A Related Work

Automated Grading Systems. Automated grading tools were first introduced in the 1960s for programming [19] and essay evaluation [29]. Since then, numerous methods have been developed, including unit testing [26], rule-based approaches [24,35], and techniques based on stacking and domain adaptation [17]. These methods typically compare student submissions with reference answers using text similarity, measurable text characteristics (e.g., sentence length, essay length, number of prepositions, and punctuation), and predefined rules. However, with the rapid advancement of deep learning, machine learning, and neural network-based methods have proven more effective than traditional non-neural approaches [31]. Consequently, various neural models have been applied to automated grading, including Long Short-Term Memory (LSTM) networks [28,31], Convolutional Neural Networks (CNNs) [32], and Prototypical Neural Networks [39]. Additionally, additive model-based methods [9] have been introduced to enhance model explainability.

Large Language Models in Automated Grading Systems. In recent years, the significant growth of LLMs in natural language processing has led to their adoption in automated grading systems for evaluation using prompts and predefined criteria (e.g., marking schemes). Compared to traditional neural models, LLMs offer the advantage of handling tasks such as question answering, reading comprehension, and summarization without requiring fine-tuning (i.e., training on task-specific datasets through supervised learning) [30]. However, in automated grading, few-shot learning—supervised learning with limited training epochs—has become the standard approach. This method has been applied to state-of-the-art LLMs, including BERT [7,33], GPT-2 [30], GPT-3 [3,5,27], LaMDA [36], and GPT-4 [8,21,37]. These studies demonstrate the effectiveness of LLMs in grading through benchmarking on grading datasets and evaluations conducted by educators. While there have been attempts [22] to apply zero-shot learning with simple prompts on higher education courses using LLMs (i.e., without further supervised learning), it nevertheless proved inadequately effective for comprehensive assessments, such as examinations.

Student Perceptions and Adaptive Feedback. While LLMs have shown effectiveness in grading and supporting student learning, understanding students' perceptions on automated grading systems is equally important. Although automated grading systems have been implemented in classroom settings [6,12,23], few studies have examined how students perceive these systems. Concerns may arise regarding the accuracy of automated grading, misunderstandings of system operations—resulting in suboptimal responses—and difficulties in adapting answers to align with evaluation criteria [2,14,20]. Moreover, effectively identifying student mistakes remains critical, as emphasized by the LLM-based feedback tool introduced in [25].

Limitations. Although previous studies offer valuable insights into Automated Grading Systems, several limitations may hinder their further development and effectiveness:

1. **Dataset Requirements**. Few-shot learning methods still require substantial datasets, especially considering the scale of LLMs. Acquiring large-scale datasets can be costly and may not be feasible for introductory courses or ad-hoc topics, where content frequently changes.
2. **Limited Focus on Student Feedback**. Automated Grading Systems often prioritize grading accuracy while neglecting additional functions such as providing meaningful feedback and comments to students.
3. **Benchmark-Oriented Evaluation**. These systems are typically evaluated based on their performance on grading datasets, overlooking the student experience and feedback.

To address these challenges, we propose a zero-shot LLM-based Automated Assignment Grading (AAG) System capable of evaluating answers that include both calculations and natural language explanations through prompt engineering. Additionally, the system provides constructive feedback to students by identifying mistakes and suggesting ways to improve. While extensive research has examined the grading effectiveness of LLMs, this study focuses on evaluating the system from students' perspectives through survey responses after grading their actual homework and delivering personalized feedback—an area that, to the best of our knowledge, remains underexplored.

B Extended Experiments and Results

B.1 AAG Grading Evaluation

While there was substantial evidence demonstrating the effectiveness of LLMs in grading (see Section A), a limitation of LLM-based grading was the distributional differences between AI scorers and human [16]. To address this issue, we compared the scores assigned by TAs and the AAG system using two open-ended questions from the "Introduction to Statistics" course. Open-ended questions were selected due to their greater flexibility in student responses compared to calculation-based questions. These two questions are summarized in Tables 2, and 3, respectively. For the comparison, both the TA and the AAG system independently graded 150 student assignments.

Figure 4 presents the grading distributions of Question 1 (left) and Question 2 (right) by both human graders and the AAG system. The Pearson correlation coefficients between human and AAG scores for Questions 1 and 2 were 0.75 and 0.82, respectively. These strong correlations ($p < 0.0001$) demonstrate a highly significant relationship, indicating that the AAG system's grading results closely align with human evaluations. Supporting the grading accuracy of the AAG system.

Despite the high correlations, noticeable differences in score distributions were observed between the two questions. Although both were open-ended, the marking schemes differed significantly. The marking scheme for Question 1 (see Table 2) provided detailed guidelines for awarding marks, whereas the marking scheme for Question 2 (see Table 3) offered only general grading criteria. As a

Table 2. Question and marking scheme for "Introduction to Statistics" question 1.

Background	You are a Statistician at the Census and Statistics Department, tasked with leading a team to collect data on the aging population in a housing estate in <anoynomised location>. The collected data will be reported to the Social Welfare Department.
Question	What kind of sampling method will you suggest using?
Marking Scheme	Award 3 marks if the student selects cluster sampling as the sampling method. Award 2 marks if the student proposes any other appropriate sampling method. Award 1 mark if the student addresses practical issues of the proposed sampling scheme (e.g., ease of implementation, availability of clustering/stratifying variables). Award 2 marks if the student considers the cost-effectiveness of the proposed sampling scheme. Award 1 mark for a strong, well-reasoned explanation of why the proposed method is cost-effective. Award 2 marks if the student considers the representativeness of the proposed sampling scheme. Award 1 mark for a strong, clear justification of how the proposed method ensures representativeness.

Table 3. Question and marking scheme for "Introduction to Statistics" question 2.

Background	You would like to conduct a survey to predict the results of the US presidential election.
Question	Find a good variable for forming strata. Explain your choice briefly.
Marking Scheme	This is an open-ended question, and the answer can be subjective. The student can choose any good variables they want. Some examples include by states and/or age group, etc. The student needs to explain that there is large between-strata variation and that a representative sample is obtained based on the chosen variable. Additionally, the student must explain that such a stratifying variable is feasible in practice.

result, both human and AAG scores for Question 1 were more diverse, while scores for Question 2 were more concentrated. This highlights the importance of the refined marking scheme function, which ensures more consistent and precise grading.

A further qualitative analysis of scoring discrepancies across all assignment questions revealed that some differences stemmed from human factors, such as grading errors (e.g., overlooking a missing zero, incorrect calculation processes, or missing explanations). These inconsistencies often arose when TAs failed to

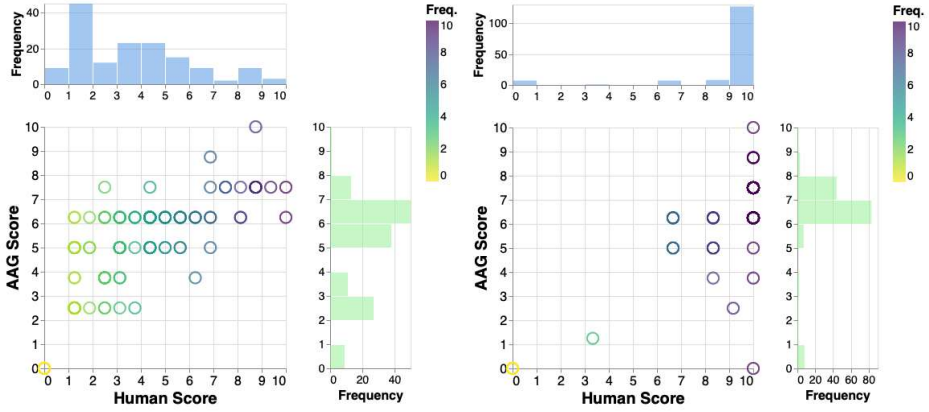

Fig. 4. Human and AAG system grading distribution of question 1 (left) and question 2 (right) in "Introduction to Statistics".

strictly follow the marking scheme, encountered vague grading guidelines, or made errors due to the large volume of assignments. These analysis highlights the importance of implementing the AAG system to enhance grading consistency and provide higher-quality feedback to students.

B.2 Student Perspective on AAG System

Survey Design: To evaluate the effectiveness of the AAG system, a voluntary survey was administered to students, all of whom received feedback from both the AAG system and TA. Participants included students from the undergraduate course "Introduction to Statistics" and the graduate courses "Data Management" and "Principles of Risk Management" at collaborating university. Participants were informed that their responses would not impact their assignment scores. The survey included 10 questions aimed at assessing the quality of the system's feedback and its influence on learning. Questions 1 to 9 employed a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree), while Question 10 asked students to choose between the traditional TA grading method and the AAG system feedback.

Survey Questions: The survey questions targeted multiple dimensions of the feedback experience (The exact survey question is available in Section E):

- **Q1-Q2:** Focused on the clarity and helpfulness of the feedback in identifying specific mistakes.
- **Q3-Q4:** Evaluated how well the feedback clarified why answers were incorrect and the degree of customization to individual responses.
- **Q5-Q6:** Measured how effectively the feedback provided actionable steps for improvement and addressed personal learning needs.
- **Q7-Q8:** Assessed the impact of feedback on future problem-solving and its insight into knowledge gaps.

– **Q9:** Investigated whether the feedback encouraged motivation to learn and correct mistakes.
– **Q10:** Asked students to choose between conventional TA grading and the AAG system feedback.

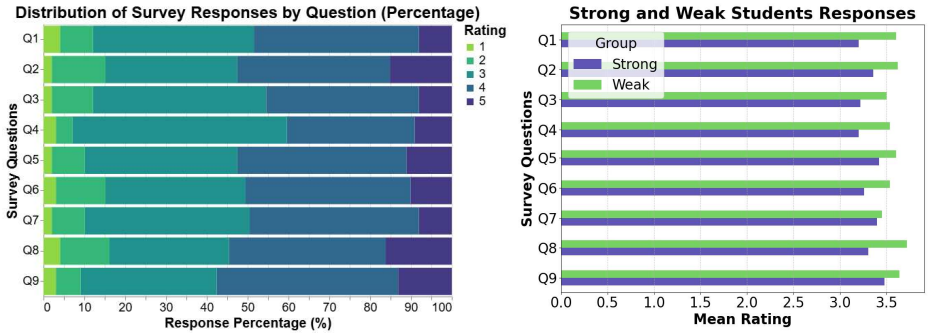

Fig. 5. Distribution of survey responses. For questions 1 through 9, responses follow a Likert scale where 5 indicates the highest rating and 1 indicates the lowest. The left plot displays the distribution of responses from all students, while the right plot shows the mean ratings for strong and weak students (defined as the top 50% and bottom 50% performers, respectively).

Results and Interpretation: In total 104 responses were collected and the survey results are summarized in Fig. 5 (left). A Wilcoxon Signed-Rank Test and a Binomial Test were conducted on the responses to Questions 1-9 and Question 10, respectively (detailed results can be found in Section C). With strong statistical significance ($p < 0.0001$), the tests revealed consistently positive responses across all survey items. Specifically, for Questions 1-9, students indicated that the AAG system significantly improved their understanding of mistakes, provided clearer explanations, and delivered highly customized feedback tailored to their learning needs. Moreover, the feedback was effective in offering actionable steps for improvement, thereby enhancing students' motivation to learn.

Specifically, the survey results indicated that students felt more prepared to tackle similar problems in the future due to the AAG system's feedback, which offered deeper insights into their learning gaps. The system's ability to provide personalized feedback contributed to a more engaging and effective learning experience. For Question 10, the majority of students (93 out of 104) preferred the AAG system over traditional TA grading, highlighting its perceived superiority in offering constructive, individualized feedback that better supports student learning.

Furthermore, the responses from weak and strong student groups, defined as the top 50% and bottom 50% performers on the corresponding assignment, were compared to identify any significant differences in perception or understanding

between the two groups. As shown in Figure 5 (right), the mean ratings indicate that weak students generally gave higher ratings to the AAG system across all aspects.

A Mann-Whitney U test[1] was also performed to assess the significance of these differences (detailed results in Section D). With strong confidence ($p < 0.05$), the analysis revealed that weaker students were more satisfied with the AAG system in four key areas.

For Q1 (Identifying Mistakes), weaker students found the detailed feedback especially helpful in pinpointing mistakes. For Q4 (Actionable Feedback for Improvement), they also felt the system guided them better toward improvement strategies. For Q6 (Future Problem-Solving), they felt more prepared for future problems, enhancing their confidence and skills. Lastly, for Q8 (Encouragement for Learning), the system effectively motivated and encouraged continuous learning, especially for struggling students.

Overall, these findings highlight the AAG system's substantial impact on enhancing feedback quality, student engagement, and learning outcomes. The system was particularly effective in supporting weaker students by improving their ability to identify mistakes, providing practical guidance for improvement, boosting their confidence in problem-solving, and fostering greater motivation to learn. These results underscore the value of personalized, actionable feedback in addressing the specific needs of students, especially those who may struggle, ultimately contributing to a more inclusive and effective learning environment.

C Hypothesis Testing for Survey Responses

This section outlines the hypothesis testing performed on the survey responses, using statistical tests to evaluate the significance of the responses for each question.

1. Wilcoxon Signed-Rank Test (for Q1 to Q9)

- **Null Hypothesis (H_0):** The median response for each question (Q1 to Q9) is less then or equal to 3.
- **Alternative Hypothesis (H_1):** The median response for each question (Q1 to Q9) is greater than 3.

2.Binomial Test (for Q10)

- **Null Hypothesis (H_0):** The proportion of binary responses is less than or equal to 0.5.
- **Alternative Hypothesis (H_1):** The proportion of binary responses greater than 0.5.

The hypotheses are determined based on the following conditions:

[1] Note that the power of the parametric t-test and the non-parametric Mann-Whitney-Wilcoxon test are similar for five-point Likert scale data [10].

- If the p-value is less than 0.05, the null hypothesis will be rejected, indicating a significant difference.
- If the p-value is greater than 0.05, the null hypothesis cannot be rejected, suggesting no significant difference.

The results of the hypothesis tests for questions Q1 to Q9 (Wilcoxon Signed-Rank Test) and Q10 (Binomial Test) are presented in Table 4.

Table 4. Wilcoxon Signed-Rank test result for Q1-Q9 and Binomial test result for Q10.

Question	**Test Statistic**	**p-value**	**Sample Size**	**Interpretation (α=0.05)**
Q1	1523	< 0.0001	103	Significant (greater than 3)
Q2	1888	< 0.0001	103	Significant (greater than 3)
Q3	1335	< 0.0001	104	Significant (greater than 3)
Q4	1000	< 0.0001	104	Significant (greater than 3)
Q5	1794	< 0.0001	103	Significant (greater than 3)
Q6	1827	< 0.0001	104	Significant (greater than 3)
Q7	1490	< 0.0001	103	Significant (greater than 3)
Q8	2131	< 0.0001	103	Significant (greater than 3)
Q9	1883	< 0.0001	103	Significant (greater than 3)
Q10	0.8942	< 0.0001	104	Significant (greater than 0.5)

For all questions Q1 to Q9, the p-values are less than 0.05, indicating that the null hypothesis is rejected, and the median response for each question is significantly greater than 3. This suggests that, on average, the respondents rated each of these questions more favorably than the neutral midpoint (3). For Q10, the p-value is also less than 0.05, leading to the rejection of the null hypothesis. This means that the proportion of positive binary responses is significantly greater than 0.5, indicating a preference for the positive response.

D Hypothesis Testing for Between-Group Comparison

This section describes the use of the **Mann-Whitney U Test** (for Q1 to Q10) for comparing the survey responses between two groups of students: weak students (Group 1) and strong students (Group 2). For the comparison between weak and strong students, the hypotheses are as follows:

- **Null Hypothesis (H_0):** There is no significant difference between the weak students (Group 1) and the strong students (Group 2) with respect to the survey responses, i.e., the distribution of survey responses for weak students is less than or equal to that of strong students.
- **Alternative Hypothesis (H_1):** The distribution of survey responses for weak students is greater than that of strong students.

The hypotheses are determined based on the following conditions:

- If the p-value is less than 0.05, the null hypothesis will be rejected, indicating a significant difference.
- If the p-value is greater than 0.05, the null hypothesis cannot be rejected, suggesting no significant difference.

The results of the Mann-Whitney U test comparing weak and strong students (Groups 1 and 2) for each survey question are summarized in Table 5. The interpretation of each result follows the decision rule outlined above.

Table 5. Comparison between weak and strong students (group 1 and group 2) with interpretations. Where n denotes the sample size of the respective group.

Survey Question	**Test Statistic**	**p-value**	$\mathbf{n}_{group1}$	$\mathbf{n}_{group2}$	**Interpretation (α=0.05)**
Q1	1687	0.0053	53	50	Significant
Q2	1537	0.0717	53	50	Not Significant
Q3	1561	0.0711	54	50	Not Significant
Q4	1603	0.0350	54	50	Significant
Q5	1482	0.1355	53	50	Not Significant
Q6	1594	0.0471	54	50	Significant
Q7	1354	0.4212	53	50	Not Significant
Q8	1655	0.0109	54	49	Significant
Q9	1515	0.0905	53	50	Not Significant
Q10	1335	0.5751	54	50	Not Significant

From the Mann-Whitney U test results, significant differences between weak and strong students were observed for Q1, Q4, Q6, and Q8, where the p-values were less than 0.05. This indicates that the survey responses of weak students were significantly higher than those of strong students for these questions. For the remaining questions (Q2, Q3, Q5, Q7, Q9, and Q10), the p-values exceeded 0.05, suggesting no significant differences in the responses between the two groups.

E Survey Questions

This section includes the set of survey questions designed to gather valuable insights into various aspects of the AAG system. The goal is to understand students' perspectives, experiences, and feedback, which will inform future development and improvements.

Survey Items

1. **Identifying Mistakes:** To what extent did the AI feedback help you identify the specific mistakes in your answers, compared to just seeing the solution alone?
   - No additional help in identifying mistakes
   - Very little additional help

   - Some additional help
   - Noticeable improvement in identifying mistakes
   - Much clearer identification of mistakes
2. **Clarity of Explanation for Errors:** Did the AI comment clarify why your answers were incorrect better than the standard solution alone could?
   - No added clarity
   - Slightly clearer but limited
   - Moderately clearer
   - Significantly clearer explanation
   - Very clear and detailed explanation
3. **Level of Customization:** How well was the AI feedback tailored to your specific answer compared to a general solution?
   - Not customized at all
   - Minimally customized
   - Somewhat customized
   - Well-customized
   - Very highly customized to my answer
4. **Benefit of Customization:** How much did the customized AI comment, tailored to your specific answer, help you address your personal learning needs compared to a general solution?
   - No additional help for my learning needs
   - Very minimal additional help
   - Somewhat helpful for my needs
   - Noticeably helpful for my needs
   - Extremely helpful and targeted to my learning needs
5. **Actionable Feedback for Improvement:** How effective was the AI feedback in providing specific steps or areas to improve, compared to the solution alone?
   - Not effective in providing improvement steps
   - Very limited effectiveness
   - Somewhat effective
   - Effective in providing actionable steps
   - Extremely effective with clear, actionable feedback
6. **Future Problem-Solving:** How prepared do you feel to handle similar problems in the future with AI feedback, as opposed to only seeing the solution?
   - Not more prepared than with solution alone
   - Slightly more prepared
   - Moderately more prepared
   - Significantly more prepared
   - Very confident in handling similar problems
7. **Depth of Insight into Learning Gaps:** To what extent did the AI feedback help you identify specific knowledge gaps, in contrast to the general solution alone?
   - No additional insight into my learning gaps
   - Minimal additional insight

   - Some additional insight
   - Clearer insight into learning gaps
   - Provided significant, valuable insight
8. **Encouragement for Learning:** How much more motivated did you feel to address your mistakes after receiving the AI comment, as compared to just the solution?
   - Not more motivated than with solution alone
   - Slightly more motivated
   - Somewhat more motivated
   - Noticeably more motivated to improve
   - Very motivated to address and learn from mistakes
9. **Comparative Learning Benefit:** To what extent did the AI feedback contribute to an improved learning experience compared to receiving only the solution?
   - No additional benefit compared to solution alone
   - Slight additional benefit
   - Moderate additional benefit
   - Significant benefit for learning
   - Greatly enhanced learning experience
10. **Preference Between Solution Types:** Overall, which did you find more helpful for understanding: the solution alone or the solution with the AI feedback?
   - Solution alone
   - Solution + AI Comment
11. **Suggestions for AI Feedback Improvement (Open-Ended):** What additional information or aspects in the AI feedback would make it even more helpful compared to the solution alone?