Human Interaction and Emerging Technologies (IHIET 2026), Vol. 236, 2026, 211–225
https://doi.org/10.54941/ahfe1008097

A Multimodal, Uncertainty-Aware, and Transparent AI Grading Tool for Scalable Automated Grading in AI and Data Science Higher Education

Olivia Dias[1], Cynthia Breazeal[2], and Kantwon Rogers[2]

[1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA
[2]Media Arts and Science Department, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA

ABSTRACT

Multimodal assignments- programming, free response, and oral - are difficult for humans to grade consistently at scale. Artificial Intelligence (AI) assisted Automated Grading Tools (AGTs) offer a path forward, but existing systems are evaluated within a single modality and lack a framework for routing assignments to human review based on AI uncertainty and modality. This paper evaluates a multimodal AGT on 95 assignments: Programming (n = 35), free response (n = 30), and oral (n = 30), each scored by two human graders and the AI. This paper presents three novel contributions to the field of AI assisted AGT. First, it contributes a per-modality validity characterization including the first directionally-resolved test on mock-interview grading. Second, it presents the exemplar-anchored harshness hypothesis as a falsifiable mechanism for AI severity on subjective modalities. Finally, it provides an empirical test of semantic-entropy escalation finding no usable routing signal, based on the sample size and experiment settings.

Keywords: Automatic grading tool, AI, multimodal, semantic entropy, Large language model, Inter-rater reliability

INTRODUCTION

Undergraduate enrollments in computer science fields continue to grow, and introductory courses in computer science and engineering now routinely enroll hundreds of students per term (Batten, 2025). At the same time, courses have shifted from single modality assignment types to multimodal tasks- such as students writing technical reports or open ended reflections, writing code, and delivering video presentation or final projects to assess students' competency in a topic (Batten, 2025; Gao et al., 2020; Patil et al., 2024). The assignments used to assess students' enrolled in these courses are hard to grade at scale due to the following reasons: open-ended writing invites subjective interpretation, code can be correct in many ways, and oral presentations must be evaluated on content and delivery simultaneously.

Received June 13, 2026; Revised June 22, 2026; Accepted July 6, 2026; Available online August 25, 2026

Human grading of this work is time-intensive, susceptible to grader bias, and inconsistent across graders, and these problems compound as enrollments rise (Deepshikha, 2025; Deeva et al., 2021; Gonzalez et al., 2026).

In order to mitigate these adverse downsides to traditional human grading, higher education seeks to use AI assisted grading to reduce the time to grade student work and upscale individual, objective student feedback and grading (Deepshikha, 2025; Deeva et al., 2021; Cavalcanti et al., 2021; Keuning et al., 2018). This paper presents a brief review of prior work in automatic grading systems in Section 2. Section 3 describes the methodology and system design. Section 4 presents the experimental results and discusses broader implications of the results. Section 5 concludes the paper.

LITERATURE REVIEW

AI assisted AGTs for a specific modality- programming, written response, or video-based- have seen substantial progress, however their synthesis remains underdeveloped. Existing AGTs address each modality, predominantly, in silos due to the course assignment needs, i.e. a computer science course with only programming assignments or a humanity course with only essays and no technical component. They do not grapple with the compounded reliability and implementation challenges that arise when evaluating across multiple modalities- coding, written response, and oral.

Studies that utilize large language models (LLMs) in their AGTs to evaluate programming assignments have shown that LLMs are good at objective grading of code functionality, however have varying results when it comes to awarding partial credit or assessing code style (Deepshikha, 2025; Deeva et al., 2021; Cavalcanti et al., 2021; Keuning et al.; Zacharis & Papadakis, 2025; Grandel et al., 2024). Automated Essay Scoring (AES) systems that utilize LLMs, such as ChatGPT or BERT, consistently reported to have sensitivity to prompt design and difficulty assessing deeper reasoning behind the LLM generated score (Seßler et al., 2025; Chen & He, 2013; Nguyen & Park, 2025). However AES systems have demonstrated the ability to provide fairly consistent scoring and achieve similar essay scores relative to a human grader, given a carefully engineered prompt, output format, and example-rich rubric (García-Varela et al., 2025; Quah et al., 2024; Seßler et al., 2025; Cisneros-González et al., 2025). Zhang et al. (2024) presents a Language-based Long-range Video Question Answering (LVQA) framework that parses a video into chunks and summarizes the shorter chunks into a succinct overview of a longer video in order to improve an LLM's ability to assess video content. Automated scoring for asynchronous video interviews (AVIs) is comparable to human graders or evaluators, given rigorous prompt engineering (Zhang et al., 2024; Uppalapati et al., 2025; Lv et al., 2024).

Each of the three major modalities-code, written, and video- has produced capable automated grading systems, and each has independently identified the core reliability challenge: performance is strong on objective, constrained evaluation and degrades on open-ended, subjective judgment. However, one critical limitation persists across the literature. All three modalities have never been integrated into a unified grading framework. Systems are designed, trained, and evaluated in isolation, with no mechanism to coordinate their

outputs, reconcile their reliability characteristics, or provide a consistent grading experience in a course with assignments of multiple modality. This paper addresses this gap by presenting a multimodal, uncertainty aware AI assisted grading system that operates across code, written, and oral assignments within a single rubric-aligned grading pipeline.

METHODOLOGY

This section describes the design and implementation of a multimodal AI assisted AGT. The AGT integrates LLMs, semantic entropy based uncertainty estimation as proposed by Farquhar et al. (2024), and retrieval augmented grading into its architecture. The paper tests the statistical significance of the results by using the average of two human grader as a ground truth and conducting a feature based disagreement analysis using Ridge and Logistic regression. The system is designed to (i) reliably assess diverse data-science assignments spanning code, written, and oral modalities; (ii) support higher education evaluation of assignments through transparent, auditable scores.

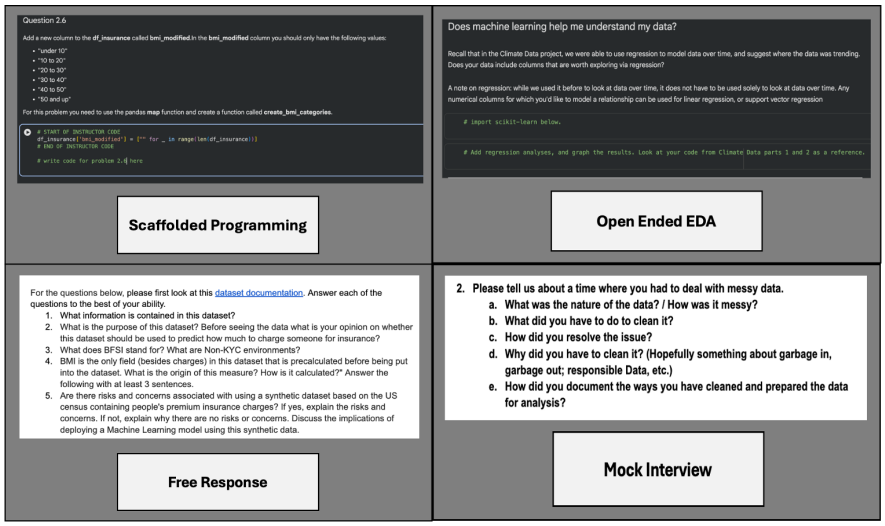
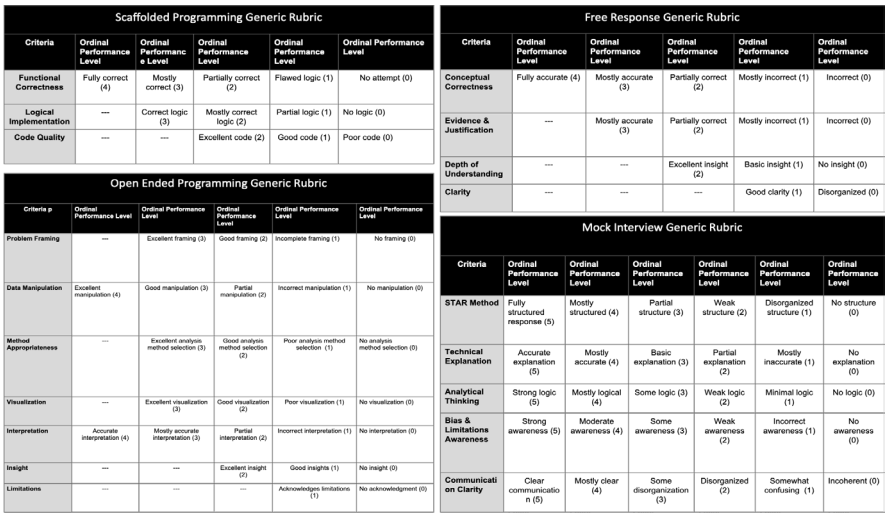


Figure 1: A screenshot of an assignment question from the Scaffolded programming, Open ended EDA programming, Free Response, and Mock Interview assignment type.

ASSIGNMENTS

The AI assisted AGT is run locally on the terminal within an IDE, and accesses assignments, rubrics, and answer keys stored locally from the computer directory. The following assignment types were input into the AI assisted AGT: scaffolded programming assignments, open ended exploratory data analysis (EDA) programming assignment, free response assignments, and mock data science interviews. For the programming assignments, the students wrote code in response to Python coding questions within a Google Colab notebook. For the free response assignments, students had to respond to a series of questions or open ended tasks that would assess their responsible AI and data science knowledge they gained throughout a given

module. Additionally, the free response assignments consist of questions that assess technical design, data analysis, and ethical questions related to data analysis and AI implementation strategies. Finally, the students did a mock data science interview, which was recorded. The mock data science interview consisted of six questions that assesses a student's data science or AI group project, internship, coursework, or work experience.



Scaffolded Programming Generic Rubric

Criteria	Ordinal Performance Level	Ordinal Performanc e Level	Ordinal Performance Level	Ordinal Performance Level	Ordinal Performance Level
Functional Correctness	Fully correct (4)	Mostly correct (3)	Partially correct (2)	Flawed logic (1)	No attempt (0)
Logical Implementation	---	Correct logic (3)	Mostly correct logic (2)	Partial logic (1)	No logic (0)
Code Quality	---	---	Excellent code (2)	Good code (1)	Poor code (0)

Open Ended Programming Generic Rubric

Criteria p	Ordinal Performance Level	Ordinal Performance Level	Ordinal Performance Level	Ordinal Performance Level	Ordinal Performance Level
Problem Framing	---	Excellent framing (3)	Good framing (2)	Incomplete framing (1)	No framing (0)
Data Manipulation	Excellent manipulation (4)	Good manipulation (3)	Partial manipulation (2)	Incorrect manipulation (1)	No manipulation (0)
Method Appropriateness	---	Excellent analysis method selection (3)	Good analysis method selection (2)	Poor analysis method selection (1)	No analysis method selection (0)
Visualization	---	Excellent visualization (3)	Good visualization (2)	Poor visualization (1)	No visualization (0)
Interpretation	Accurate interpretation (4)	Mostly accurate interpretation (3)	Partial interpretation (2)	Incorrect interpretation (1)	No interpretation (0)
Insight	---	---	Excellent insight (2)	Good insights (1)	No insight (0)
Limitations	---	---	---	Acknowledges limitations (1)	No acknowledgment (0)

Free Response Generic Rubric

Criteria	Ordinal Performance Level	Ordinal Performance Level	Ordinal Performance Level	Ordinal Performance Level	Ordinal Performance Level
Conceptual Correctness	Fully accurate (4)	Mostly accurate (3)	Partially correct (2)	Mostly incorrect (1)	Incorrect (0)
Evidence & Justification	---	Mostly accurate (3)	Partially correct (2)	Mostly incorrect (1)	Incorrect (0)
Depth of Understanding	---	---	Excellent insight (2)	Basic insight (1)	No insight (0)
Clarity	---	---	---	Good clarity (1)	Disorganized (0)

Mock Interview Generic Rubric

Criteria	Ordinal Performance Level	Ordinal Performance Level	Ordinal Performance Level	Ordinal Performance Level	Ordinal Performance Level	Ordinal Performance Level
STAR Method	Fully structured response (5)	Mostly structured (4)	Partial structure (3)	Weak structure (2)	Disorganized structure (1)	No structure (0)
Technical Explanation	Accurate explanation (5)	Mostly accurate (4)	Basic explanation (3)	Partial explanation (2)	Mostly inaccurate (1)	No explanation (0)
Analytical Thinking	Strong logic (5)	Mostly logical (4)	Some logic (3)	Weak logic (2)	Minimal logic (1)	No logic (0)
Bias & Limitations Awareness	Strong awareness (5)	Moderate awareness (4)	Some awareness (3)	Weak awareness (2)	Incorrect awareness (1)	No awareness (0)
Communicati on Clarity	Clear communicatio n (5)	Mostly clear (4)	Some disorganization (3)	Disorganized (2)	Somewhat confusing (1)	Incoherent (0)

Figure 2: The criteria and corresponding ordinal point values for the Scaffolded Programming, Open Ended EDA, Free Response, and Mock Interview Generic Rubrics.

RUBRIC DESIGN

In order to quantitatively measure performance on these assignments the paper utilizes rubrics to coordinate grading alignment across the human and LLM evaluators (Jonsson & Svingby, 2007; Reddy & Andrade, 2010). The researcher created four generic rubrics, tailored to the contents of the course, that corresponded to the four assignment types- scaffolding programming assignment, open ended EDA programming assignment, free response assignment, and mock data science interview assignment. There are four generic rubric types: Scaffolded Programming Generic Rubric used to evaluate Scaffolded Programming assignment types, Open Ended EDA Generic Rubric used to evaluate Open Ended EDA assignment types, Free Response Generic Rubric used to evaluate Free Response assignment types, and Mock Interview Generic Rubric used to evaluate Mock Interview assignment types. Each generic rubric type has different criteria for evaluating its corresponding assignment type. For each rubric criteria there is a series of ordinal point values to score the performance or evaluate the student's assignment. Each criterion is scored on different ordinal point values. A short description is provided for each ordinal point value to aid the LLM and human grader in understanding the difference between the different ordinal point values. This ordinal point values methodology is utilized to enable partial credit scoring for a response.

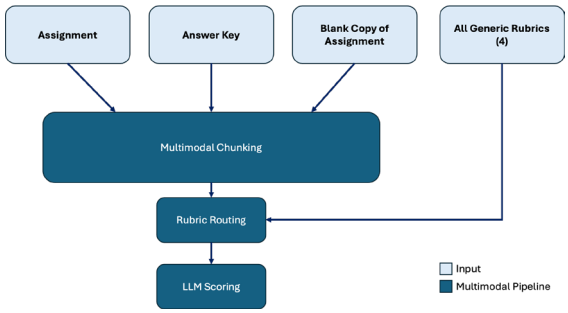


Figure 3: The input and multimodal grading pipeline for scoring an assignment.

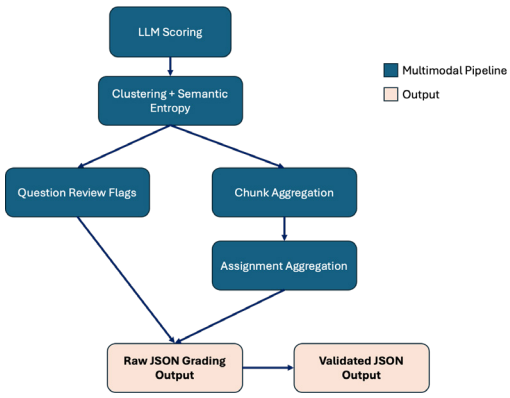


Figure 4: The multimodal grading pipeline and output for enforcing an auditable and transparent AI system.

AI MULTIMODAL GRADING PIPELINE

The AI assisted AGT is a multimodal grading pipeline designed as a staged, auditable procedure. As shown in Figure 3, a single completed assignment is inputted into the pipeline as a raw assignment artifact. The raw assignment artifact is one of the four assignment types described in Section 3.1 of the paper methodology. Additionally, supplementary artifacts are ingested into the pipeline with the student assignment to aid the LLM in scoring the assignment. As shown in Figure 3, the following supplementary artifacts are ingested to aid the grading and evaluation process: a blank copy of the assignment, all the generic rubrics, and the answer key for the assignment which is used as a reference for an exemplar assignment, i.e. an assignment that would earn the highest ordinal point value for every criteria on the generic rubric. The assignment exits the multimodal grading pipeline as a structured JSON grading record containing per-criterion scores, justifications for each criterion score, the per question score, final aggregate assignment score, and a confidence measure for each score produced. The pipeline is composed of eight principal stages: (1) Multimodal Chunking, (2) Rubric Routing (3) LLM Scoring (4) Clustering + Semantic Entropy (5) Question Review Flags

(6) Chunk Aggregation (7) Assignment Aggregation, and (8) Json output and validation. Figure 3 and 4 summarizes the architecture.

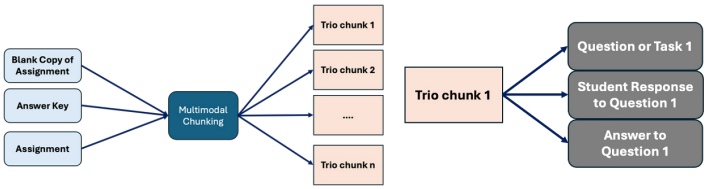


Figure 5: Illustrates the Multimodal Chunking process, stage 1 of the multimodal grading pipeline.

The multimodal chunking process ingests three raw artifacts- a blank assignment copy (for instructional context), the student submission (the material to be graded), and the answer key (an exemplar reference) - which are parsed by OpenAI's chatgpt5.4-nano model and then decomposed by the larger claude-opus-4-7 model into a list of uniquely identified GradingChunk objects (referred to as trio chunks in Figure 5), each consisting of a question, student response, and reference answer. This triad structure is applied across all assignment types: Scaffolded Programming and Open-Ended EDA are chunked by task, student code, and reference code; Free Response submissions are reflowed into question, response, and exemplar segments; and Mock Interview audio files are transcribed via OpenAI's Whisper model before being segmented into interviewer prompt, interviewee response, and reference response. This per-question decomposition is essential because it lets the chatgpt5.4-nano grading model score each question independently against its exemplar reference, creating an audit trail of the LLM's scoring logic that reflects the transparent and auditable design of the AI-assisted AGT.

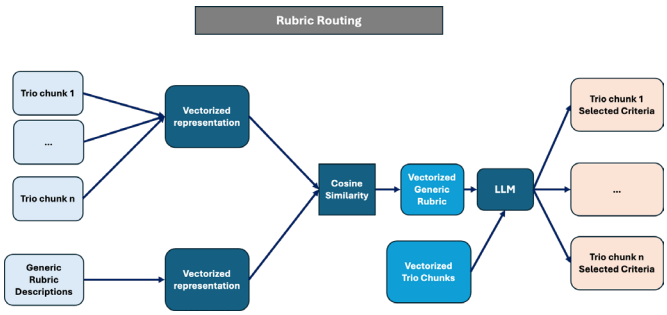


Figure 6: Illustrates the Rubric routing process, stage 2 of the multimodal grading pipeline.

To enable a retrieval-augmented generation (RAG) approach in which the pipeline is not told the assignment type, the per-question trio chunks are vectorized and reduced to an L2-normalized mean embedding that captures the assignment's modality-specific components-coding, oral, or written-which is then compared via cosine similarity against the vector embeddings of a simple description for the different generic rubric types. The generic rubric type with the greatest cosine similarity is selected for grading the assignment as shown in Figure 6. The claude-opus-4-7 model next takes the

vectorized trio chunks and the selected generic rubric and determines which of its criteria apply to each question, producing a custom per-question rubric that omits inapplicable criteria—for instance, excluding visualization and interpretation criteria from an Open-Ended EDA data-cleaning question so the student is not erroneously penalized for omitting a visualization they were never asked to produce. Each custom rubric is serialized as JSON and persisted alongside the assignment record, so that every assignment-student pair retains a reproducible rubric after grading, demonstrating the transparent and auditable design of the AI-assisted AGT.

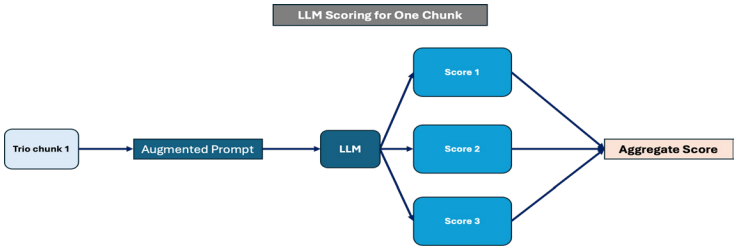


Figure 7: Illustrates the LLM scoring process, stage 3 of the multimodal grading pipeline.

In the LLM scoring stage, a single model, chatgpt5.4-nano, is prompted with each trio chunk's question, student response, and answer context. It then scores each chunk against its custom question-level rubric K = 3 times at temperature T = 0.3, a value within the conventionally low range [0.0, 0.3] recommended for factual accuracy, code evaluation, and structured data extraction, chosen so the model retains just enough stochasticity for a meaningful semantic-entropy estimate without producing qualitatively different grades on well-specified criteria (CodeSignal, n.d.). Generating three sample scores per chunk emulates having three independent graders and supports uncertainty estimation by revealing how consistently the LLM reproduces the same score for a given trio chunk.

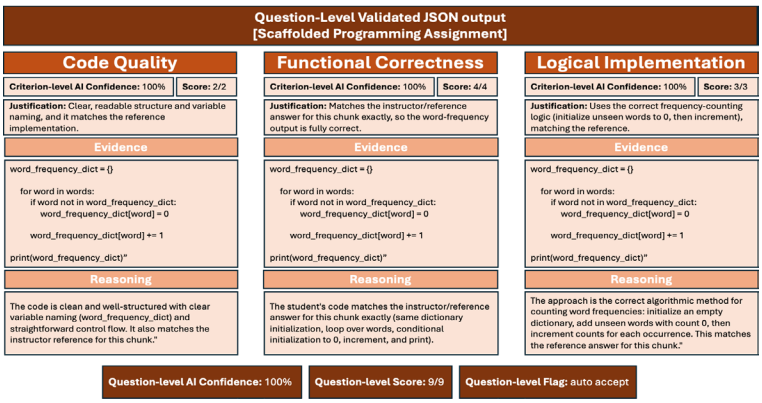


Figure 8: The multimodal grading pipeline's final output is a structured JSON file based on the validation schema that enforces specific fields to be present in the JSON file output- i.e. justification field, evidence field, reasoning field, etc. This is a visual representation of a single question score from an actual output from one of the Scaffolded Programming assignments.

After the LLM produces K = 3 sample scores for a trio chunk, the pipeline clusters semantically equivalent scores, computes each cluster's probability (Equation 1), and derives the semantic entropy (SE) (Equation 2)- where a low SE indicates the scores concentrate on one cluster and imply high confidence and a high SE indicates self-disagreement for the LLM and implies low confidence. Therefore, this measures how consistently the LLM evaluates the task rather than the correctness or inherent difficulty of the student's answer, with the resulting quantity feeding the question review flag stage. The SE is then normalized (Equation 3) and inverted into an AI confidence score (Equation 4) that is scale-invariant and therefore comparable across questions with differing numbers of possible scoring outcomes (a consequence of the per-question custom rubrics from Section \ref{sec:3_4_2}); this confidence is computed at the question level to preserve localized uncertainty and then aggregated into a final assignment-level confidence.

Equation 1. $P(c) = \frac{1}{K} \sum_{k=1}^{K} 1 [s\_k \in c]$, *where c is a specific score cluster and s_k is a specific sample score*

Equation 2. $SE = -\sum_{c=1}^{C} P(c) * \ln P(c)$, *where C is the set of all score clusters*

Equation 3. $SE\_norm = \frac{SE}{log(|C|)}$

Equation 4. $AI\_confidence = 1 - SE\_norm$

In the final pipeline stages, each trio chunk's score, AI confidence, and SE pass through question-level flag checks (auto-accept above 85% AI confidence, caution at 50–85% AI confidence, human review required at less than 50% AI confidence, flagged when the parse-fail rate exceeds 35%, and escalate below 50%) and JSON validation checks (using a validation schema for the structured JSON output), after which the per-question scores are averaged into an assignment-level score S_AI and validated against a schema, with records that fail validation or fall below 50% AI confidence flagged for mandatory human review rather than silently discarded so low-reliability cases remain auditable and the human retains ultimate authority over the final score. One human grader was responsible for determining which generic rubric type to apply to an assignment and how to section out each assignment. Then the two human graders independently grade the assignment and the average of the two human scores is used as ground truth score. The human graders used the same answer key and generic rubrics as the AI assisted AGT. For the systematic disagreement analysis, the study extracts and normalizes modality-specific features—text (response length, lexical diversity, evidence markers), code (line count, cyclomatic-complexity proxy, function count), and transcribed audio (speaking tempo, filler-word rate, Situation Task Action Result (STAR) completeness) to predict the paired disagreement (Equation 5).

Equation 5. D_score = S_AI – S_human, *where S_human is the average of the two human grader scores*

A Ridge regression with standardized coefficients quantifies how much and in which direction each feature shifts the AI score relative to the human score, while a Logistic regression predicts the high-disagreement indicator (Equation 6).

Equation 6. D_high = 1[| D_score | ≥ 10%], *where D_high = 1 means there is an absolute disagreement greater than 10% and D_high = 0 means there is an absolute disagreement less than 10%*.

This identifies the most influential features, together revealing where the AI grader is reliable, where it requires human oversight, and how future prompts, rubrics, and review policies should be improved.

RESULTS AND DISCUSSION

This section evaluates how closely AI grading matched human grading across three assignment modalities - Programming, Free Response, and Oral - and examines the characteristics of assignments that tend to produce the largest differences between the AI generated score and the human generated score. All analyses are performed at the assignment level, since every row in the dataset is a single assignment scored by both an AI grader and a human grader. The dataset contains 95 assignments. Of these 95 assignments, 35 are Programming assignments, 30 are Free Response assignments, and 30 are Oral assignments (captured as mock data science interviews). Two-way random-effects, absolute- agreement intraclass correlation coefficients (ICC) between the two human graders give ICC(2, k) = 0.649 [95% Confidence Interval: 0.42, 0.81] overall - moderate reliability - with substantial variation by modality (Programming has an ICC of 0.48, Free Response has an ICC of 0.82, and Oral video has an ICC of 0.62).

Table 1: Wilcoxon signed-rank tests of D_score by modality. The rank-biserial correlation is the effect size that complements the signed-rank test.

Modality	n	Median Paired Difference	Test Statistic W	P Value Raw	P Value Holm Bonferroni	Rank Biserial Effect Size	95% Median CI
Programming	35	–0.019	228.0	0.158	0.158	–0.276	[–0.124, 0.033]
Oral video	30	–0.149	0.0	1.86×10^{-9}	5.59×10^{-9}	–1.000	[–0.201, –0.116]
Free Response	30	–0.106	114.0	0.014	0.027	–0.510	[–0.136, –0.003]
POOLED	95	–0.115	902.0	3.14×10^{-7}	n/a (pooled)	–0.604	[–0.132, –0.084]

Table 2: Ridge and logistic regression model performance on signed disagreement (Repeated K Fold, 5 splits × 10).

Model	Metric	Mean	Std
ridge	CV MAE (test)	0.1437	0.0305
ridge	CV MAE (train)	0.1238	
ridge	CV R^2 (test)	−0.2711	0.3483
ridge	R^2 (train)	0.1489	
logistic	CV Accuracy (test)	0.4358	0.1069
logistic	CV ROC AUC (test)	0.4080	0.1249

Table 3: Ridge standardized coefficients on signed disagreement. Positive values indicate the feature pushes the prediction toward human > AI.

Feature	Standardized Coef
Programming: code length	−0.042153
Text: length of written response	−0.026863
Programming: code style metric	0.022984
Oral video: conceptual explanation	−0.022134
Text: sentiment	−0.018855
Programming: test coverage	−0.013980
Oral video: tempo	−0.010081
Oral video: fluency	0.005186
Text: lexical sophistication	−0.004371

Table 4: Logistic standardized coefficients on the high-disagreement flag, which is based on D_high as defined in Equation 6. Positive values raise the probability of high disagreement.

Feature	Standardized Coef
Text: length of written response	0.546153
Text: lexical sophistication	0.285169
Oral video: conceptual explanation	0.256210
Programming: code length	0.218685
Text: sentiment	0.128736
Programming: code style metric	−0.099140
Oral video: fluency	0.095351
Oral video: tempo	−0.084957
Programming: test coverage	0.033010

Table 5:. Per-modality Mann–Whitney U tests comparing SE between high- and low-disagreement groups (i.e. D_high = 1 to). One-sided alternative: SE higher in D_high = 1. None significant at $\alpha = 0.05$. Per-modality Spearman correlations between SE and |D_score|. Only Programming reaches one-sided significance (highlighted).

Scope	Mean SE: High-Disagreement Group	Mean SE: Low-Disagreement Group	Mann-Whitney U	Mann-Whitney p (Two-Sided)	Mann-Whitney p (one-Sided, Hypothesized Direction)	Spearman ρ	Spearman p (Two-Sided)	Spearman p (One-Sided, Hypothesized Direction)
Programming (n = 35; 19 high, 16 low)	0.3313	0.2497	181.0	0.3412	0.1706	+0.3026	0.0773	0.0386
Free response (n = 30; 22 high, low)	0.4875	0.5258	69.0	0.3932	0.8161	+0.0661	0.7287	0.3643
Oral video (n = 30; 20 high, 10 low)	0.3367	0.4067	77.5	0.3326	0.8445	-0.2256	0.2306	0.8847

The Wilcoxon signed-rank tests (Table 1) confirm a systematic AI-stingy bias that varies sharply by modality. Oral video shows the largest and most consistent gap (median D_score = −0.149, rank biserial r = −1.000, Holm corrected $p < 0.001$), meaning the AI scored below the human grader ground truth on essentially every assignment, followed by free response (median = −0.106, r = −0.510, p = 0.027); programming alone is statistically indistinguishable from zero (median = −0.019, r = −0.276, p = 0.158). The pooled effect is significantly negative (median = −0.115, r = −0.604, $p < 0.001$). The feature-based models, however, fail to explain this disagreement: the Ridge regression generalizes worse than the mean baseline (CV test R^2 = −0.271 despite train R^2 = 0.149), and the logistic classifier performs at or below chance (accuracy = 0.436, ROC AUC = 0.408) (Table 2). Standardized coefficients are correspondingly weak, no Ridge feature exceeds |0.042| (Table 3), though the logistic model assigns the largest positive weights to text response length (+0.546), lexical sophistication (+0.285), and oral-video conceptual explanation (+0.256) (Table 4). SE provides little discriminative signal: no modality shows a significant Mann–Whitney separation in SE between high- and low-disagreement groups, i.e. scores that are classified to have absolute disagreement above or below 10%, respectively. Only programming reaches one-sided significance in the SE, |D_score| Spearman correlation (ρ = +0.303, p = 0.039) (Table 5).

Together these results separate two distinct claims. The direction of AI–human disagreement is robust, modality-dependent, and substantial, particularly for oral video, where the perfect rank-biserial effect indicates the harshness is not noise but a consistent property of how the AI grades transcribed speech. Yet the predictability of that disagreement from observable assignment features is poor: the near-zero generalization of both models implies that surface characteristics such as response length, code complexity, or speaking tempo do not capture the mechanism driving the gap, which is more consistent with an exemplar-anchored harshness effect: the AI penalizes brevity relative to elaborated exemplar even when the rubric is satisfied. The weakness of SE as a disagreement flag is equally consequential: because SE reflects how consistently the model reproduces its own score rather than whether that score matches a human, it should not be treated as a proxy for grading validity outside programming, where the only modest SE signal appears. The practical implication reinforces the tiered deployment policy: programming scores can be auto-accepted, provisionally accept free response after diversifying exemplar prompts, and require mandatory human review on oral video assignments.

CONCLUSION

This paper presented a multimodal, uncertainty-aware, and transparent AI assisted AGT that grades programming, free-response, and oral assignments within a single rubric aligned pipeline, and evaluated it against averaged human ground truth across 95 assignments. Three findings stand out. First, AI-human agreement is strongly modality dependent: programming scores are statistically indistinguishable from human grading and can be

auto-accepted, whereas free-response should be provisionally accepted, and since oral assignments exhibit a large, consistent AI-stingy bias (rank-biserial $r = -1.00$) this warrants mandatory human review. Second, this disagreement is not predictable from surface features of the assignments, both the Ridge and logistic models failed to generalize, which is consistent with the proposed exemplar-anchored harshness mechanism, whereby the AI penalizes responses that are correct but less elaborated than the reference exemplar. Third, SE escalation provided no usable routing signal at this sample size, indicating that self consistency reflects how reliably the model reproduces its own score rather than whether that score is valid, and should not be used as a proxy for grading validity outside programming. Together these results argue for a tiered deployment policy in which modality, rather than automated confidence, determines the level of human oversight. The principal limitations are the modest sample size and the reliance on two human graders whose own agreement was only moderate (ICC = 0.649). Future work should test the exemplar-anchored harshness hypothesis directly through controlled rubric and exemplar manipulations, expand the dataset, and explore alternative uncertainty signals better aligned with human AI disagreement.

ACKNOWLEDGMENT

This work was supported by the PATH initiative through the Personal Robots Group at the MIT Media Lab. The author thanks Grace Lin and Kate Moore for their mentorship throughout this research.

REFERENCES

Batten, J. (2025, August 1). Infographic: Computing bachelor's enrollment continues to grow even as the field evolves. Computing Research News. https://cra.org/crn/2025/08/infographic-computing-bachelors-enrollment-continues-to-grow-even-as-the-field-evolves/

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.-S., Gašević, D., & Mello, R. F. (2021). Automatic feedback in online learning environments: A systematic literature review. Computers and Education: Artificial Intelligence, 2, Article 100027. https://doi.org/10.1016/j.caeai.2021.100027

Chen, H., & He, B. (2013). Automated essay scoring by maximizing human-machine agreement. In D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu, & S. Bethard (Eds.), Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (pp. 1741–1752). Association for Computational Linguistics. https://aclanthology.org/D13-1180/

Cisneros-González, J., Gordo-Herrera, N., Barcia-Santos, I., & Sánchez-Soriano, J. (2025). JorGPT: Instructor-aided grading of programming assignments with large language models (LLMs). Future Internet, 17(6), Article 265. https://doi.org/10.3390/fi17060265

CodeSignal. (n.d.). Exploring temperature sensitivity in LLM outputs. In Behavioral benchmarking of LLMs. CodeSignal Learn. Retrieved June 13, 2026, from https://codesignal.com/learn/courses/behavioral-benchmarking-of-llms/lessons/exploring-temperature-sensitivity-in-llm-outputs

Deepshikha, D. (2025). A comprehensive review of AI-powered grading and tailored feedback in universities. Discover Artificial Intelligence, 5, Article 251. https://doi.org/10.1007/s44163-025-00517-0

Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., & De Weerdt, J. (2021). A review of automated feedback systems for learners: Classification framework, challenges and opportunities. Computers & Education, 162, Article 104094. https://doi.org/10.1016/j.compedu.2020.104094

Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. Nature, 630(8017), 625–630. https://doi.org/10.1038/s41586-024-07421-0

Gao, X., Li, P., Shen, J., & Sun, H. (2020). Reviewing assessment of student learning in interdisciplinary STEM education. International Journal of STEM Education, 7, Article 24. https://doi.org/10.1186/s40594-020-00225-4

García-Varela, F., Nussbaum, M., Mendoza, M., et al. (2025). ChatGPT as a stable and fair tool for automated essay scoring. Education Sciences, 15(8), 946.https://www.mdpi.com/2227-7102/15/8/946.

Gonzalez, C., Donahue, K., Goldstein, D. G., Heidari, H., Jalali, M. S., Schelble, B., Singh, A., & Woolley, A. W. (2026). Toward a science of human–AI teaming for decision making: A complementarity framework. PNAS Nexus, 5(3), Article pgag030. https://doi.org/10.1093/pnasnexus/pgag030

Grandel, S., Schmidt, D. C., & Leach, K. (2024). Applying large language models to enhance the assessment of parallel functional programming assignments. In Proceedings of the 1st International Workshop on Large Language Models for Code (LLM4Code '24). Association for Computing Machinery. https://doi.org/10.1145/3643795.3648375

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. Educational Research Review, 2(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

Keuning, H., Jeuring, J., & Heeren, B. (2018). A systematic literature review of automated feedback generation for programming exercises. ACM Transactions on Computing Education, 19(1), Article 3. https://doi.org/10.1145/3231711

Lv, J., Chen, C., & Liang, Z. (2024). Automated scoring of asynchronous interview videos based on multi-modal window-consistency fusion. IEEE Transactions on Affective Computing, 15(3), 799–814. https://doi.org/10.1109/TAFFC.2023.3294335

Nguyen, H., & Park, S. (2025). Providing automated feedback on formative science assessments: Uses of multimodal large language models. In Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK '25) (pp. 803–809). Association for Computing Machinery. https://doi.org/10.1145/3706468.3706480

Patil, R., Kulkarni, A. A., Ghatage, R., Endait, S., Kale, G., & Joshi, R. (2024). Automated assessment of multimodal answer sheets in the STEM domain. arXiv. https://arxiv.org/abs/2409.15749

Quah, B., Zheng, L., Sng, T. J. H., Yong, C. W., & Islam, I. (2024). Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations. BMC Medical Education, 24, Article 962. https://doi.org/10.1186/s12909-024-05881-6

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. Assessment & Evaluation in Higher Education, 35(4), 435–448. https://doi.org/10.1080/02602930902862859

Ross, J., Curwood, J. S., & Bell, A. (2020). A multimodal assessment framework for higher education. E-Learning and Digital Media, 17(4), 290–306. https://doi.org/10.1177/2042753020927201

Seßler, K., Fürstenberg, M., Bühler, B., & Kasneci, E. (2025). Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK '25) (pp. 462–472). Association for Computing Machinery. https://doi.org/10.1145/3706468.3706527

Uppalapati, P. J., Dabbiru, M., & Kasukurthi, V. R. (2025). AI-driven mock interview assessment: Leveraging generative language models for automated evaluation. International Journal of Machine Learning and Cybernetics, 16, 10057–10079. https://doi.org/10.1007/s13042-025-02529-9

Zacharis, G., & Papadakis, S. (2025). Can AI grade like a human? Validity, reliability, and fairness in university coursework assessment. Educational Process: International Journal, 19, Article e2025591. https://doi.org/10.22521/edupij.2025.19.591

Zhang, C., Lu, T., Islam, M. M., Wang, Z., Yu, S., Bansal, M., & Bertasius, G. (2024). A simple LLM framework for long-range video question-answering. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 21715–21737). Association for Computational Linguistics. https://aclanthology.org/2024.emnlp-main.1209/