Volume: 11 Issues: 80 Special Issue [January, 2026] pp. 117 – 130
Journal of Islamic, Social, Economics and Development (JISED)
eISSN: 0128-1755
Journal website: academicinspired.com/jised
DOI: 10.55573/JISED.118009

USING AI TO ANALYZE STUDENT SCIENTIFIC WRITING IN CHEMISTRY (CER FRAMEWORK)

Siti Hajar Abd Hamid [1*]
Norah Md Noor [2]
Noor Dayana Abd Halim [3]
Azelina Abdul Rahman [4]

[1] Faculty of Social Sciences and Humanities, Universiti Teknologi Malaysia, Johor Bahru, Malaysia
(E-mail: siti83@graduate.utm.my)

Article history
Received date : 22-11-2025
Revised date : 23-11-2025
Accepted date : 28-12-2025
Published date : 15-1-2026

To cite this document:
Abd Hamid, S. H., Md Noor, N., Abd Halim, N. D., & Abdul Rahman, A. (2026). Using AI to analyze student scientific writing in chemistry (CER framework). *Journal of Islamic, Social, Economics and Development (JISED),* 11 (80), 117 – 130.

Abstract: *The rapid development of artificial intelligence technology has opened up new opportunities in educational assessment methods. This study examines the use of artificial intelligence in analysing chemistry students' scientific writing based on the Claim-Evidence-Reasoning (CER) framework. Using the Nominal Group Technique (NGT) with five expert assessors, ten implementation strategies were evaluated in terms of their suitability and priority. The analysis results showed a high level of agreement (86.67%–100%) with four strategies achieving full consensus: developing a chemistry-specific AI rubric aligned with the CER framework, integrating a chemistry misunderstanding database into AI training, providing structured step-by-step CER feedback and ensuring a safety and ethics mechanism for chemistry content. These findings highlight the need for domain-specific adaptation, human–AI collaboration, multimodal analysis capabilities and robust error-checking systems to ensure the validity and reliability of AI assessments of students' chemistry arguments. Overall, this study offers evidence-based guidelines for responsible AI implementation in chemistry education, thereby enhancing the ability to provide high-quality, scalable scientific writing feedback.*

Keywords: *Artificial Intelligence, Chemistry Education, CER Framework, Scientific Argumentation, AI-Assisted Assessment, Educational Technology, Expert Validation*

Volume: 11 Issues: 80 Special Issue [January, 2026] pp. 117 – 130
Journal of Islamic, Social, Economics and Development (JISED)
eISSN: 0128-1755
Journal website: academicinspired.com/jised
DOI: 10.55573/JISED.118009

Introduction

The landscape of artificial intelligence (AI) in education has changed significantly since the public release of ChatGPT in 2022 and the emergence of other generative AI systems. These developments have greatly enhanced the technology's capacity to analyze student work and learning processes (Kasneci et al., 2023). Powered by large-scale language models using transformer architectures, generative AI demonstrates exceptional abilities in understanding language, generating coherent responses and performing complex reasoning tasks. These capabilities far surpass those of earlier automated assessment systems, which were largely limited to rule-based scoring and surface-level pattern recognition (Chin & Brown, 2000). Unlike early AI applications that relied mainly on statistical correlations, contemporary generative AI systems can engage in contextual interaction, explain complex concepts and produce human-like written responses that are often difficult to distinguish from expert-generated text (Ruff et al., 2024). In chemistry education specifically, AI systems have shown the ability to explain chemical phenomena, solve problems across various subdisciplines, generate instructional materials, and provide feedback on students' thinking with a level of sophistication approaching that of human educators (Talanquer, 2023; Yik & Dood, 2024). Despite these advances, significant concerns remain. These include challenges related to academic integrity and assessment authenticity, potential biases embedded in AI training data and fundamental questions about what it means for students to learn and demonstrate understanding in an era where AI can instantly generate plausible scientific explanations (Feldman-Maggor et al., 2025). Consequently, the growing use of AI for feedback and assessment raises critical issues regarding the validity, rigor, and fairness of evaluations, as well as how educators can ensure that students' learning experiences remain meaningful and authentic (Feldman-Maggor et al., 2025).

Automated assessment systems powered by artificial intelligence are no longer limited to analysing surface-level features such as grammar and text structure, they are increasingly capable of evaluating scientific accuracy and content validity. However, substantial challenges remain in achieving authentic assessments of discipline-specific reasoning (Fleckenstein et al., 2023; Meyer et al., 2024; Z. Wang, 2024). Findings from multiple meta-analyses indicate that automated feedback can produce small to moderate positive effects on writing quality when implemented appropriately, although effect sizes vary considerably depending on the specificity and quality of the feedback provided (Fleckenstein et al., 2023). These findings raise concerns about whether automated systems genuinely assess students' conceptual understanding or merely detect linguistic features that correlate with expert scores in training data (Gao et al., 2025). Comparative studies examining feedback generated by ChatGPT and traditional automated writing evaluation (AWE) systems have shown that large language models can provide more specific, context-sensitive and actionable guidance for revision. Several studies have reported improved student outcomes when AI-generated feedback is used, compared to conventional AWE tools (Wang, 2025). Despite these advantages, several important limitations of AI systems have been identified. These include a tendency to generate *hallucinations*, where explanations may sound scientifically reasonable but are factually incorrect, inconsistencies in scoring standards, the potential replication of biases found in training data, and limited transparency in how assessment criteria are applied. Together, these issues reduce the educational value of AI-based systems (Mizumoto & Eguchi, 2023; Xiao et al., 2025).

Research on automated scoring of scientific argumentation further indicates that machine learning models achieve higher accuracy when evaluating less complex argument components,

Volume: 11 Issues: 80 Special Issue [January, 2026] pp. 117 – 130
Journal of Islamic, Social, Economics and Development (JISED)
eISSN: 0128-1755
Journal website: academicinspired.com/jised
DOI: 10.55573/JISED.118009

such as identifying simple claims. In contrast, their performance declines significantly when assessing higher-order reasoning tasks, including comparative arguments or the integration of multiple sources of evidence (Li & Wilson, 2025). However, the potential of generative AI in supporting authentic assessment of structured chemistry writing cannot be ignored especially in the Claim-Evidence-Reasoning (CER) framework (Karunarathne et al., 2023; Yildirim & Akcan, 2024; Yuriev et al., 2024). The ability of generative AI to produce well-structured, logical chemical arguments with clear scientific reasoning introduces new challenges to academic integrity. As AI-generated CER responses become increasingly difficult to distinguish from those produced by students, the validity of writing-based assessments as measures of student understanding is at risk (Clark et al., 2024). Effective implementation of the CER framework requires careful validation of each component: claim, evidence and reasoning to ensure that the claim is chemically accurate and appropriately focused, the evidence is relevant, sufficient and derived from credible sources, and the reasoning correctly applies chemical principles to explain the relationship between submicroscopic mechanisms and observable phenomena (Fergus et al., 2023). Current research has yet to adequately address how AI systems for chemistry CER analysis can be designed, validated, and implemented in ways that genuinely support the development of students' disciplinary argumentation skills while avoiding superficial assessment, academic integrity violations, and inequities in access (Lu et al. , 2024)

Responsible Integration of AI for Chemistry CER Analysis

The challenges of using AI to analyze chemistry writing and scientific argumentation indicate a clear need for a well-structured and responsible AI integration framework. Such a framework should prioritize educational validity, transparency and fairness, while simultaneously leveraging the genuine capabilities of AI to support meaningful chemistry learning that is grounded in conceptual understanding (Berber et al., 2025). In the absence of clear guidance, AI use risks contributing to superficial assessment practices, overreliance on automated systems or misalignment with core pedagogical goals. Consequently, chemistry educators and researchers must adopt integrated strategies for AI implementation. Key strategies include ensuring that AI systems are designed to assess chemical understanding and scientific reasoning rather than surface-level textual features; developing transparent and explainable AI systems so that assessment criteria are accessible and understandable to students; and adopting hybrid human–AI assessment approaches that preserve teachers' professional judgment in evaluative decision-making (Barredo Arrieta et al., 2020; Gao et al., 2025; Holstein et al., 2019). Within educational contexts, AI is also more appropriately positioned as a formative feedback tool that supports continuous improvement rather than as a direct instrument for high-stakes summative assessment that carries significant consequences for students' academic progression (Fleckenstein et al., 2023)

The need for responsible AI integration becomes even more critical when analyzing chemistry argumentation based on the Claim–Evidence–Reasoning (CER) framework. Evidence in chemical arguments encompasses multiple forms of representation, including macroscopic observations, molecular diagrams, chemical equations, graphs, and thermodynamic calculations, each requiring distinct evaluative criteria. The complexity of these multi-level representations cannot be adequately captured through generic argumentation rubrics, thereby necessitating AI systems that are specifically designed for the chemistry domain (Ruff et al., 2024). Accordingly, an AI integration framework for chemistry CER analysis must be grounded in chemistry education research, learning science principles and ethical AI guidelines and developed through iterative processes involving chemistry educators, students, and assessment

Volume: 11 Issues: 80 Special Issue [January, 2026] pp. 117 – 130
Journal of Islamic, Social, Economics and Development (JISED)
eISSN: 0128-1755
Journal website: academicinspired.com/jised
DOI: 10.55573/JISED.118009

experts (Feldman-Maggor et al., 2025; UNESCO, 2024). At the same time, emerging professional development research highlights the importance of preparing chemistry teachers to integrate AI critically and responsibly. Yildirim & Akcan (2024) propose a teacher competency framework encompassing AI literacy, pedagogical integration strategies and the ability to evaluate the chemical accuracy of AI-generated content. However, this framework still requires more extensive empirical validation, particularly within the context of scientific argumentation analysis and chemistry essay writing. This underscores the need for further research to examine how AI can be ethically and effectively integrated to support chemistry CER analysis without compromising academic integrity or the quality of students' reasoning.

Validity, Transparency, and Ethics: Critical Considerations for AI Assessment

Recent studies indicate that the effectiveness of AI in educational assessment depends not only on its technical capabilities but also on critical issues of validity, transparency, and ethics (Feldman-Maggor et al., 2025). Although AI systems are capable of achieving acceptable levels of scoring reliability, several studies have shown that the criteria and features used by these systems often lack a clear alignment with underlying learning constructs, thereby raising concerns about the validity of the assessments they produce (Gao et al., 2025; Hannah et al., 2023). Issues of transparency become particularly pronounced with the use of large language models (LLMs), which typically operate as black-box systems in which decision-making processes are difficult for users and educators to interpret (Barredo Arrieta et al., 2020). In the context of chemistry education, Feldman-Maggor et al., (2025) emphasize the need for explainable AI systems that allow teachers to examine the rationale behind assessment decisions and to validate outcomes based on their professional judgment. Such an approach is essential to ensure that AI functions as a supportive tool rather than a replacement for pedagogical decision-making. Ethical and fairness considerations are also central to the use of AI in assessment. Risks such as data bias, inequitable access, misinformation and threats to academic integrity have been identified as major challenges that can compromise fairness in student evaluation (Feldman-Maggor et al., 2025; Ruff et al., 2024). Accordingly, responsible AI frameworks emphasize principles of human agency, transparency, fairness, and continuous monitoring, while also highlighting the need for institutional support and sustained professional development for teachers to ensure that AI is used ethically and meaningfully within chemistry education contexts (Berber et al., 2025).

Research Aim

To investigate the effectiveness and validity of using artificial intelligence to analyze and assess chemistry students' scientific writing structured using the Claim–Evidence–Reasoning (CER) framework.

Methodology

This study employed the Nominal Group Technique (NGT) as the primary research approach. A total of five experts with expertise in chemistry education, educational assessment and the application of artificial intelligence in education participated in the study. As face-to-face expert gatherings were not feasible at the time of the study, the NGT session was conducted online using the Google Meet platform. The session lasted approximately two hours. During this period, the experts actively engaged in the structured NGT procedures to generate ideas, share perspectives, and propose solutions based on their respective areas of expertise. Following the session, the researcher performed data computation and analysis in accordance with established NGT procedures to derive findings that aligned with the research objectives.

Volume: 11 Issues: 80 Special Issue [January, 2026] pp. 117 – 130
Journal of Islamic, Social, Economics and Development (JISED)
eISSN: 0128-1755
Journal website: academicinspired.com/jised
DOI: 10.55573/JISED.118009

NGT technique steps

The Nominal Group Technique (NGT) is a research method used to identify shared perspectives among a group of individuals on a specific topic. Originally developed as a participatory technique for social planning contexts (Delbecq, Van de Ven, & Gustafson, 1975), NGT was designed to support exploratory research, community participation, the involvement of experts from multiple disciplines, and proposal review processes. Since its introduction, NGT has been widely adopted in various group discussion contexts, including empirical research within the social sciences.

NGT is a highly structured process comprising four main phases:

1. Individual idea generation in which each participant responds independently to a stimulus question.
2. Round-robin sharing of ideas where ideas are presented sequentially without discussion.
3. Clarification and consolidation during which each idea is explained to ensure shared understanding and similar ideas are merged.
4. Individual voting or ranking, used to determine the priority of the proposed ideas.

NGT sessions are typically conducted over a period of 90 minutes to two hours and involve between five and ten participants. In this process, the researcher assumes the role of facilitator and administrator to ensure smooth discussion and to minimize researcher influence on the data collected. Some research methods may be influenced by researchers' assumptions through question framing or response coding; in contrast, the Nominal Group Technique (NGT) reduces such influence by allowing participants to generate, organize and prioritize ideas independently. Nevertheless, the formulation of the stimulus question remains critical in determining the effectiveness of the NGT process. Accordingly, researchers must be explicit about the type of information they intend to elicit through the session. In the first step of the NGT implementation in this study, participants were asked to propose actions or strategies that could be collaboratively implemented to enhance the effectiveness of analyzing and assessing chemistry students' scientific writing based on the Claim–Evidence–Reasoning (CER) framework. Participants were informed that the proposed strategies should be small-scale and practical, allowing them to be realistically designed and implemented within the available resources. The researcher acted as the facilitator and participated in the session to demonstrate the proper execution of the NGT procedure.

Each participant was provided with paper and a pen and asked to write down their ideas silently and independently. Once all participants had completed this stage, the ideas were collected and displayed in list form using an Excel spreadsheet projected on a shared screen. Each idea was then explained by the participant who proposed it to ensure mutual understanding, and similar ideas were merged where appropriate. After the idea generation, listing and clarification stages were completed, participants were asked to identify priority ideas using a five-card rating system. Each participant received five small coloured cards representing scores from one to five, which were used to rate their selected ideas. Although standard NGT procedures typically require participants to rate all generated ideas, the researcher's prior experience suggested that rating an excessive number of ideas may lead to confusion and scoring errors. Accordingly, in line with the session's objective of identifying a single primary action for implementation, only selected ideas were rated. This adaptation was intended to streamline the process and reduce the likelihood of error while preserving the integrity of the NGT methodology.

Volume: 11 Issues: 80 Special Issue [January, 2026] pp. 117 – 130
Journal of Islamic, Social, Economics and Development (JISED)
eISSN: 0128-1755
Journal website: academicinspired.com/jised
DOI: 10.55573/JISED.118009

Sampling

According to Booker and McNamara (2004), an expert is an individual who has acquired formal qualifications, professional training, practical experience, membership in professional bodies, and peer recognition through sustained effort and high levels of dedication (Nikolopoulos, 2004; Perera et al., 2012). In line with this view, Mullen (2003) defines an expert as a person possessing in-depth knowledge and a high level of expertise within a specific domain. Within the context of the Nominal Group Technique (NGT), expert selection is a critical methodological consideration, as inappropriate or insufficiently rigorous selection criteria may compromise the validity, credibility, and reliability of research findings (Mustapha & Darussalam, 2017). Kaynak and Macauley (1984) further emphasize that researchers involved in such studies must have adequate knowledge and understanding of the research domain to ensure accurate and appropriate expert selection. Accordingly, this study selected experts with a minimum of seven years of relevant professional experience and demonstrated high levels of domain-specific knowledge. Expert selection was guided by stringent criteria aligned with the objectives of the study, ensuring that only individuals meeting this experience threshold were included.

Overall, the expert panel comprised individuals from diverse yet complementary backgrounds: chemistry education, secondary school chemistry teaching, educational technology and artificial intelligence in education. All experts possessed extensive experience in teaching, educational assessment or the application of AI within science learning contexts. Participation in the study was voluntary and contingent upon the experts' willingness to engage actively throughout the NGT session. In instances where substantial divergence emerged between expert perspectives and the researcher's views, the researcher retained the option to consider the inclusion of alternative experts with equivalent qualifications and professional experience. Active and sustained participation during the NGT session was treated as a core inclusion criterion and a critical factor in ensuring the effective implementation of the technique.

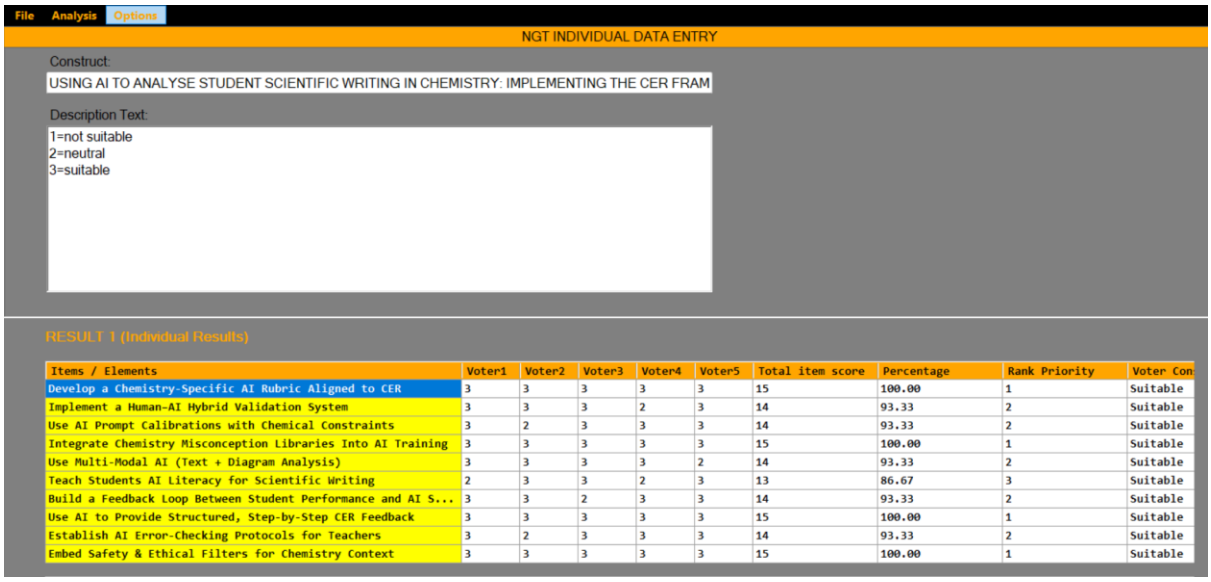
Table 1: Participant profile

Position / Area of Expertise	Number	Years of Experience	Institution
Chemistry Education Experts	1	15	Public University
Expert Chemistry Teachers (Secondary Level)	2	20	Secondary School
Educational Technology Experts	2	10	Public University

Data analysis

In this study, the Nominal Group Technique (NGT) was employed for both data collection and data analysis. All data generated during the NGT session were systematically entered into the NGT-PLUS software in stages as the session progressed. After each item had been discussed and refined by the expert panel, a voting process was conducted in real time with all participating experts present. Once the experts had provided their ratings for each item, the NGT-PLUS software was used to analyze the data and generate the study findings.

Volume: 11 Issues: 80 Special Issue [January, 2026] pp. 117 – 130
Journal of Islamic, Social, Economics and Development (JISED)
eISSN: 0128-1755
Journal website: academicinspired.com/jised
DOI: 10.55573/JISED.118009



Items / Elements	Voter1	Voter2	Voter3	Voter4	Voter5	Total item score	Percentage	Rank Priority	Voter Con
Develop a Chemistry-Specific AI Rubric Aligned to CER	3	3	3	3	3	15	100.00	1	Suitable
Implement a Human-AI Hybrid Validation System	3	3	3	2	3	14	93.33	2	Suitable
Use AI Prompt Calibrations with Chemical Constraints	3	2	3	3	3	14	93.33	2	Suitable
Integrate Chemistry Misconception Libraries Into AI Training	3	3	3	3	3	15	100.00	1	Suitable
Use Multi-Modal AI (Text + Diagram Analysis)	3	3	3	3	2	14	93.33	2	Suitable
Teach Students AI Literacy for Scientific Writing	2	3	3	2	3	13	86.67	3	Suitable
Build a Feedback Loop Between Student Performance and AI S...	3	3	2	3	3	14	93.33	2	Suitable
Use AI to Provide Structured, Step-by-Step CER Feedback	3	3	3	3	3	15	100.00	1	Suitable
Establish AI Error-Checking Protocols for Teachers	3	2	3	3	3	14	93.33	2	Suitable
Embed Safety & Ethical Filters for Chemistry Context	3	3	3	3	3	15	100.00	1	Suitable

Figure 1: Data Analysis from NGT Plus Software

Table 2: Data Result

Items / Elements	Voter					Total item score	%	Rank Priority	Voter Consensus
	1	2	3	4	5				
Develop a Chemistry-Specific AI Rubric Aligned to CER	3	3	3	3	3	15	100	1	Suitable
Implement a Human–AI Hybrid Validation System	3	3	3	2	3	14	93.33	2	Suitable
Use AI Prompt Calibrations with Chemical Constraints	3	2	3	3	3	14	93.33	2	Suitable
Integrate Chemistry Misconception Libraries into AI Training	3	3	3	3	3	15	100	1	Suitable
Use Multi-Modal AI (Text + Diagram Analysis)	3	3	3	3	2	14	93.33	2	Suitable
Teach Students AI Literacy for Scientific Writing	2	3	3	2	3	13	86.67	3	Suitable
Build a Feedback Loop Between Student Performance and AI Scoring	3	3	2	3	3	14	93.33	2	Suitable
Use AI to Provide Structured, Step-by-Step CER Feedback	3	3	3	3	3	15	100	1	Suitable
Establish AI Error-Checking Protocols for Teachers	3	2	3	3	3	14	93.33	2	Suitable
Embed Safety & Ethical Filters for Chemistry Context	3	3	3	3	3	15	100	1	Suitable

*** NGT data must exceed >75% agreement*

The Nominal Group Technique (NGT) analysis involving five experts revealed strong consensus on ten proposed strategies for implementing AI in the analysis of chemistry students' Claim–Evidence–Reasoning (CER) writing. All items achieved a status of "Appropriate," with levels of agreement ranging from 86.67% to 100%. Four critical elements emerged as the highest priorities (Priority Level 1), each receiving a perfect score of 15/15 (100%). These included developing chemistry-specific AI rubrics aligned with the CER framework, integrating chemistry misconception libraries into AI training, using AI to provide structured,

Volume: 11 Issues: 80 Special Issue [January, 2026] pp. 117 – 130
Journal of Islamic, Social, Economics and Development (JISED)
eISSN: 0128-1755
Journal website: academicinspired.com/jised
DOI: 10.55573/JISED.118009

step-by-step CER feedback and embedding safety and ethical filters tailored to the chemistry context. Five strategies were classified as second-level priorities (Priority Level 2), each attaining a score of 14/15 (93.33%). These strategies comprised implementing a hybrid human–AI validation system; applying AI prompt calibration with chemistry-specific constraints; utilizing multimodal AI for combined text and diagram analysis; establishing feedback loops between student performance and AI scoring; and developing AI error-checking protocols for teachers. The item "Teaching Students AI Literacy for Scientific Writing" received the lowest score of 13/15 (86.67%) and was categorized as a third-level priority (Priority Level 3), indicating that although it is considered important, it was perceived as slightly less critical than other implementation strategies. Overall, the high level of expert consensus (all items ≥ 86.67%) demonstrates strong agreement on the necessity and suitability of this comprehensive approach to effectively integrating AI into chemistry education assessment. The findings emphasize the importance of preserving disciplinary accuracy, ensuring ethical safeguards, and maintaining human oversight within AI-mediated assessment processes.

Table 3: Final output

Develop a Chemistry-Specific AI Rubric Aligned to CER	AI weaknesses often come from using generic language rubrics. • Create a rubric explicitly aligned to chemical accuracy, evidence usage, and reasoning logic. • Train AI prompts to detect: correct chemical equations, stoichiometric logic, particle-level explanations, misconceptions.
Integrate Chemistry Misconception Libraries Into AI Training	Students commonly misunderstand: • acid–base neutralisation • mole concept • chemical bonding ✓ Feed AI with misconception lists so it can *detect, label, and explain* misconceptions in CER writing.
Use AI to Provide Structured, Step-by-Step CER Feedback	Unstructured feedback overwhelms students. ✓ AI should break down feedback into: 1. Claim accuracy 2. Evidence quality 3. Reasoning coherence 4. Chemical accuracy ✓ Helps students focus on incremental improvement.
Embed Safety & Ethical Filters for Chemistry Context	AI must avoid producing dangerous or inaccurate chemical recommendations. • Ensure filters block unsafe lab suggestions • Verify AI does not mislead students with false chemical mechanisms • Maintain strict curriculum alignment Creates safe, reliable scientific communication.
Implement a Human–AI Hybrid Validation System	LLMs often misinterpret student reasoning. • Teacher validates AI scoring of CER components. • AI does the initial coding → teacher verifies → corrections improve the model. This reduces error and increases reliability.
Use AI Prompt Calibrations with Chemical Constraints	Generic prompts lead to hallucinations. • Add constraints such as: *"Only evaluate CER based on evidence provided by the student."* *"If uncertain, state uncertainty instead of guessing."* • Improves precision and avoids fake chemical explanations.
Use Multi-Modal AI (Text + Diagram Analysis)	Students' CER reasoning often includes diagrams or chemical equations. • Use AI that can interpret images of particle diagrams, equations, graphs. • Allows more valid analysis, especially in kinetics, titration, and thermochemistry CER tasks.

Volume: 11 Issues: 80 Special Issue [January, 2026] pp. 117 – 130
Journal of Islamic, Social, Economics and Development (JISED)
eISSN: 0128-1755
Journal website: academicinspired.com/jised
DOI: 10.55573/JISED.118009

Build a Feedback Loop Between Student Performance and AI Scoring	AI becomes more accurate when you refine it using local student samples. • Supply anonymised student CER responses as training data • Adjust AI scoring to local curriculum (e.g., DSKP/KSSM Chemistry) • Increases contextual relevance
Establish AI Error-Checking Protocols for Teachers	Teachers need simple checklists to catch AI errors: • Check if AI interpreted data/equations correctly • Compare AI grading with rubric benchmarks • Use 5–10 sample CER responses to calibrate marking This strengthens trust in AI scoring.
Teach Students AI Literacy for Scientific Writing	Students must understand how to critically use AI feedback: • Verify AI suggestions • Check chemical equations and data • Compare AI feedback with rubric This prevents blind reliance and reinforces metacognitive skills.

Table 2 demonstrates that the effective use of AI for assessing students' scientific writing in chemistry cannot depend on generic AI systems alone. The findings show that general language-based rubrics are unable to capture key chemistry-specific elements of CER writing, including the accuracy of chemical equations, stoichiometric reasoning, particle-level explanations, and common misconceptions. To address this limitation, the table highlights the importance of developing chemistry-specific AI rubrics aligned with the CER framework and integrating structured misconception libraries. These strategies enable AI to interpret students' reasoning more accurately and provide feedback that targets underlying conceptual errors rather than surface-level language quality. In addition, safety and ethical filters are essential to prevent AI from generating misleading or unsafe chemical information, thereby ensuring curriculum alignment and maintaining trust in AI-supported assessment.

Furthermore, Table 2 indicates that AI functions most effectively when embedded within a pedagogically guided system rather than used as a fully automated assessor. Human–AI hybrid validation, calibrated chemistry-specific prompts, multimodal analysis of text and visual representations, and continuous feedback loops using authentic student work all contribute to improved accuracy, reliability, and contextual relevance of AI assessment. Teacher led error-checking protocols and the development of student AI literacy further support responsible and critical use of AI feedback. Overall, the findings suggest that AI can meaningfully enhance chemistry assessment only when it is carefully designed to be discipline-specific, ethically controlled, and integrated with strong teacher oversight to support high-quality learning outcomes.

Volume: 11 Issues: 80 Special Issue [January, 2026] pp. 117 – 130
Journal of Islamic, Social, Economics and Development (JISED)
eISSN: 0128-1755
Journal website: academicinspired.com/jised
DOI: 10.55573/JISED.118009

Using AI to Analyze Student Scientific Writing in Chemistry (CER Framework)

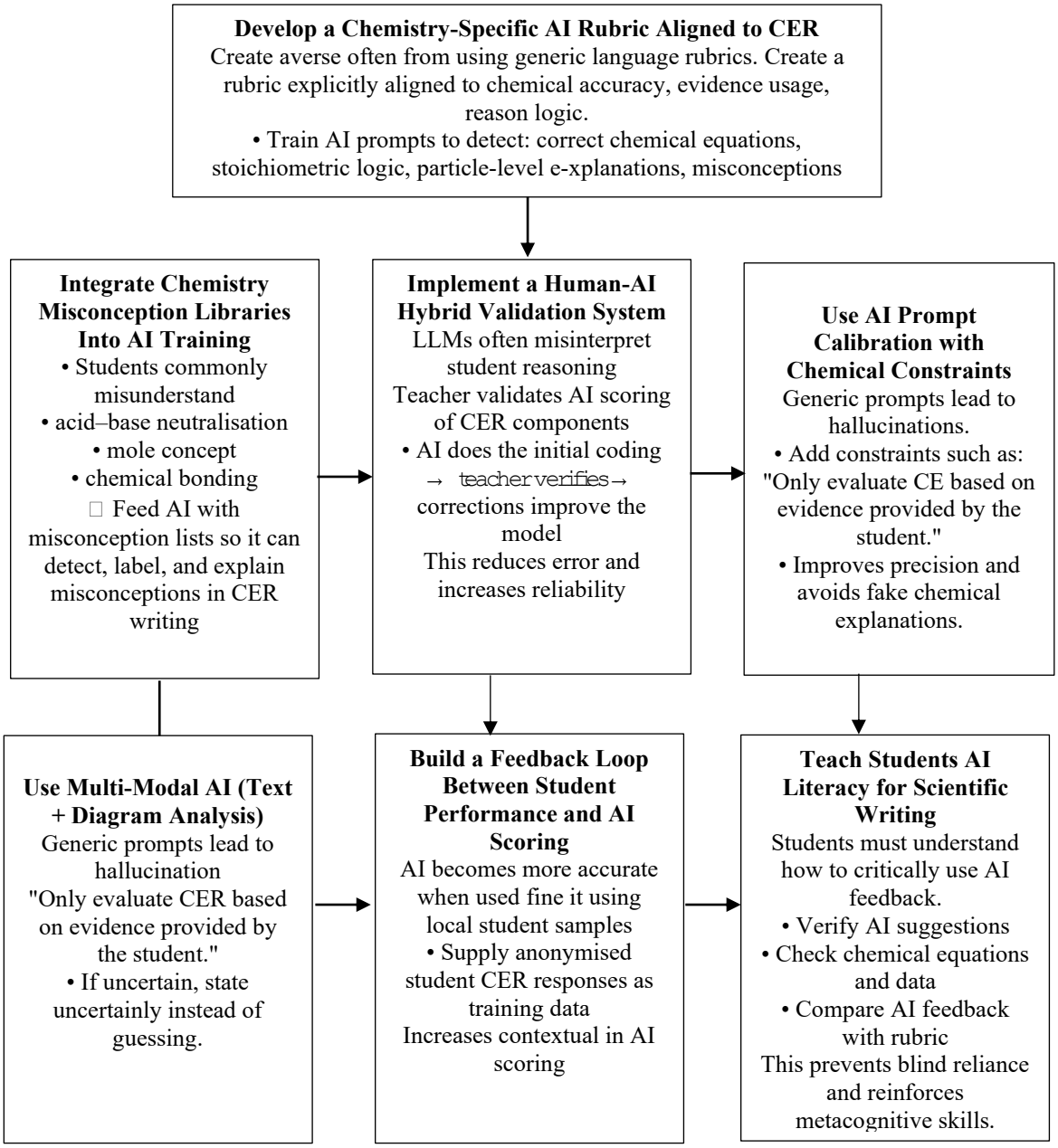


Figure 2: The Final Model for the AI Evaluation Framework in Chemistry

Discussion

The NGT analysis involving five experts demonstrates a very high level of consensus on the key strategies required for implementing AI-based analysis in the assessment of chemistry students' Claim–Evidence–Reasoning (CER) writing. All ten strategies presented in Table 2 were rated as appropriate, with agreement levels ranging from 86.67% to 100%, indicating strong expert endorsement of the proposed framework. Notably, four strategies achieved unanimous agreement (100%): the development of chemistry-specific AI rubrics aligned with CER, the integration of chemistry misconception libraries, the provision of structured step-by-step CER feedback, and the incorporation of safety and ethical filters. This unanimous prioritisation strongly suggests that generic AI systems are inadequate for chemistry assessment

Volume: 11 Issues: 80 Special Issue [January, 2026] pp. 117 – 130
Journal of Islamic, Social, Economics and Development (JISED)
eISSN: 0128-1755
Journal website: academicinspired.com/jised
DOI: 10.55573/JISED.118009

contexts, as they fail to capture the disciplinary complexity of chemical reasoning. These findings are consistent with prior research showing that large language models can generate chemically plausible explanations that nonetheless contain systematic conceptual and mechanistic errors (Yik & Dood, 2024), and that AI systems often privilege surface-level linguistic coherence over mechanistic chemical understanding (Talanquer, 2023). Collectively, the results highlight that effective AI-supported assessment must be grounded in discipline-specific chemical knowledge, including recognised misconceptions, valid forms of chemical evidence, and accurate reasoning that links macroscopic observations to submicroscopic and symbolic representations. The perfect agreement on ethical and safety filters further reflects growing expert concern over the risks of AI-generated chemistry content, reinforcing calls for responsible AI design that prioritises safety, scientific accuracy, and curriculum alignment (Feldman-Maggor et al., 2025).

In addition to these core strategies, a high level of agreement was also observed for the remaining approaches, including human–AI hybrid validation systems, chemistry-constrained prompting, multimodal AI analysis, feedback loops based on authentic student work, and AI error-checking protocols. This pattern of results indicates a shared expert perspective that AI should function as a complement to, rather than a replacement for, teacher professional judgement. Such findings align with the teacher–AI complementarity model proposed by Holstein et al., (2019), which emphasises combining AI's computational efficiency with teachers' disciplinary expertise to enhance assessment quality. These hybrid and constrained approaches also directly address long-standing validity concerns in AI-based assessment, where acceptable inter-rater reliability has not necessarily translated into valid measurement of conceptual understanding (Gao et al., 2025; Li & Wilson, 2025). In particular, the strong endorsement of chemistry-constrained prompting strategies reflects expert awareness of AI hallucination risks, whereby incorrect information is produced with high confidence (Kasneci et al., 2023). The inclusion of multimodal AI capabilities further strengthens assessment validity by enabling analysis of chemical diagrams, equations, and representations alongside text, which is essential for capturing the multi-level nature of chemical understanding (Modolo et al., 2023). Similarly, the emphasis on feedback loops grounded in local student data underscores the importance of contextualising AI systems rather than adopting generic, one-size-fits-all solutions.

Although the strategy of teaching AI literacy to students received a comparatively lower priority ranking (86.67%), it was still regarded as appropriate and necessary within the overall framework. This lower prioritisation may reflect experts' immediate focus on ensuring assessment accuracy, safety, and validity during implementation. Nevertheless, existing literature suggests that student AI literacy plays a critical role in determining whether AI feedback leads to meaningful learning gains. Studies by Wang, (2024) indicate that students often struggle to act productively on AI-generated feedback without explicit instruction, while Bucol & Sangkawong (2025) report that limited AI literacy can lead to over-reliance on AI outputs and reduced engagement in genuine scientific reasoning. These findings suggest that, alongside technical and pedagogical infrastructures such as validated rubrics and hybrid validation systems, developing students' capacity to critically evaluate AI feedback is essential for ensuring that AI integration in chemistry assessment supports rather than undermines, deep learning.

Volume: 11 Issues: 80 Special Issue [January, 2026] pp. 117 – 130
Journal of Islamic, Social, Economics and Development (JISED)
eISSN: 0128-1755
Journal website: academicinspired.com/jised
DOI: 10.55573/JISED.118009

Further Research

Future research should place greater emphasis on comprehensive validation of AI-based assessment systems, moving beyond evaluations of inter-rater reliability to include construct validity, consequential validity, and fairness across diverse student populations. This is essential to ensure that AI systems genuinely assess students' conceptual understanding of chemistry and the quality of their scientific argumentation, rather than merely analysing superficial linguistic patterns (Gao et al., 2025; Hannah et al., 2023). In addition, design-based research (DBR) is critically needed to develop and iteratively refine chemistry-specific AI tools through sustained collaboration with teachers in authentic classroom contexts. Such studies should investigate which types of AI systems are most appropriate and effective for chemistry education, recognizing that different AI architectures such as fine-tuned small models, prompt-based large language models, or hybrid systems do not function equivalently nor produce the same learning effects. Research should also examine which forms of AI-generated feedback most effectively support student learning, as well as how teachers can be prepared to use AI-based assessment critically and ethically in instructional practice (Fleckenstein et al., 2023; Yildirim & Akcan, 2024). Randomized controlled trials are further required to determine whether AI-generated feedback produces learning outcomes that are comparable to or superior to teacher-provided feedback across different chemistry topics and student populations. Such research is crucial for identifying optimal combinations of human and AI assessment that leverage technological efficiency without compromising educational quality (Wang, 2025). In parallel, future studies should prioritize the development of open-access chemistry misconception databases, validated CER rubrics, and shared benchmark datasets. These efforts would reduce institutional resource burdens, enable fairer comparisons across AI systems, and support collaborative progress in AI-driven chemistry education research (Li & Wilson, 2025).

Equity and bias should be key considerations in the use of AI for assessment. Future studies need to examine whether AI assessment systems show bias when evaluating writing from different groups of students, such as English as a Second Language (ESL) learners, students from underrepresented backgrounds, and those from different socioeconomic contexts. At the same time, research should explore AI design features and implementation strategies that support fair assessment and prevent the widening of existing educational inequalities.

Research on students' AI literacy is also important. Studies should investigate effective teaching approaches that help students critically evaluate AI-generated feedback. In particular, future research should examine how explicit instruction on assessing AI suggestions affects students' revision quality, metacognitive awareness, self-regulated learning, and their ability to identify accurate versus inaccurate AI feedback. In addition, longitudinal studies are needed to follow students across multiple chemistry courses to understand the long-term impact of AI-based assessment. Such studies can provide insights into how AI influences students' argumentation skills, conceptual understanding, development of scientific identity, and interest in science-related careers. This line of research can also help explain how institutional policies and assessment practices adapt as AI technologies continue to develop (Bearman et al., 2024; Huwer et al., 2024). Finally, technical research should focus on the development of multimodal AI systems capable of analysing text alongside chemical diagrams, molecular structures, and equations, while integrating structured domain knowledge and explainable AI approaches. These directions are critical to ensuring that AI assessment systems are valid, reliable, fair, transparent and pedagogically robust, thereby making a meaningful contribution to the advancement of chemistry education.

Volume: 11 Issues: 80 Special Issue [January, 2026] pp. 117 – 130
Journal of Islamic, Social, Economics and Development (JISED)
eISSN: 0128-1755
Journal website: academicinspired.com/jised
DOI: 10.55573/JISED.118009

References

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Berber, S., Brückner, M., Maurer, N., & Huwer, J. (2025). Artificial Intelligence in Chemistry Research─Implications for Teaching and Learning. *Journal of Chemical Education*, *102*(4), 1445–1456. https://doi.org/10.1021/acs.jchemed.4c01033

Bucol, J. L., & Sangkawong, N. (2025). Exploring ChatGPT as a writing assessment tool. *Innovations in Education and Teaching International*, *62*(3), 867–882. https://doi.org/10.1080/14703297.2024.2363901

Chin, C., & Brown, D. E. (2000). Learning in Science: A Comparison of Deep and Surface Approaches. *Journal of Research in Science Teaching*, *37*(2), 109–138. https://doi.org/10.1002/(SICI)1098-2736(200002)37:2%253C109::AID-TEA3%253E3.0.CO;2-7

Clark, M. J., Reynders, M., & Holme, T. A. (2024). Students' Experience of a ChatGPT Enabled Final Exam in a Non-Majors Chemistry Course. *Journal of Chemical Education*, *101*(5), 1983–1991. https://doi.org/10.1021/acs.jchemed.4c00161

Feldman-Maggor, Y., Blonder, R., & Alexandron, G. (2025). Perspectives of Generative AI in Chemistry Education Within the TPACK Framework. *Journal of Science Education and Technology*, *34*(1), 1–12. https://doi.org/10.1007/s10956-024-10147-3

Fergus, S., Botha, M., & Ostovar, M. (2023). Evaluating Academic Answers Generated Using ChatGPT. *Journal of Chemical Education*, *100*(4), 1672–1675. https://doi.org/10.1021/acs.jchemed.3c00087

Fleckenstein, J., Liebenow, L. W., & Meyer, J. (2023). Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence*, *6*, 1162454. https://doi.org/10.3389/frai.2023.1162454

Gao, X., Karumbaiah, S., Dalal, A., Dey, I., Gnesdilow, D., & Puntambekar, S. (2025). A Comparative Analysis of LLM and Specialized NLP System for Automated Assessment of Science Content. In A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, & S. Isotani (Eds.), *Artificial Intelligence in Education* (Vol. 15882, pp. 76–82). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-98465-5_10

Hannah, L., Jang, E. E., Shah, M., & Gupta, V. (2023). Validity Arguments for Automated Essay Scoring of Young Students' Writing Traits. *Language Assessment Quarterly*, *20*(4–5), 399–420. https://doi.org/10.1080/15434303.2023.2288253

Holstein, K., McLaren, B. M., & Aleven, V. (2019). Co-Designing a Real-Time Classroom Orchestration Tool to Support Teacher–AI Complementarity. *Journal of Learning Analytics*, *6*(2). https://doi.org/10.18608/jla.2019.62.3

Karunarathne, P. G. D. R. V., Somarathna, H. K. H. N., Liyanage, N. D. H., Vithanage, M. R., Bandara, P. S., & Wijesiri, P. (2023). *AI-Integrated Single Platform to Enhance Personal Wellbeing*. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85186142064&partnerID=40&md5=c76a0f5a7e211b4da8952e07ab2123b6

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Volume: 11 Issues: 80 Special Issue [January, 2026] pp. 117 – 130
Journal of Islamic, Social, Economics and Development (JISED)
eISSN: 0128-1755
Journal website: academicinspired.com/jised
DOI: 10.55573/JISED.118009

Li, M., & Wilson, J. (2025). AI-Integrated Scaffolding to Enhance Agency and Creativity in K-12 English Language Learners: A Systematic Review. *Information*, *16*(7), 519. https://doi.org/10.3390/info16070519

Luan, L. (n.d.). Bridging the Gap: ChatGPT's Role in Enhancing STEM Education. *Open Praxis*.

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, *6*, 100199. https://doi.org/10.1016/j.caeai.2023.100199

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

Modolo, L., Carvalho, S., & Dias, T. (2023). Digital health issues for the SUS: "mobile health" and the algorithmic automation of medical knowledge-power. *Saude e Sociedade*. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85176925522&partnerID=40&md5=2ca0a6c5bd86f564417aca7ecaa8d73b

Ruff, E. F., Engen, M. A., Franz, J. L., Mauser, J. F., West, J. K., & Zemke, J. M. O. (2024). ChatGPT Writing Assistance and Evaluation Assignments Across the Chemistry Curriculum. *Journal of Chemical Education*, *101*(6), 2483–2492. https://doi.org/10.1021/acs.jchemed.4c00248

Talanquer, V. (2023). Interview with the Chatbot: How Does It Reason? *Journal of Chemical Education*, *100*(8), 2821–2824. https://doi.org/10.1021/acs.jchemed.3c00472

UNESCO. (2024). AI competency framework for teachers. UNESCO Publishing.

Wang, S. (2025). Hybrid models of piano instruction: How combining traditional teaching methods with personalized AI feedback affects learners' skill acquisition, self-efficacy, and academic locus of control. *Education and Information Technologies*. https://www.scopus.com/inward/record.uri?eid=2-s2.0-

Wang, Z. (2024). Artificial intelligence in dance education: Using immersive technologies for teaching dance skills. *Technology in Society*. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85193437010&partnerID=40&md5=b9cedb16fc85035af77beaae148d6b14

Xiao (肖斐文), F., Zhu (朱思宇), S., & Xin (辛闻), W. (2025). Exploring the Landscape of Generative AI (ChatGPT)-Powered Writing Instruction in English as a Foreign Language Education: A Scoping Review. *ECNU Review of Education*, 20965311241310881. https://doi.org/10.1177/20965311241310881

Yik, B. J., & Dood, A. J. (2024). ChatGPT Convincingly Explains Organic Chemistry Reaction Mechanisms Slightly Inaccurately with High Levels of Explanation Sophistication. *Journal of Chemical Education*, *101*(5), 1836–1846. https://doi.org/10.1021/acs.jchemed.4c00235

Yildirim, B., & Akcan, A. T. (2024). AI-Professional Development Model for Chemistry Teacher: Artificial Intelligence in Chemistry Education. *Journal of Education in Science, Environment and Health*, 161–182. https://doi.org/10.55549/jeseh.741

Yuriev, E., Wink, D. J., & Holme, T. A. (2024). The Dawn of Generative Artificial Intelligence in Chemistry Education. *Journal of Chemical Education*, *101*(8), 2957–2959. https://doi.org/10.1021/acs.jchemed.4c00836