Kashihara, A. et al. (Eds.) (2024). Proceedings of the 32nd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

# Leveraging Generative AI for Automatic Scoring in Chemistry Education: A Web Based Approach to Assessing Conceptual Understanding of Colligative Properties

**Sri YAMTINAH[a*], Dimas Gilang RAMADHANI[b], Antuni WIYARSI[c], Hayuni Retno WIDARTI[d] & Ari Syahidul SHIDIQ[a]**

[a]*Chemistry Education, Universitas Sebelas Maret, Indonesia*
[b]*Chemistry Education, Universitas Negeri Semarang, Indonesia*
[c]*Chemistry Education, Universitas Negeri Yogyakarta, Indonesia*
[d]*Chemistry Education, Universitas Negeri Malang, Indonesia*
*jengtina@staff.uns.ac.id

**Abstract:** The integration of artificial intelligence (AI) into educational assessment offers promising advancements in automating the grading process, particularly in complex subjects like chemistry. This study focuses on implementing the Gemini 1.5 AI model to evaluate student responses in a web-based chemistry assessment. The study aimed to assess the effectiveness and accuracy of Gemini 1.5 in grading questions related to stoichiometry, a fundamental concept in chemistry. The assessment involved 320 students who answered five questions—one conceptual and four computational—focused on calculating molar quantities and applying related formulas. The AI system was utilised to evaluate the responses, providing scores based on criteria such as the correct application of formulas, calculation accuracy, and the proper use of scientific units. The study's findings indicate that Gemini 1.5 demonstrated high accuracy, with precision and recall metrics consistently ranging from 0.87 to 0.93 across the different questions. These results suggest that the AI system effectively delivered consistent and objective grading, minimising errors such as false positives and negatives. The AI's ability to provide immediate and detailed feedback highlights its potential to enhance learning by reinforcing key concepts and addressing areas where students may struggle. The conclusion drawn from this study is that integrating Gemini 1.5 into the educational assessment process improves grading efficiency and supports personalised learning by offering tailored feedback. This integration has significant implications for reducing the workload on educators while ensuring fair and accurate assessments, ultimately contributing to a more effective educational experience for students.

**Keywords:** Artificial Intelligence in Education, Automated Grading, Chemistry Assessment, Gemini 1.5 AI Model, Stoichiometry Evaluation

## 1. Introduction

Assessment is crucial in chemistry education, particularly for complex topics like colligative properties. However, traditional assessment methods often struggle with subjectivity, time consumption, and inconsistencies, which can negatively impact the quality and fairness of grading. Integrating generative AI for automatic scoring can enhance this process by providing more precise and consistent evaluation(Analita, 2023). Research indicates that AI-driven scoring systems reduce the manual workload, improve consistency, and deliver immediate feedback, essential for student learning (Fu et al., 2020; Horbach & Zesch, 2019). Additionally, AI-enabled scoring supports continuous learning by promptly addressing students' needs. It offers significant time and cost savings in developing scoring algorithms, making these tools more accessible in educational settings(Xuansheng et al., 2023). The increasing use of Automated Essay Scoring (AES) tools, driven by AI advancements, further highlights the potential of these technologies to streamline and enhance the assessment of complex chemistry concepts (Hahn et al., 2021; Zhai, 2021).

Manual grading in chemistry education often presents challenges such as subjectivity, leading to inconsistencies and potentially unfair evaluations, as different graders may interpret student responses differently (Huang, 2024). Manual grading is time-consuming, burdens educators, and delays feedback crucial for student learning and progress (Zhai et al., 2022). These delays can hinder students' comprehension of intricate concepts like colligative properties. This topic integrates multiple foundational chemistry principles, making it ideal for evaluating the effectiveness of AI-driven assessments. This delay can hinder students' understanding of complex concepts and is exacerbated by the risk of errors in repetitive data entry and calculations, further affecting the accuracy of grades (Yin et al., 2022). Intelligent grading methods based on machine learning algorithms have been proposed to address these issues, offering solutions to enhance grading efficiency and reliability by minimising subjectivity and providing timely feedback(Hiremath, 2024). Automated grading streamlines assessment processes and improves the overall learning experience by ensuring fair and consistent evaluations, particularly for challenging topics like colligative properties in chemistry.

Generative AI offers a promising solution for automating the assessment process in education by improving efficiency, consistency, and accuracy in grading student performance (Harry, 2023). By utilising AI algorithms, educators can streamline grading, provide timely and personalised feedback, and ensure objective evaluations, thus overcoming the subjectivity and time-intensive nature of manual grading (Huang, 2024). Specifically, using AI to grade complex chemistry topics like colligative properties can transform traditional educational practices by reducing grading errors and offering real-time insights into student performance, helping bridge the gap between teaching and learning outcomes. AI-enabled systems also enhance the learning experience by offering individualised learning strategies and saving educators' time, allowing them to focus on more personalised instruction (Hiremath, 2024; Palmer, 2023). Integrating AI in assessment aligns with current technological advancements in education, as tools like automated grading systems are increasingly used to deliver consistent and reliable evaluations, ultimately enhancing educational outcomes (Irham et al., 2017; Popenici & Kerr, 2017).

This study uniquely addresses a critical gap in the literature by focusing on the application of generative AI, specifically in chemistry education. Automated grading has been less explored in this field, particularly for complex, concept-heavy topics like colligative properties. By demonstrating how the Gemini 1.5 AI model can provide accurate, real-time evaluations and constructive feedback, this study highlights the transformative potential of AI in enhancing both teaching efficiency and student learning experiences. Unlike previous research that broadly addresses AI in education, this study specifically tackles the practical challenges of manual gradings—such as subjectivity and time inefficiency—by demonstrating how AI can enhance traditional educational practices, offering a transformative approach to chemistry pedagogy.

## 2. Method

### *2.1 System Design*

The methodology begins by integrating the Gemini 1.5 generative AI model into a web-based platform for seamless student interaction. This platform allows students to log in and submit their answers to chemistry questions directly online. These questions focus specifically on colligative properties, a complex area in chemistry that requires a deep understanding of various underlying principles. The platform was developed with a user-friendly interface to ensure students can easily navigate the system and submit their responses without technical difficulties. Once a student submits a response, the system captures and processes the input immediately, ensuring that the transition from submission to evaluation is efficient and secure.

The processing of these inputs involves several stages. First, the submitted text is analyzed for relevance to the question, checking for critical concepts and accurate scientific reasoning. The Gemini 1.5 AI model, trained on a vast dataset of chemistry-related texts, is then used to evaluate the depth of conceptual understanding demonstrated in the response. The AI model applies advanced natural language processing (NLP) techniques to assess the student’s answer’s logical coherence, accuracy, and completeness. The model uses machine learning algorithms optimized for chemistry content, enhancing its ability to recognize context-specific terms and concepts. Based on this analysis, the system generates a score reflecting the student’s grasp of colligative properties (see Table 1) and constructive feedback to guide further learning. This entire process is automated, enabling immediate scoring and feedback, which is crucial for reinforcing learning and providing timely student support.

Table1. *Instruments on Automated Assessment Application*

| No | Question | Concept Tested | Expected Response |
|---|---|---|---|
| **1** | What is meant by the freezing point of a solution? | Conceptual Understanding of Freezing Point | Definition of freezing point in solutions. |
| **2. a** | What is the freezing point depression of the ice cream solution made by the chef if he mixes 250 grams of sugar (Molar mass of sugar = 342 g/mol) into 500 grams of water? | Freezing Point Depression Calculation | Calculation of freezing point depression. |
| **2. b** | Calculate the molality of the solution. | Molality Calculation | Correct calculation of molality. |
| **2. c** | Determine the freezing point of the ice cream solution. | Freezing Point Determination | Calculation of the freezing point. |
| **2.d** | Discuss how the presence of sugar affects the freezing point of the solution compared to pure water. | Conceptual Understanding of Colligative Properties | Explanation of the effect of sugar on freezing point. |

### *2.2 Data Collection*

The data collection process involved gathering responses from 320 students who participated in an assessment focused on colligative properties, a key topic in chemistry. The students were presented with five questions to evaluate their understanding from multiple perspectives. The selection of participants aimed to ensure a diverse representation of skill levels, enhancing the generalizability of the results. The first question was conceptual, requiring students to explain fundamental principles related to colligative properties, such as the definition and significance of freezing point depression in solutions—this question aimed to assess their theoretical understanding and ability to articulate complex scientific concepts clearly.

The remaining four questions were computational, each addressing different aspects of colligative properties, such as calculating freezing point depression, determining molality, and applying relevant formulas to solve solute and solvent interaction problems. These questions evaluated students’ ability to apply theoretical knowledge to practical scenarios, testing their problem-solving skills and understanding of the mathematical relationships underlying colligative properties. As a result, the dataset comprised a diverse range of responses, providing a comprehensive view of each student’s conceptual understanding and computational proficiency in this area of chemistry.

### *2.3 Assessment Procedure*

The assessment begins once a student submits their response on the web-based platform. Each response is immediately processed by the Gemini 1.5 generative AI model, specifically trained on chemistry-related content to evaluate the student’s understanding of colligative properties. The AI model first analyses the response for critical concepts, logical coherence, and the accuracy of scientific reasoning. For computational questions, it also checks the correctness of the calculations and the application of relevant formulas. To ensure fairness, the AI model incorporates multiple levels of validation, including cross-referencing answers with an extensive database of correct responses. The AI then generates a score based on predefined criteria—such as the completeness and depth of the response and the accuracy of computations.

These were compared with manual scores provided by expert chemistry educators to ensure the reliability and validity of the AI-generated scores. These educators independently graded a representative sample of student responses using the same criteria as the AI model. The manual scores were then compared to the AI-generated scores to assess the consistency and accuracy of the AI's performance. This comparison involved calculating metrics like precision, recall, and Cohen's Kappa to quantify the level of agreement between the AI and human evaluators. A rigorous statistical analysis confirmed that the AI's scoring was accurate and replicable across different test conditions, enhancing confidence in the AI's reliability. The close alignment of scores confirmed the AI model's ability to accurately and consistently assess student understanding, validating its effectiveness as a reliable tool for automated assessment in chemistry learning.

.

### *2.4 Evaluation Metrics*

To evaluate the AI-driven assessment system's accuracy and reliability, key metrics like Precision, Recall, F1-Score, Confusion Matrix, and Cohen's Kappa were used. Precision measures how well the AI correctly identifies positive results, while Recall assesses its ability to capture all actual positives, both crucial for student assessments. The F1-Score balances precision and recall, ensuring a comprehensive performance view. The Confusion Matrix breaks down true and false results, highlighting areas for improvement, and Cohen's Kappa compares AI scores with human grading to assess agreement. Together, these metrics ensure the AI aligns closely with human grading standards, making it a reliable tool for automated assessments in chemistry education.

## 3. Result

### *3.1 Performance Metrics*

The performance metrics for the AI model, as summarised in the provided data in Table 2, indicate a strong overall performance in assessing student responses to questions on colligative properties. The metrics include Precision, Recall, F1-Score, and Cohen's Kappa for each of the five questions (1, 2. A, 2. B, 2. C, 2.d), providing a comprehensive view of how well the AI system is functioning across different types of questions.

For Question 1, which focuses on conceptual understanding, the AI achieved a Precision of 0.87, a Recall of 0.89, and an F1-Score of 0.88. These figures reflect the model's high accuracy in identifying correct responses and its ability to capture nearly all correct answers. Cohen's Kappa value of 0.85 further indicates a strong agreement between the AI-generated scores and those of human educators, underscoring the model's reliability in evaluating more abstract, conceptual questions. This suggests that, despite slight variability, the AI is proficient at understanding and interpreting student responses that require critical thinking and explanation, highlighting its potential to handle complex educational assessments effectively.

Table 2. *Performance Metrics Automatic Scoring Application*

| Question | Precision | Recall | F1-Score | Cohen's Kappa |
|---|---|---|---|---|
| 1 | 0.87 | 0.89 | 0.88 | 0.85 |
| 2. a | 0.92 | 0.93 | 0.92 | 0.88 |
| 2. b | 0.90 | 0.92 | 0.91 | 0.89 |
| 2. c | 0.92 | 0.92 | 0.92 | 0.88 |
| 2.d | 0.92 | 0.90 | 0.91 | 0.89 |

The AI model performs even more in the computational questions (2. a to 2.d). For example, in Question 2. a, the AI achieved a Precision of 0.92, a Recall of 0.93, and an F1-Score of 0.92, with a Cohen's Kappa of 0.88. The consistently high scores across these questions demonstrate the AI's exceptional capability in evaluating structured, formula-based problems, confirming its effectiveness in accurately assessing areas where criteria for correct answers are well-defined. Similar high metrics are observed across the other computational questions, with Precision and Recall consistently around 0.90 to 0.92 and Cohen's Kappa values close to 0.89.

These results highlight the AI's exceptional capability in accurately evaluating structured, formula-based problems, where the criteria for correct answers are clearly defined. The high F1 scores indicate a balanced performance in capturing correct answers and minimising false positives. At the same time, the strong Cohen’s Kappa values reflect a high level of agreement with human grading, confirming the AI's effectiveness in delivering consistent and reliable assessments. Such metrics underscore the model's reliability, not just as an automated tool but as a robust system capable of replicating expert-level grading across varied question types.

The data suggest that the AI model is particularly well-suited for computational assessments, where its precision and recall are nearly optimal, and it can reliably replicate the judgment of human educators. However, the slightly lower metrics in the conceptual question point to areas where refinement is needed, particularly in enhancing the AI's capacity to understand more nuanced, open-ended responses. Improving these capabilities could make the AI an even more valuable tool for comprehensive educational assessment, ensuring it can handle conceptual and computational tasks equally proficiently.

### *3.2 Confusion Matrices*

The confusion matrices in Figure 1 offer a detailed analysis of the Gemini 1.5 AI model’s performance in evaluating student responses to five questions on colligative properties. The AI’s effectiveness is assessed through true positives, false positives, and false negatives, providing a clear picture of its accuracy and reliability. For Question 1, which focuses on conceptual understanding, the AI correctly identified 131 true positives and 154 true negatives but also made 19 false positives and 16 false negatives. These results indicate that while the AI performs well overall, it encounters challenges with more abstract, less structured responses, common in conceptual questions.

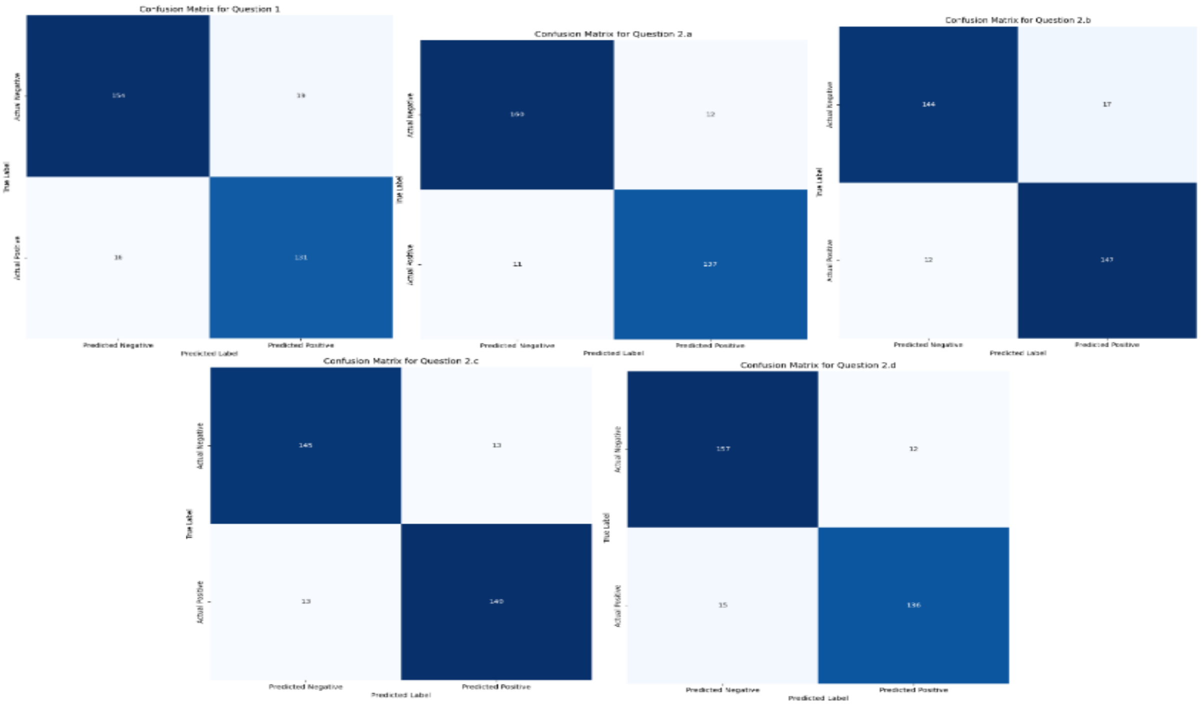


*Figure 1.* Confusion Matric Automatic Scoring Application

In contrast, the AI model performs significantly more on the computational questions (2. a to 2.d). For instance, in Question 2. a, the AI achieved 137 true positives and 160 true negatives, with only 12 false positives and 11 false negatives. Similar patterns are observed across the other computational questions, where the AI consistently delivers high true positive and true negative rates with minimal errors. This indicates the AI's proficiency in handling clear, rule-based assessments, validating its utility in grading formula-driven questions accurately and efficiently.

The key takeaway from these results is that the Gemini 1.5 AI model is particularly effective for automating the grading of computational tasks in chemistry education. Its ability to deliver precise and reliable assessments in these areas makes it a valuable tool for educators. It allows them to offload the time-consuming task of grading calculations and instead focus on more complex instructional activities. However, the slightly higher error rates in the conceptual question highlight an area for improvement, specifically in developing the AI’s capacity for nuanced interpretation and flexible assessment strategies beyond formulaic correctness. Enhancing the AI’s ability to understand and evaluate more abstract, open-ended responses would further increase its utility, ensuring it can handle both computational and conceptual assessments with equal proficiency. Such balanced capabilities would extend the AI’s functionality and solidify its role as a comprehensive support system in educational settings, capable of addressing various assessment needs.

## 4. Discussion

### *4.1. Evaluation of System Performance*

Generative AI has emerged as a transformative solution for automating the assessment process in education, particularly in scoring student responses with impressive precision and recall metrics. For instance, systems like Gemini 1.5 have demonstrated precision ranging from 0.87 to 0.92 and recall from 0.89 to 0.93 across various question types (Zhai et al., n.d.; Zheng et al., 2020). These consistent metrics across conceptual and computational questions highlight the AI's robust capability to handle diverse assessment types effectively. The high precision rates indicate that the AI rarely misidentifies incorrect answers as correct. In contrast, high recall values demonstrate thoroughness in capturing all correct responses, ensuring a comprehensive evaluation of student performance(Korkmaz & Correia, 2019).

The ability of AI models like Gemini 1.5 to deliver accurate and consistent scoring is particularly valuable in educational settings where large volumes of student assessments must be processed quickly and fairly. Unlike human graders, whose performance can be influenced by fatigue, bias, or inconsistency, the AI maintains a uniform standard of evaluation, which is crucial for maintaining academic integrity and fairness. By integrating generative AI models such as Gemini 1.5, educators can significantly enhance the grading process's efficiency, reliability, and consistency, providing timely feedback that enriches the learning experience. Immediate feedback enables students to quickly understand their mistakes and make corrections, which is particularly beneficial in iterative learning environments where continuous improvement is emphasised. This integration represents a significant advancement in assessment methodologies, enabling more personalised learning experiences and precise evaluations tailored to a wide range of question types (Bewersdorff et al., 2023).

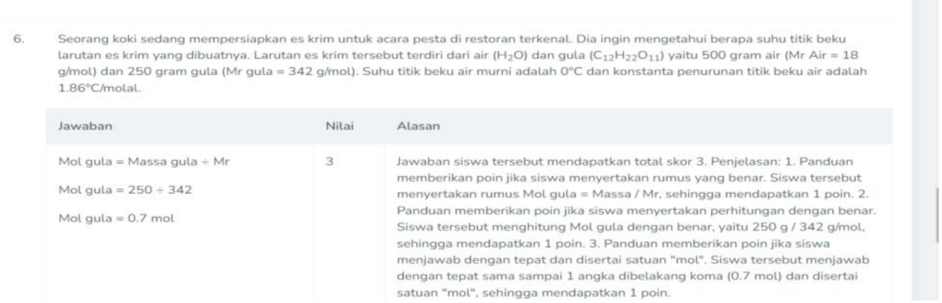
Additionally, AI’s integration into the grading process allows for more detailed analytics on student performance at individual and class levels. This data can help educators identify common areas of difficulty, inform targeted instructional strategies, and enable more effective interventions. For example, suppose AI data shows many students struggle with a specific problem. In that case, educators can adjust their teaching approach or provide additional resources to address these gaps(Klauschen et al., 2018; Ofer et al., 2021). These findings align with broader research on AI's role in education, which shows that AI-driven platforms like Gemini 1.5 can significantly improve student engagement and learning outcomes by adapting to individual learning styles and providing real-time feedback (Maestrales et al., 2021). The strong positive correlations between attitudes toward AI and its perceived impact on education further validate the potential of systems like Gemini 1.5 to revolutionise educational practices.

By automating routine grading tasks, AI allows educators to shift their focus from administrative duties to more impactful teaching activities, such as one-on-one tutoring, developing interactive lessons, or engaging in professional development. This shift enhances teaching quality and promotes a more dynamic and engaging learning environment for students. By automating routine grading tasks and offering precise evaluations, Gemini 1.5 not only alleviates the workload on educators but also ensures that students receive the timely and accurate feedback necessary for their academic growth(Swindell, 2024). However, while AI's integration into education offers numerous benefits, it also presents challenges that must be addressed to maximise its positive impact.

One critical aspect is the ethical implications of AI use in education, such as concerns about data privacy, the potential for algorithmic bias, and the need for transparency in AI decision-making processes. There is a risk that AI models trained on biased data could perpetuate or even amplify existing disparities, particularly for underrepresented or marginalised student groups. Therefore, AI systems like Gemini 1.5 must be continuously monitored, updated, and trained on diverse datasets to ensure fairness and accuracy across all student demographics. As AI systems like Gemini 1.5 continue to be integrated into educational settings, it is crucial to consider the ethical implications and ensure these technologies are implemented to maximise their benefits while mitigating potential risks (Rane, 2024). Educators and administrators must be adequately trained to understand AI's capabilities and limitations so that they can use these tools effectively and responsibly in the classroom.

### *4.2 AI Performance in Grading Chemistry Problems*

In analysing student responses to chemistry problems, the AI system exhibited a strong ability to evaluate each response's correctness and completeness accurately. The AI's structured grading approach, which includes evaluating key aspects such as correct formula application, computational accuracy, and proper scientific unit usage, ensures a comprehensive assessment aligned with academic standards. This systematic approach minimises the likelihood of grading inconsistencies that can occur with human evaluators, particularly in repetitive or large-scale assessments(Artrith et al., 2021; Kolachalama, 2018). Based on Figure 2, when tasked with grading a student's solution to a stoichiometric problem involving the calculation of the moles of sugar in a solution, the AI systematically assessed the response based on key criteria: correct application of the formula, accurate calculation, and proper use of scientific units.



6. Seorang koki sedang mempersiapkan es krim untuk acara pesta di restoran terkenal. Dia ingin mengetahui berapa suhu titik beku larutan es krim yang dibuatnya. Larutan es krim tersebut terdiri dari air ($H_2O$) dan gula ($C_{12}H_{22}O_{11}$) yaitu 500 gram air (Mr Air = 18 g/mol) dan 250 gram gula (Mr gula = 342 g/mol). Suhu titik beku air murni adalah 0°C dan konstanta penurunan titik beku air adalah 1.86°C/molal.

| Jawaban | Nilai | Alasan |
|---|---|---|
| Mol gula = Massa gula ÷ Mr<br>Mol gula = 250 ÷ 342<br>Mol gula = 0.7 mol | 3 | Jawaban siswa tersebut mendapatkan total skor 3. Penjelasan: 1. Panduan memberikan poin jika siswa menyertakan rumus yang benar. Siswa tersebut menyertakan rumus Mol gula = Massa / Mr, sehingga mendapatkan 1 poin. 2. Panduan memberikan poin jika siswa menyertakan perhitungan dengan benar. Siswa tersebut menghitung Mol gula dengan benar, yaitu 250 g / 342 g/mol, sehingga mendapatkan 1 poin. 3. Panduan memberikan poin jika siswa menjawab dengan tepat dan disertai satuan "mol". Siswa tersebut menjawab dengan tepat sama sampai 1 angka dibelakang koma (0.7 mol) dan disertai satuan "mol", sehingga mendapatkan 1 poin. |

*Figure 2.* how the Automatic Scoring Application automatically analyses answers.

Figure 2 visually represents the AI's grading process, highlighting how it breaks down complex student responses into assessable components. This detailed evaluation method allows the AI to score the response accurately and provide constructive feedback on specific areas of improvement, which can be instrumental in enhancing student understanding. The figure displays a student's response to a chemistry problem that involves calculating the moles of sugar in a solution. The student correctly applied the formula "Mol gula = Massa gula ÷ Mr," calculating the molar mass of sugar (342 g/mol) and the mass of sugar (250 g) to determine that the number of moles is 0.7 mol. The AI system accurately evaluated this response, awarding a total score of 3 points. The points were assigned based on specific criteria: 1 point for correctly stating the formula, 1 point for performing the calculation accurately, and 1 point for including the proper scientific unit ("mol") in the final answer.

This example demonstrates the AI's proficiency in applying a consistent grading rubric that identifies errors and reinforces correct methods and practices. By providing feedback on why certain steps were correct or incorrect, the AI supports a learning environment where students can actively engage with their mistakes and correct them, leading to deeper learning and retention of material(Ramadhani, 2023). The AI's analysis of the student's work demonstrates a thorough understanding of the required steps and the importance of correctly applying scientific principles and units. The accuracy and fairness of the AI's grading are evident from the detailed feedback, which aligns with the expected scoring guidelines (Lodge, 2023; Rane, 2024). The detailed nature of the feedback is particularly valuable in subjects like chemistry, where understanding procedural nuances and the correct application of scientific methods are critical to mastering the subject(Pearl & Boone, 2013; Saltz et al., 2019).

When related to the Gemini 1.5 AI model, the performance highlighted in the image showcases how generative AI, like Gemini 1.5, excels in evaluating student responses by systematically checking for key components such as formula application, calculation accuracy, and unit usage. The high precision and recall metrics observed in the Gemini 1.5 model ensure that student responses are assessed with minimal errors, providing reliable and consistent grading. This reliability is especially important in STEM education, where the consequences of grading errors can significantly impact students' understanding of fundamental concepts and their confidence in the subject matter. AI, like Gemini 1.5 in educational contexts, streamlines the grading process and enhances the quality of feedback, contributing to more personalised and effective learning experiences (Lodge, 2023; Zhai, 2021).

By incorporating AI-driven assessments, educational institutions can develop more adaptive and student-centred learning environments where feedback is immediate and highly relevant to individual student needs. This approach supports a cycle of continuous improvement, helping students build confidence in their abilities while providing educators with the insights needed to tailor their instructional strategies effectively. By leveraging AI's strengths in these areas, educational institutions can offer more tailored and impactful learning environments, better-supporting student success in complex subjects like chemistry.

## 5. Conclusion

The integration of Gemini 1.5 AI in educational assessments demonstrates significant potential, particularly in STEM education, where traditional grading methods often fall short. The system provides reliable assessments by accurately evaluating student responses, applying relevant formulas, and ensuring the correct use of scientific units. Its ability to deliver immediate feedback enhances student learning by allowing them to quickly identify and address mistakes, promoting a deeper understanding of the material. Additionally, Gemini 1.5 reduces the workload for educators while maintaining fair and uniform assessments. Although the system currently faces challenges in handling open-ended conceptual responses, further development could broaden its capabilities, making it an even more valuable tool for improving educational efficiency and supporting both educators and students in STEM contexts.

## Acknowledgments

We would like to acknowledge the financial support provided by the Research and Community Service Institute (LPPM) of Universitas Sebelas Maret (UNS) through the Indonesia Collaboration Research (RKI) grant, contract number 285/UN27.22/PT.01.03/2024. This support has been essential in enabling the successful completion of our research

## References

Analita, R. N. (2023). The Learners' Conceptual Understanding: Literature Review of Vapor-Pressure Lowering and Boiling-Point Elevation. *Journal of Education and Learning (Edulearn)*, *17*(4), 641–651. https://doi.org/10.11591/edulearn.v17i4.20805

Artrith, N., Butler, K. T., Coudert, F. X., Han, S., Isayev, O., Jain, A., & Walsh, A. (2021). Best practices in machine learning for chemistry. In *Nature Chemistry* (Vol. 13, Issue 6, pp. 505–508). Nature Research. https://doi.org/10.1038/s41557-021-00716-z

Bewersdorff, A., Zhai, X., Roberts, J., & Nerdel, C. (2023). Myths, mis- and preconceptions of artificial intelligence: A review of the literature. *Computers and Education: Artificial Intelligence*, *4*, 100143. https://doi.org/10.1016/j.caeai.2023.100143

Fu, S., Gu, H., & Yang, B. (2020). The Affordances of AI-enabled Automatic Scoring Applications on Learners' Continuous Learning Intention: An Empirical Study in China. *British Journal of Educational Technology*, *51*(5), 1674–1692. https://doi.org/10.1111/bjet.12995

Hahn, M. G., Navarro, S. M. B., de-la-Fuente-Valentín, L., & Burgos, D. (2021). A Systematic Review of the Effects of Automatic Scoring and Automatic Feedback in Educational Settings. *Ieee Access*, *9*, 108190–108198. https://doi.org/10.1109/access.2021.3100890

Harry, A. (2023). Role of AI in Education. *Interdiciplinary Journal and Hummanity (Injurity)*, *2*(3), 260–268. https://doi.org/10.58631/injurity.v2i3.52

Hiremath, A. (2024). *Transforming Handwritten Answer Assessment: A Multi-Modal Approach Combining Text Detection, Handwriting Recognition, and Language Models*. https://doi.org/10.21203/rs.3.rs-4301899/v1

Horbach, A., & Zesch, T. (2019). The Influence of Variance in Learner Answers on Automatic Content Scoring. *Frontiers in Education*, *4*. https://doi.org/10.3389/feduc.2019.00028

Huang, L. (2024). Research on the Application of Intelligent Grading Method Based on Improved ML Algorithm in Sustainable English Education. *Scalable Computing Practice and Experience*, *25*(1), 451–463. https://doi.org/10.12694/scpe.v25i1.2333

Irham, S. M., Mawardi, M., & Oktavia, B. (2017). *The Development of Guided Inquiry-Based Worksheet on Colligative Properties of Solution for Chemistry Learning*. https://doi.org/10.2991/icmsed-16.2017.9

Klauschen, F., Müller, K. R., Binder, A., Bockmayr, M., Hägele, M., Seegerer, P., Wienert, S., Pruneri, G., de Maria, S., Badve, S., Michiels, S., Nielsen, T. O., Adams, S., Savas, P., Symmans, F., Willis, S., Gruosso, T., Park, M., Haibe-Kains, B., … Denkert, C. (2018). Scoring of tumor-infiltrating lymphocytes: From visual estimation to machine learning. In *Seminars in Cancer Biology* (Vol. 52, pp. 151–157). Academic Press. https://doi.org/10.1016/j.semcancer.2018.07.001

Kolachalama, V. B. (2018). Machine learning and medical education. *Npj Digital Medicine*, *July*, 2–4. https://doi.org/10.1038/s41746-018-0061-1

Korkmaz, C., & Correia, A. P. (2019). A review of research on machine learning in educational technology. *Educational Media International*, *56*(3), 250–267. https://doi.org/10.1080/09523987.2019.1669875

Lodge, J. (2023). Adapting Assessment for/Despite Generative Artificial Intelligence. *Ascilite Publications*. https://doi.org/10.14742/apubs.2023.554

Maestrales, S., Zhai, X., Touitou, I., Baker, Q., Schneider, B., & Krajcik, J. (2021). Using Machine Learning to Score Multi-Dimensional Assessments of Chemistry and Physics. *Journal of Science Education and Technology*, *30*(2), 239–254. https://doi.org/10.1007/s10956-020-09895-9

Ofer, D., Brandes, N., & Linial, M. (2021). The Language of Proteins: NLP, Machine Learning &Amp; Protein Sequences. *Computational and Structural Biotechnology Journal*. https://doi.org/10.1016/j.csbj.2021.03.022

Palmer, E. (2023). Findings From a Survey Looking at Attitudes Towards AI and Its Use in Teaching, Learning and Research. *Ascilite Publications*. https://doi.org/10.14742/apubs.2023.537

Pearl, D., & Boone, W. J. (2013). *Assessing Scientific Practices Using Machine-Learning Methods : How Closely Do They Match Clinical Interview Performance ?* https://doi.org/10.1007/s10956-013-9461-9

Popenici, S., & Kerr, S. (2017). Exploring the Impact of Artificial Intelligence on Teaching and Learning in Higher Education. *Research and Practice in Technology Enhanced Learning*, *12*(1). https://doi.org/10.1186/s41039-017-0062-8

Ramadhani, D. G. (2023). Analysis of the Relationship Between Students' Argumentation and Chemical Representational Ability: A Case Study of Hybrid Learning Oriented in the

Environmental Chemistry Course. *Chemistry Teacher International*. https://doi.org/10.1515/cti-2023-0047

Rane, N. (2024). Enhancing the Quality of Teaching and Learning Through ChatGPT and Similar Large Language Models: Challenges, Future Prospects, and Ethical Considerations in Education. *Tesol and Technology Studies*, *5*(1), 1–6. https://doi.org/10.48185/tts.v5i1.1000

Saltz, J., Skirpan, M., Fiesler, C., Gorelick, M., Yeh, T. O. M., Heckman, R., Dewar, N., & Beard, N. (2019). *Integrating Ethics within Machine-learning Courses*. *19*(4).

Swindell, A. (2024). Against Artificial Education: Towards an Ethical Framework for Generative Artificial Intelligence (AI) Use in Education. *Online Learning*, *28*(2). https://doi.org/10.24059/olj.v28i2.4438

Xuansheng, W., He, X., Li, T., Liu, N., & Zhai, X. (2023). *Matching Exemplar as Next Sentence Prediction (MeNSP): Zero-Shot Prompt Learning for Automatic Scoring in Science Education*. https://doi.org/10.48550/arxiv.2301.08771

Yin, Y., Khaleghi, S., Hadad, R., & Zhai, X. (2022). Developing effective and accessible activities to improve and assess computational thinking and engineering learning. *Educational Technology Research and Development*, *70*(3), 951–988. https://doi.org/10.1007/s11423-022-10097-w

Zhai, X. (2021). Practices and Theories: How Can Machine Learning Assist in Innovative Assessment Practices in Science Education. *Journal of Science Education and Technology*, *30*(2), 139–149. https://doi.org/10.1007/s10956-021-09901-8

Zhai, X., He, P., & Krajcik, J. (2022). Applying machine learning to automatically assess scientific models. *Journal of Research in Science Teaching*, *59*(10), 1765–1794. https://doi.org/10.1002/tea.21773

Zhai, X., Nyaaba, M., & Ma, W. (n.d.). *Can AI Outperform Humans on Cognitive-demanding Tasks in Science?*

Zheng, Y., Bao, H., Shen, J., & Zhai, X. (2020). Investigating Sequence Patterns of Collaborative Problem-Solving Behavior in Online Collaborative Discussion Activity. *Sustainability*. https://doi.org/10.3390/su12208522