2026 ASEE ANNUAL CONFERENCE June 21–24, 2026 | Charlotte Convention Center, NC
Host Sponsor UNC Charlotte
Paper ID #52059

Assessing the Effectiveness and Efficiency of Pensieve Grader: An AI-Supported System for Grading Handwritten Work

Dr. Yan Tang, Embry-Riddle Aeronautical University - Daytona Beach

Yan Tang is an Associate Professor of Mechanical Engineering at Embry-Riddle Aeronautical University in Daytona Beach, Florida, where she also serves as the Undergraduate Program Coordinator for the Bachelor of Science in Mechanical Engineering (BSME) program. She received her Ph.D. in Mechanical Engineering from the University of Central Florida (UCF) in 2009 and her M.S. and B.S. degrees in Automatic Control Theory and Application from Nanjing University of Science and Technology (NJUST), China. Her research has evolved from control systems to engineering education, supported by the NSF-funded Research Initiation in Engineering Formation (RIEF) project, What Practice Makes Perfect. Her current work focuses on personalized instruction and AI-driven learning in engineering education, with an emphasis on developing intelligent systems that enhance student mastery and instructional effectiveness.

Madison Warner, Embry-Riddle Aeronautical University - Daytona Beach

Embry-Riddle Aeronautical University Student

Vasundhra Lathar, Embry-Riddle Aeronautical University

Vasundhra Lathar is an undergraduate student pursuing a B.S. in Aerospace Engineering with a concentration in Astronautics at Embry-Riddle Aeronautical University. Her interests include aerodynamics, fluid dynamics, and propulsion systems, as well as applied computational methods. She is also involved in research in engineering education and experimental aerodynamics, including wind tunnel testing to study the effects of wing geometry on aerodynamic performance. She plans to pursue graduate study in Aerospace Engineering and focus on spacecraft and hypersonic vehicle design

Assessing the Effectiveness and Efficiency of Pensive: An AI-Supported System for Grading Handwritten Work

Abstract

This study examines the effectiveness and efficiency of Pensive (formerly Pensieve Grader), an AI-assisted grading system that extends the capabilities of Gradescope for evaluating handwritten student work. Built on large language model (LLM) technology, Pensive automates the entire grading workflow—from transcription to rubric-based evaluation and feedback—within a human-in-the-loop interface. Although the system has been deployed at more than 20 institutions and has graded over 300,000 student responses across computer science, mathematics, physics, and chemistry, prior work shows that its adoption in physics-related fields remains limited (approximately 6.7%). This study investigates whether such limited use reflects concerns about effectiveness and efficiency or simply a lack of familiarity within subjects similar to physics.

The effectiveness of Pensive was evaluated by comparing its grading outcomes with instructor assessments across multiple problem types, including drawings (e.g., free-body diagrams (FBDs)), multiple-choice questions, and multi-step problem-solving tasks. Agreement metrics were analyzed to determine how well the system captured both handwriting recognition and grading accuracy. For diagrammatic and open-ended responses, qualitative rubrics were applied to assess Pensive's recognition of key features such as correct force representation and labeling. For quantitative problems, accuracy was examined through AI-extracted expressions and numerical results relative to instructor solutions. The study also explored how the worksheet outline, rubric design, problem structure, and calibration procedures influence grading performance.

The efficiency of Pensive was assessed in terms of grading time reduction and feedback turnaround compared to traditional manual grading. Building on the findings of Yang et al.[1], who reported an average 65% reduction in grading time across STEM disciplines, this study investigates whether similar levels of efficiency can be achieved in the context of engineering dynamics. The analysis focuses on how factors such as the worksheet outlines, rubric design, calibration, and problem format may influence the degree of time savings and usability for instructors.

Findings indicate that grading effectiveness and efficiency varied by task type. Pensive achieved high agreement with instructor grading on multiple choice and short answer questions, and reached approximately 98% agreement on analytical problems after calibration. Drawing tasks showed the greatest variability, with agreement differing substantially across structurally similar problems. These results suggest that AI-assisted grading shifts rather than uniformly reduces instructor effort, with benefits depending on task type and the reusability of calibration across assessments. By providing empirical data from handwritten dynamics exams, this work contributes to understanding how LLM-assisted grading performs in symbol-heavy, spatially complex problem types rarely represented in prior Pensive studies. The results will clarify whether barriers to broader adoption stem from technical limitations or awareness gaps, offering evidence-based insights for integrating AI-assisted grading tools into STEM education.

1. Introduction

Grading handwritten assessments in engineering dynamics is time-consuming, cognitively demanding, and difficult to scale, particularly in courses that emphasize symbolic reasoning and diagrammatic representations. Mechanics problems often require students to demonstrate problem interpretation, diagram construction, symbolic formulation, and structured equation setup, all of which demand nuanced human judgment during grading. As a result, grading consistency and turnaround time remain persistent challenges in undergraduate mechanics courses, especially as enrollments increase.

Digital grading platforms such as Gradescope have improved logistical efficiency by supporting scanned handwritten submissions and rubric-based grading workflows, enabling instructors to manage large volumes of student work more effectively [2]. However, these platforms still rely heavily on human graders to interpret student intent, evaluate symbolic reasoning, and assign partial credit, particularly for open-ended and diagrammatic responses common in engineering dynamics.

Recent advances in large language models (LLMs) have enabled AI-assisted grading systems that aim to automate portions of the grading workflow while maintaining instructor oversight. Pensive (Yang et al.[1]) is an LLM-powered grading platform that extends this workflow by automating the grading pipeline from scanned submission to final feedback within a human-in-the-loop interface. The system processes handwritten student work through sequential stages: transcription of scanned images into machine-readable text using OCR and language model post-processing, rubric-aligned evaluation of each transcribed response, and generation of per-response confidence levels—high, low, or error—that guide the degree of human review required. Instructors can construct rubrics within the platform using either additive or subtractive schemes, and a calibration process allows instructor corrections on a small number of responses to propagate as grading guidance, called grading wisdoms, across the full submission set. Deployed across more than 20 institutions and applied to over 300,000 student responses in computer science, mathematics, physics, and chemistry, Pensive has been reported to reduce grading time by an average of 65% without meaningful loss in accuracy [1].

Prior research has demonstrated the potential of AI-based tools to support grading of handwritten and open-ended responses in STEM disciplines, including mathematics and physics, with promising levels of accuracy and efficiency [3], [4]. Despite these results, adoption in physics-adjacent disciplines accounts for only approximately 6.7% of current Pensive usage [1]. Instructors therefore lack discipline-specific guidance on when such systems are effective, what types of mechanics problems can be reliably graded, and how instructor effort shifts when AI assistance is introduced.

Existing studies on AI-assisted grading, including Yang et al. [1], report aggregate efficiency metrics that treat grading time as uniform across problem types. In engineering mechanics, where assessment commonly involves diagrammatic representations, symbolic problem formulation, and structured equation setup, grading demands vary substantially across problem formats. The 65% average time reduction reported by Yang et al. [1] is derived from a model that omits preparation costs such as rubric design and calibration time, and does not account for task-level variation in AI performance. This study addresses these gaps by evaluating Pensive within the context of an undergraduate engineering dynamics course, focusing on grading effectiveness and efficiency across distinct problem types. Rather than benchmarking AI performance in isolation, the goal is to inform instructional decision-making by characterizing how AI-assisted grading behaves across the range of assessment formats typical in engineering mechanics.

2. Study Context and Grading Tasks

Course Context Pensive was implemented in two sections of ES204 Dynamics, a sophomore-level Engineering Dynamics course at Embry-Riddle Aeronautical University, Daytona Beach

Campus, an ABET-accredited institution during the Fall 2025 semester. The combined enrollment across the two sections was approximately 60 students.

Student learning was assessed using a combination of formative and summative assessments. Eight in-class formative assessments, each lasting approximately 20–30 minutes, were administered throughout the semester, followed by a comprehensive final exam serving as the summative assessment. To improve the validity and reliability of assessment, both formative and summative assessments emphasized problem interpretation and equation setup rather than numerical computation [5]. Most problems required students to formulate governing equations and representations while omitting final numerical calculations. Each assessment typically included two diagrammatic questions, such as free-body diagrams (FBDs) or mass–acceleration diagrams (MADs), along with five to six analytical problems focused on symbolic formulation and equation setup. Prior to the adoption of Pensive, all assessments were graded using Gradescope.

Grading Tasks Pensive supports five assignment formats: Homework, Exam, Programming Assignment, Online Assignment, and Bubble Sheet. This study focused exclusively on the Exam format, which aligned with the in-class formative and summative assessments used in the course. Within the Exam format, grading tasks examined in this study were selected to align with the question formats supported by Pensive and those used in the Engineering Dynamics course. The analysis focused on the following task categories:

- Single-Select Multiple Choice (SSMC) tasks, used for conceptual or intermediate checkpoints within assessments (see Figure 8 in Appendix).
- Short Answer tasks, used for simple answers to evaluate students' conceptual understanding (see Figure 9 in Appendix).
- Text/Code tasks, encompassing both problem formulation tasks and multi-step analytical solution setup. Problem formulation tasks required students to represent givens and unknowns symbolically using instructor-specified notation directly beneath the corresponding problem statements. Analytical tasks focused on deriving and organizing governing equations without completing numerical calculations (see Figure 10 in Appendix).
- Drawing tasks, including free-body diagrams (FBDs) and mass–acceleration diagrams (MADs) (see Figure 11 in Appendix), which require students to produce diagrammatic representations of forces and motion.

Multi-Select Multiple Choice (MSMC) questions were not used during the term and were therefore excluded from this study. Framing grading tasks according to Pensive's supported formats allows the evaluation to reflect realistic instructional use while enabling comparisons across tasks with differing representational and cognitive demands.

Pensive Integration and Grading Workflow After transitioning to Pensive, the grading workflow was modified to incorporate AI-assisted grading while preserving instructor oversight. The Pensive workflow consisted of four sequential stages: assignment setup, solution specification, rubric construction, and AI-assisted grading with human review. Student submissions could be uploaded at any time prior to grading and did not need to be combined into a single batch.

In the assignment setup stage, the instructor first uploaded the exam worksheet to Pensive, which auto-generated the assignment outline by identifying question boundaries from the worksheet.

Manual adjustment was occasionally required when the auto-generated outline did not correctly capture all questions or sub-parts.

In the solution specification stage, the instructor uploaded a PDF of the exam worksheet containing handwritten or typed solutions, which Pensive could convert into markdown automatically. If no solution was provided, Pensive generated one on its own. When the AI-generated solution did not align with instructor expectations, the instructor used the solution regeneration prompt to refine and regenerate it. Figure 1 and Figure 2 show the generated solution for the analytical problem in Figure 10 and the FBD/MAD problem in Figure 11, respectively.

P2b (1 pt) At the instant shown ($\theta = 80^\circ$), the 20-ft-long, uniform 100-lb pole ABC has an angular velocity of $1\ \mathrm{rad/s}$ counter-clockwise and point C is sliding to the right on a non-smooth surface. A 120-lb horizontal force P acts at B. Knowing μ_k between the pole and the ground is 0.3, determine the normal force between the pole and the ground at this instant. $I_G = \frac{mL^2}{12}$.

Givens: $W = 100$ lb, $m = \frac{W}{32.2}$, $L = 20$ ft, $\theta = 80^\circ$, $\mu_k = 0.3$

Implicit Givens:

$r_P = \frac{L}{2}\sin 80^\circ - 6\ \mathrm{ft}, r_N = \frac{L}{2}\cos\theta, r_f = \frac{L}{2}\sin\theta, \vec{a}_G = a_{Gx}\,\hat{\imath} + a_{Gy}\,\hat{\jmath}, \vec{a}_C = a_{Cx}\,\hat{\imath}, \vec{\alpha} = \alpha\,\hat{k},$
$\vec{r}_{G/C} = -\frac{L}{2}\left(\cos 80^\circ\ \hat{\imath} - \sin 80^\circ\ \hat{\jmath}\right)$

Find: N lb

Equations	Unknowns
$\sum F_x = ma_{Gx}\ (\rightarrow +)$: $P - \mu_k N = ma_{Gx}$	$N,\ a_{Gx}$
$\sum F_y = ma_{Gy}\ (\uparrow +)$: $N - W = ma_{Gy}$	a_{Gy}
$\sum M = I\alpha\ (\circlearrowleft +)$: $P \cdot r_P + N \cdot r_N - \mu_k N \cdot r_f = I\alpha$	α
Kinematics Equation: $\vec{a}_G = \vec{a}_C + \vec{\alpha} \times \vec{r}_{G/C} - \omega^2\,\vec{r}_{G/C}$	a_{Cx}

Figure 1 Generated Solution for the problem in Figure 10 (Appendix)

FBD and MAD Description

Setup: A 7.5-kg box slides down a circular chute of radius $r = 2$ m with kinetic friction coefficient $\mu_k = 0.3$. At point A (the lowest point of the chute), the box has speed $v = 2.5$ m/s. Determine the normal force and the rate of change of speed at A.

FBD Forces:

- Weight mg acting vertically downward at the box
- Normal force N acting on the box from the chute surface, directed radially inward toward the center of curvature (vertically upward at point A, in the $+\hat{e}_n$ direction)
- Kinetic friction force $\mu_k N$ acting tangent to the chute surface, opposing the motion of the box (directed in the $-\hat{e}_t$ direction, opposite to the direction of sliding)

MAD (Coordinate System and Inertia Vectors):

- Use normal-tangential coordinates: $\hat{e}_n$ pointing radially inward (toward the center of curvature, upward at A) and $\hat{e}_t$ pointing tangent to the path in the direction of motion
- Nonzero normal inertia term: ma_n acting in the $+\hat{e}_n$ direction (centripetal, toward center of curvature)
- Nonzero tangential inertia term: ma_t acting in the $\hat{e}_t$ direction

Figure 2 Generated Solution for the problem in Figure 11 (Appendix)

In the rubric construction stage, instructors created rubrics aligned with the reference solution. Rubrics could follow either an additive or subtractive scheme. In this study, the same instructor-defined rubrics used in prior Gradescope-based grading were applied within Pensive to isolate the effect of AI assistance from rubric differences. Figure 3 and Figure 4 illustrate representative rubrics used in Gradescope and Pensive, respectively. A recent update to Pensive allows instructors to test a constructed rubric on a sample of student responses, with the sample size determined by the instructor. If the AI-generated grading results are satisfactory, no further changes are needed. Otherwise, instructors review the notes accompanying each graded response to understand why specific rubric items were or were not selected, and then add calibration instructions to reduce disagreement between AI and instructor judgments. Figure 5 illustrates how grading for the analytical problem in Figure 10 changed after calibration instructions were added based on the AI-generated notes. Following calibration, the AI grader correctly selected the previously overlooked rubric item for the moment equation.

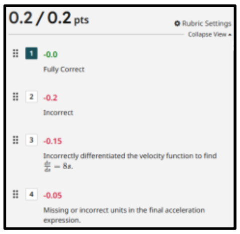


Figure 3 Rubrics in Gradescope

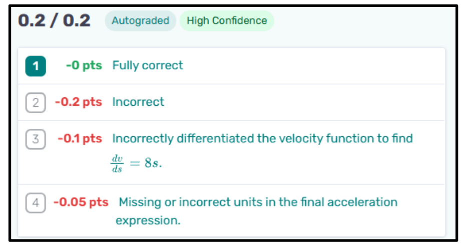


Figure 4 Rubric in Pensive

In the AI-assisted grading and review stage, Pensive generated rubric-aligned scores, confidence levels (high, medium, or low), and brief notes for the reviewer. Although Pensive supports an automatic review mode in which high-confidence AI grades are accepted without further inspection, occasional misgradings were observed even at high confidence. To preserve instructor oversight, automatic review was disabled in this study, and graders manually reviewed every AI-graded response to verify accuracy and ensure alignment with instructor expectations. This shift altered the role of graders from uniformly grading all student work to reviewing and validating AI-assisted grading outcomes. Graders typically skim all responses and intervene only on flagged or clearly incorrect cases, which is considerably faster than grading every problem in full. Table 1 summarizes key differences between the Gradescope-based and Pensive-based grading workflows.

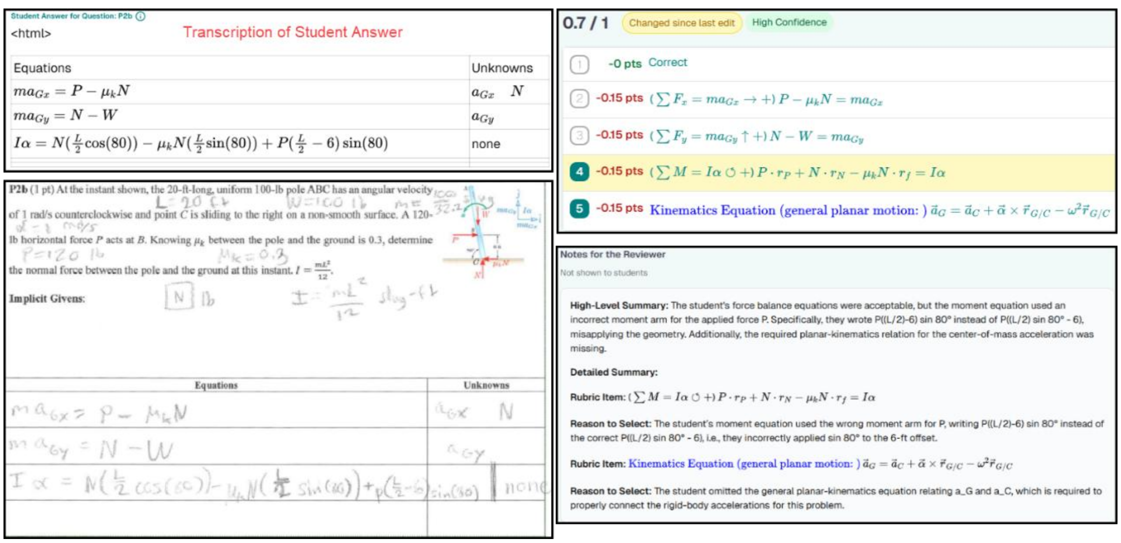


Figure 5 Example of Student Submission and Notes for the Reviewer

Table 1 Grading Workflow Comparison between Gradescope and Pensive

	Gradescope	Pensive
Submission	Scanned handwritten work	Same
Outline	Manual	AI-assisted auto-generation
Initial Grading	Graders grade with rubrics created by the instructor	AI-assisted grading with either rubrics created by the instructor or auto-generated rubrics
Review	None	Graders review all AI graded results

Although Pensive was deployed in the course during the Fall 2025 semester, the analyses reported in this paper were conducted using a more recent version of the platform applied retrospectively to the original student submissions. As an actively developed system, Pensive's underlying models and calibration mechanisms may differ from those in use during the original course delivery.

3. Effectiveness and Efficiency in AI-Assisted Grading

Evaluating AI-assisted grading systems requires constructs that extend beyond traditional measures of grading accuracy or total time saved. In the context of handwritten engineering dynamics assessments, both grading effectiveness and efficiency depend on the nature of the grading task, the role of instructor-defined rubrics, and how human effort is distributed across the grading workflow.

Effectiveness is defined as the extent to which Pensive accurately interprets and evaluates student work in alignment with instructor expectations, operationalized as the degree of agreement between AI-assigned and instructor-verified rubric item selections across different grading tasks. Because rubric item selection directly determines the score awarded, disagreement in rubric selection is equivalent to a grading error. Effectiveness is examined as a task-dependent construct across SSMC, Short Answer, Text/Code, and Drawing tasks, which place differing demands on transcription accuracy, symbolic interpretation, and rubric application. Instructor-defined rubrics serve as the reference standard for evaluation.

Efficiency is defined in terms of instructor and grader effort across the grading workflow rather than total grading time alone. Efficiency is examined across four phases: rubric design and preparation, calibration and tuning, grading execution, and review of AI-generated grades. This framing explicitly acknowledges that AI-assisted grading may shift or introduce effort rather than simply eliminate it, and that efficiency gains reported in aggregate models—such as the 65% reduction reported by Yang et al. [1]—may not transfer uniformly across problem types or to first-time adopters of the platform.

4. Evaluation Approach and Results

This study adopts an exploratory, comparative evaluation approach to examine the effectiveness and efficiency of Pensive for grading handwritten engineering dynamics assessments. Comparisons are made between traditional human-assisted grading using Gradescope and AI-assisted grading using Pensive applied to the same student submissions. The evaluation focuses on task-specific grading behavior and instructor workflow rather than benchmarking AI performance in isolation. Only initial assessment attempts were included in the analysis to ensure a consistent basis for comparison between grading platforms.

Data Collection and Comparison Baseline Data were collected from handwritten student submissions corresponding to the four grading task categories described in Section 2 and illustrated in the Appendix: Single-Select Multiple Choice (SSMC), Short Answer, Text/Code, and Drawing tasks. For each task category, a single representative problem was selected from course assessments to reflect typical grading demands encountered in the Engineering Dynamics course. Although only one problem per task category was included, each problem generated multiple student responses, allowing task-level grading behavior to be examined across the full class.

All grading was conducted within the Pensive platform using the same instructor-defined rubrics. Two grading modes were used: (1) manual grading, in which graders applied rubrics directly in Pensive without AI assistance (functionally equivalent to grading in Gradescope), and (2) AI-assisted grading, in which Pensive's AI grader generated rubric-aligned scores and confidence levels that were subsequently reviewed by human graders. Pensive-generated rubrics were not used in this study; instead, the same instructor-defined rubrics were applied consistently across both grading modes to isolate the impact of AI assistance rather than rubric differences.

The recorded data consisted of the AI grader's assigned score, associated confidence level, and whether human intervention was required during review. Individual student response logs and per-response manual grading records were not stored. Analysis therefore focused on aggregate grading behavior at the task level rather than on per-student comparisons.

AI Grader Calibration Calibration was conducted using Pensive's Calibrate AI Grader workflow and represents an instructor-driven alignment process rather than modification of the underlying AI model [6]. During calibration, instructors reviewed a small number of AI-graded responses and corrected rubric selections when necessary. These corrections informed assignment-level adjustments to how the AI grader interpreted rubric criteria, symbolic notation, and partial credit rules. Calibration was limited to rubric interpretation and did not involve changes to rubric structure or AI model parameters.

Calibration was applied selectively and primarily for task categories involving symbolic reasoning (Text/Code tasks). No calibration was applied to SSMC tasks, while calibration for Drawing tasks was limited in effectiveness due to the interpretive nature of diagrammatic representations.

Task-Level Grading Results The following paragraphs report grading results for each of the four task categories examined in this study. For each task, the reported observations include the number of student responses analyzed, AI grading agreement with instructor judgment, confidence-level patterns, primary error modes, and the role of calibration where applicable.

Single-Select Multiple Choice (SSMC) SSMC tasks consisted of conceptual checkpoint questions embedded within larger problems (e.g., P4g, P4h, P4i in Figure 8 in Appendix), with 50 student responses per question. Pensive completed grading for each 50-response set in under 30 seconds, and all AI-generated grades were assigned at high confidence. Across the three SSMC questions examined, two responses were misgraded for P4g, one for P4h, and two for P4i, corresponding to roughly 96–98% agreement with instructor grading. In every misgraded case, the AI grader recorded the response as "no selection" despite the transcription correctly identifying the student's selected answer, indicating a disconnect between transcription and rubric application rather than a recognition failure. This error mode was not persistent: SSMC questions administered in a subsequent Spring 2026 offering achieved 100% agreement, and the conditions under which the earlier misgrading occurred remain unclear. Because all AI grades were high confidence, instructors had limited signal to flag these cases proactively, suggesting that even highly accurate task categories may benefit from spot-checking when stakes are high. Overall, SSMC tasks produced the largest efficiency gain of any task type examined.

Short Answer Short Answer tasks asked students to provide a brief symbolic or numerical answer evaluating a single conceptual point (see Figure 9 in Appendix). Out of 55 student responses examined, Pensive achieved 100% agreement with instructor grading. Three of the 55 responses (approximately 5%) were assigned low confidence due to low transcription confidence rather than grading uncertainty. As shown in Figure 6, these low-confidence cases typically involved heavily edited handwriting—such as crossed-out intermediate work preceding the final answer—where the transcription system flagged ambiguity even though the final boxed or rightmost answer was correct and correctly graded. Because the AI grader still selected the correct rubric item in all cases, the low-confidence flag functioned as a useful prompt for human review without indicating an actual grading error. Calibration was not required for this task category. Short Answer tasks therefore produced high effectiveness with modest efficiency cost: graders verified the three flagged responses but otherwise accepted AI grades without intervention.

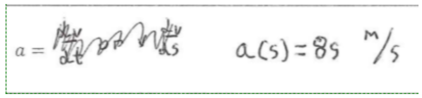


Figure 6 Example of Low Transcription Confidence for Short Answer Question in Figure 9 (Appendix)

Text/Code Text/Code tasks required students to formulate and organize governing equations for multi-step analytical problems, such as deriving force balance, moment, and kinematics equations for a rigid body in planar motion (see Figure 10 in Appendix). A total of 51 student responses were examined, and grading was conducted in two passes to evaluate the impact of calibration. In the initial pass without calibration, Pensive flagged only 6 responses as high confidence—5 of which were skipped submissions correctly identified as blank—while the remaining 45 responses were assigned low confidence. Six of the 45 low-confidence responses were misgraded, corresponding to approximately 88% overall agreement. After calibration instructions were added based on the AI-generated grading notes (as illustrated in Figure 7), grading was repeated on the same 51 responses. High-confidence coverage increased substantially to 37 responses, all of which were graded correctly, and only 1 of the remaining 14 low-confidence responses was misgraded, raising overall agreement to approximately 98%. Calibration primarily addressed how the AI grader recognized the moment equation and kinematics equation, reducing missed rubric item selections that had occurred in the uncalibrated pass. These results indicate that calibration meaningfully improved both grading accuracy and the proportion of responses Pensive could grade with high confidence, narrowing the scope of required human review without introducing new error modes.

-0.15 pts $(\sum F_x = ma_{Gx} \rightarrow +)\, P - \mu_k N = ma_{Gx}$

-0.15 pts $(\sum F_y = ma_{Gy} \uparrow +)\, N - W = ma_{Gy}$

-0.15 pts $(\sum M = I\alpha \circlearrowleft +)\, P \cdot r_P + N \cdot r_N - \mu_k N \cdot r_f = I\alpha$

- Accept equivalent notation for the moment arms as long as they are consistently used
- Select this item if the moment of each force is incorrect.
- Select this item if the moment arm of P is not $\frac{L}{2}\sin 80^\circ-6$ ft

-0.15 pts Kinematics Equation (general planar motion:) $\vec{a}_G = \vec{a}_C + \vec{\alpha} \times \vec{r}_{G/C} - \omega^2 \vec{r}_{G/C}$

- Select this item if \vec{\alpha}\times\vec{r}_{G/C} is set to zero
- Select this item if each vector is not correctly defined as implicit givens.
- kinematics relation should relate points G and C; No point other than C is acceptable.

Figure 7 Calibrated Rubric for the Problem in Figure 10 (Appendix)

Drawing (FBD/MAD) Drawing tasks required students to construct free-body and mass–acceleration diagrams for two distinct problems: P4a, involving a box sliding along a curved chute, and P4b, involving a ladder released from rest against a wall (see Figure 11 in Appendix). Each problem received 49 student responses. Grading effectiveness differed substantially between the two problems despite their nominally similar task type. For P4a, only 2 responses were assigned high confidence and were graded correctly; among the remaining 47 low-confidence responses, 39 were misgraded, corresponding to approximately 20% overall agreement. For P4b, 7 responses were assigned high confidence and 42 were assigned low confidence, with no misgraded responses in either group, yielding 100% overall agreement. The specific factors driving this contrast were not clearly identifiable from the AI grading notes, and the divergence between the two problems suggests that AI grading effectiveness for diagrammatic tasks may be sensitive to problem-specific features that are not yet well understood. Calibration was applied but produced only marginal improvement, consistent with the interpretive nature of diagrammatic representations and the difficulty of encoding spatial reasoning rules through rubric clarification. As a result, human review remained extensive for Drawing tasks, particularly for P4a, where graders effectively re-graded the majority of responses despite AI assistance.

Table 2 summarizes grading effectiveness, review requirements, and tuning actions across grading task categories.

Table 2 Summary of grading effectiveness, efficiency, and calibration by task type.

Task Type	Agreement Level	Review Required	Primary Error Modes	Calibration Applied	Impact of Calibration
SSMC	High	Low	Ambiguous markings	None	Not applicable
Short Answer	High–Moderate	Low–Moderate	Transcription, phrasing	Minimal rubric clarification	Reduced review
Text/Code – Problem Formulation	Moderate	Moderate	Symbolic notation mismatch	Calibrated notation interpretation	Improved consistency
Text/Code – Analytical Setup	Moderate–High	Moderate	Partial credit assignment	Refined rubric interpretation	Reduced discrepancies
Drawing (FBD/MAD)	Low	High	Diagram interpretation	Limited calibration effectiveness	Marginal improvement

Grading effectiveness varied substantially by task category. SSMC tasks exhibited high agreement with minimal review requirements. Short Answer tasks showed slightly lower agreement, primarily due to transcription challenges rather than conceptual errors. Text/Code tasks demonstrated moderate effectiveness, with calibration improving consistency for symbolic notation and partial credit assignment. Drawing tasks exhibited the lowest agreement due to challenges in interpreting spatial representations, requiring substantial human judgment despite AI assistance.

Efficiency was evaluated qualitatively by examining how grader effort shifted between manual grading and AI-assisted grading across the four phases defined in Section 3. AI assistance did not uniformly reduce instructor effort; rather, it redistributed effort across the workflow. Calibration emerged as the most significant source of added effort, particularly for Text/Code tasks, where instructors invested time reviewing AI grading notes, identifying systematic disagreements, and adding grading wisdoms to refine rubric interpretation. Once calibration was completed, however, the grading and review phases were substantially faster than manual grading. For example, SSMC tasks required no calibration and completed grading for 50 responses in under 30 seconds with high confidence across all responses. For Text/Code tasks, calibration increased the proportion of high-confidence responses from 6 of 51 to 37 of 51, allowing graders to verify most responses quickly and concentrate review effort on the smaller low-confidence subset. For tasks with persistently low effectiveness, particularly P4a in the Drawing category, human review remained extensive regardless of AI assistance, and calibration did not meaningfully reduce grader workload.

Importantly, calibration functions as a one-time upfront investment rather than a recurring cost. In engineering dynamics, where assessments often involve recurring problem types and a relatively stable set of rubric items (e.g., force balance equations, moment equations, kinematics relations), calibration effort developed for one assessment can carry over to subsequent assessments featuring similar problems. New calibration is generally only required when problem structure or rubric items differ substantially from prior work. This characteristic shifts the efficiency calculation: while initial adoption of AI-assisted grading involves nontrivial calibration overhead, instructors who teach the same course across multiple semesters or who build assessment banks can expect calibration effort to decline sharply after the first deployment. Overall, AI assistance produced clear efficiency gains for structured and symbol-based tasks once initial calibration was completed, while diagrammatic tasks yielded limited efficiency benefit even after calibration. This pattern suggests that the efficiency of AI-assisted grading is best characterized not as a single time-savings figure but as a trade-off between upfront calibration effort and downstream grading speed, with the favorability of that trade-off depending on both task type and the degree to which calibration can be reused across assessments.

5. Discussion

This study examined AI-assisted grading through the lens of instructional practice in an engineering dynamics course, with particular attention to how grading effectiveness and efficiency vary across problem types. Rather than treating AI grading as a uniform replacement for human grading, the results demonstrate that its value depends strongly on task structure, representational demands, and rubric alignment.

Across grading tasks, Pensive was most effective and efficient for structured tasks with constrained response formats, such as Single-Select Multiple Choice and Short Answer questions. For these tasks, AI assistance reduced repetitive grading actions and allowed graders to shift their effort toward selective validation guided by confidence levels. Text/Code tasks, particularly those involving symbolic problem formulation and analytical setup, showed moderate effectiveness that was sensitive to rubric clarity and notation conventions. Instructor-driven calibration played an important role in improving grading consistency for these tasks, suggesting that AI-assisted grading is especially responsive to alignment between instructional intent and rubric design.

In contrast, Drawing tasks such as free-body diagrams and mass–acceleration diagrams required substantial human judgment despite AI assistance. While Pensive supported rubric organization and error identification, diagram interpretation remained challenging due to spatial ambiguity and varied student representations. These findings are consistent with prior work in physics and mechanics education indicating that diagrammatic reasoning poses persistent challenges for automated assessment tools.

From an instructional perspective, these results suggest that AI-assisted grading tools are best understood as workflow amplifiers rather than graders in isolation. Their effectiveness and efficiency emerge from how instructors design tasks, structure rubrics, and integrate calibration into grading practice. When used thoughtfully, AI-assisted grading can reduce grading burden for symbol-heavy tasks while preserving instructor control over evaluative judgment.

6. Limitations and Future Work

This study has several limitations. First, the evaluation was conducted within a single course context, which may limit generalizability to other institutions or instructional settings. Second, only one representative problem per grading task category was included. Although each problem generated multiple student responses, future studies should examine a broader range of problems to further validate task-level patterns.

Third, effectiveness and efficiency were evaluated using aggregate grading behavior and workflow observations rather than fine-grained time-on-task measurements or per-student response logs. While this approach aligns with instructional practice, future work could incorporate automated time tracking and more detailed logging to strengthen quantitative analysis. Finally, this study focused on grading behavior and workflow rather than on student learning outcomes or feedback quality.

Future work will expand data collection across multiple offerings of engineering dynamics and related courses, include additional instructors to examine variability in calibration practices, and investigate how AI-assisted grading influences feedback timeliness and student learning. Further exploration of hybrid workflows for diagrammatic tasks may also help identify ways to better support grading of spatial representations.

7. Conclusion

This study evaluated the effectiveness and efficiency of Pensive for grading handwritten engineering dynamics assessments across multiple problem types. By examining AI-assisted grading within a realistic instructional workflow, the study provides discipline-specific insights into how AI tools interact with common mechanics grading tasks.

The results demonstrate that AI-assisted grading effectiveness and efficiency are task dependent. Structured and symbol-based tasks benefited most from AI assistance, particularly when combined with instructor-driven calibration and well-aligned rubrics. In contrast, diagrammatic tasks required substantial human oversight, limiting efficiency gains despite partial AI support. These findings highlight that AI-assisted grading tools do not offer uniform benefits across all assessment formats but instead shift instructor effort in task-specific ways.

Overall, this work contributes practical evidence to inform instructional decision-making around AI-assisted grading in engineering mechanics. By clarifying where AI tools are most effective, where human judgment remains essential, and how calibration supports alignment with instructional intent, this study offers a grounded framework for integrating AI-assisted grading into STEM education.

Appendix

Below are examples of different question types.

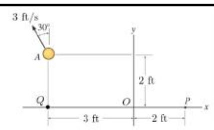
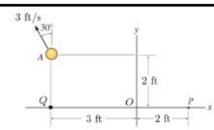
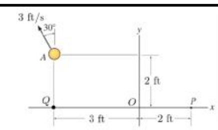
		
4g (0.1 pts) Select the direction of the angular momentum of the particle about point O. [ ] (Write A or B)	4h (0.1 pts) Select the direction of the angular momentum of the particle about point P. [ ] (Write A or B)	4i (0.1 pts) Select the direction of the angular momentum of the particle about point Q. [ ] (Write A or B)
A Clockwise B Counterclockwise	A Clockwise B Counterclockwise	A Clockwise B Counterclockwise

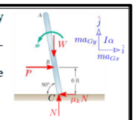
Figure 8 Example of SSMC (Single-Select Multiple Choice) Tasks

4. A point is moving along a straight line and its velocity is given by $v = 4 + 4s^2$ m/s. Represent the point's acceleration.

$a =$

Figure 9 Example of Short Answer Tasks

P2b (1 pt) At the instant shown, the 20-ft-long, uniform 100-lb pole ABC has an angular velocity of 1 rad/s counterclockwise and point C is sliding to the right on a non-smooth surface. A 120-lb horizontal force P acts at B. Knowing μ_k between the pole and the ground is 0.3, determine the normal force between the pole and the ground at this instant. $I = \frac{mL^2}{12}$.



Implicit Givens:

Equations	Unknowns

Figure 10 Example of Text/Code Tasks

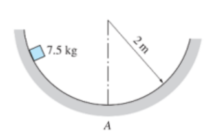
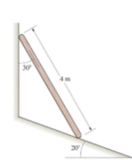
P4a (1 pt) The 7.5-kg box that is sliding down a circular chute ($\mu_k = 0.3$) reaches point A with a speed of 2.5 m/s. When the box is at A, calculate the normal force acting between it and the chute and its rate of change of speed.		**P4b** (1 pt) The 18-kg ladder is released from rest in the position shown. Model it as a slender bar and neglect friction. Determine its angular acceleration at the instant of release.	

Figure 11 Example of Drawing Tasks

References

[1] Y. Yang *et al*, "Pensieve Grader: An AI-Powered, Ready-to-Use Platform for Effortless Handwritten STEM Grading," 2025. . DOI: 10.48550/arxiv.2507.01431.

[2] A. Singh *et al*, "Gradescope: A fast, flexible, and fair system for scalable assessment of handwritten work," in *Proceedings of the Fourth (2017) Acm Conference on Learning@ Scale,* 2017, .

[3] G. Kortemeyer, J. Nöhl and D. Onishchuk, "Grading assistance for a handwritten thermodynamics exam using artificial intelligence: An exploratory study," *Physical Review Physics Education Research,* vol. 20, *(2),* pp. 020144, 2024. .

[4] G. Kortemeyer and J. Nöhl, "Assessing confidence in AI-assisted grading of physics exams through psychometrics: An exploratory study," *Physical Review Physics Education Research,* vol. 21, *(1),* pp. 010136, 2025. .

[5] Y. Tang *et al*, "Work in progress: Evaluating the effect of symbolic problem solving on testing validity and reliability," in *ASEE Annual Conference Exposition Proceedings,* 2023, .

[6] Anonymous . *Calibrate AI grader*. Available: https://help.pensieve.co/intercom-help-center/en/articles/11455901-calibrate-ai-grader.