*Corresponding Author's Email: gcling@hkbu.edu.hk
Proceedings of the World Conference on Research in Teaching and Education
Vol. 5, Issue. 1, 2026, pp. 19-37
DOI: https://doi.org/10.33422/worldte.v5i1.1999

ISSN: 2783-7750 online

Diamond Scientific Publishing

# Comparative Effectiveness of AI Tools in Enhancing Feedback Quality and Efficiency in Undergraduate Mathematics Assessments

**Shun Ling Chiang**
Department of Mathematics, Hong Kong Baptist University, Hong Kong

## Abstract

High-quality, timely feedback is essential for students' learning in mathematics; yet large class sizes and heavy workloads for teachers often result in students receiving brief, generic comments that fail to address misconceptions effectively. While artificial intelligence tools show promise for improving grading efficiency and feedback quality in STEM education, empirical evidence specific to undergraduate mathematics remains limited. This study conducted a rigorous comparative evaluation of two AI platforms — Gradescope (AI answer grouping and batch feedback) and StarGrader (generative AI with custom prompts for personalised feedback) — against traditional manual grading. The study analysed 6,922 student-question responses from three core undergraduate courses in Multivariable Calculus, Linear Algebra, and Differential Equations. Results demonstrated exceptionally high inter-rater reliability. Gradescope achieved 99.91% exact agreement with manual grading and 99.95% of scores within ±5 marks, while StarGrader attained 95.57% exact agreement and 96.44% practical reliability. Time savings were modest (average for Gradescope 22.2% and StarGrader 19.8%), but reached 30% for another course in Mathematical Analysis, and up to 35.1% in computation-heavy worksheets with refined prompts. Prompt engineering turned out to be critical especially for StarGrader, with accuracy improving from 43.75% to 94.17% through systematic refinement. Both student surveys and instructor ratings indicated more positive perceptions of AI-generated feedback regarding clarity, personalisation, and usefulness. The findings indicate that AI tools, when supported by effective prompt strategies and human oversight, can substantially enhance feedback quality and efficiency in mathematics assessments. Two evidence-based user manuals were developed to support wider adoption of the tools.

**Keywords:** AI grading, Mathematics Education, Hand-written Mathematics, Higher Education, Teaching Efficiency

# 1. Introduction

High-quality, timely, and personalised feedback is widely recognised as one of the most powerful influences on student learning, particularly in mathematics where conceptual understanding and problem-solving skills are developed through iterative practice and reflection (Hattie & Timperley, 2007). Effective feedback enables students to identify misconceptions, refine their reasoning, and build confidence in applying mathematical concepts. Despite its critical importance, providing detailed and individualised feedback remains one of the most persistent challenges in undergraduate mathematics education, especially in large-enrolment courses.

At Hong Kong Baptist University, students in the Department of Mathematics have repeatedly voiced concerns during Student-Staff Consultation Meetings about the quality of feedback on written assignments. They frequently describe the feedback as brief, generic, and insufficient for clarifying errors or misconceptions. Teaching assistants (TAs), who are primarily responsible for grading large volumes of handwritten submissions under tight deadlines, often prioritise completion speed over depth. This results in repetitive, surface-level comments that limit students' ability to learn from their mistakes and improve future performance.

Traditional manual grading, while pedagogically sound, is inherently labour-intensive and difficult to scale. In large classes, the time required for thorough marking frequently leads to trade-offs between speed and quality, often at the expense of feedback depth. Recent advances in artificial intelligence offer promising solutions to this long-standing problem. Tools such as Gradescope by Turnitin leverage computer vision for handwritten mathematical recognition and AI-assisted answer grouping, while generative AI systems like StarGrader can produce detailed, rubric-aligned individualised feedback at scale.

Although studies have begun to demonstrate the potential of these technologies in STEM education (Singh et al., 2017; Hansel et al., 2024), empirical evidence specific to undergraduate mathematics — particularly involving symbolic notation, proofs, and partial-credit reasoning — remains limited. Most existing research has focused on computer science or general STEM disciplines rather than the unique demands of core mathematics courses. There is also a notable lack of direct comparative studies between different AI approaches (answer-grouping systems versus generative AI systems) within the same mathematics context.

Mathematics presents unique difficulties for AI-assisted feedback compared to other disciplines. Symbolic notations, logical proofs, partial-credit reasoning, and handwritten expressions require sophisticated interpretation that goes beyond simple text or numerical matching. These characteristics make automated grading and feedback generation more complex than in fields dominated by multiple-choice or short-answer formats.

This study was designed to address these challenges. The primary objectives were to:

(1) evaluate the effectiveness of two AI tools — Gradescope (AI answer grouping and batch feedback) and StarGrader (generative AI with custom prompts for personalised feedback) — against traditional manual grading;

(2) assess their impact on grading accuracy, inter-rater reliability, time efficiency, and feedback quality; and

(3) examine student perceptions of the different feedback types.

The study examined 6,922 student-question responses collected from three core undergraduate courses — MATH2205 Multivariable Calculus, MATH2207 Linear Algebra I,

and MATH3405 Differential Equations I — over two academic years. Quantitative metrics focused on grading accuracy, inter-rater reliability, time efficiency, and feedback quality, while student perceptions were gathered through anonymous surveys.

The project directly supports the University's Institutional Strategic Plan (ISP 2018–2028), which emphasises innovative teaching practices and the integration of technology to enhance learning outcomes. It also contributes to the development of HKBU Graduate Attributes, particularly critical thinking, problem-solving, and effective communication of mathematical ideas. By systematically evaluating the strengths, limitations, and practical implications of AI-assisted grading in real mathematics classrooms, this research aims to provide evidence-based guidance for the Department of Mathematics and the wider university community on the sustainable adoption of AI tools in assessment and feedback practices.

The remainder of this paper presents the literature review, detailed methodology, key findings, discussion of implications, limitations, and practical recommendations for institutional implementation.

## 2. Literature Review

The integration of artificial intelligence into educational assessment has accelerated significantly over the past decade. Early automated grading systems were primarily restricted to objective or short-answer questions. However, advances in computer vision and large language models have enabled AI to process complex handwritten mathematical work, opening new possibilities for scalable and high-quality feedback in undergraduate mathematics courses (Singh et al., 2017).

One of the most widely adopted AI-supported grading platforms is Gradescope by Turnitin. Singh et al. (2017) introduced Gradescope as a fast, flexible, and fair system designed specifically for the scalable assessment of handwritten assignments and examinations. The platform uses computer vision to recognise handwritten mathematical expressions and automatically groups similar student answers, allowing instructors to provide batch feedback efficiently. Subsequent empirical research has confirmed these benefits in real classroom settings. Hansel et al. (2024) conducted a multi-case study at Indiana University involving large lecture courses and found that Gradescope helped instructors overcome key challenges, including high grading volume, timely feedback delivery, and grading consistency. Instructors reported significant efficiency gains, and course data showed reduced DFW rates alongside improved average grades.

While Gradescope excels at answer grouping and batch feedback, generative AI tools have recently emerged as a complementary approach capable of producing detailed, personalised explanatory feedback. Lee and Moore (2024) conducted a systematic review of ten empirical studies on the use of generative AI for automated feedback in higher education. Their synthesis revealed that large language model-based systems can deliver timely and comprehensive feedback at scale, often matching or exceeding the depth of traditional human feedback in certain contexts. However, the authors emphasised that the accuracy and pedagogical quality of generative AI feedback are highly dependent on careful prompt engineering and human oversight.

Recent studies have also begun to explore the application of AI in mathematics-specific contexts. For example, research on automated feedback for open-ended math responses has shown promising results using multimodal models that process both text and images (Li et al., 2024). Another line of work has examined the use of large language models for providing step-by-step feedback on mathematical reasoning tasks, highlighting the importance of chain-

of-thought prompting (Wei et al., 2022). These studies suggest that while AI can achieve high accuracy on routine computational problems, its performance on open-ended mathematical reasoning tasks remains variable and heavily reliant on well-structured rubrics and prompts.

Despite these promising developments, the existing literature reveals important limitations when applied to undergraduate mathematics education. Most studies have focused on computer science, general STEM disciplines, or objective assessment formats rather than the unique demands of symbolic mathematics, proofs, and partial-credit reasoning that characterise core mathematics courses (Zhao et al., 2025; Morris et al., 2025). While recent work has explored automated scoring of constructed-response items in mathematics using large language models, challenges remain in reliably handling open-ended proofs, logical reasoning, and partial credit allocation (Morris et al., 2025). Furthermore, few studies have conducted direct comparisons between different types of AI tools (such as answer-grouping systems versus generative AI systems) within the same mathematics courses or examined long-term impacts on student learning outcomes and instructor workload. There is also limited research addressing institutional barriers to widespread adoption. The majority of published work tends to focus on technical performance rather than pedagogical effectiveness or student perceptions of the feedback received (Meylani, 2025; Lindsay et al., 2025).

This Teaching Development Grant (TDG) project addresses these gaps by providing a direct, head-to-head comparison of two distinct AI approaches — Gradescope (AI answer grouping and batch feedback) and StarGrader (generative AI with custom prompts for personalised feedback) — against traditional manual grading. The study analysed 6,922 student-question responses from three core undergraduate mathematics courses: MATH2205 Multivariable Calculus, MATH2207 Linear Algebra I, and MATH3405 Differential Equations I. By combining quantitative metrics of grading accuracy, inter-rater reliability, and time efficiency with student perception data collected through anonymous surveys, this research offers new empirical evidence on the practical effectiveness of AI tools in mathematics assessment. It also contributes actionable insights for future implementation in higher education, particularly in contexts where handwritten symbolic work remains the dominant assessment format.

## 3. Methodology

This study adopted a mixed-methods comparative design to evaluate the effectiveness of two AI-assisted grading tools against traditional manual grading in undergraduate mathematics assessments.

### 3.1 Research Context and Participants

The research was conducted in the Department of Mathematics at Hong Kong Baptist University and focused on three core courses: MATH2205 Multivariable Calculus, MATH2207 Linear Algebra I, and MATH3405 Differential Equations I. These courses were selected because they represent a range of mathematical topics involving both computational problems and proof-based reasoning, and they typically enrol large numbers of students. The primary dataset consisted of 6,922 student-question responses collected across multiple worksheets and in-class exercises during Academic Years 2024/25 and 2025/26.

### 3.2 Grading Procedures

To enable a fair comparison, the same student submissions were graded using three methods in randomised order. Randomisation was performed manually using a six-sided die, with each of the six possible orders of the three methods (Manual, Gradescope, and StarGrader)

assigned a number from 1 to 6. This approach ensured that no single grading method consistently preceded or followed another.

The three grading methods were as follows:

(1) Traditional manual grading served as the ground-truth reference. Experienced graders (including the principal investigator and research assistant) followed detailed solution keys and marking schemes, assigning scores and providing written feedback according to standard departmental practices.

(2) Gradescope by Turnitin was used for AI-assisted answer grouping and batch feedback. Student PDFs were uploaded, answer regions were defined, and the system automatically grouped similar handwritten answers using its built-in AI. Graders then applied rubrics and comments to entire groups rather than to individual submissions, significantly streamlining the process.

(3) StarGrader (a generative AI grading assistant) was employed for fully automated, individualised grading and feedback generation. Custom prompts were developed and iteratively refined for each worksheet. These prompts explicitly included full marking schemes, partial-credit rules, handling of common edge cases (e.g., missing domain, correct steps but wrong final answer), and instructions on feedback tone and style.

For StarGrader, a systematic prompt-engineering process was implemented. The most extensive experimentation and iterative refinement occurred in MATH2205, where multiple prompt versions were tested for each worksheet and accuracy improvements were systematically tracked against manual grading. Prompt refinements focused on increasing specificity, embedding solution-key instructions directly, and clarifying partial-credit policies. For every worksheet, both the initial generation time and subsequent manual adjustment time were recorded. All time data were self-reported by the graders and included the time spent on prompt refinement.

In addition to the main dataset graded by the research assistant, the principal investigator personally implemented Gradescope for all in-class exercises in MATH2215 Mathematical Analysis during AY 2025/26 Semester 2. This allowed direct comparison of grading time with previous semesters when the same instructor graded identical exercises manually. Student perceptions of the Gradescope feedback in this real classroom setting were also collected through a separate anonymous survey.

### 3.3 Data Analysis

All scores were standardised to a 100-mark scale for direct comparison. Quantitative metrics calculated included exact agreement (percentage of cases where all three methods produced identical scores), Mean Absolute Error (MAE), Pearson correlation with manual grading, practical inter-rater reliability (percentage of AI scores within ±5 marks of manual scores).

To provide a more rigorous assessment of inter-rater reliability, the Intraclass Correlation Coefficient (ICC) was calculated. Specifically, ICC(2,1) together with its 95% confidence interval was computed to evaluate the absolute agreement between manual grading (treated as the reference) and each of the two AI-assisted grading methods. ICC(2,1) was selected because it assumes that both raters and subjects are random samples, which aligns with the study design in which the three grading methods were applied to the same set of student work.

Time savings were calculated by comparing manual grading time with the combined time for AI-assisted grading and manual adjustment. It should be noted that the reported time savings

are likely to increase as graders gain experience and establish libraries of well-performing prompts, thereby reducing the time required for prompt refinement in future applications.

### 3.4 Student Perceptions

In addition to the quantitative grading data, student perceptions were gathered through two anonymous surveys. The first was a comparative survey that presented students with identical work samples graded by all three methods. Participants rated how much they liked each feedback type on a 10-point scale and selected all feedback types that applied to 11 statements concerning clarity, helpfulness, personalisation, and usefulness (n = 15 complete responses). The second survey focused specifically on Gradescope group feedback in MATH2215 (Semester 2, AY 2025/26) and used a multi-select format to capture student experience with in-class exercises.

### 3.5 Practical Deliverables

As part of this project's commitment to translating research findings into practical teaching resources, two comprehensive user manuals were developed alongside the empirical analysis. The Gradescope User Manual provides detailed guidance on assignment creation, optimal answer box design for improved AI clustering accuracy, rubric development, and efficient batch grading workflows. The StarGrader User Manual focuses on systematic prompt engineering, prompt library construction, interpretation and refinement of AI-generated feedback, and best practices for hybrid human-AI grading. Summaries of these user manuals are presented in Appendix A.1 and A.2.

Both manuals incorporate evidence-based recommendations drawn directly from the analysis of 6,922 student-question responses and iterative prompt experiments, including a dedicated section on "Effective Design of Answer Boxes for AI Grading" (Appendix A.3). These deliverables were created to support immediate adoption by lecturers and teaching assistants in the Department of Mathematics and to facilitate wider implementation of the studied AI tools.

### 3.6 Ethical Considerations

All data collection and analysis were conducted in accordance with HKBU's research ethics guidelines. Because the project involved the collection of student feedback data, ethics/safety clearance was obtained from the Human (non-clinical) Research Ethics Panel as required by the TDG funding scheme. Student survey responses were fully anonymous, and no identifiable information was collected.

This methodology allowed for a direct, controlled comparison of the three grading approaches on the same student work while capturing both objective performance metrics and subjective student perceptions. The combination of large-scale quantitative data and targeted qualitative feedback provides a comprehensive basis for evaluating the practical effectiveness of AI tools in mathematics assessment.

This section presents the key quantitative and qualitative findings from the analysis of 6,922 student-question responses across MATH2205 Multivariable Calculus, MATH2207 Linear Algebra I, and MATH3405 Differential Equations I. The same student submissions were graded independently using three methods: traditional manual grading, Gradescope (AI answer grouping + batch feedback), and StarGrader (generative AI with custom prompts)

### 4.1 Overall Performance

Both AI tools demonstrated exceptionally high agreement with manual grading, indicating strong reliability for practical use in mathematics assessment. As shown in Tab. 4.1, Gradescope and StarGrader achieved high levels of exact agreement and practical inter-rater reliability with manual grading. Mean Absolute Error was low, and both Pearson correlations and inter-rater reliability (assessed using the Intraclass Correlation Coefficient) were near-perfect for both tools.

*Table 1: Overall Performance Metrics of Manual, Gradescope, and StarGrader Grading*

| Metric | Gradescope | StarGrader | Both |
|---|---|---|---|
| Exact Agreement with Manual | 99.91% | 95.57% | 95.57% |
| Practical Inter-rater Reliability (±5 marks) | 99.95% | 96.44% | — |
| Mean Absolute Error (MAE) | 0.007 | 0.671 | — |
| Pearson Correlation with Manual | 0.99997 | 0.98996 | — |
| Intraclass Correlation Coefficient, ICC(2,1) | 0.999972 | 0.989949 | — |
| 95% CI of ICC(2,1) | [0.999969, 0.999974] | [0.989075, 0.990753] | — |
| Average Time Savings | +22.2% | +19.8% | — |

*Source: Author's own analysis.*

These metrics suggest that both AI systems can produce scores highly consistent with experienced human graders. The scatter plots (Fig. 4.1 and Fig. 4.2) further confirm the strong linear relationship between manual and AI-generated scores, with very few outliers. On average, Gradescope provided a 22.2% time saving and StarGrader a 19.8% time saving compared with manual grading.

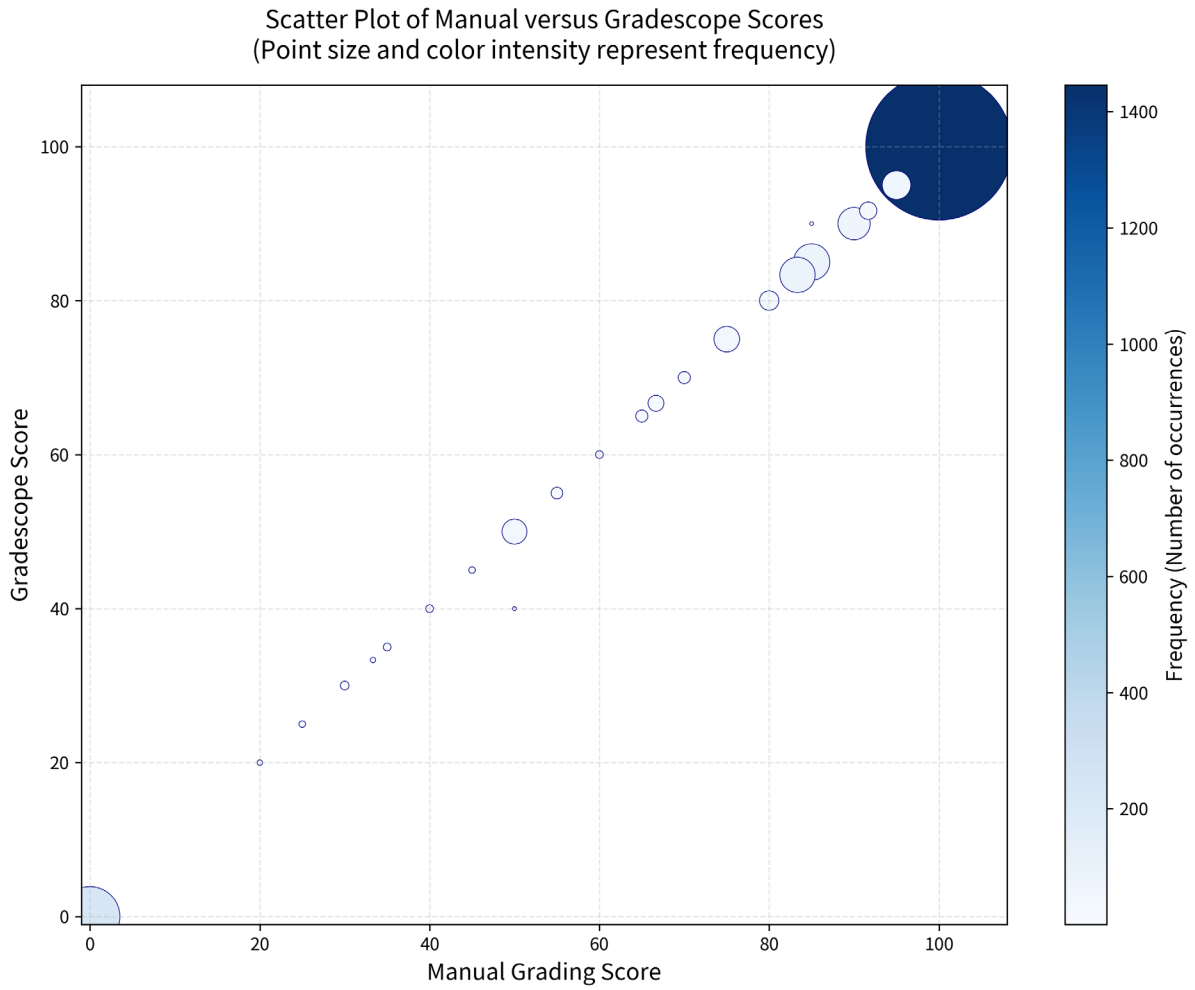


*Figure 1: Scatter Plot of Manual versus Gradescope Scores (All Courses)*

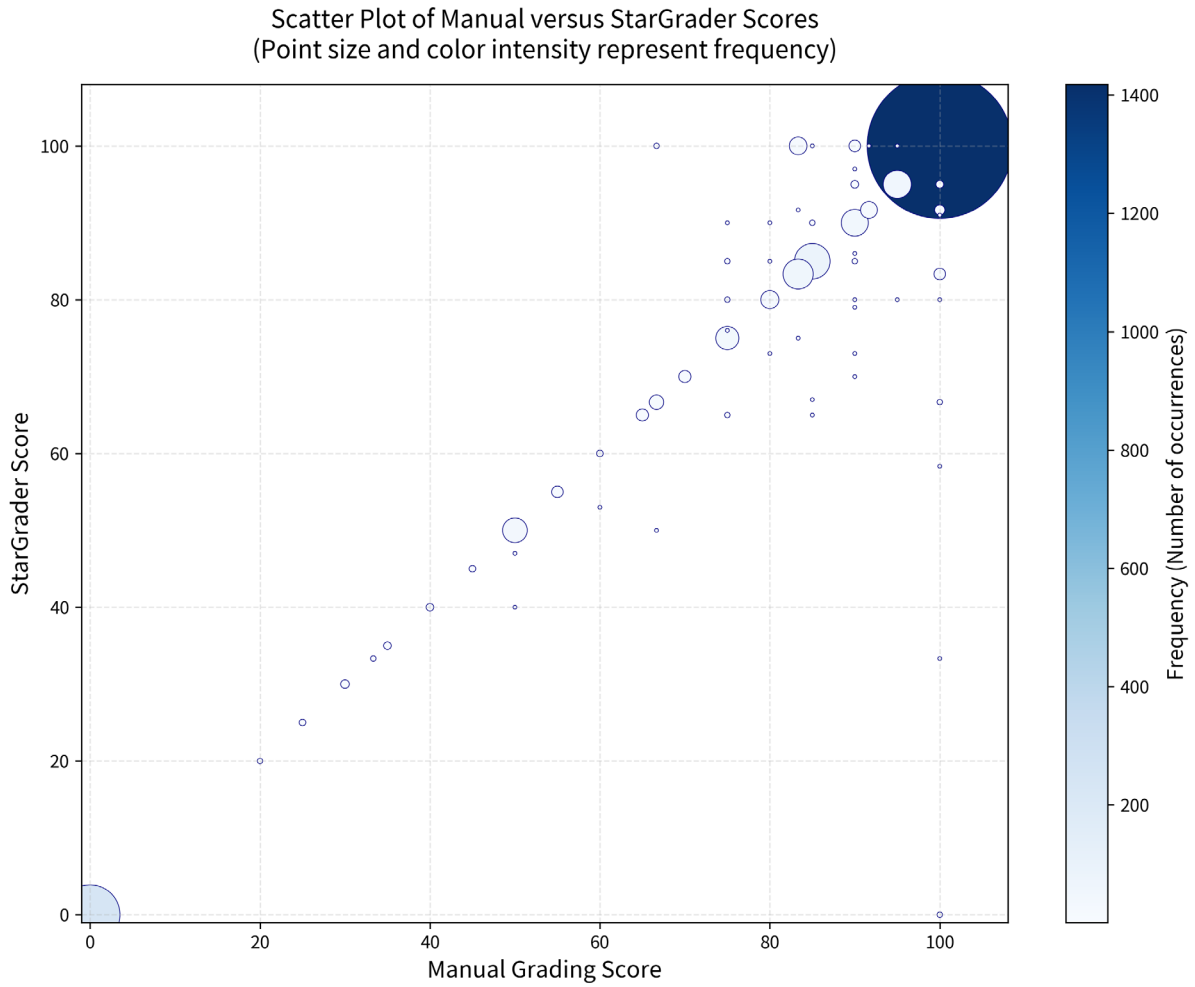


*Figure 2: Scatter Plot of Manual versus StarGrader Scores (All Courses)*

### 4.2 Performance by Course

There are noticeable variations in the performance in courses with different natures. In MATH2205 (Multivariable Calculus) and MATH2207 (Linear Algebra I), questions generally involved shorter solution steps, and alternative valid approaches typically converged to the same final answer. In contrast, questions in MATH3405 (Differential Equations I) frequently allowed for multiple acceptable formulations, where correct solutions could differ substantially. This is further complicated by the fact that general solutions to differential equations could take various forms which may look very different but are in fact equivalent. Such characteristic significantly increased the difficulty for AI tools to accurately match student responses with the expected answers, resulting in lower agreement rates and reduced time savings for MATH3405 compared with the other two courses (see Table 2). These findings suggest that the effectiveness of AI-assisted grading is highly sensitive to the degree of structural variability in acceptable solutions.

*Table 2: Performance Summary by Course: Exact Agreement, Reliability, and Time Savings*

| Course | Exact Agreement (%) | Gradescope Reliability (±5) | StarGrader Reliability (±5) | Time Savings (Gradescope) | Time Savings (StarGrader) |
|---|---|---|---|---|---|
| MATH2205 | 100.00 | 100.00% | 100.00% | +16.3% | +35.1% |
| MATH2207 | 96.08 | 100.00% | 96.08% | +43.96% | -4.8% |
| MATH3405 | 91.77 | 99.83% | 95.03% | –1.6% | –15.2% |

*Source: Author's own analysis.*

### 4.3 Effective Prompt Strategies for StarGrader

StarGrader's performance was highly sensitive to prompt quality. As shown in Fig. 4.3, accuracy in MATH2205 Worksheet 1 improved dramatically from 43.75% in the initial basic prompt to 94.17% by Iteration 3. Similar gains were observed across multiple worksheets, with the best-performing prompts reaching 98–99% accuracy on calculation-heavy tasks. Indeed, the majority of inaccuracies stemmed from the lack of or ambiguous prompts rather than OCR errors. These results clearly demonstrate that prompt engineering is a critical factor in determining the reliability and usefulness of generative AI for mathematics grading.

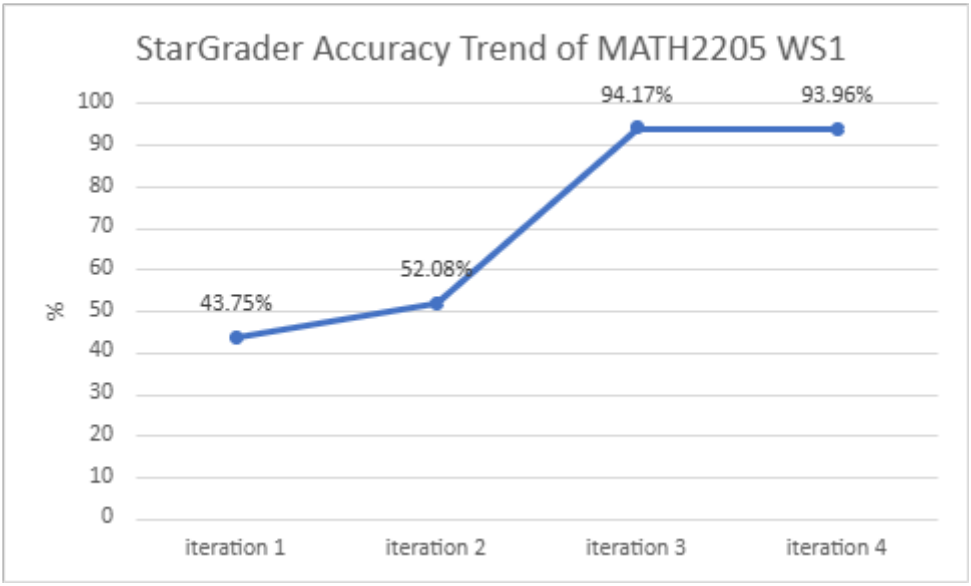


*Figure 3: StarGrader Accuracy Trend across Prompt Iterations in MATH2205 Worksheet 1*

### 4.4 Common Misreads and Error Sources

The most frequent misreads by StarGrader involved mathematical symbols (37 cases, especially partial derivatives), vectors (31 cases), numbers (18 cases), and equations (15 cases). Detailed, solution-key-aligned prompts substantially reduced these discrepancies. Fig. 4.4 and 4.5 illustrate the distribution of misreads and overall accuracy trends across courses.

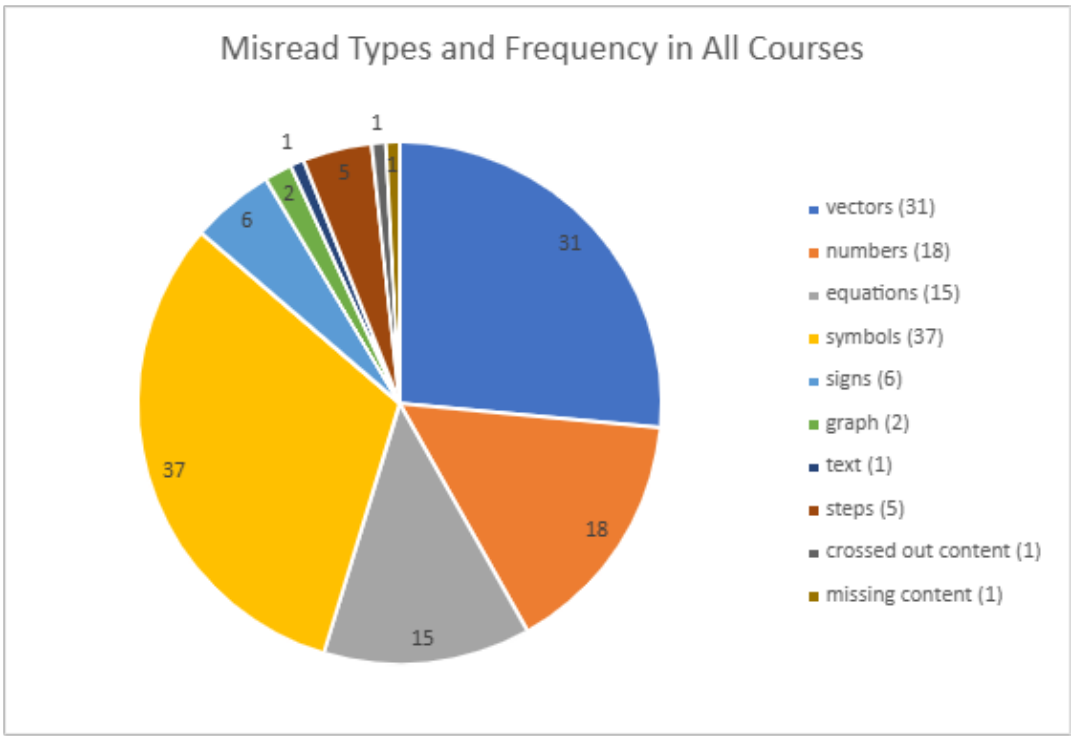


*Figure 4: Frequency of Common Misreads by StarGrader across All Courses*

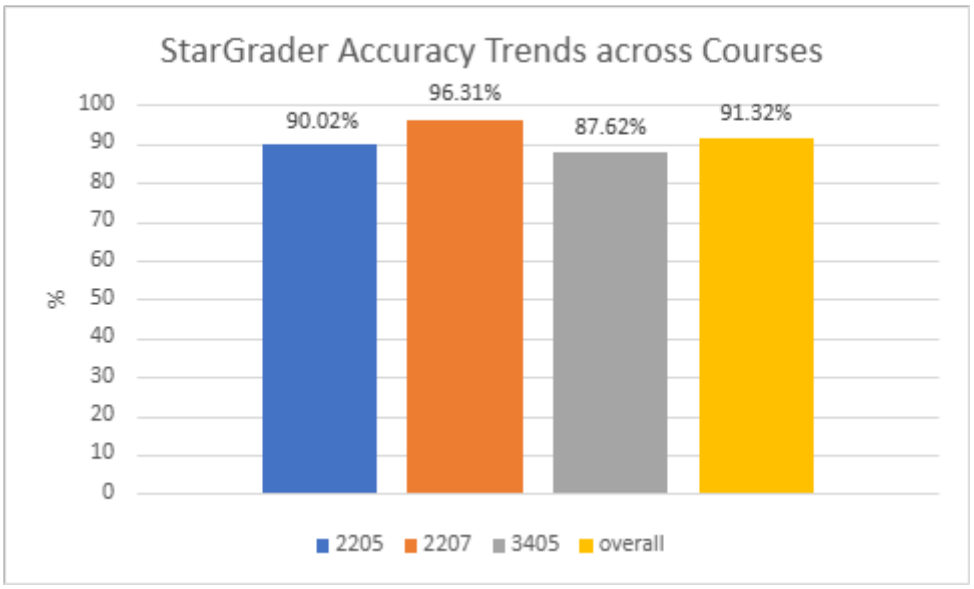


*Figure 5: StarGrader Accuracy Trends across Courses*

### 4.5 Student Perceptions and Instructor Experience

#### 4.5.1 Gradescope Feedback in MATH2215 (Mathematical Analysis)

In AY 2025/26 Semester 2, the principal investigator implemented Gradescope for all in-class exercises in MATH2215 Mathematical Analysis. Compared with previous semesters during which the same instructor graded identical exercises manually, Gradescope reduced grading time by an average of 30%.

A follow-up anonymous survey was distributed to all students in the course, achieving a 32% response rate. Feedback was overwhelmingly positive. Five students reported that the

Gradescope feedback "identified my mistakes more clearly than feedback in other math courses," while three students noted that it "helped me better understand my errors and provided constructive suggestions." No students selected any negative statements. These responses, though based on a modest response rate, reflect students' actual experience with real Gradescope feedback on their own submitted work and provide ecologically valid insights.

### 4.5.2 Comparative Student Survey

A separate comparative survey presented undergraduate mathematics students with identical work graded by all three methods (Manual, Gradescope group feedback, and StarGrader individual feedback). With 19 complete responses (response rate = 12.3%), students consistently rated both AI-generated feedback types higher than traditional manual feedback on a 10-point scale (Tab. 4.3).

*Table 3: Student Satisfaction Ratings by Grading Method (10-point scale)*

| | **Manual Grading** | **Gradescope** | **StarGrader** |
|---|---|---|---|
| **Mean** | 4.44 | 7.22 | 6.94 |

*Source: Author's own analysis.*

AI feedback (both group and individual) was also selected more frequently on positive statements such as "clear, easy to read," "helped me understand concepts better," and "provided constructive suggestions." Negative statements were rarely selected.

Some open-ended comments include:

> *"detailed feedback is useful, manual grading is not detailed"*
>
> *"The third one is too long"*

"more clear for the marks given (why deduct marks), not just marks deduction for the question"

> *"I've got the second one in a course. very useful! Good!"*

These findings, though based on a modest sample, show a consistent preference for AI-enhanced feedback and align with the positive responses from the survey conducted for MATH2215.

## 4.6 Instructor Satisfaction and Feedback Quality

To complement quantitative reliability metrics and student perceptions, the principal investigator rated the quality of grading and feedback for each worksheet on a 10-point scale and recorded qualitative observations.

As shown in Tab. 4.4, StarGrader received the highest overall mean rating (M = 6.60), followed by Gradescope (M = 6.07), while manual grading received the lowest rating (M = 4.01). Instructor ratings for Gradescope showed a clear improvement as the grader got more experience. The research assistant graded all worksheets in MATH3405 first using the three methods. Following feedback from the principal investigator on the quality of the feedback provided, the research assistant subsequently refined their approach to providing feedback when using Gradescope. As a result, instructor ratings for Gradescope improved notably, from 4.18 in MATH3405 to 7.84 in MATH2207. This improvement reflects a learning curve in the use of Gradescope.

These findings indicate that while AI-assisted tools such as Gradescope can improve grading efficiency, the quality of feedback produced is also dependent on user experience and the provision of clear guidance on feedback practices.

*Table 4: Instructor Satisfaction Ratings by Course and Grading Method (10-point scale)*

| Course | Manual Grading | Gradescope | StarGrader |
|---|---|---|---|
| MATH3405 | 4.00 | 4.18 | 5.36 |
| MATH2205 | 4.13 | 5.38 | 7.50 |
| MATH2207 | 3.97 | 7.84 | 7.25 |
| **Overall Mean** | **4.01** | **6.12** | **6.71** |

*Source: Author's own analysis.*

Qualitative comments provided rich insights into the strengths and weaknesses of each method.

*Table 5: Selected Instructor Comments by Grading Method*

| Method | Representative Comments |
|---|---|
| Manual | *"Correct scores, but no feedback at all other than the score";*<br>*"Messy handwriting, hard to read";*<br>*"Misgraded in several cases"* |
| Gradescope | *"Rubrics could be shown to student, even if that student did not make the particular mistake"* |
| StarGrader | *"Clear explanation on what's wrong, hardly seen in manual grading";*<br>*"Comprehensive but too long";*<br>*"Very long feedback even if all correct, overloading students"* |

*Source: Author's own analysis.*

### 5.1 Grading Reliability and Accuracy

The results provide strong evidence that AI tools can enhance both feedback quality and grading efficiency in undergraduate mathematics while maintaining high reliability. Across 6,922 student-question responses, both Gradescope and StarGrader demonstrated excellent inter-rater reliability with manual grading. Gradescope achieved near-perfect exact agreement (99.91%), while StarGrader reached 95.57%. Over 96% of scores from both AI tools fell within ±5 marks of manual grading. These figures exceed the 85% reliability threshold specified in the original TDG proposal and compare favourably with previous studies on AI-supported grading in STEM (Singh et al., 2017; Hansel et al., 2024). The joint exact agreement across all three methods was 95.57%, with StarGrader being the main limiting factor.

### 5.2 Time Efficiency and Contextual Interpretation

Time savings were modest but consistent across the study. Gradescope reduced grading time by 22.2% overall, while StarGrader achieved 19.8%. Larger gains were observed in MATH2205, where StarGrader saved up to 35.1% on calculation-heavy worksheets. With actual manual grading time data available for MATH2207, Gradescope demonstrated strong efficiency gains of +43.9%. Some negative time savings in MATH3405 and parts of MATH2207 reflect the additional human adjustment required for complex proof-based questions or for answers that could take different but equivalent forms.

These figures should be interpreted in context. The StarGrader results included multiple prompt development iterations, during which early versions had relatively low accuracy (as

low as 43.75% in MATH2205 WS1). This led to higher adjustment times in the initial phase. Once refined prompts achieved over 94% accuracy, adjustment time decreased substantially. Consequently, the reported time savings are conservative. With mature, well-engineered prompts and the establishment of a prompt library, StarGrader has strong potential to deliver significantly greater efficiency gains—potentially exceeding 30–40% in courses with repetitive problem types (Lee & Moore, 2024).

The practical value of Gradescope was further demonstrated in MATH2215 Mathematical Analysis, where the principal investigator achieved an average 30% reduction in grading time during AY 2025/26 Semester 2 compared with previous semesters using traditional manual grading. Student feedback from this real classroom implementation was also highly positive. These findings suggest that experienced instructors can achieve more substantial efficiency gains once they become familiar with the tool.

### 5.3 Prompt Engineering and Theoretical Contribution

A particularly significant contribution of this study is the demonstration of prompt engineering's critical role in unlocking StarGrader's effectiveness.

#### 5.3.1 Impact on StarGrader Performance

StarGrader's performance proved highly sensitive to prompt quality. As shown in Figure 3, accuracy in MATH2205 Worksheet 1 improved dramatically from 43.75% in the initial basic prompt to 94.17% by Iteration 3. This substantial improvement highlights the critical importance of systematic prompt engineering in generative AI grading for mathematics.

The best-performing prompt is presented in Appendix A.4. Its high accuracy can be attributed to several key features: clear specification of total marks and per-component allocation at the beginning, explicit and consistent deduction rules (e.g., 5 marks for missing steps or wrong final answer), and balanced feedback instructions. These elements reduced ambiguity and prevented the AI from applying its own inconsistent grading logic. In contrast, earlier prompts with only brief instructions and no detailed marking schemes resulted in much lower accuracy because the AI defaulted to its own criteria.

This finding demonstrates that well-designed prompts can transform generative AI from an unreliable tool into a highly accurate grading assistant in symbolic mathematics. The detailed prompt engineering process and the resulting high-performing prompts represent a practical contribution of this study, with the full set of prompts documented in the StarGrader User Manual.

#### 5.3.2 Theoretical Contribution

The best-performing prompts were those that explicitly incorporated detailed marking schemes, partial-credit rules, and edge-case handling directly from the solution key. This provides concrete, mathematics-specific guidance that can be readily adopted by other instructors and aligns closely with Lee and Moore's (2024) emphasis on prompt quality as the dominant factor in generative AI feedback effectiveness. From a theoretical perspective, these refined prompts also better support Hattie and Timperley's (2007) feedback model by delivering clearer "feed forward" information to students.

### 5.4 Student Perceptions

Student perceptions toward AI-enhanced feedback were consistently positive across both surveys conducted in this study. In the comparative survey (n = 15), students rated both Gradescope group feedback and StarGrader individual feedback higher than traditional

manual feedback on a 10-point scale. AI-generated feedback was also selected more frequently on positive statements such as "clear, easy to read," "helped me understand concepts better," and "provided constructive suggestions." Written comments further highlighted appreciation for detailed explanations and clearer indication of why marks were deducted.

This trend was reinforced by the MATH2215 Gradescope survey, which was administered in a real classroom setting after students received actual Gradescope feedback on their own in-class exercises. With a 32% response rate — modest but typical for voluntary end-of-semester surveys — responses were overwhelmingly positive. Five students reported that the feedback "identified my mistakes more clearly than feedback in other math courses," while three students highlighted that it "helped me better understand my errors and provided constructive suggestions." Notably, no students selected any negative statements.

Although both surveys had relatively small samples, the consistency of positive feedback across hypothetical comparisons and real classroom use complements the strong quantitative reliability results. Together, these findings suggest that students particularly value the clarity, timeliness, and specificity offered by AI tools when applied to mathematics assessment.

### 5.5 Instructor Perspectives

Instructor satisfaction ratings offered valuable insights into the practical usability of the three grading approaches. StarGrader consistently received the highest ratings (overall $M = 6.60$), reflecting its strength in providing detailed and constructive feedback. Gradescope also performed well ($M = 6.07$), with ratings improving markedly as the instructor gained experience with the platform — as reflected from the improvement from 4.00 in MATH3405 to 7.84 in MATH2207. This upward trend suggests a significant learning curve associated with effective use of Gradescope.

These findings indicate that while initial adoption may require investment in time and training, experienced users can derive substantial benefits from AI tools in terms of both efficiency and perceived feedback quality. The lower ratings for manual grading ($M = 4.01$) were primarily attributed to limited feedback depth and occasional inconsistencies, further highlighting the potential advantages of AI-assisted approaches when properly implemented.

When combined with student perceptions and the strong quantitative reliability results, these instructor perspectives reinforce the value of hybrid AI-human grading workflows in undergraduate mathematics courses.

### 5.6 Limitations

Several limitations should be acknowledged when interpreting the findings of this study. First, time savings may have been understated because the dataset included prompt development iterations with higher adjustment times. Second, although 6,922 student-question responses were analysed quantitatively, the comparative student perception survey had a very small sample size ($n = 15$). While the separate MATH2215 Gradescope survey provided additional supportive evidence, the qualitative insights remain preliminary. Third, all grading and data analysis were conducted by a single researcher and one research assistant, which may introduce individual grading style bias and limit generalisability. Fourth, manual grading time records were unavailable for MATH2207 Linear Algebra I, requiring the use of Gradescope time as a proxy in some comparisons and potentially affecting time-saving estimates. Fifth, significant institutional delays in the approval and setup process for Gradescope substantially restricted the scale of the pilot and prevented the collection of instructor and teaching assistant satisfaction data. Finally, StarGrader's performance was

highly dependent on prompt quality, indicating that generative AI grading requires careful engineering and continued human oversight.

### 5.7 Implications and Considerations

Despite these limitations, the findings suggest that both Gradescope and StarGrader can serve as viable and reliable tools for undergraduate mathematics assessment when supported by human oversight and well-designed prompts. The study provides practical, evidence-based guidance for departmental adoption of AI-assisted grading.

However, several considerations should be noted for wider implementation. The generalisability of the findings is limited, as the study was conducted at a single institution and focused primarily on computational mathematics problems. Potential risks include algorithmic bias in grading and the generation of overly long feedback, which may reduce its effectiveness for students. Additionally, the use of external AI platforms raises data privacy concerns, and institutions should ensure compliance with relevant data protection policies. Finally, while this study focused on AI for grading, the growing use of generative AI also has broader implications for academic integrity that warrant careful institutional attention.

This study provides robust empirical evidence that artificial intelligence tools can significantly enhance feedback quality and grading efficiency in undergraduate mathematics courses while maintaining high reliability. Across 6,922 student-question responses from three core mathematics courses, both Gradescope and StarGrader achieved strong inter-rater reliability with manual grading. Gradescope reached 99.95% and StarGrader 96.44% of scores within ±5 marks of the manual standard. Modest time savings were recorded overall (Gradescope +22.2%, StarGrader +19.8%), with notably larger gains in calculation-heavy worksheets. Prompt engineering proved to be a decisive factor for StarGrader, raising accuracy from 43.75% to over 94% through systematic refinement.

Preliminary student surveys indicated more positive perceptions of AI-generated feedback in terms of clarity, personalisation, and usefulness. Instructor feedback further reinforced these findings, with StarGrader receiving the highest satisfaction ratings for feedback quality and Gradescope demonstrating clear improvement as instructors gained experience. In addition to the empirical findings, the project produced two practical deliverables—the Gradescope and StarGrader User Manuals—to support wider adoption.

These findings have important implications for undergraduate mathematics education. AI tools can help address the long-standing challenge of providing timely and detailed feedback in large classes. However, the results also highlight that AI cannot yet fully replace human judgment; careful prompt design and human oversight remain essential for consistent and pedagogically sound outcomes.

Based on the evidence, the following actions are recommended:

(1) Adopt a hybrid workflow that combines AI grading with targeted human review.

(2) Implement the recommended answer box design principles to maximise AI accuracy.

(3) Develop and share a library of high-performing prompts for common mathematics topics.

(4) Conduct larger-scale surveys to gather more robust student and instructor perception data as adoption expands.

In conclusion, AI-assisted grading represents a promising approach to improving assessment practices in mathematics education. With continued refinement and institutional support, these tools have the potential to reduce instructor workload, enhance feedback quality, and ultimately improve student learning outcomes.

**Acknowledgment**

This work was supported by a Teaching Development Grant from Hong Kong Baptist University. The author thanks the research assistant for data collection and analysis.

## References

Hansel, C. A., Ottenbreit-Leftwich, A., Quick, J. D., Greene, A. H., & Ricci, M. (2024). Gradescope in Large Lecture Classrooms: A Case Study at Indiana University: How an online grading platform enhanced student learning and instructor feedback in large-scale courses. *Journal of Teaching and Learning with Technology, 13*(1). https://doi.org/10.14434/jotlt.v13i1.38519

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Lee, S. S., & Moore, R. L. (2024). Harnessing generative AI (GenAI) for automated feedback in higher education: A systematic review. *Online Learning, 28*(3), 82–106. https://doi.org/10.24059/olj.v28i3.4593

Li, H., Li, C., Xing, W., Baral, S., & Heffernan, N. (2024). Automated feedback for student math responses based on multi-modality and fine-tuning. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. ACM. https://doi.org/10.1145/3636555.3636860

Lindsay, E., Zhang, M., Johri, A., & Bjerva, J. (2025). The responsible development of automated student feedback with generative AI. *2025 IEEE Global Engineering Education Conference*, London, United Kingdom, 2025, pp. 1-10. https://doi.org/10.1109/EDUCON62633.2025.11016572

Meylani, R. (2025). AI-powered assessments in mathematics education: A systematic review of contemporary research literature. *Buca Eğitim Fakültesi Dergisi, 66*, 3642–3674. https://doi.org/10.53444/deubefd.1527781

Morris, W., Holmes, L., Choi, J. S., & Crossley, S. (2025). Automated scoring of constructed response items in math assessment using large language models. *International Journal of Artificial Intelligence in Education, 35*(2), 559-586. https://doi.org/10.1007/s40593-024-00418-w

Singh, A., Karayev, S., Gutowski, K., & Abbeel, P. (2017). Gradescope: A Fast, Flexible, and Fair System for Scalable Assessment of Handwritten Work. In *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale*. Association for Computing Machinery, New York, NY, USA, 81-88. https://doi.org/10.1145/3051457.3051466

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35*, 24824–24837. https://doi.org/10.48550/arXiv.2201.11903

Zhao, C., Silva, M., & Poulsen, S. (2025). Autograding mathematical induction proofs with natural language processing. *International Journal of Artificial Intelligence in Education, 35*(5), 3202-3232. https://doi.org/10.1007/s40593-025-00498-2

## Appendix

This appendix summarises the two AI grading tools used in the study and provides an exemplar prompt along with guidance on effective question design.

### Gradescope User Guide Summary

Gradescope uses AI primarily for handwritten answer grouping, while human graders assign marks and provide feedback. Its key strength lies in automatically grouping similar student answers, enabling efficient batch grading. The recommended workflow involves defining clear answer regions when setting up the assignment outline, uploading submissions, and applying rubrics to answer groups. For mathematics assignments, placing answer boxes tightly around the final answer (rather than over large blank spaces) significantly improves grouping accuracy. A limitation is that a complete student roster must be provided before submissions can be processed.

### StarGrader User Guide Summary

StarGrader is a generative AI tool that performs fully automated grading and provides personalised feedback. Its effectiveness depends heavily on prompt quality. Important prompt design principles include stating mark allocations clearly, maintaining consistent marking criteria, and specifying the desired feedback tone. The tool supports iterative refinement, allowing graders to adjust feedback for individual submissions or across all responses to a question.

### Effective Design of Answer Boxes for AI Grading

Effective use of both tools requires thoughtful question design. For Gradescope, concise final answer boxes improve clustering accuracy, where students should write only the final answer inside the box and provide working outside it. For StarGrader, providing sufficient space for steps and explanations helps reduce misgrading.

Clear instructions (e.g., “Write only Y or N”) also improve consistency. Overly large answer boxes often lead students to write full sentences, which reduces grouping performance in Gradescope and increases the risk of misreads in StarGrader.

### Exemplar Prompt

The following prompt achieved 94.17% accuracy on a sample worksheet:

“10 marks for each component. Full marks are 100 marks. For each component, if no steps, deduct 5 marks. If the final answer is incorrect, deduct 5 marks. Grade fairly at university level. Marks can be not deducted if the calculation mistakes do not affect the final answer. Remind the student to write down steps and compute the final answer instead of leaving an equation. Check for existence of steps before giving marks; steps can be anywhere in the

submission. If the answer in the box is wrong, 5 marks should be deducted even if the steps are all correct. Provide balanced feedback."