Large Language Models are Catalyzing Chemistry Education

Yuanqi Du[1,*], Chenru Duan[2,3,*], Andres M. Bran[4,5,*], Anna Sotnikova[6], Yi Qu[3], Heather Kulik[2,3], Antoine Bosselut[6], Jinjia Xu[7,†], and Philippe Schwaller[4,5,†]

[1]Department of Computer Science, Cornell University, NY, USA

[2]Department of Chemical Engineering, Massachusetts Institute of Technology, MA, USA

[3]Department of Chemistry, Massachusetts Institute of Technology, MA, USA

[4]Laboratory of Artificial Chemical Intelligence (LIAC), Institut des Sciences et Ingénierie Chimiques, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

[5]National Centre of Competence in Research (NCCR) Catalysis, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

[6]Natural Language Processing Lab (NLP Lab), Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

[7] Department of Chemistry and Biochemistry, University of Missouri, MO, USA

[*]*Co-first authors*

[†]Corresponding authors

June 22, 2024

Abstract

Large language models (LLMs) have demonstrated outstanding capabilities in general problem-solving and been shown to improve productivity in certain domains. Thanks to their flexibility, recent work has leveraged them for diverse scientific applications, ranging from predictive modeling, scientific Q&A, and even as autonomous agents towards automation in chemistry. The democratization of high-quality chemistry education faces several challenges, including heterogeneity among sub-fields, limited access to personalized guidance, and an uneven distribution of resources. Additionally, hands-on laboratory experiments, a crucial component of chemistry education, are difficult to scale due to inherent safety risks that necessitate close supervision. We propose that LLMs can help overcome these obstacles by providing scalable solutions that tailor educational content to individual needs, enhancing the overall learning experience. In this perspective, we discuss how LLMs can catalyze chemistry education across multiple dimensions, from preparing and delivering lectures and tackling guidance in both wet lab and computational experiments, to re-thinking evaluation methodologies in the classroom. We also discuss some potential risks of this technology, such as the possibility of generating inaccurate or biased content, and emphasize the need for further development to ensure the successful integration of LLMs in the chemistry classroom.

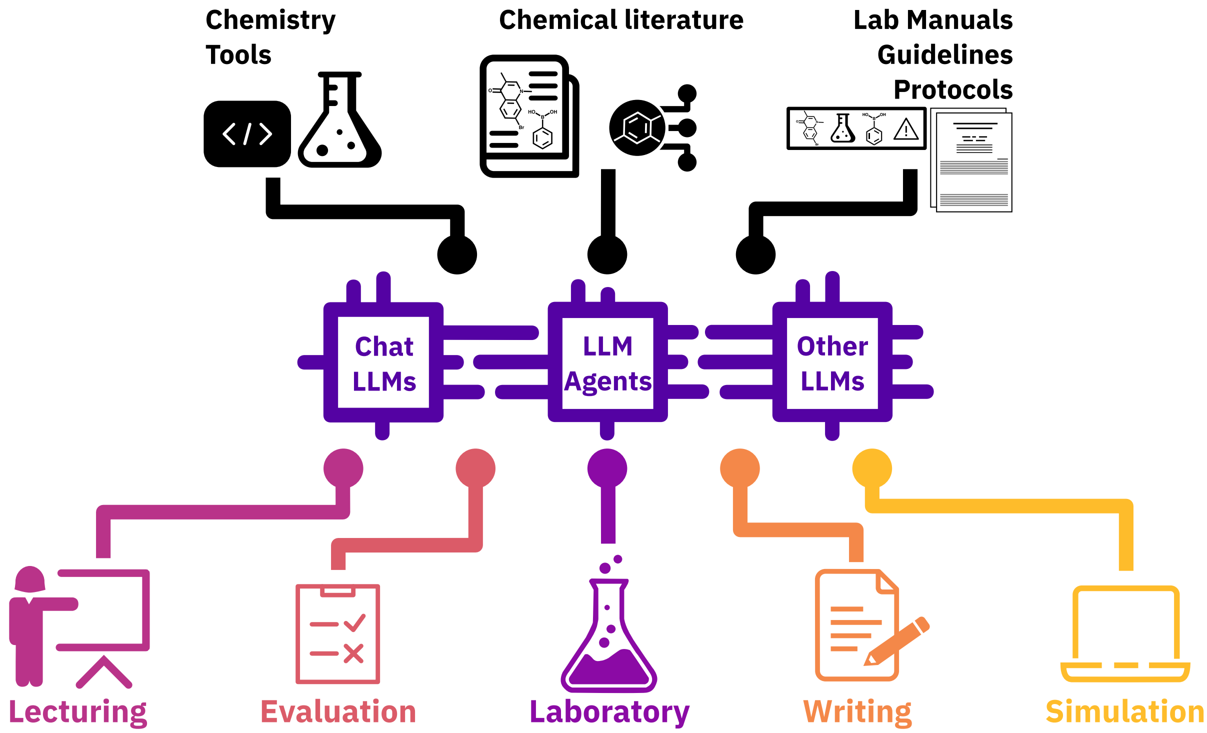

Figure 1: **Opportunities for LLMs in Chemical Education.** Large Language Models (LLMs) promise to augment chemical education across multiple dimensions (bottom). Multiple LLM-based systems such as Agents, Chat interfaces, and other systems (middle) can leverage multiple sources of chemical knowledge, such as software, literature, and laboratory manuals and protocols (top).

1 Introduction

Chemistry education and research have historically evolved closely together, each influencing the other [1]. In the early days, chemistry education relied heavily on the repetition of experimental procedures, which, while lacking theoretical insights, was useful for the industrial needs of the time [2]. The advent of unifying theoretical frameworks, such as structural theory and valence bond theory imported from research in chemistry, in turn shaped education, which by the 1950s started to lean toward more theoretical content [3, 4]. Not only has education been drastically influenced by groundbreaking theories but also by the introduction of new tools along with new ways of learning and experimenting [5, 6]. The introduction of computers came with a new way of interacting with knowledge, which started a revolution with long-standing effects [4]. From molecular visualization [7] to simulation-guided experimentation [8] and built-in data processing equipment in analytical instruments [9], chemists have fully embraced computers as a key part of their lives [10], permeating education along the way [11, 12].

New technologies in the field of Artificial Intelligence (AI) promise to fundamentally change science, much like the introduction of computers did [13, 14]. Popularized in the recent years, Large Language Models (LLMs) have demonstrated outstanding capabilities in tasks requiring reasoning and understanding across fields, including programming [15], knowledge extraction [16], experiment design [17, 18] and question-answering [19]. These very capable systems are based on neural networks, trained using massive amounts of text, and scaled to comparably massive amounts of parameters.

These capabilities have made them promising tools in research and education [20, 21, 22]. From the emergence of the most capable models [20, 23, 24], researchers have derived and built chemistry-specific applications tackling predictive modeling [25], optimization [26], assistance with experimental design [27], and automatic experimentation [18, 17], among others [28, 29]. These models also encode large amounts of chemistry knowledge, showing that they are capable of passing college-level exams and answering basic chemical questions [30, 20], as well as writing code with chemistry-relevant applications [31]. These works highlight the potential for future tools powered by LLMs, and how chemists will, similar to the adoption of computers, embrace them.

Following these advances in LLMs and more generally, generative AI, more attention has been paid to education and learning in chemistry [32, 33, 34, 35]. Promising potential applications have been discussed and implemented [36, 37, 38, 39, 40, 41, 42, 43], that highlight promising opportunities while also discussing potential negative effects of already publicly-accessible tools [44, 33]. Indeed, authors cite issues such as threats to the traditional essay writing for evaluation of understanding [44], plagiarism concerns [33], and more broadly the effect of Generative AI on student's learning and understanding [45, 46, 47, 48].

In this perspective piece, we propose a new LLM-augmented paradigm that encompasses multiple scenarios in chemistry education, from the classroom to the wet lab, computational lab, and the student evaluation stage, among others.

In the vision we introduce here, LLMs can enhance the lecturing experience in chemistry by aiding teachers in content creation as well as students with personalized Q&A and assistance in problem-solving, among others. In the lab, LLMs can help guide and assist students through manuals to ensure the correct implementation of protocols, troubleshoot experimental setups, and build connections between experiences in the lab and the underlying chemical principles. Similarly, education in computer simulations will benefit from LLMs that can write code, and as interfaces to an array of software used in computational chemistry. Furthermore, LLMs can also aid in cultivating critical thinking in students by aiding with scientific writing, helping to organize ideas, and offering pertinent resources, all with real-time feedback.

We recognize the opportunities that these new technologies offer for democratized access to high-quality educational resources in chemistry. Still, we also highlight the potential risks of misuse that accompany these benefits, as well as potential avenues for mitigation from the educator's perspective. We predict LLMs will open an entirely new world for learning and education in chemistry, and write this piece to prepare both educators and chemists to take the best from these developments.

2 Interactive learning systems are revolutionizing lecturing.

Lectures play a pivotal role in education, serving as the primary medium through which educators present specific topics from a curriculum. Although lecturing is usually a linear and one-sided process, where teachers serve as knowledge givers and students as receivers [49, 50], some pedagogical models suggest that more productive approaches can involve more active roles form the student's side, mediated by different levels of intervention [51, 52]. Under these views, lecturing is a more dynamic and bi-directional process, where the advent of LLMs offers unprecedented possibilities in the lecture planning and delivery stages [53] as shown in Figure 2. These systems potentially offer a host of benefits that extend beyond mere content enhancement. From providing instant feedback and generating thought-provoking questions, to adapting discussions, LLMs will assist educators through several stages of lecture preparation.

- **Lecture delivery**. Thanks to their capacity to provide instant feedback, generate thought-provoking questions, and adapt discussions, LLMs will assist educators in generating and refining discourse ideas, as well as providing a platform to experiment with pedagogical concepts [32]. LLMs also provide interactive interfaces to a multitude of tools, allowing educators to prepare improved visualizations and adapt their

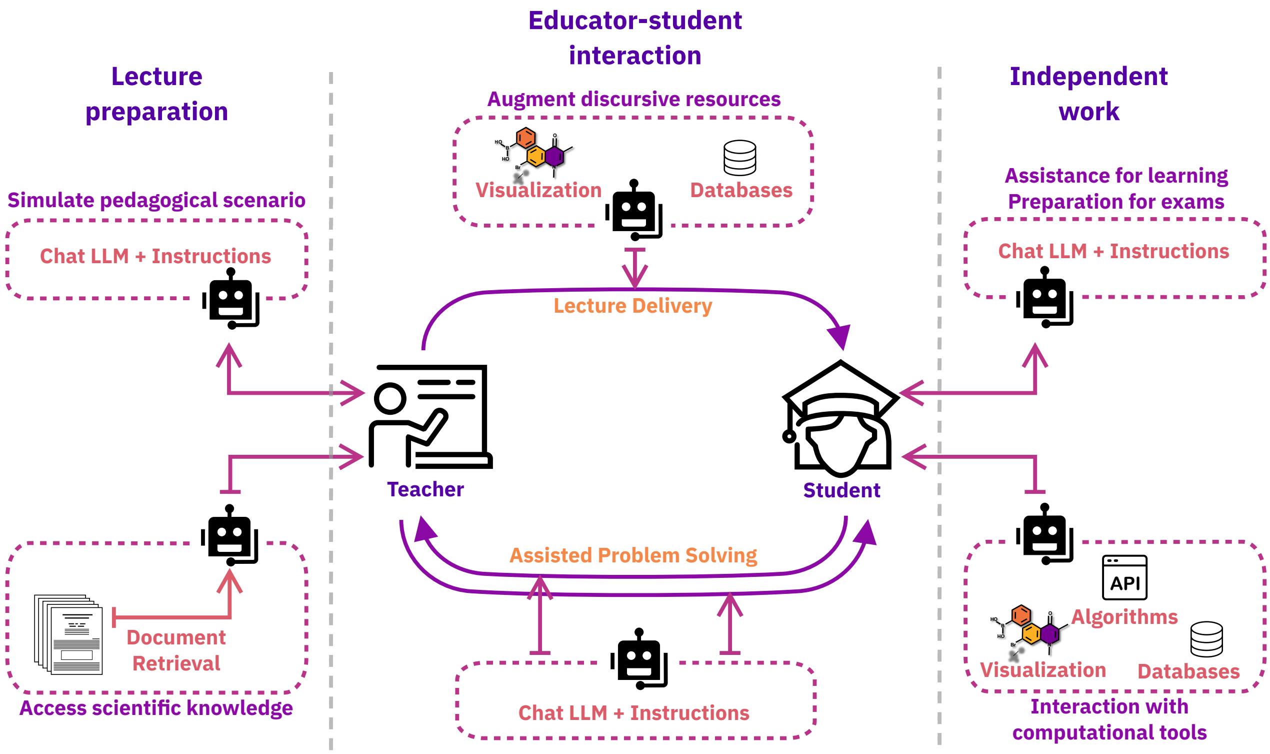

Figure 2: **LLMs in lecturing.** Multiple applications of LLMs can directly impact different steps of the lecturing process in chemistry. (center) shows the teacher-student interaction as single- and bi-directional relationships. The bi-directional relation encompasses assisted problem-solving skills, which can be augmented with instruction-tuned LLMs to follow human-defined instructions. The single-directional, mainly encompassing the lecturing sessions themselves, can also benefit from augmented visualization opportunities, along with access to tools and databases. (left) shows ways in which teachers can benefit from LLMs by using them to access scientific knowledge and experiment with pedagogical situations, while (right) shows how learning resources can be augmented for students by offering assistance with self-study and accessibility to chemistry-specific software tools.

lectures with interactive demonstrations, and enabling students to explore virtual models, all without the burden of mastering new frameworks.

- **Guidance in problem-solving**. Beyond memorizing facts, students are expected to develop problem-solving skills. A major limitation is the scalability of this process. In contrast to Massive Open Online Courses (MOOCs) for software skills, students in the natural sciences often receive limited feedback when solving problems. LLMs are regarded as a potential solution to this challenge, as they can efficiently provide students with immediate responses to questions, offer feedback on problem solutions, and suggest problem-solving strategies [54, 55, 56, 57]. These capabilities position LLMs to serve as virtual teaching assistants, augmenting the learning experience for students [58, 59].

- **Gateways to scientific knowledge**. The in-context learning and summarization capabilities of LLMs [60, 61, 62], when paired with efficient and accurate retrieval systems [63, 64], can serve as interactive gateways to scientific knowledge. They will provide literature-grounded responses to scientific inquiries and links to relevant sources, enabling students and educators to access the latest research related to their lectures [65, 66, 67].

These technologies in their current state, however, remain hampered by issues regarding the handling of references and the spread of misinformation. Potential solutions have been proposed recently, including fine-tuning for citations [68], and retrieval-augmented generation [69, 70]. These approaches partially mitigate these issues [19], paving the way for the future tools we envision. In addition, some reports show that reliance on such technologies may hinder learning while boosting confidence in acquired knowledge [46], providing devious illusions of learning. Given the ubiquitousness of such systems, addressing these issues while leveraging technologies will become an important topic in chemical education research.

Overall, LLM-based technologies have the potential for strengthening the connections between the classroom and current scientific questions, techniques, and solutions, which significantly enriches the educational experience.

3 Bridging theory and experiments: LLMs in lab sessions.

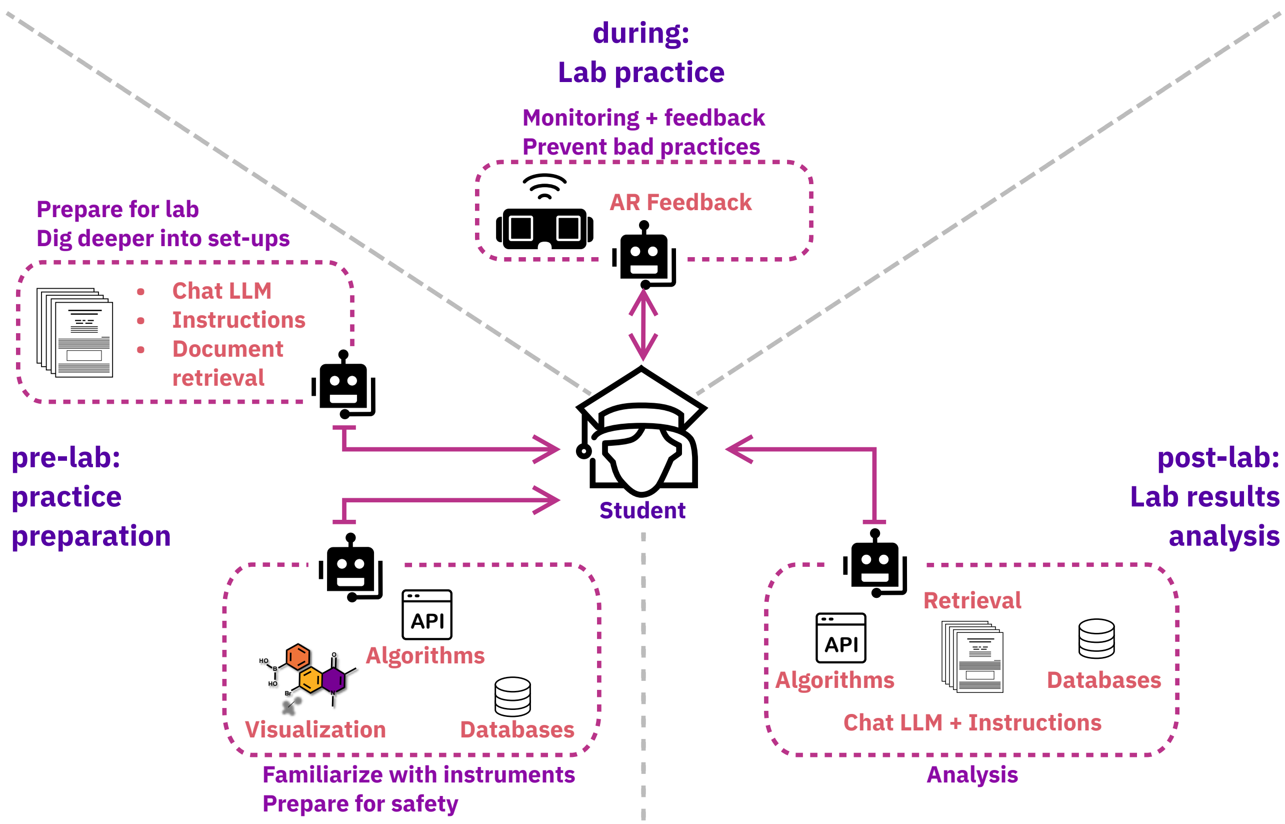

Figure 3: **LLMs in the wet lab.** LLMs will influence students' laboratory experience in all three stages: before, during, and after the lab. Before the lab (left), students can benefit from enhanced lab preparation through LLM-mediated interaction with lab manuals and other documents, and interaction with computational tools. During the lab (top), LLMs can facilitate personalized monitoring and assistance, particularly, when augmented with Augmented Reality tools. Such a setting can enhance safe lab practices. Finally, post-lab (right), students can leverage LLMs for data analysis and for support in understanding the experimental results they obtain.

Chemistry's foundations are deeply rooted in empirical observations, giving experimentation and laboratory training a key role in chemical education. As shown in Figure 3, the chemistry laboratory involves three main stages that can, to some extent, benefit from LLM applications. These can provide invaluable and scalable assistance in various aspects of laboratory work, from pre-session preparation and safety protocols to real-time feedback and post-session data analysis. This potential for integration of LLMs into the chemistry laboratory represents a significant step forward in the evolution of scientific education, offering a blend of traditional hands-on learning with this cutting-edge technology.

- **Laboratory preparation**. For laboratory sessions to be productive, students need to adequately prepare. Therefore, LLMs can provide personalized assistance in the preparation process, ensuring students derive maximum benefit from lab sessions [71]. Specifically, LLMs can clarify procedures detailed in lab manuals, offer comprehensive explanations of why certain setups are designed in particular ways, and augment safety resources [72, 73], making the student fully aware of the many design choices, risks, and learning objectives of experiments, without the entry barrier of lengthy lab manuals. By doing so, LLMs can bridge the gap between lab manuals and classroom teaching, ensuring a cohesive learning experience without imposing additional burdens on educators [74].

- **Personalized guidance & augmented reality**. Success in experimental sessions depends heavily on how well manual instructions are understood and followed by students, while failure in this respect carries serious consequences. LLMs can play a pivotal role in this aspect by providing accurate and timely feedback based on session-specific manuals and standardized procedures. Additionally, they can offer specific recommendations and safety advisories regarding the substances in use.

  The integration of Augmented Reality (AR) technology and LLMs has the potential to transform laboratory procedures [75, 76, 75]. AR technology can gather real-time data from experimental actions, and when processed through LLMs, students can receive immediate feedback, suggestions, and warnings. This integration not only enhances learning by offering personalized guidance, but also significantly improves the implementation of safety measures. By continuously monitoring actions, LLM-powered systems can be designed to identify and correct unsafe practices in real time, preventing accidents and ensuring a safer learning environment. This innovative integration of AR and LLMs represents a significant shift in laboratory training, merging the immediacy of real-time feedback with the immersive, enhanced experience of AR. Assistance in the lab is, nevertheless, of the uttermost importance, as minor mistakes in the lab can have important consequences. The implementation of such an application should thus be accompanied with other forms of regulation, such as rule-based systems carefully designed by experts.

- **Analysis of empirical results**. Laboratory training requires analyzing empirical results and exploring their connection to fundamental chemistry principles [77]. Here, LLMs can serve as powerful bridges to scientific knowledge. They can serve as assistants in analyzing and facilitating discussions centered on data that students have gathered over the course of the laboratory sessions. Furthermore, LLMs could connect students to relevant experimental results from the scientific literature, enhancing their understanding of the broader context of their findings. However, to fully harness this potential, the development of advanced retrieval systems specifically tailored for chemistry is of great importance. This integration of LLMs into the process of data analysis and interpretation could significantly enhance students' comprehension and application of chemistry principles in a practical context.

4 Augmenting chemistry simulations

Computational chemistry provides a way to explore and understand the properties, behavior, and interactions of chemical species at the atomic level [78]. Complementary to experiments, computation helps chemists gain insights into structures that are difficult to capture and characterize through experimentation (e.g., transition

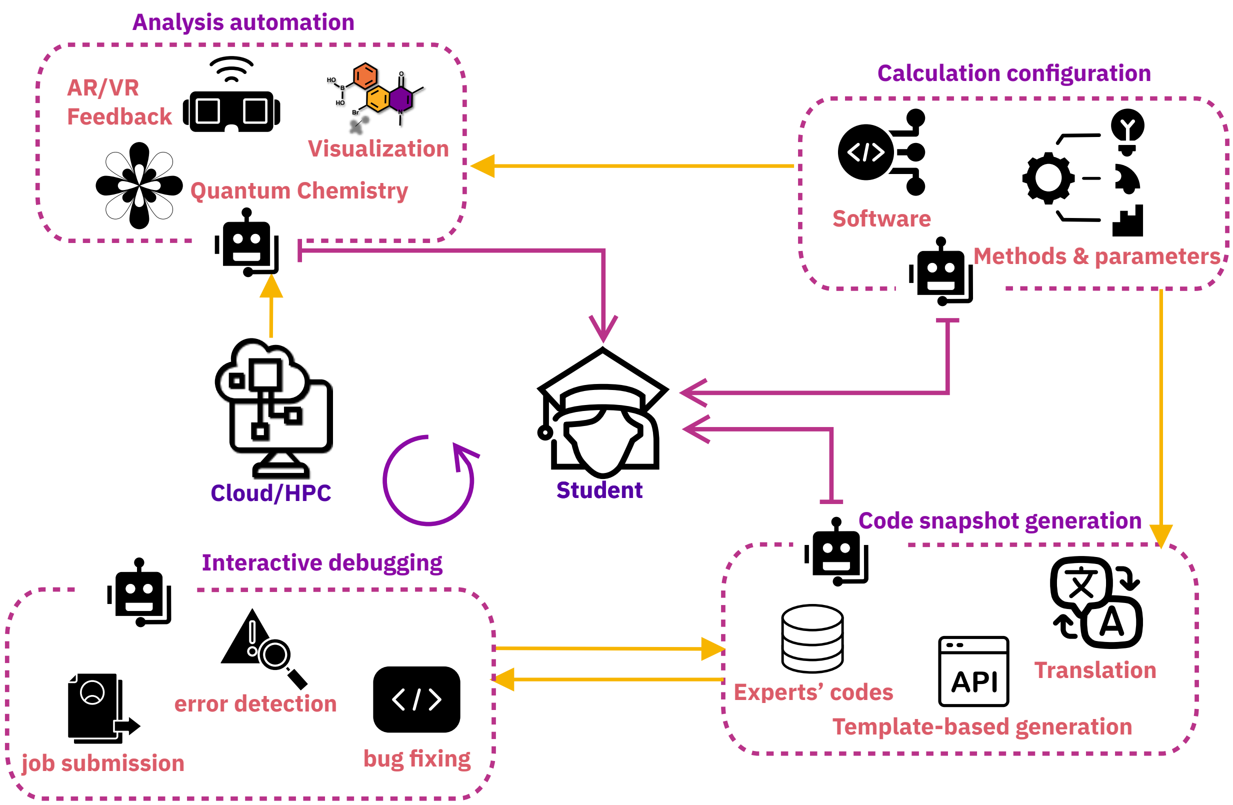

Figure 4: **LLMs in computer simulation.** LLMs will aid students for computational simulations in four parts: Starting with helping students configure their calculations (top right), LLMs can also generate code snapshots for students to revise (bottom right). After a calculation is submitted, students conduct an interactive process of debugging with the assistance of LLMs and cloud computing (bottom left). Lastly, automated analysis was performed, delivering more vivid and diverse types of visualization to students (top left).

states in chemical reactions [79]). Until as recently as 1956, however, there had not yet been calculations of a single molecule's properties on a computer [80]. The first systematic study in 1956 on diatomic molecules (for example, N_2) was performed using Hartree-Fock theory at various basis sets [81]. By 1971, the largest molecules computed were naphthalene and azulene [82]. Even in the early 2000s, many works were published with a focus on comparing the energies and orbitals of a small number of molecules. This situation shifted drastically in the 2010s with the development of both quantum chemistry software packages and high-performance computing infrastructure [83], accelerating large-scale computation for chemistry simulation and advancing chemical discovery in catalysis and drug design [84].

Recognized as an emerging discipline in the 1970s, computational chemistry aids hands-on experience in concretizing abstract quantities in chemistry, such as reaction barrier, electron density, and wavefunction. However, due to the organization of chemistry curricula, chemistry students typically lack a strong programming background by the time they finish chemistry-focused classes. Thanks to their natural language interface, LLMs can bridge this gap for chemistry students, alleviating the burden of coding for students while still delivering the central knowledge and hands-on experience of computational chemistry. Similar to how students and researchers in the 1970s could not foresee the growth of computational chemistry over the next 50 years, we are only scratching the surface of use cases for LLMs in computation. In the future, we envision LLMs cultivating a new generation of chemists more capable of computation on daily experiments validation empowered by LLMs.

- **Choosing the right configuration**. Part of the complexity of computational chemistry lies in its approximation to the exact solution of the Schrödinger equation. This necessity of making an approximation leads, in practice, to the user's choice of software, electronic structure method, wavefunction basis, and other details of the calculation setup. As these are all details that impact the final outcome of a calculation, the appropriate choices rely heavily on "know-how" that is often present but not adequately summarized in the literature [85]. With LLMs learning from open-source repositories, books, and papers, this latent knowledge in computational chemistry configuration can, in principle, be extracted and better democratized to everyone who wants to do computer simulations of chemistry. Nevertheless, as in other areas where LLMs will be applied, publication bias or knowledge source could influence such summarized recommendations.

- **Generating code snapshot**. As each quantum chemistry software has its own advantages and domain of application, becoming proficient in all of them is a challenging task. From simple natural language queries, LLMs can synthesize functional pieces of code demonstrating the use of various types of simulation on a variety of computational chemistry packages, reducing the labor of learning the syntax needed to interact with quantum chemistry software [86, 87, 88, 89, 90].

- **Debugging**. LLMs, as generative models, may not be completely error-free when generating code, leading to bugs in execution. However, this shortcoming can be viewed as a feature, rather than a bug, in learning and education [91, 92, 93, 94, 95]. The capability of the LLM to correct itself by interacting with error messages provides a natural way of debugging, which helps students learn coding gradually with chains of reasoning steps. The requirement of user interaction with LLMs prior to executing code also lowers the risk of over-reliance on LLMs when chemistry students learn to understand the setup of quantum chemistry calculations and basic rules of coding.

- **Automating analysis**. An important use case of computation in chemistry is to concisely present abstract concepts (for example, visualizing electron density to learn about covalent bonding). However, calculating these properties would normally require users develop significant familiarity with electronic structure theory and quantum chemistry packages. This separation in domain knowledge has led to reluctance for experimentalists to apply computations during the design of experiments or might not know what the feasible and and most efficient calculations are. LLMs provide a natural solution to bridge this gap, by helping experimentalists automate more advanced computational analysis and visualization to fulfill the daily needs of simulation.

As multimodal LLMs develop, one would imagine executing chemistry simulations more easily even without writing any code or input files[83]. For example, molecules drawn by hand may be recognized via image recognition, after which calculations can be directly set up by interfacing with LLM agents and run in the cloud. In terms of visualization, ball-and-stick models that are introduced in high school chemistry teaching may no longer be needed. Augmented or virtual reality would visualize the molecules and computational results, providing a more vivid experience for students to interact with atoms in a molecule and, better, its electron density or vibrational spectra[96, 97].

5 Cultivating scientific thinking through scientific writing

Scientific writing is a key facilitator of scientific thinking [98], a fundamental goal of chemistry education [99]. The traditional approach to teaching scientific writing, primarily through interactive lecturing, is undergoing a significant transformation with the integration of LLMs [100, 101, 102]. LLMs are elevating the standards of scientific writing, but they are also instrumental in facilitating critical thinking during writing [103].

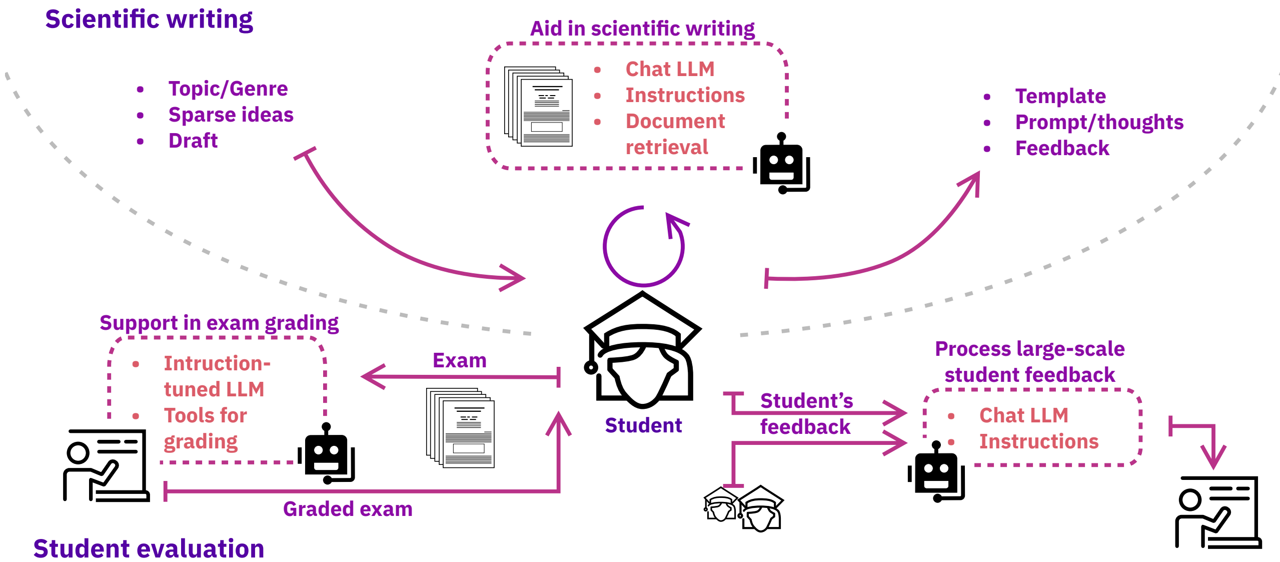

Figure 5: **LLM-assisted scientific writing and learning outcome evaluation.** (a) LLMs can bridge topic/genre with templates, prompt sparse ideas into stories, and provide feedback to drafts. (b) LLMs can propose variants to old exam questions, grade exams, and enable interactive exam formats. (c) LLMs can prompt and process student feedback at scale in real time.

- **Planning**. Planning is a critical step to structure and organize contents in scientific writing. This process involves analyzing the purpose and audience of scientific writing, understanding the importance of clear and logical structuring, and recognizing the impact of different writing styles on the reader. Through this process, students develop a deeper understanding and appreciation of the nuances of scientific writing. LLMs can serve as a rich repository of templates, examples, and criteria that define high-quality scientific writing. LLMs will also encourage students to engage in critical thinking, prompting them to question why certain approaches are effective.
- **Writing**. After planning, writing is an iterative process to concretize thoughts and ideas. This process not only improves their writing skills but also cultivates a habit of self-reflection and critical evaluation, which are essential for developing scientific writing skills. LLMs can provide real-time feedback and suggestions, playing a crucial role in shaping student writing. This feedback mechanism is complemented by exercises in critical thinking as students are encouraged to reflect on the feedback and consider how different word choices or structural changes alter the clarity and persuasiveness of their arguments.

Despite being a promising tool, ethical concerns of LLMs remain challenging to overcome [104]. Schools have taken action to [105, 106, 107] navigate a co-creation between human and generative AI tools: mainly by having some transparency about the ways such tools are used or should be used. Academic publishing groups have also taken actions to request reports on the use of LLMs in publications [108, 109]. We discuss more ethical and legal concerns in Sec. 7.

6 Getting creative with evaluations: how to assess learning outcomes?

Evaluating learning outcomes is a pivotal step in the education cycle, assessing the quality of education, and providing essential feedback for refining earlier stages [110, 111]. This evaluation encompasses two primary areas: (1) assessment of student performance [112], and (2) assessment of educator effectiveness [113]. Traditional evaluation methods often rely on educator-assigned tasks such as homework assignments, group projects, and exams. Meanwhile, student-provided feedback includes such procedures as course evaluations and student reports. The advent of LLMs is revolutionizing this process by introducing innovative, dynamic, and effective assessment strategies [114, 115].

- **Automated exam workflow**. LLMs will be instrumental in aiding instructors to formulate exam questions and variations of homework assignments based on patterns in previous materials. This variety will not only ensure a comprehensive assessment of the curriculum, but it will also keep the evaluation process up-to-date and adaptive to students. LLMs will also aid in grading assignments, thereby alleviating the routine grading workload on teaching assistants [116, 117, 118, 119]. Finally, automated grading will mitigate potential biases or human errors in grading.
- **Enhancing creative problem-solving**. LLMs can pose unique, scenario-based problems that require creative thinking and problem-solving, going beyond traditional assessment paradigms [120, 121, 122, 123]. Specifically, students can be asked to teach the LLMs to complete a lab or simulation run, fostering students' ability to apply learned concepts in novel situations, a key indicator of a deep understanding of the learned concepts rooted in practice.
- **Personalized feedback for educators**. Moving beyond traditional course evaluation forms, LLMs can prompt students for more timely, personalized, and insightful feedback. Furthermore, they will help monitor long-term learning trends and suggest immediate attention based on the feedback. This monitoring could include analyzing student responses for common trends, generating comprehensive and personalized feedback reports, and offering targeted suggestions for course improvements.

7 Challenges

Despite the promising capabilities demonstrated by the current generation of LLMs, specific risks need to be mitigated for their effective application in educational settings.

- **Hallucination.** LLMs, in their current state, often generate inaccurate and misleading information, even though it may seem plausible [124]. Mitigating these hallucinations remains a large and active research area with proposed solutions ranging from improved data quality [125], enhanced interpretability [126], constrained generation [127], to adversarial training, etc [128]. Consequently, students and teachers must be aware of the possibility of LLMs providing convincing, but false, responses when prompted, and use these tools carefully with appropriate verification. This may be particularly challenging for students learning new fields, such as chemistry or more specifically advanced computational chemistry.
- **Over-reliance**. Despite their capabilities, LLMs should not only be treated as a machine to ease exam question design. To alleviate the risk of LLMs eroding educational outcomes may require lecturers and professors to rethink exams, or even more, whether memorization-based recollection of knowledge as an educational goal will still hold in the age of LLMs. There is also a concern that LLMs might lead to surface-level learning, where students may prioritize quick answers over a deep understanding of the subject matter.
- **Ethical and legal concerns**. When students interact with LLMs, their data, including potentially sensitive information, might be collected and used in ways that are not transparent, raising privacy

concerns [129, 130]. Furthermore, the ownership of the work created with the assistance of LLMs may be subject to patent or copyright. The introduction of LLMs could change the traditional roles of teachers and students, with potential implications for the teaching profession and the nature of learning. In addition, the training data of LLMs from the internet may lead to potential copyright infringement, and the misinformation spread by LLMs also lack of responsibility. Finally, the regulation for the proper use of LLMs by students has been experimented with in schools and publication groups (Sec. 5).

- **Over-standardization**. Unified opinions offered by LLMs can hardly be expected to lead to breakthroughs in science and engineering. But can LLMs really provide diverse perspectives and ways of thinking? Over-standardization of LLMs for information and problem-solving could impede the development of critical thinking and research skills in students and thus influence our next-generation researchers.

- **Inequality.** LLMs are recognized to exhibit reduced efficacy when operating with low-resource languages or certain dialects like African American vernacular English [131, 132, 133, 134, 135]. As a result, users representing these linguistic variations may encounter sub-optimal performance compared to native English speakers. This diminished user experience can manifest in various ways, potentially leading to outputs of inferior quality, increased biases, and even model-generated hallucinations.

References

1. Gabel, D. Improving teaching and learning through chemistry education research: A look to the future. *Journal of Chemical education* **76**, 548 (1999).

2. Cooper, M. M. & Stowe, R. L. Chemistry education research—from personal empiricism to evidence, theory, and informed practice. *Chemical reviews* **118**, 6053–6087 (2018).

3. Cooper, M. M. & Stowe, R. L. Chemistry Education Research—From Personal Empiricism to Evidence, Theory, and Informed Practice. *Chemical Reviews* **118**, 6053–6087 (2018). URL `https://doi.org/10.1021/acs.chemrev.8b00020`. Publisher: American Chemical Society.

4. Pièrre, L. On the Self-Image of Chemists, 1950-2000. *HYLE–International Journal for Philosophy of Chemistry* **12**, 99–130 (2006). URL `https://www.hyle.org/journal/issues/12-1/laszlo.htm`.

5. Tondeur, J. *et al.* Preparing pre-service teachers to integrate technology in education: A synthesis of qualitative evidence. *Comput. Educ.* **59**, 134–144 (2012). URL `https://api.semanticscholar.org/CorpusID:19825319`.

6. Bellou, I., Papachristos, N. M. & Mikropoulos, T. Digital learning technologies in chemistry education: A review (2018). URL `https://api.semanticscholar.org/CorpusID:189138062`.

7. DeLano, W. L. *et al.* Pymol: An open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr* **40**, 82–92 (2002).

8. Walters, W. P., Stahl, M. T. & Murcko, M. A. Virtual screening—an overview. *Drug discovery today* **3**, 160–178 (1998).

9. Strohalm, M., Kavan, D., Novák, P., Volny, M. & Havlicek, V. mmass 3: a cross-platform software environment for precise analysis of mass spectrometric data. *Analytical chemistry* **82**, 4648–4651 (2010).

10. Bellou, I., Papachristos, N. M. & Mikropoulos, T. A. Digital learning technologies in chemistry education: A review. *Digital technologies: Sustainable innovations for improving teaching and learning* 57–80 (2018).

11. Wohlfart, O., Wagner, A. L. & Wagner, I. Digital tools in secondary chemistry education – added value or modern gimmicks? In *Frontiers in Education* (2023). URL `https://api.semanticscholar.org/CorpusID:259288834`.

12. Hrynevych, L., Morze, N. V., Vember, V. & Boiko, M. Use of digital tools as a component of stem education ecosystem. *Educational Technology Quarterly* (2021). URL `https://api.semanticscholar.org/CorpusID:250930721`.

13. Wang, H. *et al.* Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023).

14. Back, S. *et al.* Accelerated chemical science with ai. *Digital Discovery* **3**, 23–33 (2024).

15. Romera-Paredes, B. *et al.* Mathematical discoveries from program search with large language models. *Nature* **625**, 468–475 (2024).

16. Petroni, F. *et al.* Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Association for Computational Linguistics, 2019).

17. Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570–578 (2023).

18. Bran, A. M., Cox, S., White, A. D. & Schwaller, P. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376* (2023).

19. Lála, J. *et al.* Paperqa: Retrieval-augmented generative agent for scientific research (2023). `2312.07559`.

20. Achiam, J. *et al.* Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

21. Extance, A. Chatgpt has entered the classroom: how llms could transform education. *Nature* **623**, 474–477 (2023).

22. Tu, X., Zou, J., Su, W. & Zhang, L. What Should Data Science Education Do With Large Language Models? *Harvard Data Science Review* **6** (2024). Https://hdsr.mitpress.mit.edu/pub/pqiufdew.

23. Touvron, H. *et al.* Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

24. Team, G. *et al.* Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

25. Jablonka, K. M., Schwaller, P., Ortega-Guerrero, A. & Smit, B. Leveraging large language models for predictive chemistry. *Nat. Mach. Intell.* **6**, 161–169 (2024).

26. Ramos, M. C., Michtavy, S. S., Porosoff, M. D. & White, A. D. Bayesian optimization of catalysts with in-context learning. *Preprint at https://arxiv.org/abs/2304.05341* (2023).

27. Zheng, Z. *et al.* A GPT-4 Reticular Chemist for Guiding MOF Discovery**. *Angewandte Chemie International Edition* **62**, e202311983 (2023). URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.202311983`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.202311983.

28. Jablonka, K. M. *et al.* 14 examples of how llms can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital Discovery* **2**, 1233–1250 (2023). URL http://dx.doi.org/10.1039/D3DD00113J.

29. AI4Science, M. R. & Quantum, M. A. The impact of large language models on scientific discovery: a preliminary study using gpt-4. *arXiv preprint arXiv:2311.07361* (2023).

30. Guo, T. *et al.* What can large language models do in chemistry? a comprehensive benchmark on eight tasks (2023). 2305.18365.

31. D. White, A. *et al.* Assessment of chemistry knowledge in large language models that generate code. *Digital Discovery* **2**, 368–376 (2023). URL https://pubs.rsc.org/en/content/articlelanding/2023/dd/d2dd00087c. Publisher: Royal Society of Chemistry.

32. Kasneci, E. *et al.* Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences* **103**, 102274 (2023).

33. Alasadi, E. A. & Baiz, C. R. Generative ai in education and research: Opportunities, concerns, and solutions. *Journal of Chemical Education* **100**, 2965–2971 (2023).

34. Latif, E. *et al.* Artificial general intelligence (agi) for education. *arXiv preprint arXiv:2304.12479* **1** (2023).

35. Emenike, M. E. & Emenike, B. U. Was this title generated by chatgpt? considerations for artificial intelligence text-generation software programs for chemists and chemistry educators. *Journal of Chemical Education* **100**, 1413–1418 (2023).

36. Rojas, A. J. An Investigation into ChatGPT's Application for a Scientific Writing Assignment. *Journal of Chemical Education* **101**, 1959–1965 (2024). URL https://doi.org/10.1021/acs.jchemed.4c00034. Publisher: American Chemical Society.

37. Exintaris, B., Karunaratne, N. & Yuriev, E. Metacognition and Critical Thinking: Using ChatGPT-Generated Responses as Prompts for Critique in a Problem-Solving Workshop (SMARTCHEMPer). *Journal of Chemical Education* **100**, 2972–2980 (2023). URL https://doi.org/10.1021/acs.jchemed.3c00481. Publisher: American Chemical Society.

38. Humphry, T. & Fuller, A. L. Potential ChatGPT Use in Undergraduate Chemistry Laboratories. *Journal of Chemical Education* **100**, 1434–1436 (2023). URL https://doi.org/10.1021/acs.jchemed.3c00006. Publisher: American Chemical Society.

39. Araújo, J. L. & Saúde, I. Can ChatGPT Enhance Chemistry Laboratory Teaching? Using Prompt Engineering to Enable AI in Generating Laboratory Activities. *Journal of Chemical Education* **101**, 1858–1864 (2024). URL https://doi.org/10.1021/acs.jchemed.3c00745. Publisher: American Chemical Society.

40. Clark, T. M., Fhaner, M., Stoltzfus, M. & Queen, M. S. Using ChatGPT to Support Lesson Planning for the Historical Experiments of Thomson, Millikan, and Rutherford. *Journal of Chemical Education* **101**, 1992–1999 (2024). URL https://doi.org/10.1021/acs.jchemed.4c00200. Publisher: American Chemical Society.

41. Lee, G.-G. & Zhai, X. Using ChatGPT for Science Learning: A Study on Pre-service Teachers' Lesson Planning (2024). URL http://arxiv.org/abs/2402.01674. ArXiv:2402.01674 [physics].

42. Kortemeyer, G. Toward AI grading of student problem solutions in introductory physics: A feasibility study. *Physical Review Physics Education Research* **19**, 020163 (2023). URL `https://link.aps.org/doi/10.1103/PhysRevPhysEducRes.19.020163`. Publisher: American Physical Society.

43. Kieser, F., Wulff, P., Kuhn, J. & Küchemann, S. Educational data augmentation in physics education research using ChatGPT. *Physical Review Physics Education Research* **19**, 020150 (2023). URL `https://link.aps.org/doi/10.1103/PhysRevPhysEducRes.19.020150`. Publisher: American Physical Society.

44. Yeadon, W., Inyang, O.-O., Mizouri, A., Peach, A. & Testrow, C. P. The death of the short-form physics essay in the coming AI revolution. *Physics Education* **58**, 035027 (2023). URL `https://dx.doi.org/10.1088/1361-6552/acc5cf`. Publisher: IOP Publishing.

45. Mai, D. T. T., Da, C. V. & Hanh, N. V. The use of ChatGPT in teaching and learning: a systematic review through SWOT analysis approach. *Frontiers in Education* **9** (2024). URL `https://www.frontiersin.org/articles/10.3389/feduc.2024.1328769`. Publisher: Frontiers.

46. Messeri, L. & Crockett, M. J. Artificial intelligence and illusions of understanding in scientific research. *Nature* **627**, 49–58 (2024). URL `https://www.nature.com/articles/s41586-024-07146-0`. Publisher: Nature Publishing Group.

47. Essel, H. B., Vlachopoulos, D., Essuman, A. B. & Amankwa, J. O. ChatGPT effects on cognitive skills of undergraduate students: Receiving instant responses from AI-based conversational large language models (LLMs). *Computers and Education: Artificial Intelligence* **6**, 100198 (2024). URL `https://www.sciencedirect.com/science/article/pii/S2666920X23000772`.

48. Lo, C. K., Hew, K. F. & Jong, M. S.-y. The influence of ChatGPT on student engagement: A systematic review and future research agenda. *Computers & Education* **219**, 105100 (2024). URL `https://www.sciencedirect.com/science/article/pii/S0360131524001143`.

49. Taber, K. S. Chemistry lessons for universities?: a review of constructivist ideas (2000). URL `https://api.semanticscholar.org/CorpusID:146491205`.

50. Fox, D. Personal theories of teaching. *Studies in Higher Education* **8**, 151–163 (1983). URL `https://api.semanticscholar.org/CorpusID:144162037`.

51. Ahmad, Z., Ammar, M., Sellami, A. & Al-Thani, N. J. Effective Pedagogical Approaches Used in High School Chemistry Education: A Systematic Review and Meta-Analysis. *Journal of Chemical Education* **100**, 1796–1810 (2023). URL `https://doi.org/10.1021/acs.jchemed.2c00739`. Publisher: American Chemical Society.

52. Sjöström, J., Eilks, I. & Talanquer, V. Didaktik Models in Chemistry Education. *Journal of Chemical Education* **97**, 910–915 (2020). URL `https://doi.org/10.1021/acs.jchemed.9b01034`. Publisher: American Chemical Society.

53. Clark, T. M., Fhaner, M., Stoltzfus, M. & Queen, M. S. Using chatgpt to support lesson planning for the historical experiments of thomson, millikan, and rutherford. *Journal of Chemical Education* **101**, 1992–1999 (2024).

54. Lewkowycz, A. *et al.* Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems* **35**, 3843–3857 (2022).

55. Yao, S. *et al.* Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems* **36** (2024).

56. Orrù, G., Piarulli, A., Conversano, C. & Gemignani, A. Human-like problem-solving abilities in large language models using chatgpt. *Frontiers in artificial intelligence* **6**, 1199350 (2023).

57. Tsai, M.-L., Ong, C. W. & Chen, C.-L. Exploring the use of large language models (llms) in chemical engineering education: Building core course problem models with chat-gpt. *Education for Chemical Engineers* **44**, 71–95 (2023).

58. Liu, M. & M'Hiri, F. Beyond traditional teaching: Large language models as simulated teaching assistants in computer science. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, 743–749 (2024).

59. Zhang, D., Cao, Q., Guo, Y. & Wang, L. Assistant teaching system for computer hardware courses based on large language model. In *International Conference on Computer Science and Education*, 301–313 (Springer, 2023).

60. Min, S. *et al.* Rethinking the role of demonstrations: What makes in-context learning work? (2022). 2202.12837.

61. Wei, J. *et al.* Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846* (2023).

62. Tang, L. *et al.* Evaluating large language models on medical evidence summarization. *npj Digital Medicine* **6**, 158 (2023).

63. Asai, A. *et al.* Reliable, adaptable, and attributable language models with retrieval. *ArXiv* **abs/2403.03187** (2024). URL https://api.semanticscholar.org/CorpusID:268248911.

64. Khattab, O. & Zaharia, M. A. Colbert: Efficient and effective passage search via contextualized late interaction over bert. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020). URL https://api.semanticscholar.org/CorpusID:216553223.

65. Gao, T., Yen, H., Yu, J. & Chen, D. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627* (2023).

66. Zhang, Y. *et al.* When large language models meet citation: A survey. *arXiv preprint arXiv:2309.09727* (2023).

67. Yik, B. J. & Dood, A. J. Chatgpt convincingly explains organic chemistry reaction mechanisms slightly inaccurately with high levels of explanation sophistication. *Journal of Chemical Education* **101**, 1836–1846 (2024).

68. Introducing PPLX Online LLMs. URL https://www.perplexity.ai/hub/blog/introducing-pplx-online-llms.

69. Alawwad, H. A., Alhothali, A. M., Naseem, U., Alkhathlan, A. & Jamal, A. F. Enhancing textbook question answering task with large language models and retrieval augmented generation. *ArXiv* **abs/2402.05128** (2024). URL https://api.semanticscholar.org/CorpusID:267547633.

70. Soong, D. *et al.* Improving accuracy of gpt-3/4 results on biomedical data using a retrieval-augmented language model (2023). 2305.17116.

71. Uchiyama, S., Umemura, K. & Morita, Y. Large language model-based system to provide immediate feedback to students in flipped classroom preparation learning. *arXiv preprint arXiv:2307.11388* (2023).

72. Leinonen, J. *et al.* Comparing code explanations created by students and large language models. In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1*, 124–130 (2023).

73. Deng, Y., Lei, W., Liao, L. & Chua, T.-S. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. *arXiv preprint arXiv:2305.13626* (2023).

74. Kung, T. H. *et al.* Performance of chatgpt on usmle: potential for ai-assisted medical education using large language models. *PLoS digital health* **2**, e0000198 (2023).

75. Maier, P. & Klinker, G. J. Augmented chemical reactions: 3d interaction methods for chemistry. *Int. J. Online Biomed. Eng.* **9**, 80–82 (2013). URL https://api.semanticscholar.org/CorpusID:16201911.

76. Núñez, M. R. Á., Quirós, R., Núñez, I., Carda, J. B. & Camahort, E. Collaborative augmented reality for inorganic chemistry education (2008). URL https://api.semanticscholar.org/CorpusID:17160050.

77. Humphry, T. & Fuller, A. L. Potential chatgpt use in undergraduate chemistry laboratories. *Journal of Chemical Education* **100**, 1434–1436 (2023).

78. Cramer, C. J. *Essentials of Computational Chemistry: Theories and Models* (Wiley, 2014).

79. Prozument, K. *et al.* Photodissociation transition states characterized by chirped pulse millimeter wave spectroscopy. *Proceedings of the National Academy of Sciences* **117**, 146–151 (2020). URL https://www.pnas.org/doi/abs/10.1073/pnas.1911326116. https://www.pnas.org/doi/pdf/10.1073/pnas.1911326116.

80. Roothaan, C. C. J. New developments in molecular orbital theory. *Rev. Mod. Phys.* **23**, 69–89 (1951). URL https://link.aps.org/doi/10.1103/RevModPhys.23.69.

81. BOYS, S. F., COOK, G. B., REEVES, C. M. & SHAVITT, I. Automatic fundamental calculations of molecular structure. *Nature* **178**, 1207–1209 (1956). URL https://doi.org/10.1038/1781207a0.

82. Buenker, R. & Peyerimhoff, S. Ab initio scf calculations for azulene and naphthalene. *Chemical Physics Letters* **3**, 37–42 (1969). URL https://www.sciencedirect.com/science/article/pii/000926146980014X.

83. Raucci, U. *et al.* Interactive quantum chemistry enabled by machine learning, graphical processing units, and cloud computing. *Annual Review of Physical Chemistry* **74**, 313–336 (2023). URL https://www.annualreviews.org/content/journals/10.1146/annurev-physchem-061020-053438.

84. Nandy, A. *et al.* Computational discovery of transition-metal complexes: From high-throughput screening to machine learning. *Chemical Reviews* **121**, 9927–10000 (2021). URL `https://doi.org/10.1021/acs.chemrev.1c00347`.

85. Mardirossian, N. & Head-Gordon, M. Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Molecular Physics* **115**, 2315–2372 (2017). URL `https://doi.org/10.1080/00268976.2017.1333644`. `https://doi.org/10.1080/00268976.2017.1333644`.

86. AI4Science, M. R. & Quantum, M. A. The impact of large language models on scientific discovery: a preliminary study using gpt-4 (2023). `2311.07361`.

87. Poesia, G. *et al.* Synchromesh: Reliable code generation from pre-trained language models. *arXiv preprint arXiv:2201.11227* (2022).

88. Vaithilingam, P., Zhang, T. & Glassman, E. L. Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. In *Chi conference on human factors in computing systems extended abstracts*, 1–7 (2022).

89. Xu, F. F., Alon, U., Neubig, G. & Hellendoorn, V. J. A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, 1–10 (2022).

90. Liu, J., Xia, C. S., Wang, Y. & Zhang, L. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *Advances in Neural Information Processing Systems* **36** (2024).

91. Ribeiro, F., de Macedo, J. N. C., Tsushima, K., Abreu, R. & Saraiva, J. Gpt-3-powered type error debugging: Investigating the use of large language models for code repair. In *Proceedings of the 16th ACM SIGPLAN International Conference on Software Language Engineering*, 111–124 (2023).

92. Kang, S., Chen, B., Yoo, S. & Lou, J.-G. Explainable automated debugging via large language model-driven scientific debugging. *arXiv preprint arXiv:2304.02195* (2023).

93. MacNeil, S. *et al.* Decoding logic errors: A comparative study on bug detection by students and large language models. In *Proceedings of the 26th Australasian Computing Education Conference*, 11–18 (2024).

94. Al-Hossami, E., Bunescu, R., Smith, J. & Teehan, R. Can language models employ the socratic method? experiments with code debugging. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, 53–59 (2024).

95. Tian, R. *et al.* Debugbench: Evaluating debugging capability of large language models. *arXiv preprint arXiv:2401.04621* (2024).

96. O'Connor, M. *et al.* Sampling molecular conformations and dynamics in a multiuser virtual reality framework. *Science Advances* **4**, eaat2731 (2018). URL `https://www.science.org/doi/abs/10.1126/sciadv.aat2731`. `https://www.science.org/doi/pdf/10.1126/sciadv.aat2731`.

97. Wang, Y. *et al.* Interachem: Exploring excited states in virtual reality with ab initio interactive molecular dynamics. *Journal of Chemical Theory and Computation* **18**, 3308–3317 (2022). URL `https://doi.org/10.1021/acs.jctc.2c00005`.

98. Lindsay, D. *Scientific writing= thinking in words* (Csiro Publishing, 2020).

99. Kuhn, D. Science as argument: Implications for teaching and learning scientific thinking. *Science education* **77**, 319–337 (1993).

100. Lee, M., Liang, P. & Yang, Q. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, 1–19 (2022).

101. Salvagno, M., Taccone, F. S. & Gerli, A. G. Can artificial intelligence help for scientific writing? *Critical care* **27**, 75 (2023).

102. Rojas, A. J. An investigation into chatgpt's application for a scientific writing assignment. *Journal of Chemical Education* **101**, 1959–1965 (2024).

103. Gero, K. I., Liu, V. & Chilton, L. Sparks: Inspiration for science writing using language models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*, 1002–1019 (2022).

104. Prillaman, M. Is chatgpt making scientists hyper-productive? the highs and lows of using ai. *Nature* (2024).

105. Perkins, M. Academic integrity considerations of ai large language models in the post-pandemic era: Chatgpt and beyond. *Journal of university teaching & learning practice* **20**, 07 (2023).

106. Fütterer, T. *et al.* Chatgpt in education: global reactions to ai innovations. *Scientific reports* **13**, 15310 (2023).

107. Cotton, D. R., Cotton, P. A. & Shipway, J. R. Chatting and cheating: Ensuring academic integrity in the era of chatgpt. *Innovations in Education and Teaching International* **61**, 228–239 (2024).

108. Editorial policies: Artificial intelligence (ai). `https://www.nature.com/nature-portfolio/editorial-policies/ai`. [Online; accessed 19-March-2024].

109. Change to policy on the use of generative ai and large language models. `https://www.science.org/content/blog-post/change-policy-use-generative-ai-and-large-language-models`, (2023). [Online; accessed 19-March-2024].

110. Kellaghan, T. & Greaney, V. *Using assessment to improve the quality of education* (Unesco, International Institute for Educational Planning, 2001).

111. Harris, D. N. & Sass, T. R. Teacher training, teacher quality and student achievement. *Journal of public economics* **95**, 798–812 (2011).

112. Bloom, B. S. *et al.* Handbook on formative and summative evaluation of student learning. (1971).

113. Becker, W. E. & Watts, M. How departments of economics evaluate teaching. *American Economic Review* **89**, 344–349 (1999).

114. Dai, W. *et al.* Can large language models provide feedback to students? a case study on chatgpt. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, 323–325 (IEEE, 2023).

115. Fergus, S., Botha, M. & Ostovar, M. Evaluating academic answers generated using chatgpt. *Journal of Chemical Education* **100**, 1672–1675 (2023).

116. Pinto, G. *et al.* Large language models for education: Grading open-ended questions using chatgpt. In *Proceedings of the XXXVII Brazilian Symposium on Software Engineering*, 293–302 (2023).

117. Gao, R., Thomas, N. & Srinivasa, A. Work in progress: Large language model based automatic grading study. In *2023 IEEE Frontiers in Education Conference (FIE)*, 1–4 (IEEE, 2023).

118. Alqahtani, T. *et al.* The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. *Research in Social and Administrative Pharmacy* (2023).

119. Yan, L. *et al.* Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology* **55**, 90–112 (2024).

120. Stevenson, C., Smal, I., Baas, M., Grasman, R. & van der Maas, H. Putting gpt-3's creativity to the (alternative uses) test. *arXiv preprint arXiv:2206.08932* (2022).

121. Sinha, R., Song, Z. & Zhou, T. A mathematical abstraction for balancing the trade-off between creativity and reality in large language models. *arXiv preprint arXiv:2306.02295* (2023).

122. Organisciak, P., Acar, S., Dumas, D. & Berthiaume, K. Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity* **49**, 101356 (2023).

123. Exintaris, B., Karunaratne, N. & Yuriev, E. Metacognition and critical thinking: using chatgpt-generated responses as prompts for critique in a problem-solving workshop (smartchemper). *Journal of Chemical Education* **100**, 2972–2980 (2023).

124. Clark, T. M. Investigating the use of an artificial intelligence chatbot with general chemistry exam questions. *Journal of Chemical Education* **100**, 1905–1916 (2023).

125. Marion, M. *et al.* When less is more: Investigating data pruning for pretraining llms at scale. *arXiv preprint arXiv:2309.04564* (2023).

126. Wellawatte, G. P. & Schwaller, P. Extracting human interpretable structure-property relationships in chemistry using xai and large language models. *arXiv preprint arXiv:2311.04047* (2023).

127. Geng, S., Döner, B., Wendler, C., Josifoski, M. & West, R. Sketch-guided constrained decoding for boosting blackbox large language models without logit access. *arXiv preprint arXiv:2401.09967* (2024).

128. Zhang, Y. *et al.* Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219* (2023).

129. Lund, B. D. *et al.* Chatgpt and a new academic reality: Artificial intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology* **74**, 570–581 (2023).

130. Kirk, H. R., Vidgen, B., Röttger, P. & Hale, S. A. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence* **6**, 383–392 (2024). URL https://doi.org/10.1038/s42256-024-00820-y.

131. Koenecke, A. *et al.* Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences* **117**, 7684–7689 (2020).

132. Ziems, C., Chen, J., Harris, C., Anderson, J. & Yang, D. Value: Understanding dialect disparity in nlu. *arXiv preprint arXiv:2204.03031* (2022).

133. Yong, Z.-X., Menghini, C. & Bach, S. H. Low-resource languages jailbreak gpt-4. *arXiv preprint arXiv:2310.02446* (2023).

134. Ghosh, S. & Caliskan, A. Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 901–912 (2023).

135. Hofmann, V., Kalluri, P. R., Jurafsky, D. & King, S. Dialect prejudice predicts ai decisions about people's character, employability, and criminality. *arXiv preprint arXiv:2403.00742* (2024).