Clinical Trial Patient Matching with LLMs

Michael Wornow

October 29, 2024

Bio

- **Michael Wornow** (mwornow@stanford.edu)
- 5th year computer science PhD student @ Stanford
- Advised by Nigam Shah and Chris Ré
- Research: Developing foundation models for healthcare, with the goal of improving patient outcomes and hospital operations

Talk Outline

1. **Problem**
    a. What is clinical trial patient recruitment, and why is it hard?
2. **Prior Work**
    a. What did people try before LLMs?
3. **Papers**
    a. Zero-shot patient matching with off-the-shelf LLMs
    b. PRISM: Fine tuning an LLM for clinical trial matching
4. **Future Work**

Talk Outline

1. **Problem**
    a. What is clinical trial patient recruitment, and why is it hard?
2. **Prior Work**
    a. What did people try before LLMs?
3. **Papers**
    a. Zero-shot patient matching with off-the-shelf LLMs
    b. PRISM: Fine tuning an LLM for clinical trial matching
4. **Future Work**

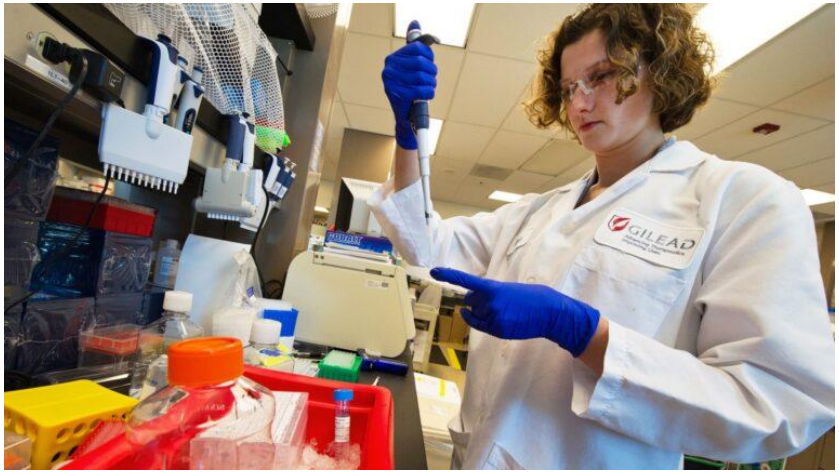
Background

The drug discovery process + clinical trials

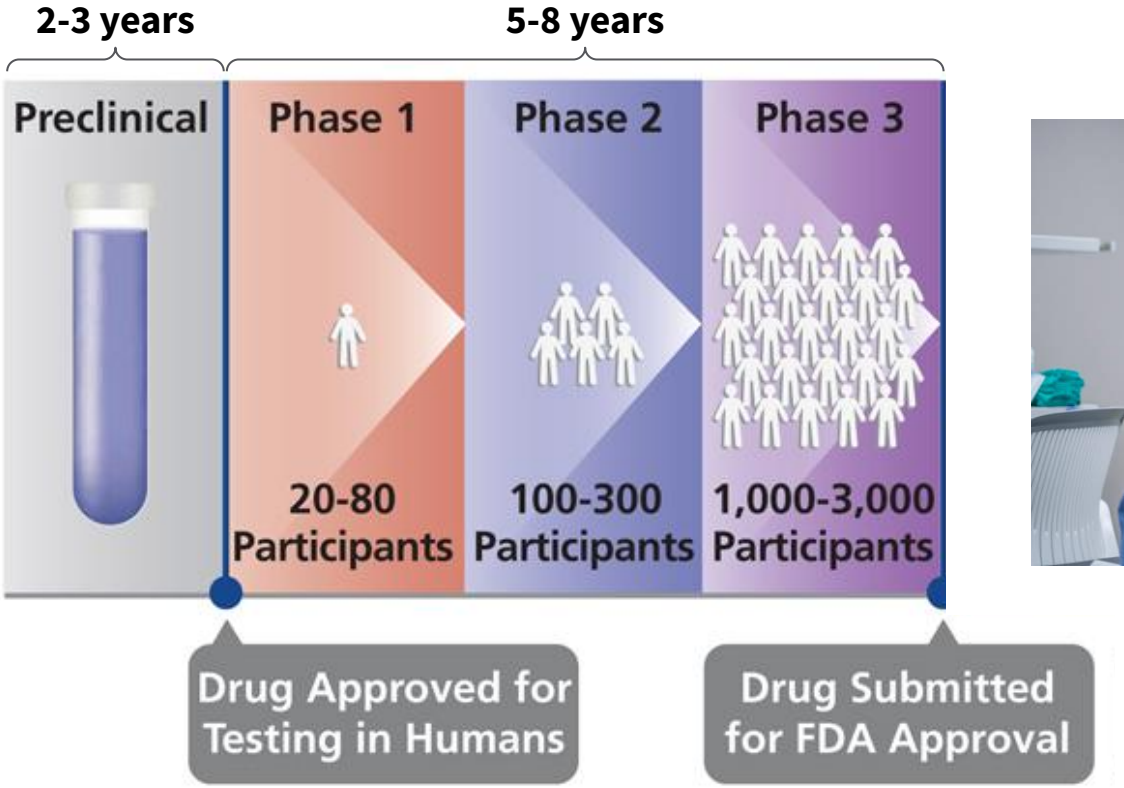
A drug must pass **3 phases** of **clinical trials** before reaching patients

A drug must pass **3 phases** of **clinical trials** before reaching patients

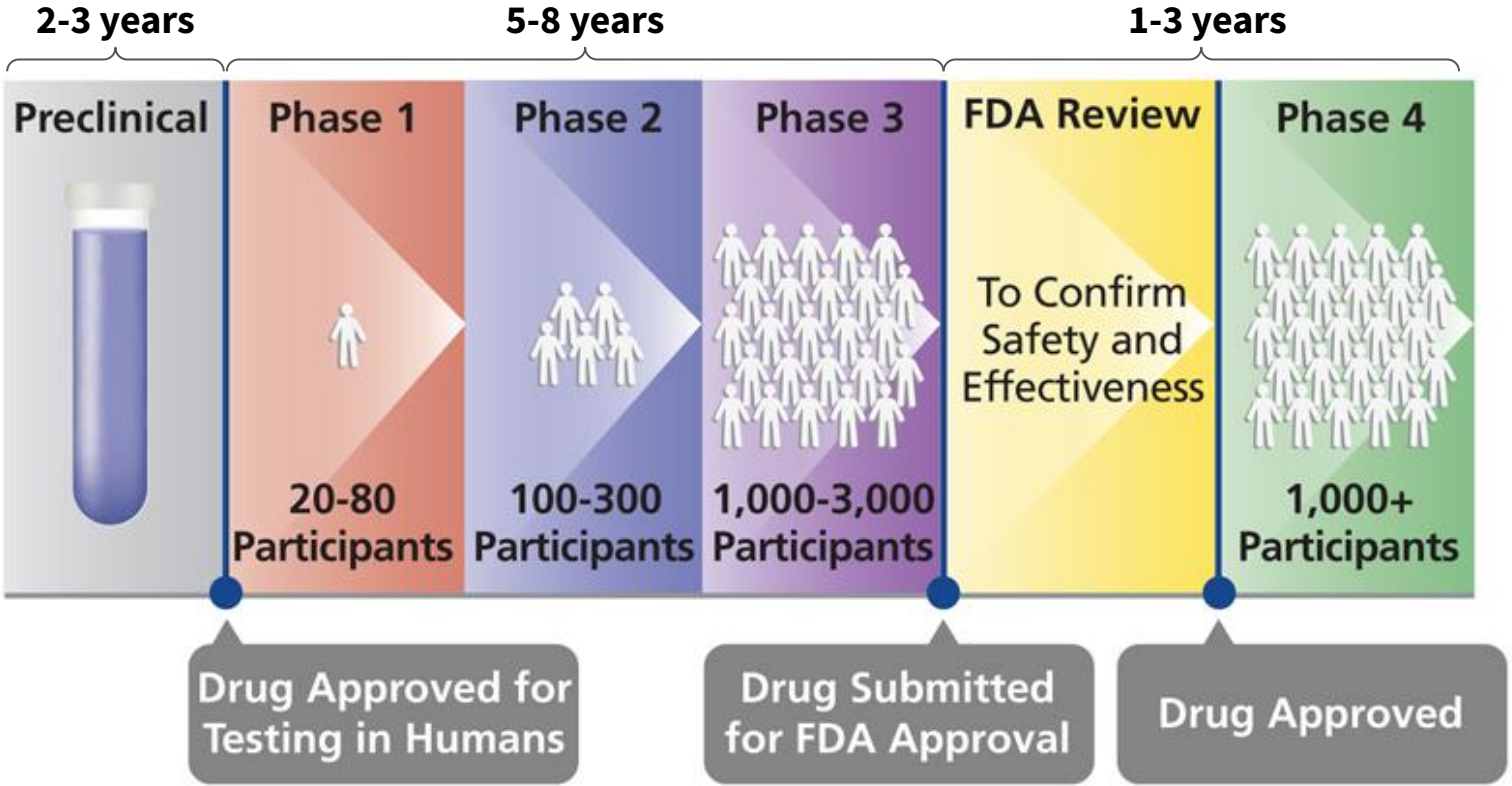
2-3 years



A drug must pass **3 phases** of **clinical trials** before reaching patients

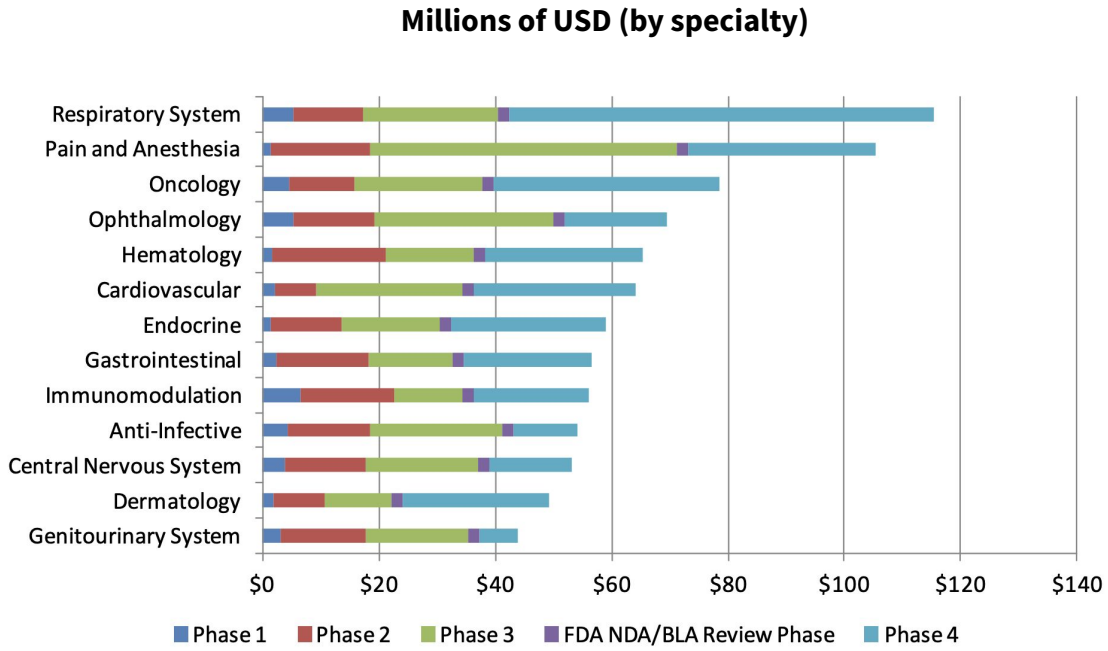


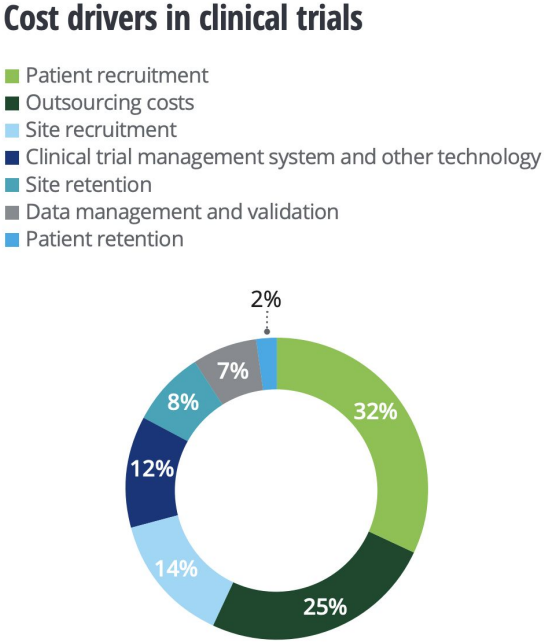
A drug must pass **3 phases** of **clinical trials** before reaching patients



https://ocra.fsu.edu/study-support-services/clinicaltrialsgov-administration/navigating-clinical-trials/

Clinical trials are **long** and **expensive**

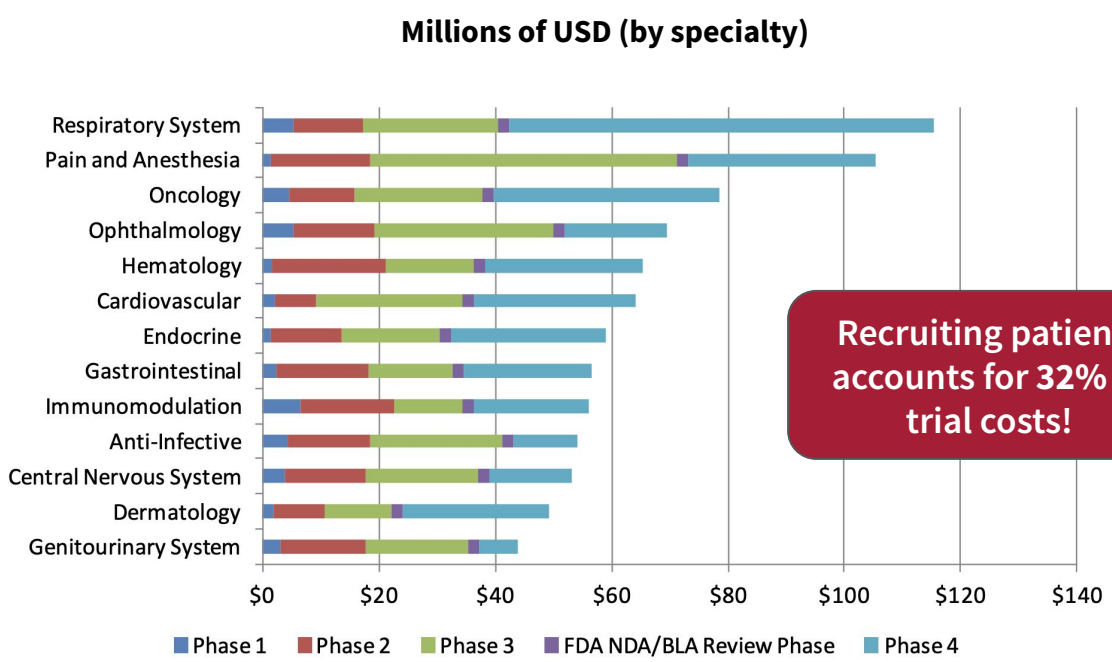


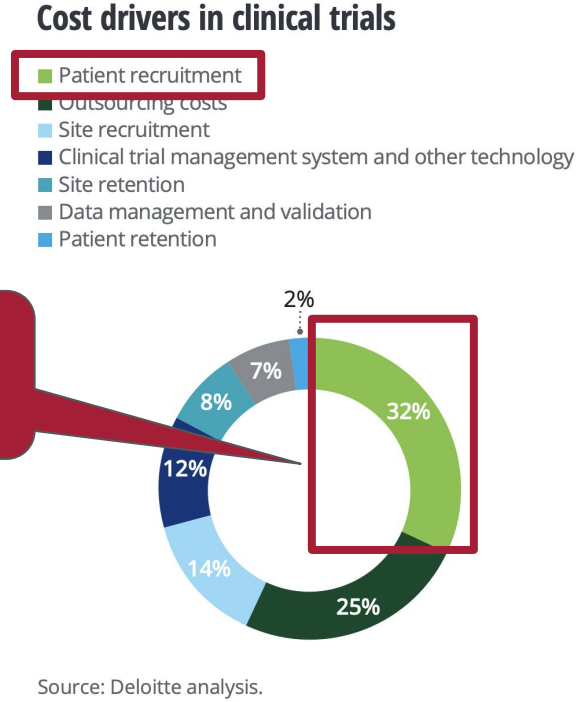


Source: Deloitte analysis.

https://aspe.hhs.gov/sites/default/files/private/pdf/77166/rpt_erg.pdf
https://www2.deloitte.com/content/dam/insights/us/articles/22934_intelligent-clinical-trials/DI_Intelligent-clinical-trials.pdf

Clinical trials are **long** and **expensive**





Recruiting patients accounts for 32% of trial costs!

Source: Deloitte analysis.

https://aspe.hhs.gov/sites/default/files/private/pdf/77166/rpt_erg.pdf
https://www2.deloitte.com/content/dam/insights/us/articles/22934_intelligent-clinical-trials/DI_Intelligent-clinical-trials.pdf

Problem

Finding eligible patients for clinical trials is hard

Finding patients to enroll in clinical trials is an unsolved **challenge**

Finding patients to enroll in clinical trials is an unsolved **challenge**

1 in 3

Clinical trials **outright fails**
due to a lack of patients

Finding patients to enroll in clinical trials is an unsolved **challenge**

1 in 3

Clinical trials **outright fails** due to a lack of patients

80%

of trials **are delayed 1+ months** due to slow patient enrollment

Finding patients to enroll in clinical trials is an unsolved **challenge**

1 in 3

Clinical trials **outright fails** due to a lack of patients

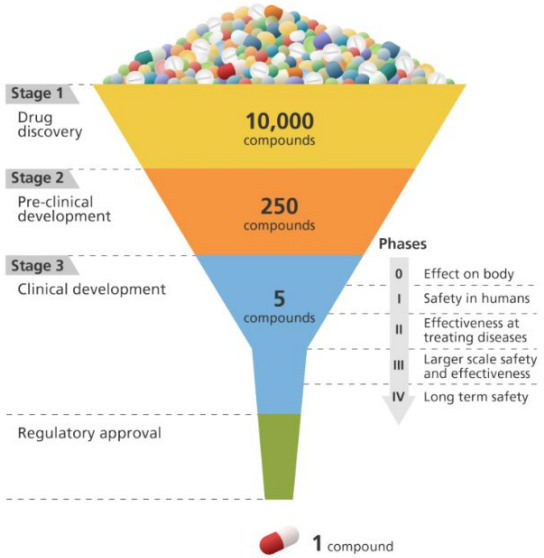
80%

of trials **are delayed 1+ months** due to slow patient enrollment

94%

Of patients who are **eligible** and interested in a trial **are never considered**

Why should you care? Faster recruitment means…



More Successful Drug Trials

More Diverse Trials

Improved Patient Access to Therapies

Why is recruitment difficult?

Why is recruitment difficult?

Trial Protocol

Assessing the Impact of Lipoprotein (a) Lowering With TQJ230 on Major Cardiovascular Events in Patients With CVD (Lp(a)HORIZON)

Eligibility Criteria ICMJE

Key Inclusion Criteria

- Lp(a) ≥ 70 mg/dL at the screening visit, measured at the Central laboratory
- Myocardial infarction: ≥ 3 months from screening and randomization to ≤ 10 years prior to the screening visit
- Ischemic stroke: ≥ 3 months from screening and randomization to ≤ 10 years prior to the screening visit
- Clinically significant symptomatic peripheral artery disease

Why is recruitment difficult?

Trial Protocol

Assessing the Impact of Lipoprotein (a) Lowering With TQJ230 on Major Cardiovascular Events in Patients With CVD (Lp(a)HORIZON)

Eligibility Criteria ICMJE

Key Inclusion Criteria

- Lp(a) ≥ 70 mg/dL at the screening visit, measured at the Central laboratory
- Myocardial infarction: ≥ 3 months from screening and randomization to ≤ 10 years prior to the screening visit
- Ischemic stroke: ≥ 3 months from screening and randomization to ≤ 10 years prior to the screening visit
- Clinically significant symptomatic peripheral artery disease

EHR at a hospital

Structured Data:
Age, Diagnosis, Medications, ...

Unstructured Text:

History of Present Illness:
54 year old female with recent diagnosis of ulcerative colitis on 6-mercaptopurine, prednisone 40-60 mg daily, who presents with a new onset of headache and neck stiffness. The patient is in distress, rigoring and has aphasia and only limited history is obtained. She reports that she was awaken 1AM the morning of [**2147-11-16**] with a headache which she describes as bandlike. She states that headaches are unusual for her. She denies photo- or phonophobia. She did have neck stiffness. On arrival to the ED at 5:33PM, she was afebrile with a temp of 96.5, however she later spiked with temp to 104.4 (rectal), HR 91, BP 112/54, RR 24, O2 sat 100 %. Head CT was done and relealved attenuation within the subcortical white matter of the right medial frontal lobe. LP was performed showing opening pressure 24 cm H2O WBC of

Why is recruitment difficult?

Trial Protocol

Assessing the Impact of Lipoprotein (a) Lowering With TQJ230 on Major Cardiovascular Events in Patients With CVD (Lp(a)HORIZON)

Eligibility Criteria ICMJE

Key Inclusion Criteria

- Lp(a) ≥ 70 mg/dL at the screening visit, measured at the Central laboratory
- Myocardial infarction: ≥ 3 months from screening and randomization to ≤ 10 years prior to the screening visit
- Ischemic stroke: ≥ 3 months from screening and randomization to ≤ 10 years prior to the screening visit
- Clinically significant symptomatic peripheral artery disease

EHR at a hospital

Structured Data:
Age, Diagnosis, Medications, ...

Unstructured Text:

History of Present Illness:
54 year old female with recent diagnosis of ulcerative colitis on 6-mercaptopurine, prednisone 40-60 mg daily, who presents with a new onset of headache and neck stiffness. The patient is in distress, rigoring and has aphasia and only limited history is obtained. She reports that she was awaken 1AM the morning of [**2147-11-16**] with a headache which she describes as bandlike. She states that headaches are unusual for her. She denies photo- or phonophobia. She did have neck stiffness. On arrival to the ED at 5:33PM, she was afebrile with a temp of 96.5, however she later spiked with temp to 104.4 (rectal), HR 91, BP 112/54, RR 24, O2 sat 100 %. Head CT was done and relealved attenuation within the subcortical white matter of the right medial frontal lobe. LP was performed showing opening pressure 24 cm H2O WBC of

80%

Of relevant data is **unstructured**, making **automation difficult**

Why is recruitment difficult?

Trial Protocol

Assessing the Impact of Lipoprotein (a) Lowering With TQJ230 on Major Cardiovascular Events in Patients With CVD (Lp(a)HORIZON)

Eligibility Criteria ICMJE

Key Inclusion Criteria

- Lp(a) ≥ 70 mg/dL at the screening visit, measured at the Central laboratory
- Myocardial infarction: ≥ 3 months from screening and randomization to ≤ 10 years prior to the screening visit
- Ischemic stroke: ≥ 3 months from screening and randomization to ≤ 10 years prior to the screening visit
- Clinically significant symptomatic peripheral artery disease

"For every trial, we manually review **1000s of records**."

- Stanford Research Coordinator

Clinical research coordinator

EHR at a hospital

Structured Data:
Age, Diagnosis, Medications, ...

Unstructured Text:

History of Present Illness:
54 year old female with recent diagnosis of ulcerative colitis on 6-mercaptopurine, prednisone 40-60 mg daily, who presents with a new onset of headache and neck stiffness. The patient is in distress, rigoring and has aphasia and only limited history is obtained. She reports that she was awaken 1AM the morning of [**2147-11-16**] with a headache which she describes as bandlike. She states that headaches are unusual for her. She denies photo- or phonophobia. She did have neck stiffness. On arrival to the ED at 5:33PM, she was afebrile with a temp of 96.5, however she later spiked with temp to 104.4 (rectal), HR 91, BP 112/54, RR 24, O2 sat 100 %. Head CT was done and relealved attenuation within the subcortical white matter of the right medial frontal lobe. LP was performed showing opening pressure 24 cm H2O WBC of

80%

Of relevant data is **unstructured**, making **automation difficult**

Why is recruitment difficult?

Trial Protocol

Assessing the Impact of Lipoprotein (a) Lowering With TQJ230 on Major Cardiovascular Events in Patients With CVD (Lp(a)HORIZON)

Eligibility Criteria ICMJE

Key Inclusion Criteria

- Lp(a) ≥ 70 mg/dL at the screening visit, measured at the Central laboratory
- Myocardial infarction: ≥ 3 months from screening and randomization to ≤ 10 years prior to the screening visit
- Ischemic stroke: ≥ 3 months from screening and randomization to ≤ 10 years prior to the screening visit
- Clinically significant symptomatic peripheral artery disease

"For every trial, we manually review **1000s of records**."

- Stanford Research Coordinator

Clinical research coordinator

EHR at a hospital

Structured Data:
Age, Diagnosis, Medications, ...

Unstructured Text:

History of Present Illness:
54 year old female with recent diagnosis of ulcerative colitis on 6-mercaptopurine, prednisone 40-60 mg daily, who presents with a new onset of headache and neck stiffness. The patient is in distress, rigoring and has aphasia and only limited history is obtained. She reports that she was awaken 1AM the morning of [**2147-11-16**] with a headache which she describes as bandlike. She states that headaches are unusual for her. She denies photo- or phonophobia. She did have neck stiffness. On arrival to the ED at 5:33PM, she was afebrile with a temp of 96.5, however she later spiked with temp to 104.4 (rectal), HR 91, BP 112/54, RR 24, O2 sat 100 %. Head CT was done and relealved attenuation within the subcortical white matter of the right medial frontal lobe. LP was performed showing opening pressure 24 cm H2O WBC of

80%

Of relevant data is **unstructured**, making **automation difficult**

10-40 min

Time to review a **single EHR**

Why is recruitment difficult?

Trial Protocol

Assessing the Impact of Lipoprotein (a) Lowering With TQJ230 on Major Cardiovascular Events in Patients With CVD (Lp(a)HORIZON)

Eligibility Criteria ICMJE

Key Inclusion Criteria

- Lp(a) ≥ 70 mg/dL at the screening visit, measured at the Central laboratory
- Myocardial infarction: ≥ 3 months from screening and randomization to ≤ 10 years prior to the screening visit
- Ischemic stroke: ≥ 3 months from screening and randomization to ≤ 10 years prior to the screening visit
- Clinically significant symptomatic peripheral artery disease

"For every trial, we manually review **1000s of records**."

- Stanford Research Coordinator

Clinical research coordinator

EHR at a hospital

Structured Data:
Age, Diagnosis, Medications, ...

Unstructured Text:

History of Present Illness:
54 year old female with recent diagnosis of ulcerative colitis on 6-mercaptopurine, prednisone 40-60 mg daily, who presents with a new onset of headache and neck stiffness. The patient is in distress, rigoring and has aphasia and only limited history is obtained. She reports that she was awaken 1AM the morning of [**2147-11-16**] with a headache which she describes as bandlike. She states that headaches are unusual for her. She denies photo- or phonophobia. She did have neck stiffness. On arrival to the ED at 5:33PM, she was afebrile with a temp of 96.5, however she later spiked with temp to 104.4 (rectal), HR 91, BP 112/54, RR 24, O2 sat 100 %. Head CT was done and relealved attenuation within the subcortical white matter of the right medial frontal lobe. LP was performed showing opening pressure 24 cm H2O WBC of

80%

Of relevant data is **unstructured**, making **automation difficult**

10-40 min

Time to review a **single EHR**

"I spend **over 80% of my time manually reading EHR** for eligibility"

- Stanford Research Coordinator

Goal

Speed up patient recruitment by **automatically matching** eligible patients to trials

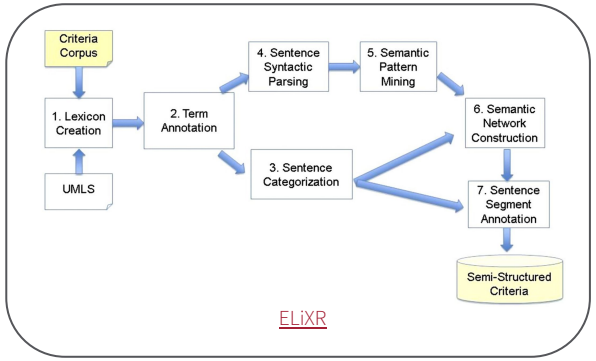
Talk Outline

1. **Motivation**
    a. What is clinical trial patient recruitment, and why is it hard?
2. **Prior Work**
    a. What did people try before LLMs?
3. **Papers**
    a. Zero-shot patient matching with off-the-shelf LLMs
    b. PRISM: Fine tuning an LLM for clinical trial matching
4. **Future Work**

Many people have tried solving this problem

Many people have tried solving this problem

Rule-based



ELiXR

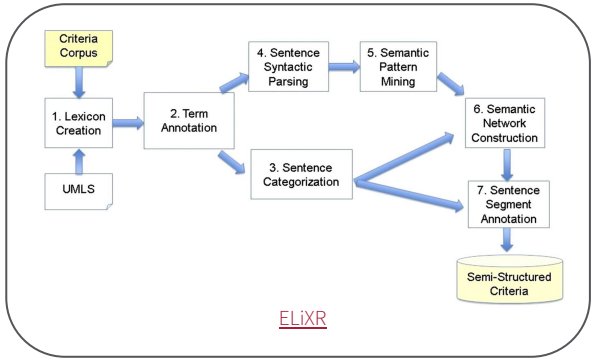
```
public class Abdominal extends BaseClassifiable {
    private static final List<Pattern> POSITIVE_MARKERS = new ArrayList<>();

    static {
        POSITIVE_MARKERS.add(Pattern.compile("bowel surgery", Pattern.CASE_INSENSITIVE));
        //POSITIVE_MARKERS.add(Pattern.compile("polypectomy", Pattern.CASE_INSENSITIVE)); // Disabled by @kasac
//        POSITIVE_MARKERS.add(Pattern.compile("resection", Pattern.CASE_INSENSITIVE));   // Disabled by @kasac
        POSITIVE_MARKERS.add(Pattern.compile("splenectomy", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("intestine resection", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("intestinal resection", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("bowel resection", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("hysterectomy", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("liver transplant", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("pancreatectomy", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("liver surgery", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("gastric resection", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("gastrectomy", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("hepatectomy", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("appendectomy", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("colostomy", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("cholecystectomy", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("colectomy", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("TAH"));   // 18 times
```

n2c2 RBC

Many people have tried solving this problem

Rule-based



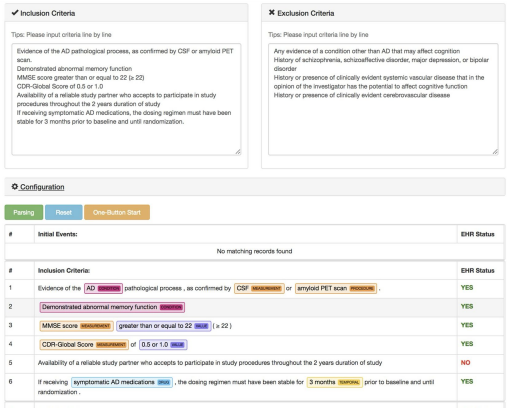
ELiXR

```
public class Abdominal extends BaseClassifiable {
    private static final List<Pattern> POSITIVE_MARKERS = new ArrayList<>();

    static {
        POSITIVE_MARKERS.add(Pattern.compile("bowel surgery", Pattern.CASE_INSENSITIVE));
        //POSITIVE_MARKERS.add(Pattern.compile("polypectomy", Pattern.CASE_INSENSITIVE)); // Disabled by @kasac
//        POSITIVE_MARKERS.add(Pattern.compile("resection", Pattern.CASE_INSENSITIVE));   // Disabled by @kasac
        POSITIVE_MARKERS.add(Pattern.compile("splenectomy", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("intestine resection", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("intestinal resection", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("bowel resection", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("hysterectomy", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("liver transplant", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("pancreatectomy", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("liver surgery", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("gastric resection", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("gastrectomy", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("hepatectomy", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("appendectomy", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("colostomy", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("cholecystectomy", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("colectomy", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("TAH"));   // 18 times
```

n2c2 RBC

Text to Query via NLP



Criteria2Query

Example 1	- Has a known history of HIV , multiple or severe drug allergies , or severe post-treatment hypersensitivity reactions .

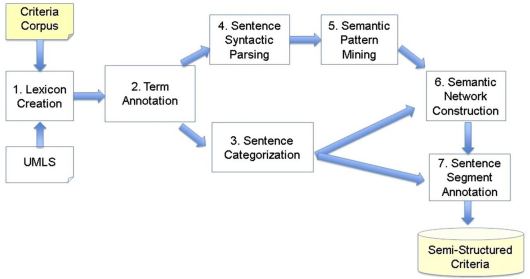
ELiIE

```
<sent>
    <text>Has a known history of HIV , multiple or severe drug allergies , or severe post-treatment
hypersensitivity reactions .</text>
    <entity class="Condition" encoding="4008881" index="T1" negation="N" relation="None" start="5" >
HIV </entity>
    <attribute class="Qualifier" index="T2" start="7"> multiple </attribute>
    <attribute class="Qualifier" index="T3" start="9"> severe </attribute>
    <entity class="Condition" encoding="439224" index="T4" negation="N" relation="T2:modified_by|
T3:modified_by" start="10"> drug allergies </entity>
    <attribute class="Qualifier" index="T5" start="14"> severe </attribute>
    <entity class="Condition" encoding="43021226" index="T6" negation="N" relation="T5:modified_by"
start="15"> post-treatment hypersensitivity reactions </entity>
</sent>
```

ELiIE

Many people have tried solving this problem

Rule-based



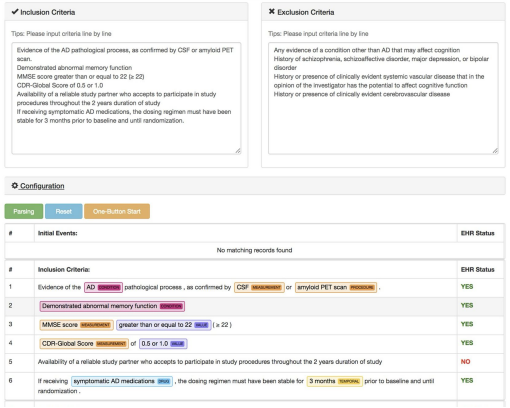
ELiXR

```
public class Abdominal extends BaseClassifiable {
    private static final List<Pattern> POSITIVE_MARKERS = new ArrayList<>();

    static {
        POSITIVE_MARKERS.add(Pattern.compile("bowel surgery", Pattern.CASE_INSENSITIVE));
        //POSITIVE_MARKERS.add(Pattern.compile("polypectomy", Pattern.CASE_INSENSITIVE)); // Disabled by @kasac
//          POSITIVE_MARKERS.add(Pattern.compile("resection", Pattern.CASE_INSENSITIVE));   // Disabled by @kasac
        POSITIVE_MARKERS.add(Pattern.compile("splenectomy", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("intestine resection", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("intestinal resection", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("bowel resection", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("hysterectomy", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("liver transplant", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("pancreatectomy", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("liver surgery", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("gastric resection", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("gastrectomy", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("hepatectomy", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("appendectomy", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("colostomy", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("cholecystectomy", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("colectomy", Pattern.CASE_INSENSITIVE));
        POSITIVE_MARKERS.add(Pattern.compile("TAH"));   // 18 times
```

n2c2 RBC

Text to Query via NLP



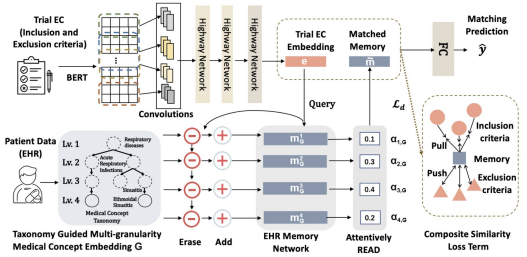
Criteria2Query

Example 1	- Has a known history of HIV , multiple or severe drug allergies , or severe post-treatment hypersensitivity reactions .

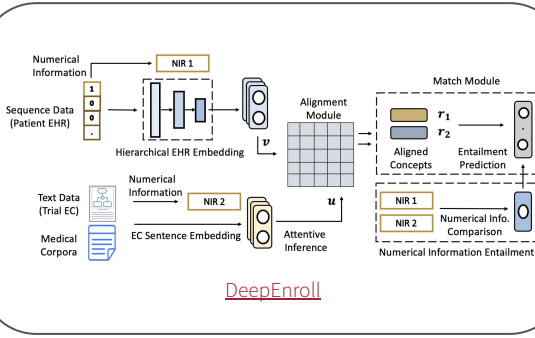
```
<sent>
    <text>Has a known history of HIV , multiple or severe drug allergies , or severe post-treatment
hypersensitivity reactions .</text>
    <entity class="Condition" encoding="4008881" index="T1" negation="N" relation="None" start="5" >
HIV </entity>
    <attribute class="Qualifier" index="T2" start="7"> multiple </attribute>
    <attribute class="Qualifier" index="T3" start="9"> severe </attribute>
    <entity class="Condition" encoding="439224" index="T4" negation="N" relation="T2:modified_by|
T3:modified_by" start="10"> drug allergies </entity>
    <attribute class="Qualifier" index="T5" start="14"> severe </attribute>
    <entity class="Condition" encoding="43021226" index="T6" negation="N" relation="T5:modified_by"
start="15"> post-treatment hypersensitivity reactions </entity>
</sent>
```

EliIE

End-to-End Deep Learning



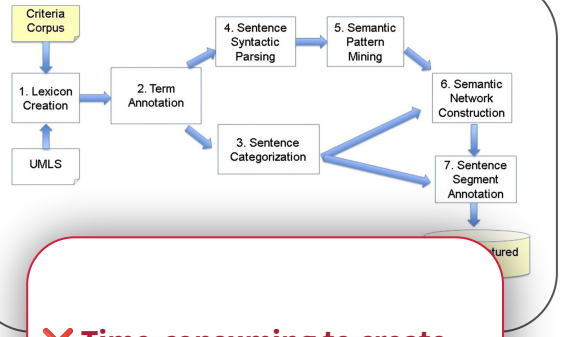
COMPOSE



DeepEnroll

But clinical text has proven difficult to model

Rule-based



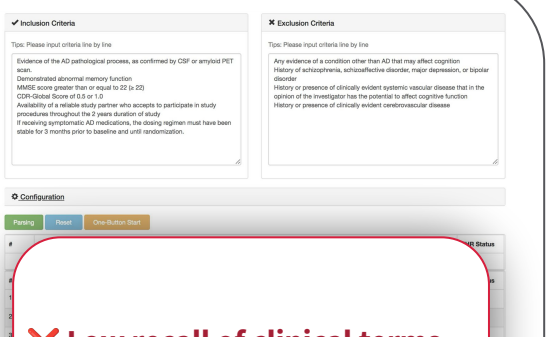
❌ Time-consuming to create

❌ Can't adapt to new trials

```
POSITIVE_MARKERS.add(Pattern.compile("bowel resection", Pattern.CASE_INSENSITIVE));
POSITIVE_MARKERS.add(Pattern.compile("hysterectomy", Pattern.CASE_INSENSITIVE));
POSITIVE_MARKERS.add(Pattern.compile("liver transplant", Pattern.CASE_INSENSITIVE));
POSITIVE_MARKERS.add(Pattern.compile("pancreatectomy", Pattern.CASE_INSENSITIVE));
POSITIVE_MARKERS.add(Pattern.compile("liver surgery", Pattern.CASE_INSENSITIVE));
POSITIVE_MARKERS.add(Pattern.compile("gastric resection", Pattern.CASE_INSENSITIVE));
POSITIVE_MARKERS.add(Pattern.compile("gastrectomy", Pattern.CASE_INSENSITIVE));
POSITIVE_MARKERS.add(Pattern.compile("hepatectomy", Pattern.CASE_INSENSITIVE));
POSITIVE_MARKERS.add(Pattern.compile("appendectomy", Pattern.CASE_INSENSITIVE));
POSITIVE_MARKERS.add(Pattern.compile("colostomy", Pattern.CASE_INSENSITIVE));
POSITIVE_MARKERS.add(Pattern.compile("cholecystectomy", Pattern.CASE_INSENSITIVE));
POSITIVE_MARKERS.add(Pattern.compile("colectomy", Pattern.CASE_INSENSITIVE));
POSITIVE_MARKERS.add(Pattern.compile("TAH"));   // 18 times
```

n2c2 RBC

Text to Query via NLP



❌ Low recall of clinical terms

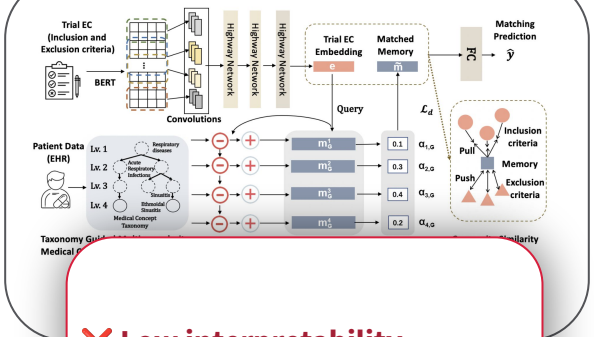
❌ Ignores EHR clinical text

Example 1	**- Has a known history of HIV , multiple or severe drug allergies , or severe post-treatment hypersensitivity reactions .**

```
<sent>
    <text>Has a known history of HIV , multiple or severe drug allergies , or severe post-treatment
hypersensitivity reactions .</text>
        <entity class="Condition" encoding="4008881" index="T1" negation="N" relation="None" start="5" >
HIV </entity>
    <attribute class="Qualifier" index="T2" start="7"> multiple </attribute>
    <attribute class="Qualifier" index="T3" start="9"> severe </attribute>
    <entity class="Condition" encoding="439224" index="T4" negation="N" relation="T2:modified_by|
T3:modified_by" start="10"> drug allergies </entity>
    <attribute class="Qualifier" index="T5" start="14"> severe </attribute>
    <entity class="Condition" encoding="43021226" index="T6" negation="N" relation="T5:modified_by"
start="15"> post-treatment hypersensitivity reactions </entity>
</sent>
```

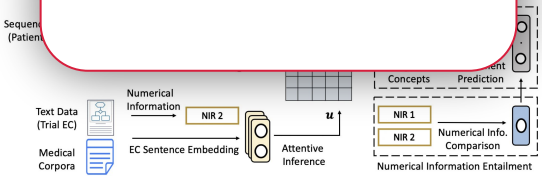
EliIE

End-to-End Deep Learning



❌ Low interpretability

❌ Ignores EHR clinical text



DeepEnroll

Goal

Speed up patient recruitment by **automatically matching** eligible patients to trials

Proposal

Speed up patient recruitment by **automatically matching** eligible patients to trials **with LLMs!**

Talk Outline

1. **Motivation**
    a. What is clinical trial patient recruitment, and why is it hard?
2. **Prior Work**
    a. What did people try before LLMs?
3. **Papers**
    a. Zero-shot patient matching with off-the-shelf LLMs
    b. PRISM: Fine tuning an LLM for clinical trial matching
4. **Future Work**

In this talk, we'll cover two LLM-based approaches

Use the most powerful general-purpose LLM to match patients to trials

Zero-Shot Clinical Trial Patient Matching with LLMs

Michael Wornow* MWORNOW@STANFORD.EDU
Alejandro Lozano* LOZANOE@STANFORD.EDU
Dev Dash DEV@STANFORD.EDU
Jenelle Jindal JJINDAL@STANFORD.EDU
Kenneth W. Mahaffey KMAHAF@STANFORD.EDU
Nigam H. Shah NSHAH@STANFORD.EDU
Stanford University, Palo Alto, CA, USA

In this talk, we'll cover two LLM-based approaches

Use the most powerful general-purpose LLM to match patients to trials

Zero-Shot Clinical Trial Patient Matching with LLMs

Michael Wornow* MWORNOW@STANFORD.EDU
Alejandro Lozano* LOZANOE@STANFORD.EDU
Dev Dash DEV@STANFORD.EDU
Jenelle Jindal JJINDAL@STANFORD.EDU
Kenneth W. Mahaffey KMAHAF@STANFORD.EDU
Nigam H. Shah NSHAH@STANFORD.EDU
Stanford University, Palo Alto, CA, USA

Finetune a special-purpose LLM to match patients to trials

PRISM: **P**atient **R**ecords **I**nterpretation for **S**emantic Clinical Trial **M**atching using Large Language Models

Shashi Gupta[1†], Aditya Basu[1†], Mauro Nievas[1], Jerrin Thomas[1], Nathan Wolfrath[2], Adhitya Ramamurthi[2], Bradley Taylor[2], Anai N. Kothari[2*], Regina Schwind[1], Therica M. Miller[3], Sorena Nadaf-Rahrov[4], Yanshan Wang[5], Hrituraj Singh[1*]

[1] Triomics Research, San Francisco, USA.
[2] Medical College of Wisconsin, Milwaukee, USA.
[3] Icahn School of Medicine at Mount Sinai, New York, USA.
[4] Cancer Informatics For Cancer Centers, Los Angeles, USA.
[5] University of Pittsburgh, Pittsburgh, USA.

Talk Outline

1. **Motivation**
    a. What is clinical trial patient recruitment, and why is it hard?
2. **Prior Work**
    a. What did people try before LLMs?
3. **Papers**
    a. Zero-shot patient matching with off-the-shelf LLMs
    b. PRISM: Fine tuning an LLM for clinical trial matching
4. **Future Work**

Zero-Shot Clinical Trial Patient Matching with LLMs

Michael Wornow*, Alejandro Lozano*, Dev Dash, Jenelle Jindal, Kenneth W. Mahaffey, Nigam H. Shah
February 2024

Open research questions

1. **Accuracy:** Can we do **zero-shot matching** of patients?

Open research questions

1. **Accuracy:** Can we do **zero-shot matching** of patients?

2. **Efficiency:** Can we reduce the **cost / time / data / token usage** of LLM-based approaches?

Open research questions

1. **Accuracy:** Can we do **zero-shot matching** of patients?

2. **Efficiency:** Can we reduce the **cost / time / data / token usage** of LLM-based approaches?

3. **Deployment:** How can we **scale** these approaches to health systems?

Dataset: 2018 n2c2 cohort selection benchmark

- **86** patients (test set only)
- **377** deidentified clinical notes (~4 per patient)
- **13** inclusion criteria based on a trial for diabetes

Criteria	Definition	Prevalence
ABDOMINAL	History of intra-abdominal surgery, small or large intestine resection, or small bowel obstruction	0.35
ADVANCED-CAD	Advanced cardiovascular disease (CAD). For the purposes of this annotation, we define "advanced" as having 2 or more of the following: • Taking 2 or more medications to treat CAD • History of myocardial infarction (MI) • Currently experiencing angina • Ischemia, past or present	0.52
ALCOHOL-ABUSE	Current alcohol use over weekly recommended limits	0.03
ASP-FOR-MI	Use of aspirin to prevent MI	0.79
CREATININE	Serum creatinine > upper limit of normal	0.28
DIETSUPP-2MOS	Taken a dietary supplement (excluding vitamin D) in the past 2 months	0.51
DRUG-ABUSE	Drug abuse, current or past	0.03
ENGLISH	Patient must speak English	0.85
HBA1C	Any hemoglobin A1c (HbA1c) value between 6.5% and 9.5%	0.41
KETO-1YR	Diagnosis of ketoacidosis in the past year	0.00
MAJOR-DIABETES	Major diabetes-related complication. For the purposes of this annotation, we define "major complication" (as opposed to "minor complication") as any of the following that are a result of (or strongly correlated with) uncontrolled diabetes: • Amputation • Kidney damage • Skin conditions • Retinopathy • nephropathy • neuropathy	0.50
MAKES-DECISIONS	Patient must make their own medical decisions	0.97
MI-6MOS	MI in the past 6 months	0.09

Task: Given a patient's clinical notes, assess if she meets a criterion

Associated Diagnoses: None .
Subjective:
11/30/15: 80 who presented to the hospital with 3 days history of fever and cough. She was diagnosed with CAP and was started on antibiotics. Unfortunately, she had a significant episode of hypoxemia and had to be intubated. Pinkish frothy sputum was reported after intubation. Patient has a remote history of smoking.
.....
11/30/2015 06:00 Transparent Physical Examination General: intubated and sedated. Eye: Pupils are equal, round and reactive to light, Extraocular movements are intact. HENT: intubated and sedated. Neck: Supple, No lymphadenopathy. Respiratory: bilateral rales. Cardiovascular: Normal rate, Regular rhythm, No murmur. Gastrointestinal: Soft, Non-distended. Musculoskeletal: intubated and sedated. Integumentary: Warm, Dry. Neurologic: intubated and sedated. Results Review Labs Last 24 Hrs SELECT Labs ONLY

Clinical notes for Patient A

https://www.researchgate.net/figure/Example-of-clinical-notes-in-the-EHR_fig2_345508476

Task: Given a patient's clinical notes, assess if she meets a criterion

Associated Diagnoses: None .
Subjective:
11/30/15: 80 who presented to the hospital with 3 days history of fever and cough. She was diagnosed with CAP and was started on antibiotics. Unfortunately, she had a significant episode of hypoxemia and had to be intubated. Pinkish frothy sputum was reported after intubation. Patient has a remote history of smoking.
.....
11/30/2015 06:00 Transparent Physical Examination General: intubated and sedated. Eye: Pupils are equal, round and reactive to light, Extraocular movements are intact. HENT: intubated and sedated. Neck: Supple, No lymphadenopathy. Respiratory: bilateral rales. Cardiovascular: Normal rate, Regular rhythm, No murmur. Gastrointestinal: Soft, Non-distended. Musculoskeletal: intubated and sedated. Integumentary: Warm, Dry. Neurologic: intubated and sedated. Results Review Labs Last 24 Hrs SELECT Labs ONLY

Clinical notes for Patient A

ABDOMINAL	History of intra-abdominal surgery, small or large intestine resection, or small bowel obstruction

Inclusion Criteria X

Task: Given a patient's clinical notes, assess if she meets a criterion

Associated Diagnoses: None .
Subjective:
11/30/15: 80 who presented to the hospital with 3 days history of fever and cough. She was diagnosed with CAP and was started on antibiotics. Unfortunately, she had a significant episode of hypoxemia and had to be intubated. Pinkish frothy sputum was reported after intubation. Patient has a remote history of smoking.
.....
11/30/2015 06:00 Transparent Physical Examination General: intubated and sedated. Eye: Pupils are equal, round and reactive to light, Extraocular movements are intact. HENT: intubated and sedated. Neck: Supple, No lymphadenopathy. Respiratory: bilateral rales. Cardiovascular: Normal rate, Regular rhythm, No murmur. Gastrointestinal: Soft, Non-distended. Musculoskeletal: intubated and sedated. Integumentary: Warm, Dry. Neurologic: intubated and sedated. Results Review Labs Last 24 Hrs SELECT Labs ONLY

ABDOMINAL	History of intra-abdominal surgery, small or large intestine resection, or small bowel obstruction

Clinical notes for Patient A

Inclusion Criteria X

"Yes, patient A meets criteria X"

Our goal is to do “zero-shot” learning, i.e. using no labeled data

Our goal is to do “zero-shot” learning, i.e. using no labeled data

Few-shot

In addition to the task description, the model sees a few examples of the task.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ← examples
4  plush girafe => girafe peluche      ← examples
5  cheese =>                           ← prompt
```

ChatGPT

“fromage”

Brown et al. 2020

Our goal is to do “zero-shot” learning, i.e. using no labeled data

Few-shot

In addition to the task description, the model sees a few examples of the task.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ←
4  plush girafe => girafe peluche      ←
5  cheese =>                           ← prompt
```

ChatGPT

“fromage”

Zero-shot

The model predicts the answer given only a natural language description of the task.

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

ChatGPT

“fromage”

Brown et al. 2020

Results

LLMs can **accurately** do **zero-shot matching** of patients to eligibility criteria

LLMs achieve SOTA with **zero training data** or **data labeling**

Model	Open Source	Prec.	Rec.	Overall Macro-F1	Overall Micro-F1
Baseline	✓	0.69	0.78	0.43	0.76
Prior SOTA	✓	0.88	0.91	0.75	0.91

Higher numbers are better!

LLMs achieve SOTA with **zero training data** or **data labeling**

Model	Open Source	Prec.	Rec.	Overall Macro-F1	Overall Micro-F1
Baseline	✓	0.69	0.78	0.43	0.76
Prior SOTA	✓	0.88	0.91	0.75	0.91
Llama-2-70b	✓	$0.82_{(0.77,0.87)}$	$0.41_{(0.36,0.46)}$	$0.46_{(0.44,0.48)}$	$0.67_{(0.64,0.70)}$
Mixtral-8x7b	✓	$0.72_{(0.68,0.75)}$	$0.83_{(0.79,0.86)}$	$0.64_{(0.59,0.67)}$	$0.79_{(0.77,0.82)}$
Llama-3-70b	✓	$0.76_{(0.72,0.79)}$	$0.88_{(0.85,0.91)}$	$0.68_{(0.64,0.71)}$	$0.83_{(0.81,0.85)}$
Qwen-2-72b	✓	$0.80_{(0.76,0.83)}$	$\mathbf{0.94}_{(0.92,0.96)}$	$0.74_{(0.68,0.77)}$	$0.88_{(0.86,0.90)}$
GPT-3.5		$0.74_{(0.70,0.77)}$	$0.80_{(0.77,0.84)}$	$0.59_{(0.54,0.63)}$	$0.80_{(0.77,0.82)}$
GPT-4		$\mathbf{0.91}_{(0.88,0.93)}$	$0.92_{(0.89,0.94)}$	$\mathbf{0.81}_{(0.77,0.84)}$	$\mathbf{0.93}_{(0.91,0.94)}$

Higher numbers are better!

LLMs achieve SOTA with **zero training data** or **data labeling**

Model	Open Source	Prec.	Rec.	Overall Macro-F1	Overall Micro-F1
Prior SOTA	✓	0.88	0.91	0.75	0.91
GPT-4		**0.91** (0.88,0.93)	0.92 (0.89,0.94)	**0.81** (0.77,0.84)	**0.93** (0.91,0.94)

GPT-4 beats prior methods with **zero labeled data**, which means it can be applied to any trial with **minimal reconfiguration.**

Higher numbers are better!

To apply an LLM-based system to new trials/patients, swap the text.

Associated Diagnoses: None .
Subjective:
11/30/15: 80 who presented to the hospital with 3 days history of fever and cough. She was diagnosed with CAP and was started on antibiotics. Unfortunately, she had a significant episode of hypoxemia and had to be intubated. Pinkish frothy sputum was reported after intubation. Patient has a remote history of smoking.
.....
11/30/2015 06:00 Transparent Physical Examination General: intubated and sedated. Eye: Pupils are equal, round and reactive to light, Extraocular movements are intact. HENT: intubated and sedated. Neck: Supple, No lymphadenopathy. Respiratory: bilateral rales. Cardiovascular: Normal rate, Regular rhythm, No murmur. Gastrointestinal: Soft, Non-distended. Musculoskeletal: intubated and sedated. Integumentary: Warm, Dry. Neurologic: intubated and sedated. Results Review Labs Last 24 Hrs SELECT Labs ONLY

Clinical notes for Patient A

ABDOMINAL	History of intra-abdominal surgery, small or large intestine resection, or small bowel obstruction

Inclusion Criteria X

"Yes, patient A meets criteria X"

To apply an LLM-based system to new trials/patients, swap the text.

Associated Diagnoses: None .
Subjective:
11/30/15: 80 who presented to the hospital with 3 days history of fever and cough. She was diagnosed with CAP and was started on antibiotics. Unfortunately, she had a significant episode of hypoxemia and had to be intubated. Pinkish frothy sputum was reported after intubation. Patient has a remote history of smoking.
.....
11/30/2015 06:00 Transparent Physical Examination General: intubated and sedated. Eye: Pupils are equal, round and reactive to light, Extraocular movements are intact. HENT: intubated and sedated. Neck: Supple, No lymphadenopathy. Respiratory: bilateral rales. Cardiovascular: Normal rate, Regular rhythm, No murmur. Gastrointestinal: Soft, Non-distended. Musculoskeletal: intubated and sedated. Integumentary: Warm, Dry. Neurologic: intubated and sedated. Results Review Labs Last 24 Hrs SELECT Labs ONLY

ALCOHOL-ABUSE	Current alcohol use over weekly recommended limits

Clinical notes for Patient A

Inclusion Criteria Y

"No, patient A meets criteria Y"

To apply an LLM-based system to new trials/patients, swap the text.

Associated Diagnoses: None .
Subjective:
11/30/15: 80 who presented to the hospital with 3 days history of fever and cough. She was diagnosed with CAP and was started on antibiotics. Unfortunately, she had a significant episode of hypoxemia and had to be intubated. Pinkish frothy sputum was reported after intubation. Patient has a remote history of smoking.
.....
11/30/2015 06:00 Transparent Physical Examination General: intubated and sedated. Eye: Pupils are equal, round and reactive to light, Extraocular movements are intact. HENT: intubated and sedated. Neck: Supple, No lymphadenopathy. Respiratory: bilateral rales. Cardiovascular: Normal rate, Regular rhythm, No murmur. Gastrointestinal: Soft, Non-distended. Musculoskeletal: intubated and sedated. Integumentary: Warm, Dry. Neurologic: intubated and sedated. Results Review Labs Last 24 Hrs SELECT Labs ONLY

Clinical notes for Patient B

ALCOHOL-ABUSE	Current alcohol use over weekly recommended limits

Inclusion Criteria Y

"Yes, patient B meets criteria Y"

We evaluate the **cost efficiency** of different prompting strategies

Model	Prompt Strategy		Performance			
	Criteria	Notes	Prec.	Rec.	Overall Macro-F1	Overall Micro-F1
GPT-4	All	All	0.89	0.87	0.80	0.90
	All	Individual	0.91	**0.92**	0.81	**0.93**
	Individual	All	**0.94**	0.86	**0.85**	0.92
	Individual	Individual	0.92	0.89	0.82	0.92

Almost identical performance

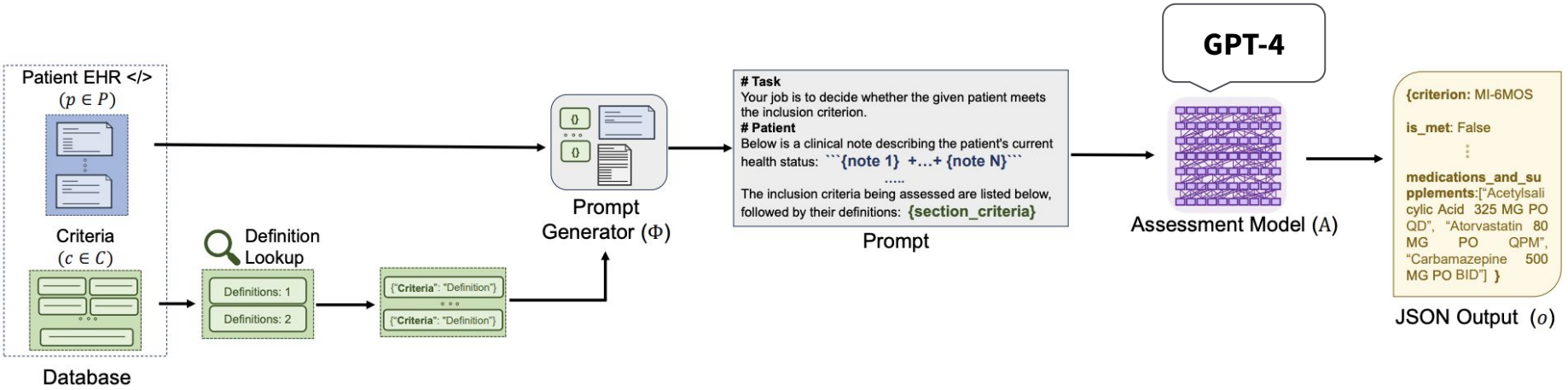
We evaluate the **cost efficiency** of different prompting strategies

Model	Prompt Strategy		Performance				Cost per Patient		
	Criteria	Notes	Prec.	Rec.	Overall Macro-F1	Overall Micro-F1	Dollars	API Calls	Tokens
GPT-4	All	All	0.89	0.87	0.80	0.90	$0.87	1	8.0k
	All	Individual	0.91	**0.92**	0.81	**0.93**	$1.55	4.4	15.3k
	Individual	All	**0.94**	0.86	**0.85**	0.92	$9.08	13	76.9k
	Individual	Individual	0.92	0.89	0.82	0.92	$11.88	57	103.8k

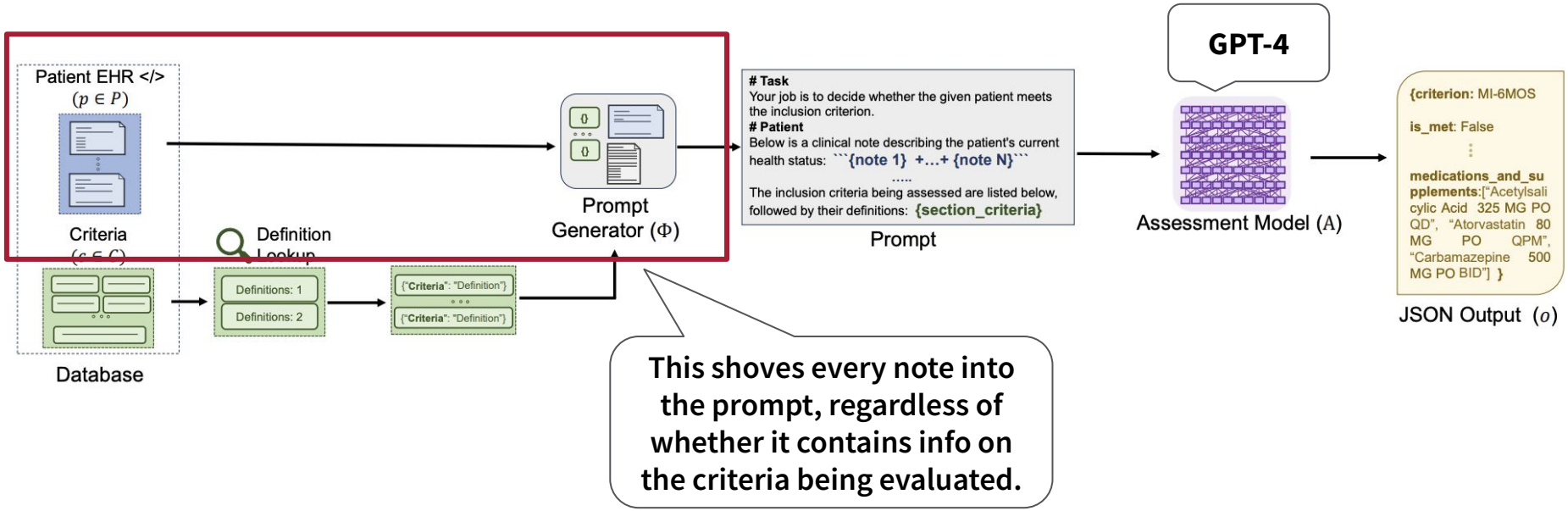
Almost identical performance

But huge cost differences!

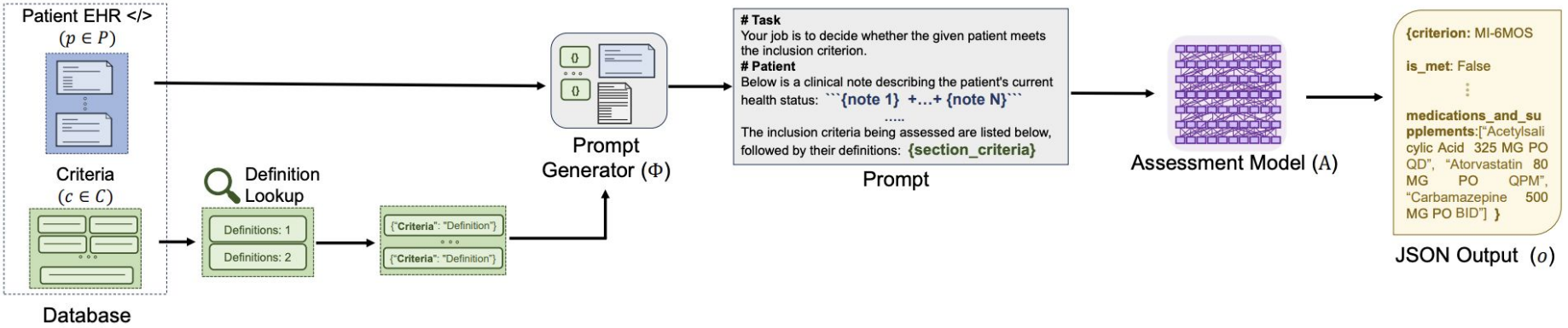
What if instead of feeding all clinical notes into the LLM at once, we only fed in the **most relevant notes?**



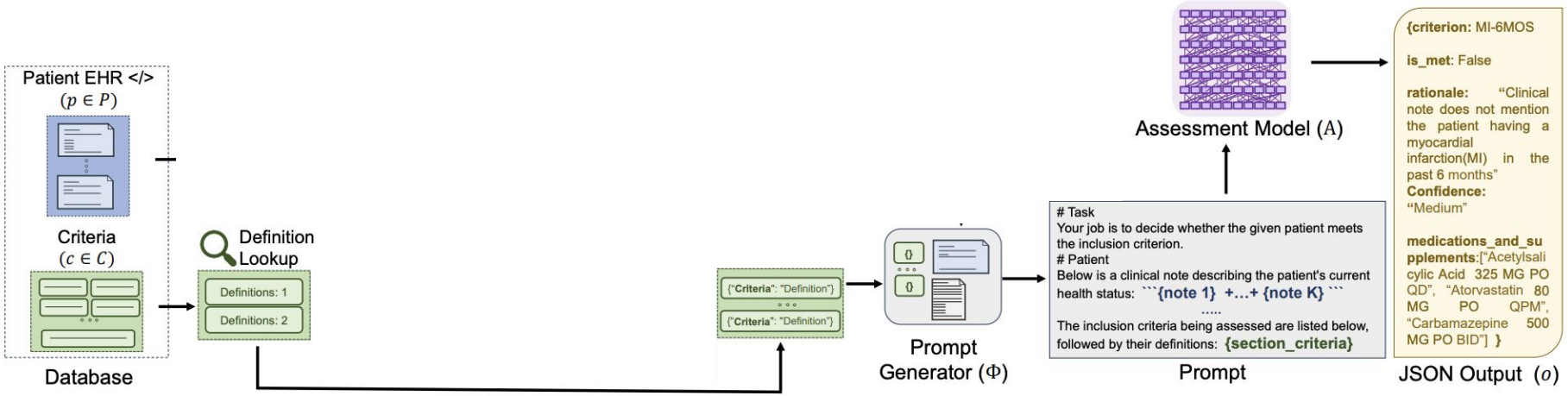
What if instead of feeding all clinical notes into the LLM at once, we only fed in the **most relevant notes?**



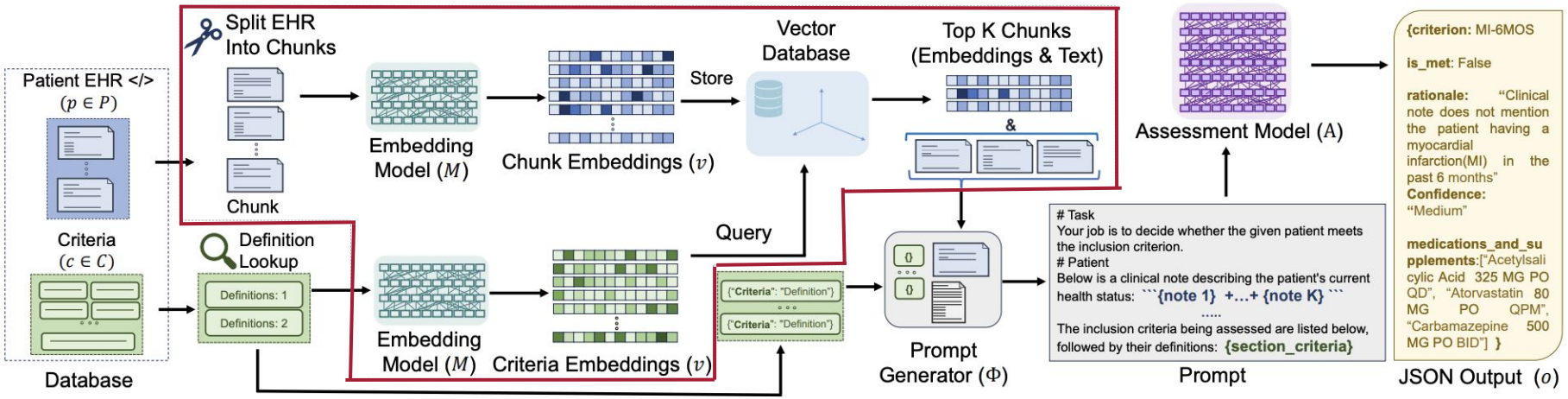
We design a 2-stage retrieval pipeline to select only relevant notes



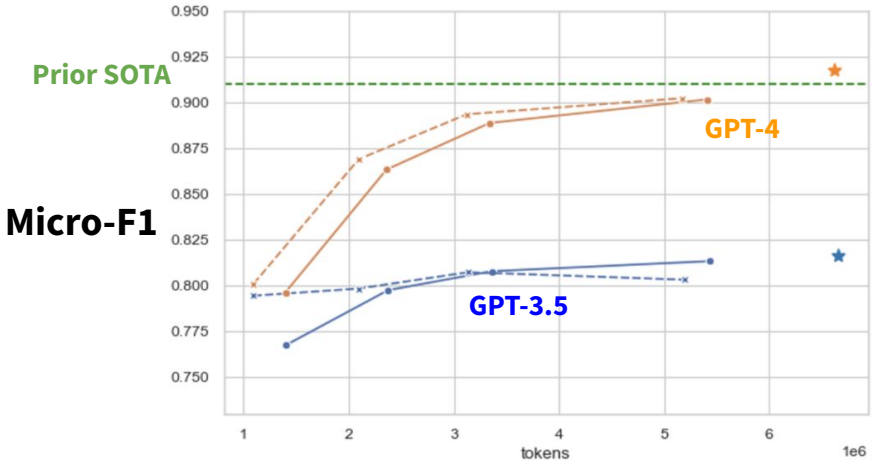
We design a 2-stage retrieval pipeline to select only relevant notes

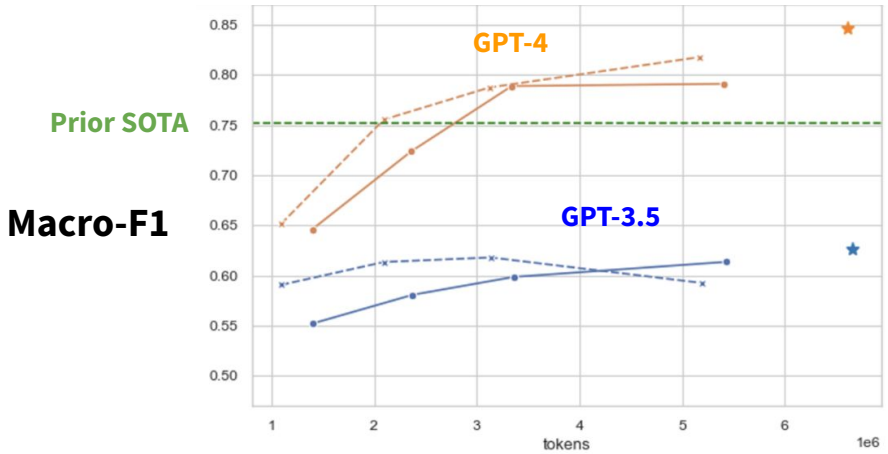


We design a 2-stage retrieval pipeline to select only relevant notes



We can approach SOTA performance using a fraction of the tokens





We can potentially scale to **hundreds of notes** using document retrieval.

We have GPT-4 output a rationale for each eligibility decision

Associated Diagnoses: None .
Subjective:
11/30/15: 80 who presented to the hospital with 3 days history of fever and cough. She was diagnosed with CAP and was started on antibiotics. Unfortunately, she had a significant episode of hypoxemia and had to be intubated. Pinkish frothy sputum was reported after intubation. Patient has a remote history of smoking.
.....
11/30/2015 06:00 Transparent Physical Examination General: intubated and sedated. Eye: Pupils are equal, round and reactive to light, Extraocular movements are intact. HENT: intubated and sedated. Neck: Supple, No lymphadenopathy. Respiratory: bilateral rales. Cardiovascular: Normal rate, Regular rhythm, No murmur. Gastrointestinal: Soft, Non-distended. Musculoskeletal: intubated and sedated. Integumentary: Warm, Dry. Neurologic: intubated and sedated. Results Review Labs Last 24 Hrs SELECT Labs ONLY

ADVANCED-CAD	Advanced cardiovascular disease (CAD). For the purposes of this annotation, we define "advanced" as having 2 or more of the following: • Taking 2 or more medications to treat CAD • History of myocardial infarction (MI) • Currently experiencing angina • Ischemia, past or present

Clinical notes for Patient A

Inclusion Criteria X

"Yes, patient A meets criteria X"

We have GPT-4 output a rationale for each eligibility decision

Associated Diagnoses: None .
Subjective:
11/30/15: 80 who presented to the hospital with 3 days history of fever and cough. She was diagnosed with CAP and was started on antibiotics. Unfortunately, she had a significant episode of hypoxemia and had to be intubated. Pinkish frothy sputum was reported after intubation. Patient has a remote history of smoking.
.....
11/30/2015 06:00 Transparent Physical Examination General: intubated and sedated. Eye: Pupils are equal, round and reactive to light, Extraocular movements are intact. HENT: intubated and sedated. Neck: Supple, No lymphadenopathy. Respiratory: bilateral rales. Cardiovascular: Normal rate, Regular rhythm, No murmur. Gastrointestinal: Soft, Non-distended. Musculoskeletal: intubated and sedated. Integumentary: Warm, Dry. Neurologic: intubated and sedated. Results Review Labs Last 24 Hrs SELECT Labs ONLY

ADVANCED-CAD	Advanced cardiovascular disease (CAD). For the purposes of this annotation, we define "advanced" as having 2 or more of the following: • Taking 2 or more medications to treat CAD • History of myocardial infarction (MI) • Currently experiencing angina • Ischemia, past or present

Clinical notes for Patient A

Inclusion Criteria X

Decision: "***Yes, patient A meets criteria X***"

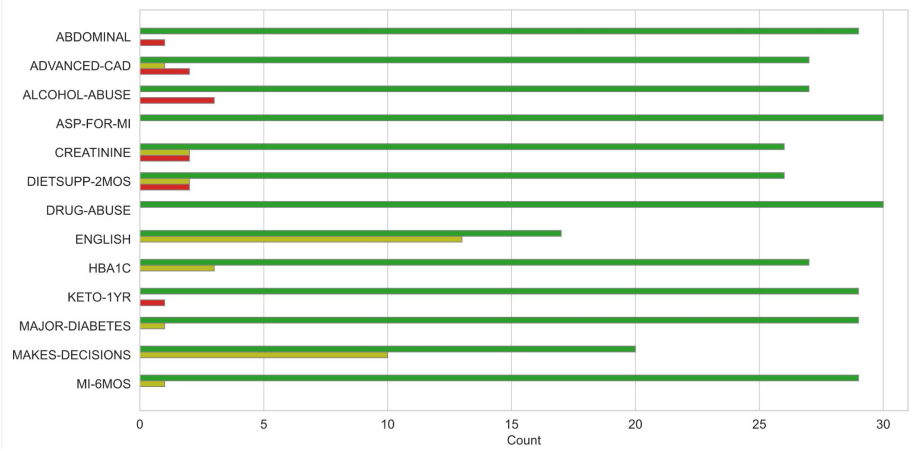
Rationale: The patient is currently taking two or more medications to treat CAD (Asa, Atenolol, Lisinopril, Imdur, and Furosemide). The patient also has a history of stable angina. Therefore, the patient meets two of the categories (a and c) required to meet the ADVANCED-CAD criterion.

Two clinicians reviewed the “rationales” generated by GPT-4

Two clinicians reviewed the “rationales” generated by GPT-4

97%

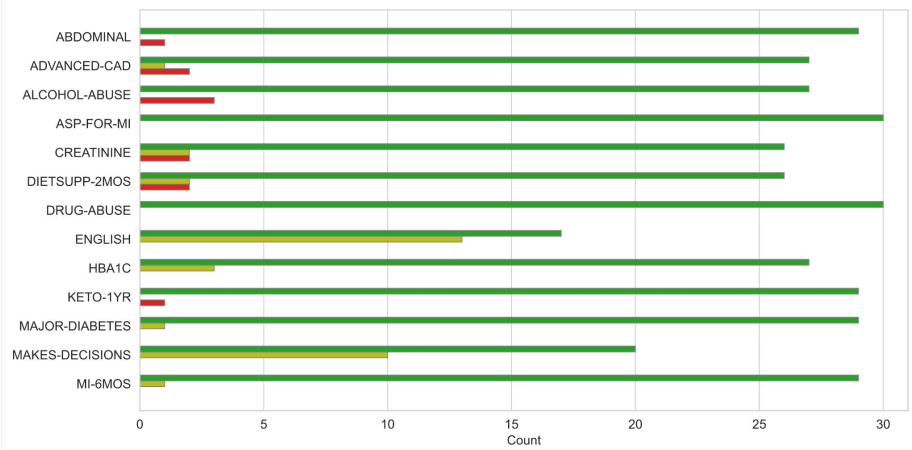
of “**correct**” decisions
were **fully or partially justified.**

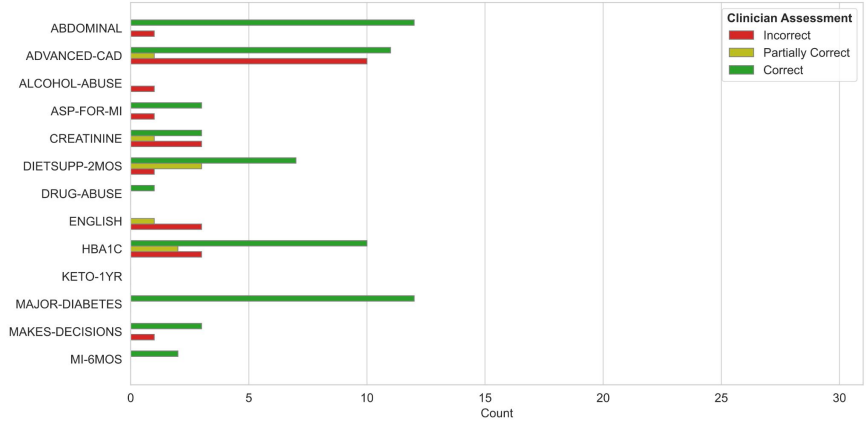


Two clinicians reviewed the “rationales” generated by GPT-4

97%
of “**correct**” decisions
were **fully or partially justified.**

75%
of “**incorrect**” decisions
were **fully or partially justified.**





Our Findings

1. **Accuracy:** Can we do **zero-shot matching** of patients?

Our Findings

1. **Accuracy:** Can we do **zero-shot matching** of patients?
    a. **Yes --** GPT-4 achieves SOTA performance on n2c2 2018 cohort selection benchmark
        i. But not real EHR data :(

Our Findings

1. **Accuracy:** Can we do **zero-shot matching** of patients?
    a. **Yes --** GPT-4 achieves SOTA performance on n2c2 2018 cohort selection benchmark
        i. But not real EHR data :(

2. **Efficiency:** Can we reduce the **cost / time / data / token usage** of LLM-based approaches?

Our Findings

1. **Accuracy:** Can we do **zero-shot matching** of patients?
    a. **Yes --** GPT-4 achieves SOTA performance on n2c2 2018 cohort selection benchmark
        i. But not real EHR data :(

2. **Efficiency:** Can we reduce the **cost / time / data / token usage** of LLM-based approaches?
    a. **Yes --** retrieval + better prompting strategies can dramatically reduce cost
        i. **~30x cheaper** than manual screening ($1.55 v. $34.75 per patient)
        ii. **100x's faster** than manual screening (seconds v. 1 hr per patient for Phase III cancer trial)

Our Findings

1. **Accuracy:** Can we do **zero-shot matching** of patients?
   a. **Yes --** GPT-4 achieves SOTA performance on n2c2 2018 cohort selection benchmark
      i. But not real EHR data :(

2. **Efficiency:** Can we reduce the **cost / time / data / token usage** of LLM-based approaches?
   a. **Yes --** retrieval + better prompting strategies can dramatically reduce cost
      i. **~30x cheaper** than manual screening ($1.55 v. $34.75 per patient)
      ii. **100x's faster** than manual screening (seconds v. 1 hr per patient for Phase III cancer trial)

3. **Deployment:** How can we **scale** these approaches to health systems?

Our Findings

1. **Accuracy:** Can we do **zero-shot matching** of patients?
   a. **Yes --** GPT-4 achieves SOTA performance on n2c2 2018 cohort selection benchmark
      i. But not real EHR data :(

2. **Efficiency:** Can we reduce the **cost / time / data / token usage** of LLM-based approaches?
   a. **Yes --** retrieval + better prompting strategies can dramatically reduce cost
      i. **~30x cheaper** than manual screening ($1.55 v. $34.75 per patient)
      ii. **100x's faster** than manual screening (seconds v. 1 hr per patient for Phase III cancer trial)

3. **Deployment:** How can we **scale** these approaches to health systems?
   a. **Feasible --** retrieval seems promising; calculate embeddings once over entire EHR
   b. **Human-in-the-loop potential --** model-generated explanations are generally correct

Talk Outline

1. **Motivation**
    a. What is clinical trial patient recruitment, and why is it hard?
2. **Prior Work**
    a. What did people try before LLMs?
3. **Papers**
    a. Zero-shot patient matching with off-the-shelf LLMs
    b. PRISM: Fine tuning an LLM for clinical trial matching
4. **Future Work**

PRISM: Patient Records Interpretation for Semantic Clinical Trial Matching using Large Language Models

Shashi Kant Gupta*, Aditya Basu*, Mauro Nievas, Jerrin Thomas, Nathan Wolfrath, Adhitya Ramamurthi, Bradley Taylor, Anai N. Kothari, Regina Schwind, Therica M. Miller, Sorena Nadaf-Rahrov, Yanshan Wang, Hrituraj Singh
April 2024

Open research questions

1. **Beating OpenAI:** Can we **outperform proprietary models** using a smaller, open source LLM?

Open research questions

1. **Beating OpenAI:** Can we **outperform proprietary models** using a smaller, open source LLM?

2. **Real-World Trials:** Can LLMs do trial matching on **real-world cancer patients** / trials?

Open research questions

1. **Beating OpenAI:** Can we **outperform proprietary models** using a smaller, open source LLM?

2. **Real-World Trials:** Can LLMs do trial matching on **real-world cancer patients** / trials?

3. **End-to-End Pipeline:** Can LLMs match patients to trials **without human intervention**?

Task: Given a patient's clinical notes, assess if she meets a criterion

Associated Diagnoses: None .
Subjective:
11/30/15: 80 who presented to the hospital with 3 days history of fever and cough. She was diagnosed with CAP and was started on antibiotics. Unfortunately, she had a significant episode of hypoxemia and had to be intubated. Pinkish frothy sputum was reported after intubation. Patient has a remote history of smoking.
.....
11/30/2015 06:00 Transparent Physical Examination General: intubated and sedated. Eye: Pupils are equal, round and reactive to light, Extraocular movements are intact. HENT: intubated and sedated. Neck: Supple, No lymphadenopathy. Respiratory: bilateral rales. Cardiovascular: Normal rate, Regular rhythm, No murmur. Gastrointestinal: Soft, Non-distended. Musculoskeletal: intubated and sedated. Integumentary: Warm, Dry. Neurologic: intubated and sedated. Results Review Labs Last 24 Hrs SELECT Labs ONLY

ABDOMINAL	History of intra-abdominal surgery, small or large intestine resection, or small bowel obstruction

Clinical notes for Patient A

Inclusion Criteria X

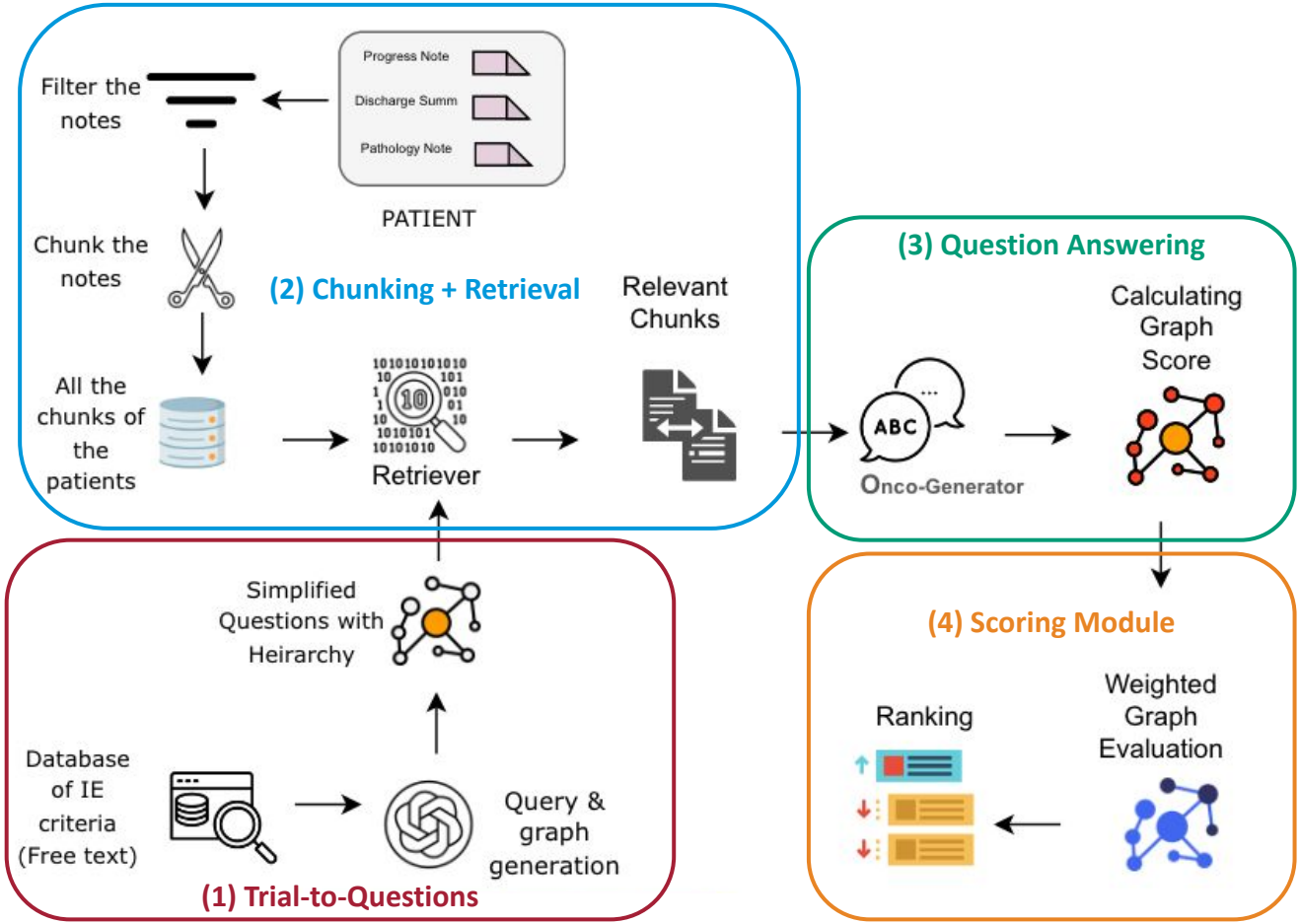
Model

"Yes, patient A meets criteria X"

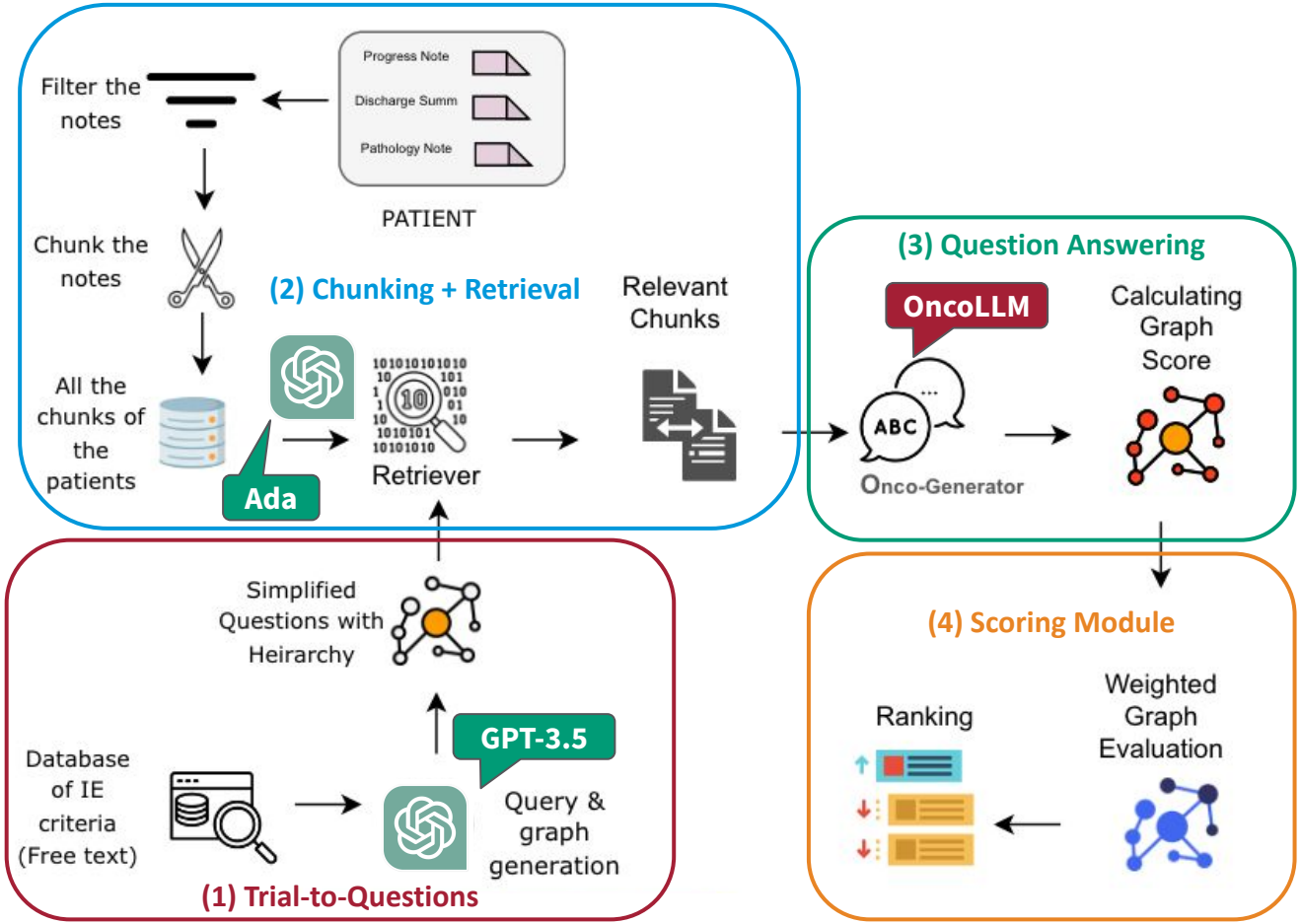
Models

- Proprietary
  - GPT-3.5 (175B)
  - GPT-4 (1000B)
- Open Source
  - Mistral (7B)
  - Mixtral (8x7B)
  - Qwen-1.5 (14B)
- Custom
  - **OncoLLM (14B) --** Qwen-1.5 finetuned on a single cancer center's EHR

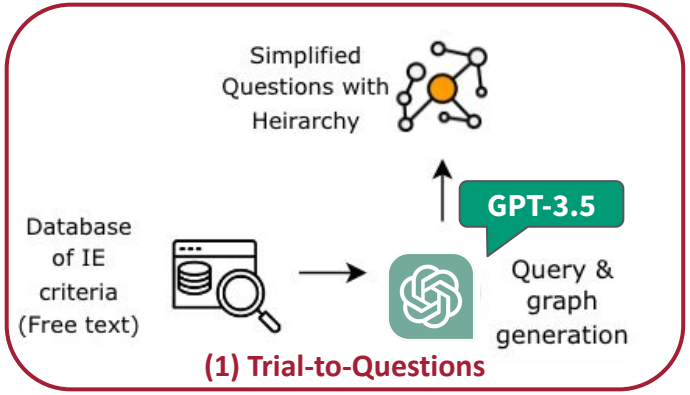
PRISM: End-to-end trial matching pipeline



PRISM: End-to-end trial matching pipeline



PRISM: End-to-end trial matching pipeline



First, **map** each criterion into 1+ **distinct Yes/No questions** and assign it a “**Tier**” based on importance

First, **map** each criterion into 1+ **distinct Yes/No questions** and assign it a "**Tier**" based on importance

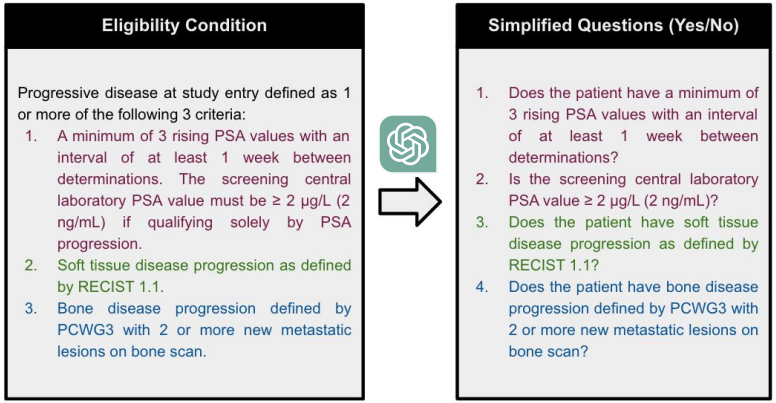
Eligibility Condition

Progressive disease at study entry defined as 1 or more of the following 3 criteria:

1. A minimum of 3 rising PSA values with an interval of at least 1 week between determinations. The screening central laboratory PSA value must be ≥ 2 µg/L (2 ng/mL) if qualifying solely by PSA progression.
2. Soft tissue disease progression as defined by RECIST 1.1.
3. Bone disease progression defined by PCWG3 with 2 or more new metastatic lesions on bone scan.

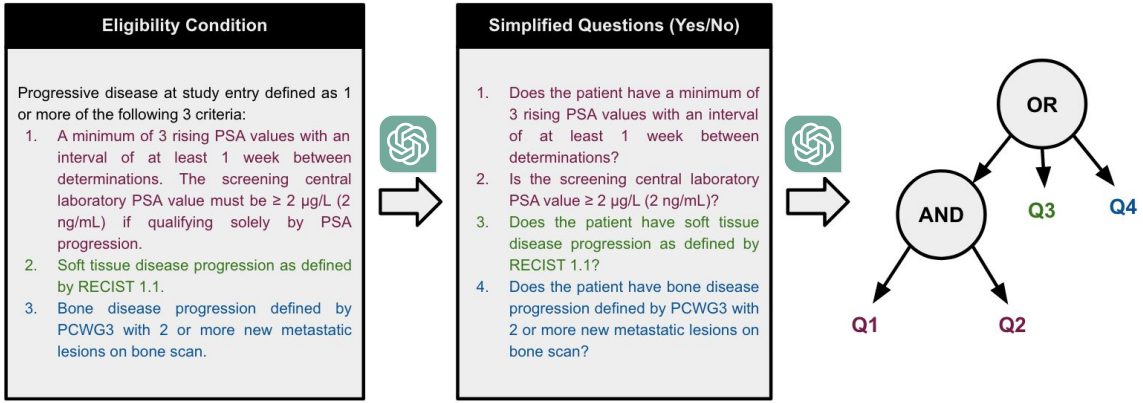
Use **GPT-3.5** to convert textual trial criteria from CT.gov (**"Eligibility Condition"**) into independent Yes/No questions (**"Simplified Questions"**) with boolean relationships

First, **map** each criterion into 1+ **distinct Yes/No questions** and assign it a "**Tier**" based on importance



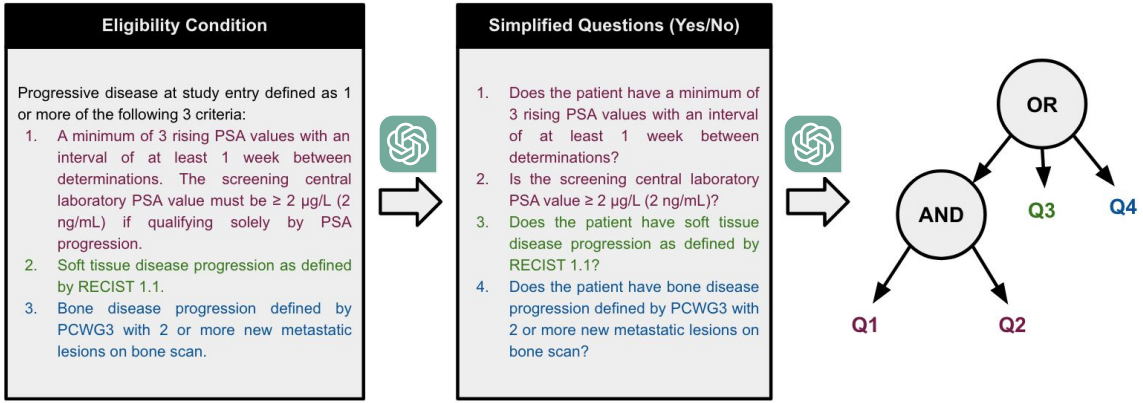
Use **GPT-3.5** to convert textual trial criteria from CT.gov (**"Eligibility Condition"**) into independent Yes/No questions (**"Simplified Questions"**) with boolean relationships

First, **map** each criterion into 1+ **distinct Yes/No questions** and assign it a "**Tier**" based on importance



Use **GPT-3.5** to convert textual trial criteria from CT.gov (**"Eligibility Condition"**) into independent Yes/No questions (**"Simplified Questions"**) with boolean relationships

First, **map** each criterion into 1+ **distinct Yes/No questions** and assign it a "**Tier**" based on importance



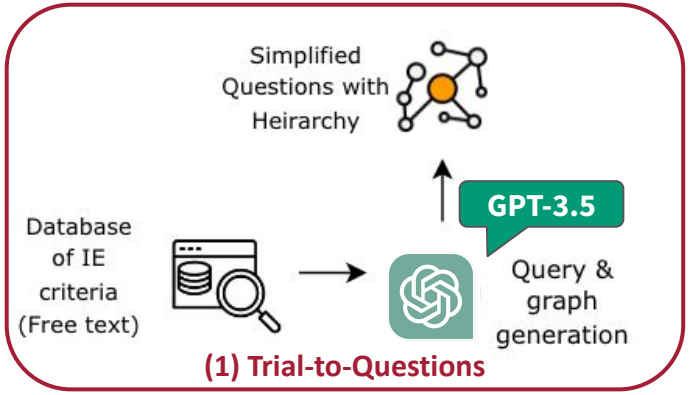
Concept	Tier
Cancer Type	1
Cancer Subtype	1
Cancer Stage	1
Cancer Grade/Histology	1
Genetic & Biologic Markers	2
Lab/Imaging Criteria	2
Prior treatment/surgery	2
Comorbidities	3
Functional Status	4
Others	4

Fig. 2: Concept-Tier Mapping

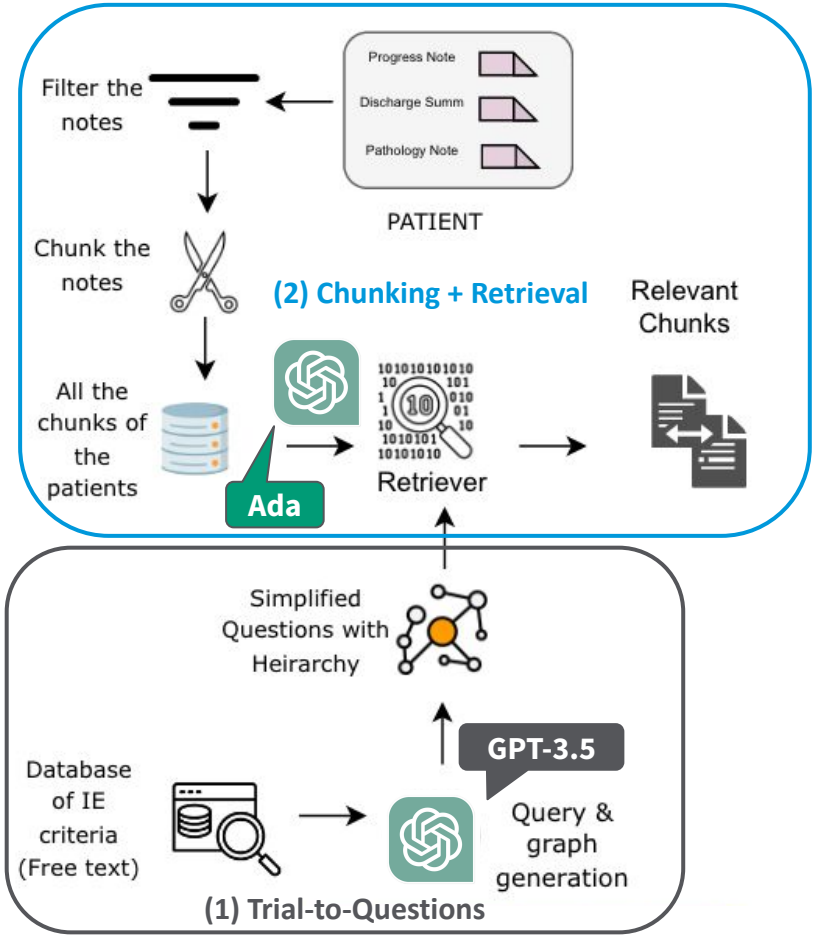
Use **GPT-3.5** to convert textual trial criteria from CT.gov (**"Eligibility Condition"**) into independent Yes/No questions (**"Simplified Questions"**) with boolean relationships

Use **GPT-4** to classify each criterion into a **Tier** from **1** (least) to **4** (most important)

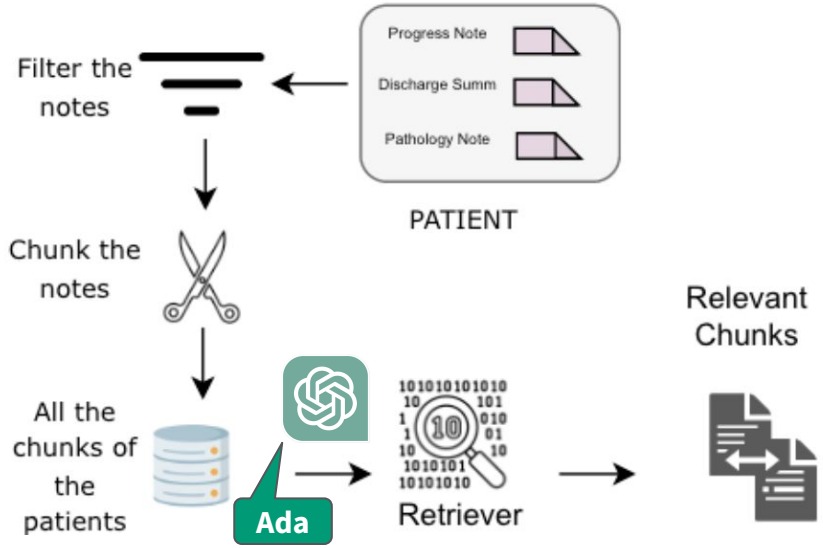
PRISM: End-to-end trial matching pipeline



PRISM: End-to-end trial matching pipeline



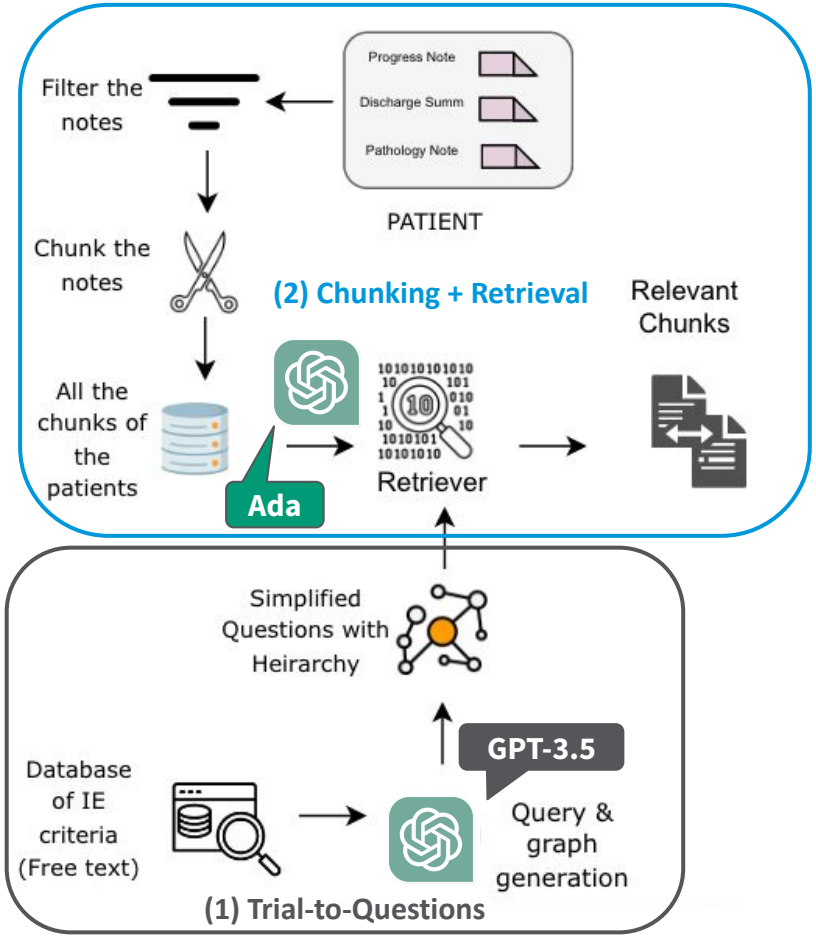
We **embed** each note and **retrieve** the most relevant chunks



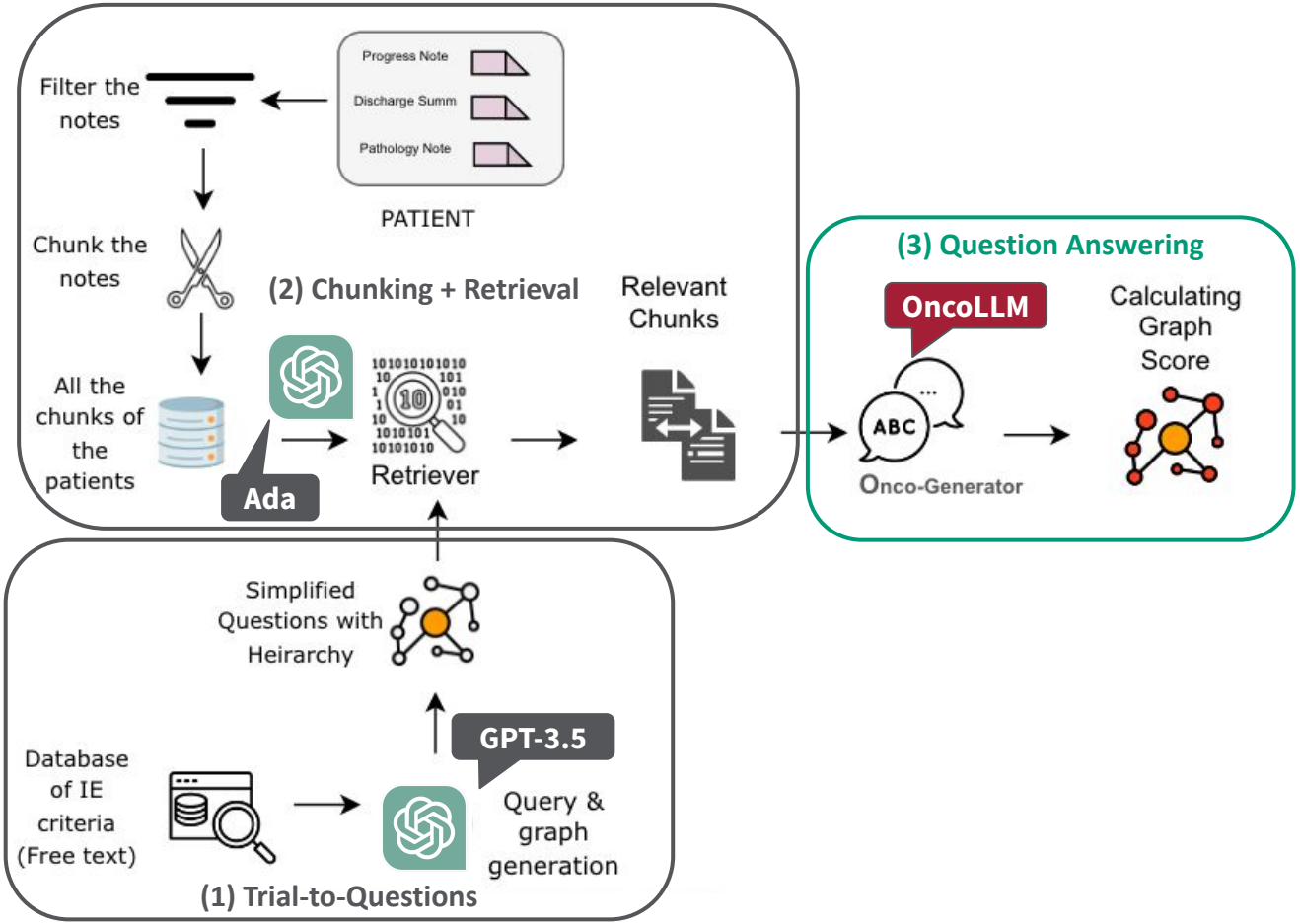
Category	# of Patients	Average # of Documents
Breast Cancer	37	49
Lung Cancer	20	76
Prostrate Cancer	29	57
Colorectal Cancer	7	283
Skin Cancer	4	52

We use **OpenAI's Ada** model to **embed** each note **chunk** and retrieve relevant chunks based on **cosine similarity** to the query

PRISM: End-to-end trial matching pipeline



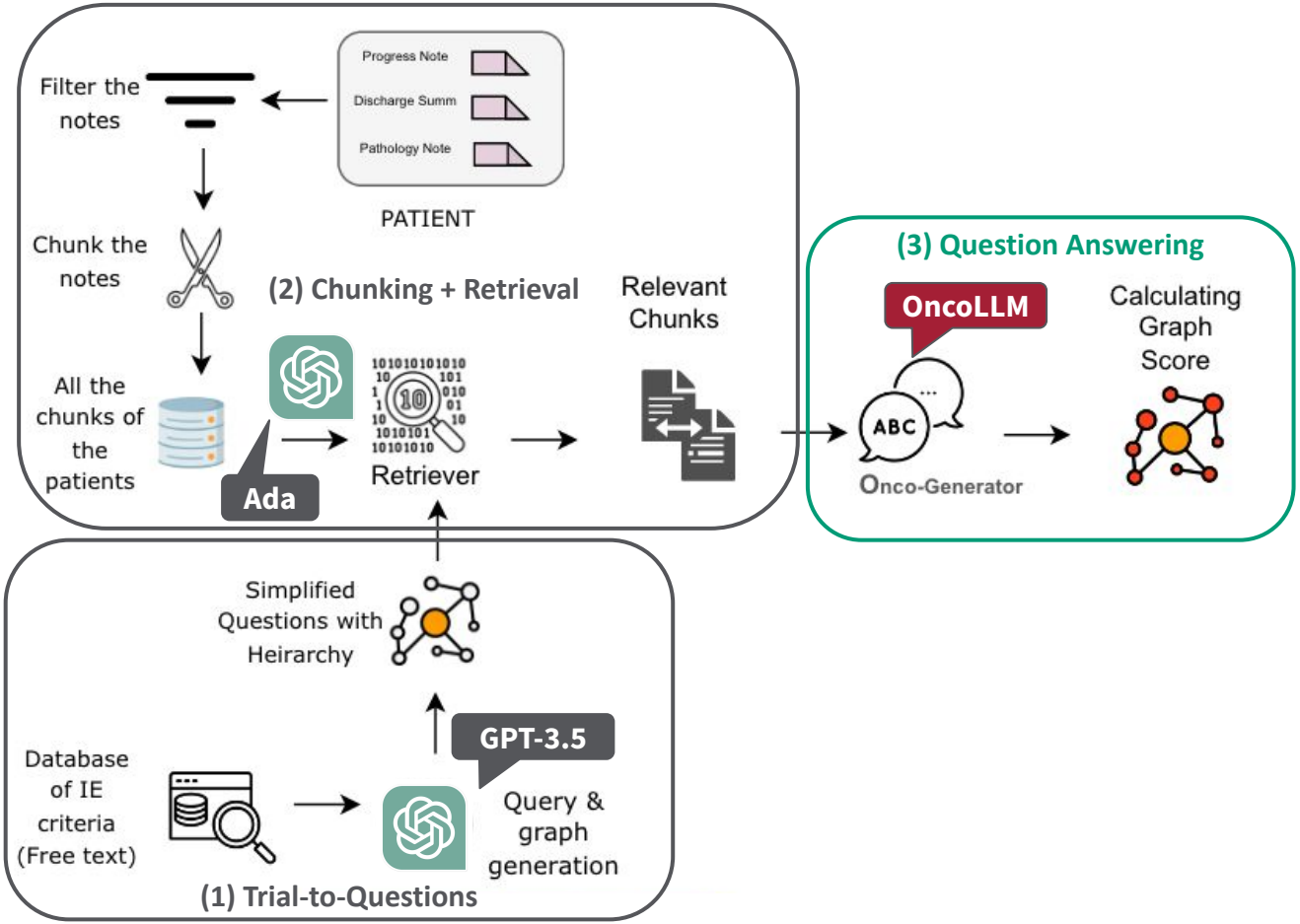
PRISM: End-to-end trial matching pipeline



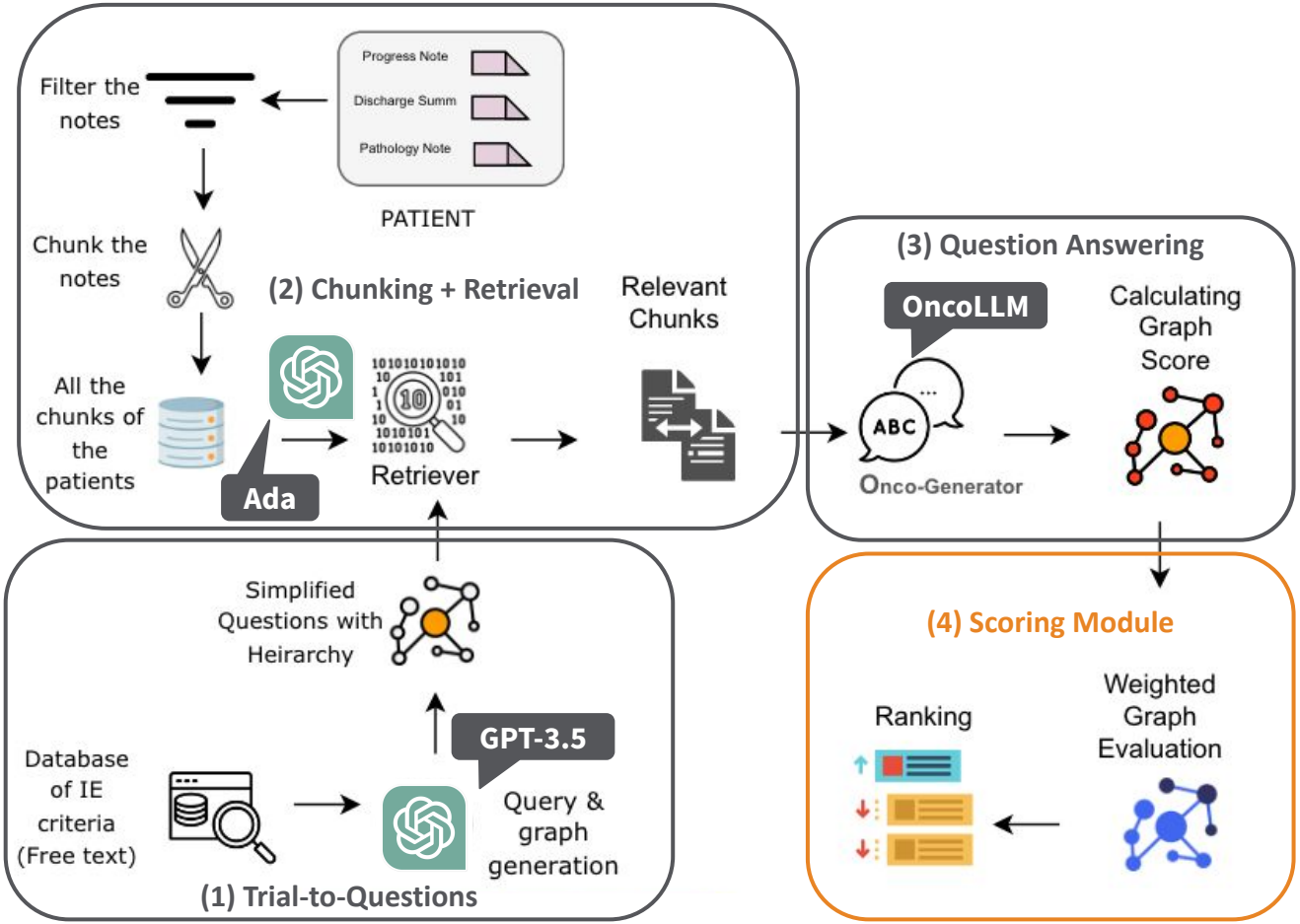
Ask LLM to provide **4 outputs** using **retrieved chunks**

1. **Question Explanation**
   a. Requirements of the question, strategy for answering, and what additional information may be required
2. **Answer Explanation**
   a. Chain-of-thought thinking
3. **Answer**
   a. Yes/No/NA
4. **Confidence**
   a. Score from 1 (low) - 5 (high)

PRISM: End-to-end trial matching pipeline

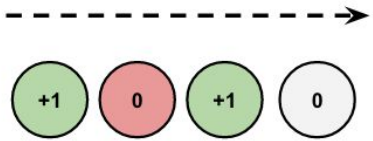


PRISM: End-to-end trial matching pipeline



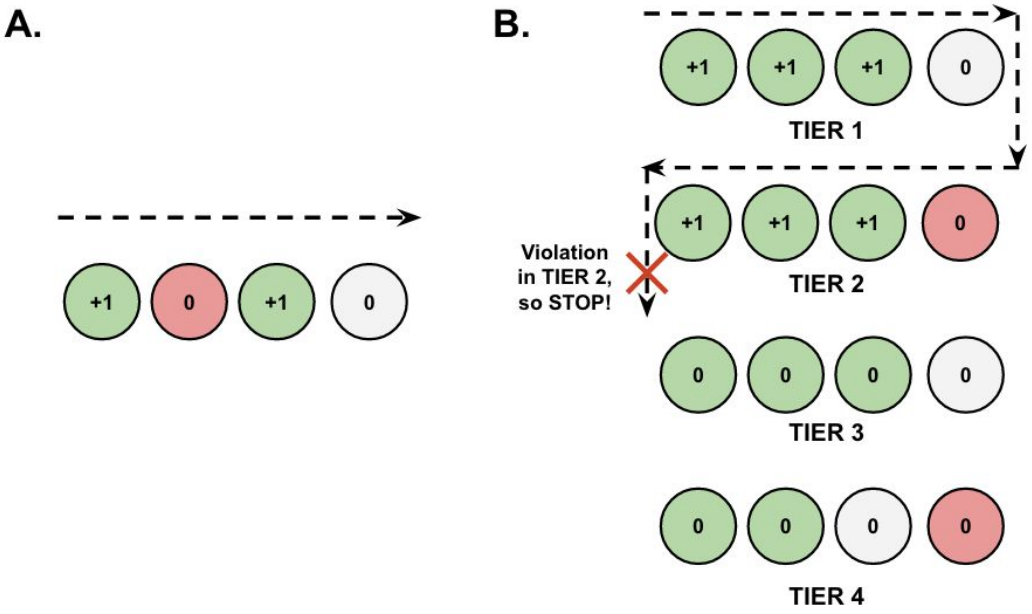
They evaluate **3 methods** for aggregating criterion-level decisions into a **single trial <> patient score**

A.



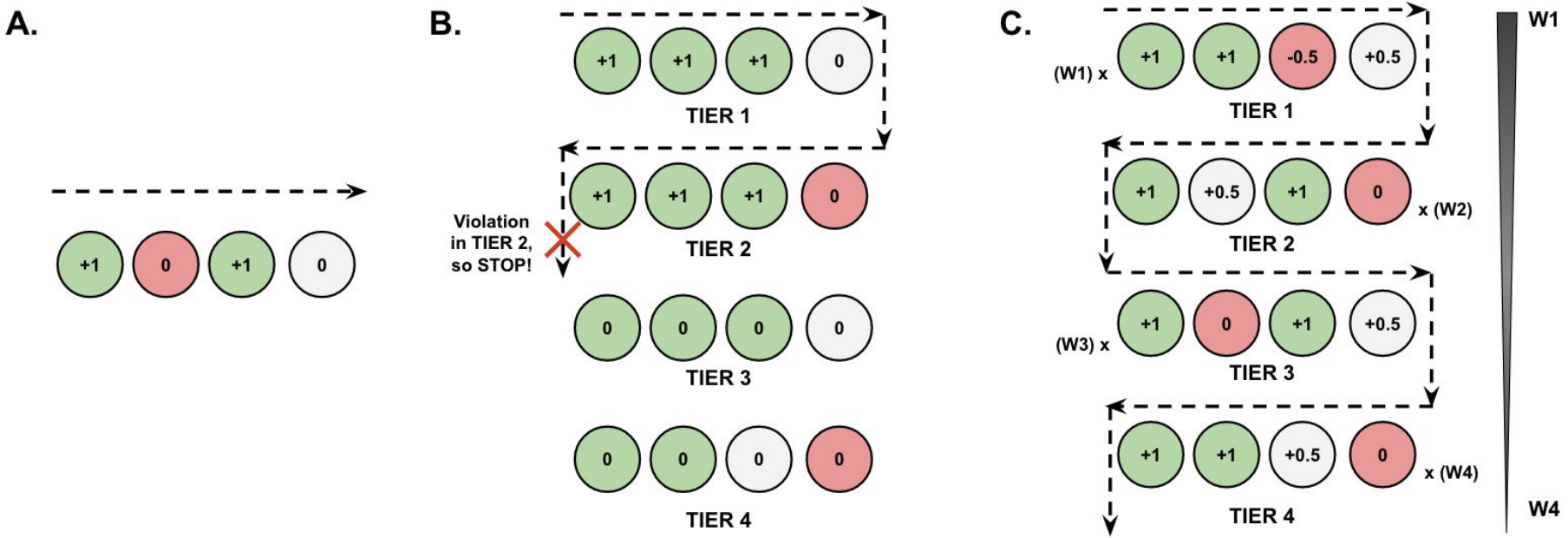
Simple

They evaluate **3 methods** for aggregating criterion-level decisions into a **single trial <> patient score**



Simple **Iterative Tier**

They evaluate **3 methods** for aggregating criterion-level decisions into a **single trial <> patient score**



Simple **Iterative Tier** **Weighted Tier**

Dataset 1: Question-level accuracy

- **Data**
  - **10k** notes from **50** cancer patients
  - **720** total Yes/No/NA questions based on E/I criteria
- **Evaluation**
  - Accuracy (i.e. percent of questions answered correctly)
- **Source**
  - Real-world data from one cancer center

PRISM beats GPT-3.5 on question answering, nears GPT-4 with **100x fewer** parameters

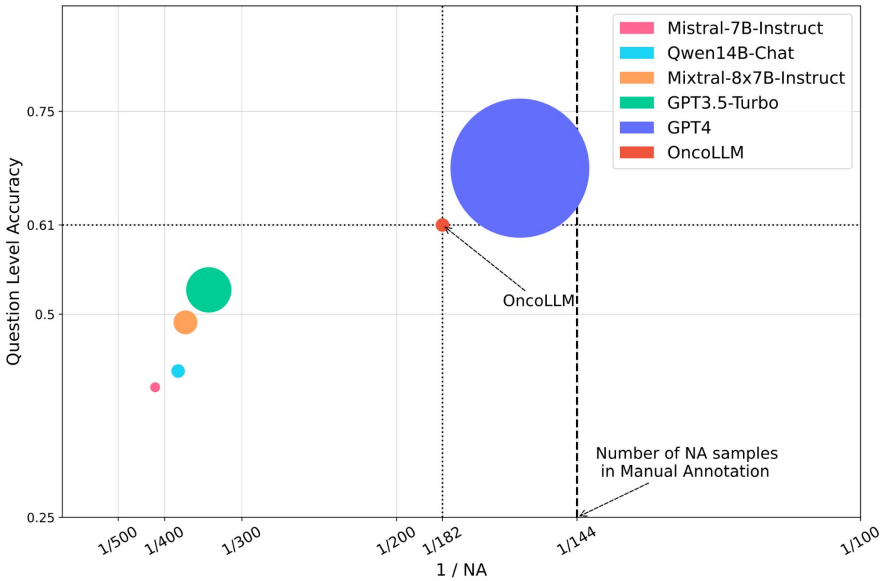
Model Name	All	Without *N/A* samples
GPT3.5-Turbo	53%	48%
Mistral-7B-Instruct	41%	32%
Mixtral-8x7B-Instruct	49%	43%
Qwen14B-Chat	43%	34%
OncoLLM	**63%**	**66%**
GPT4	68%	72%
Expert Doctors*	70%	-

Question-level accuracy on 720 eligibility questions derived from 10k notes and 50 cancer patients

PRISM beats GPT-3.5 on question answering, nears GPT-4 with **100x fewer** parameters

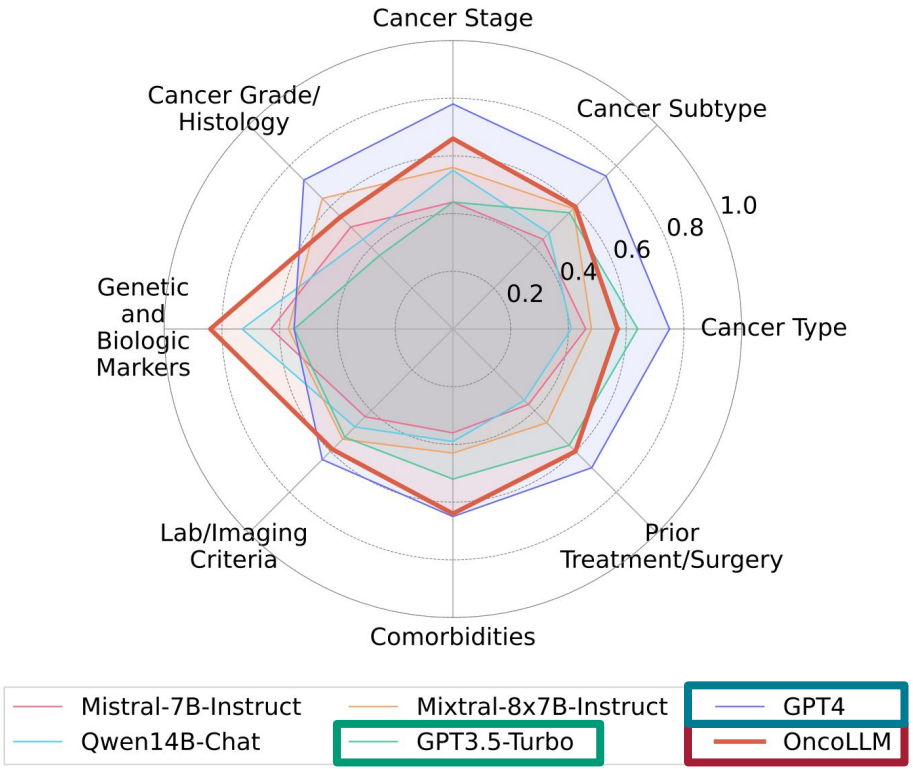
Model Name	All	Without *N/A* samples
GPT3.5-Turbo	53%	48%
Mistral-7B-Instruct	41%	32%
Mixtral-8x7B-Instruct	49%	43%
Qwen14B-Chat	43%	34%
OncoLLM	**63%**	**66%**
GPT4	68%	72%
Expert Doctors*	70%	-

Question-level accuracy on 720 eligibility questions derived from 10k notes and 50 cancer patients



X-axis = frequency of N/A answers (farther right is less useful)
Circle = size of model
Y-axis = question-level accuracy

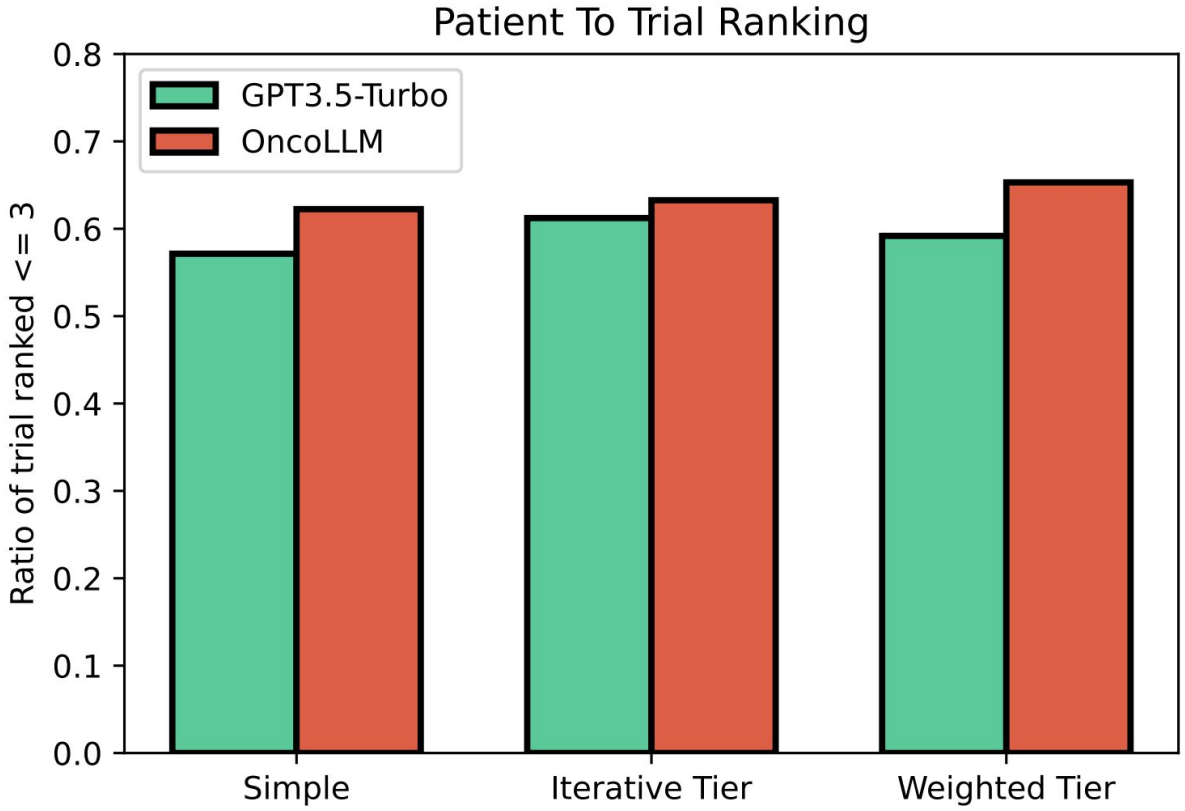
PRISM performs well across all types of criteria



Dataset 2: Patient-to-trial ranking

- **Task**
  - Given a patient, rank their most relevant trials
- **Data**
  - **98** patient-trial positive matches (i.e. patient enrolled in trial)
  - **980** patient-trial negatives (i.e. patient did not enroll in trial)
- **Evaluation**
  - **Precision@3** (i.e. real trial is top-3 out of 10 ranked by model)
- **Source**
  - Real-world data from one cancer center

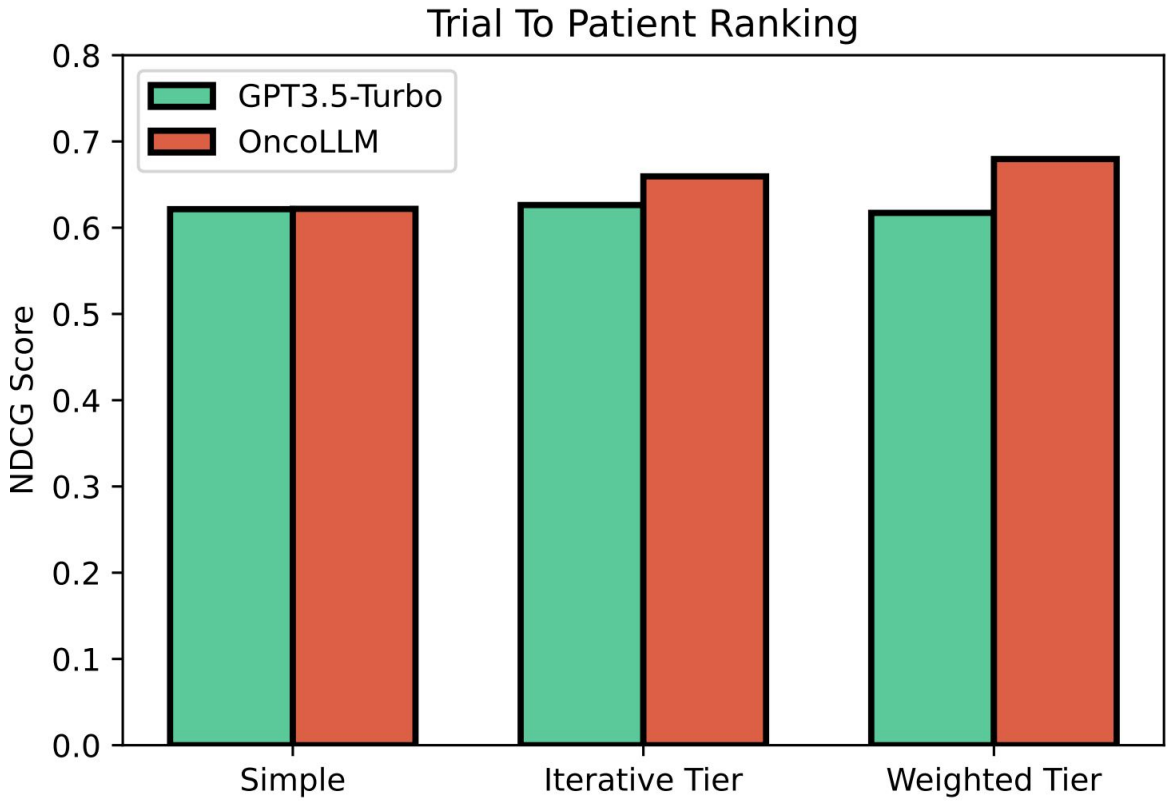
PRISM beats GPT-3.5 on patient-to-trial ranking



Dataset 3: Trial-to-patient ranking

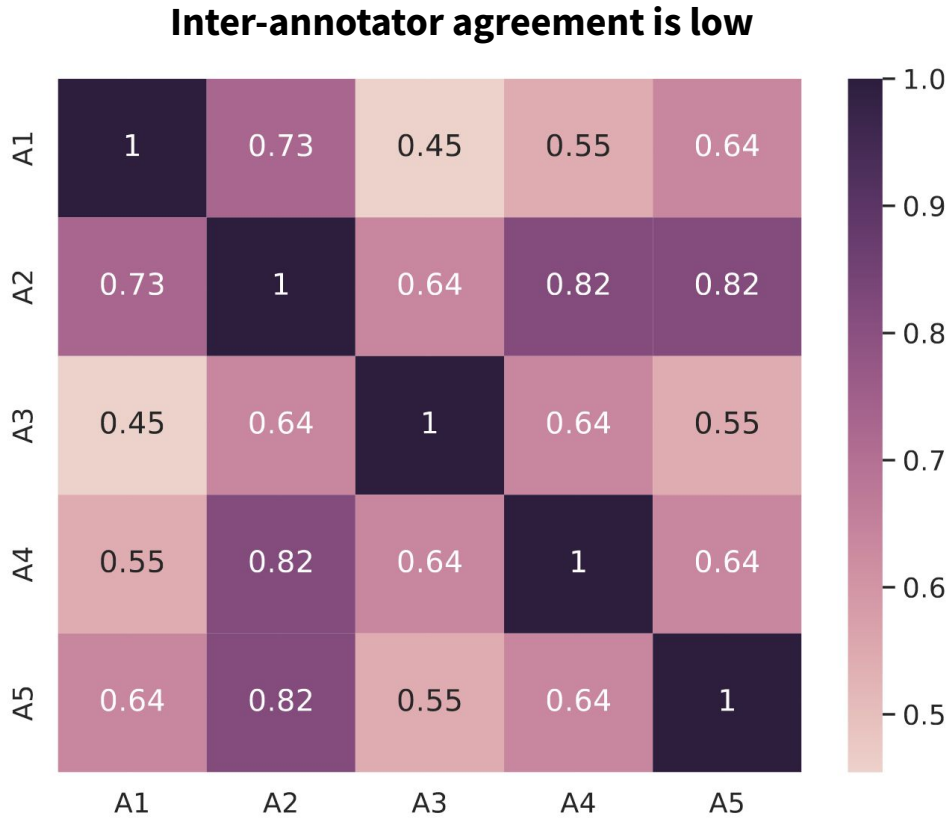
- **Task**
  - Given a trial, rank patients by potential eligibility
- **Data**
  - **36** clinical trials
  - Each trial has **1-3 enrolled** and **5-21 not enrolled** patients
- **Evaluation**
  - **NDCG** with binary relevance score
- **Source**
  - Real-world data from one cancer center

PRISM beats GPT-3.5 on trial-to-patient ranking



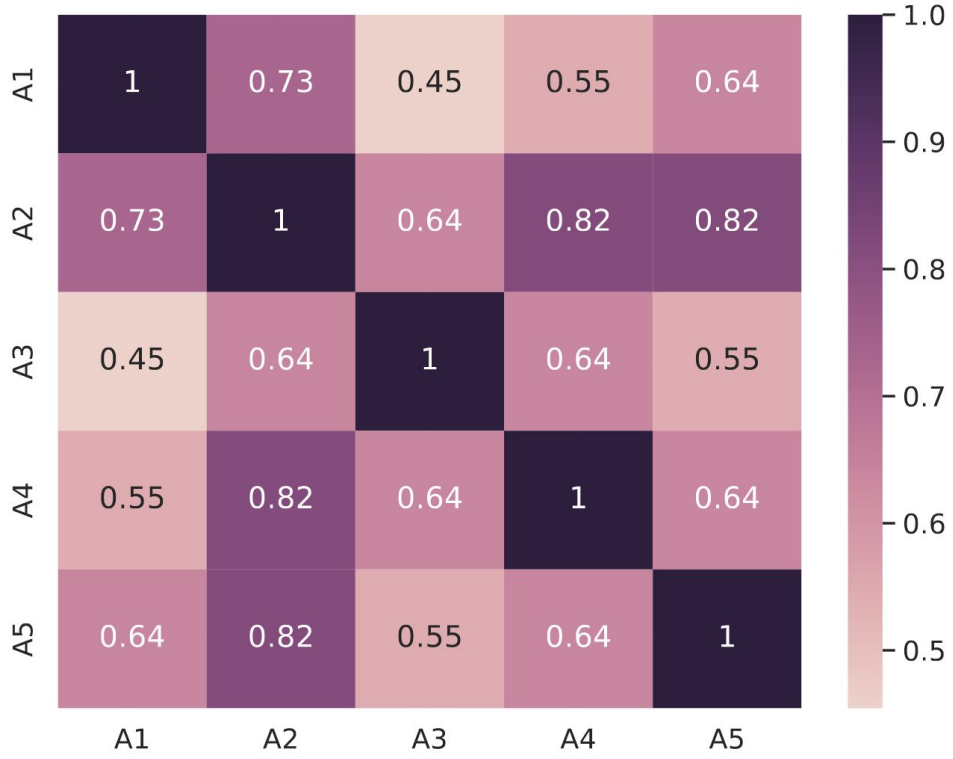
Clinical trial matching is **hard**, even for **humans**

Clinical trial matching is **hard**, even for **humans**



Clinical trial matching is **hard**, even for **humans**

Inter-annotator agreement is low

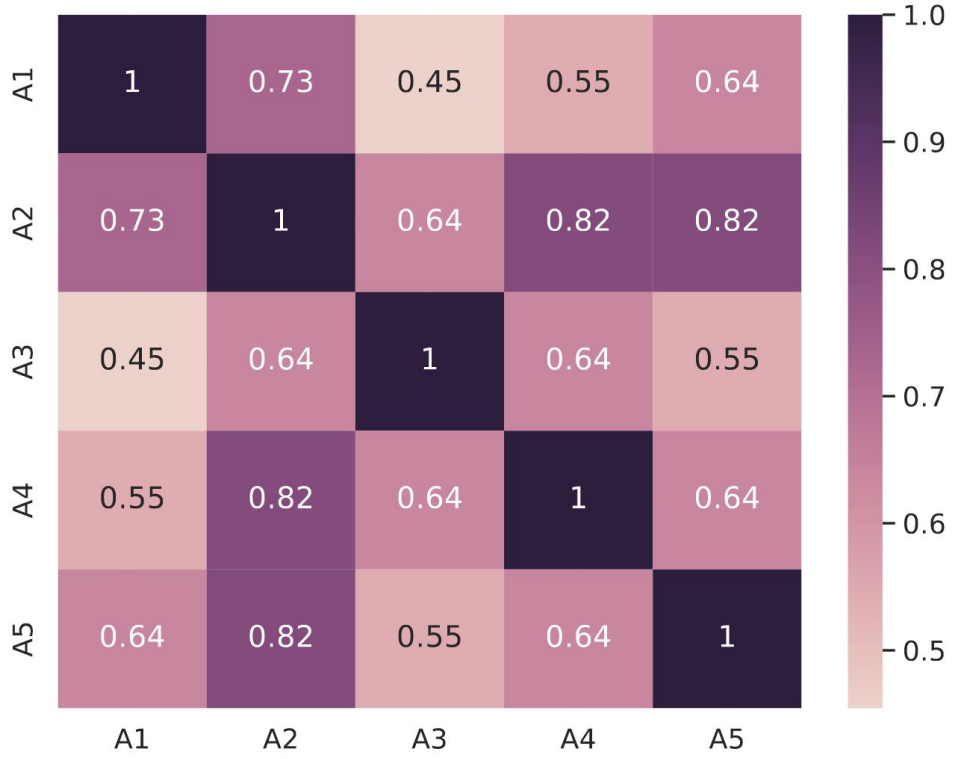


We underestimate model performance

- Manual review found that **patients** were **eligible** for **OncoLLM's top-1** ranked trial in **9/10 trials**

Clinical trial matching is **hard**, even for **humans**

Inter-annotator agreement is low



We underestimate model performance

- Manual review found that **patients** were **eligible** for **OncoLLM's top-1** ranked trial in **9/10 trials**

- Manual review of **criterion-level questions** found correctness of…
  - **75% of answers** (63% reported previously)
  - **90% of explanations** for correct answers
  - **86% of citations**

OncoLLM was **36x cheaper** than GPT-4…

- **OncoLLM** = **$0.17** per patient-trial
- **GPT-4** = **$6.18** per patient-trial (April 2024 pricing)

OncoLLM was **36x cheaper** than GPT-4…(in April 2024)

- **OncoLLM** = **$0.17** per patient-trial
- **GPT-4** = **$6.18** per patient-trial (April 2024 pricing)
  - Today, this would be…
    - **~$0.26** per patient-trial with **GPT-4o**
    - **~$0.02** per patient-trial with **GPT-4o-mini**

April 2024
- GPT-4-32k (Azure) = $60/1M input tokens

October 2024
- GPT-4o-2024-08-06 (Azure) = $2.50/1M input tokens (24x cheaper than GPT-4)

October 2024
- GPT-4o-mini-2024-07-18 (OpenAI) = $0.150/1M input tokens (400x cheaper than GPT-4)

Open research questions

1. **Beating OpenAI:** Can we **outperform proprietary models** using a smaller, open source LLM?

Open research questions

1. **Beating OpenAI:** Can we **outperform proprietary models** using a smaller, open source LLM?
    a. **Mixed --** OncoLLM beats GPT-3.5, but lags GPT-4
        i. **~36x cheaper** than GPT-4 (in April 2024)
        ii. **100x smaller** model than GPT-4
        iii. More **privacy preserving** + enables on-premise hosting

Open research questions

1. **Beating OpenAI:** Can we **outperform proprietary models** using a smaller, open source LLM?
   a. **Mixed --** OncoLLM beats GPT-3.5, but lags GPT-4
      i. **~36x cheaper** than GPT-4 (in April 2024)
      ii. **100x smaller** model than GPT-4
      iii. More **privacy preserving** + enables on-premise hosting

2. **Real-World Trials:** Can LLMs do trial matching on **real-world cancer patients** / trials?

Open research questions

1. **Beating OpenAI:** Can we **outperform proprietary models** using a smaller, open source LLM?
   a. **Mixed --** OncoLLM beats GPT-3.5, but lags GPT-4
      i. **~36x cheaper** than GPT-4 (in April 2024)
      ii. **100x smaller** model than GPT-4
      iii. More **privacy preserving** + enables on-premise hosting

2. **Real-World Trials:** Can LLMs do trial matching on **real-world cancer patients** / trials?
   a. **Yes --** LLMs are capable of reasoning over real-world EHR data; retrieval enables scaling to hundreds of notes

Open research questions

1. **Beating OpenAI:** Can we **outperform proprietary models** using a smaller, open source LLM?
   a. **Mixed --** OncoLLM beats GPT-3.5, but lags GPT-4
      i. **~36x cheaper** than GPT-4 (in April 2024)
      ii. **100x smaller** model than GPT-4
      iii. More **privacy preserving** + enables on-premise hosting

2. **Real-World Trials:** Can LLMs do trial matching on **real-world cancer patients** / trials?
   a. **Yes --** LLMs are capable of reasoning over real-world EHR data; retrieval enables scaling to hundreds of notes

3. **End-to-End Pipeline:** Can LLMs match patients to trials **without human intervention**?

Open research questions

1. **Beating OpenAI:** Can we **outperform proprietary models** using a smaller, open source LLM?
   a. **Mixed --** OncoLLM beats GPT-3.5, but lags GPT-4
      i. **~36x cheaper** than GPT-4 (in April 2024)
      ii. **100x smaller** model than GPT-4
      iii. More **privacy preserving** + enables on-premise hosting

2. **Real-World Trials:** Can LLMs do trial matching on **real-world cancer patients** / trials?
   a. **Yes --** LLMs are capable of reasoning over real-world EHR data; retrieval enables scaling to hundreds of notes

3. **End-to-End Pipeline:** Can LLMs match patients to trials **without human intervention**?
   a. **Yes --** PRISM demonstrates ability to go from raw clinical notes + trial E/I criteria to a final patient-to-trial or trial-to-patient ranking; PRISM outperforms human annotators

Talk Outline

1. **Motivation**
    a. What is clinical trial patient recruitment, and why is it hard?
2. **Prior Work**
    a. What did people try before LLMs?
3. **Papers**
    a. Zero-shot patient matching with off-the-shelf LLMs
    b. PRISM: Fine tuning an LLM for clinical trial matching
4. **Future Work**

Lots of opportunities to push research forward!

1. **Scaling to all patients**
   a. All actively enrolling interventional trials on CT.gov (**~20k**)
   b. All patients in a health system (**10M's of notes**)

2. **Real-world deployments**
   a. Evaluations on **actively enrolling** trials (Unlu et al. 2024)

3. **Patient-facing tools**
   a. Can we **disintermediate doctors** and directly help patients? antidote SubjectWell

Thank you to everyone in the Shah Lab + beyond!

People

- Nigam Shah
- Chris Ré
- Jason Fries
- Dev Dash
- Jenelle Jindal
- Kenneth Mahaffey
- Alison Callahan
- Frazier Huo
- Akshay Swaminathan
- Ethan Steinberg
- Suhana Bedi
- Hejie Cui
- Miguel Fuentes
- Alyssa Unell
- Mehr Kashyap
- Jonathan Chen
- Keith Morse
- Duncan McElfresh
- Nikesh Kotecha
- Aditya Sharma
- Alejandro Lozano
- Lionel Jeremiah

Research

https://clinicaltrialmatch.stanford.edu

Contact: mwornow@stanford.edu

Funding

Stanford HEALTH CARE
STANFORD MEDICINE

Thanks!