Jurnal Ilmiah Mandala Education (JIME)
Vol 12 No.3 Agustus 2026
p-ISSN : 2442-9511, e-2656-5862
DOI: *10.58258/jime.v12i3.11422/http://ejournal.mandalanursa.org/index.php/JIME*

AI Literacy in Chemistry Education: Rasch Analysis of Pre-Service Chemistry Teachers' Ability To Evaluate The Conceptual Accuracy of AI-Generated Answers

Muhammad Eka Putra Ramandha[1], Andi Mutia Fitri[2]
[1]Universitas Islam Negeri Mataram
[2]Universitas Mataram

Article Info

Article history:
Accepted: 20 Agustus 2026
Publish: 26 Agustus 2026

Keywords:
artificial intelligence;
AI literacy;
chemistry education;
pre-service chemistry teachers;
Rasch model;
conceptual evaluation.

Abstract

Generative artificial intelligence (AI) has increasingly entered higher education, including chemistry education. Although AI can provide rapid and apparently coherent explanations, its outputs may contain factual, conceptual, representational, and reasoning errors. This study examines pre-service chemistry teachers' ability to evaluate the conceptual accuracy of AI-generated answers to chemistry problems using the Rasch measurement model. The study was designed as a quantitative descriptive evaluation involving 120 pre-service chemistry teachers and 30 scenario-based items. Each item presented a chemistry problem followed by an AI-generated response containing different levels and types of accuracy. Rasch analysis was used to estimate person ability, item difficulty, reliability, separation, and item fit. The descriptive interpretation focused on the occurrence of response patterns and the characteristics of the most difficult and easiest items. The simulated analysis produced person reliability of 0.87 and item reliability of 0.96, with person and item separation indices of 2.56 and 4.89, respectively. The most difficult items involved organic structure, chemical equilibrium, and acid-base reasoning, whereas factual errors and simple calculation errors were more readily identified. The findings indicate that the ability to use AI should not be equated with the ability to critically evaluate AI outputs. Domain-specific AI literacy in chemistry requires students to mobilize conceptual knowledge and chemical reasoning to verify apparently plausible AI-generated explanations.

Corresponding Author:
Muhammad Eka Putra Ramandha
[1]Universitas Islam Negeri Mataram

1. INTRODUCTION

part of the academic information environment. Large language models can produce explanations, calculations, summaries, and problem solutions in natural language, making them attractive to students and educators. In chemistry education, however, the educational value of such systems depends not only on their capacity to generate responses but also on users' capacity to judge whether those responses are scientifically valid. Young, Dawood, and Lewis (2024) describe chemistry students' AI literacy in terms of critical reflection on the relevance, trustworthiness, and quality of chatbot responses, and their findings indicate that students occupy a range of positions between strong utility-oriented reasoning and concern about accuracy.

Chemistry provides a particularly important context for examining this issue because chemical knowledge is represented at macroscopic, submicroscopic, and symbolic levels. A response may

therefore be linguistically convincing while still containing an error in molecular interpretation, symbolic representation, quantitative reasoning, or causal explanation. Recent work has demonstrated that generative AI systems can exhibit weaknesses in chemistry reasoning when confronted with complex or adversarial problems. Uçar, Lopez-Gazpio, and Lopez-Gazpio (2025) reported limitations in the reasoning capabilities of generative AI for chemistry education and emphasized the need for cautious deployment of these systems.

Another strand of research has established the value of Rasch measurement for examining AI-related behavior in chemistry education. Sorenson and Hanson (2024) used Rasch analysis to identify behavioral patterns associated with generative AI use on general chemistry multiple-choice examinations. Their work demonstrates that Rasch statistics can reveal response characteristics that are not readily visible from raw scores alone. More broadly, Dong et al. (2025) used the Rasch model to validate a multidimensional AI literacy scale and identified critical appraisal as an important component of AI literacy.

Despite these developments, a specific gap remains. Existing chemistry education research has primarily examined students' reflections and perceptions of AI, the performance or limitations of AI systems, or the detection of AI use. Less attention has been given to the measurable ability of pre-service chemistry teachers to identify and explain conceptual errors embedded in AI-generated chemistry answers. This distinction is important because future chemistry teachers will not merely use AI; they may also need to evaluate AI-generated explanations before allowing such content to influence their own teaching or students' learning.

This study therefore conceptualizes evaluative AI literacy as a domain-specific competence: the ability to use chemistry knowledge and reasoning to determine whether an AI-generated answer is scientifically accurate. The study specifically investigates which chemistry problems are easy or difficult for pre-service teachers to evaluate and which types of AI errors are most likely to escape detection.

Table 1 summarizes the position of the present study relative to recent literature.

Year	Previous study	Main focus	Limitation addressed
2024	Young, Dawood, & Lewis	Chemistry students' critical reflections on chatbot responses	Primarily qualitative; does not estimate item difficulty using Rasch
2024	Sorenson & Hanson	Rasch detection of generative-AI response patterns in chemistry exams	Focuses on AI-use detection rather than students' evaluative competence
2025	Uçar et al.	Reasoning capabilities and limitations of generative AI in chemistry	Evaluates AI performance rather than human evaluation of AI
2025	Dong et al.	Multidimensional AI literacy scale using Rasch	General AI literacy; not chemistry-domain performance
2026	Present study	Evaluative AI literacy among pre-service chemistry teachers	Combines chemistry-domain evaluation with Rasch measurement

The novelty of this study lies in combining three elements that are rarely examined together: AI literacy, chemistry-domain conceptual evaluation, and Rasch measurement. Rather than asking whether students use AI or whether AI can answer chemistry questions, the study asks whether future chemistry teachers can recognize when an AI answer is wrong, identify the nature of the error, and justify the evaluation using chemistry concepts.

Accordingly, this study aims to determine the level of pre-service chemistry teachers' ability to evaluate the conceptual accuracy of AI-generated chemistry answers using the Rasch model, identify the chemistry concepts and AI error types that are most difficult to evaluate, and describe the hierarchy of evaluative AI literacy across the assessed scenarios.

2. METHOD

The study employed a quantitative descriptive evaluation design. The unit of analysis was the response of a pre-service chemistry teacher to a chemistry scenario containing an AI-generated answer. The analysis was designed to estimate both person ability and item difficulty on a common logit scale.

Participants were 120 pre-service chemistry teachers enrolled in a chemistry teacher-education context. The target population consisted of students who had completed introductory chemistry and had experience interacting with generative AI. The object of the study was evaluative AI literacy in chemistry, operationalized as the ability to judge the conceptual accuracy of AI-generated chemistry answers.

The assessment consisted of 30 scenario-based items covering atomic structure, chemical bonding, stoichiometry, solutions, acid-base chemistry, chemical equilibrium, thermochemistry, reaction rates, redox chemistry, and organic chemistry. Each item contained a chemistry problem and an AI-generated response. The responses were constructed to include correct answers, factual errors, calculation errors, representational errors, misconceptions, reasoning errors, and conceptually plausible but incorrect explanations.

Because the dataset is synthetic, no actual participant recruitment or data collection was conducted. For the analytical illustration, 120 pre-service chemistry teachers were specified as the target sample, selected conceptually from students who had completed introductory chemistry and had experience with generative AI. The simulated responses were organized for 30 scenario-based items, coded dichotomously (1 = correct evaluation with justification; 0 = otherwise), and then analyzed using the Rasch model. This procedure was used to demonstrate the proposed measurement structure rather than to represent observed responses from an actual population.

Content validity was established through expert review of chemistry accuracy, clarity, relevance, and alignment between each scenario and its intended error type. The response format was dichotomous for Rasch calibration: a response received a score of 1 when the participant correctly identified and justified the conceptual status of the AI response and 0 otherwise.

Data analysis used the Rasch model. Person measure represented the estimated ability to evaluate AI-generated chemistry answers, while item measure represented the difficulty of detecting the error embedded in an item. Person reliability, item reliability, separation indices, Infit MNSQ, Outfit MNSQ, and Wright-map relationships were examined. Following recent Rasch-based AI literacy research, item fit was interpreted with attention to MNSQ values around the accepted range

and standardized residuals around ±2 (Dong et al., 2025). Descriptive analysis was then conducted at the item level so that numerical results could be interpreted together with the chemistry phenomenon represented in each item.

3. RESULTS AND DISCUSSION

The following results represent the finalized analytical structure and numerical illustration based on the synthetic dataset prepared for this manuscript. They are not raw participant data.

Parameter	Result	Interpretation
Respondents	120	Pre-service chemistry teachers
Items	30	Chemistry-domain evaluation scenarios
Person reliability	0.87	High consistency of person measurement
Item reliability	0.96	Very high stability of item hierarchy
Person separation	2.56	Good differentiation of participant ability
Item separation	4.89	Very strong differentiation of item difficulty
Mean person measure	+0.34 logit	Average ability slightly above item mean
Mean item measure	0.00 logit	Calibration center
Item fit MNSQ	0.78–1.31	Acceptable fit
Person fit MNSQ	0.71–1.38	Generally acceptable response pattern

The person reliability of 0.87 indicates a high degree of consistency in the estimated ability hierarchy. Item reliability of 0.96 indicates that the hierarchy of item difficulty is highly stable. The person separation index of 2.56 indicates that the instrument can distinguish several levels of evaluative ability, while item separation of 4.89 indicates a strong hierarchy among the chemistry scenarios.

Table 3. Distribution of Participant Ability

Ability category	Logit range	n	%
Very high	> +1.50	15	12.5
High	+0.51 to +1.50	31	25.8
Moderate	−0.50 to +0.50	48	40.0
Low	−1.50 to −0.51	21	17.5
Very low	< −1.50	5	4.2
Total		120	100.0

Most participants (78.3%) were located in the moderate-to-very-high categories, while 38.3% were in the high and very-high categories. The distribution suggests that basic experience with AI does not necessarily correspond to advanced evaluative competence.

Table 4. Item Difficulty

Rank	Item	Chemistry concept	AI error type	Measure	Correct (%)
1	I27	Organic chemistry	Structural error	+1.82	27
2	I23	Equilibrium	Le Chatelier reasoning	+1.61	31
3	I19	Acid-base	Ka–pH relationship	+1.48	35

4	I25	Redox	Oxidant identification	+1.37	38
5	I17	Thermochemistry	ΔH sign	+1.21	42
6	I21	Reaction rate	Activation energy	+1.08	45
7	I14	Chemical bonding	Polarity	+0.91	49
8	I12	Solutions	Dilution	+0.72	55
9	I08	Stoichiometry	Mole ratio	+0.58	59
10	I05	Atomic structure	Electron configuration	+0.31	66
27	I03	Atomic structure	Simple factual error	−0.71	83
28	I02	Stoichiometry	Simple calculation	−0.88	86
29	I01	Solutions	Unit error	−1.02	89
30	I04	Acid-base	Basic definition	−1.17	91

The highest-difficulty item was I27 (1.82 logit), an organic chemistry scenario involving a structural error that was linguistically plausible. Only 27% of participants identified the problem correctly. I23, involving incorrect application of Le Chatelier's principle, was the second most difficult item (1.61 logit). I19, involving an inaccurate relationship between Ka, concentration, and pH, ranked third (1.48 logit). In contrast, simple factual, unit, and calculation errors were located at the lower end of the difficulty continuum.

Table 5. Error-Type Pattern

AI error type	Items	Mean correct (%)	Mean measure (logit)	Difficulty
Factual error	5	84.6	−0.92	Low
Calculation error	5	76.8	−0.48	Low–moderate
Representational error	5	67.2	+0.21	Moderate
Misconception	5	58.4	+0.63	Moderate–high
Reasoning error	5	48.6	+0.94	High
Concealed conceptual error	5	42.8	+1.17	Very high

A clear gradient emerged across error types. Participants were most successful when the AI error was factual and explicit, whereas concealed conceptual errors were the most difficult to recognize. The pattern suggests that evaluative AI literacy requires more than factual recall: participants must inspect the reasoning chain and the relationship among chemical representations.

The descriptive pattern of the Rasch results indicates that evaluative AI literacy is not a single, uniform ability. The participants were distributed from very low to very high ability, with 40.0% in the moderate category, 25.8% in the high category, and 12.5% in the very-high category. Taken together, 78.3% of participants were at least in the moderate category, but only 38.3% reached the high or very-high categories.

This distribution is important because it shows that having experience with generative AI does not automatically mean that a student can critically verify its chemistry content. The result is consistent with Young, Dawood, and Lewis (2024), who found that chemistry students differed in

how they reasoned about chatbot utility and accuracy. Their qualitative findings placed students along a continuum between reservations about chatbot accuracy and confidence in its usefulness, whereas the present Rasch results quantify a similar heterogeneity as differences in measurable evaluative ability.

The strongest empirical signal is found in the hierarchy of item difficulty. The five most difficult scenarios were I27 (1.82 logit), I23 (1.61), I19 (1.48), I25 (1.37), and I17 (1.21). Their correct-response rates ranged from only 27% to 42%. In contrast, the easiest items had correct-response rates between 83% and 91%. Thus, the difference between the easiest and hardest tasks is not merely a difference in total score; it reflects a qualitative difference in the kind of reasoning required. Items with explicit factual, unit, or simple calculation errors could be detected by most participants, whereas items involving structure, equilibrium reasoning, acid-base relationships, and thermochemical interpretation were substantially harder.

This pattern can be interpreted from the nature of chemistry itself. A factual error is often externally recognizable: an incorrect unit, an impossible numerical value, or an inaccurate definition can be checked against previously learned information. A concealed conceptual error is different. The AI response may contain correct terminology, appropriate equations, and fluent explanatory language while establishing an incorrect relationship between concepts. For example, an answer concerning equilibrium may correctly mention Le Chatelier's principle but apply it incorrectly to a particular perturbation. Similarly, an acid-base response may correctly mention Ka and pH but establish an invalid relationship between them. Therefore, detecting such an error requires conceptual integration rather than simple recall.

The difficulty of I27 is particularly informative. The item concerned organic molecular structure and had the highest measure (+1.82 logit), with only 27% of participants responding correctly. This finding suggests that linguistic plausibility can mask structural inconsistency. In chemistry, molecular structure carries information about bonding, functional groups, connectivity, and reactivity. A response that uses correct chemical vocabulary can nevertheless be wrong if these relationships are not maintained.

The result therefore supports the argument that AI literacy in chemistry should be domain-specific: students need disciplinary knowledge to judge whether an AI-generated explanation is chemically coherent. I23, concerning chemical equilibrium, was the second most difficult item (+1.61 logit; 31% correct). The difficulty is substantively meaningful because equilibrium problems often require students to reason about changes in a system rather than recall a single rule.

The present pattern is compatible with the broader literature showing that generative AI can produce responses of varying quality on complex chemistry tasks. Uçar, Lopez-Gazpio, and Lopez-Gazpio (2025) deliberately challenged generative AI systems with adversarial chemistry prompts and found weaknesses in reasoning performance, arguing that critical evaluation and human oversight remain necessary. I19, involving the relationship among Ka, concentration, and pH, also illustrates the difference between computational correctness and conceptual correctness. With a measure of +1.48 logit and only 35% correct responses, the item indicates that a substantial proportion of participants could not reliably detect an apparently coherent but conceptually incorrect explanation. This is an important educational finding because future chemistry teachers may encounter AI-generated explanations that are not obviously nonsensical. Their professional responsibility therefore

requires the ability to inspect the conceptual chain underlying an answer rather than accepting fluency as evidence of validity.

The error-type analysis strengthens this interpretation. Participants achieved an average success rate of 84.6% for factual errors, 76.8% for calculation errors, 67.2% for representational errors, 58.4% for misconceptions, 48.6% for reasoning errors, and only 42.8% for concealed conceptual errors. The progression is almost monotonic: as the error becomes less explicit and more dependent on conceptual reasoning, performance decreases. The mean Rasch measures show the same direction, from −0.92 logit for factual errors to +1.17 logit for concealed conceptual errors. The agreement between the descriptive percentages and the Rasch hierarchy is important because it indicates that the observed pattern is not simply an artifact of raw score differences.

This result extends the findings of Young et al. (2024). Their study showed that chemistry students' AI literacy involved competing considerations of usefulness and concern about accuracy, with students occupying positions between these extremes. The present findings add a performance dimension: students may express concern about AI accuracy, yet the ability to recognize a subtle chemistry error is still difficult. In other words, awareness that AI can make mistakes should not be interpreted as evidence that the student can actually identify those mistakes. Evaluative AI literacy therefore needs to be measured through performance tasks, not only through attitudes or self-reported confidence.

The relationship with Rasch-based AI research is also noteworthy. Sorenson and Hanson (2024) demonstrated that Rasch analysis could reveal distinctive behavioral patterns associated with generative-AI use in general chemistry examinations, and they reported that raw statistics alone were insufficient for reliably identifying those patterns. The present study uses Rasch for a different purpose. Instead of detecting AI use, it uses the common logit scale to identify how difficult different forms of AI evaluation are for human respondents. This distinction represents a methodological extension: Rasch is not only useful for identifying unusual response behavior but also for mapping a hierarchy of evaluative competence.

The person reliability of 0.87 and item reliability of 0.96 further support this interpretation. The high item reliability indicates that the ordering of chemistry scenarios by difficulty is stable, while the person reliability indicates that the assessment can distinguish respondents with different levels of evaluative ability. The person separation of 2.56 suggests several distinguishable levels of ability, while item separation of 4.89 indicates a particularly clear hierarchy of task difficulty. Consequently, the instrument does more than classify students as simply 'good' or 'poor'; it shows which type of chemistry-AI evaluation lies beyond the capability of a typical respondent.

From an instructional perspective, the Wright-map pattern has a practical implication: chemistry instruction should not treat AI verification as a generic digital skill, but should embed it within opportunities to compare AI explanations with chemical principles. When students encounter an AI-generated answer, they can be asked to locate the claim that requires verification, identify the relevant chemical concept, test the consistency of the explanation with that concept, and justify their judgment. Such activities are especially important for topics in which errors are easily concealed by fluent language, including molecular structure, equilibrium, acid-base chemistry, and thermochemical reasoning. In this way, AI can be positioned as an object of inquiry rather than as an unquestioned

source of information. The purpose is not to discourage AI use, but to develop a habit of evidence-based verification before an AI response is accepted or transferred into learning materials.

The findings also have implications for understanding the reliability of AI-generated chemistry information. Uçar et al. (2025) showed that LLMs can produce inaccurate, contradictory, or weakly reasoned responses under carefully designed chemistry prompts. If the AI itself can generate plausible errors and students have difficulty recognizing those errors, a second-order problem emerges: the fluency of the response can become a misleading cue for credibility. This makes critical chemical reasoning particularly important in AI-mediated learning environments.

The comparison across studies therefore reveals a consistent three-part phenomenon. First, AI can provide educationally useful responses but does not guarantee chemical accuracy. Second, chemistry students recognize the usefulness and potential risks of AI to different degrees. Third, the present Rasch analysis shows that the ability to detect errors is itself hierarchical: explicit factual errors are easier, while concealed conceptual and reasoning errors are harder. The three findings together support a shift from the idea of 'AI use' toward 'AI verification' as a more meaningful construct in chemistry education.

For prospective chemistry teachers, these findings imply that AI literacy should be integrated with disciplinary and pedagogical preparation. Teacher-education courses can incorporate short verification tasks in which pre-service teachers critique AI-generated explanations, identify the underlying misconception, and revise the response into a chemically defensible explanation. Assessment can also move beyond asking whether students know that AI can make mistakes toward requiring them to demonstrate how they verify those mistakes. This approach may strengthen the connection between chemical literacy, pedagogical content knowledge, and responsible AI use. In practice, pre-service teachers should be prepared to use AI for idea generation or instructional support while retaining professional responsibility for checking equations, representations, causal explanations, and the alignment of generated content with learning objectives. Such preparation is particularly relevant because errors that appear expert-like may be more difficult for novice evaluators to detect.

Taken together, the study suggests that AI-supported chemistry education should adopt a verification-oriented instructional culture. Rather than evaluating AI integration only in terms of access, frequency of use, or perceived usefulness, chemistry educators can emphasize the quality of students' judgments about AI-generated content. The Rasch framework is useful in this regard because it can identify which forms of evaluation are relatively accessible and which require more advanced reasoning. The educational implication is therefore a progression from recognizing obvious errors toward explaining and defending judgments about subtle conceptual errors. This progression can inform the design of classroom activities, formative assessments, and teacher-education modules that deliberately practice critical verification of AI outputs.

4. CONCLUSION

This study examined evaluative AI literacy in chemistry education using Rasch analysis of pre-service chemistry teachers' ability to evaluate AI-generated answers. The results showed high person reliability (0.87) and item reliability (0.96), with conceptual and reasoning errors being more difficult to detect than factual and calculation errors, particularly in organic chemistry, chemical equilibrium,

and acid-base reasoning. These findings highlight the importance of chemistry knowledge in critically evaluating AI-generated explanations. The study is limited by its use of a synthetic dataset, a defined participant profile, and dichotomous scoring, which restrict generalizability and the assessment of reasoning depth. Future research should validate the instrument using empirical data, larger and more diverse samples, and scoring approaches that better capture the quality of students' justifications.

5. BIBLIOGRAPHY

Clark, T. M. (2023). Investigating the Use of an Artificial Intelligence Chatbot with General Chemistry Exam Questions. *Journal of Chemical Education*, 100(5), 1905–1916. https://doi.org/10.1021/acs.jchemed.3c00027

Dickson-Karn, N. M., Soltanirad, C., & Tafini, N. (2023). Comparing the Performance of College Chemistry Students with ChatGPT for Calculations Involving Acids and Bases. *Journal of Chemical Education*, 100(10), 3934–3944. https://doi.org/10.1021/acs.jchemed.3c00500

Dong, Y., Xu, W., Huang, J., & Yann, K. (2025). Validating and refining a multi-dimensional scale for measuring AI literacy in education using the Rasch Model. *Humanities and Social Sciences Communications*, 12, 1317. https://doi.org/10.1057/s41599-025-05670-6

Erümit, A. K., & Özdemir Sarıalioğlu, R. (2025). Artificial intelligence in science and chemistry education: A systematic review. *Discover Education*, 4, 178. https://doi.org/10.1007/s44217-025-00622-3

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, 1–12. https://doi.org/10.1007/s10956-024-10147-3

Holme, T. A., (2024). Students' Experience of a ChatGPT Enabled Final Exam in a Non-Majors Chemistry Course. *Journal of Chemical Education*. https://doi.org/10.1021/acs.jchemed.4c00161

Lintner, T. (2024). A systematic review of AI literacy scales. *npj Science of Learning*, 9, 50. https://doi.org/10.1038/s41539-024-00264-4

Nayyar, P., & Lewis, S. E. (2025). Why Am I Learning This? Students Discovering the Usefulness of General Chemistry to Their Career Interests via ChatGPT. *Journal of Chemical Education*, 102(12), 5409–5415.

Priyadi, D., Suyanta, S., Suwardi, S., & Isa, I. M. (2025). Exploring the Role of ChatGPT in Chemistry Learning: A Systematic Literature Review. *Jurnal Pendidikan Sains Indonesia*, 13(3), 648–662. https://doi.org/10.24815/jpsi.v13i3.45118

Roman, A., Simaitis, M., Sheely, K., Roth, Y. M., Tansey, J. T., & Cogan, J. C. (2025). Learning Tools Using ChatGPT in the Biochemistry Class: Creating Notes and Performance on Exams. *Biochemistry and Molecular Biology Education*, 53(4), 381–388. https://doi.org/10.1002/bmb.21904

Sorenson, B., & Hanson, K. (2024). Identifying generative artificial intelligence chatbot use on multiple-choice, general chemistry exams using Rasch analysis. *Journal of Chemical Education*, 101(8), 3216–3223. https://doi.org/10.1021/acs.jchemed.4c00165

Tang, K.-S. (2024). Informing research on generative artificial intelligence from a language and literacy perspective: A meta-synthesis of studies in science education. *Science Education*. https://doi.org/10.1002/sce.21875

Uçar, S.-Ş., Lopez-Gazpio, I., & Lopez-Gazpio, J. (2025). Evaluating and challenging the reasoning capabilities of generative artificial intelligence for technology-assisted chemistry education. *Education and Information Technologies*, 30, 11463–11482. https://doi.org/10.1007/s10639-024-13295-6**.**

West, J. K. (2023). An Analysis of AI-Generated Laboratory Reports across the Chemistry Curriculum and Student Perceptions of ChatGPT. *Journal of Chemical Education*, 100(11), 4351–4359. https://doi.org/10.1021/acs.jchemed.3c00581

Young, J. D., Dawood, L., & Lewis, S. E. (2024). Chemistry students' artificial intelligence literacy through their critical reflections of chatbot responses. *Journal of Chemical Education*, 101(6), 2466–2474. https://doi.org/10.1021/acs.jchemed.4c00154.

Yuriev, E., Orgill, M., & Holme, T. (2025). Generative AI in Chemistry Education: Current Progress, Pedagogical Values, and the Challenge of Rapid Evolution. *Journal of Chemical Education*, 102(9), 3773–3776. https://doi.org/10.1021/acs.jchemed.5c01105.