Ask a Better Question: How an AI Grader Sharpened Hypothesis-Driven History-Taking in a Medical Student Objective Structured Clinical Examination

Lindsey A Stockton, MD[1] Andrea Cooley, DO[1] Ameer Hamza Shakur, PhD[2] Andrew Jamieson, PhD[2] Judah Gruen, MD[2]

[1] UT Tyler School of Medicine [2] UT Southwestern Medical Center

2026 UT REAL Health AI Symposium

Background

End-of-clinical clerkship OSCEs commonly assess medical students' history-taking, physical exam skills, communication, and clinical reasoning with checklist rubrics scored as complete, partially complete, or incomplete. Broad items leave substantial room for grader interpretation.

Hidden ambiguity: Goes unnoticed with a single consistent grader.
Liability: Becomes a problem with multiple raters or shared rubrics across institutions.
Lack of specificity: Broad items can't distinguish hypothesis-driven history-taking from rote checklist completion.

Approach

This work grew out of a broader project comparing AI grading of OSCE encounters with human scoring, using a purpose-built tool from the Jamieson Bioinformatics Lab at UT Southwestern.

- Ran an Internal Medicine end-of-clerkship OSCE encounter through the tool against our existing history-taking rubric.
- The rubric-clarity finding emerged incidentally and was not the original aim.
- The tool graded several broad items inconsistently, often disagreeing sharply with the historical grade.
- If a model applying the rubric literally couldn't grade consistently, a second human grader would likely face the same ambiguity.

Lessons Learned

- Revised the rubric item by item, replacing broad prompts with explicit, behaviorally anchored language that the AI couldn't interpret multiple ways.
- Unplanned benefit: several new items became diagnosis-specific probes instead of generic symptom prompts.
- Clarifying language for a machine grader pushed the rubric toward rewarding hypothesis-driven history-taking, where students build toward specific differential diagnoses

1 AI grader run
Internal Medicine OSCE encounter scored against the existing history-taking rubric

2 Inconsistency exposed
Broad items graded unevenly, with large disagreements versus the historical grade

3 Rubric rewritten
Broad prompts replaced with explicit, behaviorally anchored language

4 Hypothesis-driven probes
New items target specific diagnoses instead of generic symptom prompts

Table 1. Before-and-After Comparison of Rubric Items (Fatigue Case)

#	Rubric Item	Original Language	Revised Language	Refinement Type
1	**Diet**	"Diet (assessing for contributing risk factors)" — no criteria	"Must ask about diet with specifics related to dietary restrictions or what it consists of. Incomplete if they only ask 'how is your diet' without any follow-up questions."	Added explicit completion criteria
2	**OB/Gyn / Menstrual History**	"OB/Gyn History (LMP, G/P, menstruation pattern)" — bundled, no sub-criteria	"Menstrual cycle pattern (has it changed, how many pads/tampons, how many days). Must ask all three to get full credit."	Split into named sub-items + explicit count requirement
3	**Weight-related symptoms**	"Related Symptoms must ask: weight loss"	"Related Symptoms must ask: weight loss or gain"	Closed a scope gap
4	**Sleep**	(absent)	NEW: "Related Symptoms must ask: sleep patterns"	New hypothesis-linked item
5	**Thyroid screen**	(absent)	NEW: "Related Symptoms must ask: cold/hot intolerance"	New item, targets hypothyroidism
6	**Mood/depression screen**	(absent)	NEW: "Related Symptoms must ask: about their mood"	New item, targets depression
7	**Pregnancy**	(absent)	NEW: "Must ask if there is a possibility of pregnancy"	New hypothesis-linked item
8	**Post-surgical vitamin compliance**	(absent)	NEW: "Must ask if they are taking their Vitamins after Roux-en-Y surgery"	New item, targets B12/folate deficiency
9	**Family History**	(absent as standalone item)	NEW: "Family Medical History"	New standalone item

Highlighted rows: items most substantially refined, either newly added hypothesis-linked "must ask" questions or items rewritten with explicit completion criteria.

Conclusion

Using an AI grader as a diagnostic tool, not only an evaluator, surfaced latent ambiguity in a rubric already in active use and weighted heavily in the clerkship grade. Its inconsistency acted as an early-warning signal for discrepancies that human graders would have hidden.

Rubric language precise enough for an LLM to grade without clarification also encouraged hypothesis-driven history-taking, which may help institutions standardize grading when sharing OSCE rubric templates.

References

1. Holmes A, Singh B, McColl G. Revisiting the hypothesis-driven interview in a contemporary context. Australas Psychiatry. 2011;19(6):488–492.
2. Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M. OSCE checklists do not capture increasing levels of expertise. Acad Med. 1999;74(10):1129–1134.
3. Yudkowsky R, Otaki J, Lowenstein T, Riddle J, Nishigori H, Bordage G. A hypothesis-driven physical examination learning and assessment procedure for medical students: initial validity evidence. Med Educ. 2009;43(8):729–740.
4. Effect of clinically discriminating, evidence-based checklist items on the reliability of scores from an Internal Medicine residency OSCE. Adv Health Sci Educ Theory Pract. 2014;19:369–380.

THE UNIVERSITY OF TEXAS SYSTEM
REAL Health AI
Harnessing AI to Transform Health Care