Pharmacy Education (2026) 26(1) 514 - 524
https://doi.org/10.46542/pe.2026.261.514524

RESEARCH ARTICLE

Customising ChatGPT for chemistry education: A domain-grounded chatbot to support pharmacy students in medicinal chemistry

Federico A. Fookes[1], Elangeni A. Gilbert[1,2], Federico L. Baroni[1], María Eugenia Taverna[2,3]

[1] Universidad Católica de Santa Fe (UCSF), Santa Fe, Argentina
[2] Instituto de Desarrollo Tecnológico para la Industria Química (INTEC), Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) - Universidad Nacional del Litoral (UNL), Santa Fe, Argentina
[3] Universidad Tecnológica Nacional (UTN), San Francisco, Córdoba, Argentina

Keywords
Artificial intelligence
Education
Internet-based intervention
Medicinal chemistry
Pharmacy
Problem solving

Correspondence
María Eugenia Taverna
Instituto de Desarrollo Tecnológico para la Industria Química (INTEC)
CONICET - Universidad Nacional del Litoral (UNL)
Santa Fe
Argentina
metaverna@santafe-conicet.gov.ar

Abstract

Background: This study develops and evaluates ChemIA, a course-specific GPT that supports pharmacy students in undergraduate Medicinal Chemistry. Customised with instructor-authored materials, it better matches the syllabus and reduces generic-model inaccuracies. **Methods:** The exploratory evaluation assessed technical performance and student perceptions. Technical performance was tested with 15 assessment questions from course assessments. Across repeated interactions, ChemIA largely matched instructor-validated answers; discrepancies mainly involved information found only in images. It also handled a small set of questions based on English-translated materials. These findings indicate basic multilingual functionality in the tested conditions, not comprehensive multilingual capability. ChemIA supported course activities including quizzes, pre-class study, podcast-style content, and a pharmacological case study. **Results:** Survey responses indicated that students perceived ChemIA as useful for clarifying medicinal chemistry concepts, preparing for examinations, and obtaining supplementary explanations. Students also reported greater confidence in ChemIA's responses than in those generated by a generic GPT model. In the case-study activity, students rated ChemIA's explanations more often as very clear than those from the non-custom model. **Conclusion:** These findings offer exploratory evidence that a domain-grounded GPT can complement Medicinal Chemistry education. Its simple setup and adaptability may support similar applications in other disciplines. However, findings are limited by the exploratory design and the absence of direct measures of learning or academic performance.

Introduction

Artificial intelligence (AI), particularly large language models (LLMs) such as ChatGPT, has the potential to reshape education by providing personalised support, answering specific questions, and generating detailed explanations across diverse knowledge areas (Caccavale *et al.,* 2024). These systems can facilitate learning, reduce instructors' routine workload, and allow educators to focus on higher-order pedagogical goals. Prior research has shown that LLMs can support scientific writing, provide automated feedback, and foster critical thinking by encouraging students to evaluate AI-generated content (Ruff *et al.,* 2024). Moreover, AI-powered tools can emulate adaptive pedagogy, promoting metacognition, problem-solving, and scientific communication skills (Ferreira-Mello *et al.,* 2019; Lu *et al.,* 2021; Ramesh & Sanampudi, 2022; Exintaris *et al.,* 2023; Guo & Lee, 2023; Pence *et al.,* 2024).

The educational role of conversational AI can be understood through the lenses of scaffolding and self-regulated learning. Previous studies have suggested

that educational chatbots can support self-regulated learning by facilitating access to learning resources, encouraging active engagement with content, and providing immediate feedback (Guan *et al.,* 2025). Similarly, conversational AI systems may serve as scaffolding tools by offering guidance and explanations that assist learners in developing conceptual understanding (Hilliger *et al.,* 2026). In this context, ChemIA represents a course-specific educational chatbot designed to support medicinal chemistry learning. The higher confidence reported by students in the present study may be consistent with predictions derived from these theoretical frameworks, although further research is needed to directly evaluate the underlying learning mechanisms. These perspectives are not intended to provide a comprehensive educational framework for the present study. Rather, they serve to situate the potential pedagogical role of domain-grounded conversational AI within established learning theories. Several studies have explored the role of LLMs in science and health-profession education. For example, recent work has shown that machine learning and natural language processing (NLP) tools can assist instructors by easing the implementation of adaptive support, particularly through automation of tasks such as grading, analysing students' argumentation in open-ended questions, and generating tailored feedback. These applications highlight the potential of LLMs to individualise instruction by offering timely, context-sensitive explanations that respond to learners' specific difficulties. In the context of medicinal chemistry education—where students must integrate complex chemical, biological, and pharmacological concepts—such adaptive support may be especially valuable, providing opportunities to clarify mechanisms of action, scaffold conceptual understanding, and reduce cognitive load during self-directed study (Dood *et al.,* 2020; Ariely *et al.,* 2024; Martin *et al.,* 2024). To effectively employ large language models such as ChatGPT in adaptive learning environments, it is crucial to recognise both their capabilities and their limitations. In chemistry, West et al. (2023) evaluated ChatGPT (GPT-3.5) for generating lab reports, noting strengths but also weaknesses in accuracy and style. Ruff et al. (2024) reported that revising AI-generated chemistry texts improved structure but often lacked references and background information. Rojas (2024) showed benefits for idea generation and proofreading in inorganic chemistry, while emphasising the importance of verifying AI-generated outputs against primary sources (Rojas, 2024). Kim et al. (2024) examined the impact of ChatGPT on engineering laboratory reports, highlighting both improvements in clarity and the need for careful verification of AI-generated content. Kunitsu (2023) found that GPT-4 performed well on general chemistry but less effectively on specialised queries. Other initiatives include exploratory chatbots for general chemistry (Lolinco & Holme, 2023) and reviews of publicly available custom GPTs in chemistry (Sigot & Tassoti, 2025), which emphasised both potential and the importance of careful instruction design and copyright considerations (Kenthapadi *et al.,* 2024).

Despite their potential, large language models also present important limitations. Prior studies have documented the generation of inaccurate or hallucinated information, overconfident responses, and difficulties in distinguishing reliable from unreliable content. These challenges are particularly relevant in health and science education, where factual accuracy and disciplinary precision are essential. Consequently, concerns regarding epistemic reliability have motivated growing interest in grounded approaches that constrain responses to validated sources.

While these studies demonstrate the promise of LLMs in science education, they also reveal limitations: general-purpose models often provide imprecise or superficial responses in specialised domains such as medicinal chemistry. Customisation through grounding—aligning outputs with explicit, verifiable sources—emerges as a necessary step to ensure both accuracy and content alignment (Kenthapadi *et al.,* 2024). Recent research has begun to explore grounded implementations in higher education. For instance, Tascioglu (2025) examined GPT-based assistants as reusable templates, while Sehri and Dumond (2025) highlighted their potential to reduce teaching workload. Collectively, these studies indicate that LLMs can support a variety of educational tasks, including scientific writing, feedback generation, problem solving, and content explanation. Nevertheless, they also highlight important limitations. Most investigations have focused on general-purpose models, which may produce inaccurate, superficial, or inconsistent responses when applied to specialised disciplinary content. In addition, many studies have been conducted in controlled settings or have evaluated isolated activities, leaving open questions about implementing domain-grounded systems in authentic classroom contexts and their perceived value for students in discipline-specific learning environments.

Taken together, previous studies indicate that LLMs can support content explanation, feedback generation, scientific communication, and problem-solving activities. However, the literature also converges on several recurring limitations, including hallucinations, lack of disciplinary specificity, variability in response quality, and the scarcity of evaluations conducted in

authentic classroom settings. These limitations highlight the need for studies examining domain-grounded systems specifically designed around instructor-developed materials and implemented within real educational contexts.

Despite growing interest in the use of large language models in science education, most reported studies have focused on general-purpose systems, controlled demonstrations, or theoretical discussions of their educational potential. Consequently, there is still limited evidence regarding the implementation of domain-grounded GPTs built from instructor-authored materials and deployed in authentic classroom settings.

The literature highlights technical reliability and hallucinations as the main challenges of LLM adoption in education, alongside concerns regarding student over-reliance, assessment fairness, and data privacy. Together, these issues emphasise the need for careful consideration of technical, pedagogical, and ethical factors when implementing AI-based educational tools (Shi *et al.,* 2026).

Furthermore, little is known about how students perceive the usefulness, clarity, and trustworthiness of responses generated by such systems in specialised disciplines such as medicinal chemistry.

To address this gap, the present exploratory study describes the development and classroom implementation of ChemIA, a domain-grounded GPT designed for a Medicinal Chemistry course, and examines its technical performance as well as students' perceptions of its use. Grounded in instructor-authored lectures, ChemIA was designed to improve alignment with course content and support accurate responses within the intended educational scope. Implementation included (i) custom instruction design, (ii) accuracy validation via multiple-choice tests, and (iii) integration into authentic tasks—a podcast and a case study under three conditions (traditional resources (lecture notes only), ChemIA, and a generic non-custom GPT model). Evaluation comprised comparative and satisfaction surveys assessing LLM use and students' experiences with ChemIA.

Development and technical evaluation of the chatbot

Bot development

ChemIA customisation process

The development of ChemIA—an acronym for Chemistry and Artificial Intelligence—began with the selection of GPT-5, chosen for its advanced natural language processing capabilities. The complete GPT instructions, uploaded course materials, representative prompts, examples of student–ChemIA interactions, and the questions used for the preliminary accuracy assessment are provided in the Supplementary. ChemIA was developed using the ChatGPT Custom GPT interface rather than through an API-based implementation. Consequently, platform-level processes such as document chunking, embedding generation, retrieval algorithms, context management, multilingual processing, and internal model configuration parameters are not accessible to users and cannot be directly modified or reported. The customisation process required grounding through instructor-authored materials, predefined system restrictions, and access to a paid OpenAI plan. This implementation approach was intentionally selected because it enables the development of domain-grounded educational tools that can be readily deployed by instructors without specialised AI or programming expertise, thereby supporting evaluation under authentic classroom conditions rather than optimisation of model parameters.

The chatbot instructions were refined through several iterative rounds of testing conducted by the Medicinal Chemistry course instructors. Responses were evaluated with respect to factual accuracy, alignment with course content, and adherence to the intended scope of the course. Based on these evaluations, instructions and constraints were progressively adjusted to improve performance within the educational context. This refinement process formed part of the practical development of the tool rather than a formal validation study. Consequently, the exact number of iterations was not systematically recorded, and no inter-rater agreement procedures were implemented.

ChemIA was structured into two modules: (1) system instructions, acting as high-priority rules that guide responses and constrain sources (Figure 1), and (2) subject-specific documents aligned with Medicinal Chemistry resources (see Supplementary).

Both the instructors and the students were native Spanish speakers; therefore, most course materials were delivered in Spanish. However, one lecture (on antilipidemic drugs) was delivered in English to evaluate the model’s multilingual capabilities.

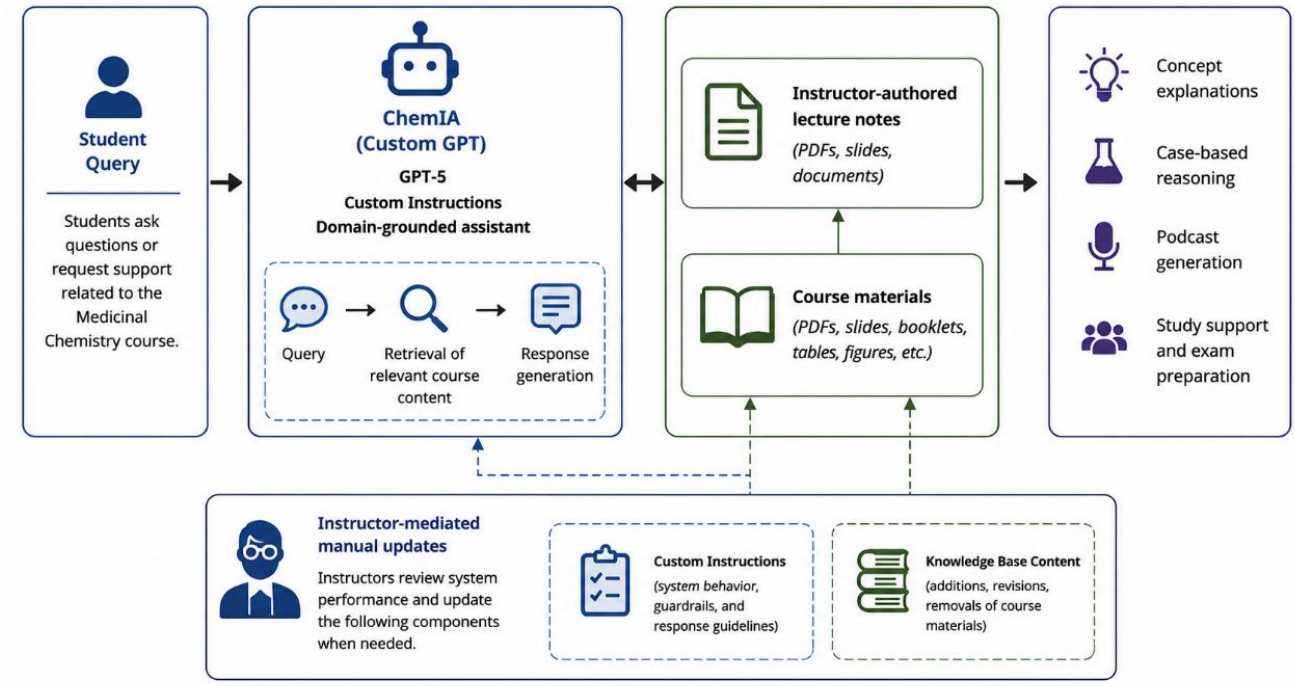


Figure 1: Workflow and architecture of ChemIA, a customised GPT developed for medicinal chemistry education. The diagram illustrates how student queries are processed by a GPT-5–based system through system instructions, a course-specific knowledge base, context retrieval, and response generation to produce course-aligned, context-specific responses for educational applications.

ChemIA instruction setup

A tailored set of system instructions was developed to align ChatGPT with the Medicinal Chemistry syllabus (see Supplementary). These instructions specified the knowledge base (namely, the lecture materials uploaded by the course instructors), directed the model to prioritise and ground its responses in this content, and incorporated explicit constraints intended to minimise the inclusion of information beyond the scope of the course materials. The directives were iteratively refined through repeated testing, during which phrasing was adjusted to reduce ambiguity and promote consistent, syllabus-aligned responses. The purpose of this process was not to optimise the underlying language model or develop a formally validated prompt-engineering framework. Rather, it was intended to adapt the system to the pedagogical requirements of the Medicinal Chemistry course by improving alignment with course content, reducing responses outside the intended scope, and enhancing the usefulness of the tool as an educational resource for students and instructors. In addition, students' responses may have been influenced by social desirability bias, novelty effects associated with exposure to a new technology, and instructor expectancy effects inherent to classroom-based educational research.

Performance was assessed through consistency and accuracy tests based on past exams (see Supplementary), with directives iteratively refined to enhance clarity. This promoted response consistency across repeated interactions while keeping the system adaptable to future updates.

Empirical study design

This study should be understood as an exploratory classroom implementation of a custom GPT designed to provide both instructors and students with tools to enhance their performance. The evaluation focused on students' perceptions and the technical accuracy of a domain-grounded chatbot. It does not aim to measure learning outcomes or establish causal effects.

An initial orientation session allowed students to explore ChemIA, recognising both its potential and its limitations. In later tasks, the tool was applied to analyse drug mechanisms of action, with outputs compared against official course materials in collaborative sessions to detect inconsistencies. Seminar discussions then used ChemIA-generated answers as prompts for critical reflection on accuracy and relevance. These preparatory activities familiarised pharmacy students with AI-assisted inquiry before participating in the empirical components of the study.

Two innovative applications were incorporated. First, a podcast project in which ChemIA generated draft scripts summarising medicinal chemistry topics, fostering scientific communication and offering an alternative learning modality (Frydenberg, 2008; Gribbins, 2007). The following prompt was used:

"Please create a 2-minute podcast introduction on nonsteroidal anti-inflammatory drugs (NSAIDs). The podcast should cover the basic concepts, mechanisms of action, and common uses of NSAIDs. It should be clear, engaging, and suitable for an audience with a basic understanding of medicinal chemistry."

Representative examples of ChemIA educational applications used throughout the course are presented in Table I.

Table I: Representative examples of ChemIA educational applications used in the Medicinal Chemistry course

Educational application	Representative prompt	Educational purpose
Podcast generation	Create a 2-minute podcast introduction on NSAIDs.	Content summarisation and scientific communication
Concept explanation	Explain the mechanism of action of proton pump inhibitors.	Clarification of medicinal chemistry concepts
Case-based reasoning	Compare the mechanisms of action of H_2 receptor antagonists and proton pump inhibitors in peptic ulcer disease.	Pharmacological reasoning and application of knowledge

Complete prompts and representative outputs are provided in the Supplementary.

Second, a case study on antiulcer drugs was implemented, in which three groups addressed the same pharmacological problem under different conditions: (i) traditional resources (lecture notes only), (ii) ChemIA as the primary tool, and (iii) a generic non-custom GPT model. This structure enabled comparison of domain-grounded AI, general-purpose AI, and non-AI approaches within a pharmacy learning context.

ChemIA evaluation

The empirical work was guided by three research questions: how pharmacy students perceived the usefulness of ChemIA for supporting medicinal chemistry learning (RQ1); how their confidence in ChemIA's responses compared to confidence in those generated by a generic AI model (RQ2); and how students rated the clarity of explanations provided in a case-based activity across ChemIA, a generic AI model, and no-AI conditions (RQ3).

The study was conducted in a Medicinal Chemistry course within an undergraduate Pharmacy program at a university in Argentina. The ChemIA-user group (n = 36) consisted of fourth-year pharmacy students (mean age = 22.1 ± 1.3 years; 72% women, 28% men), reflecting the typical gender distribution of the program. A second group of students from a different cohort (n = 24, aged 21–35; 75% women, 25% men) participated in the structured case study conducted under the three conditions (ChemIA, generic GPT model, or traditional resources). The two groups corresponded to different student cohorts enrolled in the course during different academic semesters and therefore did not overlap. Because participants were drawn from naturally occurring classroom cohorts, no random allocation procedure was performed.

Evaluation of ChemIA proceeded through accuracy testing and student experience measures. Accuracy was assessed using fifteen assessment questions developed from instructor-authored lecture materials. ChemIA was prompted with each question five times, followed by five additional iterations in which it was instructed to provide an alternative correct explanation. Accuracy was defined as the proportion of responses that matched the instructor-validated answer, whereas response consistency referred to the consistency of responses across repeated prompts. The consistency observed across the ten trials indicated low variability and supported the ChemIA's responses.

Student perceptions were examined using two complementary surveys. The survey instruments consisted of single-item, ad hoc measures tailored to the exploratory nature of the study. All items were reviewed by two content-expert instructors to establish content validity and ensure correspondence with course goals. The constructs measured—perceived usefulness, confidence in accuracy, clarity of explanations, and perceived support—were intentionally limited to narrow judgments directly tied to the course activities rather than broader psychological constructs (e.g., motivation, engagement, self-efficacy) that would require validated multi-item scales. Future studies will incorporate validated instruments when examining more complex constructs.

The ChemIA-user survey (n = 36) included Likert-scale items assessing perceived usefulness, confidence in the accuracy of AI-generated responses, and self-reported study strategies. Using distinct instruments ensured validity while maintaining comparability through shared baseline constructs. During the case-study activity, students rated the clarity and helpfulness of the explanations produced under the three conditions and provided short qualitative comments that contextualised their judgments.

All quantitative analyses were conducted using R (R Core Team, 2025). Descriptive statistics were used to summarise survey responses. Confidence ratings obtained from students who evaluated both the generic AI tool and ChemIA were compared using the Wilcoxon signed-rank test following assessment of normality with the Shapiro–Wilk test. Exact p-values, 95% confidence intervals, and rank-biserial correlation effect sizes are reported where appropriate. Because the analyses were exploratory and addressed a limited number of predefined comparisons, no formal adjustment for multiple testing was applied. Qualitative comments collected during the case-study activity were reviewed and coded inductively by the instructors to identify recurring themes related to clarity, alignment with course content, and perceived value for understanding mechanisms of action. Responses were grouped according to common patterns and used to contextualise the quantitative findings. Because the qualitative component was exploratory in nature, no formal thematic analysis framework, inter-coder agreement procedure, or qualitative validation protocol was applied.

Results

Findings are presented along two dimensions: ChemIA's technical performance and students' perceptions of its usefulness. ChemIA was used as a teaching support tool to generate self-assessment quizzes and short pre-class readings, although all materials required instructor validation before use. To address RQ1 (perceived usefulness), survey data showed that 86% of ChemIA users reported prior experience with AI tools, although this was generally limited to non-academic applications. As summarised in Table II, most students rated ChemIA as very useful for consulting questions at any time and valued having a tool explicitly aligned with the Medicinal Chemistry course. Students reported using ChemIA primarily for clarifying concepts (24%), preparing for examinations (24%), seeking supplementary explanations (19%), and completing assignments (19%); 14% used it mainly for exploration. Open-ended comments indicated that the main perceived advantage was immediate access to explanations consistent with lecture content, although a minority of students continued to prefer in-person consultations

Table II summarises the main quantitative outcomes obtained from the survey and case-study activities.

Table II: Summary of survey and case-study outcomes

Outcome	Sample size (n)	Result
Previous use of AI tools for learning	36	31/36 (86%)
Perceived usefulness of podcast-generation feature	36	33/36 (92%) rated it as very useful
Perceived usefulness of multilingual responses	36	36/36 (100%) rated it as very useful
Perceived usefulness of ChemIA for answering course questions	36	34/36 (94%) rated it as very useful
Students skeptical of ChatGPT before using ChemIA	36	22/36 (61%)
Support for similar subject-specific GPTs in other courses	36	35/36 (97%)
Mean confidence in generic GPT responses (1–5 scale)	23	3.39 ± 0.65
Mean confidence in ChemIA responses (1–5 scale)	21	4.05 ± 0.59
Students rating ChemIA explanations as "very clear"	24	21/24 (88%)
Students rating generic GPT explanations as "very clear"	24	15/24 (62%)
Preferred tool if repeating the activity: ChemIA	24	14/24 (58%)
Preferred tool if repeating the activity: traditional materials	24	9/24 (38%)
Preferred tool if repeating the activity: generic GPT	24	1/24 (4%)

Note: The survey sample (n = 36) and the case-study sample (n = 24) correspond to different cohorts and educational activities.

A second indicator of usefulness was obtained from the multimodal podcast activity. As summarised in Table II, most students rated the option of generating podcast-style scripts as very useful. In the NSAID pilot task,

ChemIA produced concise introductory scripts that students later edited and used for individual review or group presentations. Students frequently mentioned the use of ChemIA for summarising content and generating accessible explanations. Addressing RQ2 (confidence in accuracy), confidence was operationalised as students' self-reported confidence in the accuracy of AI-generated responses. These ratings were compared using the Wilcoxon signed-rank test. Normality assessment using the Shapiro–Wilk test indicated that the confidence ratings departed significantly from normality (W = 0.757, $p < 0.001$); therefore, the Wilcoxon signed-rank test was used. Because confidence ratings were obtained from the same groups of students (the analysis included only the 21 of 24 participants who provided complete paired responses for both the generic AI and ChemIA conditions), comparisons were performed using the Wilcoxon signed-rank test. Students reported significantly higher confidence levels when using ChemIA compared with the generic AI tool (Wilcoxon signed-rank test, V = 4.5, p = 0.0093). The estimated median difference was approximately one point on the confidence scale (pseudo-median = −1.00, 95% CI [−2.00, −1.00]), indicating consistently higher confidence ratings for ChemIA. The effect size was large (rank-biserial correlation = 0.86, 95% CI [0.67, 0.95]). The test compared median confidence scores for parallel pharmacological questions answered under both AI conditions (see Fig. 2a and the Supplementary).

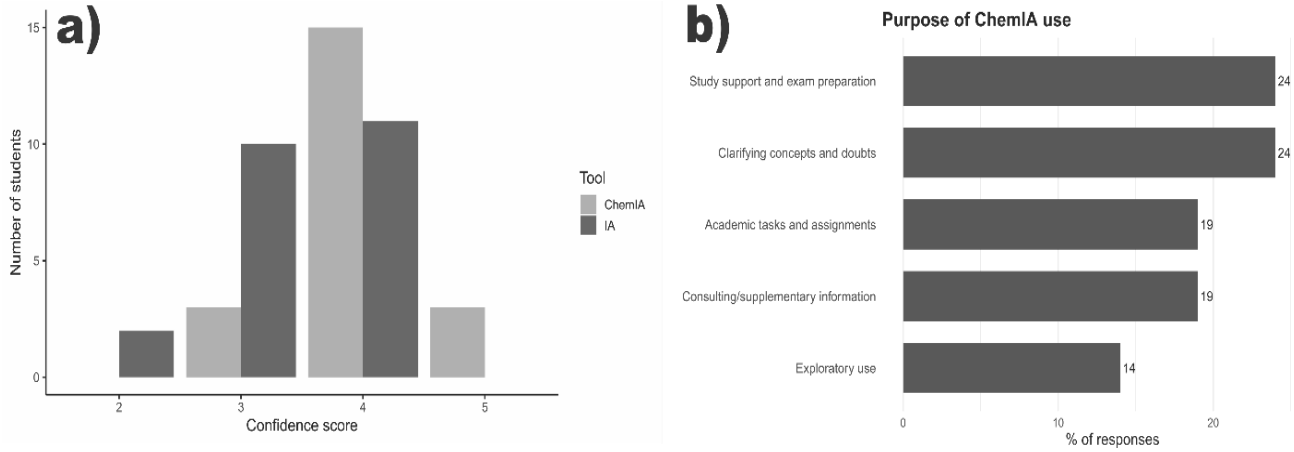


Figure 2: (a) Students' confidence in responses generated by ChemIA and the generic GPT model. (b) Reported educational uses of ChemIA among participating students

Findings for RQ3 (clarity of explanations across study conditions) come from the antiulcer drug case study, where students worked under three conditions: ChemIA, a generic GPT model, or traditional resources. Eighty-eight percent of students rated ChemIA's explanations as "very clear," compared with 62% for the generic model. Students reported that ChemIA's responses helped differentiate mechanistic differences between H_2 antagonists and proton pump inhibitors, noting that alignment with course terminology reduced confusion in interpreting drug classes.

Finally, ChemIA's domain-specific performance was evaluated through an exploratory accuracy assessment using fifteen instructor-validated assessment questions. Across ten interaction cycles, during which the chatbot answered each question five times in its original form and five times in paraphrased form (150 responses), ChemIA achieved an overall accuracy of 93.3% (140/150 correct responses). Errors were restricted to a single question regarding the oral administration of estrogens. The relevant information in the uploaded course material was available only in image format rather than as machine-readable text. Additionally, ChemIA correctly answered questions derived from English-translated lecture materials.

Discussion

This study examined the implementation of a domain-grounded chatbot, ChemIA, in an undergraduate Medicinal Chemistry course. The findings suggest that a customised conversational agent aligned with instructor-authored materials may serve as a technically sound and potentially useful educational resource when integrated into authentic learning activities. These findings are consistent with previous studies reporting that chatbot-based learning

environments can effectively support teaching and learning activities (Al-Hafdi & AlNajdi, 2024).

Within the limited accuracy assessment performed, ChemIA showed a high level of agreement with instructor-validated answers (93%) across repeated interactions. This level of consistency contrasts with reports that general-purpose GPT models often produce superficial or inconsistent answers in specialised scientific domains, underscoring the value of discipline-specific grounding (Schulze Balhorn *et al.,* 2024). These results highlight the importance of careful content preparation and verification when implementing AI tools in pharmacy education. ChemIA also responded correctly to the limited set of questions derived from English-translated course materials. Although this finding suggests that the system functioned appropriately under these specific conditions, it should not be interpreted as evidence of general multilingual capability (Lee *et al.,* 2013).

Regarding students' perceptions, ChemIA was rated as highly useful for supporting the study of medicinal chemistry. Students used the chatbot to clarify concepts, prepare for examinations, and generate alternative explanations when course materials were insufficient. The positive reception of the podcast-generation feature further illustrates that students value flexible formats for revisiting core concepts. In this study, usefulness should be interpreted specifically as students' perceived value of ChemIA for supporting understanding of medicinal chemistry content, rather than as a broader motivational construct.

A key finding relates to confidence in the accuracy of AI-generated responses. Students reported significantly higher confidence in ChemIA than in a generic GPT model. The large effect size indicates a substantial difference in students' reported confidence between the two systems and suggests that ChemIA was perceived as a more trustworthy source of course-related information. This perception may reflect the fact that ChemIA's responses were grounded in validated course content. Confidence in this context reflects students' perceptions of the credibility and usefulness of the responses rather than objective measures of correctness. Consequently, higher confidence should not be interpreted as evidence of greater accuracy, as it may also reflect overreliance on AI-generated information or misplaced trust. Importantly, higher confidence should not be interpreted as evidence of greater factual accuracy. Students may express confidence in AI-generated responses even when they are unable to independently verify their correctness, highlighting the importance of fostering critical evaluation skills when integrating AI into educational settings. The case-study activity further demonstrated the benefits of domain grounding. Students rated ChemIA's explanations of antiulcer drugs as clearer than those generated by the generic model. Prior work has shown that clarity is a critical factor in how learners evaluate AI-generated scientific explanations (Lombrozo, 2006). In pharmacological topics where mechanistic distinctions must be precise, alignment with course terminology may be particularly important for comprehension.

Despite the potential benefits of domain-grounded AI systems, their educational use also raises important challenges. Grounding may reduce the likelihood of hallucinated responses, but it cannot eliminate inaccuracies entirely. Students may also develop excessive reliance on AI-generated information, potentially reducing opportunities for critical evaluation and independent reasoning. In addition, high confidence in AI responses may contribute to confirmation bias if learners preferentially accept information that aligns with their expectations without adequate verification. Ethical considerations related to transparency, accountability, and responsible use of AI in educational settings should therefore remain central when implementing these technologies. Consequently, AI-based tools should be viewed as complementary learning supports rather than authoritative sources of knowledge, and their use should be accompanied by activities that foster critical thinking and source evaluation skills.

Overall, the results indicate that ChemIA provided factual information and was perceived as a useful complement to traditional study methods. These findings should be interpreted within the limits of the constructs assessed—usefulness, confidence, and technical accuracy—and not generalised to outcomes such as motivation, engagement, or metacognition, which were not measured. Future research using validated instruments and experimental designs could clarify how domain-specific AI tools affect broader learning outcomes in pharmacy education.

Limitations

This study has several limitations. First, the use of the ChatGPT interface rather than API-based access limits reproducibility due to the inability to control parameters such as temperature. Second, although the Medicinal Chemistry instructors gradually refined ChemIA instructions by incorporating additional restrictions until the responses were judged to be accurate and aligned with the course content, no formal inter-rater evaluation procedure was performed and no inter-rater agreement metrics were calculated.

Third, the exploratory design relied on single-item measures, which restricts psychometric robustness and generalisability. Fourth, the accuracy assessment consisted of 15 instructor-authored questions, which, while aligned with course content, represents a narrow sampling of the domain. In addition, accuracy was evaluated using a binary correct/incorrect classification, which does not capture other relevant dimensions of response quality such as completeness, clarity, depth of explanation, or pedagogical usefulness. Furthermore, the accuracy assessment was performed only on ChemIA and did not include an equivalent evaluation of a general-purpose GPT model using the same question set and evaluation protocol. Consequently, the study cannot determine the extent to which the observed performance is attributable to the grounding strategy itself. The reported accuracy should therefore be interpreted as a preliminary evaluation of ChemIA's performance within the specific educational context examined rather than as a comprehensive validation of the chatbot. Because the evaluation questions were derived from the same instructional materials used to construct the knowledge base, the assessment primarily reflects content alignment and response consistency rather than independent validation of the system. Accordingly, the findings should not be considered representative of ChemIA's performance across all medicinal chemistry topics or educational contexts. Future work will incorporate API-controlled testing, validated multi-item instruments, and expanded item banks to enhance methodological rigor. This study involved a small number of students drawn from naturally occurring classroom cohorts rather than through random assignment. Consequently, selection bias cannot be excluded, and differences in prior knowledge, motivation, or other unmeasured characteristics between groups may limit the generalisability of the findings. In addition, no formal baseline measures of academic performance or digital literacy were collected. Although previous experience with AI tools was recorded descriptively, it was not incorporated as an analytical variable. Consequently, potential differences in participants' baseline characteristics could not be controlled for and may have influenced the observed perceptions of ChemIA. As this study was exploratory in nature, sample size was determined by course enrollment and participation in the educational activities. The small sample size also restricts statistical power, echoing similar challenges reported by Pence et al. (2024). Therefore, the findings should be interpreted as preliminary and hypothesis-generating rather than confirmatory. Together, these results highlight the importance of providing accurate AI tools and fostering active learner participation, while future studies should employ larger and better-matched samples to more comprehensively assess the educational impact of AI-driven learning.

Accordingly, survey findings should be interpreted exclusively as exploratory perception data and should not be generalised as evidence of educational effectiveness or broader psychological constructs.

Conclusion

This exploratory study described the development, implementation, and preliminary evaluation of ChemIA, a domain-grounded GPT designed to support students in a Medicinal Chemistry course. The findings suggest that students perceived the tool as useful, accessible, and aligned with course content, and reported higher confidence when using ChemIA compared with a general-purpose GPT in the specific activities evaluated.

The preliminary accuracy assessment indicated a high level of agreement with instructor-validated responses within the limited set of course-related questions examined. However, this evaluation was based on a small number of questions derived from the same educational materials used to develop the system and should therefore be interpreted cautiously.

Importantly, this study did not assess learning outcomes, academic performance, knowledge retention, or other indicators of educational effectiveness. Consequently, no conclusions can be drawn regarding the impact of ChemIA on student learning. Rather, the results provide exploratory evidence regarding students' perceptions of the tool and its alignment with course materials in an authentic classroom setting.

Future research should involve larger and more diverse student populations, independent evaluation datasets, objective learning measures, and more rigorous comparative designs to better understand the educational value, limitations, and broader applicability of domain-grounded AI systems in higher education.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical considerations

This study was reviewed and approved by the Research Bioethics Committee of the Universidad Católica de Santa Fe (Resolution No. 3/2025). All participants provided informed consent prior to participation. Participation was entirely voluntary, and students could decline participation without any academic consequences. Survey responses were collected and analysed confidentially using anonymised data. To minimise potential influences associated with the instructor–student relationship, participation or non-participation had no effect on course grades, academic standing, or access to educational activities.

Financial disclosure statement

This research received support from the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), the Universidad Nacional del Litoral (UNL), and the Universidad Católica de Santa Fe (UCSF). The funding sources had no involvement in the study design, data collection, analysis, interpretation of data, writing of the report, or in the decision to submit the article for publication.

Acknowledgments

The authors gratefully acknowledge the support provided by the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), the Universidad Nacional del Litoral (UNL), Universidad Tecnológica Nacional (UTN) and the Universidad Católica de Santa Fe (UCSF). The authors also thank the pharmacy students who participated in the classroom activities and surveys associated with this study.

Associated content

The Supplementary contains: ChemIA system instructions; Medicinal Chemistry course syllabus; Accuracy test (15 validated questions); Student surveys and anonymised responses; Case study data and assignments; Sample ChemIA outputs (e.g., NSAID podcast script).

References

Al-Hafdi, F. S., & AlNajdi, S. M. (2024). The effectiveness of using chatbot-based environment on learning process, students' performances and perceptions: A mixed exploratory study. *Education and Information Technologies*, **29**(15), 20633–20664. https://doi.org/10.1007/s10639-024-12671-6

Ariely, M., Nazaretsky, T., & Alexandron, G. (2024). Causal-mechanical explanations in biology: Applying automated assessment for personalized learning in the science classroom. *Journal of Research in Science Teaching*, **61**(8), 1858–1889. https://doi.org/10.1002/tea.21929

Caccavale, F., Gargalo, C. L., Gernaey, K. V., & Krühne, U. (2024). Towards Education 4.0: The role of large language models as virtual tutors in chemical engineering. *Education for Chemical Engineers*, **49**, 1–11. https://doi.org/10.1016/j.ece.2024.07.002

Dood, A. J., Dood, J. C., Cruz-Ramírez de Arellano, D., Fields, K. B., & Raker, J. R. (2020). Analyzing explanations of substitution reactions using lexical analysis and logistic regression techniques. *Chemistry Education Research and Practice*, **21**(1), 267–286. https://doi.org/10.1039/C9RP00148D

Exintaris, B., Karunaratne, N., & Yuriev, E. (2023). Metacognition and critical thinking: Using ChatGPT-generated responses as prompts for critique in a problem-solving workshop (SMARTCHEMPer). *Journal of Chemical Education*, **100**(8), 2972–2980. https://doi.org/10.1021/acs.jchemed.3c00481

Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., & Romero, C. (2019). Text mining in education. *WIREs Data Mining and Knowledge Discovery*, **9**(6), e1332. https://doi.org/10.1002/widm.1332

Frydenberg, M. (2008). Principles and pedagogy: The two Ps of podcasting in the information technology classroom. *Information Systems Education Journa*l, **6**(6), 1-11.

Gribbins, M. (2007). The perceived usefulness of podcasting in higher education: A survey of students' attitudes and intention to use. In Proceedings of the Second Midwest United States Association for Information Systems (MWAIS 2007) (Paper 6).

Guan, R., Raković, M., Chen, G., & Gašević, D. (2025). How educational chatbots support self-regulated learning? A systematic review of the literature. *Education and Information Technologies*, **30**(4), 4493–4518. https://doi.org/10.1007/s10639-024-12881-y

Guo, Y., & Lee, D. (2023). Leveraging ChatGPT for enhancing critical thinking skills. *Journal of Chemical Education*, **100**(12), 4876–4883. https://doi.org/10.1021/acs.jchemed.3c00505

Hilliger, I., Pérez-Sanagustín, M., González, C., Villalobos, E., & Celis, S. (2026). A GAI-based chatbot for scaffolding self-regulated learning. *Journal of Learning Analytics*, **13**(1), 74–88. https://doi.org/10.18608/jla.2026.9143

Kenthapadi, K., Sameki, M., & Taly, A. (2024). Grounding and evaluation for large language models: Practical challenges and lessons learned (survey). Proceedings of the ACM

SIGKDD International Conference on Knowledge Discovery and Data Mining, 6523–6533. https://doi.org/10.1145/3637528.3671467

Kim, D., Majdara, A., & Olson, W. (2024). A pilot study inquiring into the impact of ChatGPT on lab report writing in introductory engineering labs. *International Journal of Technology in Education*, **7**(2), 259–289. https://doi.org/10.46328/ijte.691

Kunitsu, Y. (2023). The potential of GPT-4 as a support tool for pharmacists: Analytical study using the Japanese National Examination for Pharmacists. *JMIR Medical Education*, **9**, e48452. https://doi.org/10.2196/48452

Lee, O., Quinn, H., & Valdés, G. (2013). Science and language for English language learners in relation to Next Generation Science Standards and with implications for Common Core State Standards for English Language Arts and Mathematics. *Educational Researcher*, **42**(4), 223–233. https://doi.org/10.3102/0013189X13480524

Lolinco, A. T., & Holme, T. A. (2023). Developing a curated chatbot as an exploratory communication tool for chemistry learning. *Journal of Chemical Education*, **100**(10), 4092–4098. https://doi.org/10.1021/acs.jchemed.3c00520

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, **10**(10), 464–470. https://doi.org/10.1016/j.tics.2006.08.004

Lu, Y., Pian, Y., Chen, P., Meng, Q., & Cao, Y. (2021). RadarMath: An intelligent tutoring system for math education. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**(18), 16087–16090. https://doi.org/10.1609/aaai.v35i18.18020

Martin, P. P., Kranz, D., Wulff, P., & Graulich, N. (2024). Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry. *Journal of Research in Science Teaching*, **61**(8), 1757–1792. https://doi.org/10.1002/tea.21903

Pence, H. E., Hightower, G., Forlenza, J., Leonard, K., McLellan, A., Suero, A., Amoah, B., Mbow, M., Borner, S., Castillo, A., & Pence, L. E. (2024). Using generative AI systems for critical thinking engagement in an advanced chemistry course: A case study. *Journal of Chemical Education*, **101**(9), 3789–3794. https://doi.org/10.1021/acs.jchemed.4c00242]

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, **55**(3), 2495–2527. https://doi.org/10.1007/s10462-021-10068-2

Rojas, A. J. (2024). An investigation into ChatGPT's application for a scientific writing assignment. *Journal of Chemical Education*, **101**(5), 1959–1965. https://doi.org/10.1021/acs.jchemed.4c00034

Ruff, E. F., Engen, M. A., Franz, J. L., Mauser, J. F., West, J. K., & Zemke, J. M. O. (2024). ChatGPT writing assistance and evaluation assignments across the chemistry curriculum. *Journal of Chemical Education*, **101**(6), 2483–2492. https://doi.org/10.1021/acs.jchemed.4c00248

Schulze Balhorn, L., Weber, J. M., Buijsman, S., Hildebrandt, J. R., Ziefle, M., & Schweidtmann, A. M. (2024). Empirical assessment of ChatGPT's answering capabilities in natural science and engineering. *Scientific Reports*, **14**, 4998. https://doi.org/10.1038/s41598-024-54936-7

Sehri, M., & Dumond, P. (2025). Integrating ChatGPT as an academic assistant. In Proceedings of the 2025 Canadian Engineering Education Association Conference (CEEA-ACÉG25) (Paper 301). Canadian Engineering Education Association. https://doi.org/10.24908/pceea.2025

Shi, Y., Yu, K., Dong, Y., & Chen, F. (2026). Large language models in education: A systematic review of empirical applications, benefits, and challenges. *Computers and Education: Artificial Intelligence*, **10**, 100529. https://doi.org/10.1016/j.caeai.2025.100529

Sigot, M., & Tassoti, S. (2025). A missed opportunity for no-code chatbots? Current challenges in publicly available chemistry GPTs. *Journal of Chemical Education*, **102**(5), 2151–2159. https://doi.org/10.1021/acs.jchemed.4c01552

Tascioglu, A. (2025). From prototype to reusable template: Designing GPT-based teaching assistants for higher education. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5467466v

West, J. K., Franz, J. L., Hein, S. M., Leverentz-Culp, H. R., Mauser, J. F., Ruff, E. F., & Zemke, J. M. (2023). An analysis of AI-generated laboratory reports across the chemistry curriculum and student perceptions of ChatGPT. *Journal of Chemical Education*, **100**(11), 4351–4359. https://doi.org/10.1021/acs.jchemed.3c00581