Can AI Evaluate Assessment? A Study of Large Language Model Meta-Assessment Performance

Authors:

Kyle Green, M.A.
Columbia University

Yu Bao, Ph.D.
James Madison University

Stephanie LeRoy, M.A.
James Madison University

Megan Good, Ph.D.
James Madison University

ABSTRACT

Meta-assessment—the practice of evaluating the quality of assessment reports—summarizes institutional measurement efforts and, when used well, offers formative feedback to improve future processes. This study examines whether artificial intelligence (AI), specifically the large language model-based AI systems ChatGPT-5 and Microsoft Copilot Pro, can assist in this process. Using a binary checklist and a multi-criteria rubric, we compared AI-generated ratings to those of a human expert across three versions of a fictional assessment report (strong, moderate, weak). Each condition was replicated five times to assess response variability. AI aligned with human ratings in 59% of individual cases across all report versions, evaluation formats, assessment elements, AI platforms, and replications, but performance significantly varied depending on these factors, with stronger agreement when assessing high quality reports and when using the checklist format. Both platforms tended to struggle with evaluating measurement. Analysis of AI-generated rationale revealed inconsistencies and a tendency to overlook methodological flaws in weaker reports. While AI is not a replacement for human expertise, it may serve as a supplemental tool for reviewing high-quality reports or specific assessment elements. We discuss practical implications and future directions for AI-supported meta-assessment.

Correspondence E-mail: kdg2140@tc.columbia.edu

Keywords: Large language models, Meta-assessment, Higher education assessment, Artificial intelligence, Assessment quality

Meta-assessment helps institutions evaluate and improve their own assessment practices, but it is often time-consuming and subjective. This study explores whether AI can help. Specifically, we examine whether systems based on large language models (LLMs) can generate meta-assessment ratings that align with those of a human expert. While AI has been used in various educational settings, from predicting student success to automating grading, its potential role in evaluating assessment reports appears to be largely unexplored. Given the growing demands on institutional effectiveness professionals, understanding whether AI can meaningfully assist with meta-assessment could have important implications for the efficiency and consistency of assessment practices.

Meta-Assessment

Meta-assessment is the practice of evaluating the quality of assessment (Ory, 1992). It typically involves using a rubric or checklist to define key elements of quality, such as clearly stated learning outcomes and alignment between outcomes and measures, and providing constructive feedback to faculty who develop assessment reports.

Fulcher and Good (2013) identified several practical benefits of meta-assessment, two of which are highlighted here. First, meta-assessment enables practitioners to quantify assessment quality across the institution, providing a comprehensive snapshot for accreditors. Assessment professionals can calculate average scores based on key elements to evaluate institutional progress and compliance. One accreditor, the Southern Association of Colleges and Schools (SACSCOC) (2024), expects evidence of (1) well-defined student learning outcomes, (2) direct measures aligned with those outcomes, and (3) use of results for improvement. These elements can be systematically aggregated across assessment reports to produce institution-level insights.

Second, meta-assessment plays a critical role in fostering improvement by offering feedback on the effectiveness of assessment practices (Fulcher & Good, 2013; Rodgers et al., 2013). Assessment reports, as well as the practices they reflect, can vary widely in quality. Fulcher and colleagues (2024), for example, describe several "report archetypes" (p. 16-17), which range from "lazy reports" to those showing "focused improvement." Constructive feedback helps authors recognize and address shortcomings, which can move programs toward more intentional and effective practice.

Though valuable, meta-assessment is resource intensive. If assessment reports are submitted annually, they must be evaluated—typically by assessment professionals, an assessment committee, or faculty. When raters outside the assessment office are involved, training is needed, which can take up to two days (as was the practice at our institution, James Madison University). In addition to the time commitment, some institutions provide financial compensation for raters. Our institution's meta-assessment practice provides an illustration of how this process can evolve over time in light of practical demands.

Meta-Assessment at James Madison University

Meta-assessment at JMU began in 2010 with the development of a rubric featuring 14 criteria for defining assessment quality (Fulcher et al., 2015). Each year, approximately 120 assessment reports were collected and evaluated. Teams of trained raters comprised of paid faculty members and graduate assistants independently assessed 12–15 reports

over five days, adjudicating scores and compiling qualitative comments daily. Following the rating period, extensive quality control procedures ensured accuracy before individualized feedback reports were generated at the degree, department, and college levels. In total, we estimate that assessment staff dedicated at least 1,000 hours annually to this process, including report collection and organization, rating, quality control, and feedback dissemination.

Over the first six years of this process, we observed notable improvement in assessment report quality across campus. However, this growth eventually plateaued as the institution reached a level of assessment maturity, with most degree programs implementing robust assessment processes. Recognizing this shift, our central assessment office transitioned to a leaner approach in 2023, freeing capacity for more innovative assessment initiatives, including a newly developed partnership model focused on institutional analyses related to learning and development. While we still collect assessment reports annually, our meta-assessment practice now uses an abbreviated checklist centered on the three elements required by accreditors: (1) well-defined student learning outcome statements, (2) direct measures aligned with those outcomes, and (3) use of results to seek improvement. Considering these developments, we began exploring whether emerging artificial intelligence technologies could help further streamline meta-assessment without compromising quality.

Artificial Intelligence in Assessment

AI has been used in a broad array of educational contexts, such as prediction of academic success (Abu Saa et al., 2019), early drop-out warnings (Howard et al., 2018), and automated grading (Ouyang et al., 2022). For assessment and measurement practitioners, AI presents exciting opportunities to enhance efficiency and expand capacity by automating time-intensive tasks such as item generation (e.g., Bezirhan & von Davier, 2023) and qualitative analysis (e.g., Slotnick & Boeing, 2024). However, to our knowledge, no studies have examined how AI could streamline the valuable but time-intensive meta-assessment process.

The Present Study

Given the largely unexplored potential of AI in meta-assessment, this study investigates the ability of large language models to analyze variations of an academic program assessment report and generate feedback comparable to that of a human assessment expert. We focus on three key questions:

1. Assessment report quality: How does the alignment between AI-generated assessment quality ratings and human ratings vary across strong, moderate, and weak assessment reports?
2. Assessment quality elements: How accurate are AI-generated ratings across different assessment elements (student learning outcomes, measures, and use of results)?
3. Evaluation format: Does agreement between AI and human ratings differ depending on whether a checklist or rubric is used?

By exploring these questions, we aim to better understand AI's potential to increase the efficiency of meta-assessment in higher education. To strengthen reliability and generalizability, we address the three research questions by comparing the performance of two LLM-based platforms, OpenAI's ChatGPT-5 and Microsoft's Copilot Pro, and by conducting five replications for each model.

Method

Study Design

This exploratory comparative design incorporated manipulations of report quality, assessment elements, and evaluation format to assess ChatGPT's and Copilot Pro's ability to simulate human meta-assessment. Report quality was manipulated directly within a single fictitious assessment report, resulting in strong, moderate, and weak versions. Performance was assessed across three core elements of assessment quality: student learning outcomes, measurement, and use of results. Evaluation format was defined by the use of either a checklist or a rubric. This structure enabled comparisons across realistic conditions that mirror how many institutions approach assessment. The following sections further describe how each variable was operationalized.

Assessment Report Quality

To evaluate AI's performance across varying report quality, we created a fictitious assessment report for a "Survival Studies, B.S." program. Drawing on our years of experience, we systematically manipulated it to produce strong, moderate, and weak versions. The strong version reflected best practices, with clearly articulated student learning outcomes (SLOs), strong outcome-measure alignment, and meaningful use of results for improvement. The moderate version included three degradations: less precise SLO language, removal of a column displaying outcome-measure alignment, and a vaguer Use of Results section.

The weak version reflected further degradations in clarity and quality: one SLO focused on faculty rather than students, the measurement table and key context were removed, and the Use of Results section included an inappropriate improvement plan ("We need to recruit better students"). Full copies of these fictitious report versions are available in Appendices A, B, and C.

Assessment Quality Elements

Although assessment cycles differ across institutions, three foundational elements are commonly present: student learning outcomes, measurement, and use of results for improvement. To maintain a standardized evaluation framework, this study focused exclusively on these core components. All report versions contained these three sections, while the specificity level for assessment of each element varied based on evaluation format.

Evaluation Format

We applied two scoring approaches drawn from institutional practice: a streamlined checklist and a more detailed rubric. Comparison of these formats allowed us to examine whether the degree of scoring detail influenced AI's ability to approximate human judgment. In the checklist condition, each of the three foundational elements

received a single binary rating (0 or 1), with a score of 1 assigned if two conditions were satisfied: the element was present and of at least moderate quality.

For the rubric condition, we used a subset of our former 14-point rubric, selecting only the criteria that were directly related to the three foundational elements. The rubric included two criteria for SLOs (specificity and student-centered orientation), four for measurement (relationship to SLOs, types of measures, presentation of results, and interpretation of results), and one for use of results for improvement. Each criterion was rated on a four-point scale ranging from Beginning (1) to Exemplary (4), allowing for a more nuanced evaluation.

Procedures

Before we generated AI output, a human assessment expert evaluated all three versions of the Survival Studies report using both the checklist and rubric. The expert ratings and corresponding commentary served as the benchmark for evaluating AI performance.

Our first step towards generating AI output was translating evaluation criteria into prompts for a text-to-text LLM. For the checklist format, prompt development was grounded in a metacognitive reflection task in which two higher education assessment experts independently identified the specific criteria that informed their evaluation of the strong version of the Survival Studies report. These criteria were then embedded in prompts following best practices, e.g., step-by-step instructions (Mollick, 2023), ensuring that prompt language reflected the natural cognitive processes of expert raters. Language for the rubric prompt was taken verbatim from the rubric used in our institution's evaluation format prior to 2023.

Prompts were refined over several rounds of preliminary trials for both the checklist and rubric formats (see Appendices D and E for full versions of final prompts). With many AI tools available, we ultimately selected ChatGPT-5 and Microsoft Copilot Pro for their wide accessibility and ease of use across experience levels, supporting replication studies.

Analysis

Finalized prompts were entered separately with each report into the free versions of ChatGPT-5 and Copilot Pro, with memory and sharing features turned off. For every replication and report, we opened a new chat to ensure independence across analyses. We then recorded both the numerical ratings and the accompanying rationale. This produced 60 unique AI-generated numerical ratings across all combinations of report versions, evaluation formats, assessment elements, and LLMs; each of these was then replicated five times, resulting in 300 individual scores (see Tables 1 and 2). AI-generated ratings were then compared to human expert ratings to calculate percent agreement for each condition (see Tables 3 and 4). Additionally, because the LLMs were prompted to provide rationale and commentary alongside each rating, an assessment expert reviewed this qualitative output to evaluate the accuracy and usefulness of the justifications.

Results

Tables 1 and 2 present human and AI ratings, score differences, and replication variability (standard deviations in parentheses) across report quality and assessment elements for the checklist and rubric formats, respectively. Across comparisons, the two LLM platforms and the human rater demonstrated stronger alignment on the checklist format than on the rubric format. Across five replications in the checklist format, ChatGPT-5 yielded 87% absolute agreement and Copilot Pro 89% (Table 3); in the rubric format, both displayed significantly less alignment (ChatGPT-5 = 44%, Copilot Pro = 50%), as presented in Table 4. Agreement generally decreased as report quality declined and was lowest for rubric items that required more interpretation.

Table 1. *Ratings for Checklist Evaluation Format Across Replications*

Report Quality	General Element	Human Score	ChatGPT-5 Score	Copilot Score	Human and ChatGPT5 Difference	Human and Copilot Difference
Strong	SLO	1	1(0)	1(0)	0	0
	Measure	1	1(0)	1(0)	0	0
	Improvement	1	1(0)	1(0)	0	0
Moderate	SLO	1	1(0)	1(0)	0	0
	Measure	1	1(0)	1(0)	0	0
	Improvement	1	1(0)	1(0)	0	0
Weak	SLO	1	0.8(0.45)	1(0)	0.2	0
	Measure	1	0.8(0.45)	1(0)	0.2	0
	Improvement	0	0.8(0.45)	1(0)	-0.8	-1

Note. Checklist ratings used a binary scale: 0 (*not satisfied*) or 1 (*satisfied*). Standard deviations are reported in parentheses

Assessment Report Quality

Across both evaluation formats, alignment with human ratings was highest for strong reports. When reports featured clearly articulated student learning outcomes, aligned measures, and specific use-of-results statements, both AI systems generally matched human judgment. For strong and moderate reports, both reproduced human ratings exactly in the checklist condition. When using the checklist for the weak report, ChatGPT-5 produced lower mean ratings (≈ 0.8 with SD ≈ 0.45) and greater variability, while Copilot Pro produced stable but inflated scores (always 1 on the Improvement element).

When using the rubric format, AI model performance for moderate and weak reports diverged. Copilot Pro results demonstrated a level of agreement that matched report quality (i.e., strong for the strong report and progressively weaker as quality declined). On weaker reports, Copilot scores were consistently inflated, as AI frequently failed to penalize missing or misclassified evidence. ChatGPT Results were less consistent, with most disagreement on individual scores occurring with the moderate report.

Table 2. *Ratings for Rubric Evaluation Format Across Replications*

Report Quality	General Element	Specific Element	Human Score	ChatGPT 5 Score	Copilot Score	Human and ChatGPT5 Difference	Human and Copilot Pro Difference
Strong	SLO	Clarity/Specificity	4	2.8(0.45)	3.6(0.55)	-1.2	-0.4
		Orientation	4	3.6(0.55)	4(0)	-0.4	0
	Measure	Relationship to SLOs	4	3.6(0.55)	3.6(0.55)	-0.4	-0.4
		Types	4	4(0)	4(0)	0	0
		Presentation of Results	4	3.2(0.45)	3.6(0.55)	-0.8	-0.4
		Interpretation of Results	4	3.4(0.55)	4(0)	-0.6	0
	Improvement	Program Modification Plan	3	3(0)	3.4(0.55)	0	0.4
Moderate	SLO	Clarity/Specificity	3	2(0)	2.6(0.55)	-1	-0.4
		Orientation	4	2.6(0.55)	3.6(0.55)	-1.4	-0.4
	Measure	Relationship to SLOs	3	3.2(0.45)	3(0)	0.2	0
		Types	3	4(0)	4(0)	1	1
		Presentation of Results	2	2.8(0.45)	3.2(0.45)	0.8	1.2
		Interpretation of Results	3	3.4(0.89)	4(0)	0.4	1
	Improvement	Program Modification Plan	3	3.2(0.45)	3(0)	0.2	0
Weak	SLO	Clarity/Specificity	2	2(0)	2.2(0.45)	0	0.2
		Orientation	1	2.2(0.45)	2.2(0.45)	1.2	1.2
	Measure	Relationship to SLOs	3	2.8(0.45)	3(0)	-0.2	0
		Types	3	3.4(0.55)	4(0)	0.4	1
		Presentation of Results	2	2.8(0.45)	3(0)	0.8	1
		Interpretation of Results	2	2.8(0.45)	3.2(0.45)	0.8	1.2
	Improvement	Program Modification Plan	2	2.4(0.55)	3(0)	0.4	1

Note. Rubric ratings were as follows: 1=Beginning, 2=Developing, 3=Good, 4=Exemplary. Standard deviations are reported in parentheses.

Table 3. *Agreement Between LLM and Human Ratings – Checklist*

Prompt Version	Report Version	General Element	Specific Element	Human & ChatGPT-5	Human & Copilot
Checklist	Strong	SLO	SLO	100%	100%
		Measure	Measure	100%	100%
		Improvement	Improvement	100%	100%
Checklist	Moderate	SLO	SLO	100%	100%
		Measure	Measure	100%	100%
		Improvement	Improvement	100%	100%
Checklist	Weak	SLO	SLO	80%	100%
		Measure	Measure	80%	100%
		Improvement	Improvement	20%	0%
Overall				87%	89%

Table 4. *Agreement Between LLM and Human Ratings – Rubric*

Prompt Version	Report Version	General Element	Specific Element	Human & ChatGPT-5	Human & Copilot
Rubric	Strong	SLO	Clarity/Specificity	0%	60%
			Orientation	60%	100%
		Measure	Relationship to SLOs	60%	60%
			Types	100%	100%
			Presentation of Results	20%	60%
			Interpretation of Results	40%	100%
		Improvement	Program Modification Plan	100%	60%
Rubric	Moderate	SLO	Clarity/Specificity	0%	60%
			Orientation	0%	60%
		Measure	Relationship to SLOs	80%	100%
			Types	0%	0%
			Presentation of Results	20%	0%
			Interpretation of Results	20%	0%
		Improvement	Program Modification Plan	80%	100%
Rubric	Weak	SLO	Clarity/Specificity	100%	80%
			Orientation	0%	0%
		Measure	Relationship to SLOs	80%	100%
			Types	60%	0%
			Presentation of Results	20%	0%
			Interpretation of Results	20%	0%
		Improvement	Program Modification Plan	60%	0%
Overall				44%	50%

However, despite this case-by-case misalignment, it is interesting to note that when averaging all scores and replications across the seven rubric elements, ChatGPT's overall score for the moderate report (3.03) closely matched that of the human rater (3.0).

Assessment Quality Elements

As a whole, elements related to Use of Results for Improvement showed the highest level of absolute agreement between human and AI (68%), followed by SLOs (61%) and Measures (54%). However, there was significant variability in the number of specific elements comprising these general categories and in the level of complexity related to each. A deeper look suggests that these metrics may underestimate performance for SLOs and overestimate it for Measures.

In the checklist condition, both SLO and Measure performance displayed near perfect alignment with human ratings. In the rubric condition, results for these elements again appeared similar, this time presenting moderate alignment with human ratings. However, AI had consistent difficulties judging outcome–measure alignment and distinguishing between direct and indirect evidence, trends related to measurement that did not always result in misalignment with human scores. This was especially the case in the binary checklist condition, where scores often matched but rationale differed.

In contrast, much of the disagreement in human and AI scores for SLOs when using the rubric seems to be the result of model differences, specifically with how ChatGPT adopted an unnecessarily strict linguistic standard when assessing SLOs in the strong report. This resulted in twice the amount of misalignment (27% agreement, .87 average difference with human ratings) compared to CoPilot (60% agreement, .43 average difference). ChatGPT seemed to have a consistent issue with the word "articulate," arguing that it was not sufficiently action oriented. This issue highlights ChatGPT's tendency to be more conservative in its ratings, while at least demonstrating a somewhat coherent line of reasoning. Overall, both AI platforms consistently identified vague or non-student-centered phrasing in weaker reports, but their numerical ratings were more lenient than those of the human rater.

For use of results, both models performed relatively well for strong and moderate reports, often identifying missing specificity in improvement plans. However, in the weak report condition, both overgeneralized the link between results and actions, leading to inflated scores. This mirrors the human rater's observation that the models sometimes credited intent rather than evidence.

Evaluation Format

Performance significantly varied by evaluation format. The checklist produced higher overall agreement (ChatGPT-5 = 87%, Copilot Pro = 89%) than the rubric (ChatGPT-5 = 44%, Copilot Pro = 50%). This difference reflects the binary checklist's simpler decision structure, which limits room for disagreement. The only major checklist discrepancy occurred with the Use of Results element of the weak report, where both models displayed an obvious inability to recognize that this report lacked evidence that improvement plans were informed by results.

The rubric format introduced more variability, especially when using ChatGPT, because it required finer differentiation across seven sub-criteria. Both models showed less variation across replications for objective criteria such as Types of Measures compared

to more subjective criteria like Orientation and Presentation of Results. Overall, ChatGPT-5 displayed stricter but more variable scoring, whereas Copilot Pro demonstrated consistent but more lenient evaluation.

Discussion

This study explored the potential of AI, specifically using ChatGPT-5 and Copilot Pro, to evaluate academic program assessment reports by comparing AI-generated ratings to those of a human expert. Across five replications per condition, AI aligned with human ratings at high levels (88% agreement) in the checklist condition and moderate levels (47% agreement) in the rubric condition. Agreement was strongest for high quality reports across evaluation formats. The most pronounced differences between human and AI ratings emerged in lower-quality reports and when evaluating the Measures component in the rubric condition, which included four specific elements (compared to two for SLOs and one for Improvement). This suggests that the more granular criteria of the rubric may challenge AI when it is used to evaluate reports that contain multiple errors. These findings highlight both the promise and limits of AI assistance, particularly in evaluating more complex components like measurement quality.

While rubric-based evaluations produced lower overall agreement, this format revealed important AI system differences. ChatGPT-5 tended to produce more conservative, variable scores, whereas Copilot Pro produced highly stable but inflated ratings. While ChatGPT's variability may reflect the capacity for nuanced sensitivity to report quality, it also leaves many questions about the underlying rationale being used when AI and human scores seem to align. For example, ChatGPT and the human rater arrived at a very similar average overall rating for the moderate report, but the way these overall scores were generated were very different, as ChatGPT matched the human rating for individual elements in only 10 out of 35 cases.

Both AI platforms displayed a midrange scoring tendency compared to human rating when using the rubric format, assigning slightly lower scores to the strong report and higher scores to the weak report. Whereas the decline from the strong to moderate to weak report was represented in human ratings by a .86 overall average score difference between each, both AI platforms produced approximately a .4 average difference between the reports. It is possible that this gap would be reduced with a few-shot prompting approach in which an LLM would benefit from multiple examples of human rationale in distinguishing between reports of varying quality.

From a practical standpoint, it is important to understand the types of errors AI may display when assessing reports. Theoretically, these could include 1) hallucinations (e.g., fabricated SLOs), which are common with LLMs (Özer, 2024); 2) misses, when AI fails to detect the presence or absence of key information; 3) misplaced attention, when AI focuses on present but irrelevant information; and 4) misapplication of criteria, when AI may identify relevant content but apply faulty judgment. In the current study, there was evidence of all these error types, and they often intersected to produce misaligned scores.

In one instance in the rubric condition, ChatGPT provided a rating of 2 for the strong report's SLOs, reasoning that the verbs in "understand key concepts of survival science" and "demonstrate knowledge of survival skills" were vague and hard to measure. While ChatGPT's logic is sound, neither of these SLOs exist on the strong report (or on any of the reports, for that matter). This was one of two true hallucinations we found in the

present study. Misses occurred most frequently with weaker reports, where the AI models occasionally failed to identify that elements were missing or of low quality, often inferring high overall quality from a single strong example while disregarding deficiencies. Another example of one of the few two-point gaps between human and AI ratings was partially due to an error in misplaced attention. In the rubric condition, Copilot rated the weak report's Interpretation of Results a perfect 4, providing rationale that was entirely based on the Use of Results section. In this instance, not only did Copilot extract information from the wrong area of the report, but also its application of that information seemed misguided. A clearer instance of a "faulty judgment" error was provided when ChatGPT correctly identified a problematic SLO on the moderate report ("Understand the essential survival skills…") but took issue with the SLO being "less student centered" rather than using a vague verb. This resulted in an overly low rating on the Orientation element.

Even when scores were in agreement, the underlying rationale provided by AI often conflicted with that of the human expert, an insight that became apparent after comparing qualitative commentary. For example, LLM and human scores on outcome-measure alignment frequently agreed, but rationales for these scores differed. When the human expert assigned less than full credit due to missing or incorrect alignment, the comment clearly articulated this nuance. AI, by contrast, tended to cite a general lack of specificity without identifying the root issue. This is especially consequential because other assessment elements are dependent on accurate evaluation of the relationship between measures and learning outcomes. Errors in this area seemed to cascade, distorting AI's judgment across other areas of the report (e.g., Presentation of Results).

Given the binary nature of the checklist ratings, which increased the likelihood that levels of agreement were inflated by chance (50% baseline vs. 25% for the rubric), this format masked minor errors that appeared in the AI commentary. Analysis revealed that AI often displayed unwarranted confidence in weaker report elements, whereas human commentary expressed hesitation when assigning ratings of 1 for elements that passed the threshold of moderate quality but nonetheless contained flaws. For instance, AI claimed that outcomes were mapped to measures across all report versions, while the human expert correctly flagged missing mappings in the medium- and low-quality versions. This did vary to some degree by model: ChatGPT-5 occasionally expressed uncertainty, reflected by non-integer averages across replications (e.g., 0.8 ± 0.45), signaling ambiguity in borderline cases, whereas Copilot Pro reported identical scores across replications.

Both LLM platforms frequently accepted report labels at face value, failing to critically assess their accuracy. If a measure was labeled "direct," AI typically accepted it without further scrutiny. For instance, both models accepted "exit interview" as a direct measure in the weak report. ChatGPT-5 occasionally noted ambiguity and returned fractional scores to indicate uncertainty; Copilot Pro consistently gave higher marks. Both models also often treated any mention of "change" as proof of data-informed improvement. This raises an important question for practice: can report writers mislead AI by using incorrect labels that sound appropriate? Untrained LLMs appear vulnerable to such surface-level cues. This reflects a broader challenge: human assessment expertise is likely not reducible to language alone. Expert reviewers use contextual frameworks to resolve linguistic contradictions, such as when verbal cues or lack of information in one report area contradict a label in another. Additionally, accurate interpretation of tables, charts, and other visual information requires nonverbal perceptual reasoning, an area

where LLMs remain notably weak relative to verbal comprehension (Galatzer-Levy et al., 2024).

Future research should explore whether customized LLMs, trained with annotated reports and embedded in a culture of assessment, can better detect quality indicators beyond surface labels. Importantly, humans without sufficient training (e.g., new graduate assistants or faculty on committees) are also susceptible to accepting labels at face value. This suggests that accurate assessment, whether by human or AI, depends on contextual depth. Notably, a healthy byproduct of using AI as a tool is that the process encourages assessment professionals to critically reflect on their own criteria and rationale, which itself can be a highly impactful form of training.

Implications for Research and Practice

These findings suggest that AI has the potential to be a valuable supplemental tool for higher education assessment, particularly when reviewing strong assessment reports or using a straightforward checklist. However, using LLM-based systems like ChatGPT-5 and Microsoft Copilot Pro for this purpose requires careful context-specific preparation. Human – AI rating alignment for SLOs was lower than we expected in this study, given the seemingly straightforward nature of this element. Despite this, we feel confident that with minimal additional training to clarify desired standards, LLMs would be able to accurately appraise SLOs in a way that would result in high alignment with human ratings. AI reliably flagged vague language and non-student-centered phrasing in weaker reports. This type of AI feedback may be especially useful for faculty or staff new to assessment practices when submitting or reviewing reports.

AI was less effective in evaluating measurement quality, a critical aspect of assessment. Given the complexity of SLO-measure alignment, practitioners might use AI primarily for initial reviews, particularly focusing on SLOs or use of results. When report quality is expected to be high, AI can streamline feedback and improve efficiency. In a hybrid model, human expertise would remain focused on more nuanced elements like measurement quality and alignment. This approach could help offset staffing limitations in assessment offices.

Despite AI's benefits, human oversight is essential, especially for programs with weaker assessment practices. AI struggled most with low-quality reports, which are indicative of situations where faculty need both feedback and developmental support—roles AI cannot fill (yet?). Institutions considering AI for meta-assessment should establish clear guidelines for review and verification to ensure feedback aligns with professional standards.

Limitations and Future Research Directions

This study has several limitations that suggest important directions for future research. One key limitation is the inherent variability of large language models, which generate different outputs each time they are prompted. Although our study utilized two LLM-based systems and five replications per condition, our findings still provide just a snapshot of AI behavior. It should be noted that in practice, effective use of language models often requires back and forth dialogue between the user and the LLM, as prompts often require slight adjustment in relation to any apparent misunderstandings the model has of the task. While making subtle changes to our prompt wording would have

potentially affected the results in meaningful ways, we intentionally opted to capture the results generated from initial prompt sequences to ensure consistency with our existing human processes. Our overarching goal in this endeavor was to provide assessment practitioners with a snapshot of the alignment between AI and human ratings across multiple dimensions when holding constant the language used in the instructions. This has practical implications, as programs planning to assess a high volume of reports with the help of LLMs might find engaging in back-and-forth manual prompt alterations unrealistically labor intensive. Nonetheless, future studies should explore the ways that subtle but systematic prompt adjustments might improve overall performance.

A second limitation involves the scope of LLMs examined. Although this investigation compared two distinct general-purpose AI platforms, Copilot Pro frequently relies on the same GPT models that ChatGPT uses. A greater range of additional LLMs should be included in future studies. Additionally, both platforms were evaluated in their default configurations. This was intentional for the purposes of our study, which used fictional data, as institutional applications of AI involving real assessment reports may require more secure or customizable tools. Future research should include additional models (e.g., Claude, Gemini, or domain-specific educational LLMs) and both free and paid versions to examine how architecture, context windows, and training data affect performance in assessment tasks. This underscores the need for comprehensive comparisons of multiple LLMs, including paid versions, to evaluate performance and reliability in assessment contexts.

Ethical dimensions also warrant further consideration. When institutional data are involved, it is critical to understand how information entered into an LLM may be stored, reused, or used for model training. While our use of fictional reports obviated the concerns inherent in the use of actual institutional data, this challenge remains relevant for real-world application. In addition, assessment professionals should think carefully about how to disclose the use of AI in feedback processes, especially in contexts where human evaluation is expected or preferred.

Additionally, this study relied on a single expert to evaluate AI responses. While this ensured consistency, it limited diversity of perspectives and feedback. Future research should involve multiple expert reviewers and incorporate real assessment data to reflect the variety of practices across institutions and disciplines. Broader testing will help determine how well AI generalizes to different institutional contexts.

Measurement quality emerged as one of the most difficult elements for AI to evaluate. Because assessing measurement involves both alignment (e.g., between outcomes and measures) and classification (e.g., direct vs. indirect measures), it demands a level of judgment that may exceed the capacity of current general-purpose models. Refining prompts to encourage deeper analysis by providing more examples of human rating decisions may help address this gap. A more ambitious direction involves developing customized AI models specifically trained for the purpose of meta-assessment, an effort that could significantly enhance precision and awareness of educational assessment contexts.

Finally, future studies should also include structured analyses of AI-generated qualitative data alongside analyses of numerical ratings. While this study prompted AI to provide rationale with each score and informally compared it to expert feedback, we did not conduct a formal qualitative analysis. The inconsistencies observed in justifications

and recommendations point to the value of deeper review. Because written feedback often drives dialogue and improvement at the program level, understanding how AI generates and communicates rationale is critical to assessing its potential role in meta-assessment.

Conclusion

As AI continues to evolve, future models may better distinguish between direct and indirect measures, evaluate alignment more accurately, and offer actionable feedback for assessment practices. Yet even as its capabilities grow, AI is best positioned as a supplemental tool, not a substitute for human expertise. When used thoughtfully, AI has the potential to streamline routine aspects of meta-assessment, allowing professionals to focus on more complex work where their judgment is most needed. Institutions should take a cautious but proactive approach, leveraging AI where it adds value while keeping student learning and the human relationships that support it at the center of meaningful assessment practice.

Appendices

Appendicies A-E: Survival Studies Reports & ChatGPT Prompts

References

Abu Saa, A., Al-Emran, M., & Shaalan, K. (2019). Factors affecting students' performance in higher education: A systematic review of predictive data mining techniques. Technology, Knowledge and Learning, 24(4), 567–598. https://doi.org/10.1007/s10758-019-09408-7

Bezirhan, U. & von Davier, M. (2023). Automated reading passage generation with OpenAI's large language model. Computers and Education: Artificial Intelligence, 5. https://doi.org/10.1016/j.caeai.2023.100161

Fulcher, K.H., Good, M.R., & Sanchez, E.R.H. (2024). Assessment 101 in higher education: The fundamentals and how to apply them. Routledge. https://doi.org/10.4324/9781003465553

Fulcher, K.H., Good, M.R., & Smith, K.L. (2015). 2015 Update to Rubric: The assessment progress template rubric. Product developed by the Center for Assessment and Research Studies, James Madison University, Harrisonburg, VA. https://www.jmu.edu/assessment/_files/apt_rubric_sp2015.pdf

Fulcher, K. & Good, M. (2013, November). The surprisingly useful practice of meta-assessment. Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA).

Galatzer-Levy, I. R., Munday, D., McGiffin, J., Liu, X., Karmon, D., Labzovsky, I., Moroshko, R., Zait, A., & McDuff, D. (2024). The cognitive capabilities of generative AI: A comparative analysis with human benchmarks. arXiv. https://arxiv.org/abs/2410.07391

Howard, E., Meehan, M., & Parnell, A. (2018). Contrasting prediction methods for early warning systems at undergraduate level. Internet and Higher Education, 37, 66–75. https://doi.org/10.1016/j.iheduc.2018.02.001

Microsoft. (2025). Copilot (Sep 28 version) [AI assistant]. https://www.microsoft.com/copilot

Mollick, E. (2023, November 1). Working with AI: Two paths to prompting. One Useful Thing. https://www.oneusefulthing.org/p/working-with-ai-two-paths-to-prompting

OpenAI. (2025). ChatGPT-5 (Oct 15 version) [Large language model]. https://chat.openai.com/chat

Ory, J. C. (1992). Meta-assessment: Evaluating assessment activities. Research in Higher Education, 33(4), 467–481. https://doi.org/10.1007/BF00973767

Ouyang, F., Zheng, L., & Jiao, P. (2022). Artificial intelligence in online higher education: A systematic review of empirical research from 2011 to 2020. Education and Information Technologies, 27(6), 7893–7925. https://doi.org/10.1007/s10639-022-10925-9

Özer, M. (2024). Is Artificial Intelligence hallucinating? Turkish Journal of Psychiatry, 35(4), 333–335. Advance online publication. https://doi.org/10.5080/u27587

Rodgers, M., Grays, M. P., Fulcher, K. H., & Jurich, D. P. (2013). Improving academic program assessment: A mixed methods study. Innovative Higher Education, 38, 383–395. https://doi.org/10.1007/s10755-012-9245-9

SACSCOC. (2024). Principles of Accreditation: Foundation for Quality Enhancement. Southern Association of Colleges and Schools: Commission on Colleges. https://sacscoc.org/app/uploads/2024/01/2024PrinciplesOfAccreditation.pdf

Slotnick, R. C. & Boeing, J. Z. (2024). Enhancing qualitative research in higher education assessment through generative AI integration: A path toward meaningful insights and a cautionary tale. New Directions for Teaching and Learning. https://doi.org/10.1002/tl.20631