMALE | [...]

6. Recent Cardiac Symptoms (last 365 days)
- Chest pain/angina: No
- Dyspnea/shortness of breath: Yes (date unknown) [...]

12. Other Relevant Labs
- Creatinine: 1.12 (2023-12-02)
- eGFR: No data [...]

17. Known Risk Factors
- Diabetes mellitus: No (A1c date unknown)
- Hyperlipidemia: Yes
- Family history of premature CAD: Unknown [...]

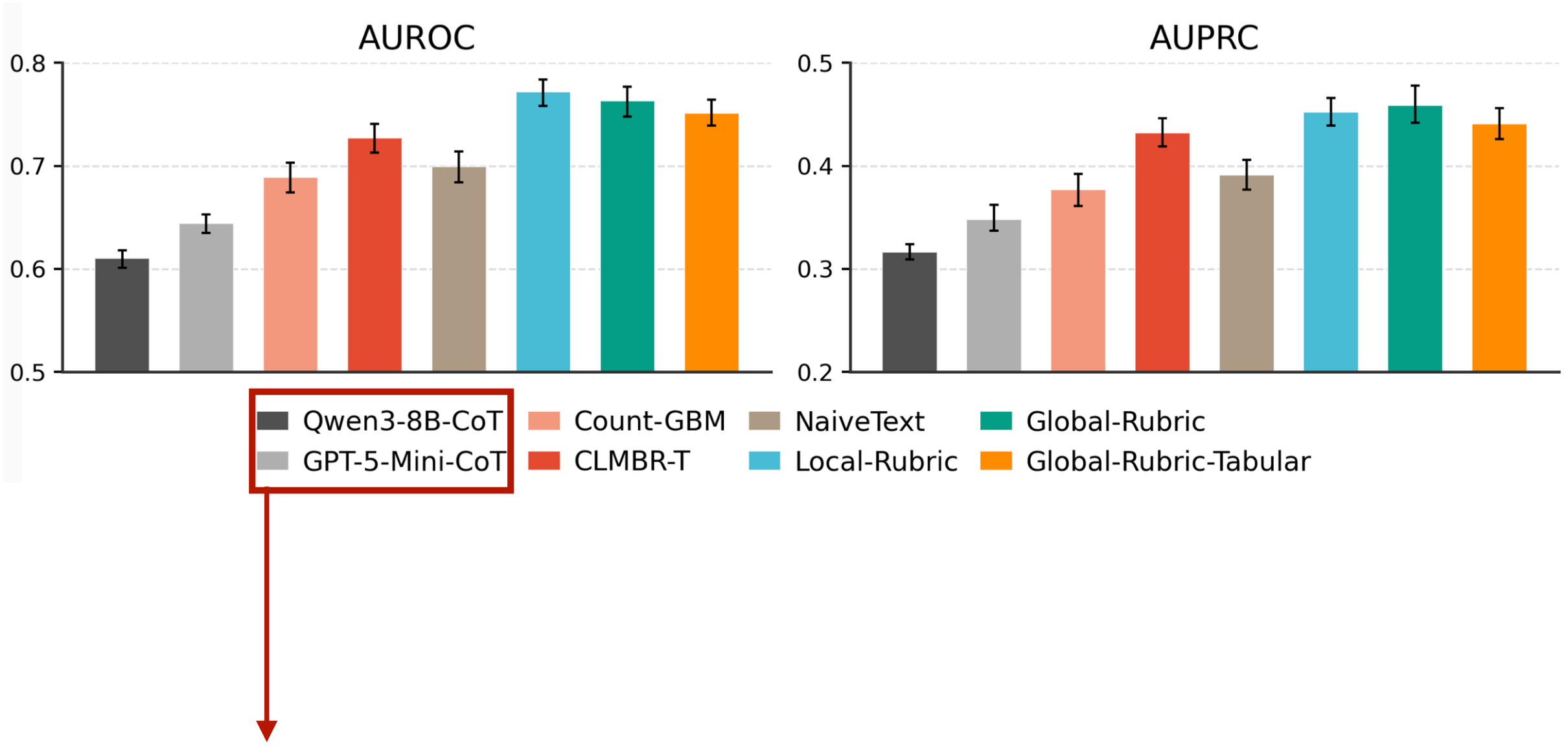
20. Non-cardiac Serious Illness That May Mimic or Alter MI Risk Interpretation
- Active malignancy: No
- Severe infection/sepsis in past 30 days: No
- Major surgery in past 30 days: Yes | multiple inpatient procedures noted in December 2023 (e.g., CPT4/00520 on 2023-12-26 [...]

Global-Rubric

(More standardized / less cost)

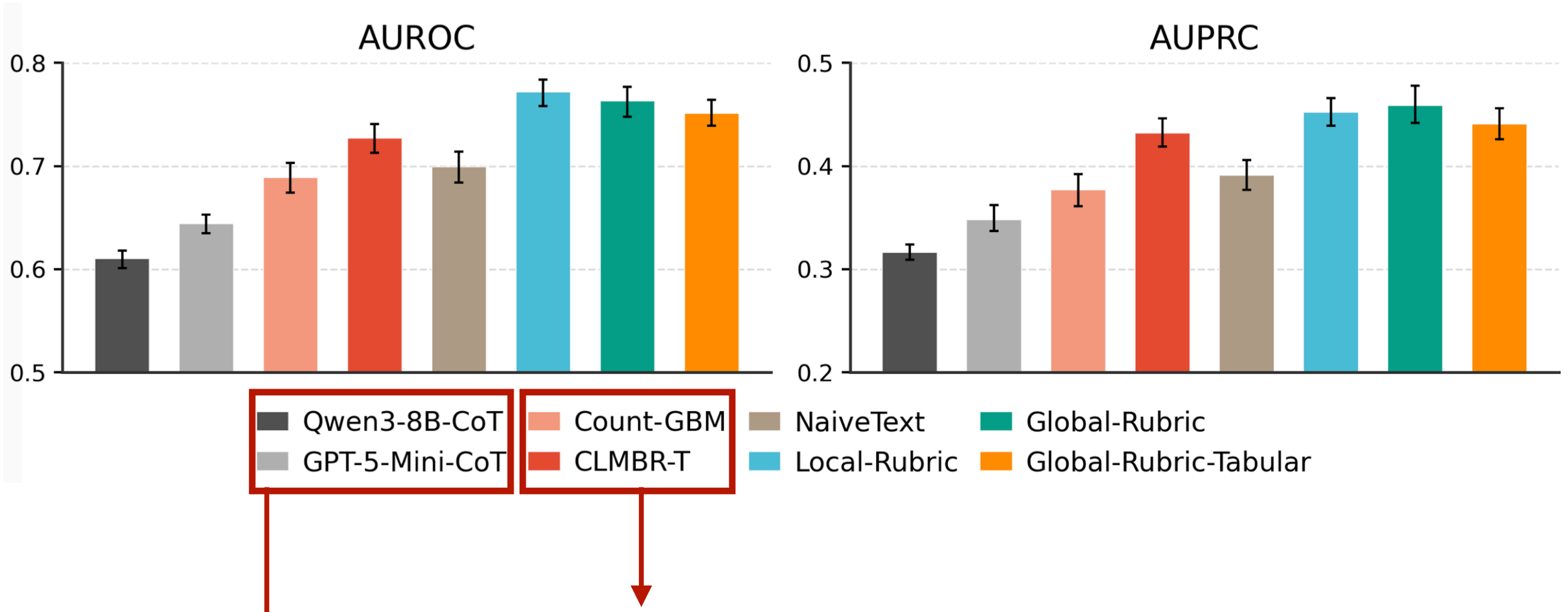
Our rubric-transformed text-serializations (Paper)

Baselines & sneak peek at results



Zero-shot. Prompted with NaiveText serializations to answer “Yes/No”. Probabilities from 10 samples

Baselines & sneak peek at results



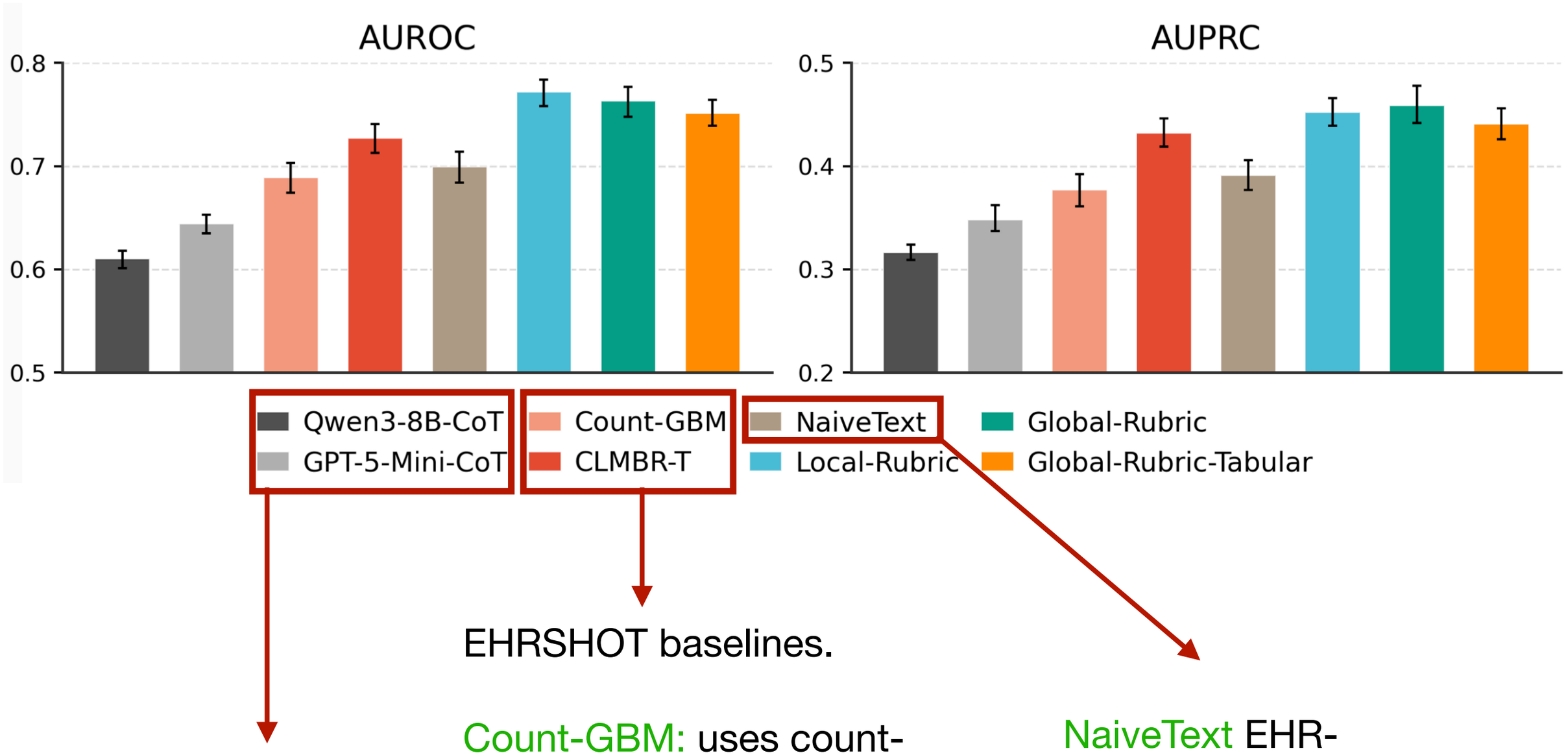
Zero-shot. Prompted with NaiveText serializations to answer “Yes/No”. Probabilities from 10 samples

EHRSHOT baselines.

Count-GBM: uses count-features + boosting machine

CLMBR-T: Autoregressive FM pre-trained on 2.57M in-distribution patient. Embedding + linear head at test-time.

Baselines & sneak peek at results



Zero-shot. Prompted with NaiveText serializations to answer “Yes/No”. Probabilities from 10 samples

EHRSHOT baselines.

Count-GBM: uses count-features + boosting machine

CLMBR-T: Autoregressive FM pre-trained on 2.57M in-distribution patient. Embedding + linear head at test-time.

NaiveText EHR-serialization are used to get text-embeddings and fit linear heads on top.

Hegselmann et al (2025)

How are (global) rubrics learned?

(B) Rubric Synthesis

 # Ask an LLM to synthesize a task-specific rubric.

Create a rubric for predicting hypertension risk in the next year by analyzing data from 40 patients.

List of EHRs (Medoids, x_{text} format):

Pt_1: [78yo, F, HTN meds, SBP=148...]
...
Pt_{40}: [27yo, M, family hx, SBP=129...]

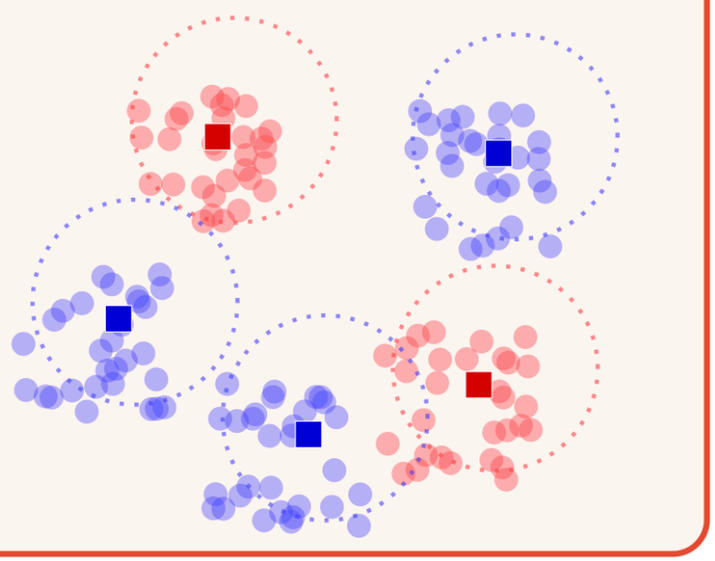
Output a structured rubric.

- Be data-driven [...]
- Be structured and consistent [...]
- Extract facts only [...]
- [...]

How are (global) rubrics learned?

(A) Diverse Cohort Selection

 # Label stratified k-means in text-serialization (x_{text}) embedding space



(B) Rubric Synthesis

 # Ask an LLM to synthesize a task-specific rubric.

Create a rubric for predicting hypertension risk in the next year by analyzing data from 40 patients.

List of EHRs (Medoids, x_{text} format):

Pt_1: [78yo, F, HTN meds, SBP=148...]
...
Pt_{40}: [27yo, M, family hx, SBP=129...]

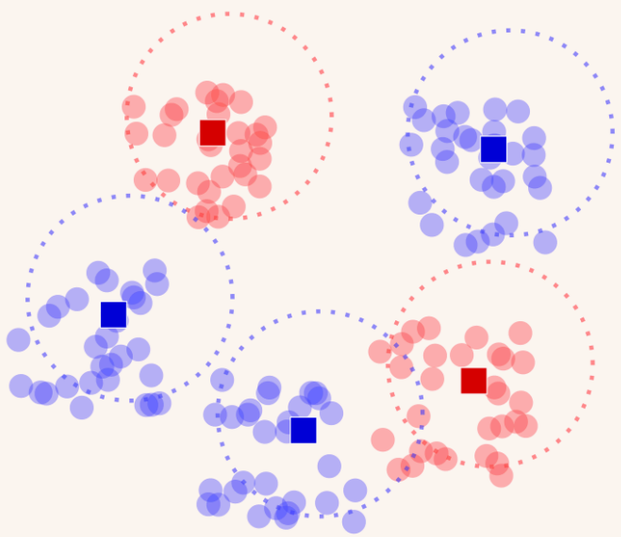
Output a structured rubric.

- Be data-driven [...]
- Be structured and consistent [...]
- Extract facts only [...]
- [...]

How are (global) rubrics learned?

(A) Diverse Cohort Selection

 # Label stratified k-means in text-serialization (x_{text}) embedding space



(B) Rubric Synthesis

 # Ask an LLM to synthesize a task-specific rubric.

Create a rubric for predicting hypertension risk in the next year by analyzing data from 40 patients.

List of EHRs (Medoids, x_{text} format):

Pt_1: [78yo, F, HTN meds, SBP=148...]
...
Pt_{40}: [27yo, M, family hx, SBP=129...]

Output a structured rubric.
- Be data-driven [...]
- Be structured and consistent [...]
- Extract facts only [...]
- [...]

(C) Task-Specific Rubric $\mathcal{R}$

 # LLM-derived rubric $\mathcal{R}$ for transforming x_{text} to x_{rubric}

§1. DEMOGRAPHICS
├ Age, sex, BMI
§2. CV RISK FACTORS
├ BP readings (SBP/DBP)
├ HTN medications
§3. COMORBIDITIES
├ Diabetes, CKD status
§4. TEMPORAL TRENDS
├ BP trajectory (6-12mo)
├ Weight changes
§5. ALERT FLAGS
├ Resistant HTN markers
├ End-organ damage
[...]

$\mathcal{R} : x_{\text{text}} \rightarrow x_{\text{rubric}}$

Full global rubric creation prompt

```
## Task
- Name:  {task_name}
- Query:  {task_query}

## Context
You will be given {40} labeled patient EHR examples ({20} positive, {20} negative).
Another model will later use your rubric to transform new patient EHRs into
structured summaries, which will then serve as input to a supervised classifier.

## What You Must Do
Study the examples below.  Combine what you observe in them with your medical
knowledge to design a rubric template -- a set of named fields that, when filled
in for any patient, produce a structured summary optimized for this prediction task.

The rubric should:
1.  **Be data-driven and discriminative.** Identify which features, patterns, and
interactions actually separate the positive and negative cases.  The rubric should
capture not just obvious indicators but also subtler or compound features you notice.
At the same time, do not overfit to these 40 cases -- use your clinical knowledge to
include factors that are generally relevant even if not prominent in this sample.

2.  **Be structured and consistent.** Every rubricified output must follow the same
field names and order.  For each field, specify what to extract from the EHR and how
to format it.  Specify what to write when data is absent.

3.  **Extract facts only.** The evaluator filling in the rubric must extract and
organize information from the EHR. It must NOT make predictions, assign risk levels,
or draw conclusions.

4.  **Be concise.** The rubric should focus on extracting information that is
relevant to the task.  It should not ask the evaluator to reproduce the entire EHR.

## Positive Examples (Ground Truth:  Yes)
{NaiveText EHR serializations of 20 positive examples concatenated (x_text format)}

## Negative Examples (Ground Truth:  No)
{NaiveText EHR serializations of 20 negative examples concatenated (x_text format)}

## Output
Output ONLY the rubric template itself -- the instructions another model will follow
to transform a patient EHR. No preamble, no explanation of your reasoning.  The
template must be self-contained and directly usable.
```

Global rubric example

 # Excerpt from Global Rubric Instructions ($\mathcal{R}$) for Hypertension Diagnosis Task

---

A. Preparation (before extracting)
1. Define the prediction window: ''next year'' relative to the EHR reference date/time.
2. Define time windows to extract:
- Very recent: last 30 days
- Recent: 31–180 days
- Baseline/remote: >180 days
3. Standardize units and formats:
- Blood pressure: mmHg (systolic/diastolic)
- Weight: kg or oz → convert to kg
- Height: cm or in → convert to meters
[...]

Step 2 – Blood pressure (BP) data extraction and normalization
Extract all systolic/diastolic BP values with timestamps and context (office, inpatient, ED, home, ambulatory, perioperative).
Normalize: remove implausible values (document them), ensure mmHg.
For each time window (very recent, recent, baseline): compute count, mean, median, SD, min, max; identify last BP; flag highest recent BP
Compute simple trend metrics (e.g., recent slope; BP variability via SD).
Categorize BP per ACC/AHA categories using aggregated recent values:
Normal ($<$120/$<$80), Elevated (120–129/$<$80), Stage 1 (130–139 or 80–89), Stage 2 ($\geq$140 or $\geq$90).
[...]

Step 9 – Synthesis per domain (structured fields and scoring)
For each domain, record presence, supporting data, recency, and confidence (High/Moderate/Low).
Domain A – BP phenotype: last BP (date/context), mean recent BP (last 30d; 31–180d), BP category, variability flag, ambulatory/home BP.
Domain B – Metabolic / vascular risk: Diabetes (Y/N) – last A1c (% and date), BMI and obesity category, Hyperlipidemia (Y/N) – LDL value and date, Smoking (current/former/never) Create a simple domain scorecard: number of High/Moderate/Minor risk features.
[...]

Rubric *transformed* input

```
# Excerpt from Global Rubric Instructions (ℛ) for Hyperten
---
A. Preparation (before extracting)
1. Define the prediction window: ''next year'' relative t
2. Define time windows to extract:
- Very recent: last 30 days
- Recent: 31-180 days
- Baseline/remote: >180 days
3. Standardize units and formats:
- Blood pressure: mmHg (systolic/diastolic)
- Weight: kg or oz → convert to kg
- Height: cm or in → convert to meters
[...]

Step 2 - Blood pressure (BP) data extraction and normalizat
Extract all systolic/diastolic BP values with timestamps ar
perioperative).
Normalize: remove implausible values (document them), ensu
For each time window (very recent, recent, baseline): comp
last BP; flag highest recent BP
Compute simple trend metrics (e.g., recent slope; BP variab
Categorize BP per ACC/AHA categories using aggregated recer
Normal (<120/<80), Elevated (120-129/<80), Stage 1 (130-1
[...]

Step 9 - Synthesis per domain (structured fields and scorir
For each domain, record presence, supporting data, recency,
Domain A - BP phenotype: last BP (date/context), mean rece
flag, ambulatory/home BP.
Domain B - Metabolic / vascular risk: Diabetes (Y/N) - las
Hyperlipidemia (Y/N) - LDL value and date, Smoking (current
number of High/Moderate/Minor risk features.
[...]
```

```
# Global Rubric Representation
---
3. Demographics
- 55 | FEMALE | [...]

6. Recent Cardiac Symptoms (last 365 days)
- Chest pain/angina: No
- Dyspnea/shortness of breath: Yes (date unknown) [...]

12. Other Relevant Labs
- Creatinine: 1.12 (2023-12-02)
- eGFR: No data [...]

17. Known Risk Factors
- Diabetes mellitus: No (A1c date unknown)
- Hyperlipidemia: Yes
- Family history of premature CAD: Unknown [...]

20. Non-cardiac Serious Illness That May Mimic or Alter MI Risk Interpretation
- Active malignancy: No
- Severe infection/sepsis in past 30 days: No
- Major surgery in past 30 days: Yes | multiple inpatient procedures noted in December 2023 (e.g., CPT4/00520 on 2023-12-26 [...]
```

How are rubric instructions *applied*?

(D) Rubric Application via LLMs

 # Ask an LLM to apply the rubric transformation $\mathcal{R}$ to each input.

[...]

Rubric $\mathcal{R}$:
{rubric_instructions}

Patient EHR:
{ehr_text (x_{text})}

Fill in every field of the rubric template above using ONLY information from this patient's EHR. Rules:

- Follow the exact field order and section structure of the rubric.
- If data for a field is not present, write "No data".
- [...]

 # Global Rubric Representation

---

3. Demographics
- 55 | FEMALE | [...]

6. Recent Cardiac Symptoms (last 365 days)
- Chest pain/angina: No
- Dyspnea/shortness of breath: Yes (date unknown) [...]

12. Other Relevant Labs
- Creatinine: 1.12 (2023-12-02)
- eGFR: No data [...]

17. Known Risk Factors
- Diabetes mellitus: No (A1c date unknown)
- Hyperlipidemia: Yes
- Family history of premature CAD: Unknown [...]

20. Non-cardiac Serious Illness That May Mimic or Alter MI Risk Interpretation
- Active malignancy: No
- Severe infection/sepsis in past 30 days: No
- Major surgery in past 30 days: Yes | multiple inpatient procedures noted in December 2023 (e.g., CPT4/00520 on 2023-12-26 [...]

Costly in time & money

Same issue with local rubrics (summaries)

Automating rubric application

Naive text ↦ Rubric text

(E) Rubric Application via Parser

 # Ask an LLM to generate a parser script to apply the learned rubric transformation $\mathcal{R}$ to each input.

Write a Python script that reads patient EHR text serializations and fills in a structured clinical rubric template using deterministic string/regex parsing only [...]

Rubric $\mathcal{R}$: {rubric_instructions}

Example EHR text serializations:
{List of medoid pairs: $(x_{\text{text}}, x_{\text{rubric}})$}

The generated script must:

- Use only Python standard libraries such as 're', 'json' [...]
- No LLM API calls, no network requests, no subprocess calls to external tools [...]
- [...]

Automating rubric application

Naive text ↦ Rubric text

(E) Rubric Application via Parser

 # Ask an LLM to generate a parser script to apply the learned rubric transformation $\mathcal{R}$ to each input.

Write a Python script that reads patient EHR text serializations and fills in a structured clinical rubric template using deterministic string/regex parsing only [...]

Rubric $\mathcal{R}$: {rubric_instructions}

Example EHR text serializations: {List of medoid pairs: $(x_{\text{text}}, x_{\text{rubric}})$}

The generated script must:

- Use only Python standard libraries such as 're', 'json' [...]
- No LLM API calls, no network requests, no subprocess calls to external tools [...]
- [...]

```python
def split_into_visit_blocks(text: str) -> List[Dict[str, Any]]:
    """
    Returns list of blocks: {setting, date (date), date_str, text, idx}
    """
```

```python
def get_demographics(text: str) -> Dict[str, str]:
    out = {"age": "NA", "sex": "NA", "race": "NA", "ethnicity": "NA"}
    m_age = re.search(r"Patient age:\s*(\d{1,3})\b", text)
```

 # Naive Text Serialization

Patient Demographics

- Patient age: 78, FEMALE [...]

Detailed Past Medical Visits

Inpatient Visit (14 days to pred. time, current visit)

Conditions

- Acute posthemorrhagic anemia
- pH measurement, venous: 7.25, 7.31, 7.31 [...]

Medications

- furosemide 20 MG Oral Tablet
- pantoprazole 20 MG Delayed Release Oral Tablet [...]

Procedures

- Chest x-ray
- Electrocardiogram report [...]

Emergency Room Visit (87 days before prediction time)

Conditions

- Benign essential hypertension
- Chest pain [...]

Medications

- 2ML ondansetron 2MG/ML inject.
- nitroglycerin 0.4 MG [...]

Procedures

- Ct angiography
- Comprehen metabolic panel[...]

Automating rubric application

Naive text ↦ Rubric text

(E) Rubric Application via Parser

 # Ask an LLM to generate a parser script to apply the learned rubric transformation $\mathcal{R}$ to each input.

Write a Python script that reads patient EHR text serializations and fills in a structured clinical rubric template using deterministic string/regex parsing only [...]

Rubric $\mathcal{R}$: {rubric_instructions}

Example EHR text serializations: {List of medoid pairs: $(x_{\text{text}}, x_{\text{rubric}})$}

The generated script must:

- Use only Python standard libraries such as 're', 'json' [...]
- No LLM API calls, no network requests, no subprocess calls to external tools [...]
- [...]

```python
def split_into_visit_blocks(text: str) -> List[Dict[str, Any]]:
    """
    Returns list of blocks: {setting, date (date), date_str, text, idx}
    """
```

```python
def get_demographics(text: str) -> Dict[str, str]:
    out = {"age": "NA", "sex": "NA", "race": "NA", "ethnicity": "NA"}
    m_age = re.search(r"Patient age:\s*(\d{1,3})\b", text)
```

```python
def tobacco_status_and_notes(text: str) -> str:
    # Determine status
    t_lower = text.lower()
    status = "unknown"
    if re.search(r"\bformer smoker\b", t_lower) \
        or re.search(r"\bex-?smoker\b", t_lower):
        status = "former smoker"
    elif re.search(r"\bnever smoker\b", t_lower) \
        or re.search(r"\bnever\b.*\bsmok", t_lower):
        status = "never"
    else:
        # In this dataset, presence of "Tobacco user"
        # often interpreted as current smoker.
        if re.search(r"^\s*-\s*Tobacco user\b", text, flags=re.M
            or re.search(r"\bTobacco user\b", text):
            status = "current smoker"
        elif re.search(r"\bTobacco use and exposure\b", text):
            status = "unknown"
```

```python
# ----------------------------
# Rubric filling
# ----------------------------

RUBRIC_FIELDS = [
    "PATIENT_ID",
    "REFERENCE_DATE",
    "DEMOGRAPHICS",
    "TOBACCO_USE",
    "RECENT_BMI",
    "RECENT_WEIGHT_HEIGHT",
    "BP_RECENT_3",
    "BP_COUNTS_OVER_TIME",
    "HISTORY_OF_HYPERTENSION",
    "ANTIHYPERTENSIVE_MEDICATIONS_CURRENT",
```

Automating rubric application

DEMOGRAPHICS:
Age: 44; Sex: F; Race: White; Ethnicity: Not Hispanic or Latino

TOBACCO_USE:
current smoker; (Tobacco user); (Chewed tobacco consumption: N, N); (Cigarette consumption: N, N)

RECENT_BMI:
BMI = 22.96 (2023-12-21, inpatient)

RECENT_WEIGHT_HEIGHT:
Weight = 2346 (ounces) (2023-12-21); Height = 67 (in) (2023-12-21)

BP_RECENT_3:
1) 105/74 mmHg (2023-12-21, inpatient); 2) NA; 3) NA

BP_COUNTS_OVER_TIME:
365d_>=140/90 = 0; 365d_>=130/80 = 0; 90d_>=140/90 = 0; 90d_>=130/80 = 0

HISTORY_OF_HYPERTENSION: NA

ANTIHYPERTENSIVE_MEDICATIONS_CURRENT:
hydrochlorothiazide — Thiazide diuretic — START: 2023-12-21 — STOP: Active
lisinopril — ACE inhibitor — START: 2023-12-21 — STOP: Active

ANTIHYPERTENSIVE_MEDICATIONS_RECENT_CHANGES:
hydrochlorothiazide — started — 2023-12-21
lisinopril — started — 2023-12-21

OTHER_MEDICATIONS_AFFECTING_BP: NA

RELEVANT_COMORBIDITIES:
- Diabetes mellitus (Type 1 / Type 2): Type 1
- Chronic kidney disease (CKD) or End-stage renal disease (ESRD) / on dialysis: Yes
- Renal transplant: No
- Coronary artery disease / prior MI: No
- Renovascular hypertension or hypertensive renal failure: No
-...

Synthetic example for rubric representation generated by script

```
# ------------------------------
# Rubric filling
# ------------------------------

RUBRIC_FIELDS = [
    "PATIENT_ID",
    "REFERENCE_DATE",
    "DEMOGRAPHICS",
    "TOBACCO_USE",
    "RECENT_BMI",
    "RECENT_WEIGHT_HEIGHT",
    "BP_RECENT_3",
    "BP_COUNTS_OVER_TIME",
    "HISTORY_OF_HYPERTENSION",
    "ANTIHYPERTENSIVE_MEDICATIONS_CURRENT",
```

Tabularizing rubrics

Naive text ↦ Rubric text

(E) Rubric Application via Parser

 # Ask an LLM to generate a parser script to apply the learned rubric transformation $\mathcal{R}$ to each input.

Write a Python script that reads patient EHR text serializations and fills in a structured clinical rubric template using deterministic string/regex parsing only [...]

Rubric $\mathcal{R}$: {rubric_instructions}

Example EHR text serializations:
{List of medoid pairs: $(x_{\text{text}}, x_{\text{rubric}})$}

The generated script must:

- Use only Python standard libraries such as 're', 'json' [...]
- No LLM API calls, no network requests, no subprocess calls to external tools [...]
- [...]

Rubric text ↦ Tabular features

(F) Rubric Tabularization

 # Ask an LLM to generate a script to transform x_{rubric} to tabular features based on $\mathcal{R}$.

Write a Python script to convert rubric-formatted patient EHRs into numeric feature vectors [...]

Example rubric-transformed EHR serializations:
{List of medoids in x_{rubric} format, obtained from x_{text} using parser in Panel (E)}

Your logic must:

- General: handle any value the rubric parser could plausibly produce [...]
- Robust: gracefully handle missing values [...]
- [...]

Tabularizing rubrics

Rubric text ↦ Tabular features

(F) Rubric Tabularization

 # Ask an LLM to generate a script to transform x_{rubric} to tabular features based on $\mathcal{R}$.

Write a Python script to convert rubric-formatted patient EHRs into numeric feature vectors [...]

Example rubric-transformed EHR serializations:
{List of medoids in x_{rubric} format, obtained from x_{text} using parser in Panel (E)}

Your logic must:

- General: handle any value the rubric parser could plausibly produce [...]
- Robust: gracefully handle missing values [...]
- [...]

Tabularizing rubrics

Rubric text ↦ Tabular features

(F) Rubric Tabularization

 # Ask an LLM to generate a script to transform x_{rubric} to tabular features based on $\mathcal{R}$.

Write a Python script to convert rubric-formatted patient EHRs into numeric feature vectors [...]

Example rubric-transformed EHR serializations:
{List of medoids in x_{rubric} format, obtained from x_{text} using parser in Panel (E)}

Your logic must:

- General: handle any value the rubric parser could plausibly produce [...]
- Robust: gracefully handle missing values [...]
- [...]

```
age_years (numeric)
age_missing (binary)
sex_missing (binary)
sex__M (categorical)
sex__F (categorical)
sex__Other (categorical)
ethnicity_missing (binary)
```

```
tobacco_status_missing (binary)
tobacco_status__never (categorical)
tobacco_status__former smoker (categorical)
tobacco_status__current smoker (categorical)
tobacco_status__unknown (categorical)
tobacco_supporting_mentions_count (numeric)
tobacco_supporting_mentions_missing (binary)
```

```
antihtn_current_count (numeric)
antihtn_current_missing (binary)
antihtn_class_current__ace_inhibitor (binary)
antihtn_class_current__arb (binary)
antihtn_class_current__beta_blocker (binary)
antihtn_class_current__calcium_channel_blocker (binary)
```

```
bp_affecting_class__sympathomimetic_vasopressor (binary)
bp_affecting_class__oral_contraceptive_progestin (binary)
bp_affecting_class__stimulant_can_raise_bp (binary)
bp_affecting_class__vegf_inhibitor_antineoplastic_can_raise_bp
comorb__diabetes_mellitus_type_1_type_2__yes (binary)
comorb__diabetes_mellitus_type_1_type_2__missing (binary)
```

275 features for HTN

Tabularizing rubrics

Rubric text ↦ Tabular features

(F) Rubric Tabularization

 # Ask an LLM to generate a script to transform x_{rubric} to tabular features based on $\mathcal{R}$.

Write a Python script to convert rubric-formatted patient EHRs into numeric feature vectors [...]

Example rubric-transformed EHR serializations:
{List of medoids in x_{rubric} format, obtained from x_{text} using parser in Panel (E)}

Your logic must:

- General: handle any value the rubric parser could plausibly produce [...]
- Robust: gracefully handle missing values [...]
- [...]

```
age_years (numeric)
age_missing (binary)
sex_missing (binary)
sex__M (categorical)
sex__F (categorical)
sex__Other (categorical)
ethnicity_missing (binary)
```

```
tobacco_status_missing (binary)
tobacco_status__never (categorical)
tobacco_status__former smoker (categorical)
tobacco_status__current smoker (categorical)
tobacco_status__unknown (categorical)
tobacco_supporting_mentions_count (numeric)
tobacco_supporting_mentions_missing (binary)
```

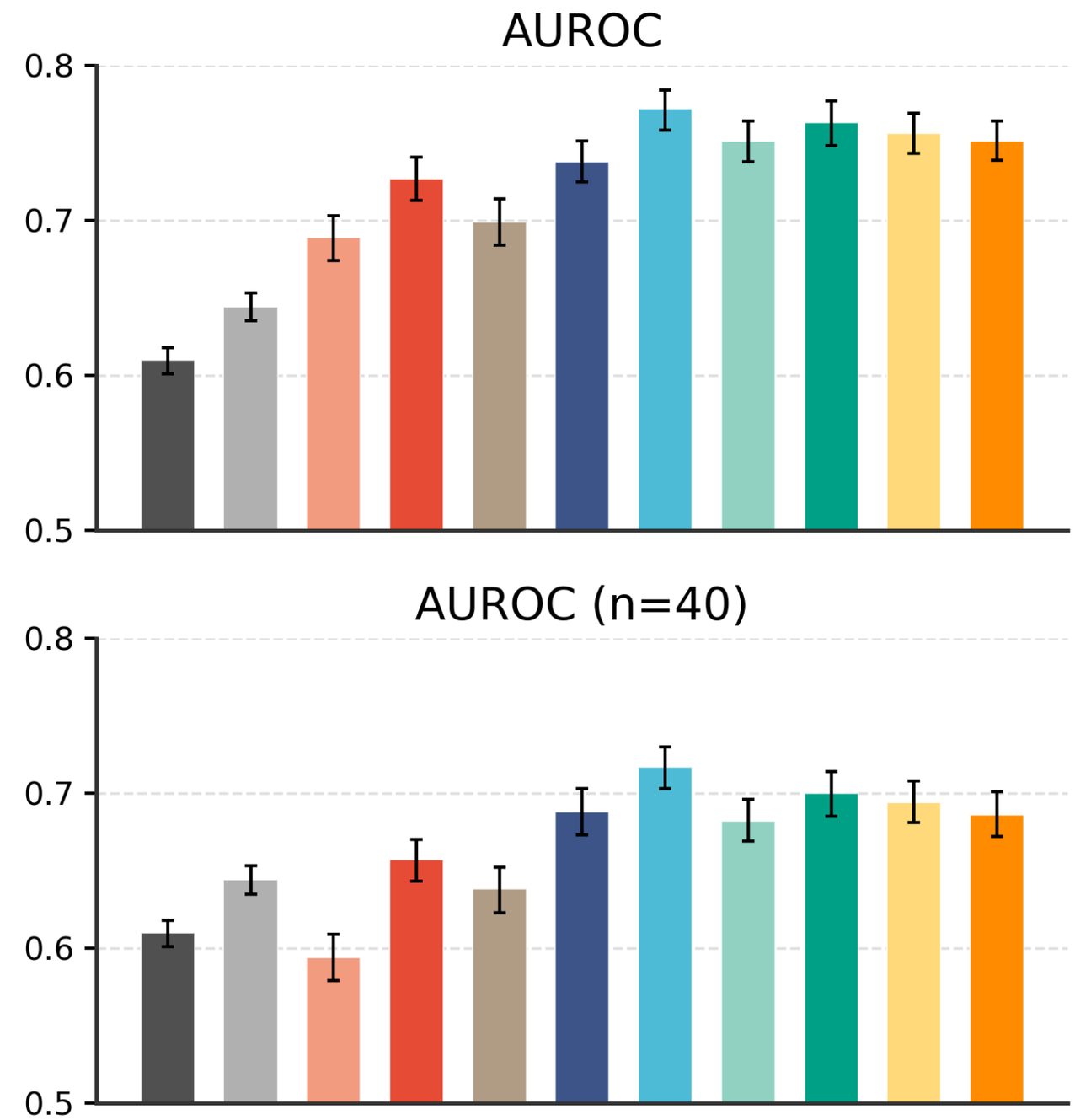
```
antihtn_current_count (numeric)
antihtn_current_missing (binary)
antihtn_class_current__ace_inhibitor (binary)
antihtn_class_current__arb (binary)
antihtn_class_current__beta_blocker (binary)
antihtn_class_current__calcium_channel_blocker (binary)
```

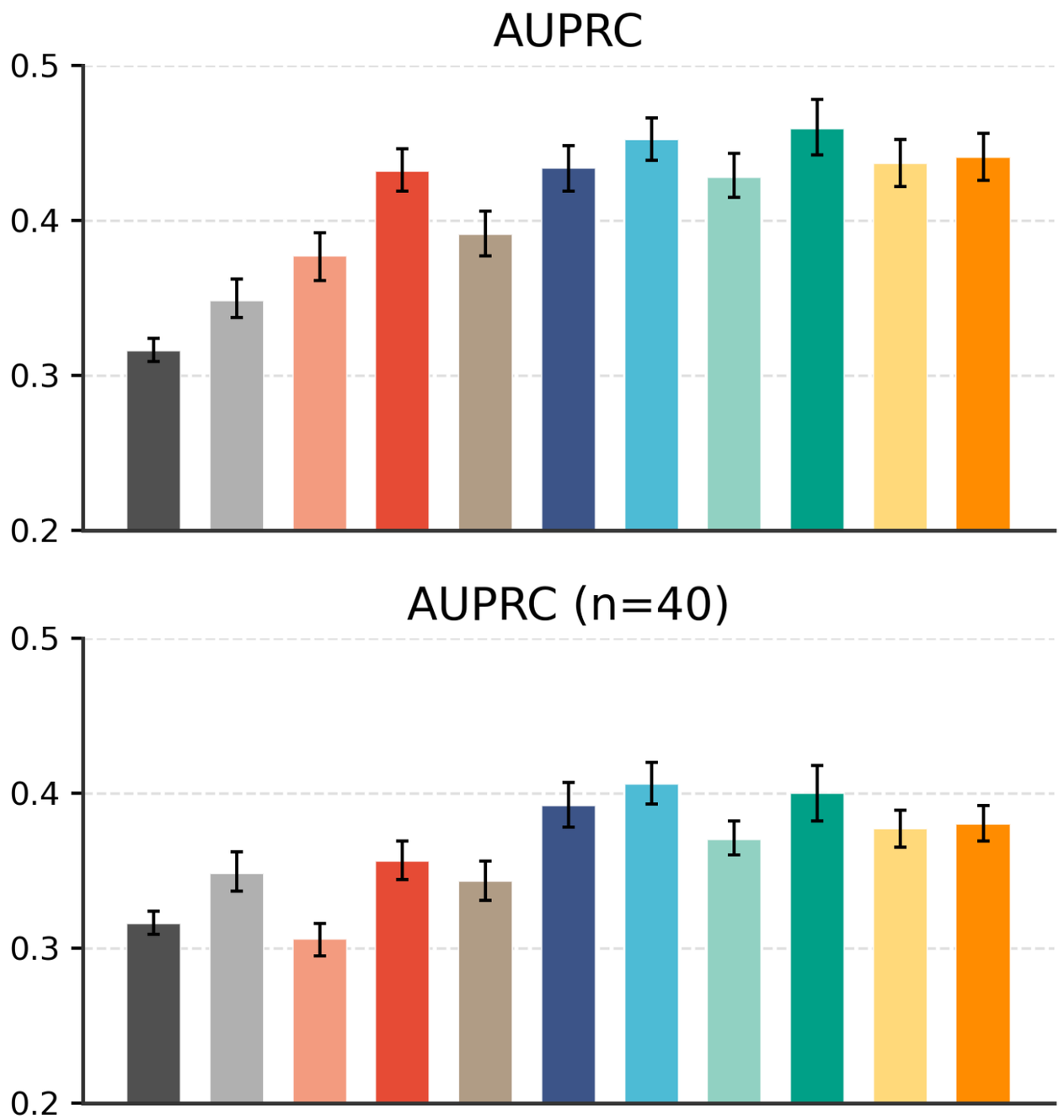
```
bp_affecting_class__sympathomimetic_vasopressor (binary)
bp_affecting_class__oral_contraceptive_progestin (binary)
bp_affecting_class__stimulant_can_raise_bp (binary)
bp_affecting_class__vegf_inhibitor_antineoplastic_can_raise_bp
comorb__diabetes_mellitus_type_1_type_2__yes (binary)
comorb__diabetes_mellitus_type_1_type_2__missing (binary)
```

275 features for HTN

Tabular representations automatically enable a whole suite of ML tools (causal, interp., etc)

Overall results & ablations





GPT5-Mini-CoT | Qwen3-8B-CoT | Count-GBM | CLMBR-T | NaiveText

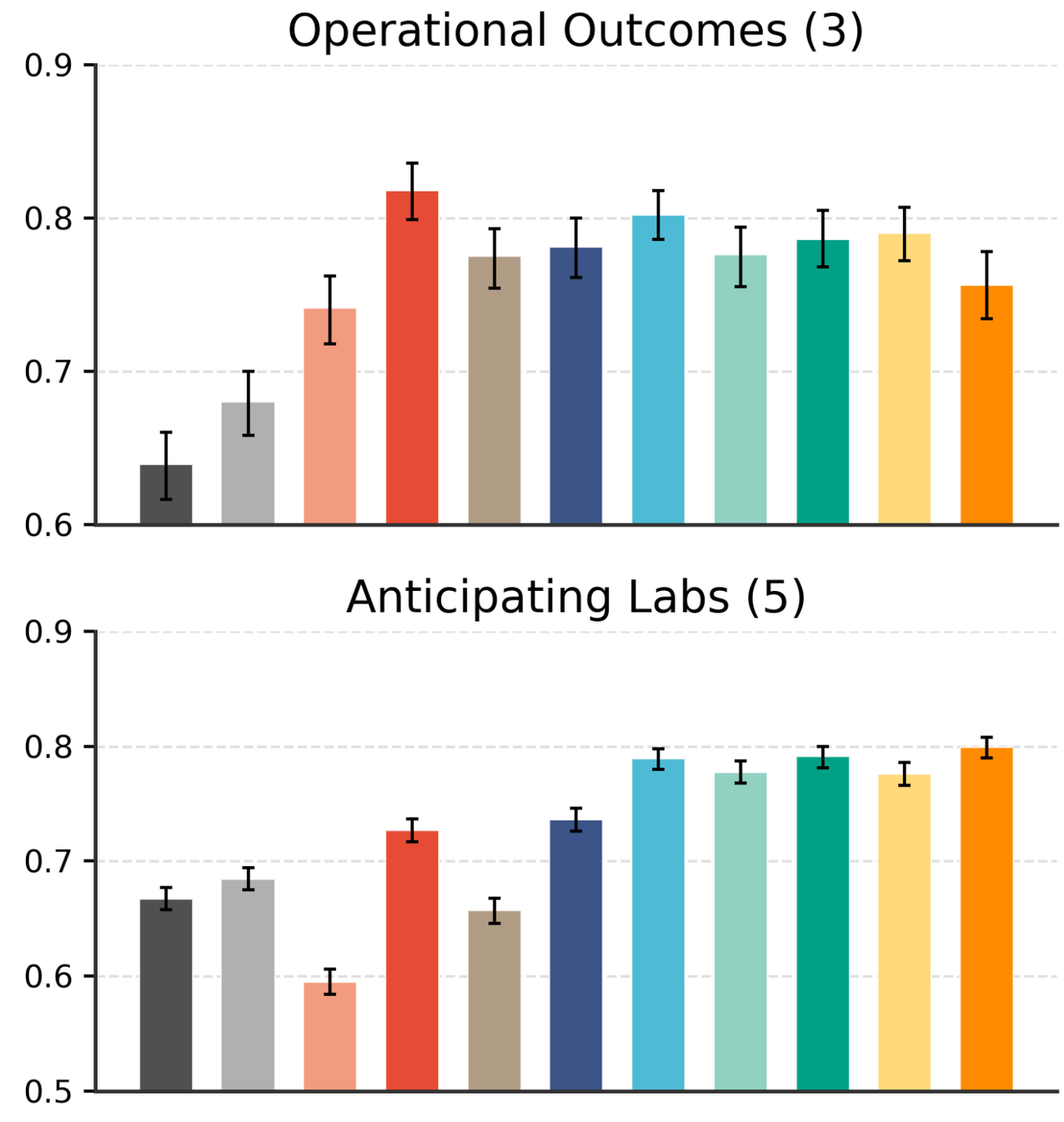
Local-Rubric-Generic | Local-Rubric | Global-Rubric-Blind | Global-Rubric | Global-Rubric-Auto | Global-Rubric-Tabular

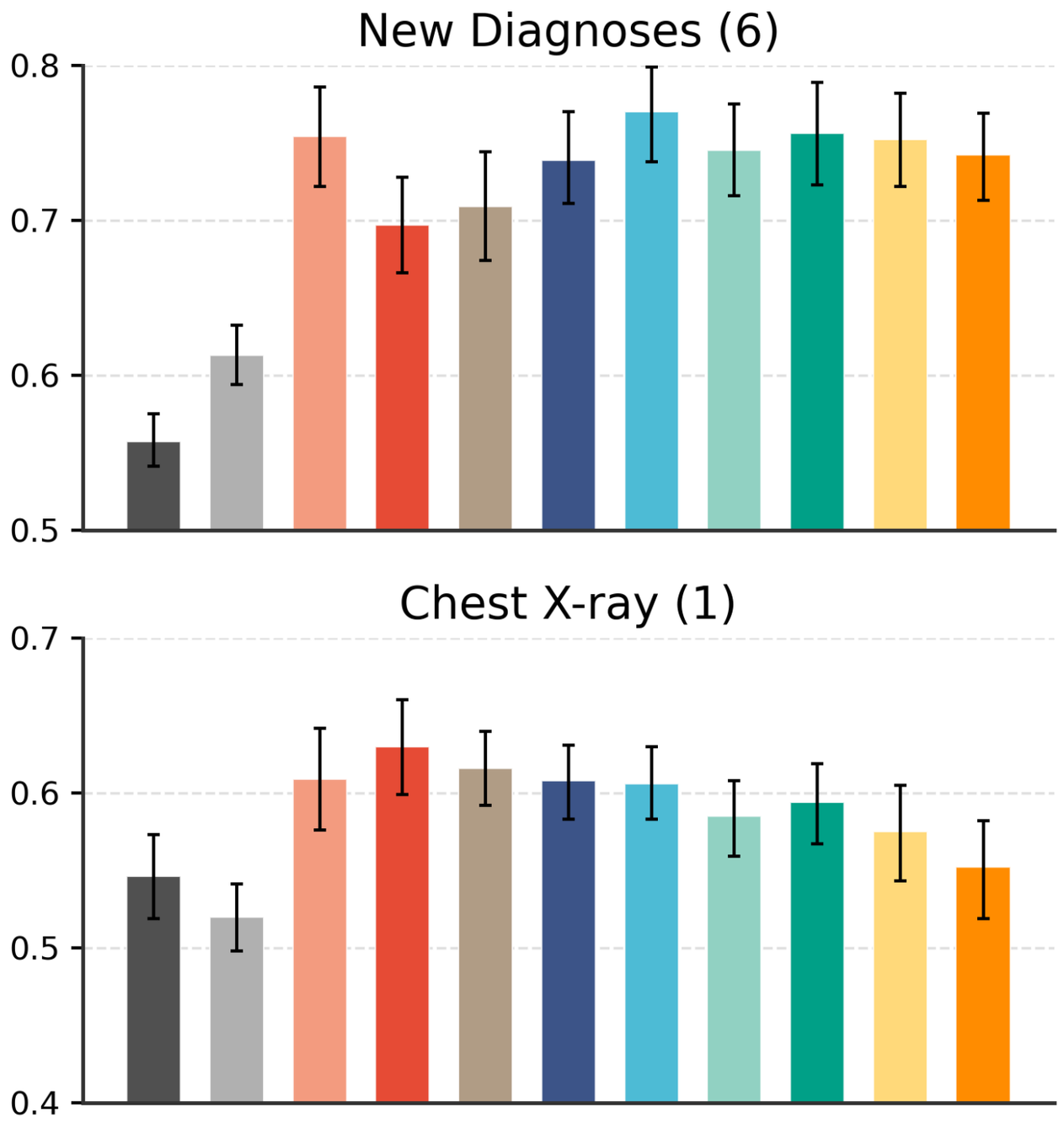
Generic clinical summaries, instead of task-conditioned

Rubrics created without inspecting any examples

Auto. rubric application & tabularization in prev. slides

Breakdown by task groups





GPT5-Mini-CoT · Qwen3-8B-CoT · Count-GBM · CLMBR-T · NaiveText
Local-Rubric-Generic · Local-Rubric · Global-Rubric-Blind · Global-Rubric · Global-Rubric-Auto · Global-Rubric-Tabular

(AUROC, n=All)

Future work & limitations

- More evaluation with different benchmarks
  - e.g., MIMIC using clinical notes
  - non-health datasets
- Rubric improvement
  - currently one-shot
  - how to use > 40 patients (current bottleneck is context size)
- An algorithm to learn the rubric from all patients?
- Learning from failures in the natural language space
  - Exploring if LLMs can help learning long-tails better, in a broad sense