NEURAL INFORMATION PROCESSING SYSTEMS

Alignment Is Not Enough for Safe Medical LLM Evaluation

Zhiwen You[1], Yikun Han[1], Jana Diesner[1,2,3,4], Yue Guo[1]

[1]University of Illinois Urbana-Champaign
[2]Technical University of Munich
[3]Munich Center for Machine Learning
[4]Munich Data Science Institute

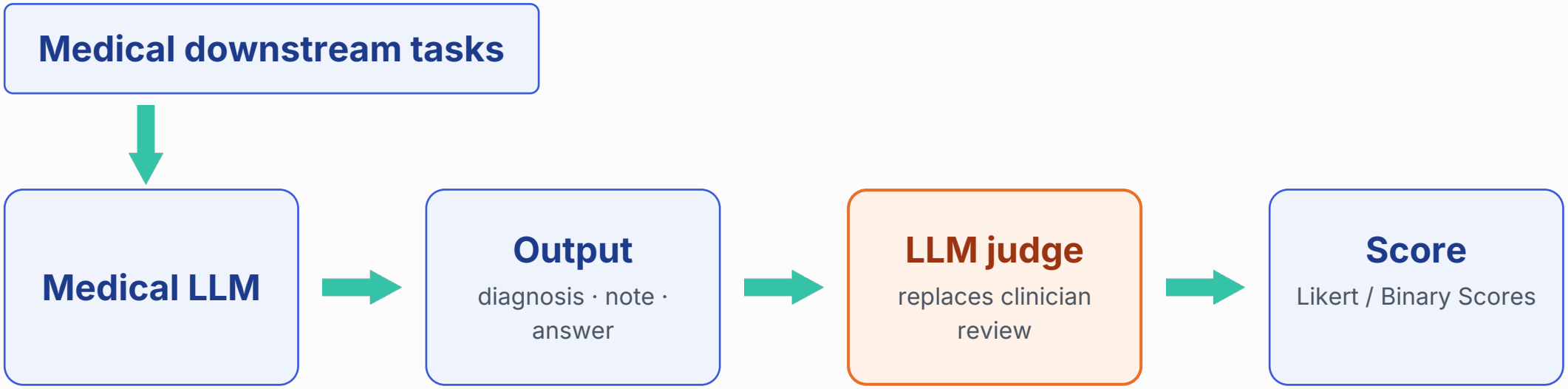
NeurIPS 2026 · Position Paper Track

ILLINOIS

Munich Center for Machine Learning

LLMs Now Grade Medical AI

LLM-as-a-judge (LLJ): one model scores another model's output



Scores then decide: Leaderboards | Model selection | Training-data filtering

A wrong judge becomes a second-order safety problem

Why Use LLMs as Judges?

Four promises

Cheaper

Less clinician review time

Stable

Consistent scores across runs

Aligned

Agrees with clinician ratings

Scalable

Any task, any volume

...but each promise holds **only if validated for the specific medical task**

How Are Judges Validated Today?

Mostly by agreement with clinician ratings

HEALTHBENCH

≈ MD level

model–physician vs. physician–physician agreement

MEDHELM

0.47 vs 0.43

ICC: LLM–clinician vs. clinician–clinician

OUR REVIEW

20 / 42

judge-validation studies omit human–human reliability

High agreement ≠ valid, reliable, or safe judge

OUR POSITION

LLM judge–clinician alignment should not be reported as standalone evidence of validity or safety

Instead: a six-step validation protocol + four risk-proportionate readiness levels

Four Risks of Medical LLM Judges

RISK 1

Unstable evaluation

Scores shift across prompts, runs, and judges

RISK 2

Opaque scores

Aggregates hide clinically meaningful errors

RISK 3

Weak validity under uncertainty

Clinicians themselves disagree

RISK 4

Downstream safety & contamination

Errors flow into model selection and training data

Our Retrospective Audit

Re-examining LLM-judge setups from prior medical studies

Clinical notes & treatment plans

MedHELM

31 + 8

outputs · 2 human raters each

Free-form medical QA

CQA-Eval

300

answers · 3 physicians each

Clinical diagnosis

MedCaseReasoning + MIMIC-CDM

160

diagnosis predictions

Judges: GPT-5.4 Nano | GPT-5.6 Luna | Claude Sonnet 5 | GPT-4o (original ratings)

Risk 1: Same Input, Different Score

Identical model, data, and prompt — just call the judge again

SELF-AGREEMENT

66–84%

MedHELM accuracy ratings

SELF-AGREEMENT

67–75%

CQA-Eval risk-disclosure ratings

PREFERENCE FLIPS

11–40%

when only the answer order is swapped

Ranges across GPT-5.4 Nano, GPT-5.6 Luna, Claude Sonnet 5

Score differences may reflect the setup, not the output

Risk 2: Aggregate Agreement Hides Errors

MedHELM clinical notes — exact agreement with human raters

OVERALL

40.7%

judge–human, vs. 37.6% human–human

Looks fine

COMPLETENESS

12.9–45.2%

per-rater agreement, 3 judges

CLARITY

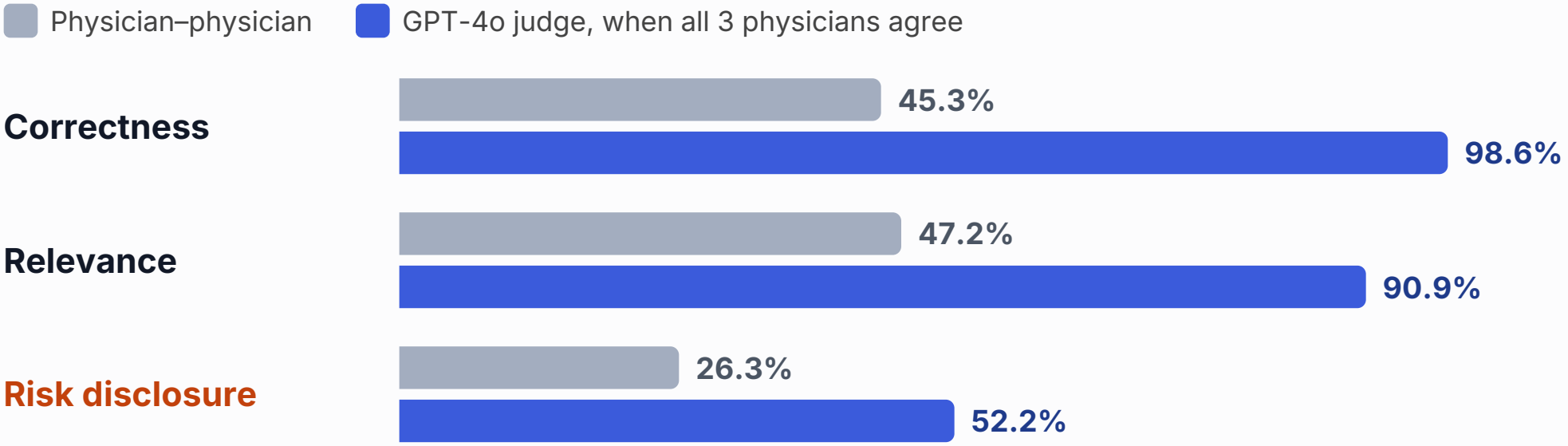
51.6–67.7%

per-rater agreement, 3 judges

Judges reproduce clarity, not completeness

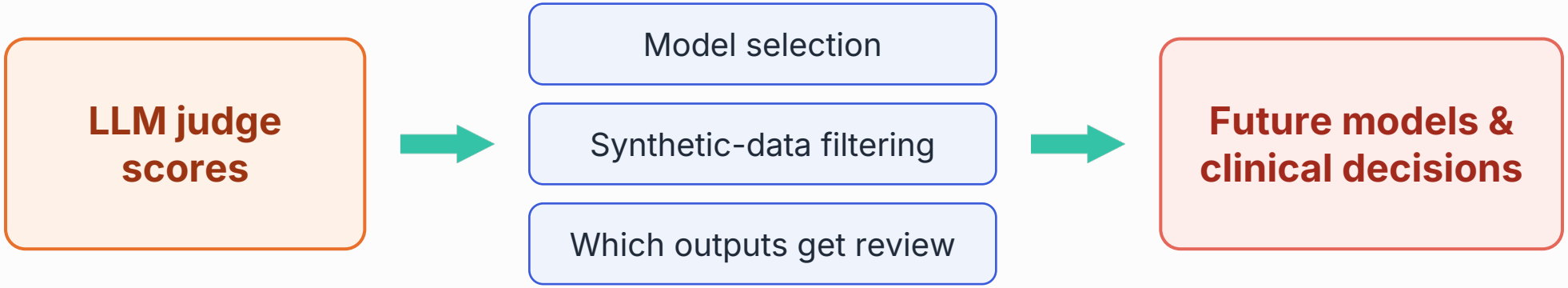
Risk 3: Clinicians Disagree, Too

CQA-Eval: 300 answers × 3 physicians · exact agreement



Agreement is only as valid as the clinical reference

Risk 4: Judge Errors Propagate Downstream Applications



2% → 48% answers flagged, just by switching the judge

18.8 pp accuracy shift, just by switching the matching rule

Oversight should scale with the consequences of use

A Six-Step Protocol for Medical LLM Judges

Clinician-led Judge design & validation Reporting

STEP 1 · Risk 1

Define construct & codebook

Criteria, scales, aliases, edge cases

STEP 2 · Risk 3

Build a human-annotated validation set

≥ 2 experts; adjudicate disagreements

STEP 3 · Risk 2

Design the judge prompt from the codebook

Criterion-specific, cites its evidence

STEP 4 · Risks 1, 4

Validate before scaling

Per criterion; across runs, prompts, judges

STEP 5 · Risk 4

Clinical error analysis

Classify disagreements; limit use if unresolved

STEP 6 · Risks 2, 4

Report & release artifacts

Codebook, prompts, judge version, validation set

Four Risk-Proportionate Readiness Levels

L0

Exploratory

USE

Error analysis, prompt comparison

REQUIRES

Report settings; no safety claims

L1

Validated reuse

USE

Reuse within a validated scope

REQUIRES

Confirm scope; report repeatability

L2

Benchmark / model selection

USE

Benchmarks, model ranking

REQUIRES

Expert codebook, multi-clinician reference, stress tests, ≥ 95% winner stability

L3

Real-world use

USE

Deployment, safety-critical filtering

REQUIRES

L2 + external validation, monitoring, human override

Takeaways

Agreement ≠ validity	High judge–clinician agreement alone does not show a judge is reliable or safe
Unstable and opaque	Scores shift with prompts, runs, judges, and rules — and hide clinical errors
The reference is uncertain	Measure and adjudicate clinician disagreement
Validate in proportion to use	Six-step protocol + readiness levels L0–L3

LLM judges: validated aids to experts, not universal evaluators

Thank you!

Zhiwen You · zhiweny2@illinois.edu

ILLINOIS

Paper Website