AI-Marking Assistant: A Web-Based Application for Human-in-the-loop GAI Assisted Assessment Marking and Feedback

Article history
Received
1 December 2025
Received in revised form
12 December 2025
Accepted
18 December 2025
Published online
27 December 2025

Paul Craig[a], Thomas Selig[a], Yu Liu[a], Ling Wang[b], Erick Purwanto[a], Wan-ting Shen[c]*
[a]Department of Computing, School of Advanced Technology, Xian Jiaotong Liverpool University, China
[b]Educational Development Unit, Xian Jiaotong Liverpool University, China
[c]Department of Biological Sciences, Xian Jiaotong Liverpool University, China
**p.craig@xjtlu.edu.cn*

Abstract
This paper introduces AI-Marking Assistant (AI-MA), a prototype application that aims to improve the efficiency and consistency of grading and grading feedback by allowing educators to integrate Generative Artificial Intelligence (GAI) assistance into the grading process. While automated grading has the advantages over traditional human marking of being efficient, timely, consistent, scalable, and objective, there are also known limitations and potential issues associated with process. Grading and feedback can lack the nuance and context that would normally come from an expert marker. Results can also be biased by the training data and there are significant ethical and legal implications of allowing a machine to grade assignments without human oversight. AI-MA aims to overcome these limitations by offering a human-in-the-loop GAI assisted interface that allows educators to leverage the power of GAI while keeping an active oversight and interactive role in the marking process. AI-MA allows human graders to manually mark assignments, or edit the output of a GAI model with results fed back to the model in order to improve its performance. A pilot study with the application demonstrates its potential to significantly improve grader performance while maintaining the quality of marking and grade feedback for students. The evaluation with four markers and 250 student submissions showed a 40% reduction in marking time after initial examples were provided to the system.

Keywords: Educational technology, Generative AI, Human computer interaction, Automated assessment, Human-in-the-loop AI.

Introduction

The advantages of providing students with timely personal feedback on assignment submissions are well known to educators and academic researchers alike (Craig et al., 2014; Williams, 2024; Simonsmeier et al., 2020). Personalized feedback helps students understand their strengths and areas for improvement (Tetzlaff et al., 2021). Feedback tailoring to individual needs can guide students toward better performance (Tetzlaff et al., 2021). It can also improve motivation and confidence. When students receive specific, personalized feedback, they feel more motivated to engage with their learning and confidence grows as they see progress and understand how to enhance their work. Timely feedback can also foster a positive connection between students and educators, demonstrating care and investment in the student's progress to enhance overall student engagement.

Despite these benefits, providing personalized feedback presents challenges, particularly in large classes (Tetzlaff et al., 2021). The time required for marking assignments, providing comments, and addressing individual needs can be substantial, creating workload pressures for educators. Additionally, balancing encouragement with constructive criticism while maintaining positive rapport can be emotionally demanding (Øen, 2024).

The current literature identifies several standards for effective personalized feedback. Feedback should encourage students to engage with the assessment process, reflect on their performance, and consider how to improve (Tetzlaff et al., 2021). It should help students understand what good performance looks like through explicit statements of what is done well or poorly and why. Effective feedback recognizes learner differences and tailors feedback to individual needs, highlighting strengths and areas for growth. It should be fair, honest, and clear, balancing positive reinforcement with areas for improvement. Timeliness is also crucial, allowing students to act on feedback to close the gap between their current performance and

Craig, P., Selig, T., Liu, Y., Wang, L., Purwanto, E., & Shen, W (2025), AI-Marking Assistant: A Web-Based Application for Human-in-the-loop GAI Assisted Assessment Marking and Feedback, *ASEAN Journal of Engineering Education*, 9(2), 162 -170. https://doi.org/10.11113/ajee2025.9n2.213

desired standards. Meeting these requirements places a significant burden on educators that is particularly challenging for larger classes.

While various methods exist to automate assignment marking and feedback, they often have limitations. Multiple-choice question grading is easily automated but limited to objective questions that may not assess higher-level learning outcomes (Tuma, 2022). Peer marking and self-marking can shift marking burdens from teachers (Simonsmeier et al., 2020), but these methods require clear criteria that students can easily understand and may suffer from inconsistency and fairness concerns. Peer feedback can be valuable but similarly relies on clear criteria and may lack consistency.

Generative Artificial Intelligence (GAI) and other automated methods can also be applied to automated marking for short answer and essay questions (Burrows et al., 2015; Souppez et al., 2023). AI has clear advantages over traditional human marking as it is known to be efficient, timely, consistent, scalable, and objective (Farrelly & Baker, 2023; Chan & Colloton, 2024). However, there are disadvantages and limitations of AI grading that should be considered. These can be summarised as follows:

- **Lack of Nuance.** Automated grading lacks the nuanced understanding that an expert human marker brings. Context, tone, and subtle nuances in student responses can be challenging for AI to capture. While it can handle routine tasks, it may miss the depth and context that a human grader would naturally consider (Johnston et al., 2024).
- **Bias.** Biases can creep into automated grading systems. These biases stem from the training data used to develop the AI model. If the training data contains inherent biases, the system may perpetuate them during grading (Li et al., 2023).
- **Ethical concerns.** Ethical considerations arise when allowing a machine to grade assignments without sufficient human oversight. While the responsibility for the marking still lies with the educator, if they do not adequately monitor the marking process, bias and inaccuracies from the AI might be left in the marks given to the student (Farrelly & Baker, 2023).

It can be concluded that while AI has clear advantages, human oversight remains vital for holistic evaluation that ensures fair, accurate, and ethical grading practices. This aligns with the Human-in-the-Loop (HITL) AI paradigm, which emphasizes maintaining human agency and oversight in AI-assisted systems (Tarun et al., 2025). Furthermore, from an assessment validity and reliability perspective, AI-assisted tools must ensure that marks accurately reflect student achievement against learning outcomes while maintaining consistency across submissions (Messick, 1995).

A number of techniques aim to overcome the limitations of automated grading by allowing for greater human involvement in the automated marking process (Kaya & Cicekli, 2024; Malik et al., 2019). These demonstrate the potential for automated grading with increased human involvement in the process to make grading more efficient (Farrelly & Baker, 2023) and consistent (Farrelly & Baker, 2023; Chan & Colloton, 2024).

The GradeAid (Kaya & Cicekli, 2024) framework supports automated marking of short text answers by allowing markers to train a Natural Language Processing (NLP) model with a set of sample answers. This model achieves high levels of predictive accuracy and is effective at providing students with timely and effective feedback. It is however limited in that it requires a training dataset that covers a suitable range of answers and may suffer from scalability issues if this is not adequate. Similarly, commercial tools like Gradescope and Turnitin's AI Feedback Studio offer AI-assisted grading features but often lack the interactive, real-time fine-tuning capabilities of a true human-in-the-loop system.

Generative grading (Malik et al., 2019) provides feedback on open-ended assignments in structured domains (e.g., computer programming, graphics, and short response questions) using generative descriptions of student cognition, expressed as probabilistic programs to synthesize labeled example solutions to infer feedback for real student solutions. This grading achieves a 50% improvement over previous best results. This method also relies on being pre-configured with a set of example solutions and also may suffer from scalability issues if this is not adequate.

The aim of the work presented in this paper is to investigate a more human-in-the-loop approach to AI assisted marking where the marker can provide input during critical stages of the process, providing marking feedback examples and editing generated examples as-and-when required if the generated samples fall short of expectations. This should remove scalability and associated quality issues, as the training set will be augmented as-and-when required during the marking procedure, and potential ethical issues as all marks should be vetted by the human marker. This approach addresses a research gap in current educational technology which is a lack of interactive, adaptable AI-assisted grading tools that maintain meaningful human oversight while improving scalability. Theoretically, it contributes by operationalizing HITL principles within an assessment context while considering factors that influence technology adoption, such as perceived usefulness and ease of use (Venkatesh et al., 2003).

AI-Marking Assistant

The AI-Marking Assistant (AI-MA) is an online application that assists educators in marking larger numbers of text-based assignments. The motivation of the application's design is to allow markers to benefit from the efficiency and consistency of AI marking while avoiding the disadvantages such as bias and misunderstanding that can affect the marking quality. The application also allows for human oversight to avoid biases or inaccuracies being left in the marks when they are given to students. Indeed, the system allows the marker to fine-tune the AI marking process by allowing the marker to provide the system with examples of their marking for reference.

The assignment page of the interface presents the user with a list of all the assignments or has the option to add a new assignment, or click on the name of an assignment to start marking.

Assignment Setup Interface

To start adding a new assignment, the user opens the assignment setup page, names the assignment, sets a numeric value for the maximum grade, and provides instructions for the AI. An AI instructions field defines the role of the AI marking assistant and tells it to expect a set of instructions for the student and a marking scheme. This field has a default value as follows.

> *You are a virtual teacher who should provide a mark and feedback according to the following instructions and marking scheme.*

The student instructions field tells the AI what instructions are given to the student for the assignment, and the marking scheme field provides the marking scheme.

After specifying the instructions for the AI, the user can drag and drop a CSV or MS Excel file to upload the data. They then specify the fields to identify each student and the fields for their assignment responses. This can be a single answer or a group of answers for which a single mark and paragraph of feedback are given.

Assignment Marking Interface

The assignment marking page of the AI-MA is shown in Figure 1. This is the area of the interface where users can mark assignment submissions and generate marks feedback. On the left-hand side of the page, the marker can see the name of the assignment, the marking criteria, and their progress in the marking. They can also configure the AI by setting the temperature.

The temperature set for AI defines how creative or predictable it is. When the temperature is low, the model tends to choose the most probable token (word) based on its training data. The output is more predictable and correct but lacks variation and creativity. With a high temperature, the model introduces more randomness. It considers tokens beyond the most probable ones. The generated text becomes more diverse, but there's a higher chance of grammar mistakes or nonsensical content. In general, marking tasks would require a low AI temperature, particularly in more scientific or engineering-oriented subjects where the instructions of the assignment are well defined. Higher AI temperatures would yield more creative feedback and could possibly be useful for formative tasks where feedback is used to trigger creative or original thought rather than more formal summative assessments.

The right-hand side panel shows the assignment submissions. The user can scroll up and down on this panel and expand or collapse individual student marks as and when required. After a student's mark is submitted, it is automatically collapsed to show just the student name, ID number, and mark. The marking

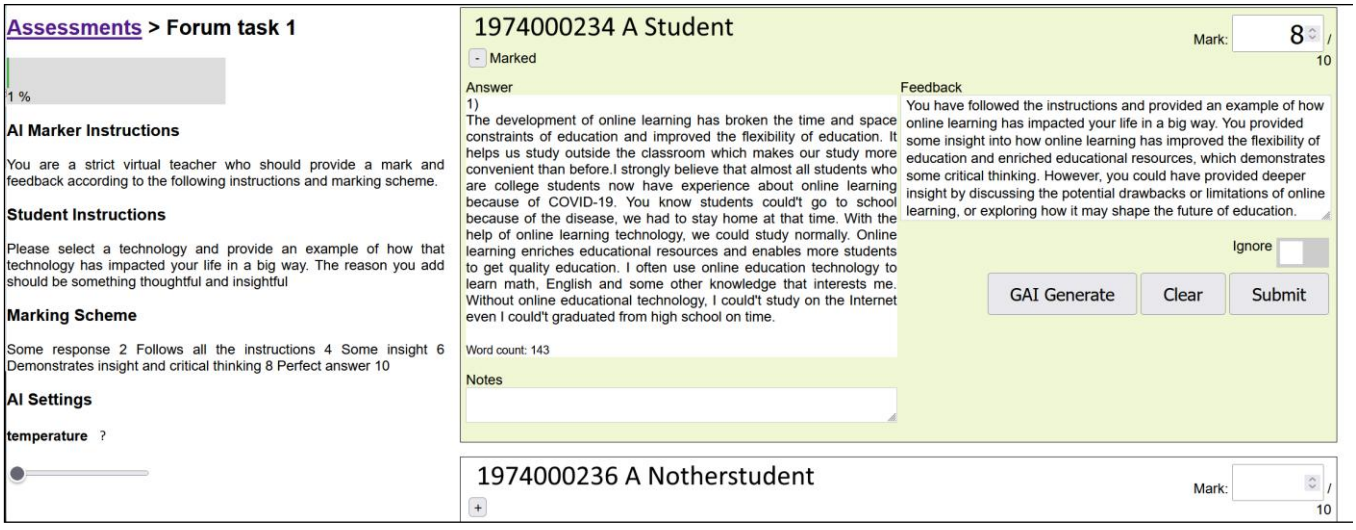

Figure 1. Assignment marking interface. The marker scrolls through student submissions of the right hand side of the screen. The Feedback and mark can be edited before being submitted to the GAI with uses it as an example for future marking.

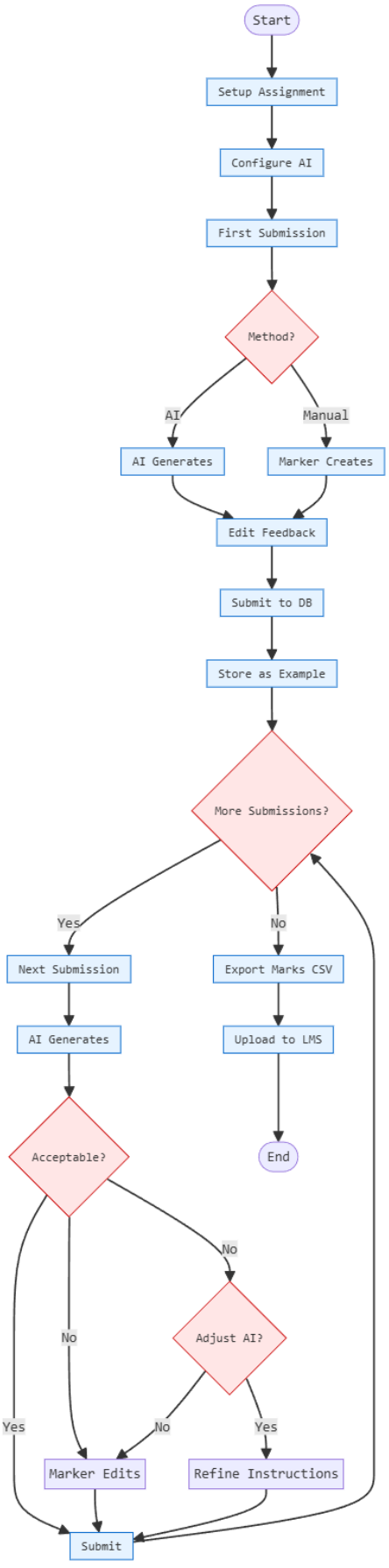

Figure 2. Workflow diagram of the AI-MA marking process

panels can be ordered either by the student's name, their ID, the mark, or the status of the marking (i.e., if a submission is marked, unmarked, or marked but not submitted).

The panels for individual student submission marks are laid out with their name and ID number at the top, the status of marking below this, the answer on the left-hand side, notes (added by the marker) below the answer, and the mark, feedback, and control buttons on the right-hand side. The notes box, feedback, and marks boxes are all editable. Clicking the ignore button indicates that the submission should be ignored. This would be the case if the student submitted their answer twice by mistake or if they have multiple chances to submit. The AI generate button uses AI to generate feedback and a mark. The clear button clears the mark and feedback. The submit button submits the mark and feedback to the database. The colour of the box will change if the marks are submitted or ignored, and a red warning will appear if feedback is typed but not submitted.

After feedback and marks are generated, they can be edited by the marker before being submitted to the database. These marks and feedback will be used, in turn, as examples to guide the AI for future marking. This allows the marker to selectively fine-tune the AI to improve its output. It also allows the marker to supervise the output of the AI and take responsibility for the marks submitted. This helps alleviate the ethical issues of automated marking, giving the human marker more input into the marking process.

Marking Workflow

The general workflow for marking student assignment submissions (see Figure 2) can be described as follows.

1. Upload the assignment submissions – The first step in marking is to specify the assignment AI instruction and upload the assignment submissions in CSV format as described in the Assignment Setup Interface above. This specifies the role of the AI, the instructions to students, and the assignment marking scheme. The data should contain one or more fields to identify each student and one or more text fields with a student's response(s) to the assignment questions or requirements.
2. Marking the first assignment submission – The next stage in marking is to proceed to the assignment marking page and begin marking with the first assignment. At this stage, it is important to understand that while the system has instructions on how to do the marking, there are still no examples of how the feedback should appear. The GAI Generate button can generate feedback, but this will not necessarily fit the marker's expectations with regard to the length of feedback, the tone of the feedback, or how much attention is paid to different marking criteria. The marker is therefore expected to edit the feedback (perhaps quite extensively) in order for it to fit

their expectations. The user will also likely need to add a mark for the first few submissions until the GAI can consistently mark according to their expectations. Once the marker is happy with the feedback and the mark, they can press submit to commit the mark to the database.

3. Marking the next assignment submissions – For the next assignments, the GAI Generate function will use submitted assignment feedback and marks as an example to generate new feedback and marks that better meet the marker's expectation. Normally these will require less editing and, as more assignments are submitted, less editing is generally required as the AI has more examples to use.
4. Refine GAI instructions (optional) – As part of the marking process, the marker can also adjust the GAI instructions. This can help adjust for the level of the AI if (if, for example, it is being too generous or strict) or account for requirements that have not been specifically stated in the AI instructions. The marker may want to tell the GAI to act as a strict marking assistant if it is too generous, or add additional criteria such as originality or insight if these are not explicitly addressed in the original marking scheme. This process can also encourage the marker to clarify points in the assignment specification or marking scheme for future assignments.
5. Export the marks – The final stage, after all the assignment submissions are marked, is to export the marks. Marks are exported in CSV format, which allows them to be uploaded into online learning platforms such as Moodle or Canvas.

System Architecture

AI-Marking Assistant's architecture is shown in Figure 3. The system is is built using PHP linked to an SQL database and the Azure Open AI API running ChatGPT 3.5. The UI makes extensive use of JavaScript for UI elements and Ajax to send and retrieve data from a server asynchronously without interfering with the display and behaviour of the existing page. The data is sent and received from ChatGPT in JSON format.

Data privacy and security measures include anonymization of student identifiers before processing through the Azure OpenAI API, encrypted data transmission, and local storage of sensitive information. The system also complies with institutional data protection policies and does not retain student submissions beyond the marking period.

Evaluation

In order to evaluate the AI-MA system, we used it for the marking of two different forum-based assignments with four different markers. These were two short forum tasks aimed at improving the student's understanding of emerging technologies, and a class test evaluating the students' knowledge of their project topic based on the contents of a report they had written. The total number of student submissions marked was 250 across two assignments, with four markers participating in the evaluation.

Before marking, each marker was given a five-minute training session to help them understand the operation of the interface. After marking, the markers participated in a structured interview where they were asked to reflect on the marking experience.

Ethical Considerations

The study was conducted in accordance with institutional ethical guidelines. Student submissions were anonymized before marking, and participants (markers) provided informed consent. The AI-MA system was deployed on a secure institutional server with access restricted to authorized markers. Student data was processed through Azure OpenAI API with appropriate data protection measures in place, including data anonymization and secure transmission protocols.

Example Assignments

Both assignments used in our evaluation were for an FHEQ level 3 (university preparation year) module introducing students to emerging technologies. The

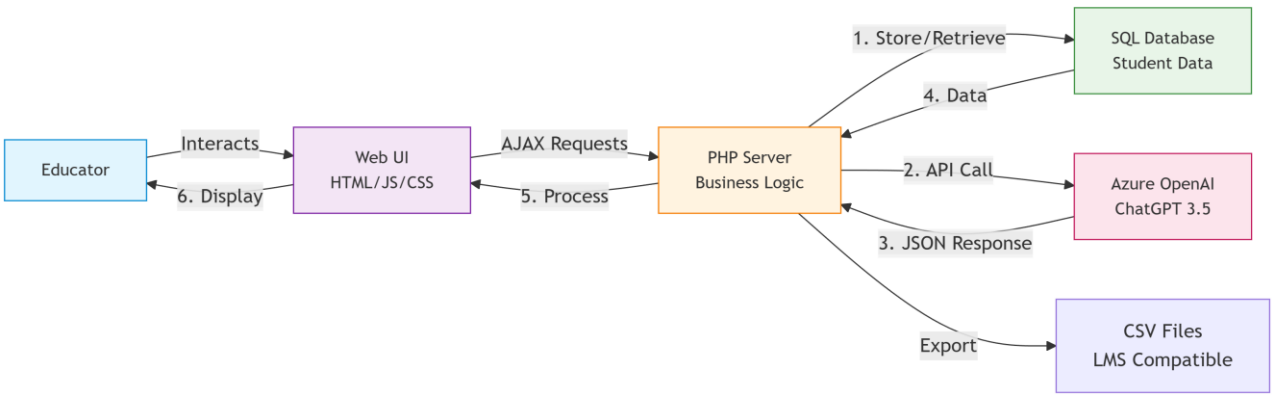

Figure 3. System architecture of AI-MA showing data flow between components

class had a total of 250 students, mostly non-native English speakers studying with English as the medium of instruction (EMI). Each assignment would contribute a small percentage (approx 5%) of the total module grade for a module with the normal credit weighting of a standard technical module.

The purpose of the first task was to introduce the topic of emerging technologies by asking them to describe a technology that had a significant impact on their life.

The instructions to students for this task are as follows.

Please select a technology and provide an example of how that technology has impacted your life in a big way. The reason you add should be something thoughtful and insightful.

The assignment marking scheme is as follows.

Some response 2, Follows all the instructions 4, Some insight 6, Demonstrates insight and critical thinking 8, Perfect answer 10.

The second marking task was also designed to help introduce the topic of emerging technologies. This time by asking them to describe an emerging technology that they expected to become the next big thing.

The instructions to students for this task are as follows.

You should identify an emerging technology you think will be big in 4 or 5 years when you graduate. Provide a single reason why you think that technology has the potential to be big. The reason you add should be something thoughtful and insightful.

The assignment marking scheme is as follows.

Some reasonable response 2, Follows all the instructions 4, Some insight 6, Demonstrates insight and critical thinking 8, Perfect answer 10.

One of the markers, the module leader, marked all submissions from each assignment. The other three markers marked a selection of twenty random submissions from each assignment. These were final year PhD students who worked as teaching assistants on the module and would also be responsible for supporting the students and answering questions for their final assignment, which would be a report outlining a proposal for a new product or service using an emerging technology. These TAs would benefit from the exercise to gauge the student's understanding of emerging technologies.

Performance

The marking performance of AI-MA is described in Table 1, with the average time taken to mark the first 10 and every subsequent 10 assignment submissions shown for AI-MA with GAI generation available and with GAI not available as a control. This shows that the performance of markers dramatically improves after the first 10 submissions marked and again significantly after the next 20 submissions marked. After this, the marker only needs a couple of seconds on average to mark each submission. This is a lot more than the normal gains in efficiency due to marker familiarity. Overall, this equates to a significant improvement in efficiency with about a 40% time saving for 60 submissions. This would likely improve to up to around 60% with larger numbers of submissions for assignments of this type.

Table 1. AI-MA Performance, Average Marking Time Per Submission (seconds) n=4 markers, measurements recorded under controlled conditions

Submissions	AI-MA	control
1-10	4.41s	4.52s
11-20	2.51s	3.87s
21-30	1.89s	3.97s
31-40	1.66s	3.66s
41-50	1.77s	3.73s
51-60	1.57s	3.64s

Statistical analysis shows a significant reduction in marking time between the first 10 submissions and subsequent batches ($p < 0.05$), with effect sizes (Cohen's d) ranging from 0.8 to 1.2 for different comparison points.

Table 2. AI-MA User Satisfaction (1 Very Negative, 3 Neutral, 5 Very Positive)

	Average Rating (1-5)	Standard Deviation
Usability	3.75	0.43
Functionality	4.75	0.43
Performance	5.0	0.0
Overall	4.5	0.17

Structured Interview

The structured interview to evaluate the application focused on the usability, functionality and performance. These questions were follows.

- Usability. Is the application easy to use? Is it east to achieve what you want to with the interface and does it operate how you would expect it to?
- Functionality. Is the functionality what you would expect from a grading assistant? Is there functionality you think is unnecessary or anything you feel needs to be added?
- Performance. What do you think of the quality of the feedback and marks suggested by the system? Does it match what you would expect from a human marker given the marking criteria supplied.

How did you have to edit the marks and feedback suggested in order to make them appropriate for the students?

- Overall. What is your overall impression of the system, would you use something like this for your marking in the future.

Each tester was scheduled a 30 minute session to answer the questions which were recorded and transcribed after the interviews were complete.

Results

The results of user satisfaction survey is shown in Table 2 with results based on analysis of 250 student submissions marked by four evaluators. Here it can be seen that users were generally very positive about the functionality and performance of the tool and less positive about the usability where they identified some minor issues. The results of our structured interviews were compiled and can be summarised as follows

Usability

The testers generally felt the usability of the application was satisfactory despite some minor issues. There was some confusion as to how to navigate to the Marking page from the Assignments page. The markers would also have preferred to be able to edit assignment AI instructions on the Assignment marking page. These could be considered as minor issues that could be easily resolved in the next iteration of development.

The import and export functions of the application had no major usability issues. This was relatively straight forward for the users and the most difficult step seemed to be exporting submissions and importing marks back into the Learning Management System used.

The usability of the marking page (where users would spend most of their time on the application) was also felt to be good and the users appreciated being able to scroll through the assignment submissions for efficient marking. Initially, the users had some difficulty knowing how to generate feedback and marks, but after they figured this out on the first submission the process became very intuitive.

Functionality

Overall the users were happy with the functionality of the system for marking short-form text based assignments. They also indicated how they would like to improve the functionality in a number of areas.

The markers would have liked to have the software include a user management system (UMS) to restrict access from other system users. As the version of the tool ran on an offline virtual server for a single module, this was not an issue. If the tool was to be made available for real life use on different modules, then a UMS would be required in order to control access and ensure student privacy.

The markers would also like to see the tool be available for different types of text based assignment including assignments with individual component marks, and assignments submitted as reports. Currently the system only supports a single mark for each submission and the length of assignment submissions is limited to around 1,200 words due to limitations of the ChatGPT API. While the first issue could be fixed by adjusting the interface to include multiple marking criteria, the second issue is a limit of the ChatGPT API and would be difficult to resolve.

Performance

The users were generally happy with the performance of the marking and thought it was to a sufficient level to justify using the tool for marking. The feedback of our markers was mostly consistent and can be summarised as follows.

- First feedback. The first time feedback generated for a particular assignment tended to be quite verbose but could be edited down to form more concise succinct feedback. After this feedback was edited and submitted it would be used as a model for future feedback.
- Subsequent feedback. After four or five examples of feedback where provided, the GAI became more reliable at providing suitable feedback. This tended to follow the same general form addressing different aspects of the marking scheme in the same order with similar language. This made the feedback easy to check to correct any issues.
- GAI Temperature. The markers kept the AI temperature low which tended to keep the GAI consistent. Raising the temperature made the text more varied but the marking less reliable.
- Tone. The tone of the feedback tended to be more positive than the testers would have preferred. In order to make it more balanced to properly account for short-comings the marking instructions needed to be edited to say 'you are a strict marking assistant'. After this there was a good balance of positive and negative as well as constructive criticism on how the answer could be improved.
- Marks. The marking was judged to be applied well and consistent with expectations. In most cases the mark matched the markers expectation. In less than 15% of cases the mark needed to be adjusted by 1 mark. In less than five percent of cases it needed to be adjusted by 2 or more marks. In these cases normally the GAI gave a mark that was too generous.
- Consistency and bias. After the initial 4 or 5 marking examples were given the marking was judged to be consistent a fair throughout the marking. Where bias was detected this could be

corrected easily be the marker. For example, after two good answers on the topic of Natural Language Processing (NLP), the GAI gave a slightly inflated mark for the next answer focusing on NLP. The marker was able to adjust the feedback and mark in a few seconds. This made the marking more reliable for subsequent answers using this topic.

- Answer length. Directions in the marking scheme related to the length of the answer would not be accounted for. For example, if a maximum word count was given or the marking scheme specified a concise answer this would be ignored.

Overall Outcome

Overall, all our test users agreed that they would be happy to use a system like the one proposed for their marking in the future. They believed that the advantages of improved efficiency and consistency would allow them to increase the amount of feedback and interaction with the students on a module to improve associated learning outcomes.

All test users agreed they would use such a system for future marking, acknowledging its potential to improve efficiency and consistency while maintaining necessary human oversight. However, they noted that further evaluation with more complex assignments and larger marker samples would strengthen these findings.

Conclusion

We have developed a web-based application for human-in-the-loop GAI assisted assessment marking and feedback and evaluated it with the assessment of two short written assignments. The interface works to help with marking and feedback for short text based answers and is evaluated to show potential to help educators to provide more consistent feedback in a more timely and efficient manner. The evaluation also demonstrates the potential to overcome limitations of fully automated marking by allowing markers to correct bias and other potential quality issues. As the marker is able to monitor and edit marks and feedback provided by the GAI and have their adjustments fed back into the GAI as marking examples, they can take responsibility for the marks and lead the GAI to be more consistant with their expectations. This corrects for GAI bias and removes the ethical problem of the GAI generating marks without supervision.

The human-in-the-loop approach implemented in AI-MA aligns with established frameworks for responsible AI in education (Medrano, 2025) as well as studies investigating human-AI partnering for other activities (Choudhury and Shamszare 2024; Lotfalian Saremi et al. 2025), balancing automation benefits with necessary human judgment. Theoretically, this work bridges HITL AI principles with assessment validity frameworks, demonstrating how AI can enhance grading consistency while preserving educator judgment. These are key factors in both assessment reliability and technology acceptance.

Our results also demonstrate some of the advantages and limitations of GAI generated feedback and marking in general. While marking with a low GAI temperature is shown to be consistent and generally reliable, it relies on precise instructions and can be overly positive. It also marks according to content without being able to judge aspects of the writing style such as conciseness. Future work will see us refine the interface and extend the functionality of the application so it can be used and evaluated with longer text based assignments. This would involve working to stretch the functionality of the ChatGPT API by chaining calls including different parts of each submission.

While the results are promising, several limitations should be acknowledged: the small sample of markers, focus on short text responses, and absence of comparison with expert human marking. Future work will extend the system's functionality for longer text-based assignments through API call chaining and explore integration with learning management systems. Additionally, research into student perceptions of AI-assisted feedback and its impact on learning outcomes will be valuable for understanding the broader implications of human-in-the-loop AI grading systems. Further theoretical exploration could examine AI-MA through the lens of comprehensive technology acceptance models (such as UTAUT) and assessment validation frameworks to strengthen its pedagogical foundations.

Acknowledgements

The research described in this paper is supported by Xi'an Jiaotong Liverpool University (XJTLU) Research Conference Fund and the XJTLU Academy of Future Education AIED research centre project AI and Learner Emotion grant number RC4AIED202401 as well as XJTLU Teaching Development Fund grant number TDF 202425S1-65 and XJTLU Academy of AI (AoA) Proof of Concept project reference number POC-25-05.

Conflict of Interest

The authors declare that there are no conflicts of interest for this paper.

References

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. International journal of artificial intelligence in education, 25(1), 60-117.

Chan, C. K. Y., & Colloton, T. (2024). Generative AI in higher education: The ChatGPT effect (p. 287). Taylor & Francis.

Choudhury, A., & Shamszare, H. (2024). The impact of performance expectancy, workload, risk, and satisfaction on trust in ChatGPT: Cross-sectional survey analysis. JMIR Human Factors, 11, e55399.

Craig, P., Roa-Seïler, N., Rosano, F. L., & Díaz, M. M. (2013, July). The role of embodied conversational agents in collaborative face to face computer supported learning games. In *Proc. 26th International Conference on System Research, Informatics & Cybernetics. Baden-Baden, Germany*.

Craig, P., Roa-Seiler, N., Diaz, M. M., & Rosano, F. L. (2014). A cognitionics approach to computer supported learning in the Mexican State of Oaxaca. *Informatica*, 38, 241.

Farrelly, Tom, and Nick Baker. "Generative artificial intelligence: Implications and considerations for higher education practice." *Education Sciences* 13.11 (2023): 1109.

Johnston, H., Wells, R. F., Shanks, E. M., Boey, T., & Parsons, B. N. (2024). Student perspectives on the use of generative artificial intelligence technologies in higher education. *International Journal for Educational Integrity*, *20*(1), 2.

Kaya, M., & Cicekli, I. (2024). A hybrid approach for automated short answer grading. *IEEE Access*, *12*, 96332-96341.

Li, M., Enkhtur, A., Yamamoto, B. A., Cheng, F., & Chen, L. (2025). Potential societal biases of ChatGPT in higher education: A scoping review. *Open Praxis*, *17*(1), 79-94.

Lotfalian Saremi, M., Ziv, I., Asan, O., & Bayrak, A. E. (2025). Trust, Workload, and Performance in Human–Artificial Intelligence Partnering: The Role of Artificial Intelligence Attributes in Solving Classification Problems. *Journal of Mechanical Design*, *147*(1), 011702.

Malik, A., Wu, M., Vasavadia, V., Song, J., Coots, M., Mitchell, J., et al. (2019). Generative grading: near human-level accuracy for automated feedback on richly structured problems. arXiv preprint arXiv:1905.09916.

Medrano, T. (2025). Co-Designing With AI: Human-Centered Approaches for. *Foundations and Frameworks for AI in Education*, 291.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, *50*(9), 741.

Øen, K., Krumsvik, R. J., & Skaar, Ø. O. (2024, January). Development of inclusive practice–the art of balancing emotional support and constructive feedback. In *Frontiers in Education* (Vol. 9, p. 1281334). Frontiers Media SA.

Simonsmeier, B. A., Peiffer, H., Flaig, M., & Schneider, M. (2020). Peer feedback improves students' academic self-concept in higher education. *Research in Higher Education*, *61*(6), 706-724.

Souppez, J. B. R., Goswami, D., & Yuen, J. (2023, December). Assessment and feedback in the generative ai era: transformative opportunities, novel assessment strategies and policies in higher education. In *International Federation of National Teaching Fellows Symposathon 2023*.

Tarun, B., Du, H., Kannan, D., & Gehringer, E. F. (2025). Human-in-the-loop systems for adaptive learning using generative AI. *arXiv preprint arXiv:2508.11062*.

Tetzlaff, L., Schmiedek, F., & Brod, G. (2021). Developing personalized education: A dynamic framework. *Educational Psychology Review*, 33(3), 863-882.

Tuma, F. (2022). Educational benefits of writing multiple-choice questions (MCQs) with evidence-based explanation. *Postgraduate Medical Journal*, *98*(1156), 77-78.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS quarterly*, 425-478.

Williams, A. (2024). Delivering Effective Student Feedback in Higher Education: An Evaluation of the Challenges and Best Practice. *International Journal of Research in Education and Science*, *10*(2), 473-501.