

Lecture 3

Accountable AI: Let AI Read, Let Logic Decide

VERDICT: Clinical Trial Matching with SMT

Cyrus Zhou and Monica Lam

[Accountable AI with Grounded, Faithful, Consistent, Actionable Rationales: A Case Study in Clinical Trial Matching with VERDICT](#)

Zikai Zhou, Yufei Jin, Yilin Xu, Yu-Chiang Wang, Chieh-Ju Chao, Monica S. Lam

In The 2026 Conference on Empirical Methods in Natural Language Processing (EMNLP) , Budapest, Hungary, 2026.

DECISION MAKING IN HIGH STAKES DOMAINS

MEDICINE

MATCHING PATIENTS WITH CLINICAL TRIALS

WHAT IS THE DIAGNOSIS AND TREATMENT FOR THIS PATIENT?

FINANCE

ARE REGULATIONS FOLLOWED FOR EACH CLIENT?

ARE CONTRACTS FOLLOWED FOR EACH TRANSACTION?

LEGAL

HOW TO DECIDE ON A DISPUTE ACCORDING TO THE LAW, CONTRACT?

TRUSTWORTHINESS: IMPROVE THE ACCURACY OF AI

AT WHAT ACCURACY CAN AI BE TRUSTED?

AI-Based Decisions

- OpenEvidence is correct 73% on MedExpertQA
- **Is it good enough?**
 - Can we deploy without human expert judgment?
 - Is it useful for an expert doctor?
 - Is it useful for a new doctor?

Why is it not enough to just say “Yes” or “No”

90% accurate → 10% of the users are mishandled!

We need **human-in-the-loop**: AI must provide a point-by-point **rationale**

- Experts need to check the argument
 - **Why** the decision?
 - What are the **assumptions**?
- Ambiguity: Experts need to control the decision with policies
 - Does the decision **adhere to the policies**?
- Experts and users need to follow up on the decision: e.g.
 - If the answer is no, an applicant wants to know how to become eligible
 - If the answer is yes, the doctor may need to change the diagnosis as certain patient condition changes.
 - What would **change** the decision? (**pivotal condition**)

Grounding
Faithfulness

Consistency

Actionability

Accountable Decision Making

Grounding: Why?

Is the stated reason the real reason?

Faithfulness: What was assumed?

Missing facts, filled in silently?

Assumptions: a check list to ensure they are true

Consistent: Following policies?

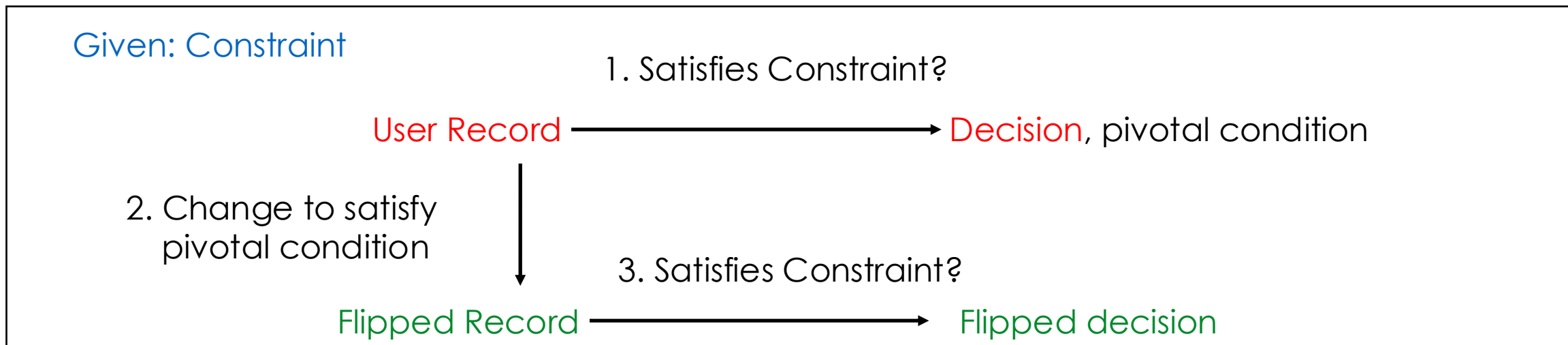
Same facts, same answer?

Actionability: What would change it?

Pivotal condition –

A path to a different outcome

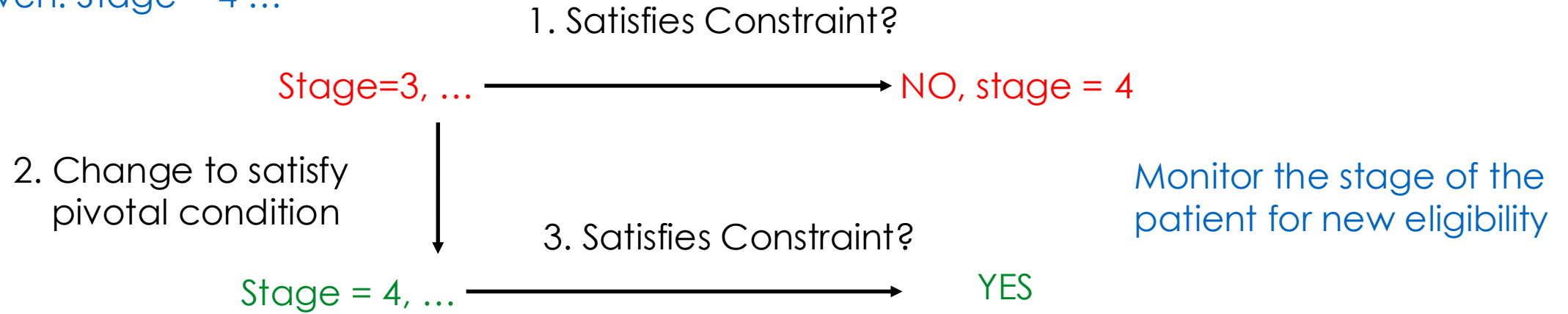
What are Pivotal Conditions?



- Pivotal Condition: "what would change the decision?"
- If we flip the user's record based on the pivotal condition then the decision should change.

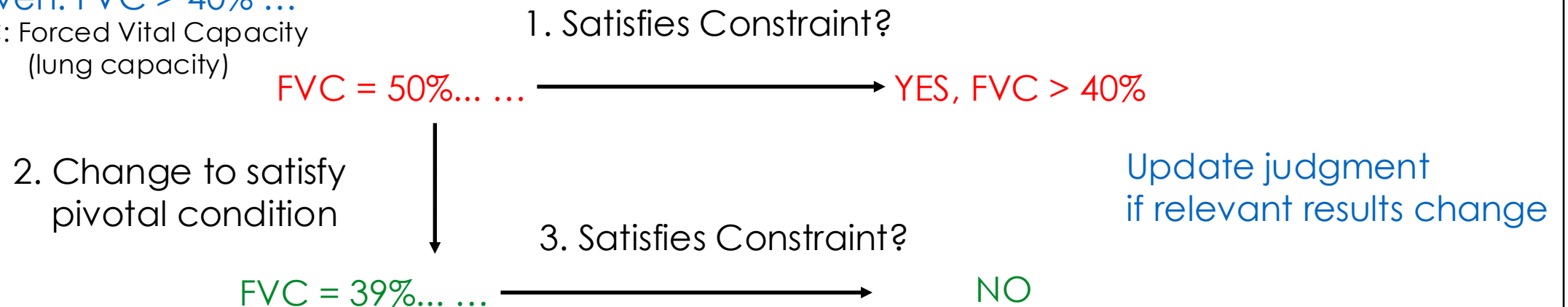
Pivotal Conditions are Useful in Practice

Given: Stage = 4 ...



Given: FVC > 40% ...

FVC: Forced Vital Capacity
(lung capacity)



Pivotal conditions are applicable to all domains

Given: safe-driver discount if
accident free for 3 yrs..

1. Satisfies Constraint?

Accident last year, ... → NO, no accident for 3 yrs

2. Change to satisfy
pivotal condition

3. Satisfies Constraint?

Three years past w/o accidents ... → YES

Remind client how
they can get discount

Given: safe-driver discount if
accident free for 3 yrs..

1. Satisfies Constraint?

No accidents in last 3 years → YES, no accident for 3 yrs

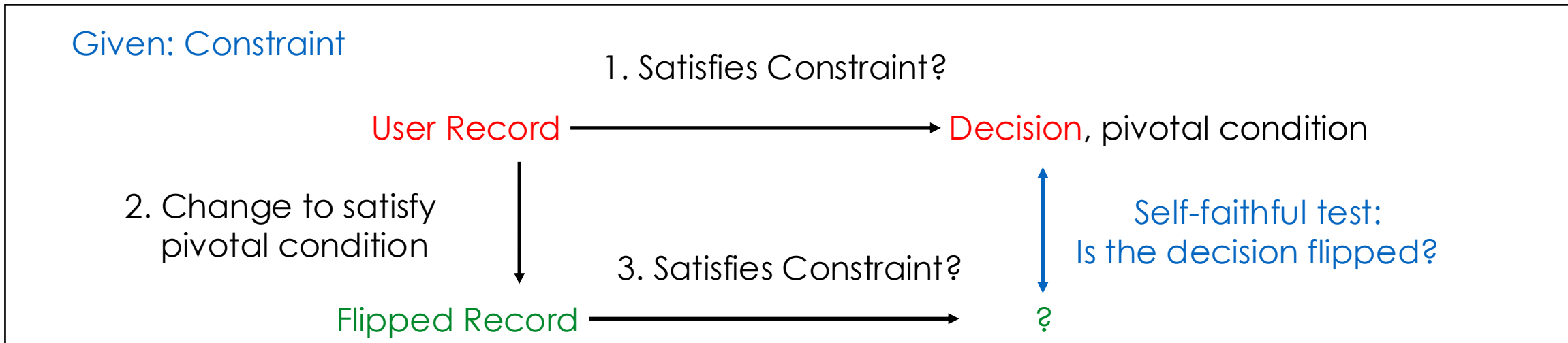
2. Change to satisfy
pivotal condition

3. Satisfies Constraint?

Accident in next year → NO

Warn client that they
can lose discount if
they get into an accident

Self-faithful matchers



- A self-faithful matcher: Flipping the reported pivotal condition flips the decision!
- However, a matcher may make mistakes
- **Self-faithfulness is a useful automatic metric for evaluating rationales**
 - Measures % time flipping the record (according to the pivotal condition) flips decision

Outline

- **Decision making with AI**
 - Introduction to accountability
- **A real use case: Clinical trial matching**
 - Complexity in missing data, handling of policies, and scale
- **LLM-based decision making: ChatGPT**
 - Retrieval has poor recall; matching accuracy is OK, poor accountability
- **VERDICT: leveraging a theorem prover**
 - Semantic parse from LLM to formal representation is challenging
- **Evaluation and comparison**
 - An automatic evaluation metric: Self-faithfulness
 - VERDICT has similar accuracy! But much better accountability

Outline

- **Decision making with AI**
 - Introduction to accountability
- **A real use case: Clinical trial matching**
 - Complexity in missing data, handling of policies, and scale
- **LLM-based decision making: ChatGPT**
 - Retrieval has poor recall; matching accuracy is OK, poor accountability
- **VERDICT: leveraging a theorem prover**
 - Semantic parse from LLM to formal representation is challenging
- **Evaluation and comparison**
 - An automatic evaluation metric: Self-faithfulness
 - VERDICT has similar accuracy! But much better accountability

Clinical trials

For drug manufacturers: trials are needed before releasing new drugs

Trial Phases	Volunteers	Main question
1	a few dozen	Is it safe? What dose?
2	a few hundred	Does it seem to work?
3	thousands	Better than prior care?

For some patients, a trial is the only route to the newest treatment.

What is the Problem?

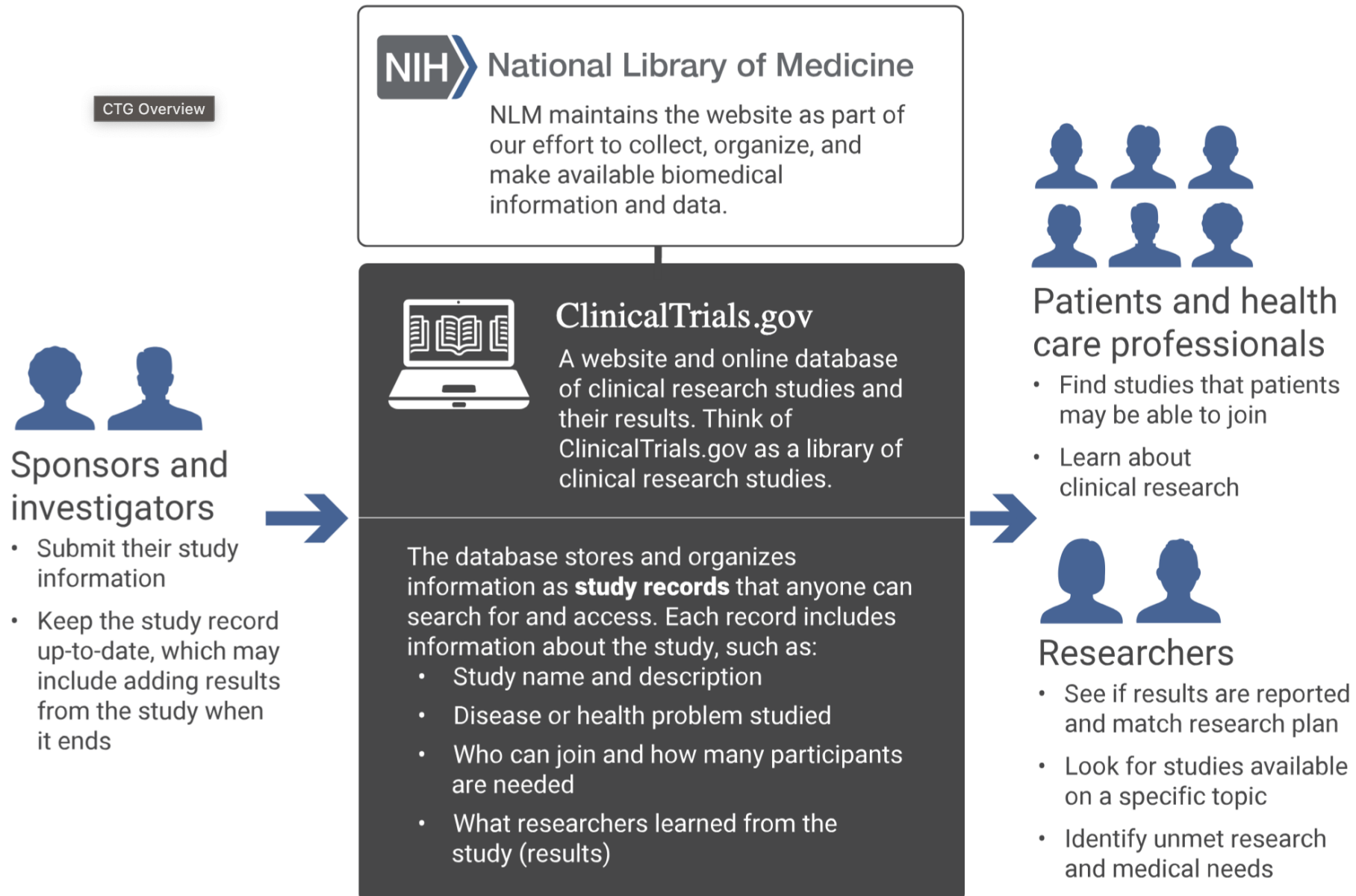
~80%

of trials miss enrollment targets or timelines

\$1.9B

spent on recruitment per year

All trials in US are in ClinicalTrials.gov!



CAVEAT

THE TRIALS USED IN PRACTICE ARE MORE COMPLICATED
THAN THOSE IN CLINICALTRIALS.GOV

WE HAVE ACCESS TO REAL CLINICAL TRIAL CONSTRAINTS
THROUGH COLLABORATION WITH STANFORD MED SCHOOL

Full Example

Eligibility Criteria	
Description	
Inclusion Criteria: • Age ≥ 18 years at time of randomization (18 years and older); • Acute myocardial infarction; • Reperfusion of the infarct-related coronary artery in terms within 12 h of symptom onset; • Written the informed consent to participate in research;	Ages Eligible for Study ⓘ 18 Years and older (Adult, Older Adult)
Exclusion Criteria: • Inability to obtain informed consent; • Patients previously undergone endovascular / surgical revascularization of coronary artery; • Severe comorbidity; • History of myocardial infarction; • History of intracranial haemorrhage; • Pulmonary edema, cardiogenic shock; • Creatinine clearance <30 mL/min or dialysis; • Unable to undergo or contra-indications for MRI; • Allergy for contrast agent; • Indication or use of oral anticoagulant therapy; • Major bleeding; • Atrio-ventricular block II and III degree; • Active gastroduodenal ulcer; • Aortic dissection; • Acute psychotic disorders	Sexes Eligible for Study ⓘ All Accepts Healthy Volunteers ⓘ No

Eligibility criteria are constraints written in NL

- Boolean structure:
 - and / or / one of / not both
- Thresholds:
 - $\text{CrCl} \geq 60 \text{ mL/min}$
- Time windows:
 - no chemo within 6 months
- Exceptions:
 - unless disease is stable
- Qualifiers:
 - documented, at screening visit

An example excerpt:

Tofersen trial, simplified (SatIR Fig. 1)

*“Weakness attributable to ALS and documented SOD1 mutation.
FVC $\geq 50\%$ of predicted;
stable FVC $< 50\%$ but $\geq 45\%$ may be considered.
If taking riluzole, dose must be stable.”*

FVC: Forced Vital Capacity (lung capacity)

3 pages of a real patient record

Hospital Medicine — Admission H&P

Patient: J.D. (F, 52) • **MRN:** 1xxxxxx • **Room:** 6W-6213
Date/Time: 2025-09-21 18:42 PT • **Admitting Service:** Hospital Medicine-B
Attending: Alex Kim, MD • **Resident:** Priya Shah, MD • **PCP:** Unknown
Code Status: Full
Isolation: None
Language: English
Insurance: Commercial PPO

Chief Complaint

“Chest pressure and nausea.”

History of Present Illness

52-year-old female with HTN and HLD presenting with acute central chest pressure starting ~12:30 PT while at rest, radiating to L arm, associated with diaphoresis and nausea, no syncope. Pain 8/10, non-pleuritic, non-positional. EMS EKG reportedly ST elevations anteriorly; ASA given by EMS? **unclear**. In ED: EKG with 2 mm STE V2–V4; initial trop-I 0.42 ng/mL ↑. Given ASA 325 mg PO, ticagrelor 180 mg PO, heparin 4,000 U IV bolus. Cath lab activated; patient transferred for emergent PCI at 13:35.

Cath Lab (brief op note): Proximal LAD 99% thrombotic lesion → drug-eluting stent ×1 with TIMI-3 flow restoration. LVDP mildly elevated. No complications. Total contrast 90 mL. Fluoro time 12.8 min.

Post-PCI pain resolved (0–1/10). Admitted to CCU then transferred to floor on 6W at 18:10 in stable condition on DAPT.

Prior episodes: Intermittent exertional chest tightness over last 2–3 months climbing stairs, self-limited; never evaluated.

Precipitating factors today: At rest; possible recent work stress; slept 4–5 h/night for past week. No recent illness, immobilization, or surgery.

ED Course (chronological): See nursing and ED MD notes copied below under *Copied Forward Content*.

Problem List (active)

- Acute anterior STEMI s/p PCI to proximal LAD (09/21/2025)
- Hypertension (2016)
- Hyperlipidemia (2018)

4. Overweight (BMI 29)

Allergies

- Penicillin — rash (non-anaphylactic)

Home Medications (patient report, med reconciliation completed by pharmacy)

- Lisinopril 10 mg PO daily (last fill 2 months ago; reports good adherence)
- Atorvastatin 40 mg PO nightly
- OTC: multivitamin daily
- Denies herbals/supplements

Social History / SDOH

- Lives with spouse in apartment (elevator access); independent in ADLs/IADLs.
- Work: retail manager; high stress; on feet most of day.
- Tobacco: never.
- Alcohol: 1–2 drinks/week.
- Illicits: denies.
- Transportation: drives.
- Food security/housing security: stable.
- Prefers afternoon appointments due to work schedule.

Family History

- Father MI at 58; deceased at 72 from stroke.
- Mother T2DM, HTN.

Review of Systems (template — positives in bold)

- Constitutional:** diaphoresis at onset; no fever/chills, no weight loss.
- Cardiac:** chest pressure, radiation L arm; no palpitations; no LE edema.
- Respiratory:** no SOB at rest, no cough, no wheeze.
- GI:** nausea; no vomiting/diarrhea, no melena/hematochezia.
- GU:** no dysuria, no hematuria.
- Neuro:** no syncope, no focal weakness, no speech changes.
- MSK:** no calf pain, no recent trauma.
- Derm:** no rashes currently.
- Endo:** no heat/cold intolerance.
- Psych:** baseline stress; no SI/HI.

Auto-imported ROS block (system default) — 14-pt ROS negative except as noted above.

Vitals (selected)

ED Triage (12:48): BP 162/94, HR 104, RR 20, Temp 36.8 °C, SpO₂ 98% RA, Wt 78 kg, Ht 164 cm (BMI 29.0)
Post-PCI (15:00): BP 138/82, HR 88, RR 16, SpO₂ 98% RA
On transfer to floor (18:15): BP 134/79, HR 82, RR 16, Temp 36.7 °C, SpO₂ 98% RA

Intake/Output (last 24 h): Intake 450 mL; Output 300 mL; Net +150 mL.

Physical Exam (on floor 18:25)

- General:** alert, oriented, comfortable, no distress.
- HEENT:** PERRL, anicteric; MMM.
- Neck:** supple; no JVD at 30°.
- CV:** RRR, no murmurs/rubs/gallops.
- Lungs:** CTAB, normal WOB.
- Abd:** soft, NT/ND, +BS.
- Ext:** warm, well-perfused; no edema; R radial access site clean/dry/intact with TR band removed at 17:30; distal pulses 2+.
- Neuro:** AO×3, non-focal; speech clear; moves all extremities.
- Skin:** no rash.

Labs (auto-imported)

CBC
09/21 13:10 — WBC 8.2, Hgb 13.6, Hct 40.3, Plt 248
BMP
09/21 13:10 — Na 139, K 4.2, Cl 103, CO₂ 24, BUN 15, Cr 0.86, Glu 146
Mg/Phos
09/21 13:10 — Mg 1.9, Phos 3.4
Cardiac markers
09/21 13:10 — Troponin-I **0.42** (↑)
09/21 17:40 — Troponin-I **2.3** (↑↑)
Lipid panel (fasting pending) ordered for 09/22 06:00
HbA1c ordered for 09/22 06:00

Note: Full lab history available in chart. Above values auto-pulled; see results tab for reference ranges.

Microbiology

- None pending. COVID/flu not indicated.

Imaging / Studies

EKG (ED 13:04): Sinus tachy 102. 2 mm ST elevations V2–V4 with reciprocal depressions II/III/aVF.

Complexities in real-life decision making

1. Objective alignment

Trial intent

treat · prevent · diagnose

Patient objectives & preferences

chief complaint, geographic, treatment choice, ...

2. Constraints can be hard or soft

Text	Kind
Must have documented SOD1 mutation	Hard
Stable FVC 45–50% may be considered	Stable: soft / permissive
Able to attend weekly visits	Operational (ignored in matching)

Complexities in real-life decision making

3. Data missingness

- **Most** facts needed are missing
- Data **imputation**: we need to guess the values
- We hypothesize a guidance: if a fact is **salient**, it will be mentioned in the record.
 - Patient record with “appendix” condition satisfies trial asking for “acute appendix” but not “burst appendix”
- Imputations become **assumptions**: must disclose to expert, in case expert disagrees

4. Control with policies

- **Experts** do not want to leave it to LLMs “probabilistic solution”
→ No consistency, different treatments to different applicants!
- Example: Unspecified biomarker reads as negative

Policies: expressed in natural language

By type

e.g.
any unreported biomarker → negative.

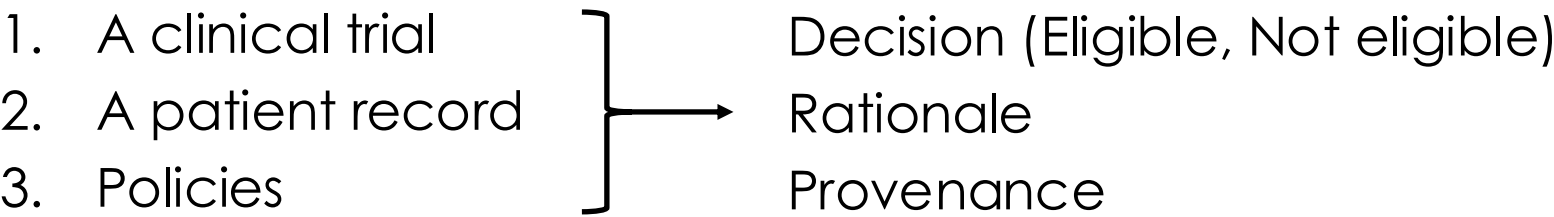
By direction of Constraints	Silence suggests...
Exclusion	probably fine (soft)
Inclusion	probably doesn't qualify (soft)

Context-dependent

e.g.
a missing lab is OK for prescreening

Problem statement for matching against constraints

Given



Rationale

- Reasons why (grounding)
- Assumptions made (faithfulness)
- Pivotal condition (actionability)

Provenance: how we derive the patient info

Info needed for matching	Value from	Source
OBSERVED	the record	evidence span
IMPUTED	a policy	which policy, how
UNRESOLVED	--	—

Why is clinical trial hard?

- **Complexities in real-life decision making**
 - Hard and soft constraints
 - Objectives and preferences
 - Data imputation, assumptions, policies
- **Logic: must hold jointly, not on average**
- **Vocabulary: synonyms and hierarchies**
- **Scale: 500k trials × dozens of rules**

~80%

of trials miss enrollment targets or timelines

\$1.9B

spent on recruitment per year

recall

precision

efficiency

controllability

interpretability

COMPLEXITIES IN CLINICAL TRIAL MATCHING

- Hard and soft constraints
- Objectives and preferences
- Data imputation, assumptions, policies

QUIZ: HOW DID WE FIGURE THEM OUT?

CLINICAL TRIAL MATCHING HAS MULTIPLE SUBPROBLEMS

Two directions of matching

Trial → patients

One constraint set × many records.
A standard database query.

Patient → trials

One record × 500,000+ constraint sets.
Find every trial the patient satisfies.

Other Examples of Constraint Matching

Constraints	Candidates
Job Requirements	Resumes
Diagnosis guidelines	Patient records
System fault diagnosis	Conditions

Quiz: Which direction is harder?

Finding trials has 2 subproblems

How to retrieve relevant trials?

- Find the likely candidates with high recall
- Goal: 1. perfect recall: 0 false negatives (missed trials)
2. as accurate as possible: few false positives
- Preview of **future lecture**:
New retrieval technique: **constraint-satisfaction based**
Much better recall and accuracy than embedding

Given a short list of trials: how to match a trial with a patient?

- High accuracy and **high accountability** (this lecture)

Outline

- **Decision making with AI**
 - Introduction to accountability
- **A real use case: Clinical trial matching**
 - Complexity in missing data, handling of policies, and scale
- **LLM-based decision making: ChatGPT**
 - Retrieval has poor recall; matching accuracy is OK, poor accountability
- **VERDICT: leveraging a theorem prover**
 - Semantic parse from LLM to formal representation is challenging
- **Evaluation and comparison**
 - An automatic evaluation metric: Self-faithfulness
 - VERDICT has similar accuracy! But much better accountability

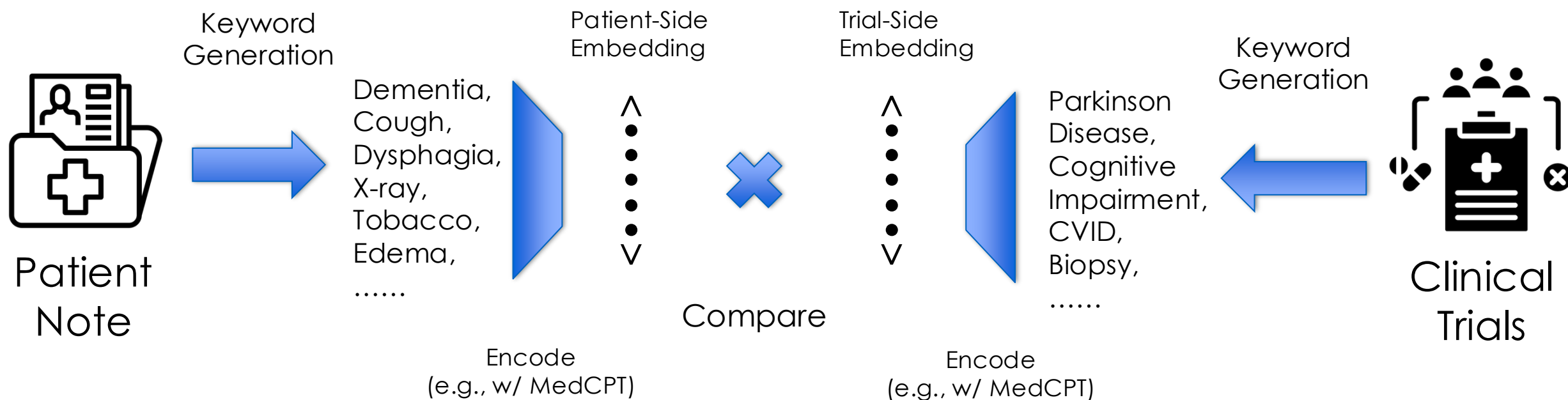
Architecture of typical LLM Matchers



Example: TrialGPT (Jin et al., Nature Communications 2024)

TrialGPT: Finding Trials for Patients

1. Retrieval based on embedding similarity



2. Matching: Use LLM to patch patient record with retrieved trials

Two Problems with TrialGPT

Retrieval: constraints as soft signals

Negation, thresholds, time barely move a similarity score.

Ineligible-but-relevant trials crowd out eligible ones.

Judgment: plausible ≠ faithful

Rationales are generated to sound right,
not to trace the derivation
(Jacovi & Goldberg 2020; Lanham et al. 2023).

Top results by similarity

✗ Similar words, ineligible patient

✗ Similar words, ineligible patient

✗ Similar words, ineligible patient

✗ Similar words, ineligible patient

✗ Similar words, ineligible patient

✓ The eligible match, ranked last

Benchmarks

Real clinical trials

Synthetic patient vignettes

SIGIR

59 synthetic patient vignettes
3,621 ClinicalTrials.gov trials

TREC 2021

50 synthetic patient vignettes
26,581 trials

Recall of Dense Retrieval

	Recall	#Retrieved
SIGIR	93%	500
TREC 2021	92%	1000
TREC 2022	88%	1000

QUIZ:
What if the life-saving trial
is not retrieved?

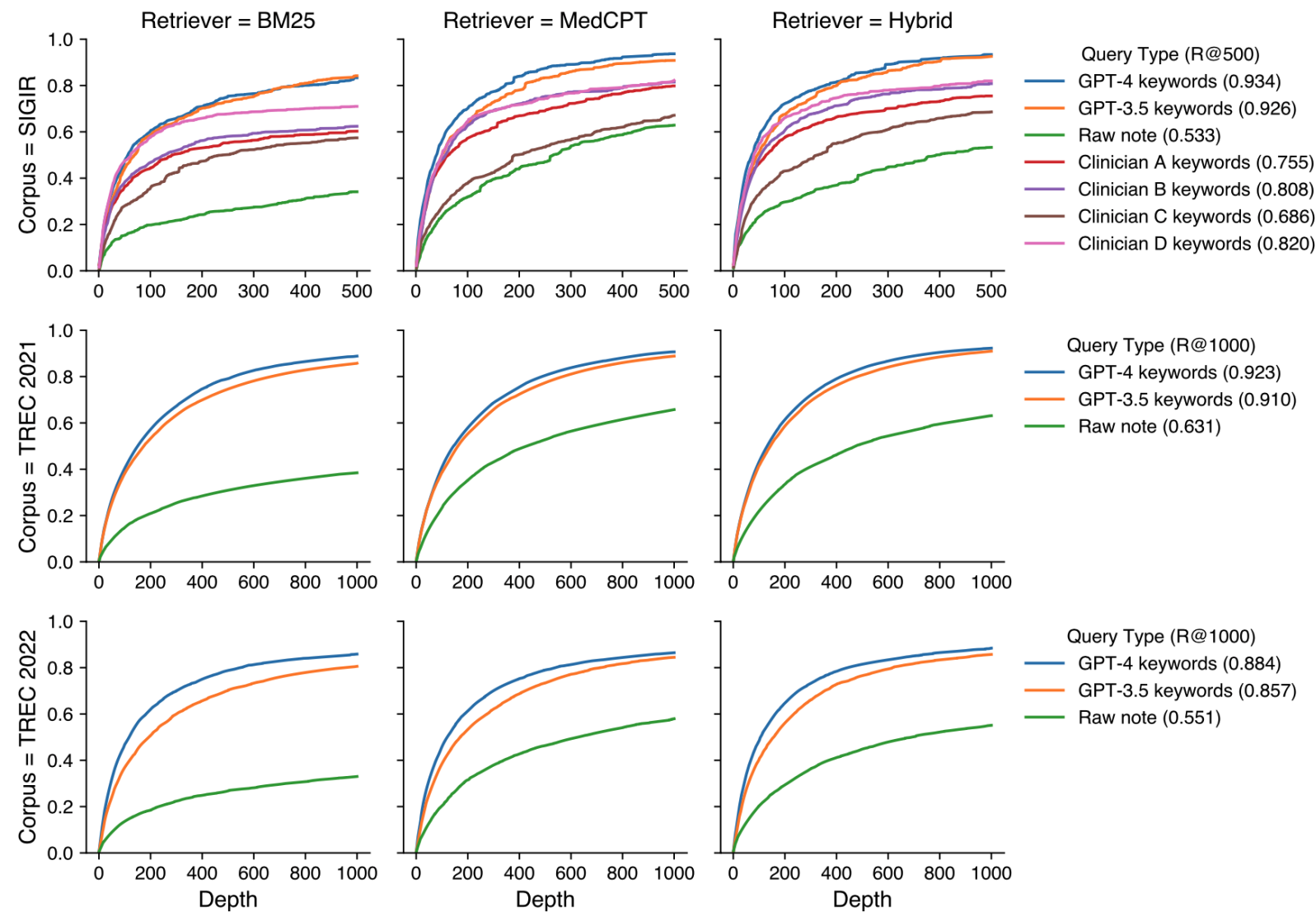


Fig. 2 | First-stage retrieval results. a Overview of TrialGPT-Retrieval. LLMs first generate a list of keywords for a given patient note. These keywords are used to derive the keyword-level relevant clinical trials, which are then fused to generate a final ranking. **b** Recalls of relevant clinical trials at different depths for various query

types and retrievers. The hybrid retriever combines the results of the BM25 (lexical matching) and the MedCPT (semantic matching) retrievers. Source data are provided as a Source Data file.

LLM-based matching

LLM-based pipelines have around 90% accuracy, but poor accountability

Examples of inaccurate matching

Invented blocker (grounding)

Trial: Obstructive Sleep Apnea
Patient: 10-year-old with sleep apnea symptoms

LLM matcher result

Decision: ineligible

Reason: His age,
since the trial does not specify
pediatric inclusion.

Problem: The trial has no age criterion.

Silent assumption (faithfulness)

Trial: Pulmonary Embolism
Patient: Suspected Pulmonary Embolism
after hip surgery.

LLM matcher result:

Decision: ineligible

Reason: Imaging and labs not in the note.

Problem: Missing treated as failed.

QUIZ:

**HOW DOES AN LLM MATCHER HANDLE THE
COMPLEXITIES IN CLINICAL TRIAL MATCHING?**

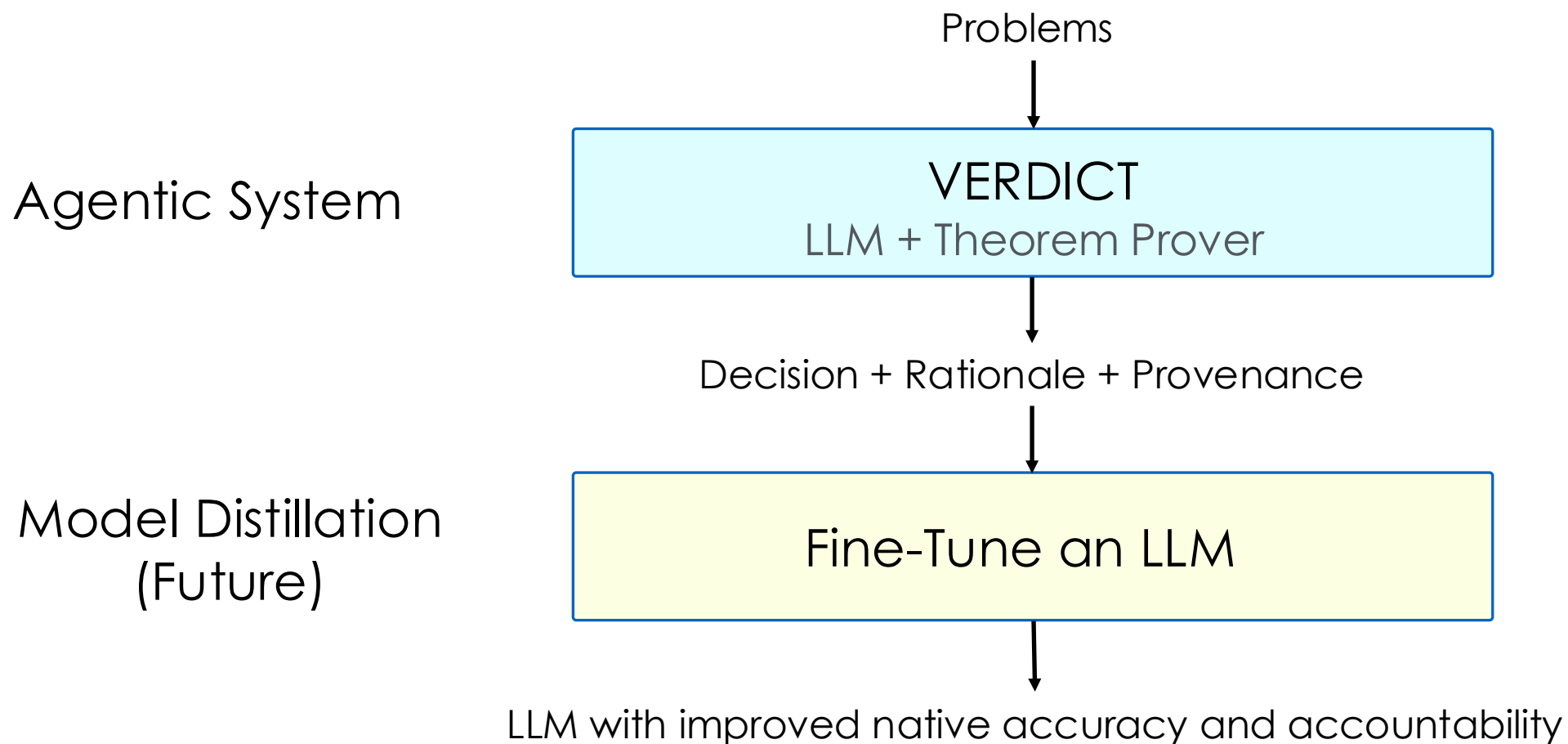
- Hard and soft constraints
- Objectives and preferences
- Data imputation, assumptions, policies

HOW TO IMPROVE ACCOUNTABILITY?

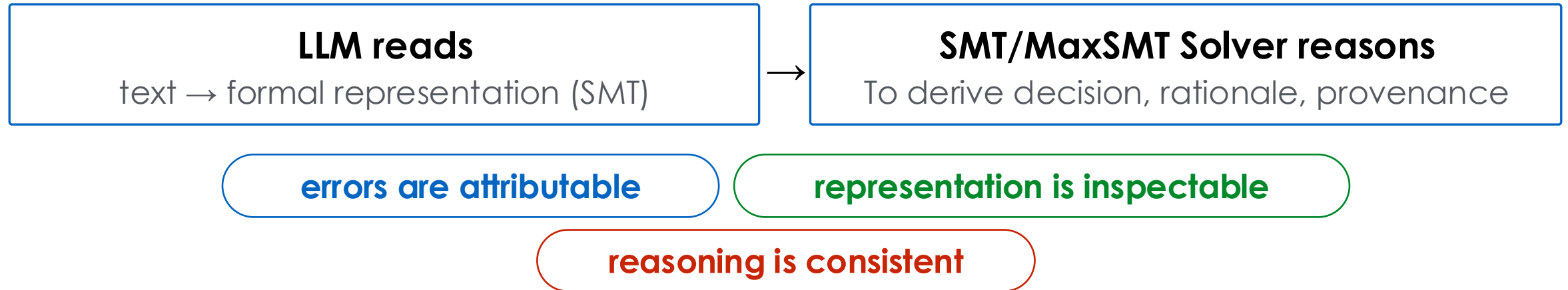
Outline

- **Decision making with AI**
 - Introduction to accountability
- **A real use case: Clinical trial matching**
 - Complexity in missing data, handling of policies, and scale
- **LLM-based decision making: ChatGPT**
 - Retrieval has poor recall; matching accuracy is OK, poor accountability
- **VERDICT: leveraging a theorem prover**
 - Semantic parse from LLM to formal representation is challenging
- **Evaluation and comparison**
 - An automatic evaluation metric: Self-faithfulness
 - VERDICT has similar accuracy! But much better accountability

Tackle complexity with a theorem prover



VERDICT: Separate understanding from reasoning



Let AI read. Let logic decide.

Logic-based reasoning has been used before

SatLM

LLM writes a declarative spec;
a SAT solver solves it

Logic-LM

Several solvers + self-refinement
from solver errors

Lang2LTL

Commands →
linear temporal logic for planners

	Previous AI Benchmarks	Clinical trial matching
Information	complete	mostly missing
Vocabulary	a few entities	Large medical ontology (350k+ terms & relations)
Setting	one answer	scale + a clinician in the loop

Introduction to Formal Notations

- **Propositional Logic (PL)**
 - Variables: Boolean
 - Operators: and, or, not
 - Conjunctive Normal Form: $A \text{ and } (B \text{ or } C) \text{ and } (A \text{ or } D) \dots$
 - SAT solver: checks if a set of PL statements are satisfiable
- **SMT (Satisfiability Modulo Theories)** ← More expressive than PL,
SMT solver more expensive than SAT solver
but cheap compared to LLM matching
 - Variables: Boolean, number, ...
 - Operators: and, or, not, arithmetic, numeric comparison
 - Conjunctive Normal Form:
 $A \text{ and } (B > 10 \text{ or } C + B > D) \text{ and } (A \text{ or } E) \dots$
 - SMT solver: checks if a set of SMT statements are satisfiable

Satisfiability: can all of these be true at once?

```
(declare-const a Bool)
(declare-const b Bool)
(declare-const c Bool)
(assert (=> a b))
(assert (not (and b c)))
(assert c)
(check-sat) ; sat
; model: c=true, b=false, a=false
```

Add:

```
(assert a)
(check-sat) ; unsat
; core: a, a→b, ¬(b ∧ c), c
```

- SAT → a model
(a witness assignment)
- UNSAT → an unsat core
(the reason of UNSAT)

Z3 -- An SMT Solver

- Open-source solver by Microsoft Research, released 2007
- Used in analysis, verification, testing for software and hardware,
 - E.g. Does there exist an input that causes errors in a program (CS 243)
- NP-complete:
 - Propositional satisfiability
 - Uninterpreted functions, linear integer arithmetic
- Highly optimized today with continuous research to this date

Matching with SMT

Trial Req: {
 (assert (or has_cough has_fever))
 (assert (>= age 18))
 (assert (not is_pregnant))

Patient Fact

Trial Req + {
 (assert (= age 67))
 (assert has_cough)
 (assert (not is_pregnant))

→ **Z3** → **SAT**

Patient Fact

Trial Req + {
 (assert (= age 67))
 (assert (not has_cough))
 (assert (not has_fever))
 (assert (not is_pregnant))

→ **Z3** → **UNSAT**

MaxSMT: Maximum SMT (solved by Z3)

Given a Boolean formula (a set of clauses)

SMT

Does there exist an assignment to variables
that makes the formula *completely* true (satisfiable)?

MaxSMT: maximize the number of satisfied clauses

Find an assignment
that makes the *maximum number of clauses true*

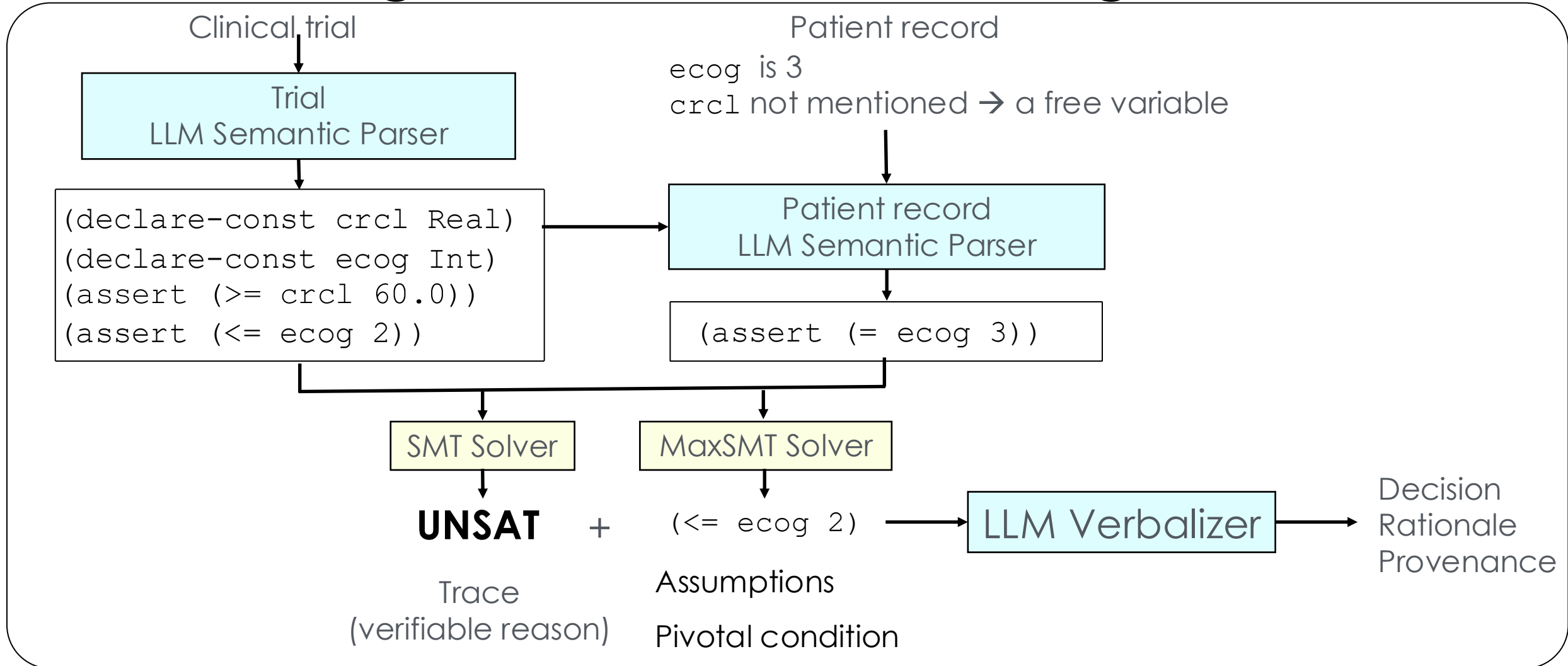
Weighted MaxSMT: maximize weighted sum of satisfied clauses

Hard constraints: infinite weight

Soft constraints: finite weights

Unsatisfied clauses of answer → Violated set (pivotal condition)

Formulating clinical trial matching



SMT decides; MaxSMT further explains

SMT returns a decision SAT/UNSAT, a trace (**verifiable rationale**), model/unsat core

Use MaxSMT to find assignment to unknown variables to maximize eligibility (**Assumptions**)

If SMT returns UNSAT (INELIGIBLE)

- Use MaxSMT to find pivotal conditions that turn INELIGIBILITY to ELIGIBILITY
 - Hard constraints in trial (infinite weight)
 - Soft constraints (weight 1)
 - **Pivotal conditions**: soft facts violated by the optimum returned by MaxSMT

SMT returns SAT (ELIGIBLE)

- Use MaxSMT to find pivotal conditions that turn ELIGIBILITY to INELIGIBILITY
 - Hard constraints: $\neg$ (original constraints) (infinite weight)
 - Soft constraints: original soft constraints and assumptions (weight 1)
 - **Pivotal conditions**: soft facts violated by the optimum returned by MaxSMT

Example: Stage IV NSCLC, EGFR trial

(VERDICT Fig. 1)

Trial constraint

$(ecog \leq 2) \wedge egfr \wedge \neg platinum$
 $\wedge (crcl \geq 60) \quad ; + \text{ stage IIIB-IV}$

Patient constraint

Condition	Value	State
ecog	1	OBSERVED
platinum	false	OBSERVED
egfr	false	IMPUTED (Policy 1)
crcl	$\perp$	UNRESOLVED

- SMT: UNSAT $\rightarrow$ INELIGIBLE,
- MaxSMT:
pivotal condition { egfr = false }
- Assumption assigned by model:
crcl = 72 (to satisfy $crcl \geq 60$)

What the clinician sees

VERDICT

Decision: Not eligible as recorded

Why	an activating EGFR mutation is required; No result is on file.
Imputation	policy P1 reads the missing EGFR test as negative, not observed.
Assumption	no creatinine clearance on file, no policy supplies one; assume ≥ 60 .
To reverse	confirm EGFR status.

Monolithic LLM

"Likely ineligible — no EGFR result on file."
No artifact to inspect or rerun.

Returning to the Examples of LLM Failures

Invented blocker (grounding)

Trial: Obstructive Sleep Apnea
Patient: 10-year-old with sleep apnea symptoms

LLM matcher result

Decision: ineligible

Reason: His age,
since the trial does not specify
pediatric inclusion.

Problem: The trial has no age criterion.

Silent assumption (faithfulness)

Trial: Pulmonary Embolism
Patient: Suspected Pulmonary Embolism
after hip surgery.

LLM matcher result:

Decision: ineligible

Reason: Imaging and labs not in the note.

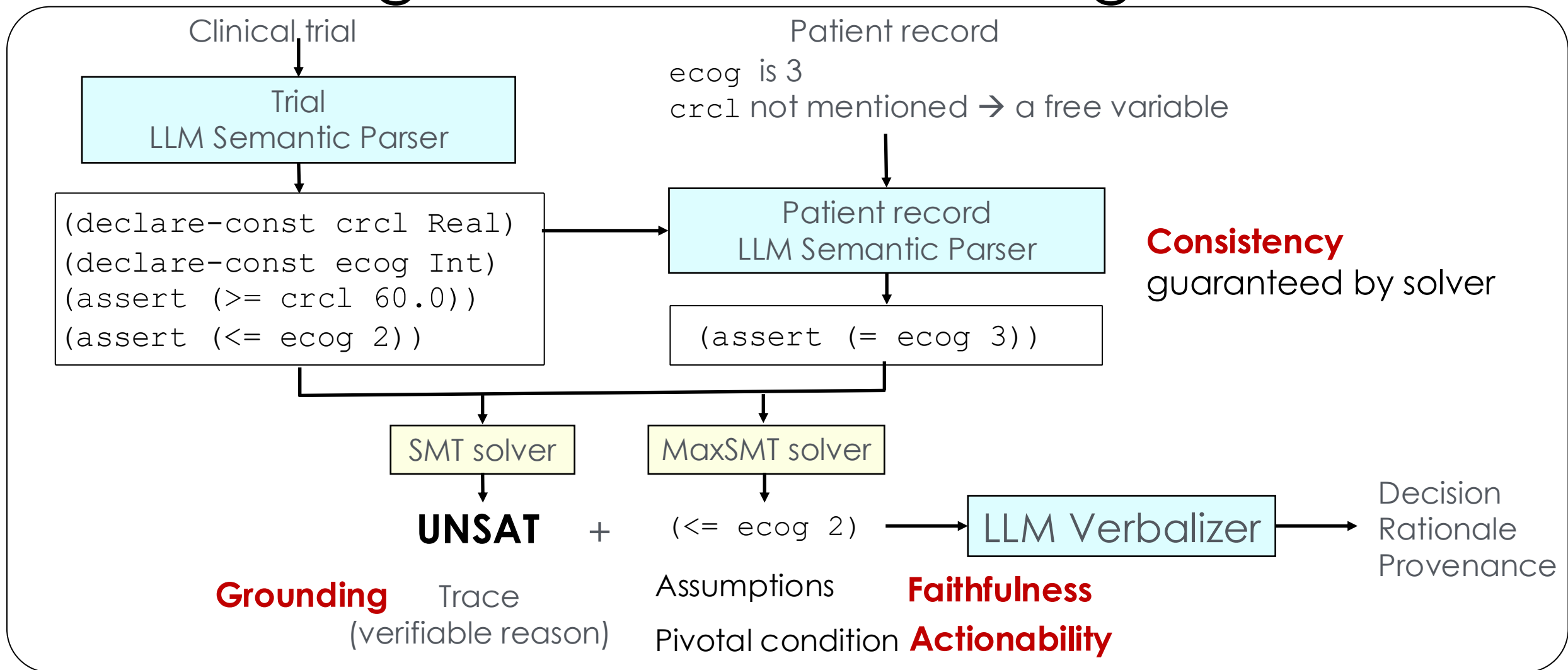
Problem: Missing treated as failed.

VERDICT

Decision: Eligible in both,

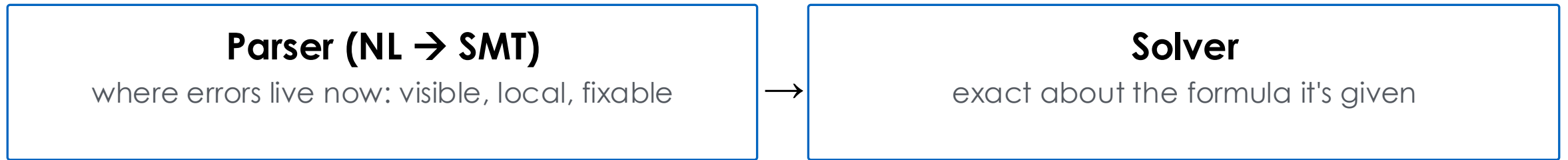
Result: Includes a list of items to confirm at the in-person visit.

Formulating clinical trial matching



Quiz: Are we 100% accurate and accountable?

VERDICT: Errors confined to LLM parser



Formalism is not easy to work with!

The solver faithfully executes whichever the parser chooses.

LLMS need to write down clearly how to interpret the NL → accountability

Example:

FVC is greater than 50%;

If FVC is stable 45–50% may be considered

```
(declare-const fvc_pct Real)
(declare-const fvc_stable Bool)
(assert (or (>= fvc_pct 50.0)
            (and (>= fvc_pct 45.0) fvc_stable)))
```

- How to handle "May be considered"
 - ignored
- If "Stable" is not in the note:
 - what do we assume?
 - treated as a free variable
 - solver choose value to get a SAT decision.
 - (The value is the assumption)
- Should we use $\geq$ or $>$?
 - simple decisions can change the final outcome completely

Details of semantic parser in the next lecture

QUIZ:

**CAN LOGIC-BASED BE WORSE IN MATCHING
THAN LLM MATCHING?**

Outline

- **Decision making with AI**
 - Introduction to accountability
- **A real use case: Clinical trial matching**
 - Complexity in missing data, handling of policies, and scale
- **LLM-based decision making: ChatGPT**
 - Retrieval has poor recall; matching accuracy is OK, poor accountability
- **VERDICT: leveraging a theorem prover**
 - Semantic parse from LLM to formal representation is challenging
- **Evaluation and comparison**
 - An automatic evaluation metric: Self-faithfulness
 - VERDICT has similar accuracy! But much better accountability

EVALUATION

HOW TO EVALUATE ACCOUNTABILITY?

- 1. HUMAN JUDGE (NECESSARY, EXPENSIVE)**
- 2. SELF-FAITHFULNESS – AN AUTOMATIC METRIC!**

Benchmarks

Real clinical trials

Synthetic patient vignettes

SIGIR-derived labels

552 patient-trial patients

5 GPT-5 judges with different review framings;

Use majority vote -- 80.8% unanimous.

Clinician audit of a stratified sample.

TREC 2021

363 patient-trial pairs

Official human judgments are based on
relevance and not eligibility

Caveat

No benchmark states a missing-data policy:
Accuracy partly measures agreement with
an unstated one.

What matching costs

SatIR: pay once per trial

Heavy one-time parsing per trial,
amortized over every patient;
then fast retrievals (0.15s)

~45 min

552 pairs · 8 threads · ≈4,000 GPT-4.1 calls

< \$ 1,000

final reported runs (≈\$2,500 incl. early runs)

Accuracy: VERDICT similar to LLM matchers

BASELINES

LLM-based matcher:

TrialGPT, CoT LLM, ZSPM, LLMMatch

Logic-based matcher: DLSC

SIGIR-derived, GPT-4.1, F1



TREC 2021 F1	VERDICT	best LLM	DLSC
GPT-5-mini	0.828	0.776	0.700
Claude Haiku 4.5	0.800	0.704	0.700
Qwen2.5-7B	0.738	0.695	0.470

*LoRA-tuned Qwen2.5-7B formalizer:
0.738 → 0.829 F1 ($\approx$ GPT-5-mini).*

Four pillars of **accountability**

Pillar	How LLMs fail it	Countermeasure
Grounding	reason $\neq$ the real reason	verified rationales
Faithfulness	gaps filled silently	assumptions (explicit ledger)
Consistency	same facts, different logic	policy adherence
Actionability	a bare "no"	pivotal conditions

Consistency and self-faithfulness

Consistency

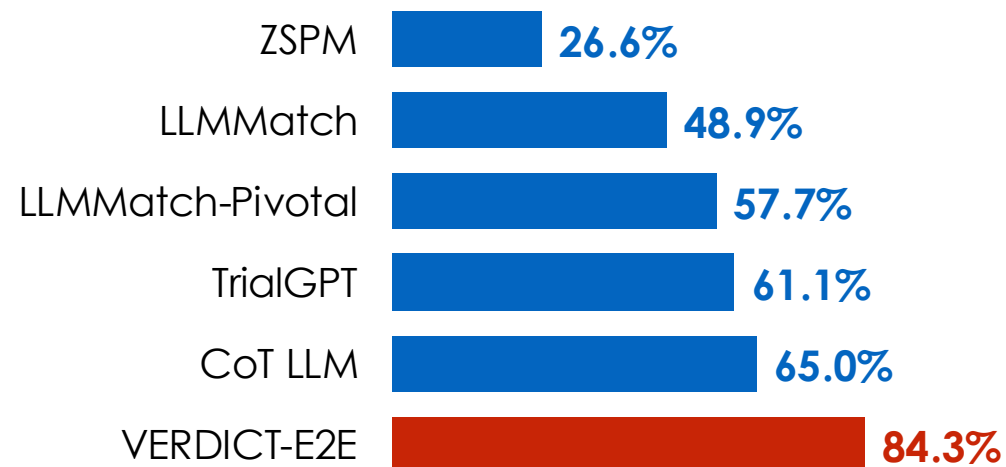
71–81%

Policy adherence of LLM matchers,
even with the policy applied per condition

*Putting the rules in the prompt lowered
LLMMatch agreement (67.5 → 64.8).*

Self-faithfulness

Pivotal flip rate (ineligible → eligible)

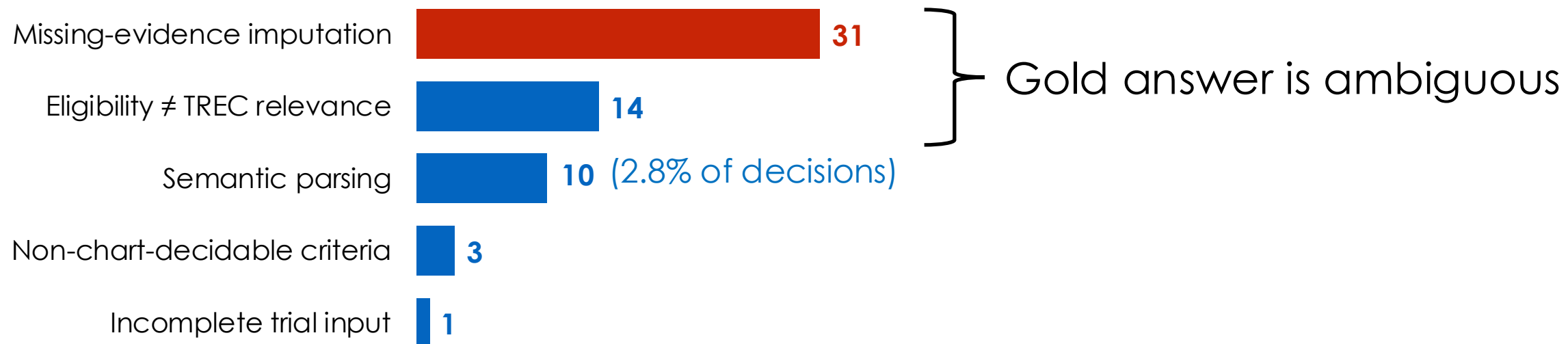


VERDICT-E22: end-to-end system including the parser

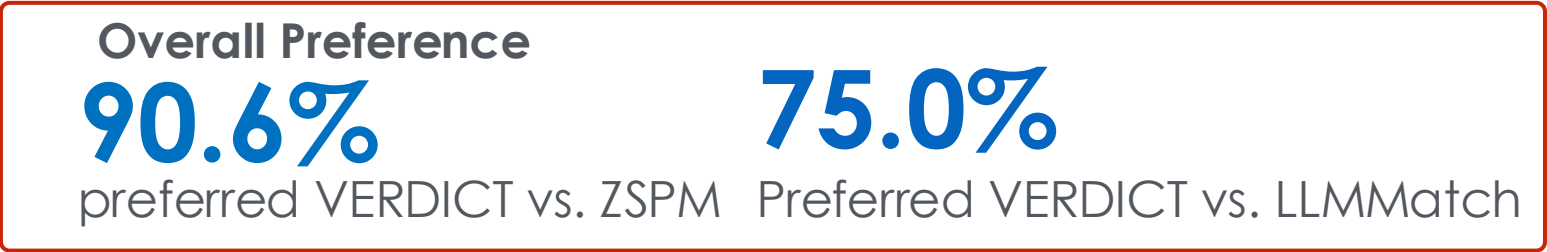
QUIZ: Why is VERDICT-E2E not 100%

Error analysis

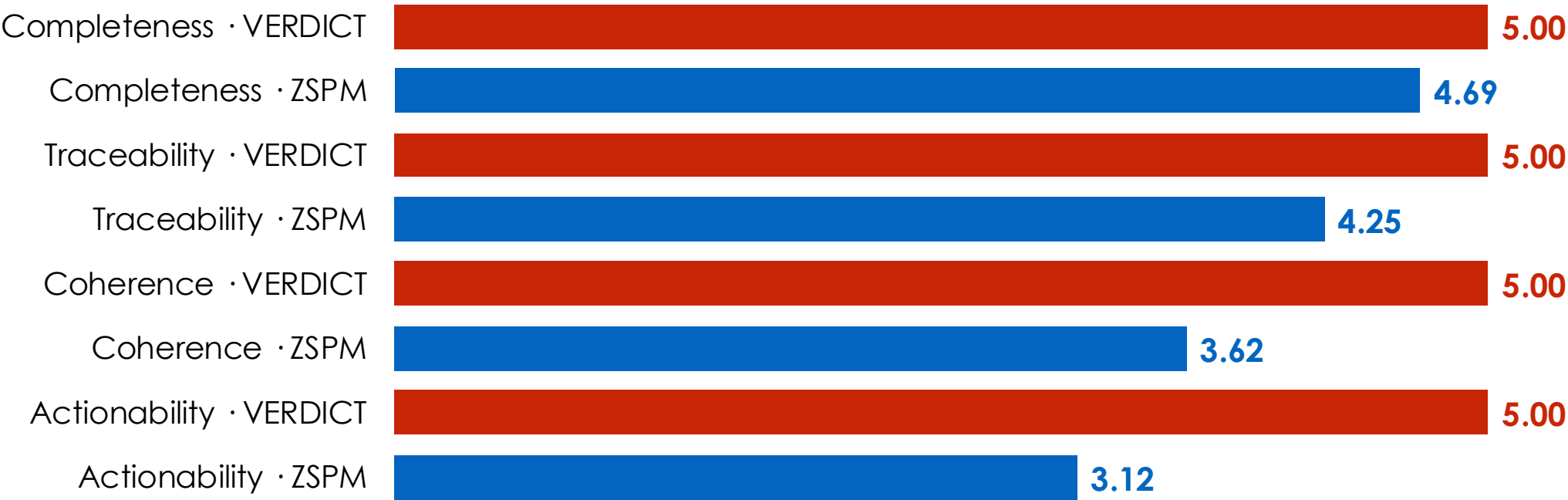
59 TREC disagreements



Clinician ratings



Mean accountability rating: VERDICT (top) vs. ZSPM (bottom)



Completeness: does it address the criteria that matter?

Traceability: does every claim trace to the record or to an explicit imputation?

Coherence: are the claims consistent with each other?

Actionability: does it say what is unresolved and what would change the decision?

Conclusions

Accountability for High-Stakes Decision Making

Grounding: Why?

Is the stated reason the real reason?

Faithfulness: What was assumed?

Missing facts, filled in silently?

Assumptions: a check list to ensure they are true

Consistent: Following policies?

Same facts, same answer?

Actionability: What would change it?

Pivotal condition –

A path to a different outcome

Complexity of real-life decision making: Missing data, policies

Let AI read. Let logic decide