

Deploying A Multimodal AI Platform for Automated OSCE Grading in Medical Education

Huong-Tra Ngo, PhD,¹ Ameer H. Shakur, PhD,¹ Shinyoung Kang, BS,¹ David Hein, MS¹, Philip Jarrett, MD MBA², Michael J. Holcomb, MS¹, Thomas O. Dalton, MD⁴, Krystle K. Campbell, DHA,³ Daniel J. Scott, MD,^{4,5} Andrew R. Jamieson, PhD,¹

¹Lyda Hill Department of Bioinformatics, ²Department of Emergency Medicine, ³Department of Internal Medicine, ⁴Simulation Center, ⁵Department of Surgery UT Southwestern Medical Center

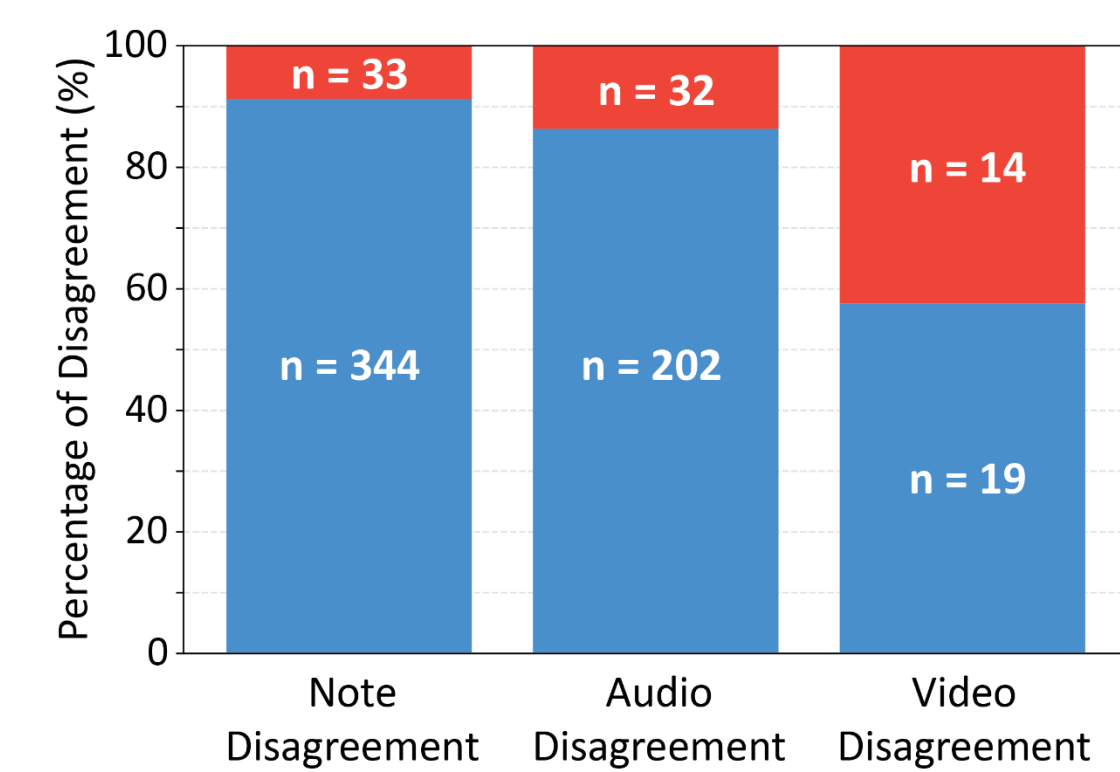
Introduction: Objective Structured Clinical Examination (OSCE)

- What?** A standardized, simulation-based assessment in medical education to evaluate clinical competency (clinical reasoning, communication skills, patient-centered care) through structured encounters with standardized patients.
- Contain?** History-taking, physical examination, verbal signposting, patient interaction, post-encounter documentation
- Artifacts?** Video recordings, notes, faculty's grading rubrics
- Challenges?** Labor-intensive, costly to grade, prone to inner/inter-rater variability, difficult to scale, delayed feedback (months to grade), simplified rubrics that may reduce educational rigor.
- Innovation?** We present the first-ever deployment of multimodal AI platform for automated OSCE grading across written notes, communication skills, and physical examination.

Methods

- Platform: **SimRubrics** for rubric design + **MAPLES** for zero-shot AI grading
- UT Southwestern Simulation Center, Fall 2023 – Fall 2025
- Bottom 5-10% flagged for standardized patient evaluator review & faculty adjudication.
- Gemini 2.5 Pro

Results: Human Review & Faculty Adjudication

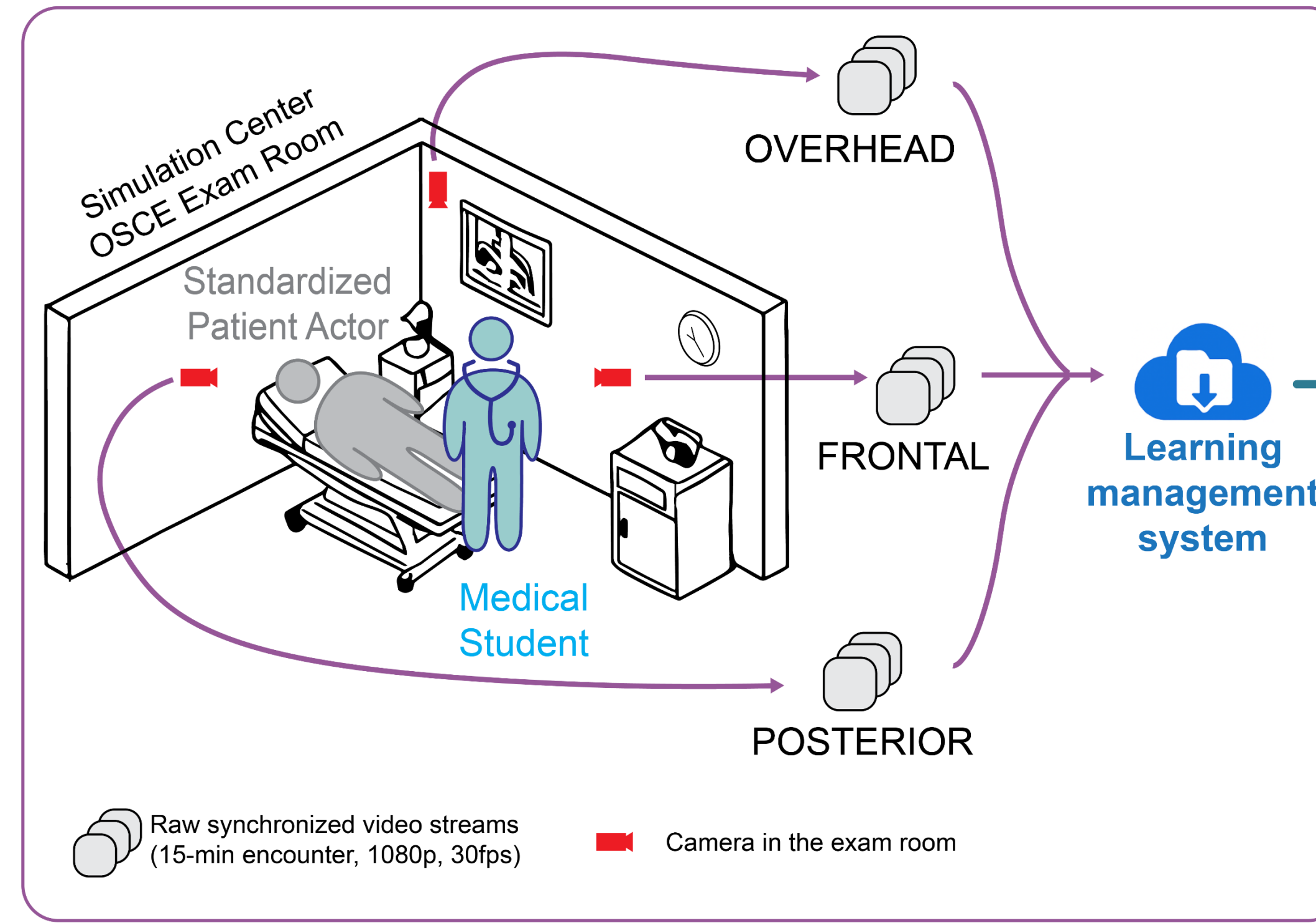


Direction of Disagreement

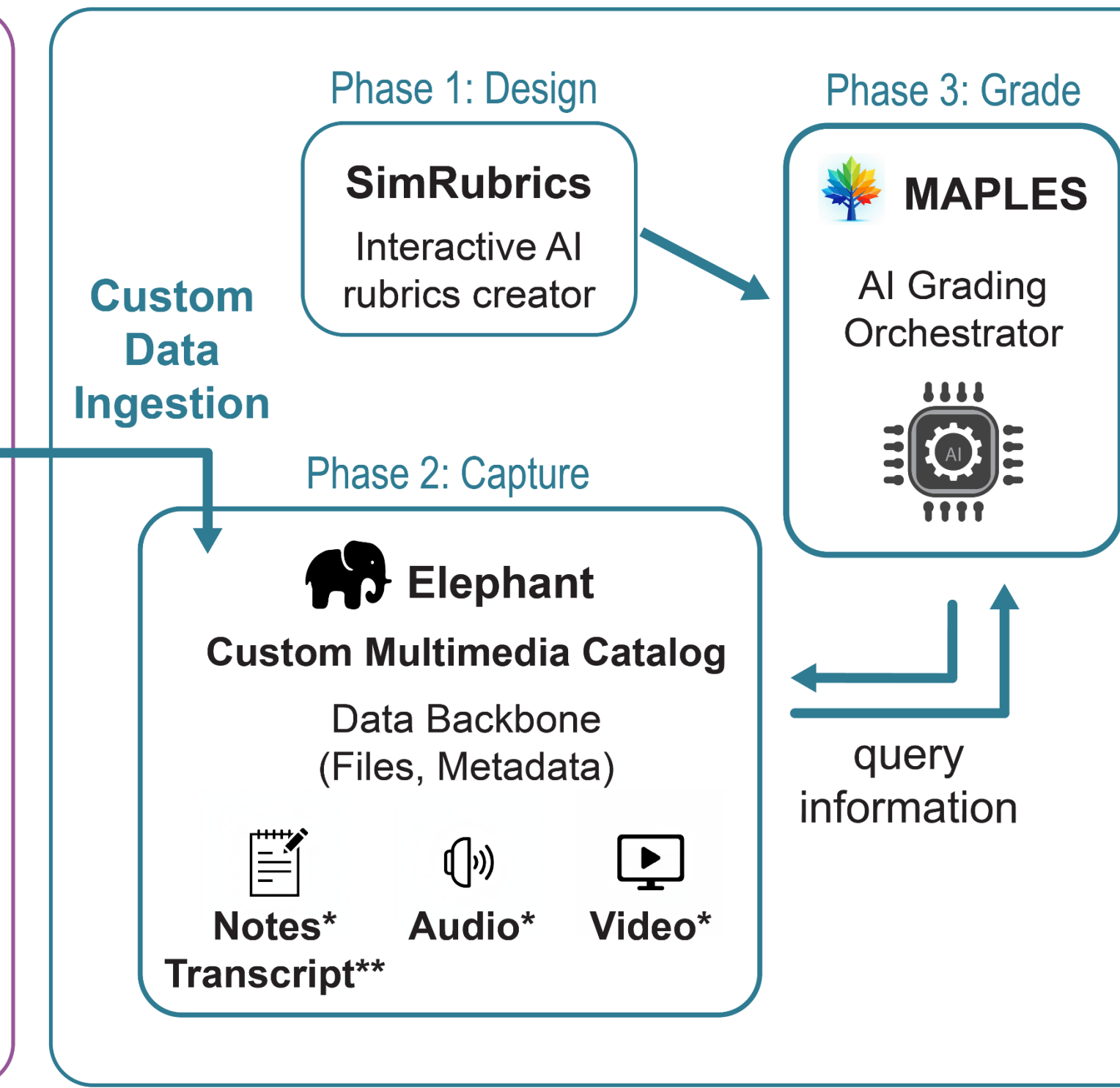
■ AI Grades Higher than SP
■ AI Grades Lower than SP

- Faculties:**
- Agree with AI: 76.4%
 - Agree with SP: 22.2%
 - Neither: 1.4%

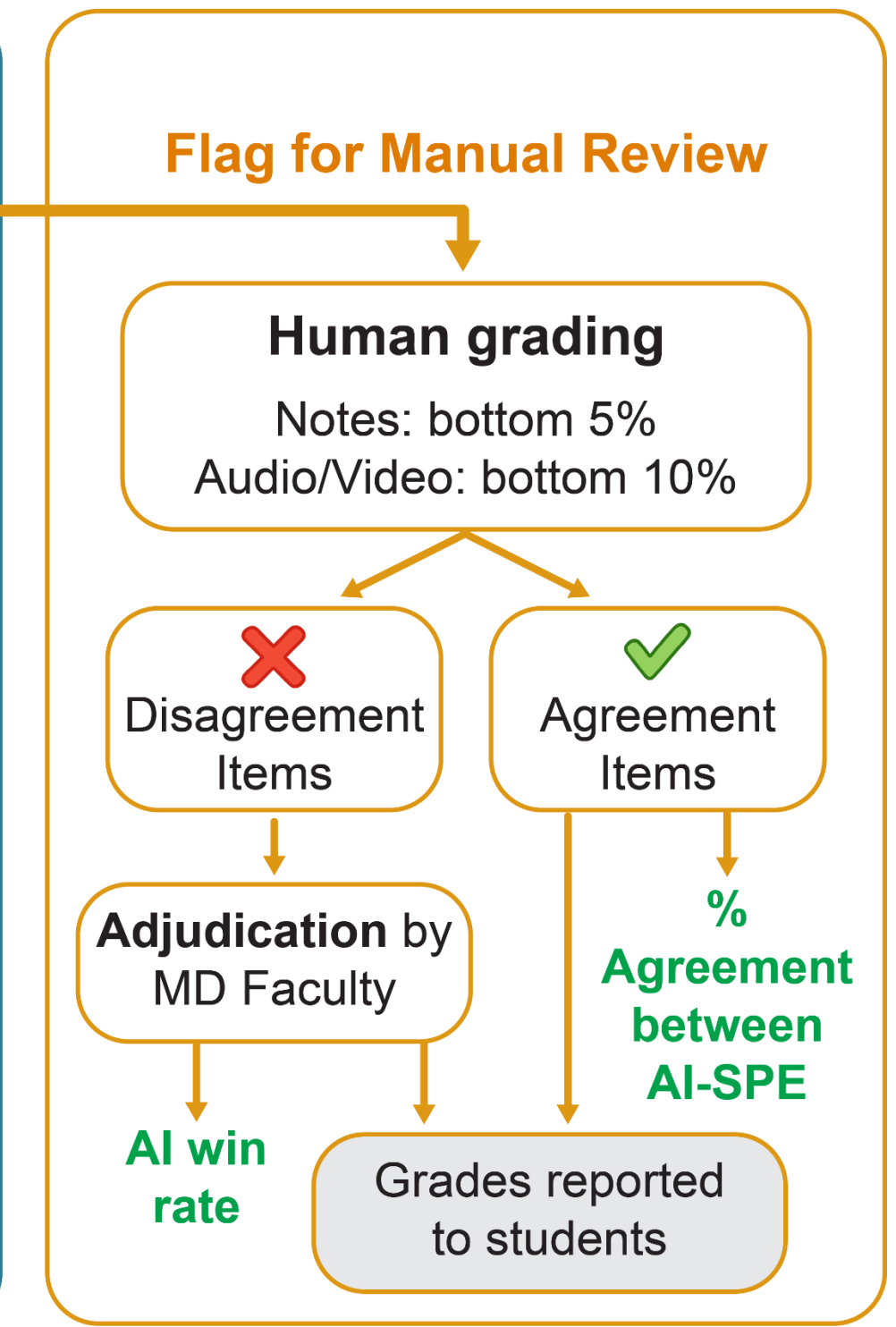
1. OSCE Encounter Captured



2. Platform Architecture



3. Human Review



MAPLES Platform Screenshots

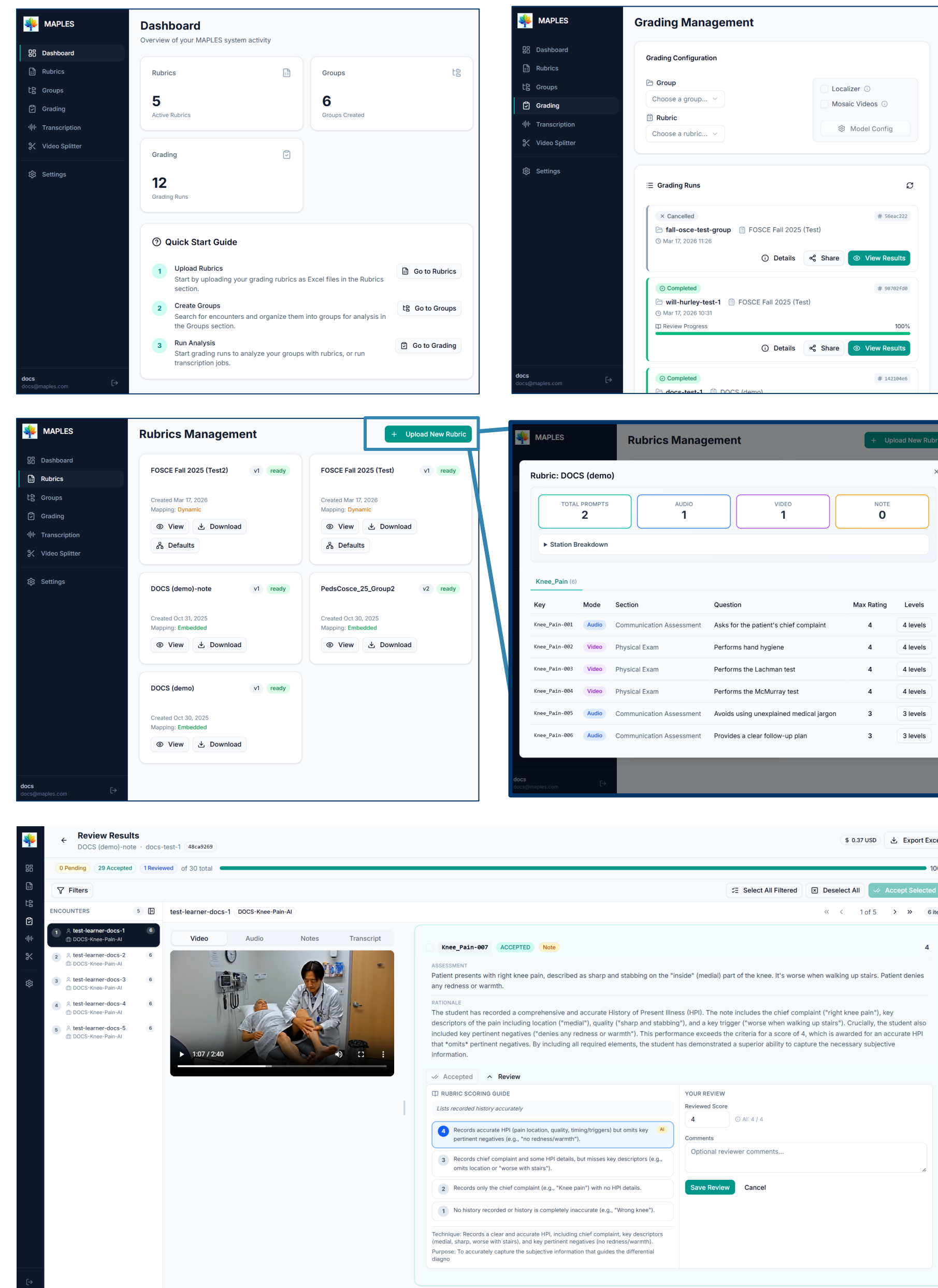


Table 1. Prospective deployment summary

OSCE Event	Learners	Data source	Total gradable items	Human graded items	AI-Human Agreement (%)	Human Effort Reduction (by items graded)
Fall 2023	245	Notes*	24,255	4,059	0.93	91%
Fall 2024	233	Notes*	23,067	1,494	0.96	97%
Spring 2025	227	Notes*	49,832	2,634	0.96	97%
		Transcript**	16,711	2,365	0.96	95%
Fall 2025	221	Notes*	50,502	2,879	0.94	97%
		Audio**	20,550	2,343	0.94	95%
		Video**	4,776	544	0.83	95%

* Textual data used to assess clinical reasoning skills. ** Video, audio, transcript data used to assess PE and communication skills

Discussion

- Lesson learned:** Successful deployment depends on robust data management practices at the simulation center, including the use of modern data management software, rigorous data capture processes, and well-designed, reliable assessment rubrics.
- Reduces turnaround time from months to days.
- Enables more detailed rubrics & more resources for feedback.

Future

- Scale up deployment to 5 more UT sites
- National multi-site collaboration
- Contact us to collaborate

Jamieson Lab



labs.utsouthwestern.edu/jamieson-lab

Project MAPLES



ut-real-ai-project-maples.com

UT Southwestern
Medical Center