International Journal on Artificial Intelligence Tools
(2025) 2550015 (25 pages)

DOI: 10.1142/S0218213025500150

# Enhancing Mobile Learning Applications with Large Language Models: Design and Evaluation of AIChemApp

Ahmed Ewais

*Computer Science Department*
*Arab American University*
*P. O. Box 240, Jenin, Palestine*
*ahmad.ewais@aaup.edu*

Received 12 February 2025
Accepted 29 July 2025
Published

Traditional mobile learning is constrained by the need to update its content manually and frequently due to rigid structures. Conversely, integrating LLMs with mobile learning applications is expected to provide a catalyst for both developers and teachers on programming and technical perspectives for such applications by considering automatic content generation, dynamic assessments, and reducing technical complexities. This study presents the feasibility and effectiveness of integrating LLMs, specifically ChatGPT, with mobile learning applications developed for chemistry courses. The provided solution leverages LLM's ability to provide content generation and real-time responses related to queries about the periodic table, electron configuration, chemical equations and game-based assessment, ultimately enhancing the effectiveness and engagement of mobile learning. An experimental evaluation was conducted to assess the accuracy of ChatGPT's responses within the mobile learning application for chemistry. The results demonstrated that ChatGPT achieves an impressive 93.33% accuracy in providing information about elements and their categories from the periodic table. However, the accuracy rates for the chemical equation balancing and electron configuration tasks were relatively low, at 75.71% and 68.91%, respectively, indicating areas for further optimization and refinement. This research highlights the technical implications of LLM integration with mobile learning applications, including potential challenges and opportunities for developers and educators in terms of API communications, prompt engineering, and the quality and accuracy of AI-generated responses.

*Keywords*: Large language models; mobile learning; ChatGPT; prompt engineering; accuracy.

## 1. Introduction

Advances in mobile technology have increased its availability, and the lower cost of mobile application development and hardware diminishes many of the obstacles associated with the adoption of mobile learning applications. In addition to improvements in mobile technology, the need to integrate mobile technology with education has become important and necessary for schools and higher educational institutions due to varying levels of lockdown in some countries worldwide for health, political, etc., issues and disasters. In recent years, the number of users accessing the internet via mobile devices — whether at home, at school, or in the workplace — has significantly increased compared with that in previous decades, largely due to the widespread availability of smartphones and mobile networks.[1] These factors make mobile learning applications interesting solutions for supporting digital transformation in education and have led many researchers to view them as an important proposition for remote learning.[2–4] Different studies have shown that mobile learning offers improvements over traditional textbooks, enhancing students' comprehension, retention and knowledge level.[5]

Among the recent advances in mobile learning, the utilization of AI to support mobile learning attracted the interest of researchers,[6] who have started to propose different solutions for smart educational learning.[7] The significance of using AI in technology-enhanced learning environments has been recognized for nearly half a century, as evidenced by conversation theory.[8] For example, the integration of AI with mobile learning has been investigated previously in different research contexts. In Ref. 9, researchers proposed an AI adaptive user interface for mobile learning applications called the Mobile Academy, which predicts cultural aspects and language preferences to offer a culturally adapted user interface. Furthermore, intelligent mobile applications have been proposed to assist visually impaired people during their daily consumption.[10] Recent studies such as Ref. 1 presented a review study that classifies automated feedback systems on the basis of features such as architecture (domain, expert knowledge, student feedback, student data, feedback generation model, and implementation), educational context (domain, level, and settings), and feedback type (adaptiveness, timing, learner control, and purpose).

Recently, the availability of large language models (LLMs) has accelerated adoption in mobile learning contexts since LLMs have the ability to perform comprehensive analyses of different queries and provide intelligent answers to support interactivity, create learning content, and support personalized learning experiences.[11,12] They also assist in enhancing automatic feedback and facilitating real-time assistance. Examples of LLMs include the Chat Generative Pretrained Transformer (ChatGPT), which was developed by OpenAI and released for public use in November 2022; Gemini, which was developed by Google and released for public use in February 2024; and Copilot, which was developed by Microsoft and released for public use in March 2023.

In previous years, various investigations have explored the advantages of using LLMs in general educational contexts.[13,14] Some studies have utilized LLMs' ability to handle natural question-and-answer exchanges with users.[15,16] However, the application of LLMs in the field of chemistry still needs more investigation. Some studies, such as Ref. 17, have shown web-based learning for the chemistry domain integrated with ChatGPT, which focused on evaluating AI-generated responses to chemistry assessment questions. Similarly, Clark[18] reported different abilities to address questions in two formats (closed-response format and open-response format) related to questions from final exams from two general chemistry courses. Previous studies concluded that LLMs as chatbots are ill-equipped to provide reliable answers, especially for application- and interpretation-based questions, as well as questions involving nontextual information.

Conversely, a study[19] revealed that GPT-4 has interesting prediction results, reaching 70%, concerning eight chemistry tasks: name prediction, molecular property prediction, yield prediction, reaction prediction, reagent selection, retrosynthesis, text-based molecule design, and molecule captioning. A study presented in Ref. 20 showed that ChatGPT reached 60% accuracy in finding the molecular point groups of the molecules, and it reached 69% accuracy in finding the octanol−water partition coefficients of essential oil components. In general, the previous studies were performed via web-based applications, and zero-shot techniques were not used to obtain answers from LLMs. Additionally, the language used was solely English. To improve the accuracy of LLMs, Ref. 21 proposed an LLM-powered chemistry engine called ChemCrow, an agent designed to assist users in synthesizing target molecules and identifying structurally or functionally similar molecules. Although the obtained accuracy results were better than those of GPT-4 in factors such as chemical accuracy, quality of reasoning, and task completion, ChemCrow fails to provide relevant conclusions for a number of chemistry tasks.

However, to our knowledge, no previous studies have investigated the use of ChatGPT for chemistry courses in the Arabic language. Most of the previous studies[17–19,21,22] investigated the use of ChatGPT as a chatbot application so that it can answer a sequence of questions via web-based applications or tend to focus on specific case studies. Furthermore, several previous studies have investigated the use of LLMs in languages other than English. For example, the work presented in Ref. 23 proposed a solution for undergraduate students who are not fluent in English to complete a formative assessment in their preferred language, such as German, in a chemistry course. Another study[24] reported the use of ChatGPT, not particularly for chemistry, in higher education for Latin American universities. Another study[25] proposed ChemLLM as the first LLM dedicated to chemistry in the Chinese language. Therefore, this study is among the first attempts to examine the integration of ChatGPT in mobile learning for chemistry courses for Arabic-speaking users. This study explores GPT-4 capabilities within the field of chemistry, which has the potential to expedite research and development activities.

The main goal of this research work is to investigate the efficacy of using ChatGPT in the mobile learning context in the Arabic language. It shows the technological improvements driving innovation in this domain, as the paper explores different technical perspectives and insights for developers, researchers and policy makers involved in the development and implementation of future mobile learning applications. This study seeks to provide a number of technical requirements related to the use of an LLM (GPT-4) to adapt and provide dynamic feedback on user interaction during the use of mobile learning applications. This work emphasizes enhancing different modules of the previously developed AIChemApp via GPT-4 by improving interactivity, personalizing learning experiences, and providing real-time support while learning chemistry topics.

## 2. Materials and Methods

Different obstacles are encountered in traditional mobile learning applications, such as providing static content fixed at the time of application development. Additionally, many of the previous applications are implemented as ad hoc solutions, making timely updates difficult for developers. Furthermore, this research builds on findings from our earlier studies,[26,27] which emphasized the need to explore more chemical equations and a broader range of exercises. Even the integration of LLMs into mobile learning applications offers a promising approach to overcoming previous limitations by enhancing the efficiency and quality of learning support. Advancements in this field are still challenging, and research gaps persist. For example, as highlighted in previous studies,[18,28] LLM-generated responses often contain inaccuracies and raise concerns about their reliability. Moreover, there have been limited efforts to integrate LLMs into mobile learning applications, highlighting the need for further exploration of technical development requirements in this area.

Motivated by previous findings and insights from prior research, this study investigates the role of LLMs in enhancing mobile learning applications, focusing on their ability to enrich resources and content and ensure seamless integration. Furthermore, this study focuses on the challenges of integrating advanced machine learning models, specifically LLMs such as GPT-4, into mobile learning applications for dynamic content generation.

In particular, this research explores the possibility of integrating ChatGPT as an LLM to allow users to explore diverse chemical equations, interactive exercises, and games. As a proof of concept, an updated version of ChemApp,[26] called AIChemApp, was developed to validate the proposed solution. This updated version introduces significant advancements, including the consolidation of two previous dynamic modules—Chemical Equations and Chemical Lab — into a single Interactive Chemical Lab module. This consolidation optimizes computational efficiency, eliminates code redundancy, and streamlines the integration of application programming interface (API)-driven LLM functionalities.

Additionally, AIChemApp leverages LLM in the four modules embedded in the application to provide AI-driven content and feedback. As reported in Ref. 19, GPT-4 outperformed other LLMs, as it had competitive results in responding to a number of chemistry tasks using zero-shot or few-shot LLMs in the chemistry domain. Therefore, GPT-4 is adopted to support AIChemApp with different intelligent feedback tailored to specific goals for each module in AIChemApp via the zero-shot technique. Following the categorized evaluation methods in Refs. 1 and 29, this research adopted the accuracy of GPT-4's results on the basis of expert evaluation.

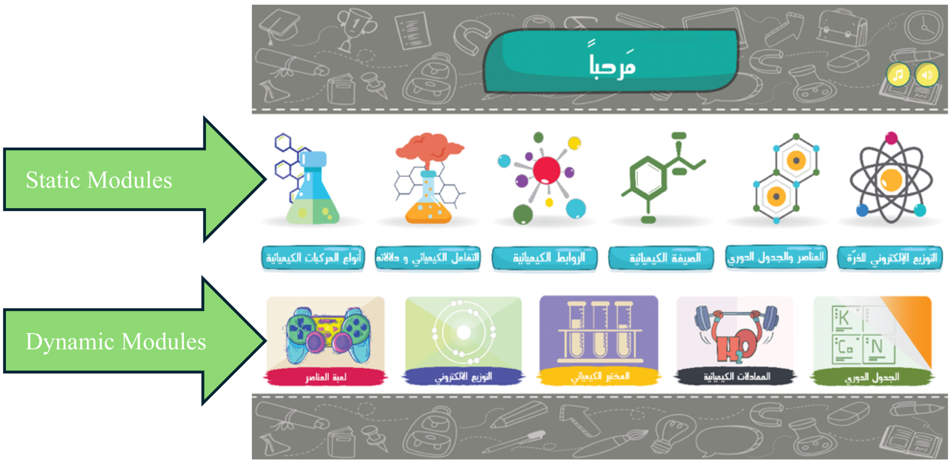



Fig. 1. Main interface of AIChemApp.

AIChemApp included materials and exercises related to lessons from chemistry course textbooks for the eighth-grade level in elementary schools in Palestine, as described in our previous research.[26,27] More technical details and requirements for the use of GPT-4 with AIChemApp are explained in the following subsections.

### 2.1. *ChatGPT integration with AIChemApp*

AIChemApp provides two types of content: static modules and interactive modules (see Fig. 1). The static module is used to provide users with educational videos generated via whiteboard animation to explain different lessons from the textbook. On the other hand, the dynamic modules are interactive periodic tables, interactive chemical laboratories, electron configurations, and molecular games. The interactive modules leverage LLM to provide users with AI-generated feedback aligned with each module's purpose.

Starting with the lessons, which are available in the upper part (specified for the static module), the main goal of the lessons is to explain the same content found in the textbook's lessons using the whiteboard animation technique. Both snapshots in Fig. 2 show two video files for separate lessons related to chemical equations. The videos primarily contain text explanations and images for each lesson. The application allows seamless integration, updating, and removal of lessons through a content provider object in the Android framework. This object organizes lesson video links in a structured key-value format and XML files, enabling efficient content management.

As mentioned earlier, to integrate AIChemApp with an LLM such as GPT-4, the use of an API facilitates communication between the different modules and the LLM. This integration ensures seamless interaction, enabling AIChemApp to harness the LLM's ability to generate contextually relevant responses tailored to


(a) (b)

Fig. 2. Whiteboard animation videos created for different lessons.

module-specific queries. The API serves as the bridge, allowing AIChemApp to send structured requests containing details about the modules and the required information from LLMs. Once the LLM processes the requests (queries), it generates responses that are returned to the module via the API connection session, fetching the answers (responses) from the LLM and extracting the required details to be displayed in the GUI.

For AIChemApp, each module has distinct goals, varying levels of detail, and unique GUI styles. Consequently, structured JavaScript Object Notation (JSON) files tailored to each module are generated to maintain consistency. These JSON outputs ensure that module-specific requirements, such as data granularity and GUI integration, are met effectively. Additionally, all queries are stored as key-value pairs in a content provider object within the Android framework. This approach simplifies future updates and facilitates the development of upgraded application versions. To support Arabic-speaking users, queries are stored in Arabic to ensure linguistic accessibility and ease of use. By presenting explanations in Arabic and maintaining technical labels in English in JSON files, AIChemApp effectively caters to its target audience, offering a bilingual educational experience.

Following the proposed prompt patterns for conversational LLMs,[30–33] this study adopts the *template pattern* for prompt engineering. This approach structures the prompts to guide the LLM's behavior and specify the format of its output. For example, the research presented in Refs. 19, 20, and 33 explored the use of prompt engineering for LLMs to obtain AI-generated feedback for chemistry topics and tasks. According to Ref. 34 the first directive statement in the prompt should define a specific context, desired response type, and output format. The proposed prompt template used in this study transforms GPT-4 responses into a structure that is consistent with the formatting needs of the AIChemApp modules and its graphical user interface. This approach is needed when the target format is not known to the LLM. This ensures that the LLM-generated content adheres to the formatting requirements and objectives of different AIChemApp modules and their graphical user interfaces (GUIs).

From a prompt engineering perspective, AIChemApp's modules send requests (queries) to GPT-4 via a contextual statement. This statement includes three key components: context, question, and output_format. The context key specifies the module's name and purpose. The question key represents the query that will be sent to GPT-4 from each module. Finally, output_format defines the elements to be included in the retrieved JSON file, ensuring compatibility with AIChemApp's GUI and functional requirements. The replaceable fields in the statement's structure, indicated through italicized and underlined text, allow for dynamic customization. This customization is implemented via a dedicated Java method, enabling user-specific interactions and tailoring outputs to the goals of each module. More details about the adopted prompt template are explained in the following subsections.

This study defines an algorithm to leverage API connectivity, JSON parsing, and user interface integration to achieve seamless interaction between AIChemApp modules and GPT-4. The design ensures robust data exchange and processing, facilitating contextually relevant responses generated by the LLM. By incorporating a computational approach that prioritizes efficiency and modularity, the algorithm focuses on meeting the application's technical and functional requirements while maintaining the reliability of dynamic data handling. Finally, the algorithm extracts the anticipated details for each module from the LLM's responses to provide users with the necessary information. The following pseudocode presents the proposed sequence of actions required to integrate LLM into AIChemApp.

Pseudo code for integrating ChatGPT with ChemApp's modules:

```
#Start
Retrieve query components depending on the selected task of the module
JSONObject = Append components to the module's JSON file
Connection = establish connection with GPT-4 using API key
If(connection)
  JSONResponse = send JSONObject to GPT using established connection
  var = Parse(JSONResponse) //to extract retrieved data from GPT-4 depending
                    on the module requirements
  displayInfo on Module_UI //to display received information on ChemApp UI
else
  Display Error_Message
#End
```

The following subsections present the adopted approach for integrating GPT-4 with each module in AIChemApp. Each subsection presents a description of the module and its GUI, primary objective and purpose of the module, specified prompt template, and a sample of obtained results from GPT-4.

#### 2.1.1. *Interactive periodic table module*

The interactive periodic table module mainly presents the element names and symbols, atomic masses, atomic numbers, and both groups and categories (see

Table 1. Prompt template for the interactive periodic table module.

| Key | Value |
|---|---|
| Context | This query is sent from "*Periodic Table*" Module which is responsible for "*displaying structured information related to elements, groups, and categories presented in periodic table*." The respond will be displayed for users whose age is between 14–16 years old. |
| Question | ماهي "*الفلزات القلوية*" في الجدول الدوري؟ |
| Output_format | يكون الرد بصيغة ملف JSON وهذا الملف يحتوي على ما يلي:<br>"*اسم المجموعة name،*<br>*ميزات هذه المجموع features،*<br>*استخداماتها في الصناعة وغيرها uses،*<br>*التحذيرات المتعلقة بالمجموعة وعناصرها ان وجد warnings،*<br>*واخيرا اسماء العناصر التابعة للمجموعة elements*".<br>يكون عناصر الملف باللغة الانجليزية والقيم باللغة العربية. |

Fig. 3(a)). While a group presents a vertical column of elements that have similar chemical properties, a category is a classification of elements on the basis of their properties (e.g., metals, nonmetals, and metalloids). By selecting an element from the periodic table, the user will be directed to a new screen that displays the property of the element in a symbolic way, as shown in Figs. 3(b) and 3(c).

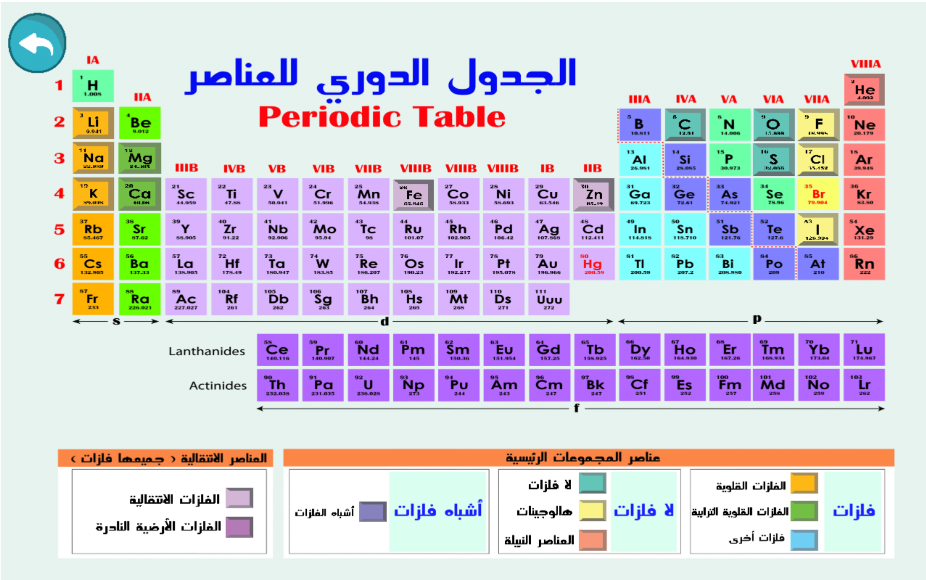
The primary objective of integrating GPT-4 with the periodic table module is to provide AI-generated textual explanations for a selected chemical element or its category. Therefore, upon selecting an element from the module's GUI, an AI-generated textual explanation is seamlessly presented in the interface (see Fig. 4) via a customized prompt template.

To meet the requirements of the interactive periodic table module, a customized contextual prompt template is sent to GPT-4. This template leverages the *context* key to specify the module's name and purpose, providing necessary background information. The key, which is dynamically generated based on user input (e.g., selecting an icon from the periodic table), targets specific elements or groups, such as alkali metals, to tailor the query. The *output_format* key organizes the response into a JSON structure containing details such as the group or element name, features, uses, warnings, and related elements. The structure of the JSON file adapts depending on the selected icon. Table 1 lists the customized prompt template used as the input message for GPT-4. The English translations of the Arabic content in the following tables and figures are provided in Appendix C for clarity.

To ensure that GPT-4 generates structured and relevant responses, the query includes the category name and specifies key details, such as a list of chemical elements, primary characteristics, industrial or practical applications, and safety-related warnings. Figure 5 presents an example of a JSON file generated by GPT-4 on the basis of this query, providing detailed information about the metal groups (e.g., alkaline metals) displayed in the periodic table.

### 2.1.2. *Interactive chemical lab module*

For the interactive chemical lab module, the interface features a series of test tubes representing chemical elements, each labeled with its name from the periodic table.




(a)


(b)


(c)

Fig. 3. Periodic table and details about Ca or K elements selected from the periodic table.

Fig. 4. Selecting a category and its corresponding elements are automatically highlighted and AI-generated textual detail is displayed.


{"group": {
  "name": "الفلزات القلوية",
  "features": [
    "فلزات ناعمة ولامعة",
    "عالية التفاعل، خاصة مع الماء",
    "لها إلكترون واحد في غلافها الخارجي",
    "تكون أيونات موجبة +1 بسهولة"
  ], "uses": [
    "تستخدم في تصنيع الصابون والمنظفات",
    "تدخل في صناعة البطاريات",
    "تستخدم في إنتاج الأسمدة الكيميائية",
    "تستخدم في الأضواء الكاشفة وأجهزة الإنارة"],
  "warnings": [
    "شديدة التفاعل مع الماء والهواء",
    "يجب تخزينها تحت الزيت أو في جو خامل",
    "يمكن أن تسبب حروقًا عند ملامستها للجلد",
    "تسبب تفاعلات انفجارية مع الماء"
  ], "elements": [
    "الليثيوم",
    "الصوديوم",
    "البوتاسيوم",
    "الروبيديوم",
    "السيزيوم",
    "الفرانسيوم" ] }}

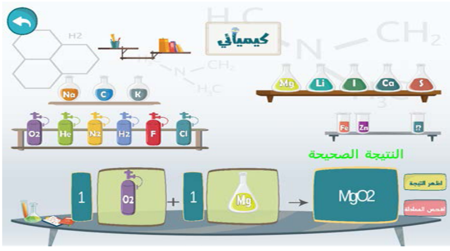

Fig. 5. Sample JSON file generated from a query sent to GPT-4, requesting information about the metals group in Arabic.

The interface also includes numerical inputs for specifying the number of atoms required to balance chemical equations. Users can drag and drop the test tubes for each chemical element and adjust the number of atoms accordingly to formulate an equation. Once the equation is constructed, the user has two options: “show result” or “check equation” (see Fig. 6). The result button shows the correct balanced equation retrieved from GPT-4. The “check equation” button provides AI feedback on the equation entered by the user. Accordingly, the user will be notified whether the answer is correct.

The primary objective of this module is to help users understand potential chemical reactions between various chemical elements and provide corrective feedback on formulated balanced equations. By doing so, users can learn the correct number of atoms required for specific reactions and gain a clear understanding of the resulting compounds.

The integration of GPT-4 into this module provides corrective feedback, directly informing users about the correctness of their formulated chemical equations. Additionally, the integration enriches the learning process by delivering dynamic expla-


(a)




(b)

Fig. 6. Formulating a balanced equation using different GUI components.

nations of why a specific reaction succeeded or failed, fostering a deeper understanding of chemical reaction principles. This functionality is implemented via a contextual statement that is based on the adopted prompt template, which includes details that are specific to the interactive chemical lab module. For example, the context key defines the module's name and primary objective, the question key includes a predefined question including the formulated equation by the user, and the output format key specifies the JSON structure, including attributes such as a compound name, a Boolean value indicating the equation's correctness, the correct balanced equation, and additional details. A sample contextual statement to query GPT-4 for the user-formulated equation ($Ca + O2 -> CaO$)? is shown in Table 2.

The module is able to process JSON files generated by GPT-4 to extract and map relevant details to corresponding components in the GUI. Figure 7 presents two sample JSON files created by GPT-4 in response to queries about specific chemical equations, demonstrating the structured output and its alignment with the module's requirements.

### 2.1.3. *Electron configuration module*

The *electron configuration* module visually represents atoms of different chemical elements, including their electrons distributed across energy levels or shells (see Fig. 8). The graphical user interface (GUI) enables users to interact with visual elements by dragging and dropping components, such as a chemical element and its

Table 2. Prompt template for the interactive chemical lab module.

| Key | Value |
|---|---|
| Context | This query is sent from "*Interactive Chemical Lab*" Module which is responsible for "*helping students in understanding possible chemical reaction between the different chemical elements and provide feedback on formulated balanced equations*." The respond will be displayed for users whose age is between 14–16 years old. |
| Question | هل المعادلة التالية موازنة وصحيحة:"*Ca + O2 –> CaO*"؟ |
| Output_format | يكون الرد بصيغة ملف JSON وهذا الملف يحتوي على ما يلي:<br>"*اسم المركب الناتج compound_name* ،<br>*صحة المعادلة is_equation_correct* ،<br>*المعادلة الصحيحة correct_equation*،<br>*تفاصيل أخرى*<br>*details: {explanation:… reaction_type:… other_info:…* }"<br>يكون عناصر الملف باللغة الانجليزية والقيم باللغة العربية. |

Table 3. Prompt template for the electron configuration module.

| Key | Value |
|---|---|
| Context | This query is sent from "*Electron Configuration*" Module which is responsible for "*helping the user to know how many electrons that are available in each energy level (orbit) around an atom for chemical element.*" The respond will be displayed for users whose age is between 14–16 years old. |
| Question | هل التوزيع الالكتروني لذّرة "*البيريليوم Be4* 1s^2 2s^2" صحيح؟ |
| Output_format | يكون الرد بصيغة ملف JSON وهذا الملف يحتوي على ما يلي:<br>"*اسم العنصر name*،<br>*الرمز symbol* ،<br>*صحة الحل is_correct*،<br>*التوزيع الالكتروني الصحيح* "1s^2 2s^2"<br>يكون عناصر الملف باللغة الانجليزية والقيم باللغة العربية. |

corresponding number of electrons, into their correct positions. Upon completing the electron configuration for a chemical element, the user can click the 'Show Result' button to receive a notification indicating whether the constructed electron arrangement for the selected element is accurate, enhancing their understanding of the atomic structure.

The primary objective of this module is to assist users in understanding how electrons are distributed across energy levels (orbits) around an atom for a given chemical element. To achieve this, the user is prompted to drag a selected chemical element to the center of the orbit and then input the corresponding number of electrons in each orbit. By integrating GPT-4, the module compares the user-entered electron configuration with the correct configuration retrieved from GPT-4 in a structured JSON response. Additionally, the *module* provides feedback explaining why certain configurations are stable or unstable, enhancing the user's understanding of atomic structures. This process is *implemented* by *considering* a customized prompt template that queries GPT-4 to generate an accurate electron configuration for *a* selected element (see Table 3).

The result obtained from GPT-4 includes comprehensive details about the name of the element, symbol, atomic number, a Boolean value indicating whether the

```
{
 "reaction": {
   "compound_name": "أكسيد الكالسيوم",
   "is_equation_correct": غير صحيح,
   "correct_equation": "2Ca + O2 → 2CaO",

"details": {
     "explanation": " المعادلة غير موزونة، حيث أن جزيء
الأكسجين (O2) واحد من يتفاعل مع ذرتين من الكالسيوم لإنتاج
جزيئين من أكسيد الكالسيوم (CaO). يجب موازنة المعادلة بإضافة
ذرتين من الكالسيوم على الجانب الأيسر.",
     "reaction_type": "تفاعل تكوين",
     "other_info": " أكسيد الكالسيوم يعرف أيضًا بالجير الحي
ويستخدم في العديد من التطبيقات الصناعية مثل إنتاج الأسمنت
ومعالجة المياه."}}}
```


(a)


```
{
 "reaction": {
   "compound_name": " نترات البوتاسيوم + ثاني أكسيد
الكربون + ماء",
   "is_equation_correct": غير صحيح,
   "correct_equation": "K2CO3 + 2HNO3 → 2KNO3
+ CO2 + H2O",
   "details": {
     "explanation": " المعادلة غير موزونة. عند تفاعل
كربونات البوتاسيوم (K2CO3) مع حمض النيتريك (HNO3) ، يتم
إنتاج نترات البوتاسيوم (KNO3) بالإضافة إلى ثاني أكسيد الكربون
(CO2) والماء (H2O). يجب موازنة المعادلة بإضافة جزيئين من
حمض النيتريك للحصول على المعادلة الصحيحة.",
     "reaction_type": "تفاعل حمض وقاعدة",
     "other_info": " هذا التفاعل يعد جزءًا من التفاعلات التي
تحدث عند استخدام كربونات البوتاسيوم كمادة قلوية لتعديل
الأحماض."}}}
```


(b)

Fig. 7. Two JSON sample files include feedback about user-formulated equation in Arabic.

user's input is correct, and the correct electron configuration. These details are structured in JSON format as shown in Fig. 9 and visually represented in the module's interface.

### 2.1.4. *Molecules game module*

The game module incorporates a stage-based design where each stage consists of multiple questions derived from exercises in the course textbook (see Fig. 10(a)). The questions are based mainly on exercises that are available in each lesson of the course textbook. Each question can be answered by selecting the different characters that form the correct answer. For correct answers, a clapping sound will be played, and award points will be stored in the user's profile, unlocking the next stage. To encourage accuracy, the game allows up to three attempts per question, after which the correct answer is automatically displayed if all the attempts fail. The GUI is

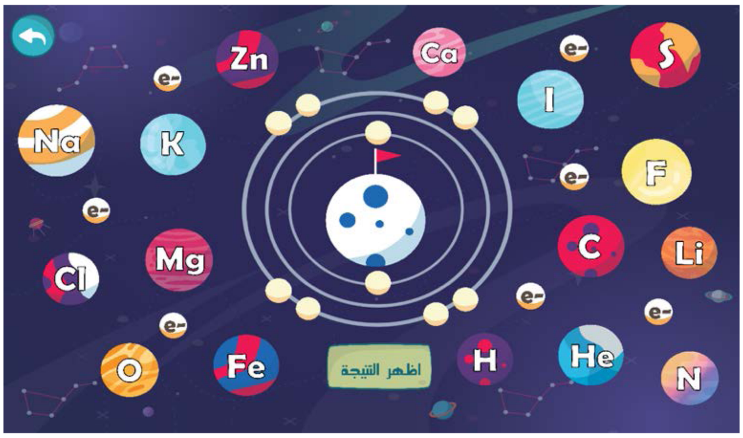



Fig. 8. Electron configuration module to allow the user to specify electrons in each orbit.

{ "element": "بيريليوم",
"symbol": "Be",
"atomic_number": 4,
“is_correct”: “نعم”,
"electronic_configuration": "1 s^2 2 s^2"}

Fig. 9. Structure of JSON file for response from GPT-4 in electron configuration module.

split into two sections: one for displaying the questions and the other for presenting answer options as characters, as shown in Fig. 10(b).

The primary goal of this module is to assess the user's knowledge and understanding level of each topic covered in the lessons. To utilize GPT-4 with this module, each game phase has its questions stored in objects instantiated from the *content provider* class in the Android framework. Questions are extracted from a *content provider* object, appended to the request, and sent to ChatGPT for contextual processing. Therefore, the customized contextual statement for this module includes details about the module (context key), three questions (question key), and JSON elements (output format) to structure the response appropriately for integration with the game module, as shown in Table 4.

The extracted feedback from the generated JSON file is displayed inside the corresponding GUI components. Figure 11 shows the JSON file contents produced in response to the prior query sent to GPT-4, showcasing three questions along with their correct answers and the letters available to construct the correct answers for each question.

## 3. Evaluation

Aligned with previous evaluation frameworks for AI-driven mobile applications,[29,35–37] this study adopted an evaluation methodology focusing on accuracy assessment based on expert evaluation. The evaluation aimed to validate the


(a)


(b)

Fig. 10. Stage-based game including three questions in each stage.

Table 4. Prompt template for the molecule game module.

| Key | Value |
|---|---|
| Context | This query is sent from "*Molecules Game*" Module which is responsible for "*assessing the user's knowledge about topics covered in the lessons*". The respond will be displayed for users whose age is between 14–16 years old. |
| Question | لعبة كلمات بالاجابة عن الثلاثة أسئلة كالتالي:<br>"*ماهي أصغر جزء في المادة؟*<br>*ما هو عدد البروتينات في الذرة؟*<br>*ماذا تسمى الجسيمات موجبة الشحنة في نواة الذرة؟*"<br>اللعبة تظهر أحرف لاستخدامها في الاجابة على هذه الثلاثة اسئلة |
| Output_format | يكون الرد بشكل ملف JSON يحتوي على التالي:<br>"*نص السؤال question*،<br>*الاجابة الصحيحة correct answer* ،<br>*والاحرف ذات الصلة letters*،"<br>يكون عناصر الملف باللغة الانجليزية والقيم باللغة العربية. |

accuracy, relevance, and completeness of GPT-4 integration across AIChemApp's modules. The process involved structured task execution, expert analysis, and quantitative accuracy calculations. The evaluation process was divided into four steps to ensure clarity and reproducibility as follows.

**Step 1 (Dataset Preparation):** Each of the four AIChemApp modules was assigned a tailored dataset based on topics covered in the textbook and other sup-

```
{"quiz_game": [
  {
   "question": "ما هو أصغر جزء في المادة؟",
   "correct_answer": "ذرة"
  },
  {
   "question": "كم عدد البروتونات في الذرة؟",
   "correct_answer": "عدد ذري" },
  {"question": "ما هي الجسيمات موجبة الشحنة في نواة الذرة؟",
    "correct_answer": "بروتونات"
   }
  ],
 "letters": ["ذ", "ر", "ة", "و", "م", "ك", "ت", "ع", "د", "ذ", "ي", "ب", "ن", "و", "ب", "ر", "ت", "و", "ن", "ا", "س", "م", "ت"]}
```


Fig. 11. JSON file for GPT-4 response in molecules game module.

plementary chemistry resources recommended by domain experts, ensuring comprehensive coverage of real-world application scenarios. For *the interactive periodic table module*, thirty-five elements and 5 categories were selected for use in queries designed to extract factual information from GPT-4, such as element properties, group information, and classifications. Concerning *the interactive chemical lab module*, seventy chemical equations were curated — extended beyond the textbook examples per the evaluators' recommendations — to assess GPT-4's ability to verify, balance, and explain chemical reactions. *Additionally, the electron configuration module* has seventy-four chemical elements, with their correct electron configurations used as a dataset to examine the accuracy of GPT-4 in generating the correct electron distribution across energy levels. Finally, the *Molecules Game Module* includes thirty open-ended questions derived from textbook exercises as a dataset to evaluate the model's capacity to generate relevant feedback in game-based assessments. All the input queries were formatted via a plain text (i.e., without subscripts or superscripts), which is consistent with the platform's technical constraints and ensures uniformity in the evaluation process.

**Step 2 (Application Demonstration):** The author used the tailored datasets for the 4 modules to simulate real user interactions. For each module, the input dataset was sent to GPT-4 via the application's GUI and API, and the returned JSON-structured responses were parsed and rendered in real time within the application's GUI. This step allowed for a complete assessment of the end-to-end system behavior, from user input through AI processing to final display, providing insight into both functionality and content quality. To manage the scale of testing, a randomly sampled validation subset was adopted for each module, maintaining a representative sample while reducing the overhead of repeated API requests.

**Step 3 (Expert Review):** A panel of five PhD-level Arabic-speaking chemistry experts reviewed GPT-4 responses across all the modules. The experts were provided with a document that included screenshots of the results obtained from the application. The screenshots, similar to Figs. 3, 4, 6, 8, and 10, were all collected inside the shared document, and an explanation of each performed interaction was added to the screenshot to clarify the context of the interaction. The experts were informed that their focus should be on content accuracy, relevance, and quality, independent of GPT-4's AI nature, usability or technical perspectives. Additionally, reference materials (e.g., electron distribution charts, balanced equation keys, and textbook excerpts) were provided to the experts to ensure consistency. To mitigate possible bias, they were explicitly told that the results were generated within a mobile learning application. This clarification ensured that their assessments were based solely on content quality and accuracy rather than preconceived notions about AI tools.

Accordingly, a structured evaluation rubric was developed on the basis of prior AI evaluation methodologies,[29,35–37] outlining module-specific criteria. The rubric was implemented as a table containing the following fields: interaction number, module name, description of the interaction, screenshot of the obtained result, evaluator's decision (true or false), and expert comments (optional). Concerning the interactive periodic table, interactive chemical label, and electron configuration modules, the accuracy of the generated AI content was evaluated on the basis of expert knowledge and predefined answer keys. In contrast, the expert's evaluation of the Molecules Game Module was based on the AI responses' relevance and completeness, using conceptual questions derived from the textbook. The experts were encouraged to provide comments justifying their decisions when the responses were false. All evaluation forms from individual experts were systematically compiled into a spreadsheet for further analysis (see Appendix B).

**Step 4: (Accuracy Measurement and Results):** Finally, on the basis of the evaluators' reviews, the accuracy rate for each module was calculated to evaluate the reliability of AIChemApp in generating specific and accurate content. This involves computing the accuracy metric for each module via the following formula:

$$\text{Accuracy} = \frac{\text{Total Number of Responses}}{\text{Number of Correct Responses}}. \tag{1}$$

This measurement metric serves to objectively compare performance across modules and determine the reliability of utilizing GPT-4 in different scenarios for mobile learning applications. The results and comparative analysis are presented in the Results and Discussion section.

## 4. Result and Discussion

The integration of GPT-4 within AIChemApp demonstrated significant advancements in enhancing the functionalities of mobile learning applications for chemistry education. Each module's performance was evaluated on the basis of its ability to generate accurate and contextually relevant content tailored to each module's objective. The periodic table module achieved a high accuracy rate of 93.33%, as shown in Table 5. The obtained details related to chemical elements, groups, and categories were general details that are relevant to information from the textbook. However, the interactive chemical lab module has a lower accuracy result (75.71%). Additionally, the accuracy of the electron configuration module decreased to 68.91%, which is the lowest ratio. The electron distribution for the first 18 elements that

Table 5. Accuracy results for the three modules in AIChemApp.

| Module name | Total queries | Accurate feedback | Accuracy ratio (%) |
|---|---|---|---|
| Interactive Periodic Table | 40 | 36 | 93.33% |
| Interactive Chemical Lab | 70 | 53 | 75.71% |
| Electron Configuration | 74 | 51 | 68.91% |

require the first three levels has 100% accuracy results. However, for elements that need more than three levels (from $_{19}$K and above), 23 elements (out of 74) were reported as false feedback from GPT-4, and some exceptions for electron distribution were not covered by GPT-4, such as $_{24}$Cr, $_{29}$Cu, $_{41}$Nb, $_{42}$Mo, and $_{47}$Ag, which have exception rules related to electron distributions at the 4th and 5th levels.

With respect to consistency results for the game module, three evaluators compared twenty predefined questions (extracted from the chemistry textbook) against responses from GPT-4. The teachers confirmed that the obtained details and responses were consistent with the answers for each question by following the rubric provided to the evaluators.

The findings of this study demonstrate the feasibility of integrating an LLM (ChatGPT) with a mobile learning application related to a chemistry course. Emerging tools such as Gemini and Copilot provide additional opportunities for integration. The ongoing advancements in LLMs offer the potential to refine feedback mechanisms, providing more accurate and contextually relevant responses. Additionally, this study emphasizes technical strategies for effective communication with LLMs, including the use of APIs, JSON, XML, and parsing classes, to ensure tailored feedback depending on the purposes of the mobile application. These technical approaches contribute to a sustainable and scalable information retrieval framework for educational applications and LLMs.[38]

In the periodic table module, integration was achieved by enabling users to interact with interconnected learning materials that support "multiple meanings and multiple concepts".[39] This was implemented by sending predefined queries to GPT-4 to retrieve factual details about the elements, categories, and groups presented in the periodic table. The obtained feedback was displayed in real time within the AIChemApp GUI. Among all the modules, this module achieves the highest accuracy rate because of GPT-4's advanced knowledge base and its ability to provide precise and reliable details about the periodic table. Similar findings were reported in Ref. 40, which highlighted GPT-4's strong performance in answering general open-ended questions, achieving accuracy rates exceeding 80%. However, challenges remain in achieving complete accuracy.[36] Furthermore, prior studies[18,19,41] have shown that LLMs sometimes generate descriptions that contradict chemical facts. Accordingly, there is a need for further refinement in ensuring factual correctness.

In the interactive chemical label module, some failures were observed, where GPT-4 produced incorrectly formulated equations. Similar issues were highlighted in previous studies,[19,20] which reported challenges in reaction prediction, with Ref. 20 reporting an accuracy rate of only 58%. Chemistry evaluators also identified minor errors, particularly regarding the states of reactants and products and the omission of reaction conditions, which should ideally be included in the equations. Examples of evaluators' feedback are provided in Appendices A and B. However, some remarks were beyond the scope of the topics covered in the textbook and were not incorporated. These findings emphasize the need to refine the ability of GPT-4 to handle complex chemical reactions accurately while aligning with curriculum-specific requirements.

In the electron configuration module, the GPT-4 responses were also sometimes inaccurate, as shown in the previous table (Table 5). The complexity of electron configurations across elements requires complex rule-based calculations aligned with the Aufbau principle, the Hund's rule, and the Pauli exclusion principle,[42] which are not always accurate with GPT-4 probabilistic model outputs. Therefore, GPT-4 can misinterpret electron configurations, especially for electron distributions based on the "SPDFG" sequence and quantum numbers. The electron distribution can also depend on the shell, subshell, and orbits, as mentioned by one of the evaluators involved in this study (as shown in Appendix A). However, the goal of this module in AIChemApp is to mention the number of electrons in each orbital, which was the basic requirement of the textbook. It was also reported that 5 exceptions were not revealed by GPT-4.

In addition to the programming effectiveness and efficacy of LLMs, the accuracy of different LLMs' responses depends on the complexity of the problem. Furthermore, GPT-4 is considered a general-purpose conversation engine that generates responses on the basis of a broad training dataset rather than dedicated rules and requirements for chemistry topics and learning concepts. Accordingly, on the basis of previously reported short-term data and the drawbacks raised in this research work and other related works,[18,43] it is important to enrich LLM datasets with factual information related to chemistry, with the exceptions to some chemical rules and a list of balanced equations. This highlights the need to provide training datasets with technical knowledge specific to chemistry, which was also raised in Refs. 13 and 32. As an initial attempt in this matter, research studies[25,44–46] presented a dedicated LLM for chemistry and its tasks. Therefore, several additional validation steps are needed to improve response accuracy and reliability not only for chemistry mobile learning applications but also for different specializations.[47]

The accuracy of responses from LLMs also depends on prompt engineering techniques and LLMs' ability to support medium-to-low-resource languages, such as the Arabic language, in multilingual mobile applications. Prompt engineering and prompt design play a significant role in obtaining accurate diagnostic outcomes, as reported in Refs. 15, 30, 33, 34, and 47. For example, a prompt template helps guide LLMs to produce outputs that meet strict formatting or linguistic requirements, which is crucial for delivering structured and meaningful feedback. Despite the use of a customized prompt template in our study, the process mitigated the risk of irrelevant or overly generalized responses.

Another important aspect to be considered is what was reported in previous studies,[32,48] which outlined that ChatGPT provides better responses for English text than other languages do.[48–50] Our study proposed the use of both Arabic and English words in prompt templates where Arabic text is used for values for both *key questions* and *output_format*, whereas English text is used for *context*. This is used to help ChatGPT better process the bilingual structure of the input while maintaining clarity in specifying the module's purpose and its requirements. By separating the languages, the context remains universally interpretable by the model, whereas the question and output requirements align with the target audience's primary language, ensuring both linguistic accuracy and module relevance details. Therefore, prompt engineering has become a tool not only for improving response accuracy but also for expanding inclusivity in multilingual mobile learning applications. Despite these advancements, continued research is needed to refine prompt engineering strategies, especially for underrepresented languages, to enhance their integration into mobile learning platforms. Additionally, the results of this study highlight the ongoing need for further research to address the challenges faced by non-Latin and low-resource languages in mobile learning contexts. These challenges include difficulties in natural language processing, content adaptation, and semantic understanding, which limit the accessibility and effectiveness of AI-based educational tools for diverse linguistic communities.

## 5. Limitations

The content and assessment of the adopted chemistry course are limited to topics covered in the textbook. The number of evaluators in this study was only five. Consequently, further studies investigating the utilization of ChatGPT for more complex topics in the chemistry domain with a larger pool of experts are needed. Additionally, investigating the use of explainable AI (XAI) techniques[51] in mobile learning to offer more details about LLM responses, predictions, and decisions remains an open area for future research.

The LLM adopted in this research work was ChatGPT, although it is not the only tool available for integration with AIChemApp. Alternative tools, such as Gemini and Copilot, are available. Conducting comparative studies among these tools could contribute to academic discourse and help identify the most accurate and effective solutions for chemistry and other STEM topics within the mobile learning domain.

As this research is related to the computer science domain and technical programming perspectives, measuring factors affecting pedagogical aspects and learning outcomes is not covered in this research. Therefore, some instructional design principles and guidelines require further investigation with pedagogical experts. Furthermore, randomized controlled experiments could be conducted to evaluate the effectiveness of AIChemApp.

## 6. Conclusion

This study validates the integration of LLMs such as ChatGPT within mobile learning applications, particularly in the domain of chemistry education. It identifies key requirements and features needed to utilize LLMs in mobile learning applications covering different topics, such as periodic tables, chemical reactions, electron configurations, and game-based assessments. Furthermore, this study extends beyond answering basic questions in chemistry; it provides nuanced answers and feedback related to elements and groups from the periodic table, corrective feedback on balancing chemical equations and electron configurations, and questions and answers for game-based assessments. This study lays the groundwork for future frameworks that utilize AI-driven learning modules, potentially broadening their application to other academic disciplines and courses beyond chemistry.

This study also highlights challenges, such as the accuracy of results related to chemical equations and electron configurations, and the limitations in support for low-resource languages such as Arabic. Importantly, the benefit obtained from the use of LLMs requires a validation process, as most LLMs are general-purpose conversation tools that still require domain-specific datasets. This is due to incorrect results related to balancing equations and electron configurations. This study can address the gap between the utilization of LLMs in mobile learning for chemistry research and inspire future research to explore enhanced integration of LLMs in mobile learning for different science topics.

Future investigations should examine ways to increase engagement through personalized, AI-driven feedback and interactive tasks in mobile learning applications using LLMs. Another future research direction could focus on using LLMs for generating adaptive learning resources and contents tailored to individual user profiles. This replaces static learning resources such as videos embedded in AIChemApp with adaptive learning paths that evolve based on user progress in mastering content and completing learning tasks and assessments successfully.


### Acknowledgments

The author thanks Sirein Jaradat, who was involved in different programming tasks during the initial phase of developing ChemApp. Furthermore, this paper utilized ChatGPT to refine the grammar and textual explanations. All content was reviewed and validated by the author. The author thanks the academic editor and anonymous reviewers for their insightful comments and suggestions.


### Supplementary Material

Appendices can be found in this link: https://github.com/aewais/IJAIT/S0218213025500150


### ORCID

Ahmed Ewais https://orcid.org/0000-0003-0035-9374

## References


1. G. Deeva, D. Bogdanova, E. Serral, M. Snoeck and J. De Weerdt, A review of automated feedback systems for learners: Classification framework, challenges and opportunities, *Computer Education* **162** (2021) 104094, doi:10.1016/j.compedu.2020.104094.
2. I. Molenaar, Towards hybrid human-AI learning technologies, *European Journal of Education* **57**(4) (2022) 632–645, doi:10.1111/ejed.12527.
3. A. Hamzah, A. F. Hidayatullah and A. G. Persada, Discovering trends of mobile learning research using topic modelling approach, *International Journal of Interactive Mobile Technologies* **14**(9) (2020) 11069, doi:10.3991/ijim.v14i09.11069.
4. A. Alshahrani, The impact of ChatGPT on blended learning: Current trends and future research directions, *International Journal of Data and Network Science* **7**(4) (2023) 2029–2040, doi:10.5267/j.ijdns.2023.6.010.
5. L. F. Motiwalla, Mobile learning: A framework and evaluation, *Computer Education* **49**(3) (2007) 581–596, doi:10.1016/j.compedu.2005.10.011.
6. V. Kuleto *et al.*, Exploring opportunities and challenges of artificial intelligence and machine learning in higher education institutions, *Sustainability* **13**(18) (2021) 10424, doi:10.3390/su131810424.
7. J. Wang and Z. Yu, Smart educational learning strategy with the internet of things in higher education system, *International Journal of Artificial Intelligence Tools* **31** (2021) 2140101, doi:10.1142/S0218213021401011.
8. G. Pask, *Conversation Theory: Applications in Education and Epistemology* (Elsevier, Amsterdam, 1976).
9. M. Miraz, M. Ali and P. S. Excell, Cross-cultural usability evaluation of AI-based adaptive user interface for mobile applications, *Acta Scientiarum Technology* **44** (2022) e61112, doi:10.4025/actascitechnol.v44i1.61112.
10. S.-C. Ng *et al.*, An intelligent mobile application for assisting visually impaired in daily consumption based on machine learning with assistive technology, *International Journal of Artificial Intelligence Tools* **30**(01) (2021) 2140002, doi:10.1142/S0218213021400029.
11. P. Kumar, Large language models (LLMs): Survey, technical frameworks, and future challenges, *Artificial Intelligence Review* **57**(10) (2024) 260, doi:10.1007/s10462-024-10888-y.
12. S. Wang, F. Wang, Z. Zhu, J. Wang, T. Tran and Z. Du, Artificial intelligence in education: A systematic literature review, *Expert Systems with Applications* **252** (2024) 124167, doi:10.1016/j.eswa.2024.124167.
13. M. A. Kuhail, N. Alturki, S. Alramlawi and K. Alhejori, Interacting with educational chatbots: A systematic review, *Education and Information Technologies* **28**(1) (2023) 973–1018, doi:10.1007/s10639-022-11177-3.
14. M. J. Parker, C. Anderson, C. Stone and Y. Oh, A large language model approach to educational survey feedback analysis, *International Journal of Artificial Intelligence in Education* **35** (2024) 444–481, doi:10.1007/s40593-024-00414-0.
15. P. Rospigliosi, Artificial intelligence in teaching and learning: What questions should we ask of ChatGPT? *Interactive Learning Environments* **31**(1) (2023) 1–3, doi:10.1080/10494820.2023.2180191.
16. R. P dos Santos, Enhancing chemistry learning with ChatGPT and Bing chat as agents to think with: A comparative case study, *Rochester* (2023) 4447416, doi:10.2139/ssrn.4447416.

17. S. Fergus, M. Botha and M. Ostovar, Evaluating academic answers generated using ChatGPT, *Journal of Chemical Education* **100**(4) (2023) 1672–1675, doi:10.1021/acs.jchemed.3c00087.
18. T. M. Clark, Investigating the use of an artificial intelligence chatbot with general chemistry exam questions, *Journal of Chemical Education* **100**(5) (2023) 1905–1916, doi:10.1021/acs.jchemed.3c00027.
19. T. Guo *et al.*, What can large language models do in chemistry? A comprehensive benchmark on eight tasks, in *Proc. of the 37th Int. Conf. on Neural Information Processing Systems* (Curran Associates Inc., Red Hook, 2024), pp. 59662–59688.
20. C. M. Castro Nascimento and A. S. Pimentel, Do large language models understand chemistry? A conversation with ChatGPT, *Journal of Chemical Information and Modeling* **63**(6) (2023) 1649–1655, doi:10.1021/acs.jcim.3c00285.
21. A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White and P. Schwaller, Augmenting large language models with chemistry tools, *Nature Machine Intelligence* **6**(5) (2024) 525–535, doi:10.1038/s42256-024-00832-8.
22. R. Ejjami, Revolutionizing Moroccan education with AI: A path to customized learning, *International Journal for Multidisciplinary Research* **6**(3) (2024), doi:10.36948/ijfmr.2024.v06i03.19462.
23. P. P. Martin and N. Graulich, Beyond language barriers: Allowing multiple languages in postsecondary chemistry classes through multilingual machine learning, *Journal of Science Education and Technology* **33**(3) (2024) 333–348, doi:10.1007/s10956-023-10087-4.
24. L. Rivadeneira, D. Bellido De Luna and C. Fernandez, Exploring the role of ChatGPT in higher education institutions: Where does Latin America stand? *Digital Government: Research and Practice* **6** (2024) 3689370, doi:10.1145/3689370.
25. D. Zhang *et al.*, ChemLLM: A chemical large language model, arXiv:2402.06852, doi:10.48550/arXiv.2402.06852.
26. A. Ewais, S. Jaradat, K. Rabaya and O. D. Troyer, Usability aspects related to the use of M-Learning in elementary schools in Palestine, *International Journal of Innovative Technology and Exploring Engineering* **9**(2) (2019) 2339–2347, doi:10.35940/ijitee.A4361.129219.
27. A. Ewais, R. Hodrob, M. Maree and S. Jaradat, Mobile learning application for helping pupils in learning chemistry, *International Journal of Interactive Mobile Technologies* **15**(1) (2021) 1865–7923, doi:10.3991/IJIM.V15I01.11897.
28. A. Przegalinska, L. Ciechanowski, A. Stroz, P. Gloor and G. Mazurek, In bot we trust: A new methodology of chatbot performance measures, *Business Horizons* **62**(6) (2019) 785–797, doi:10.1016/j.bushor.2019.08.005.
29. M. N. A. Khan, A. M. Mirza, R. A. Wagan, M. Shahid and I. Saleem, A literature review on software testing techniques for smartphone applications, *Engineering, Technology & Applied Science Research* **10**(6) (2020) 3844, doi:10.48084/etasr.3844.
30. P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi and G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, *ACM Computing Surveys* **55**(9) (2023) 1–35, doi:10.1145/3560815.
31. J. White *et al.*, A prompt pattern catalog to enhance prompt engineering with ChatGPT, arXiv:2302.11382. Available at: http://arxiv.org/abs/2302.11382
32. F. Alam, S. A. Chowdhury, S. Boughorbel and M. Hasanain, LLMs for low resource languages in multilingual, multimodal and dialectal settings, in *Proc. of the 18th Conf. of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, eds. M. Mesgar and S. Loáiciga (Association for Computational Linguistics,

St. Julian's, Malta, 2024), pp. 27–33. Available at: https://aclanthology.org/2024.eacl-tutorials.5
33. A. D. White *et al.*, Assessment of chemistry knowledge in large language models that generate code, *Digital Discovery* **2**(2) (2023) 368–376, doi:10.1039/D2DD00087C.
34. L. Reynolds and K. McDonell, Prompt programming for large language models: Beyond the few-shot paradigm, in *Extended Abstracts of the 2021 CHI Conf. on Human Factors in Computing Systems* (Association for Computing Machinery, New York, 2021), pp. 1–7, doi:10.1145/3411763.3451760.
35. R. Bommasani, P. Liang and T. Lee, Holistic evaluation of language models, *Annals of the New York Academy of Sciences* **1525**(1) (2023) 140–146, doi:10.1111/nyas.15007.
36. Y. Chang *et al.*, A survey on evaluation of large language models, *ACM Transactions on Intelligent Systems and Technology* **15**(3) (2024) 1–39, doi:10.1145/3641289.
37. S. Jalil, S. Rafi, T. D. LaToza, K. Moran and W. Lam, ChatGPT and software testing education: Promises & perils, in *2023 IEEE Int. Conf. on Software Testing, Verification and Validation Workshops (ICSTW)* (IEEE, 2023), pp. 4130–4137, doi:10.1109/ICSTW58534.2023.00078.
38. M. A. Al-Sharafi, M. Al-Emran, M. Iranmanesh, N. Al-Qaysi, N. A. Iahad and I. Arpaci, Understanding the impact of knowledge management factors on the sustainable use of AI-based chatbots for educational purposes using a hybrid SEM-ANN approach, *Interactive Learning Environments* **31**(10) (2023) 7491–7510, doi:10.1080/10494820.2022.2075014.
39. I. Harel and S. Papert, Software design as a learning environment, *Interactive Learning Environments* **1**(1) (1990) 1–32, doi:10.1080/1049482900010102.
40. C. Wang *et al.*, Evaluating open-QA evaluation, in *Proc. of the 37th Int. Conf. on Neural Information Processing Systems* (Curran Associates Inc., Red Hook, 2024), pp. 77013–77042.
41. S. Sadeghi, A. Bui, A. Forooghi, J. Lu and A. Ngom, Can large language models understand molecules? *BMC Bioinformatics* **25**(1) (2024) 225, doi:10.1186/s12859-024-05847-x.
42. C. Boyce, Using logic to define the Aufbau–Hund–Pauli relation: A guide to teaching orbitals as a single, natural, unfragmented rule-set, *Foundations of Chemistry* **16**(2) (2014) 93–106, doi:10.1007/s10698-012-9176-7.
43. A. Tlili *et al.*, What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education, *Smart Learning Environments* **10**(1) (2023) 15, doi:10.1186/s40561-023-00237-x.
44. W. Jin, C. Coley, R. Barzilay and T. Jaakkola, Predicting organic reaction outcomes with Weisfeiler–Lehman network, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2017). Available at: https://proceedings.neurips.cc/paper_files/paper/2017/hash/ced556cd9f9c0c8315cfbe0744a3baf0-Abstract.html
45. D. Tran, L. Pascazio, J. Akroyd, S. Mosbach and M. Kraft, Leveraging text-to-text pretrained language models for question answering in chemistry, *ACS Omega* **9**(12) (2024) 13883–13896, doi:10.1021/acsomega.3c08842.
46. N. C. Frey *et al.*, Neural scaling of deep chemical models, *Nature Machine Intelligence* **5**(11) (2023) 1297–1305, doi:10.1038/s42256-023-00740-3.
47. A. Shoufan, Exploring students' perceptions of ChatGPT: Thematic analysis and follow-up survey, *IEEE Access* **11** (2023) 38805–38818, doi:10.1109/ACCESS.2023.3268224.
48. Y. Bang *et al.*, A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity, in *Proc. of the 13th Int. Joint Conf. on Natural*

*Language Processing and the 3rd Conf. of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, eds. J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti and A. A. Krisnadhi (Association for Computational Linguistics, Nusa Dua, Bali, 2023), pp. 675–718, doi:10.18653/v1/2023.ijcnlp-main.45.

49. B. Y. Lin, S. Lee, X. Qiao and X. Ren, Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning, in *Proc. of the* 59*th Annual Meeting of the Association for Computational Linguistics and the* 11*th Int. Joint Conf. on Natural Language Processing* (*Volume* 1*: Long Papers*), eds. C. Zong, F. Xia, W. Li and R. Navigli (Association for Computational Linguistics, 2021), pp. 1274–1287, doi:10.18653/v1/2021.acl-long.102.
50. A. Abdelali *et al.*, LAraBench: Benchmarking Arabic AI with large language models, in *Proc. of the 18th Conf. of the European Chapter of the Association for Computational Linguistics* (*Volume* 1*: Long Papers*), eds. Y. Graham and M. Purver (Association for Computational Linguistics, St. Julian's, Malta, 2024), pp. 487–520. Available at: https://aclanthology.org/2024.eacl-long.30
51. A. Abusitta, M. Q. Li and B. C. M. Fung, Survey on explainable AI: Techniques, challenges and open issues, *Expert Systems with Applications* **255** (2024) 124710, doi:10.1016/j.eswa.2024.124710.