Journal of Artificial Intelligence and Technology, 2026, 6, 646-653
https://doi.org/10.37965/jait.2026.0942

ISTP
RESEARCH ARTICLE

The Assessment of ChatGPT and DeepSeek to Solve Chemistry Exams and the Classification of the Exam Questions According to Bloom's Taxonomy in order to Alter the Level of Exams. A Comparative Study

Ali Ghadban,[1,2†] Fatima Al Khatib,[1†] Hanan Rahal,[1†] and Sanaa Khaled[1†]

[1]Chemical Sciences Laboratory (CSL@LIU), Department of Biological and Chemical Sciences, School of Arts and Sciences, Lebanese International University, Beirut, P.O. Box 146404, Lebanon

[2]Department of Chemistry and Biochemistry, Faculty of Sciences, Lebanese University, Zahle, Lebanon

[†]Authors contributed equally to this work.

(Received 22 September 2025; Revised 11 February 2026; Accepted 27 May 2026; Published online 06 July 2026)

Abstract: This study examines the performance of ChatGPT and DeepSeek in solving undergraduate chemistry exams (general, organic, inorganic, and physical) and their ability to classify questions by Bloom's Taxonomy to adjust cognitive levels. Both models perform well on general chemistry, with DeepSeek scoring about eight points higher, and both succeed on direct calculation tasks (e.g., kinetics, atomic structure, and Bohr model). ChatGPT struggles with complex numerical problems (e.g., Arrhenius equation) and does not handle diagram based tasks (e.g., orbital diagrams). In physical chemistry, both achieve near-perfect scores on thermodynamics questions, while performance on organic chemistry is low (38% for ChatGPT vs. 23% for DeepSeek), reflecting difficulty with stereochemistry, acid-base chemistry, resonance, and spatial structures. In inorganic chemistry, both perform moderately (62% vs. 56%), solving simple but not symmetry or group theory-related problems. Statistical analyses across multiple-choice questions show that differences between models are small and not statistically significant for any exam: organic chemistry (37.5% vs. 25.0%, $p = 0.63$), inorganic (84% vs. 68%, $p = 0.72$), general (100% vs. 86.7%, $p = 0.50$), and physical chemistry (100% vs. 95.2%, $p = 1.00$), indicating only potential advantages that cannot be generalized. Both classify most questions correctly by Bloom's levels, though errors occur at the apply/analyze levels. ChatGPT is more effective in adjusting question difficulty by moving through Bloom's verbs.

Keywords: AI; Bloom's taxonomy; ChatGPT; DeepSeek; education; general chemistry; inorganic chemistry; organic chemistry; physical chemistry

I. INTRODUCTION

Since Turing's 1950 question, "Can machines think?", AI has evolved from infancy to rapid growth, with generative AI enabling tasks that once took hours or days to be completed in seconds [1,2]. Tasks that used to take hours or days now it takes few seconds with AI tools [3]. It is used in various domains starting from medicine [4,5], to engineering[6,7], and basic sciences [8–13]. Students and educators have been using these AI tools in studying [14,15], preparing exams or quizzes [16,17], and generating or evaluating courses' syllabi [18,19]. The shift to online learning during events like COVID-19 and regional conflicts exposed vulnerabilities in STEM assessments, as students increasingly used AI tools like ChatGPT, raising academic integrity concerns [20]. Research shows a nearly 200% spike in use of illicit academic help services (e.g., Chegg) for chemistry during remote learning (April–August 2020), highlighting students' reliance on third-party support even in exams [21]. Generative chatbots like ChatGPT (2022) and DeepSeek (2023) pose new challenges, as students may not see their use as cheating. Authorities advise assessment reform and AI ethics education rather than relying on pen-and-paper exams or unreliable detection [22]. For example, researchers successfully submitted AI-generated exam answers that passed human markers nearly undetected, prompting calls for redesigning assessment formats [23]. DeepSeek's chain of thought transparency and multimodal capabilities reportedly boost its performance especially in technical and science domains, sometimes surpassing ChatGPT in elaborate reasoning and chemical problem-solving [24]. Other recent studies highlight that DeepSeek outperforms ChatGPT in generating bibliographic references and retrieving accurate academic data, though no model yet guarantees full accuracy [25]. Its growth is mirrored by major Chinese universities integrating DeepSeek into curricula both to teach AI and to address security and ethical concerns [26].

Bloom's Taxonomy, developed by Benjamin Bloom in 1956 and revised in 2001, classifies educational goals by cognitive complexity [27,28]. Bloom's Taxonomy, a hierarchy from basic recall to complex creation, guides learning objectives, instruction, and assessment. It promotes higher-order thinking: analysis, evaluation, creation, ensuring alignment between curriculum, teaching,

Corresponding author: Ali Ghadban (e-mail: ali.ghadban@liu.edu.lb).

and assessment for deeper, transferable learning [29]. Despite the growing use of LLMs in education and chemistry assessment, few studies have systematically evaluated ChatGPT and DeepSeek on chemistry exam questions across multiple disciplines, and existing work has not fully examined performance according to Bloom's taxonomy. The aim of this study is to address these gaps by evaluating the effectiveness of ChatGPT vs. DeepSeek at answering chemistry exam questions (general, organic, inorganic, and physical chemistry exams), classification of the exam questions according to Bloom's verbs, rephrasing the exam questions using both AI chatbots in order to alter the level of the questions and hence lowering cheating levels for any upcoming online assessments.

The remainder of the paper is structured as follows: Section II presents a review of prior studies on the application of large language models in education and question-solving, highlighting gaps addressed by this study. Section III details the methodology, including the chemistry exams and procedures for evaluating ChatGPT and DeepSeek. Section IV presents the results and discussion of the models' performance in solving chemistry exam questions. Section V examines the ability of both chatbots to classify and rephrase exam questions according to Bloom's Taxonomy. Finally, Section VI concludes the paper and outlines directions for future work.

II. RELATED WORK

Recent machine learning approaches have automated classifying exam questions by Bloom's levels. Khalifa and Mohamed [30] used term frequency–inverse document frequency and a support vector machine to achieve ~80% accuracy on 600 questions. Reviews highlight the importance of verbs and parts of speech for better classification [31]. Transformer-based classifiers like DistilBERT have demonstrated validation accuracies near 91%, notably improving performance on higher-order levels (Apply through Create) [32]. Hwang *et al.* studied AI methods for generating and evaluating Bloom's Taxonomy-aligned multiple-choice questions (MCQs) in chemistry and biology using GPT-3.5, RoBERTa, and expert review, showing the models can create higher-order thinking questions and align with human complexity assessments [33]. Fergus *et al.* found that ChatGPT gives well-written chemistry answers but struggles with complex application, interpretation, and non-text information, suggesting assessments emphasize problem-solving and data interpretation [34]. Studies showed that ChatGPT helped chemistry students improve writing and lab reports, reduced grammar errors [35], support lesson planning [36], and assisted with general and analytical chemistry calculations [37]. A recent study found that ChatGPT correctly answered 80% of high school general chemistry entrance exam questions and aided in generating new questions through prompt engineering [38]. Clark [39] discussed ChatGPT's ability to solve general chemistry exam questions. It excelled at recognizing concepts in closed-response questions but had a problem-solving success rate of 44%, below the class average of 69%. Its open-response answers showed strong language skills but often included flawed yet convincing explanations. Sorenson and Hanson [40] used Rasch analysis on general chemistry exams and accurately identified AI-influenced answers with few false alarms, showing AI-generated patterns are detectable. In an earlier study, they found no evidence of increased cheating in unproctored online exams compared to prior in-person exams [41]. Kassem *et al.* [13] compared ChatGPT and Gemini (formerly Bard) on basic organic topics, including nomenclature, classification, molecule drawing, and identifying resonance contributors. Other studies explored general chemistry applications of ChatGPT [42], but, to our knowledge, none evaluated its performance on organic, physical, and inorganic chemistry exams. Moreover, few studies investigated DeepSeek's use in chemistry exams. To our knowledge, no peer-reviewed study compared ChatGPT and DeepSeek on chemistry exams under controlled Bloom's taxonomy conditions. Bouchra *et al.* [24] provided a broad comparison in science education but lacked peer review, DOI, and analysis by question difficulty.

III. METHODOLOGY

All data used in this study were anonymized and reported as class averages. Permission to use these data for research and publication purposes was obtained from the Lebanese International University (LIU). The study did not involve any identifiable student information, and all procedures were conducted in accordance with institutional guidelines for research ethics. The exams selected for this study were administered during the Fall 2023–2024 semester (hereafter referred to as Fall 2023). One examination from each course was analyzed (one general, one physical, one organic, and one inorganic), and the number of questions in each exam is reported in Table I.

These chemistry courses at LIU enrolled students from prepharmacy, biomedical, nutrition, chemistry, biochemistry, and biology majors. Fall 2023 was the first fully on-campus semester after COVID-19, with exams adjusted for the prior online period. A similar online/hybrid approach occurred during Lebanon's October 2024 war (Fall 2024), when students, more familiar with AI, likely used chatbots for cheating. Fall 2024 exam results were compared with Fall 2023 to assess potential cheating. The data were collected for both chatbots in December 2024 and January 2025. Every exam question was fed to the said AI chatbot, and the following prompts were asked in order:

1. Answer the question
2. Classify the question according to Bloom's taxonomy
3. Use the above example to write the question in another way using the other verbs of Bloom's taxonomy.

Sometimes, the chatbot did the paraphrasing of step 3 above for one verb only from Bloom's taxonomy so we used a clearer prompt and asked to use the other Bloom's taxonomy verbs by specifying the category as remember, understand, apply, analyze, evaluate, and create. The results of this study (grading and question paraphrasing) were tabulated so an easy comparison could be made between ChatGPT and DeepSeek. The results of all outcomes are summarized in the supporting information in tables as Table Si. For this study, DeepSeek operated in its deep-thinking mode, while ChatGPT utilized its internal reasoning mode to generate responses.

Statistical analyses were performed using IBM SPSS Statistics. To evaluate the performance of ChatGPT and DeepSeek on MCQs across the examined chemistry exams, we conducted a

Table I. Number and type of questions analyzed per exam

Exam	Number of MCQs/true or false/matching	Number of subjective questions
General	22	5
Physical	21	5
Organic	16	7
Inorganic	25	6

paired analysis at the individual question level. Each MCQ was coded as a binary outcome (1 = correct, 0 = incorrect) for both models. Because the same questions were answered by both models, differences in performance were analyzed using McNemar's exact test, which is appropriate for paired binary data and provides an exact p-value for small sample sizes. Discordant pairs, where one model answered correctly while the other did not, were used to calculate the odds ratio, serving as a measure of effect size, with 95% confidence intervals computed using exact methods. Paired outcome tables and odds ratios were reported in the Supporting Information, while summary statistics, effect sizes, and p-values were presented in the main text to provide both descriptive and inferential insights into the relative performance of the two models. Statistical analysis was restricted to MCQs because they provide objective, binary outcomes, whereas subjective questions involve grading variability that limits reliable statistical comparison in a single exam setting.

IV. RESULTS AND DISCUSSION

This study evaluates ChatGPT-4.o and DeepSeek on four undergraduate chemistry exams (general, physical, inorganic, organic) at LIU, including their ability to classify and paraphrase questions by Bloom's taxonomy. DeepSeek takes longer to respond (20 seconds to minutes) and sometimes has server issues, while ChatGPT responds instantly with no delays. DeepSeek can upload unlimited attachments, whereas ChatGPT requires a paid plan. Exam grades are summarized in Fig. 1. Both chatbots passed the general and physical chemistry exams with overall grades above 80%, but scored below 60% on inorganic and organic chemistry. On the physical chemistry exam, ChatGPT scored 98% and DeepSeek 100%. DeepSeek outperforms ChatGPT in inorganic (62% vs. 56%) and general chemistry (90% vs. 82%), while ChatGPT scores higher in organic chemistry (38% vs. 23%). Comparing with class averages in Fall 2023 and 2024, physical and general chemistry averages rise by ~10%, with exams of comparable difficulty and students coming from online or hybrid backgrounds. Although the increase coincides with chatbot availability, their impact cannot be separated from instructional or cohort factors. Organic chemistry averages drop from 80% in Fall 2023 to 57% in Fall 2024, but no conclusions about chatbot effects can be drawn due to possible differences in instruction, assessments, or cohorts. Inorganic chemistry averages remain comparable, likely reflecting on-campus exams.

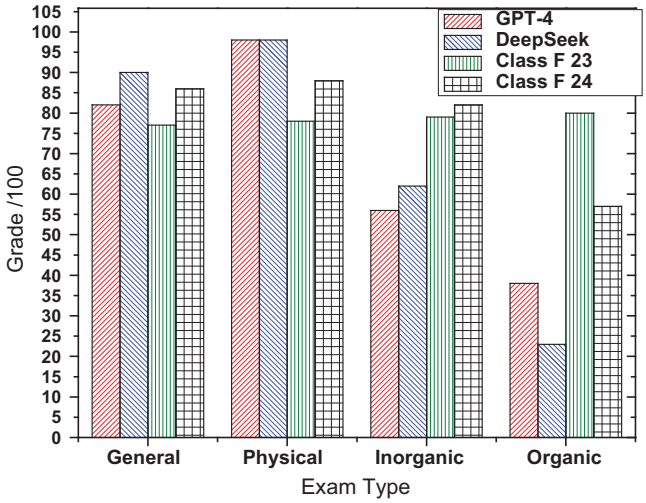

Fig. 1. The grades of ChatGPT-4.o and DeepSeek obtained for four chemistry exams of Fall 2023. A comparison with class averages of Fall 2023 and Fall 2024 is shown as well.

A. SOLVING ORGANIC CHEMISTRY EXAM

The main topics discussed in this organic chemistry exam are revision topics from general chemistry, resonance, acids and bases, stereochemistry, and reactions of alkanes. Both ChatGPT and DeekSeek perform poorly in most of the organic chemistry exam topics. The results are summarized in Table S1 (in supporting info) with the grades, and the comments on every question are shown. A summary of the main topics is shown in Table II. It has been found that both chatbots cannot handle drawing structures or analyzing molecules in space, which is reflected in lower accuracy in such problem types. Both chatbots also show difficulties with stereochemistry problems such as assigning configuration, showing relationship between structures, assigning chirality/optical activity of molecules, and in converting a perspective formula to Fischer projection. DeepSeek misanalyzes groups around the chiral center and can replace atoms by other atoms, and the result in most cases is either wrong or correct by chance. On the other hand, ChatGPT can identify clearly the groups around the chiral center, number them correctly and can assign configuration when the least priority group is at the back in perspective formulas (dash-wedge).

However, when assigning configurations in Fischer projections, ChatGPT swaps the horizontal lines with the vertical lines. In other words, according to ChatGPT the vertical and horizontal lines in Fischer projection are the towards and backward positions, respectively. The latter ends up with structures having an opposite configuration. Both chatbots cannot convert perspective formulas to Fischer projections or a Newmans projection to line angle formula. This finding is particularly concerning given research by Hsin-Kai [43] who emphasized the critical importance of visual-spatial reasoning in chemistry education. Drawing resonance structures in organic chemistry is one of the most tedious tasks a student finds and that's why too many papers have been published to aid students and instructors [44–46]. Both chatbots still have

Table II. Main topics covered in the organic chemistry exam with comments on the outcomes of the answers

Question Type	ChatGPT	DeepSeek
Stereochemistry	Weak	Weak
Resonance	Can describe movement of electrons but cannot draw	Analyze wrong structure
Acidity and basicity	Can discuss the factors but cannot manage to rank correctly in most of times	Can discuss the factors but cannot manage to rank correctly in most of times
Halogenation of Alkanes	1. Can predict product of easy molecules 2. Show mechanism but without drawing arrows 3. Failure with molecules with quaternary carbons	1. Can predict product of easy molecules 2. Show mechanism but without drawing arrows 3. Failure with molecules with quaternary carbons

problems in drawing resonance structures in spite of the fact that both chatbots state the rules of drawing resonance but fails to draw it. DeepSeek always analyzes the wrong structure, and it investigates the resonance structures of ions already found in its data base as nitrate or carbonate ions. On the other hand, ChatGPT can analyze the structure (in most cases), explain how electrons must move, but cannot provide a drawing (Table S1). Another important topic in organic chemistry is the ranking of acidity of molecules without having the pK_a values. Drawing the conjugate bases and examining their stability is the key point to compare acidity. Both chatbots do not draw conjugate bases in their answers and sometimes misinterpreted the structures being analyzed. However, the factor affecting the acidity is correctly stated in the examples for both ChatGPT and DeepSeek but the outcomes of both are partially correct: either a problem in analyzing the molecule (for both) or choosing the wrong choice in MCQs as is the case of ChatGPT. Likewise, analyzing basicity of molecules faces similar problems. Predicting the product of organic reactions is an important task that a student should learn. The free radical substitution reaction of alkanes is one of the first organic reactions that we teach to students. Students have to draw the products and show the mechanism. It has been shown that both chatbots can correctly predict the product of such reaction by providing the name of the product but without drawing it. The free radical mechanism of the reaction, showing the initiation, propagation, and termination steps, is correctly provided by both chatbots but without the use of fish-hook arrows. The radical is placed on the wrong atoms in the structure in some cases. The latter example is provided for 2-methylbutane. Both chatbots cannot manage to analyze more complicated structures such as 1,1-dimethylcyclobutane or 2,2,3-trimethylbutane where a quaternary carbon is found in the structure. The products resulting from the substitution of the quaternary carbon is provided in both examples. Hence, once asked about the number or drawing the monohalogenated products, or the most stable radical obtained after removal of hydrogen, the wrong answer is provided. Table III summarizes the performance of ChatGPT and DeepSeek on the MCQs ($n = 16$) of the chemistry exam, along with paired statistical analysis. In this study, ChatGPT achieves a higher numerical accuracy on the MCQs than DeepSeek (37.5% vs. 25.0%), but the difference is not statistically significant (McNemar's exact test, $p = 0.63$; OR = 3.0, 95% CI: 0.30–30.6). The wide confidence interval reflects the limited number of discordant items, indicating substantial uncertainty. These findings suggest a potential advantage of ChatGPT for MCQs on this exam, but they should be interpreted cautiously and cannot be generalized to other exams or question types.

B. SOLVING INORGANIC CHEMISTRY EXAM

The inorganic chemistry exam covered 25 MCQs and several subjective problems, covering major topics including hybridization, bond order, molecular orbital (MO) theory, symmetry and point group assignment, magnetic properties, and periodic trends. The results of the questions are summarized in Table S2 (in supporting info). Overall, both ChatGPT and DeepSeek handle simple recall-based questions successfully but show consistent weaknesses when higher-level reasoning, orbital visualization, or symmetry analysis is required. The results are summarized in Table IV. Hybridization and simple bonding questions are among the best performed. For example, both chatbots correctly assign sp^3 hybridization to phosphorus in PCl_3 and sp hybridization to beryllium in $BeCl_2$. However, ChatGPT frequently fails to connect hybridization with the correct geometry (e.g., describing $BeCl_2$ as bent instead of linear). DeepSeek generally offers clearer justifications, though sometimes confusing electron domain geometry with molecular geometry. These issues parallel the organic chemistry results, where the models could recall terms but misapply them in structural reasoning. MO theory reveals much greater challenges. While both models correctly describe O_2 as paramagnetic with a bond order of two, they struggle with less common diatomics such as Li_2, He_2, and Be_2. ChatGPT tends to misapply the bond order formula, inflating the stability of unstable molecules. DeepSeek occasionally draws or describes MO diagrams correctly but gave contradictory bond orders. These inconsistencies highlight a gap in stepwise reasoning, similar to the organic part where resonance rules are recalled but not applied to actual structures. Symmetry and group theory are particularly problematic. In predicting IR and Raman activity of PCl_3 vibrations, ChatGPT offers short answers

Table III. MCQ performance and paired statistical comparison between ChatGPT and DeepSeek in the organic chemistry exam. Odds ratio and confidence interval were calculated from discordant pairs (ChatGPT correct/DeepSeek incorrect = 3; ChatGPT incorrect/DeepSeek correct = 1)

Model	Correct (*n*)	Accuracy (%)	Statistical test	Effect size (95% CI)	p-value
ChatGPT	6/16	37.5	McNemar's exact test	Odds ratio = 3.0 (95% CI: 0.30–30.6)	0.63
DeepSeek	4/16	25.0	—	—	—

Table IV. Main topics covered in the inorganic chemistry exam with comments on the outcomes of the answers

Question type	ChatGPT	DeepSeek
Hybridization (PCl_3, $BeCl_2$)	Correct hybridization, weak geometry link	Correct, better geometry link
σ and π bonds (C_2H_4)	Correct: $5\sigma + 1\pi$	Correct
MO bond order (O_2, Li_2, He_2, Be_2)	O_2 correct, others misapplied	O_2 correct, others inconsistent
Symmetry/point groups (OF_2, SF_6, trans-N_2F_2)	OF_2 and SF_6 correct, trans-N_2F_2 incomplete	Slightly better in trans-N_2F_2
IR/Raman activity (PCl_3)	Short answer, no reasoning	Partial reasoning
Magnetic properties (O_2, N_2, others)	O_2 and N_2 correct, others inconsistent	O_2 and N_2 correct, others slightly better
Subjective (definitions, explanations)	Good on definitions, weak in applications	Good on definitions, better but still weak in applications

Table V. MCQ performance and paired statistical comparison between ChatGPT and DeepSeek in the inorganic chemistry exam. Discordant pairs: ChatGPT correct/DeepSeek incorrect = 3; ChatGPT incorrect/DeepSeek correct = 5

Model	Correct (*n*)	Accuracy (%)	Statistical test	Effect size (95% CI)	*p*-value
ChatGPT	17/25	68	McNemar's exact test	Odds ratio = 0.6 (95% CI: 0.12–3.0)	0.72
DeepSeek	21/25	84	—	—	—

without using character tables, while DeepSeek attempts symmetry-based reasoning but often misidentifies key symmetry elements. Both correctly recognize the inversion center in SF_6, but analysis of lower-symmetry molecules such as trans-N_2F_2 is inconsistent, with ChatGPT missing a mirror plane while DeepSeek identifies most elements correctly. This mirrors the organic stereochemistry questions, where superficial recognition led to errors in spatial interpretation.

Magnetic property questions show mixed results. Both chatbots correctly classify O_2 as paramagnetic and N_2 as diamagnetic, but ChatGPT sometimes ignores antibonding effects in borderline cases, while DeepSeek hesitates on spin states. Both recall definitions well (e.g., coordination number) but struggle with applied reasoning, such as explaining crystal field trends or coordination geometries, mirroring their organic chemistry performance.

On the 25 MCQs, ChatGPT answered 17 correctly (68.0%) and DeepSeek answered 21 correctly (84.0%). Paired comparison using McNemar's exact test shows no statistically significant difference between the models ($p = 0.72$), with an odds ratio of 0.6 (95% CI: 0.12–3.0), reflecting a small, non-significant tendency favoring DeepSeek (Table V). The wide confidence interval highlights the uncertainty due to the limited number of discordant items, suggesting that while DeepSeek achieved higher numerical accuracy on this exam, the difference should be interpreted cautiously and cannot be generalized beyond this specific assessment.

C. SOLVING GENERAL CHEMISTRY EXAM

The results of the answers to questions are summarized in Table S3 in supporting information and in Table VI (general summary of topics). Both AI systems demonstrate strong performance in kinetics problems, for instance finding the rate law of a reaction using the method of the initial rate (Questions 1a–1e and 2a–2c), with perfect or near-perfect scores on most sub-questions. However, a critical difference emerges in Question 2b related to the calculation of the rate constant of the reaction from the half-life, where ChatGPT received 0 points despite providing correct analysis methodology, while DeepSeek achieved full marks with both correct analysis and accurate final answer. This discrepancy highlights the importance of computational precision in chemistry problem-solving, where conceptual understanding alone is insufficient without accurate mathematical execution. A significant weakness identified in ChatGPT is its difficulty handling large numbers and complex calculations, particularly evident in calculations using the Arrhenius equation (MCQ-4). The evaluators noted that ChatGPT has problems dealing with large numbers, while DeepSeek has no problem dealing with calculations and large numbers, and the steps of calculations are clear. The discrepancy may reflect differences in training methodologies or computational architectures between the models. The practical implications are significant, as chemistry frequently involves calculations with scientific notation, logarithms, and exponential functions that require high precision. The most striking performance differences occurred in atomic structure problems (Questions 3a–3d), where both AI systems received zero points across all sub-questions. This consistent failure suggests fundamental limitations in visual-spatial reasoning and diagram interpretation capabilities.

The evaluators note that both systems misinterpret the orbital energy diagram, with ChatGPT providing "$1s^2 2s^2 2p^4$" instead of the correct electron configuration, while DeepSeek does not calculate the number of electrons in the p orbitals. This systematic error propagates through all related questions, highlighting the interconnected nature of chemistry concepts. The inability of both AI systems to accurately interpret orbital diagrams suggests a fundamental limitation that can significantly impact their utility in chemistry instruction and assessment. Both AI systems demonstrate strong performance on multiple-choice questions, with ChatGPT scoring 39/45 and DeepSeek scoring 42/45. This pattern aligns with previous research by Goorts *et al.* [47] who found that AI systems generally perform well on structured, multiple-choice assessments that rely on pattern recognition and knowledge recall. The consistent performance across MCQ items suggests that both systems possess robust foundational chemistry knowledge, despite the computational and visual-spatial challenges identified in other question types. This finding has important implications for the types of assessment tasks where AI systems may be most effectively employed. In calculations requiring finding the wavelength or energy from transitions in Bohr model dealing with transitions in hydrogen or hydrogen-like ions, both chatbots can successfully handle such problems. Good performance of both chatbots is observed as well in writing the electron configuration of elements or ions.

Table VI. Main topics covered in the general chemistry exam with comments on the outcomes of the answers

Question type	ChatGPT	DeepSeek
Kinetics and mechanism	1. Strong in initial rates method 2. Weak competencies in finding rate constant from half-life 3. Failure with Arrhenius equation 4. Success in integrated rate law	1. Strong in initial rates method 2. Good competencies in finding rate constant from half-life 3. Can handle calculations using Arrhenius equation 4. Success in integrated rate law
Atomic structure	Failure in analyzing orbital energy diagram	Failure in analyzing orbital energy diagram
Bohr Model	Good abilities in calculations involving calculation of energies and wavelength of transitions	Good abilities in calculations involving calculation of energies and wavelength of transitions
Electron configuration	Good capabilities in writing expanded and shorthand electron configuration	Good capabilities in writing expanded and shorthand electron configuration

Table VII. MCQ performance and paired statistical comparison between ChatGPT and DeepSeek in the general chemistry exam. Discordant pairs: ChatGPT correct/DeepSeek incorrect = 0; ChatGPT incorrect/DeepSeek correct = 2

Model	Correct (*n*)	Accuracy (%)	Statistical test	Effect size (95% CI)	*p*-value
ChatGPT	13/15	86.7	McNemar's exact test	Odds ratio = 0 (95% CI: 0–3.8)	0.50
DeepSeek	15/15	100	—	—	—

Table VIII. Main topics covered in the physical chemistry exam with comments on the outcomes of the answers

Question Type	ChatGPT	DeepSeek
Work, enthalpy, internal energy, pressure, Zeroth and first law of thermodynamics	Very good skills	Very good skills
Calculating Enthalpy using bond energies and Hess's law	Very good skills	Very good skills

Table IX. MCQ performance and paired statistical comparison between ChatGPT and DeepSeek in the physical chemistry exam. Discordant pairs: ChatGPT correct/DeepSeek incorrect = 0; ChatGPT incorrect/DeepSeek correct = 1

Model	Correct (*n*)	Accuracy (%)	Statistical test	Effect size (95% CI)	p-value
ChatGPT	20/21	95.2	McNemar's exact test	Odds ratio = 0 (95% CI: 0–5.3)	1.00
DeepSeek	21/21	100	—	—	—

On the 15 MCQs, ChatGPT answered 13 correctly (86.7%) while DeepSeek answered all 15 correctly (100%). Paired comparison using McNemar's exact test shows no statistically significant difference ($p = 0.50$), with an odds ratio of 0 (95% CI: 0–3.8), reflecting a small number of discordant items (Table VII). These results suggest a numerical advantage for DeepSeek on this exam, but the difference should be interpreted cautiously and cannot be generalized beyond this assessment.

D. SOLVING PHYSICAL CHEMISTRY EXAM

The tested physical chemistry exam was mainly discussing simple topics and fundamentals in physical chemistry including the first law of thermodynamics. The detailed study is summarized in Table S4. The main topics (MCQs and subjective questions) discussed in the exam are questions asking to calculate for work, heat, internal energy, and enthalpy (Table VIII). Both chatbots can successfully tackle such questions with the correct reasoning. Adding, both chatbots could manage to calculate enthalpy using Hess's law and bond energies (Subjective questions). It is found that both AI platforms almost got a full mark (100% DeepSeek and 98 % for ChatGPT).

On the 21 MCQs, ChatGPT answered 20 correctly (95.2%) while DeepSeek answered all 21 correctly (100%). Paired comparison using McNemar's exact test shows no statistically significant difference ($p = 1.00$), with an odds ratio of 0 (95% CI: 0–5.3), reflecting only a single discordant question (Table IX). These results suggest a numerical advantage for DeepSeek on this exam, but the difference is minimal and should be interpreted cautiously.

V. CLASSIFICATION OF THE EXAM QUESTIONS ACCORDING TO BLOOM'S TAXONOMY

To reduce cheating in online chemistry exams and help instructors quickly generate alternative questions of varying difficulty, ChatGPT and DeepSeek classified exam questions using Bloom's taxonomy and then rephrased them across categories to build a question bank. The study covered general and organic chemistry exams, with detailed results in the supporting information (Tables S5–S6).

A. CLASSIFICATION OF THE ORGANIC CHEMISTRY EXAM QUESTIONS ACCORDING TO BLOOM'S TAXONOMY

No questions were classified as "remember." GPT classified one as "understand," both chatbots classified six as "apply," and DeepSeek classified six as "evaluate." Most questions fell under "analyze" (20 by GPT, 15 by DeepSeek). Results indicate incomplete coverage of Bloom's taxonomy, suggesting improvements for future exams. The chatbots agree on ~60% of classifications; most differences are minor and likely due to image-analysis limits or classification errors. Some application-level questions (e.g., optical activity, Newman projections, and chiral configurations) are occasionally misclassified as analysis. GPT shows better ability to adjust question levels across Bloom's taxonomy, while DeepSeek often keeps levels unchanged except at higher categories. For resonance structure questions, GPT-4 appropriately rephrases items at the remember and understand levels, with little change between apply and analyze, and higher difficulty at evaluate and create levels. DeepSeek instead simplifies the structure to the nitrite ion (NO_2^-), likely misinterpreting the original question. Although difficulty increased across categories, it was based on an incorrect structure. For acidity comparisons within a molecule, GPT-4 effectively adjusts question difficulty across Bloom's categories. In contrast, DeepSeek shows little variation from remember to apply, repeatedly asking for acidity ranking and conjugate-base stability. Similar trends appear in MCQs, where GPT-4 better matches question difficulty to category, while DeepSeek largely maintains the same level.

B. CLASSIFICATION OF THE GENERAL CHEMISTRY EXAM QUESTIONS ACCORDING TO BLOOM'S TAXONOMY

In the general chemistry exam, ChatGPT-4.o outperforms DeepSeek in classifying and generating questions across Bloom's

taxonomy, especially at higher cognitive levels. Both perform similarly well at the remember and understand levels, aligning with prior findings on AI strength in recall and basic comprehension. [32]. ChatGPT shows better contextual explanations that support understanding, while DeepSeek emphasizes procedural accuracy. The largest differences appear at higher cognitive levels, where ChatGPT generates more authentic, complex tasks. It creates realistic application scenarios and stronger critical thinking questions at the analyze and evaluate levels, while DeepSeek tends toward more formulaic tasks, an important finding for AI support of higher-order thinking in education [48]. At the create level, the gap is greatest, with ChatGPT generating tasks requiring experimental design, modeling, and concept synthesis.

VI. CONCLUSION

Both ChatGPT and DeepSeek performed well on undergraduate chemistry assessments, with variation across subdisciplines. DeepSeek slightly excelled in general chemistry, while both models succeeded in physical chemistry but struggled in organic and moderately in inorganic chemistry. Weaknesses appeared in tasks requiring numerical reasoning, spatial analysis, and diagram interpretation. ChatGPT better adjusted question difficulty across Bloom's levels. These findings highlight both the potential and current limitations of LLMs in chemistry education. Future work will evaluate newer models, course-aligned problems, and comparisons with other chatbots to support higher-order learning outcomes.

CONFLICT OF INTEREST STATEMENT

The author(s) declare that they have no conflicts of interest to report regarding the present study.

SUPPLEMENTARY INFORMATION

Supplementary data are available for this work.

REFERENCES

[1] A. M. Turing, "I.—Computing machinery and intelligence," *Mind*, vol. LIX, no. 236, pp. 433–460, Oct. 1950, DOI: 10.1093/mind/LIX.236.433.

[2] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," in *Advances in Neural Information Processing Systems*, T. Leen, T. Dietterich, and V. Tresp, Eds., Cambridge, MA, USA: MIT Press, 2000. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2000/file/728f206c2a01bf572b5940d7d9a8fa4c-Paper.pdf

[3] C. K. Lo, "What is the impact of ChatGPT on education? A rapid review of the literature," *Educ. Sci.*, vol. 13, no. 4, pp. 410–424, 2023, DOI: 10.3390/educsci13040410.

[4] R. B. Kargbo, "Advancements in predictive medicine: NLRP3 inflammasome inhibitors and AI-driven predictive health analytics," *ACS Med. Chem. Lett.*, vol. 15, no. 3, pp. 331–333, Mar. 2024, DOI: 10.1021/acsmedchemlett.4c00056.

[5] P. Hamet and J. Tremblay, "Artificial intelligence in medicine," *Metabolism*, vol. 69, pp. S36–S40, Apr. 2017, DOI: 10.1016/j.metabol.2017.01.011.

[6] B. S. Chia *et al.*, "Engineering a new generation of gene editors: Integrating synthetic biology and AI innovations," *ACS Synth. Biol.*, vol. 14, no. 3, pp. 636–647, Mar. 2025, DOI: 10.1021/acssynbio.4c00686.

[7] P. Eber, Y. M. Sillmann, and F. P. S. Guastaldi, "Beyond 3D printing: How AI is shaping the future of craniomaxillofacial bone tissue engineering," *ACS Biomater. Sci. Eng.*, vol. 11, no. 6, pp. 3095–3098, June 2025, DOI: 10.1021/acsbiomaterials.5c00420.

[8] L. Morán-González *et al.*, "AI approaches to homogeneous catalysis with transition metal complexes," *ACS Catal.*, vol. 15, no. 11, pp. 9089–9105, June 2025, DOI: 10.1021/acscatal.5c01202.

[9] F. Sorourifar *et al.*, "Physics-enhanced neural ordinary differential equations: Application to industrial chemical reaction systems," *Ind. Eng. Chem. Res.*, vol. 62, no. 38, pp. 15563–15577, Sept. 2023, DOI: 10.1021/acs.iecr.3c01471.

[10] Z. Zheng, "AI and chemistry in action: Transforming crystallization for scalable water harvesting solutions," *ACS Cent. Sci.*, vol. 10, no. 12, pp. 2173–2174, Dec. 2024, DOI: 10.1021/acscentsci.4c01838.

[11] S. Jiang *et al.*, "Integrating machine learning and color chemistry: Developing a high-school curriculum toward real-world problem-solving," *J. Chem. Educ.*, vol. 101, no. 2, pp. 675–681, Feb. 2024, DOI: 10.1021/acs.jchemed.3c00589.

[12] D. R. Bunch, T. J. S. Durant, and J. W. Rudolf, "Artificial intelligence applications in clinical chemistry," *Clin. Lab. Med.*, vol. 43, no. 1, pp. 47–69, Mar. 2023, DOI: 10.1016/j.cll.2022.09.005.

[13] K. M. Hallal, R. Hamdan, and S. Tlais, "Exploring the potential of AI-chatbots in organic chemistry: An assessment of ChatGPT and bard," *Comput. Educ. Artif. Intell.*, vol. 5, p. 100170, 2023.

[14] A. Ardyansyah *et al.*, "Students' perspectives on the application of a generative pre-trained transformer (GPT) in chemistry learning: A case study in Indonesia," *J. Chem. Educ.*, vol. 101, no. 9, pp. 3666–3675, Sept. 2024, DOI: 10.1021/acs.jchemed.4c00220.

[15] P. Jackson *et al.*, "Artificial intelligence in medical education - perception among medical students," *BMC Med. Educ.*, vol. 24, no. 1, p. 804, July 2024, DOI: 10.1186/s12909-024-05760-0.

[16] J. K. Johnson *et al.*, "Computer-assisted test construction in chemistry," *J. Chem. Educ.*, vol. 58, no. 2, p. 177, Feb. 1981, DOI: 10.1021/ed058p177.

[17] S. Bedi *et al.*, "QUEST-AI: A system for question generation, verification, and refinement using AI for USMLE-style exams," in *Biocomputing 2025*, 0 vols., WORLD SCIENTIFIC, 2024, pp. 54–69. DOI: 10.1142/9789819807024_0005.

[18] D. V. Smoline and S. Butakov, "Applying artificial intelligence to the educational data: an example of syllabus quality analysis," *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 2012, [Online]. Available: https://api.semanticscholar.org/CorpusID:15781350

[19] H. Kim and T. K. B. Koo, "The impact of generative AI on syllabus design and learning," *J. Mark. Educ., 48*, pp. 20–41, 2024, [Online]. Available: https://api.semanticscholar.org/CorpusID:274750769

[20] H. Balalle and S. Pannilage, "Reassessing academic integrity in the age of AI: A systematic literature review on AI and academic integrity," *Social Sci. Humanit. Open*, vol. 11, p. 101299, Jan. 2025, DOI: 10.1016/j.ssaho.2025.101299.

[21] T. Lancaster and C. Cotarlan, "Contract cheating by STEM students through a file sharing website: a Covid-19 pandemic perspective," *Int. J. for Educ. Integr.*, vol. 17, no. 1, p. 3, Feb. 2021, DOI: 10.1007/s40979-021-00070-0.

[22] X. Lee, "Assessment reform in higher education: An ethical approach to harness the power of generative artificial intelligence," *Educ. Res. Perspect.*, vol. 51, p. 159, 2024.

[23] P. Scarfe *et al.*, "A real-world test of artificial intelligence infiltration of a university examinations system: A 'Turing Test' case study,"

PLos One, vol. 19, no. 6, p. e0305354, June 2024, DOI: 10.1371/journal.pone.0305354.

[24] A. Bouchra, K. Hanane, and L. Mohamed, "Evaluating ChatGPT and DeepSeek for science education: A comparative analysis of AI-powered learning assistants," in The Second International Symposium on Generative AI and Education (ISGAIE'2025), R. González Vallejo, G. Moukhliss, E. Schaeffer, and V. Paliktzoglou, Eds., Cham: Springer Nature Switzerland, 2025, pp. 433–442.

[25] Á. Cabezas-Clavijo and P. Sidorenko-Bautista, "Assessing the performance of 8 AI chatbots in bibliographic reference retrieval: Grok and DeepSeek outperform ChatGPT, but none are fully accurate," 2025, [Online]. Available: https://arxiv.org/abs/2505.18059

[26] Reuters, "Chinese universities launch DeepSeek courses to capitalise on AI boom," 2025. [Online]. Available: https://www.reuters.com/technology/artificial-intelligence/chinese-universities-launch-deepseek-courses-capitalise-ai-boom-2025-02-21/?utm_source=chatgpt.com

[27] B. S. Bloom *et al.*, *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain*. New York: David McKay, 1956.

[28] L. Anderson *et al.*, *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*, 1st ed. Upper Saddle River, NJ, USA: Pearson, 2001.

[29] Z. Ullah *et al.*, "Bloom's taxonomy: A beneficial tool for learning and assessing students' competency levels in computer programming using empirical analysis," *Comput. Appl. Eng. Educ.*, vol. 28, no. 6, pp. 1628–1640, Nov. 2020, DOI: 10.1002/cae.22339.

[30] F. A. Khalifa and A. H. Mohamad, "Automatic analysis of exam papers based on bloom's taxonomy: An artificial intelligence approach," *Int. J. Appl. Sci. Technol.*, vol. 6, no. 3, pp. 595–609, 2024, DOI: 10.47832/2717-8234.20.46.

[31] A. Rawat, S. Kumar, and S. Singh Samant, "A Systematic Review of Question Classification Techniques Based on Bloom's Taxonomy," in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, July 2023, pp. 1–7. DOI: 10.1109/ICCCNT56998.2023.10308403.

[32] A. Yaacoub, J. Da-Rugna, and Z. Assaghir, "Assessing AI-generated questions' alignment with cognitive frameworks in educational assessment," *Int. J. Comput. Theory Eng.*, vol. 17, no. 3, pp. 114–125, 2025, DOI: 10.7763/ijcte.2025.v17.1374.

[33] K. Hwang *et al.*, "Towards AI-Assisted Multiple Choice Question Generation and Quality Evaluation at Scale: Aligning with Bloom's Taxonomy," presented at the NeurIPS'23 Workshop on Generative AI for Education (GAIED), [Online]. Available: https://api.semanticscholar.org/CorpusID:266876687

[34] S. Fergus, M. Botha, and M. Ostovar, "Evaluating academic answers generated using ChatGPT," *J. Chem. Educ.*, vol. 100, no. 4, pp. 1672–1675, Apr. 2023, DOI: 10.1021/acs.jchemed.3c00087.

[35] E. F. Ruff *et al.*, "ChatGPT writing assistance and evaluation assignments across the chemistry curriculum," *J. Chem. Educ.*, vol. 101, no. 6, pp. 2483–2492, June 2024, DOI: 10.1021/acs.jchemed.4c00248.

[36] T. M. Clark *et al.*, "Using ChatGPT to support lesson planning for the historical experiments of Thomson, Millikan, and Rutherford," *J. Chem. Educ.*, vol. 101, no. 5, pp. 1992–1999, May 2024, DOI: 10.1021/acs.jchemed.4c00200.

[37] T. M. Clark *et al.*, "Comparing the performance of college chemistry students with ChatGPT for calculations involving acids and bases," *J. Chem. Educ.*, vol. 100, no. 10, pp. 3934–3944, Oct. 2023, DOI: 10.1021/acs.jchemed.3c00500.

[38] A. A. Fernández *et al.*, "ChatGPT as an instructor's assistant for generating and scoring exams," *J. Chem. Educ.*, vol. 101, no. 9, pp. 3780–3788, Sept. 2024, DOI: 10.1021/acs.jchemed.4c00231.

[39] T. M. Clark, "Investigating the use of an artificial intelligence chatbot with general chemistry exam questions," *J. Chem. Educ.*, vol. 100, no. 5, pp. 1905–1916, May 2023, DOI: 10.1021/acs.jchemed.3c00027.

[40] B. Sorenson and K. Hanson, "Identifying generative artificial intelligence chatbot use on multiple-choice, general chemistry exams using rasch analysis," *J. Chem. Educ.*, vol. 101, no. 8, pp. 3216–3223, Aug. 2024, DOI: 10.1021/acs.jchemed.4c00165.

[41] B. Sorenson and K. Hanson, "Statistical comparison between in-person and online general chemistry exam outcomes: A COVID-induced case study," *J. Chem. Educ.*, vol. 100, no. 9, pp. 3454–3461, Sept. 2023, DOI: 10.1021/acs.jchemed.3c00476.

[42] H. Samuel and E. Etim, "Review on ChatGPT: History, applications and limitations in chemistry," *Phy. Sci. Biophys. J. (PSBJ)*, vol. 7, no. 2, 2023, pp. 263–267, DOI: 10.23880/psbj-16000263.

[43] H.-K. Wu and P. Shah, "Exploring visuospatial thinking in chemistry learning," *Sci. Educ.*, vol. 88, no. 3, pp. 465–492, May 2004, DOI: 10.1002/sce.10126.

[44] D. R. Klein, "Organic chemistry I as a second language," 2003. [Online]. Available: https://api.semanticscholar.org/CorpusID:92645238

[45] F. Delvigne, "A visual aid for teaching the resonance concept," *J. Chem. Educ.*, vol. 66, no. 6, p. 461, June 1989, DOI: 10.1021/ed066p461.

[46] D. Xue and M. Stains, "Exploring students' understanding of resonance and its relationship to instruction," *J. Chem. Educ.*, vol. 97, no. 4, pp. 894–902, Apr. 2020, DOI: 10.1021/acs.jchemed.0c00066.

[47] L. Goorts *et al.*, "How do LLMs perform in the context of MCQs across different levels of thinking skills in a business education course at higher education? A comparison of ChatGPT, Gemini, and Copilot," *Computers Educ. Artif. Intell.*, *9*, p. 100475, Sept. 2025, DOI: 10.1016/j.caeai.2025.100475.

[48] J. Wang and W. Fan, "The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis," *Humanit. Soc. Sci. Commun.*, vol. 12, no. 1, p. 621, May 2025, DOI: 10.1057/s41599-025-04787-y.