

AI grading in a real OSCE

A prospective study at UT Southwestern

An objective structured clinical examination (OSCE), using notes, audio and video with targeted human review. Fall 2025 deployment.

222

second-year students

Each completed five OSCE stations.

72,907

retained item scores

Across notes, audio and video.

92.3%

fewer human scoring passes

Versus a modeled single-pass manual comparator.

HOW THE WORKFLOW WORKED

01 Multimodal scoring

AI scored rubric items using each item's designated source: notes, audio or video.

02 Targeted review

Low-scoring learners were routed for independent educator rescoring.

03 Physician adjudication

Selected AI-human disagreements received physician review.

What the workload result measures

5,616 human scoring passes in the deployed workflow, compared with 72,907 in a modeled one-pass manual workflow. Repeated scoring counts as additional passes. This is not a measured time or cost saving.

WHAT STILL NEEDS TESTING

This was a single-center study. Human comparisons focused on the low-scoring review subset; they do not establish cohort-wide accuracy or performance across institutions. UT-REAL multi-site validation is separate work.

Ngo, Shakur et al. | npj Digital Medicine | Published October 7, 2026

doi:10.1038/s41746-026-03352-4 | Article in Press at preparation

Collection: Transforming Medical Education through Artificial Intelligence

[Read the paper](#)

[Explore the study and workflow demo](#)