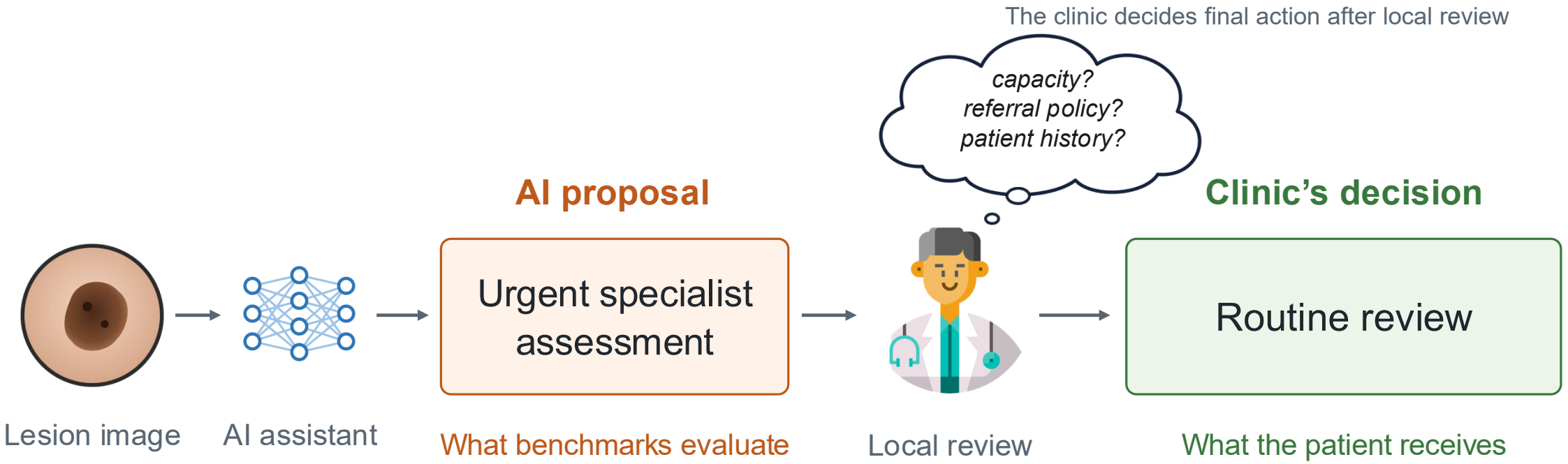
Authority-Preserving Evaluation of Medical Vision-Language Assistants

Flint Xiaofeng Fan[1,3], Cheston Tan[3], Yew-Soon Ong[2,3], Roger Wattenhofer[1]

[1]ETH Zurich, Switzerland
[2]Nanyang Technological University, Singapore
[3]Centre for Frontier AI Research, Agency for Science, Technology and Research, Singapore

ELAMI @ MICCAI 2026
Strasbourg
27/09/2026

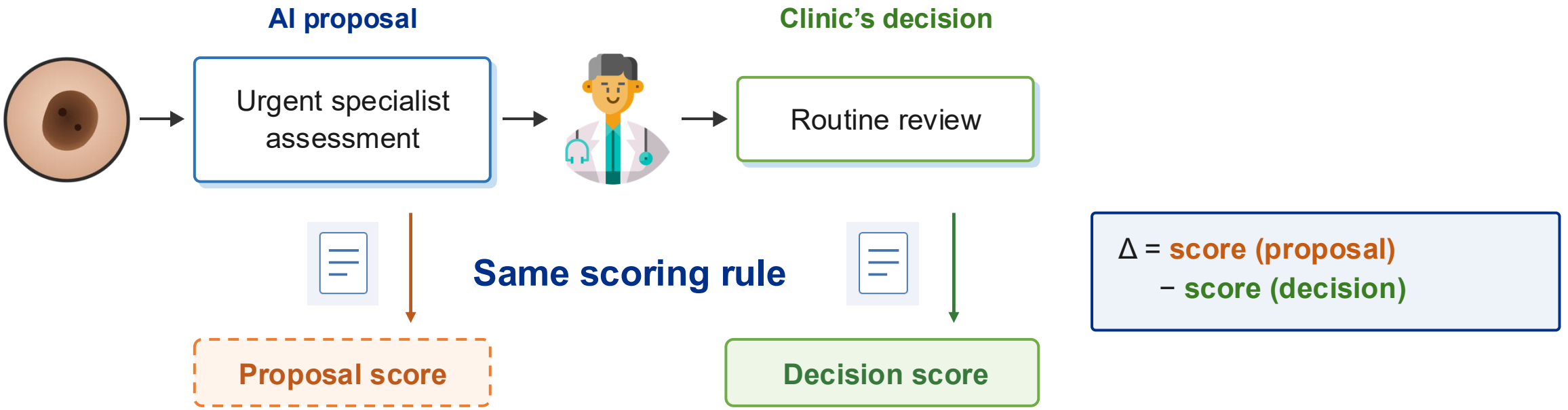
When the Clinic Changes the AI's Proposal



Illustrative case · the AI proposes one urgency levels

Q. When the clinic can overrule the AI, what should we evaluate, and what must we record to do it?

AuthEval: Score Both Under the Same Rule



Authority gap = average Δ

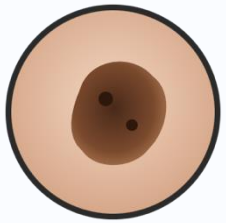
~ 0: AI-only evaluation holds;

> 0: overstates the care => check the workflow; < 0: understates the care => model needs to catch up

Same AI, Two Clinics

Real photos, known diagnoses

ISIC 2019



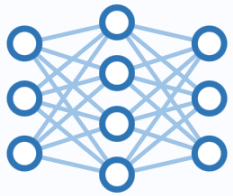
3,967 test images

Diagnosis used only to score

Real AI proposals

MedGemma

1.5 · 4B · frozen



Ranks 4 urgency levels

Top choice = proposal

Two simulated clinics

Capacity ceiling

Few slots:
pushes down

Safety floor

Risk cues:
pushes up

Neither rule sees the diagnosis

Same AI proposals · each clinic takes the AI's top allowed option

How does the gap quantifies the local context?

The gap quantifies the local context

Average scores (0 = best)

Capacity ceiling

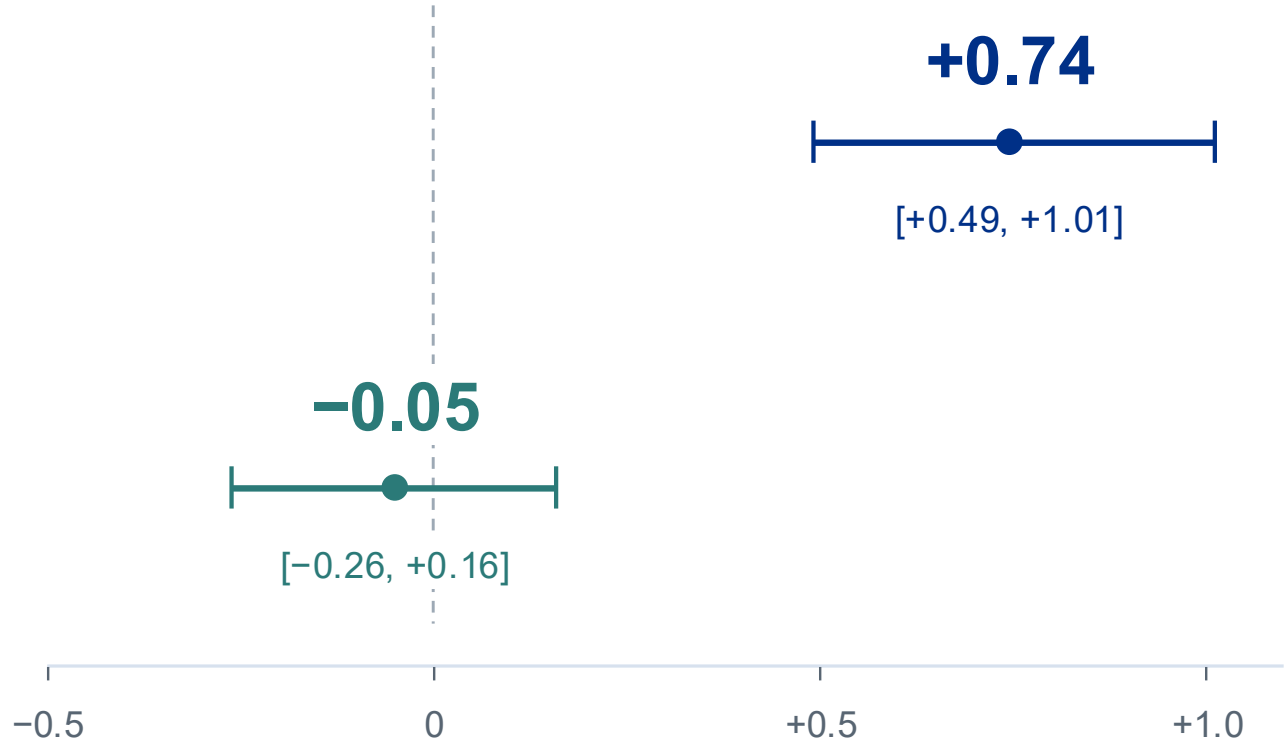
AI proposal	−2.94
Selected action	−3.68

Safety floor

AI proposal	−2.94
Selected action	−2.89

Each clinic's authority gap

Average(proposal score − selected-action score)



The sign of the gap suggest where to act.

Image-weighted · Simulator units · 95% whole-cluster bootstrap intervals

Evaluate Medical AI in Clinical Context

Match the evaluation to the question

The model

AI proposal

→ Proposal score

Selected decisions

Selected action

→ Selected-action score

What did local review change?

Compare both scores under the same local rule

Next

Validate with real clinical workflow records

Clinical partners welcome

When a proposal is declined, retain its local score.

Cannot obtain that score? Report the gap as unavailable; don’t assume it is zero.

MICCAI2026

THANK YOU

Flint Xiaofeng Fan · fxf@u.nus.edu · https://flint-xf-fan.github.io/

Centre for Frontier AI Research
A*STAR CFAR

ETH zürich

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE