Synthetic Solutions:

Comparing AI Models in Chemistry Problem Solving

Cody Mooring, Justin Smeya, Jacob Verret

CHEMH 1412: Honors General Chemistry II

May 9, 2025

Abstract

This study evaluates the chemistry problem-solving performance of three AI models, ChatGPT, DeepSeek, and Copilot, with the aim of understanding their accuracy, reasoning clarity, and types of conceptual errors. By using general chemistry questions drawn from a real assessment, this research investigates how well each AI interprets, calculates, and explains solutions across topics such as stoichiometry, thermochemistry, and electron structure. Each model's output was assessed for correctness, clarity of explanation, and logical reasoning patterns, with errors categorized into computational mistakes, misinterpretations, or flawed assumptions. The analysis reveals notable differences: while ChatGPT excels in explanation quality, DeepSeek demonstrates a strong performance in small, formulaic calculations, and Copilot frequently misinterprets questions and just attempts to provide an answer. These results suggest that no single AI currently masters all aspects of chemistry reasoning, but each has unique advantages and disadvantages. This study contributes to growing research on AI-assisted learning by identifying how specific models can aid or hinder chemistry education. Future research could expand the comparison to include visual-spatial chemistry tasks, molecular geometry, or evaluate how AI models evolve with updates and user feedback.

As artificial intelligence (AI) becomes increasingly integrated into educational and scientific fields, one major area of interest is its application in chemistry problem-solving. Tools such as ChatGPT, Copilot, and DeepSeek are popular and accessible to both students and faculty, raising important questions about AI's effectiveness in chemistry education. Do different AI models approach problem-solving in the same way, and what do their approaches tell us about how these models process chemistry knowledge? This study investigates how three different AI models perform on a standardized chemistry test. Rather than focusing purely on which model scores the highest, our aim is to identify the kinds of mistakes each AI makes and analyze what those mistakes reveal about the model's reasoning. As Yann LeCun, Turing Award winner and Chief AI Scientist at Meta, famously said, "Our intelligence is what makes us human, and AI is an extension of that quality." This research explores how far AI has come in mimicking human chemical reasoning and how far it still has to go.

The analysis centers on three distinct AI models: ChatGPT (OpenAI), Copilot (Microsoft), and DeepSeek (High-Flyer). Each model was tested using a consistent chemistry assessment across three controlled trials, with each trial modifying the resources provided. In Trial One, the models received only image-based questions from a general chemistry test and relied solely on internal model knowledge, simulating a scenario in which a student completes a test without any supporting materials. In Trial Two, a high-quality periodic table was added, testing whether external reference data would enhance performance or introduce confusion. Trial Three included the test, periodic table, and chapter notes aligned with the test's content to evaluate how structured educational content influences performance, whether it supports accurate responses or conflicts with the model's internal logic. All models answered the same

questions under these conditions. Responses were scored using an answer key, with additional evaluation focused on explanation clarity and error types across conceptual, computational, and recall-based questions.

The performance results revealed key differences. ChatGPT achieved the highest average score at 83%, demonstrating strong accuracy, consistency, and detailed explanation capabilities, particularly in multi-step reasoning. DeepSeek scored moderately at 68%, excelling in small calculations and simple conceptual problems but lacking the depth found in ChatGPT's reasoning. Copilot scored the lowest at 43%, often misinterpreting questions and prioritizing fast, surface-level answers. It struggled especially with recall and logical sequencing. These outcomes suggest performance variation and significant differences in how each model processes chemistry problems. ChatGPT adopted a methodical and human-like approach, Copilot tended toward shortcuts, and DeepSeek showed mechanical precision with limited elaboration. As Leite (2024) observed, "The quality of GenAI responses depends on the concept being evaluated," and even well-performing systems can be uneven in their understanding.

In one question, ChatGPT was asked a conceptual question comparing the atomic radius of nitrogen and oxygen. The model explained that while atomic radius generally decreases across a period, nitrogen's half-filled $2p^3$ configuration is relatively stable and experiences less electron-electron repulsion than oxygen's $2p^4$ configuration. Therefore, it correctly came to the conclusion that nitrogen has a slightly larger radius. This kind of reasoning reflects an understanding of both periodic trends and exceptions rooted in electron configuration stability, a level of depth comparable to that of a well-prepared student. Leite (2024) supports this notion, stating that "ChatGPT, Gemini, and Copilot provided satisfactory results for defining the

concepts [atom, electron, mole, molecule, chemical substance],” though only ChatGPT consistently contextualized those definitions within problem-solving.

Despite their differences, the models exhibited several shared weaknesses. Misreading question prompts was a frequent issue, especially with Copilot, indicating difficulty with visual and contextual interpretation. Overgeneralization of trends was another common error, such as assuming that periodic trends always hold without exception. Calculation errors were generally low for basic problems but increased in complexity during multi-step or unit conversion tasks. In Trial Three, the introduction of notes occasionally reduced performance consistency, suggesting that conflicting or overly detailed reference materials may confuse the models. As Blonder and Feldman-Maggor (2024) warn, “Using AI to answer questions that require critical thinking might result in students skipping over essential cognitive processes.” This highlights a central concern: while AI can provide useful guidance, it cannot replace the value of student reasoning.

The reasoning styles of the models also offer insight into their cognitive frameworks. ChatGPT approached problems similarly to how a student might articulate thoughts step-by-step, building logical frameworks to reach solutions. Copilot, by contrast, appeared to function more like an autocomplete engine, emphasizing speed over comprehension. DeepSeek provided practical responses but lacked nuanced understanding. These distinctions suggest that each AI model operates with a unique cognitive style, influencing how it solves problems. As Gasteiger (2020) notes, “Computers can be used for inductive learning: data can be put together to generate information and many pieces of information can be generalized to produce knowledge.” This inductive framework explains how certain AI models arrive at plausible conclusions without true conceptual grounding.

Several areas for future research are proposed to expand upon these findings. One avenue involves examining how AI accuracy changes when reference materials of varying quality—ranging from high-quality to deliberately misleading—are provided. Another direction includes comparing AI-generated responses with those of students to assess similarities in logic and error types. Additionally, repeated testing of AI models after a time delay or exposure to practice materials could help determine whether these systems exhibit long-term retention or adaptive learning. As Southworth et al. (2023) assert, "AI literacy is a minimum learning outcome for all post-secondary and K12 students." Therefore, understanding how students interact with AI and where they may be misled becomes a key part of chemistry education research.

Beyond immediate academic uses, the implications of AI in chemistry education extend into broader scientific practice. Abriata (2024) describes how AlphaFold and other innovations are driving advances in drug development and protein design, noting that, "These all-atoms models... stand as new ways for computers to assist drug development... likely reducing costs and experimental research time in drug development pipelines." Similarly, Iyamuremye et al. (2024) found that "The use of ChatGPT AI increased students' performance at the percentage of 16.6% in chemical bonding and atomic structure," suggesting concrete academic benefits. Still, as they caution, "the implementation of AI and ML in chemistry education is still in its juvenile stage."

As we consider this evolution, it is crucial to maintain a balanced view. AI is not a flawless tool, but rather one that requires oversight, context, and critical engagement. "AI can't teach ethics, judgment, or creativity—only humans can scaffold that kind of learning" (Blonder & Feldman-Maggor, 2024). Similarly, Yıldırım & Akcan (2024) emphasize both promise and

concern: “Most chemistry teachers believe using AI in chemistry classes may raise ethical concerns,” yet also see value in how it “will enhance students’ AI literacy, facilitate lasting learning, and encourage individualized learning.”

In conclusion, the comparative analysis of ChatGPT, Copilot, and DeepSeek demonstrates that while AI models can enhance chemistry education, their effectiveness varies widely. ChatGPT currently leads in both accuracy and reasoning depth, whereas Copilot and DeepSeek present limitations in reliability and conceptual clarity. These findings emphasize the importance of viewing AI not as a replacement for human reasoning, but as a supplementary tool that extends it. As AI technology continues to evolve, its value in educational settings will shed light not just on how well it answers questions, but on how it supports deeper thinking and learning. In this way, AI fulfills the role of an emerging technology that will contribute to the development of society, both in technological fields and, specifically, chemistry education.

References

Abriata, L. A. (2024). The Nobel Prize in Chemistry: past, present, and future of AI in biology. *Communications Biology*, *7*(1). https://doi.org/10.1038/s42003-024-07113-5

Aloys Iyamuremye, Francois Niyongabo Niyonzima, Janvier Mukiza, Innocent Twagilimana, Pascasie Nyirahabimana, Theophile Nsengimana, Jean Dieu Habiyaremye, Olivier Habimana, & Ezechiel Nsabayezu. (2024). Utilization of artificial intelligence and machine learning in chemistry education: a critical review. *Discover Education*, *3*(1). https://doi.org/10.1007/s44217-024-00197-5

Ananikov, V. P. (2024). Top 20 influential AI-based technologies in chemistry. *Artificial Intelligence Chemistry*, *2*(2), 100075–100075. https://doi.org/10.1016/j.aichem.2024.100075

Baum, Z. J., Yu, X., Ayala, P. Y., Zhao, Y., Watkins, S. P., & Zhou, Q. (2021). Artificial Intelligence in Chemistry: Current Trends and Future Directions. *Journal of Chemical Information and Modeling*, *61*(7), 3197–3212. https://doi.org/10.1021/acs.jcim.1c00619

Blonder, R., & Feldman-Maggor, Y. (2024). AI for chemistry teaching: responsible AI and ethical considerations. *Chemistry Teacher International*, *6*(4), 385–395. https://doi.org/10.1515/cti-2024-0014

Gasteiger, J. (2020). Chemistry in Times of Artificial Intelligence. *ChemPhysChem*, *21*(20), 2233–2242. https://doi.org/10.1002/cphc.202000518

Leite, B. (2024). Generative Artificial Intelligence in chemistry teaching: ChatGPT, Gemini, and Copilot's content responses. *Journal of Applied Learning & Teaching*, *7*(2). https://doi.org/10.37074/jalt.2024.7.2.13

Southworth, J., Migliaccio, K., Glover, J., Glover, J., Reed, D., McCarty, C., Brendemuhl, J., & Thomas, A. (2023). Developing a model for AI Across the curriculum: Transforming the higher education landscape via innovation in AI literacy. *Computers and Education: Artificial Intelligence*, *4*(1), 100127. https://doi.org/10.1016/j.caeai.2023.100127

Willmore, J. (2023, December 4). *AI education and AI in education | NSF - National Science Foundation*. New.nsf.gov. https://new.nsf.gov/science-matters/ai-education-ai-education

Yildirim, B., & Ahmet Tayfur Akcan. (2024). AI-Professional Development Model for Chemistry Teacher: Artificial Intelligence in Chemistry Education. *Journal of Education in Science Environment and Health*, *10*(4), 161–182. https://doi.org/10.55549/jeseh.741