OSCE AI

We put our AI marker through an exam of its own

Four doctors marked ten OSCE stations. Then our AI examiner marked the same ten, blind, on production settings, first attempt. Here's what happened, including the bits that surprised us.

QUESMED · PILOT STUDY · 2026

OSCE marking: how much do examiners agree?

Here is something every medical student suspects and nobody says out loud: part of your OSCE mark depends on which examiner you get.

That's not a criticism of examiners. It's just what happens when you ask a person to watch eight minutes of consultation, tick thirty-odd criteria, and then compress the whole thing into one judgement: fail, borderline, pass, good or excellent. Two experienced doctors can watch the same station and, in good faith, land a band apart. One weighs the missed safety-net most heavily, the other gives more credit for rapport under pressure; neither is wrong - they're just not the same person.

We wanted to know exactly how big that human difference is, because we've built an AI examiner into our OSCE practice platform, and "is the AI as good as a human?" is not answerable until you can say how good humans are **at agreeing with each other**.

So we ran a pilot: examine the examiners first, then hold the AI to that standard.

What we did

We took ten real practice-station transcripts from our voice-based OSCE stations: five station types, two performances each, spanning histories, breaking bad news, ECG interpretation and X-ray viva. Names and identifiable information were anonymised.

Four doctors (trained on how to examine the transcripts) independently marked all ten, blind to each other, using the same mark schemes: every criterion ticked yes or no, then a global score from 1 (fail) to 5 (excellent), plus free-text examiner notes. Fully crossed: forty human marksheets.

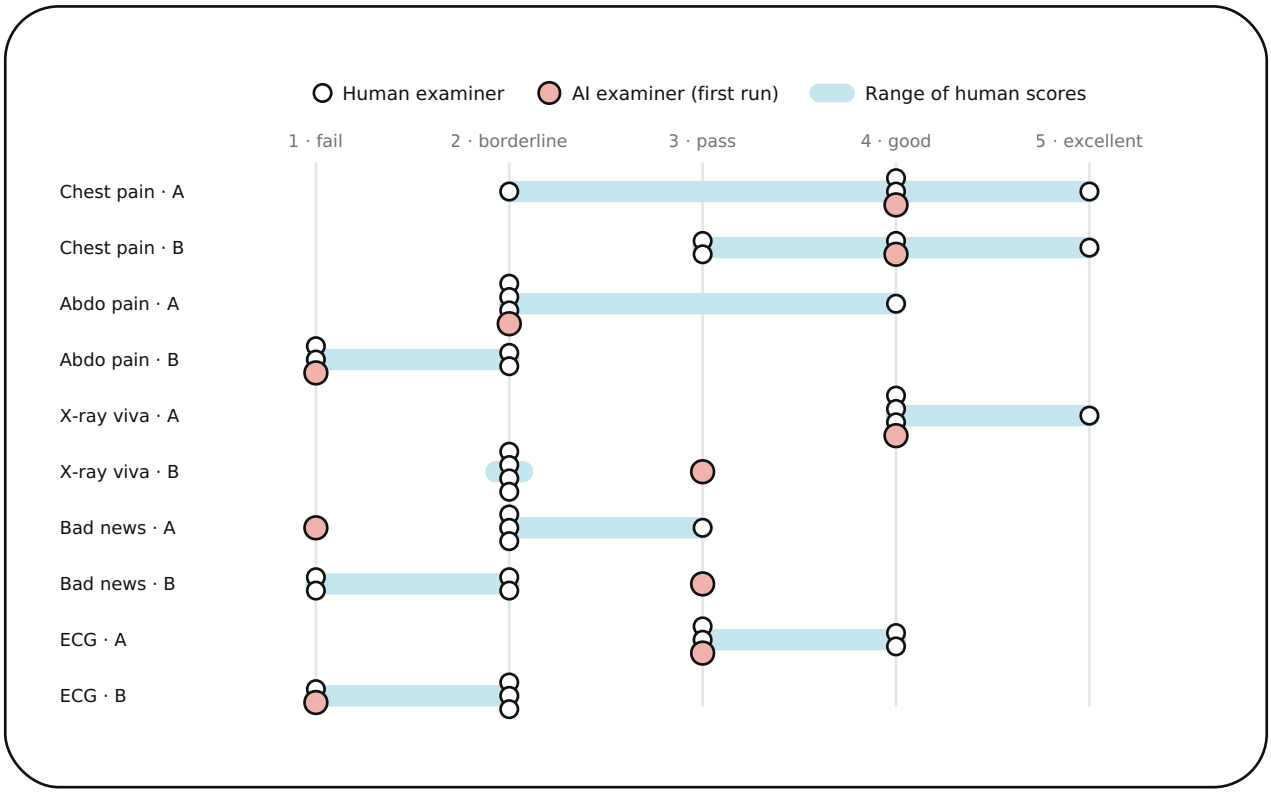
Then our AI examiner marked the same ten transcripts. No tuning, no cherry-picking, no "best of five runs": the exact configuration that marks stations on the platform today, first run counted as the headline result. We also re-ran it twice more per station to measure its consistency. More on that later.

First finding: the human examiners disagreed a lot.

Across the ten stations, all four human examiners gave the same global verdict on **just one station**. On a typical station, two examiners picked at random disagreed by about three-quarters of a band; 43% of the time they agreed exactly, and 87% of the time they were within one band of each other. One station was scored **2 (borderline) by one examiner and 5 (excellent) by another** - they agreed on almost every individual tick-box, but they disagreed about philosophy. One examiner capped the mark over a patient-safety judgement the other weighed differently.

This is the benchmark the AI is measured against, not an idealised perfect examiner, but the actual spread of agreement shown above.

So, how did the AI do?



Every global score on every station: four human examiners (open circles), the AI's first run (pink), and the span of human opinion (blue bar). On 7 of 10 stations the AI landed inside or touching the human range. Station letters A/B are the two different student performances of each station type.

The AI examiner performed well: on seven of the ten stations, the AI's verdict sat inside or touching the range of the four human opinions - and on the other three it missed that range by a single band. Averaged over all forty AI-vs-examiner comparisons, the AI was **0.07 bands harsher** than the humans: statistically, bang on zero. It agreed with a given human examiner within one band 88% of the time. The examiners managed 87% with each other.

And on the individual tick-boxes (the "did they ask about radiation of the pain?" level), the AI agreed with the examiners' majority verdict **slightly more often than the examiners agreed with each other**. It wasn't a soft touch either: the AI awarded 60.8% of available marks, sitting within the four humans' own range (57–62%).

87%

HUMAN VS HUMAN

of examiner pairs within one band of each other

88%

AI VS HUMAN

of AI–examiner pairs within one band of each other

0.77

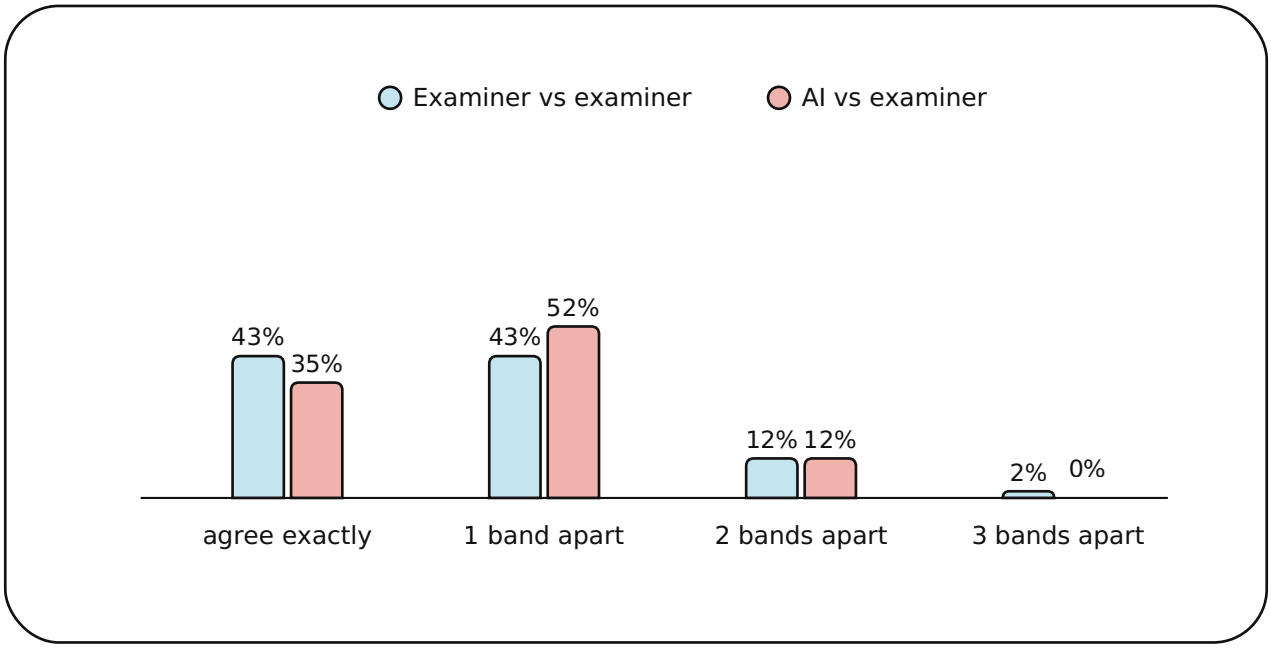
TICK-BOX AGREEMENT

AI vs examiner majority (Gwet's AC1); humans score 0.76 with each other

~14s

PER STATION

full mark scheme plus written examiner feedback, on every attempt



How far apart two verdicts on the same station land. The AI–examiner distribution looks like the examiner–examiner one, and the only 3-band gap in the whole study was between two humans.

A further detail of interest: the largest disagreement in the entire study, three whole bands, was between **two human examiners**. The AI never got further than two from anybody.

How it marks (the short version)

Our AI examiner works the way a good human one does, in two distinct passes:

1. The mark scheme pass - it reads the full transcript of your consultation against the station's mark scheme (the same criteria you can open on Quesmed) and decides, criterion by criterion, whether you did it, citing what you actually said.

2. The examiner pass - a second stage then does what the examiner in the chair does after the buzzer: it steps back, takes the completed mark scheme as evidence, and reasons about the whole performance. Crucially, that includes key marks: the criteria that carry the point of the station. You can tick thirty boxes in an abdominal-pain history, but if you never considered that the patient (female, of child-bearing age) might be pregnant, you've missed the ectopic. Recommend a penicillin-based antibiotic without asking about allergies, or take a whole low-mood history without ever asking about thoughts of self-harm, and a human examiner won't wave you through on rapport - neither will our AI examiner.

That second pass writes your feedback too: what carried the station, what to fix first, phrased the way an examiner would put it at a debrief.

The part that surprised us

We expected the interesting comparison to be the scores but it turned out to be the **reasons**.

Take the X-ray viva station that every human examiner scored 2, a borderline. Here's what they wrote, next to what the AI wrote. Neither could see the other:

HUMAN EXAMINERS

"The initial AXR description is not very detailed and stops abruptly once the dilated large bowel is identified."

"Did not use technical terms when describing the adequacy of the x-ray. No systematic review."

"Incorrect diagnosis also lowers the global rating."

AI EXAMINER

"You did not clearly commit to ulcerative colitis with toxic megacolon as the most likely diagnosis."

"Use precise radiology terms like AP film and lead-pipe colon. Comment systematically on perforation, small bowel, liver, lungs, and bones."

Same evidence, omissions and priorities, arrived at independently. Across the pilot, this held again and again: where the AI's mark differed from a given examiner's, it was almost never because it saw something different. It was because it weighed the same observations differently, which is precisely where the human examiners differed from each other too. The disagreement lives in the same place.

How consistent is the AI marker?

Ask the same examiner to re-mark a station and you won't always get the same answer. The same is true of an AI. So we re-ran every station three times and measured the wobble instead of hiding it.

The tick-box level barely moved: 92% of individual criteria got the identical mark all three times. The global verdict was steadier than a pair of human examiners, but not identical across runs, and we could see exactly why. On one breaking-bad-news station, the AI credited the warning shot in two runs and not in the third; because a warning shot is a key mark, the verdict legitimately swung with it. That's the system reasoning the way we designed it to: the input flickered, not the judgement. We know precisely which lever to pull, and tightening it is already on our roadmap.

It's still worth reinforcing that the run-to-run spread of the AI was **smaller** than the spread between two human examiners marking the same station.

What this means when you're practising

The reason we care about all this isn't a badge. It's what a trustworthy examiner in your pocket actually changes about OSCE prep:

You can rehearse the real thing. Talking to a patient, out loud, against the clock, then getting marked the way you'll actually be marked: criterion by criterion, plus the global judgement, plus feedback that tells you what to fix first.

Every attempt gets a proper debrief. The feedback isn't a canned paragraph; it's written from what you specifically said and didn't say, the way a good examiner would put it. Practice partners are wonderful, but they're not always free at a moment's notice.

The bar we set ourselves was never "is the AI perfect?" Nobody who has ever sat in an OSCE circuit believes human marking is perfect. The bar was: does the AI disagree with an examiner about as much as examiners disagree with each other, for the same reasons?

On this pilot, on first attempt, on the settings running today: yes.

QUESMED

Learn medicine the smart way. · Pilot study conducted 2026 with four practising clinicians marking anonymised voice-station transcripts. Figures show real, unadjusted data from the study.