Evaluation of LLMs Capacity to Assist in Chemistry Laboratory Courses

Sanidhya Pal[1], Anup Paul[2]*

[1] Student, Department of Computer Science and Engineering, HMR Institute of Technology and Management, Delhi - 110036, India

[2] Assistant Professor, Department of Applied Science, HMR Institute of Technology and Management, Delhi - 110036, India

* Corresponding author

doi: https://doi.org/10.21467/proceedings.178.23

ABSTRACT

Large language models (LLMs) have emerged as exceptional assistants across many domains of education. However, their efficacy in assisting with chemistry laboratory courses still needs to be explored. The present study will evaluate the capacities of several LLMs, including Claude, Gemini, ChatGPT, and others, to perform tasks relevant to chemistry education through a series of queries. The result demonstrated LLMs excel at retrieving complex information, generating well-structured experimental manuals, clarifying chemical concepts with exclusive insights, and predicting reactive sites of a given compound. However, LLMs still needed to process the chemical structure representation more successfully and identify compounds by their IUPAC names. The challenge in these domains emerge from LLMs' query reformulation and the rendering of partial response methods used to produce a response to any question. To overcome these challenges, the present study suggests integrating LLMs with cheminformatics toolkits like RDKit and providing access to standard chemical databases. By strategically combining LLMs' natural language capabilities with specialized chemical data processing functionalities, the role of LLMs in enhancing chemistry learning experiences can be fully harnessed. Therefore, the demonstrations of the present study will lay the groundwork for the future development of more effective AI-powered education tools.

Keywords: Large Language Models (LLMs), Chemistry Laboratory Courses, IUPAC Nomenclature, PubChem Database, Chemical Properties, Knowledge Retrieval

1 Introduction

The integration of artificial intelligence (AI) in educational settings have enhanced student learning experiences and also improved teaching practices [1]. Several literature highlighted the role of Large Language Models (LLMs), a form of AI technologies capable of processing and analyzing large amounts of natural language data. LLMs can perform multiple tasks, including classifying natural language, translating text, and completing input strings of text. These LLMs have the potential to benefit students of all ages in numerous ways. Some of the predicted applications are personalized learning experiences, intelligent tutoring assistance, student performance assessment, educational content generation, personalized lesson plan designing, enhanced student engagement, equal distribution of teaching resources, educational data analysis for research, problem-solving assistance, customizable and mapped learning paths, and advanced assessment capabilities [1]. Over the years many effort were made to examine success of LLMs in different perspectives of education. Any challenges existing in LLMs and the future scope were rigorously survey [2]. Many of these surveys posed a few concern relating to increase use of LLMs in education. Ethics, Plagiarism, and academic integrity are some

Proceedings DOI: 10.21467/proceedings.178; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-984081-8-1

major concerns reported in literature that documents the integration of LLMs in education [3]. In addition these, few literature reported about lack of domain expertise, unreliable results, inconsistent or nonsensical answer [4]. While another literature identified effectiveness of the technology were untested and identified an existence of limitation in data quality [5]. Therefore a demand for consistent evaluation of efficacy and scope of LLMs integration in education developed.

Educational assessment typically evaluates student learning across three primary domains. Cognitive outcomes encompass knowledge acquisition and intellectual skills. Affective outcomes pertain to emotional and attitudinal aspects. Psychomotor outcomes involve physical and motor skills. Chemical structure provides valuable insight and, therefore, serves as a comprehensive tool to assess students' knowledge and intellectual skills in chemistry and allows educators to evaluate cognitive outcomes effectively [6]. A chemical structure details the number and types of atoms in a molecule, the types of bonds among the atoms and their spatial arrangements. The chemical properties of the compound can easily predict the molecular structure of the compound [7]. The performance of LLMs largely depends on their capacity to understand molecular structure. However, the foundation of LLMs in natural language processing restricts the users from interacting with them via images of chemical structures. Since molecular formulas, IUPAC names, and line notations, specifically SMILES and InChI notation, facilitate the conversion of chemical structures into ASCII text, Tlais and coworkers have recently explored these text types in input queries to LLMs. These queries include identifying condensed structures, multiple bonds, and functional groups. All LLMs partially succeeded in generating correct answers, with the best results obtained in determining unsaturation from molecular formula and identification of functional groups from condensed structure reported [8]. Recently, Xuan Dao et al. evaluated the performance of LLMs with questions of higher chemical complexities. These questions include identifying the compounds with specific chemical properties, predicting the product of the chemical reaction, and performing the calculations. Among selected LLMs, Bard exceeds in solving all the questions. These indicated LLMs do possess intelligence toward a few aspects of chemistry [9].

Specialized AI models, AI Chatbots and LLMs such as Chemical LLMs were designed to meet the requirements of chemistry, which generally use SMILES string for input representation of a compound. However these models covers limited number of domains, require a large amount of resource and needs specialized training for users to interact [10]. Despite undergoing rapid development, specialized models for chemistry have not been optimized for non-expert users, resulting in interfaces that are often cumbersome and difficult for laypersons to utilize effectively.

The validation of theoretical knowledge with practical experiments is an essential part of chemistry education. Hands-on experiments with compounds help students understand how compounds are made and their properties. This approach enhances effective learning and increases practical application [11]. Therefore, present study will explore potentials of LLMs in delivering accurate and relevant chemical information, explicitly examining their capacity to provide comprehensive explanations of chemistry experiments and compound characteristics. The outcomes of the present study will support in the development of innovative teaching strategies to enhance the effectiveness of LLMs in chemistry education and improve the overall learning experience for students.

2 Research Methodology

Present study employed a comprehensive methodology to evaluate eight Large Language Models (LLMs) for their potential of being assistant in Chemistry Laboratory Courses. All LLMs were accessed through Google Chrome. Then their performance were systematically assessed with seven distinct

Proceedings DOI: 10.21467/proceedings.178
ISBN: 978-81-984081-8-1

Series: AIJR Proceedings
ISSN: 2582-3922

query categories. These categories ranged from basic compound name retrieval to generating detailed chemistry experiment manuals and providing advanced chemical information. Responses from the LLMs were collected and analysed using Microsoft Word and Excel. The responses were assessed on the basis of completeness, relevance, accuracy, and presentation format. Additionally, queries were categorized by difficulty and importance, allowing for a structured comparison of LLM performance across different chemistry-related tasks. Standard data for compounds were sourced from PubChem database to ensure the reliability of the information. The following section provided the details of methodology employed in present study.

2.1 Selection of Large Language Models

At the time of present study, there are only eight comprehensive and free-to-access LLMs available. Among these, three LLMs, namely Google Gemini, ChatGPT, and Bing Copilot, were recently evaluated for their contribution to chemical education [9]. In present study, Google Gemini was referred to as Google Bard, and Bing Copilot was referred to as Bing Chat. Quora POE assistant, as an LLM model, gained popularity due to its precisely structured responses, also supported by a large number of specialized versions of AI chatbox [12]. The other four, less commonly known LLMs evaluated in present study, are Claude, Perplexity AI, PI, and YouChat. Table 1 provides the list of LLMs evaluated in present study[13, 14].

Table 1*: List of LLMs were evaluated in the study*

Name of LLMs	**Developer**	**Version**
Gemini	Google	1.5
POE (Assistant)	Quora	Powered by gpt-3.5-turbo & Claude 3 Haiku.
ChatGPT	OpenAI	ChatGPT 3.5
Bing Copilot	Microsoft	Free Version
Perplexity	Andy Konwinski, Denis Yarats, Johnny Ho, and Aravind Srinivas (CEO)	1.0.21
PI	Inflection AI	(Not Disclosed)
YouChat	You.com	2.0
Claude	Anthropic	3

2.2 Additional software

All the LLM models were accessed through the Google Chrome application (Version 124.0.6367.119 (Official Build) (64-bit)) application [15]. The responses were collected and tabulated in Microsoft Word 2019 (Student Version) [16]. To analyze the performance of LLMs, Microsoft Excel 2019 (Student Version) was used. Standard data on the compounds were obtained from the online database PubChem (URL: https://pubchem.ncbi.nlm.nih.gov/) [17]. ChemDraw (Version 21.0.0.28, PerkinElmer) was used to construct the structures of the compounds [18].

2.3 Query Design

A set of queries associated with chemistry laboratory courses was prepared. Based on the level of difficulty and importance, the queries were classified into seven categories. For each query, the involved questions were input into the LLMs evaluated in present study. The responses were collected and tabulated in a Microsoft Word document. All the responses made by the LLMs were assessed, and a score was assigned based on the degree of completeness, relevance, accuracy, and format of presentation. The following set of queries was evaluated in present study.

Query One: LLM's Capacity to Retrieve Compounds Name

The first query set in present study was designed to evaluate the capacity of LLMs to retrieve compound names. In addition to this, the ability of LLMs to predict the IUPAC name of these compounds was tested. Query One was: "Write 10 common compounds and their IUPAC names." For each LLM, the input of Query One was performed twice to study the possibilities of repetition of responses [19]. The output was evaluated for four variables: first, the distinctiveness among the retrieved compounds; second, accuracy in IUPAC nomenclature; third, the presentation structure displayed in the output; and finally, the inclusion of the fundamental compound, water [20].

Query Two: LLM's Capacity to Generate IUPAC Name of Water

Water is the most fundamental compound in chemistry. The IUPAC name for water is "oxidane," although it is an uncommon name. Therefore, it can be easily used to test the deep understanding of any intelligent being [21]. Query Two was: "Write the IUPAC nomenclature of water." Since LLMs responded with more than one approved name for water, all the names were documented. A Python-based data analysis was performed to group the sampled LLMs based on the provided names.

Query Three: LLM's Capacity to Generate Manual of a Chemistry Experiment

All the LLMs' capacity to produce a full report for a basic laboratory experiment in chemistry was tested[22]. A simple experiment about the standardization of an NaOH solution with a standard oxalic acid solution was selected for the test. Query Three was: "Write an experiment manual for the standardization of an NaOH solution with 0.1 N oxalic acid solution." The structure of the manuals generated was analyzed based on the following criteria: aim, theory, materials required, procedure, calculation, disposal of chemicals, and safety precautions.

Query Four: LLMs Capacity to Retrieve Acid-Base Subject-related Information

Query Four evaluated the knowledge-retrieving capacity of LLMs with two basic questions [23]. The first query was: "Define acid and base." This query aimed to assess the possibility of all LLMs producing similar results. The second query was: "Write different theories related to acids and bases?" This query explored the capacity of LLMs to search for and acquire information related to acid-base theories. Additionally, a directed query toward the Wikipedia website was entered: "Write some lesser-known alternative theories of acids and bases, with support from Wikipedia website access?" This query tested the improvement in LLMs' responses [24].

Query Five: LLMs Capacity to Provide Identification Details of Phenolphthalein

Since the name of the compound "Phenolphthalein" was provided in the experiment manual, it was considered that details of phenolphthalein were preloaded in the memory of LLMs. The present study, therefore, requested the details of phenolphthalein from LLMs. Query Five included the following: "Find the CID of Phenolphthalein from the PubChem Database (URL: https://pubchem.ncbi.nlm.nih.gov/)," "Find the SMILES and InChI of Phenolphthalein from the

Proceedings DOI: 10.21467/proceedings.178
ISBN: 978-81-984081-8-1

Series: AIJR Proceedings
ISSN: 2582-3922

PubChem Database (URL: https://pubchem.ncbi.nlm.nih.gov/)," and "Find the IUPAC name of Phenolphthalein from the PubChem Database (URL: https://pubchem.ncbi.nlm.nih.gov/)." For control, all of the above queries were requested again, but with the name of the standard database and URL provided [17].

In addition to these, the capacity of LLMs to convert IUPAC names into common names was tested with the query: "Find the common name of '3,3-bis(4-hydroxyphenyl)-2-benzofuran-1-one' from the PubChem Database (URL: https://pubchem.ncbi.nlm.nih.gov/)." A control query without the name of the "PubChem" database and URL was also requested. Furthermore, a test was performed to check the LLMs' capacity to draw the structure of phenolphthalein with the query: "Construct the chemical structure of phenolphthalein."[17, 25]

Query Six: LLMs Foundation in Chemistry, such as for Phenolphthalein

Since advanced knowledge of compounds is a necessity for chemistry education assistants, the extent of LLMs' foundation in chemistry was tested with Query Six [26]. Query Six included the following: "Write the role of phenolphthalein," "Write the chemical equations for the synthesis of phenolphthalein," "Write the reactive sites of phenolphthalein," and "Write the chemical equations for 10 chemical properties of phenolphthalein." Finally, to test the accessibility of the generative functions of LLMs in chemistry education, the following query was used: "Write the color of '3,3-bis(4-hydroxyphenyl)-2-benzofuran-1-one' in a basic medium."

Query Seven: Expansion of Avenue of Chemistry to evaluate performance of LLMs

All the above tests were limited to only one acid-base indicator, phenolphthalein. It resulted in the evaluation of the above queries on LLMs appearing as hit-and-trial. It was further supported by the fact that on several occasions, the responses of LLMs did not match twice. Therefore, the perfect way to evaluate the foundation of LLMs is through the application of statistics. This was done with the input of Query Seven: "Write the names of ten chemical properties of acid-base indicators." For this purpose, 18 different acid-base indicators were selected [27]. For each indicator, a query was developed and run over all the evaluating LLMs. The output was analyzed for the amount of response, the information it contained, the chemistry-related information, the accuracy of the response to the query for chemical properties, and the content of the generative response.

3 Results

The most significant feature of Large Language Models (LLMs) is their capacity to understand the input question, analyze it, and generate an answer in a structured presentation [28]. For complex and uncommon queries, LLMs use the method of tokenization of the input string, which puts a restriction on conventional methods of human-computer interaction, especially in the area of chemistry education[8]. The present study explored the possibility of skipping common input methods and finding alternative routes for entering chemical education queries into LLMs. The following section discusses the output of different queries and the landmarks achieved by LLMs.

3.1 Retrieving Name of Ten Common Compounds

The retrieval of the names of common compounds was the first and basic task given to the LLMs. All the LLMs evaluated in the present study successfully provided the names of ten common compounds on their first attempt. However, on their second attempt, a few LLMs could not retrieve the names of new compounds.

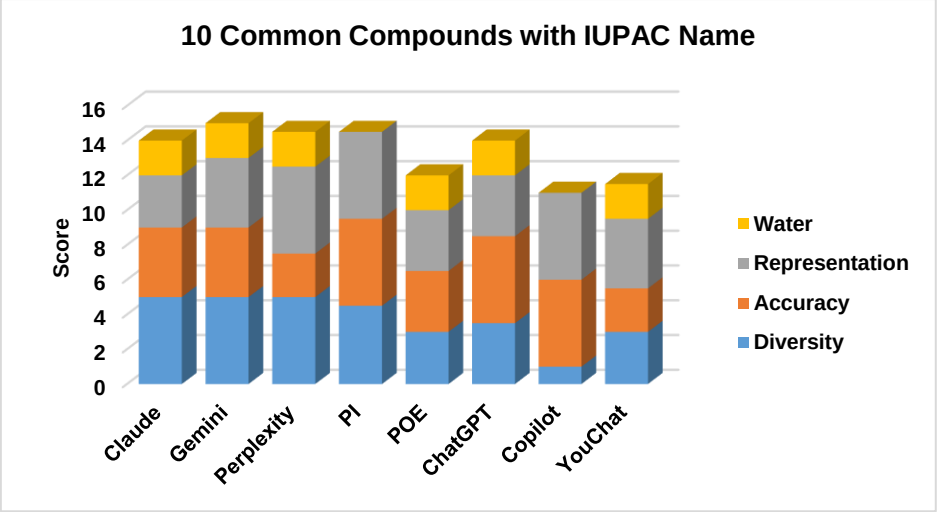

Figure 1: The performance of different LLMs to provide names of common compounds.

Figure 1 shows a bar plot of the scores each LLM obtained in the categories of diversity of retrieved compound names, accuracy of IUPAC names of the provided compounds, representation of the responses, and identification of water as a compound. As seen in the plot, the Gemini LLM displayed the highest capacity to retrieve distinctive compounds, achieved maximum accuracy of IUPAC names of the compounds, presented responses brilliantly, and successfully identified water as a compound.

3.2 Generation IUPAC Name of Water

The IUPAC name of water is "Oxidane." Although the initial study in Query One showed that many LLMs provided incorrect IUPAC names for water, all LLMs generated the correct response to the exclusive query on the IUPAC name of water. Figure 2 shows the IUPAC and additional names of water generated by the LLMs.

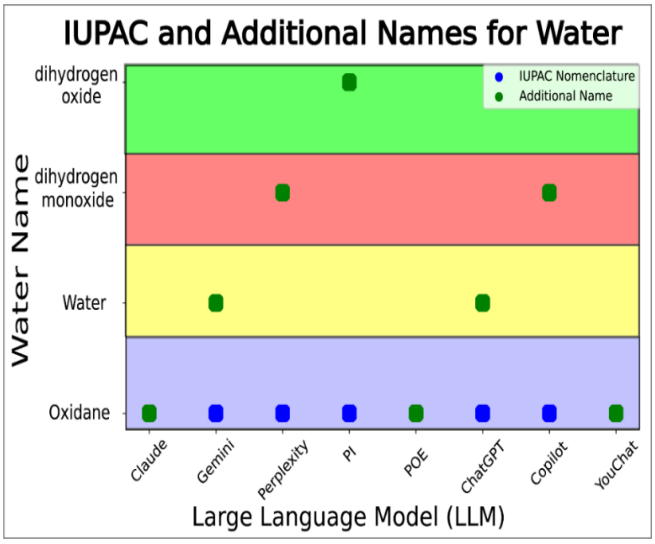

Figure 2: The IUPAC and Additional name of water provided by LLMs

3.3 Generation of Chemistry Experiment Manual

All the LLMs responded brilliantly to the input query: "Write an Experiment Manual for Standardization of NaOH solution with 0.1 N Oxalic Acid solution." The manuals generated from all the LLMs were compiled into a dataset and evaluated.

Proceedings DOI: 10.21467/proceedings.178
ISBN: 978-81-984081-8-1

Series: AIJR Proceedings
ISSN: 2582-3922

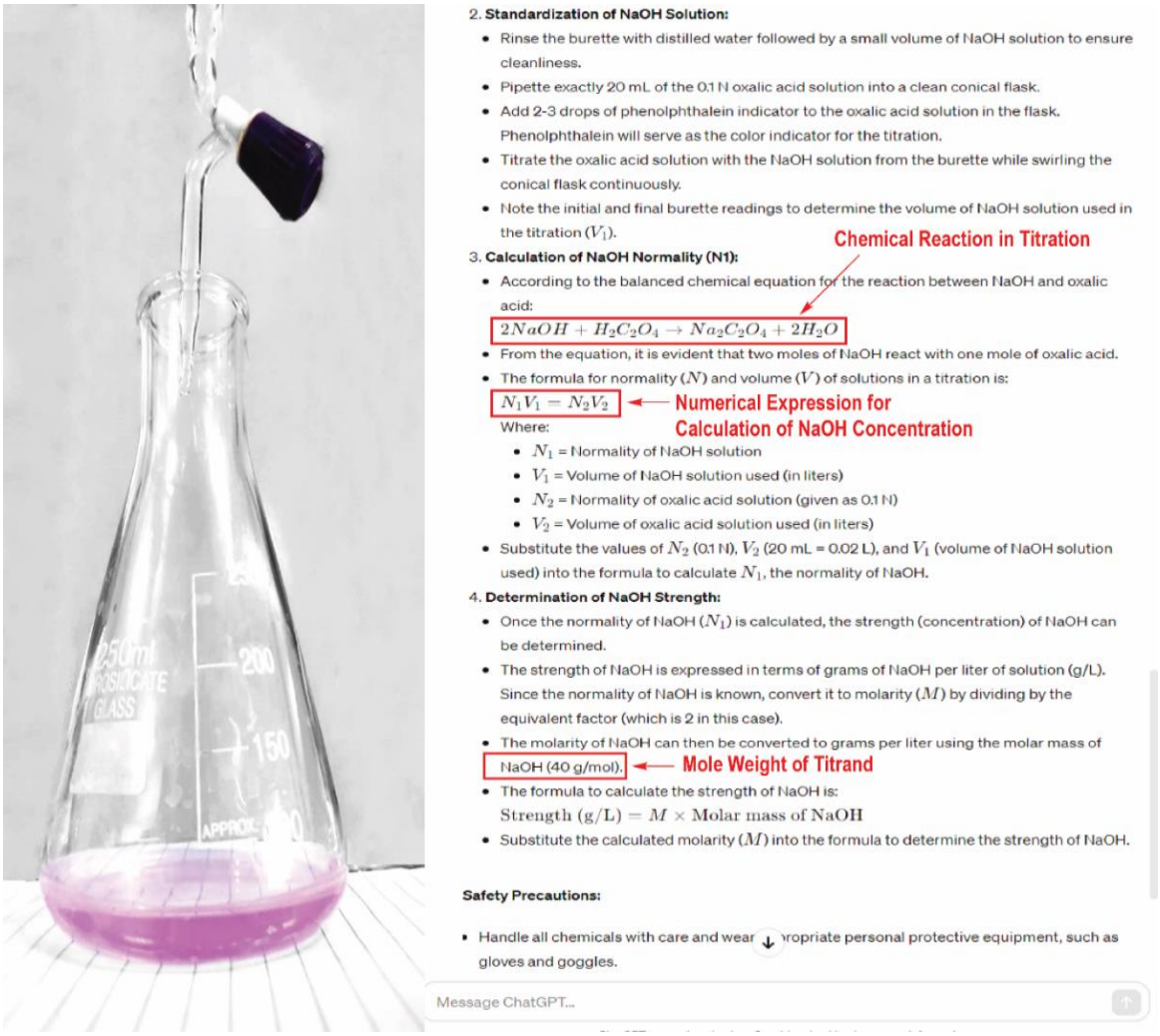

Figure 3: Format of experiment manual provided by LLMs

The manual generated by each LLM contained some unique qualities; for instance, YouChat produced a manual that contained a summary of the experiment, while POE provided a non-structured manual. Six out of the eight LLMs evaluated in present study produced manuals that satisfied the criteria to be a perfect manual for an undergraduate chemistry laboratory course [29]. Figure 3 displays the laboratory manual generated by ChatGPT LLM. Noticeably, the figure shows the capacity to provide the chemical reaction involved in the titration, accurately showing the numerical expression involved in determining the NaOH concentration and the molar weight of NaOH.

3.4 Retrieval of Acid-Base Subject-related Information

All the LLMs provided the standard definition of acids as donors and bases as acceptors of hydrogen ions. Additionally, PI and POE LLMs defined acids as acceptors of a pair of electrons, and YouChat defined bases as hydroxide ion producers. Then, a total of 18 theories explaining the concept of acid-base were retrieved from all the LLMs. Among the major theories, the Arrhenius theory, Brønsted-Lowry theory, and Lewis concept of acid-base were consistently present in all LLMs' responses. However, a large variation in LLMs' responses was observed upon queries for lesser-known alternative acid-base theories, with or without providing the name of the encyclopedia "Wikipedia" webpage for reference. Figure 4 shows the exclusive numbers of theories provided by each LLM.

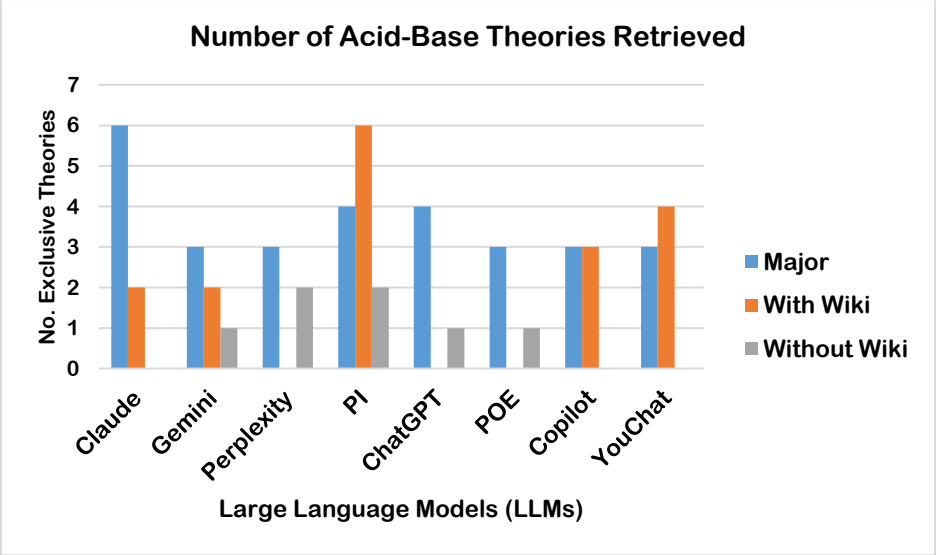

Figure 4: Total number Acid-Base theories provided by LLMs

The blue, orange, and grey bars in Figure 4 show the number of major theories and alternative theories (with or without support from the Wikipedia encyclopedia) that appeared in different LLMs. Both Claude and PI LLMs scored maximum in this test. Furthermore, it was found that some LLMs, when directed to Wikipedia websites, were able to retrieve a higher number of lesser-known theories. However, some LLMs were also found to deny responding to such queries.

3.5 LLMs Capacity to Obtain Identification Details of Phenolphthalein

Table 2 provides the validity of LLMs' responses to the query for phenolphthalein's details, retrieved from the Chemical Database website PubChem. The most common LLM response to queries containing suggestions for the website "PubChem" was: "Being an AI assistant, it does not have the capability to browse the internet or open URLs." The remaining responses were either incorrect or did not match the standard result data provided on the website. This indicates that the data used by LLMs for modeling were unable to access the websites provided in the queries.

Table 2*: Validity of LLMs responses toward queries of Phenolphthalein's details from Online PubChem Database*

LLM	PubChem CID	SMILES	InChI	IUPAC	Common Name
Claude	Access Denied	Access Denied	Access Denied	Access Denied	Access denied
Gemini	Access Denied	Access Denied	Access Denied	Access Denied	Correct
Perplexity	False	False	False	Correct	Correct
PI	Correct	Correct	Correct	Correct	Correct
ChatGPT	Access Denied	Access Denied	Access Denied	Access denied	Access Denied
POE	Access Denied	Access Denied	Correct	Access Denied	Access Denied
Copilot	Correct	Access Denied	Correct	Correct	Correct
YouChat	Correct	False	Correct	Correct	False

Proceedings DOI: 10.21467/proceedings.178
ISBN: 978-81-984081-8-1

Series: AIJR Proceedings
ISSN: 2582-3922

Figure 5 provides the structure of phenolphthalein as generated by different LLMs. Gemini denied generating the chemical structure of phenolphthalein, while Copilot and Perplexity LLMs retrieved the structure from the internet. The most noticeable aspect was the self-generation of the phenolphthalein structure, which was displayed on a prompt-like interface. Further analysis found that ChatGPT showed the name "Mathematica" at the top of the prompt.

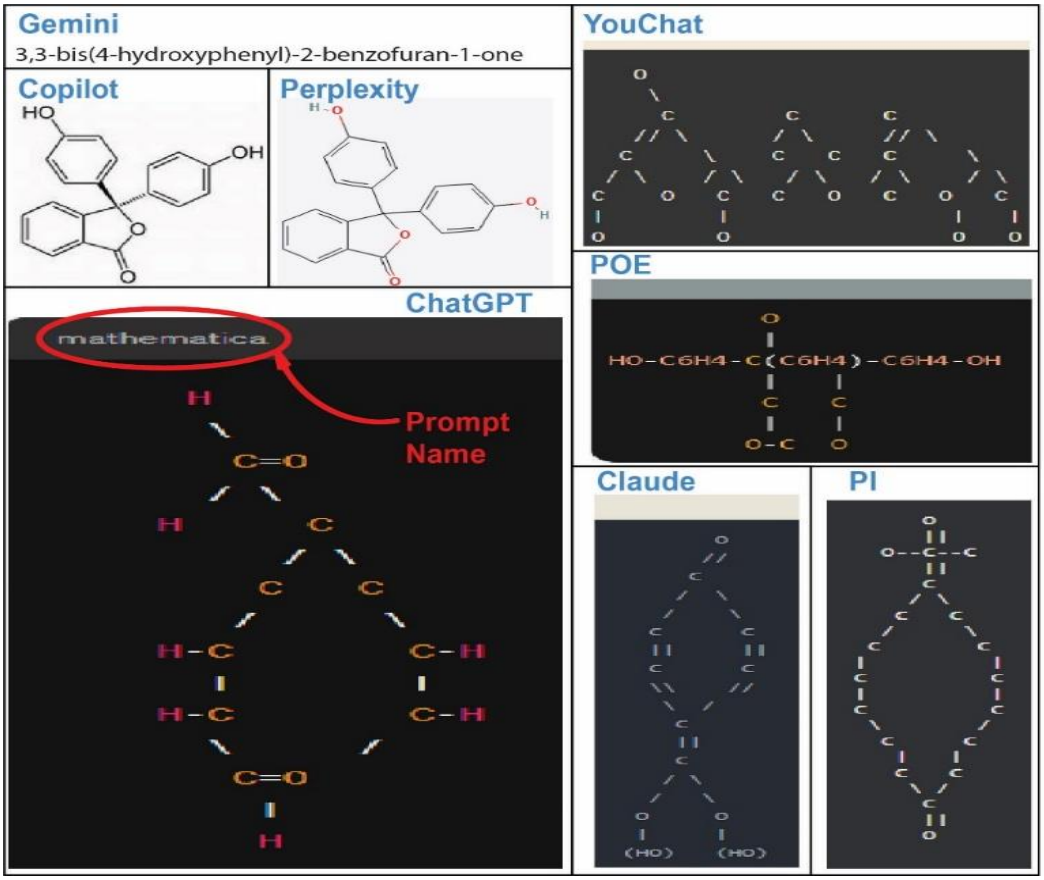

Figure 5: The structure of Phenolphthalein provided by different LLMs

3.6 Testing LLMs Foundation in Chemistry

Analyzing the LLMs' responses dataset toward queries for the "role of Phenolphthalein," it was found that all LLMs replied with pH indicator as the major function of phenolphthalein in acid-base titration. Other roles of phenolphthalein mentioned were as a "laxative" and in tests for blood. These additional responses showed the LLMs' capabilities to acquire and survey academic data.

Phenolphthalein can be synthesized from the reaction of phthalic anhydride and phenol in the presence of a strong acid catalyst like H2SO4. The method of phenolphthalein synthesis follows an electrophilic aromatic substitution (SEAr) reaction. This basic method of phenolphthalein synthesis was mentioned by all the LLMs evaluated in present study. Along with the description, LLMs also displayed the chemical reaction involved in the synthesis. It included symbols of reactants, i.e., phenol (C6H5OH) and phthalic anhydride (C8H4O3), and the products, phenolphthalein (C20H14O4) and water involved in the reaction. Additionally, POE and ChatGPT provided information regarding the formation of intermediates. Furthermore, many LLMs, such as Copilot, Perplexity, and ChatGPT, also detailed the

presence of H2SO4 over the arrows of the chemical reaction of phenolphthalein synthesis. Some LLMs provided an additional method of synthesizing phenolphthalein using zinc chloride.

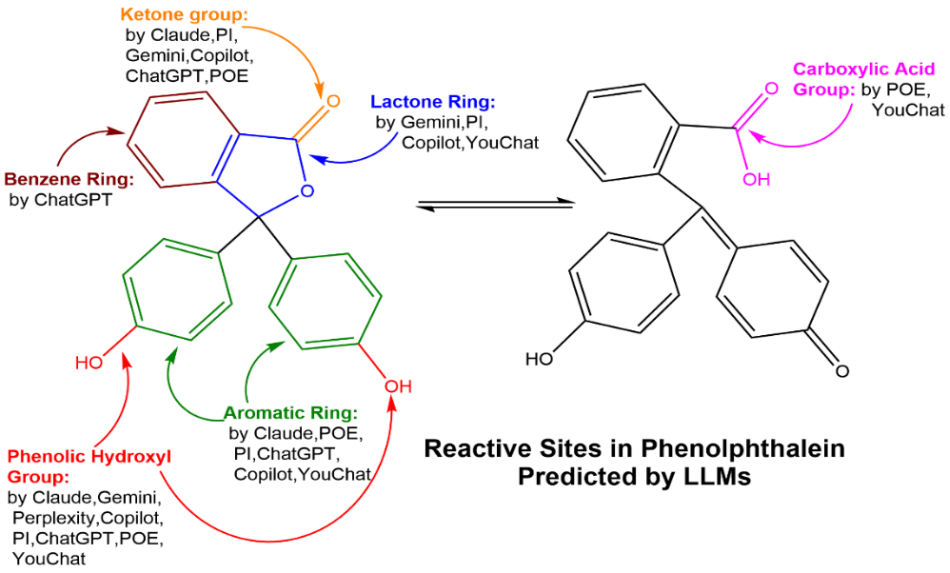

Figure 6: The reactive sites in phenolphthalein predicted by different LLMs

Figure 6 shows the reactive sites in phenolphthalein identified by different LLMs. These reactive sites include the phenolic hydroxyl group, ketone group, aromatic ring in the phenoxyl group, benzene ring at the phthalic anhydride group, lactone ring, and carboxylic acid group. The phenolic hydroxyl group was identified as the major reactive site among all LLMs. It is involved in acid-base reactions and esterification. This reactive site is responsible for the formation of pink coloration with a change in pH. The reactivity of the ketone group was predicted by Gemini, Copilot, and YouChat LLMs. This site is involved in nucleophilic addition reactions. The involvement of the aromatic ring in the phenolic group in Friedel-Crafts, nitration, and sulfonation reactions was predicted by several LLMs, such as POE and Claude. A similar reaction for the benzene ring of the phthalic group was predicted by many of the above LLMs. The reactivity of the carbon center in the lactone ring in the presence of a strong base was predicted by Copilot. Additionally, the reactivity of the carboxylic acid group in the quinoid form of phenolphthalein was predicted by ChatGPT. It is involved in the formation of salts, esterification, and decarboxylation under specific conditions.

All the LLMs evaluated in present study were successful in generating information regarding 10 chemical properties of phenolphthalein, along with the chemical reactions involved. However, there existed a wide variation in responses among the LLMs' selected chemical properties. Some LLMs responded by providing properties related to acid-base titrations and pH changes. Others generated chemical properties like reduction, oxidation, esterification, metal-complex formation, and so on. One category of LLMs provided physical properties like melting point, boiling point, molecular weight, and so on. Through surveying literature, the maximum number of Copilot responses were cited. In contrast, ChatGPT's responses appeared predictive or newly generated. Therefore, these responses were identified as generative responses. Figure 7 compares the variability versus generative score among different LLMs' responses.

Proceedings DOI: 10.21467/proceedings.178
ISBN: 978-81-984081-8-1

Series: AIJR Proceedings
ISSN: 2582-3922

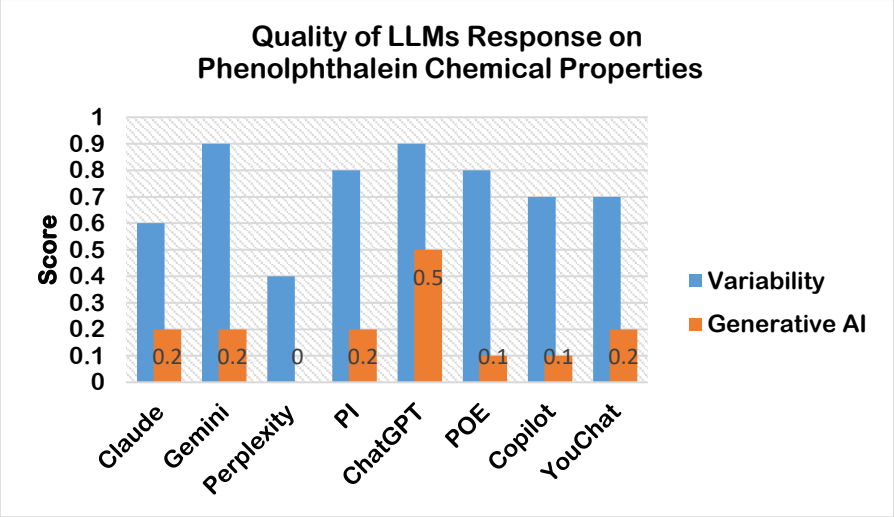

Figure 7: The overall quality of response observed for phenolphthalein properties related queries.

Table 3: The response list of LLMs for color of '3,3-bis(4-hydroxyphenyl)-2-benzofuran-1-one' in basic medium

Standard	**Claude**	**Gemini**	**Perplexity**	**PI**	**ChatGPT**	**POE**	**Copilot**	**YouChat**
Pink	Yellow	Pink	Pink	Pink	Yellow	Yellow	Pink	Yellow

Table 3, shown below, lists the color predicted by the LLMs for '3,3-bis(4-hydroxyphenyl)-2-benzofuran-1-one' in a basic medium. Fifty percent of the LLMs were unable to provide the correct result. This showed the failure of LLMs in identifying the compound from its IUPAC name.

3.7 Performance of LLMs Response after Expanding the Avenue of Chemistry

Upon expanding the number of acid-base indicators, the level of LLMs' responses did not show any significant changes. All the LLMs responded successfully to the names. The chemical properties retrieved by LLMs were concise and well-structured. However, the content varied to a large extent. Claude appeared to be the most efficient in producing "to-the-point" responses. However, it also had limited accuracy in retrieving chemical properties. In contrast, other LLMs included details of applications, synthesis, uses, safety precautions, and physical properties of the compounds. Figure 8 provides a box-whisker plot to show the amount and type of content in an LLM's response to any chemical property.

For typical queries made relating to chemical properties of a compound, a 100% response rate was obtained from LLMs. The responses also contained nearly 100% useful information. However, the amount of chemistry-related information decreased to 90%, of which only 75% of the information was related to chemical properties. Approximately 70% of the required information was correct and matched the literature. Additionally, 10% of the information was generative, meaning it had no basis in literature. Several cases were observed where the information about chemical properties of a compound provided by LLMs was incorrect.

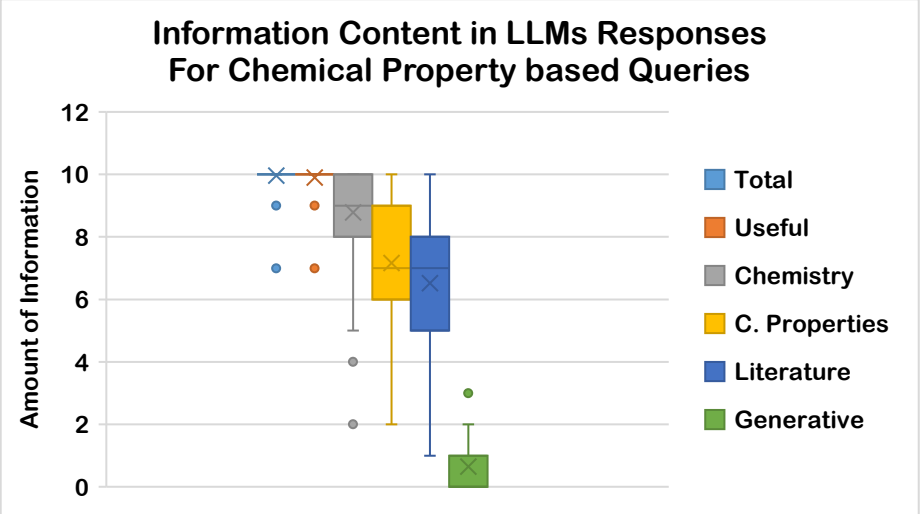

Figure 8: The type of content and its amount in LLMs response for chemistry queries

4 Discussion

The evaluation of large language models (LLMs) exposed a critical limitation in their ability to identify chemical compounds' IUPAC names accurately. Notably, LLMs failed to correctly identify the color of phenolphthalein in a basic medium, despite being provided with its IUPAC name. This shortcoming is a direct consequence of LLMs' reliance on extensive text resources for unsupervised learning, which compromises their real-time response capacity due to time-consuming query processing [30, 31].

The response system of Large Language Models (LLMs) is grounded in a set of guiding principles [31]. For common compound names, LLMs utilize cached responses to bypass model inference. In contrast, uncommon queries are addressed through query reformulation and similarity matching, which leverages cached responses and precomputed results to expedite response times [32]. Alternatively, LLMs employ partial responses and progressive rendering, generating and presenting the relevant response portions while processing remaining query components [33]. However, these approaches pose risks when processing chemical names, particularly IUPAC names, which contain repetitive symbols, syntaxes, prefixes, suffixes, and words that compromise distinguishability. Consequently, the applicability of query reformulation and partial response methods is restricted, necessitating alternative strategies for accurate chemical name processing.

A significant outcome of the present study is the development and accessing of preloaded compound data in the LLM's memory, bypassing the query processing system through a strategic sequence of queries[34]. This was evident when all LLMs successfully predicted the reactive sites of phenolphthalein, demonstrating their generative capacity in solving chemistry-related queries. Notably, this finding contradicts earlier studies, which reported difficulties in constructing chemical structures and identifying compounds from IUPAC names. Nevertheless, the results suggest that these limitations can be circumvented by leveraging the LLM's preloaded or retrieved compound names.

The present study further investigated the active learning and feedback loop method employed by LLMs. This approach involves prompting users for feedback or clarification, which is then incorporated into the training data to fine-tune the model and enhance response quality [35]. To facilitate feedback, the study introduced freely available standard chemical database websites, such as the PubChem database, in queries related to retrieving Compound Identifiers (CIDs), IUPAC and common names, SMILES, and InChI notations [36]. However, LLMs predominantly responded with denials of

Proceedings DOI: 10.21467/proceedings.178
ISBN: 978-81-984081-8-1

Series: AIJR Proceedings
ISSN: 2582-3922

interaction with these database websites, citing limitations as AI assistants, including their inability to browse the website or open URLs. However, in a generalized test assessing LLMs' capacity to browse the Wikipedia encyclopedia website revealed notable improvements in the capacity and quality of LLMs' responses, contingent upon successful interaction with specific websites. Yet, a minority of cases still yielded the denial of interaction responses from LLMs.

Another significant finding of the present study was the capacity of LLMs to integrate prompts from web-based-to-standalone application software. LLMs behave as AI chatbots, providing responses solely in text format. However, the present study observed that LLMs can produce chemical structures as response in output prompts. Further analysis revealed that these prompts were part of application software, particularly Mathematica [37]. The integration of Python-based RDKit libraries into LLMs specialized for chemistry will likely open new avenues of possibilities [38]. RDKit offers a comprehensive suite of libraries for processing intensive structural information of molecules, as well as supporting the extension of SMARTS libraries to study chemical reaction details. RDKit can address the challenge of Large Language Models (LLMs) in accurately processing chemical notations like SMILES, which arises from their string tokenization methods. By integrating RDKit with LLMs and utilizing chemical databases like PubChem, LLMs can identify compounds, retrieve SMILES strings, and analyze structures, enabling more precise query responses [39].

Applicability of Large Language Models (LLMs) in education has extensively explored in recent years. Present study widens the scope of LLMs capacity and skills to answer the queries of chemistry especially in the area of chemical lab related subjects. LLM’s extensive foundation in chemistry related information observed during the retrieval of structure, function, synthesis, and chemical properties related information of phenolphthalein. Further upon expanding the number of compounds, LLM’s maintained consistency with the quality of responses. The success in retrieving the IUPAC as well as logical name of water confirmed LLM’s understanding to chemistry related subject. Generation of well formatted chemistry experiment manual for standardization of NaOH solution proved the organization skills. Combining all these merits, LLMs displayed a suitable option for the role of chemistry lab assistant possessing requires a combination of chemistry knowledge, communication skills, interpersonal skills, organizational skills, problem-solving skills, and safety awareness. Additionally, with identification of reactive sites of phenolphthalein also showed a generative capacity in solving chemistry queries. Alternatively present study suggested to integrate RDKit with quick access of free chemical database website to increase efficiency of LLMs in solving chemistry related queries [40]. With the integration of computational chemistry tools, particularly RDKit, Large Language Models (LLMs) demonstrate significant potential in advancing molecular design, reaction prediction, and data-driven research. RDKit enhances the explainability and interpretability of LLM outputs by enabling rapid access to molecular descriptors and fingerprint data, facilitating the analysis of complex chemical compounds [41]. A critical challenge also persists in providing AI models, including LLMs, with access to high-quality and comprehensive chemical datasets [42]. The current study partially addresses this by demonstrating improved LLM performance when leveraging information from sources like the Wikipedia Encyclopedia. Addressing these challenges further could position LLMs as transformative tools in chemistry education and research, capable of conducting simulations, solving complex problems, and delivering interactive, context-sensitive responses to diverse queries. Therefore the results of present study will improve understanding of LLMs capacity to assist in chemistry lab courses and open a new avenue of research.

5 Conclusions

The present study evaluated the capacity of several large language models (LLMs) to assist in chemistry laboratory courses through a series of queries related to core chemistry concepts and skills. All LLMs exhibited a strong foundation in chemistry knowledge, which allowed them to successfully retrieve information on compound names, structures, synthesis procedures, chemical properties, and experimental methods. LLMs further displayed adept communication skills in generating well-formatted experiment manuals and provided concise, logical responses to queries. However, the study also identified certain limitations in LLMs' ability to process chemical structure representations and generation of common names from IUPAC names of compounds. These limitations stem from the method of reformulation or partial processing of uncommon queries used in natural language processing to meet the requirements of real time response. The incomplete reading of queries in these methods cause mixing up of compounds due to repetitions of symbols, syntaxes, prefix, suffix and words in IUPAC names. By strategically integrating LLMs with cheminformatics tools like RDKit and comprehensive data resources such as PubChem, these models can overcome current limitations and emerge as highly effective aids in chemistry education. These integrations would enhance LLMs' chemical data processing and structure analysis capabilities. Further, improve LLMs capacity to provide personalized assistance, boost student engagement, and facilitate meaningful and compelling learning experiences.

Declarations

Acknowledgement

Sanidhya Pal and Anup Paul would like to thank the Head of Department, Department of Applied Science, HMR Institute of Technology & Management, Delhi, for giving permission to work on this study.

Competing Interests

The author declares no conflicts of interest.

Publisher's Note

AIJR remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

How to Cite

Sanidhya Pal, Anup Paul (2025). Evaluation of LLMs Capacity to Assist in Chemistry Laboratory Courses. *AIJR Proceedings*, 213-228. https://doi.org/10.21467/proceedings.178.23

References

[1] W. Gan, Z. Qi, J. Wu, and J. C.-W. Lin, "Large language models in education: Vision and opportunities," 2023 2023: IEEE, pp. 4776-4785, doi: https://doi.org/10.1109/BigData59044.2023.10386291.

[2] Y. Chang *et al.*, "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1-45, 2024, doi: https://doi.org/10.1145/3641289.

[3] M. Perkins, "Academic Integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond," *Journal of university teaching & learning practice*, vol. 20, no. 2, p. 07, 2023, doi: https://doi.org/10.53761/1.20.02.07.

[4] Q. Ai *et al.*, "Information retrieval meets large language models: a strategic report from chinese ir community," *AI Open*, vol. 4, pp. 80-90, 2023, doi: https://doi.org/10.1016/j.aiopen.2023.08.001.

[5] R. Rejeleene, X. Xu, and J. Talburt, "Towards Trustable Language Models: Investigating Information Quality of Large Language Models," *arXiv preprint arXiv:2401.13086*, 2024, doi: https://doi.org/10.48550/arXiv.2401.13086.

[6] I. Koeper, J. Shapter, V. North, and D. Houston, "Turning chemistry education on its head: Design, experience and evaluation of a learning-centred 'Modern Chemistry'subject," *Journal of University Teaching & Learning Practice*, vol. 17, no. 3, p. 13, 2020, doi: https://doi.org/10.53761/1.17.3.13.

[7] A. Haaland, *Molecules and models: the molecular structures of main group element compounds*. Oxford University Press, 2008.

[8] K. Hallal, R. Hamdan, and S. Tlais, "Exploring the potential of AI-Chatbots in organic chemistry: An assessment of ChatGPT and Bard," *Computers and Education: Artificial Intelligence*, vol. 5, p. 100170, 2023, doi: https://doi.org/10.1016/j.caeai.2023.100170.

[9] X.-Q. Dao, "Which Large Language Model should You Use in Vietnamese Education: ChatGPT, Bing Chat, or Bard?," *Bing Chat, or Bard*, 2023, doi: https://dx.doi.org/10.2139/ssrn.4527476.

[10] K. M. Jablonka *et al.*, "14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon," *Digital Discovery*, vol. 2, no. 5, pp. 1233-1250, 2023, doi: https://doi.org/10.1039/d3dd00113j.

[11] D. G. Herrington and M. B. Nakhleh, "What defines effective chemistry laboratory instruction? Teaching assistant and student perspectives," *Journal of chemical Education*, vol. 80, no. 10, p. 1197, 2003, doi: https://doi.org/10.1021/ed080p1197.

[12] A. P. Leong, "Clause complexing in research-article abstracts: Comparing human-and AI-generated texts," *ExELL (Explorations in English Language and Linguistics)*, vol. 11, no. 2, pp. 99-132, 2023, doi: https://doi.org/10.2478/exell-2023-0008.

[13] J. G. R. Berrío, "Inteligencias artificiales generativas a 2023," *iCartesiLibri*, 2023. [Online]. Available: recursoseducativos.unam.mx.

[14] A. I. Inflection, "Inflection-1," Technical report, 2023b, 2023. [Online]. Available: https://inflection.ai/assets/Inflection-1_0622.pdf

[15] D. Rathod, "Web browser forensics: google chrome," *International Journal of Advanced Research in Computer Science*, vol. 8, no. 7, pp. 896-899, 2017, doi: http://dx.doi.org/10.26483/ijarcs.v8i7.4433.

[16] L. Foulkes, *Learn Microsoft Office 2019: A Comprehensive Guide to Getting Started with Word, PowerPoint, Excel, Access, and Outlook*. Packt Publishing Ltd, 2020.

[17] S. Kim *et al.*, "PubChem substance and compound databases," *Nucleic acids research*, vol. 44, no. D1, pp. D1202-D1213, 2016, doi: https://doi.org/10.1093/nar/gkv951.

[18] T. Brown, "ChemDraw," *The Science Teacher*, vol. 81, no. 2, p. 67, 2014. [Online]. Available: www.proquest.com.

[19] J. Yan, J. Xu, C. Song, C. Wu, Y. Li, and Y. Zhang, "Understanding in-context learning from repetitions," *arXiv preprint arXiv:2310.00297*, 2023, doi: https://doi.org/10.48550/arXiv.2310.00297.

[20] M. R. Ai4Science and M. A. Quantum, "The impact of large language models on scientific discovery: a preliminary study using gpt-4," *arXiv preprint arXiv:2311.07361*, 2023, doi: https://doi.org/10.48550/arXiv.2311.07361.

[21] R. Rheeder *et al.*, *Living Water: An interdisciplinary exploration of water as a theological theme*. AOSIS, 2023.

[22] J. Jia, T. Wang, Y. Zhang, and G. Wang, "The comparison of general tips for mathematical problem solving generated by generative AI with those generated by human teachers," *Asia Pacific Journal of Education*, vol. 44, no. 1, pp. 8-28, 2024, doi: https://doi.org/10.1080/02188791.2023.2286920.

[23] J. Schrier, "Comment on "Comparing the Performance of College Chemistry Students with ChatGPT for Calculations Involving Acids and Bases"," *Journal of Chemical Education*, 2024, doi: https://doi.org/10.1021/acs.jchemed.4c00058.

[24] L. Derksen, C. Michaud Leclerc, and P. C. L. Souza, "Searching for answers: The impact of student access to wikipedia," 2019. [Online]. Available: http://wrap.warwick.ac.uk/131653.

[25] T. Guo *et al.*, "What can large language models do in chemistry? a comprehensive benchmark on eight tasks," *Advances in Neural Information Processing Systems*, vol. 36, pp. 59662-59688, 2023, doi: https://doi.org/10.48550/arXiv.2305.18365.

[26] D. A. Boiko, R. MacKnight, B. Kline, and G. Gomes, "Autonomous chemical research with large language models," *Nature*, vol. 624, no. 7992, pp. 570-578, 2023, doi: https://doi.org/10.1038/s41586-023-06792-0.

[27] A. A. Baker, "A history of indicators," *Chymia*, vol. 9, pp. 147-167, 1964, doi: https://doi.org/10.2307/27757238.

[28] J. Kim *et al.*, "SuRe: Improving Open-domain Question Answering of LLMs via Summarized Retrieval," 2023 2023, doi: https://doi.org/10.48550/arXiv.2404.13081.

[29] F. M. Dunnivant, *Environmental laboratory exercises for instrumental analysis and environmental chemistry*. John Wiley & Sons, 2004.

[30] K. Saito, K. Sohn, C.-Y. Lee, and Y. Ushiku, "Unsupervised LLM Adaptation for Question Answering," *arXiv preprint arXiv:2402.12170*, 2024, doi: https://doi.org/10.48550/arXiv.2402.12170.

[31] M. Jovanovic and P. Voss, "Trends and Challenges of Real-time Learning in Large Language Models: A Critical Review," *arXiv preprint arXiv:2404.18311*, 2024, doi: https://doi.org/10.48550/arXiv.2404.18311.

[32] A. Anand, V. Setty, and A. Anand, "Context aware query rewriting for text rankers using llm," *arXiv preprint arXiv:2308.16753*, 2023, doi: https://doi.org/10.48550/arXiv.2308.16753.

[33] S. Müller, "Ambient Occlusion zwischen sich frei bewegenden Starrkörpern mittels Coherent Shadow Maps," Bachelor Informatik, Informatik, Universität Koblenz, Landau, 2008. [Online]. Available: https://kola.opus.hbz-nrw.de/index.php/frontdoor/index/index/year/2008/docId/223

[34] P. R. Kaveri, "ChatGPT: The power of AI," *Indian Journal of Applied Research*, vol. 13, no. 05, 2023, doi: https://doi.org/10.36106/ijar.

[35] J. Liang, L. Liao, H. Fei, B. Li, and J. Jiang, "Actively Learn from LLMs with Uncertainty Propagation for Generalized Category Discovery," 2024 2024: NAACL-HLT. [Online]. Available: https://aclanthology.org/2024.naacl-long.434. [Online]. Available: https://aclanthology.org/2024.naacl-long.434

[36] F. Suchanek and A. T. Luu, "Knowledge bases and language models: Complementing forces," 2023 2023: Springer, pp. 3-15, doi: https://doi.org/10.1007/978-3-031-45072-3_1.

[37] M. L. Abell and J. P. Braselton, *Mathematica by example*. Academic Press, 2021.

[38] G. Landrum, *Rdkit documentation*, 2013, p. 4. [Online]. Available: http://www.rdkit.org.

[39] R. R. Kotkondawar, S. R. Sutar, A. W. Kiwelekar, and V. J. Kadam, "Integrating Transformer-based Language Model for Drug Discovery," 2024 2024: IEEE, pp. 1096-1101, doi: https://doi.org/10.23919/INDIACom61295.2024.10498263.

[40] Editors, "For chemists, the AI revolution has yet to happen," *Nature*, vol. 617, no. 7961, p. 438, 2023, doi: https://doi.org/10.1038/d41586-023-01612-x.

[41] C.-H. Chen, K. Tanaka, M. Kotera, and K. Funatsu, "Comparison and improvement of the predictability and interpretability with ensemble learning models in QSPR applications," *Journal of cheminformatics*, vol. 12, pp. 1-16, 2020, doi: https://doi.org/10.1186/s13321-020-0417-9.

[42] T. Erdmann *et al.*, "ChemChat—Recent Advances in Democratizing and Facilitating Access to Domain-Specific AI/ML Through LLM-Powered Conversational Assistants," 2024. [Online]. Available: https://research.ibm.com/publications/chemchatrecent-advances-in-democratizing-and-facilitating-access-to-domain-specific-aiml-through-llm-powered-conversational-assistants. [Online]. Available: https://research.ibm.com/publications/chemchatrecent-advances-in-democratizing-and-facilitating-access-to-domain-specific-aiml-through-llm-powered-conversational-assistants

Proceedings DOI: 10.21467/proceedings.178
ISBN: 978-81-984081-8-1

Series: AIJR Proceedings
ISSN: 2582-3922