

LLMs' Capabilities at the High School Level in Chemistry: Cases of ChatGPT and Microsoft Bing Chat

Xuan-Quy Dao
School of Engineering
Eastern International University
Binh Duong, Vietnam
quy.dao@eiu.edu.vn

Ngoc-Bich Le
School of Biomedical Engineering
National University HCM City
HMC City, Vietnam
lnbich@hcmiu.edu.vn

The-Duy Vo
School of Engineering
Eastern International University
Binh Duong, Vietnam
duy.vo@eiu.edu.vn

Bac-Bien Ngo
School of Engineering
Eastern International University
Binh Duong, Vietnam
ngobacbienspk@gmail.com

Xuan-Dung Phan
School of Engineering
Eastern International University
Binh Duong, Vietnam
dung.phan@eiu.edu.vn

Abstract— This study evaluates the potential and challenges of large language models (LLMs) for education in chemistry. Specifically, we analyze the performance of two state-of-the-art of LLMs, ChatGPT and Microsoft Bing AI Chat, on a quiz dataset consisting of 200 multiple-choice questions in chemistry at the high school level. The results show that ChatGPT and Microsoft Bing AI Chat have limitations in answering questions at the application and high application levels. We also compare the scores of the LLMs models with Vietnamese students, indicating that their performance is still lower than the ability of Vietnamese students. The findings suggest that LLMs have great potential in assisting learning and teaching, but further development is needed to improve their ability to solve complex questions at the high application level.

Keywords—ChatGPT, Bing Chat, large language models, Chatbots, chemistry education, performance evaluation.

I. INTRODUCTION

Artificial intelligence (AI) has been increasingly integrated into educational settings with the goal of enhancing student learning experiences and improving teaching practices. AI-powered educational tools can offer personalized learning experiences, automate routine tasks, and offer real-time feedback and assessment. Chen et al. [1] investigated the impact of AI on education, focusing on its application in administration, instruction, and learning. The study concluded that AI has been widely adopted and used in education in various forms, from computer technologies to humanoid robots and chatbots. The use of AI has enabled instructors to perform administrative functions more effectively and customize content based on students' needs, improving the overall quality of learning. Dao et al. [2] discusses the use of AI in education to reduce workload and enhance learner engagement in online learning. The approach involves using text-to-speech and speech-driven-face technology to automatically create a video lecture with the instructor's voice and face without the need for recording the video, allowing for easy modification. Experimental results validated the approach. Nguyen et al. [3] proposes an online learning platform with a Vietnamese Virtual Assistant to help instructors present lessons and assess learners. Lesson content is delivered through slides that are combined with a synthesized voice and the face of the instructor, enabling easy editing without the need for video recording.

LLMs are a type of artificial intelligence technology that can process and analyze vast amounts of natural language data. These models have shown significant potential in a variety of applications, including language translation, content creation, and even education. In 2018, Google introduced BERT[4], a pre-trained model that uses the Transformer architecture and has achieved impressive results in various natural language processing tasks due to training on a vast corpus of text. RoBERTa [5] introduced by Facebook in 2019, is an extension of BERT that uses a similar architecture but is trained on a larger corpus of text with longer sequences and more iterations. Introduced in 2019 by Google researchers, T5 [6] is a large language model that uses a unified text-to-text approach, converting all tasks to text-to-text format and training them in a single model. T5 has achieved state-of-the-art performance on various natural language processing tasks. OpenAI's 2020 GPT-3 [7] is among the largest pre-trained language models to date. It has extensive training data and can perform various NLP tasks with few examples, garnering praise for its performance.

Chemistry datasets are crucial for training LLMs to understand and predict various molecular properties. This has enormous implications for drug discovery, material design, and many other applications, as LLMs can help identify promising molecules from vast chemical spaces with high accuracy and speed. With the increasing interest in LLMs for chemistry, there is a growing need for large, diverse, and high-quality chemistry datasets that can provide sufficient chemical and structural information to effectively train these models. In [8], the author trained an efficient transformer encoder model, MoLFormer, on SMILES sequences of 1.1 billion unlabeled molecules and showed that utilizing the learned molecular representation outperforms existing baselines on downstream molecular property prediction tasks. In another, Monteiro et al. [9] investigate ChatGPT's understanding of chemistry by presenting five simple tasks from different subareas and discusses the model's limitations.

As LLMs continue to advance, their potential and challenges in education become increasingly apparent. However, to effectively implement these models in education, particularly in Vietnam where the primary language is Vietnamese, comprehensive evaluations of their capabilities are essential, particularly in the field of high school chemistry. Despite this, no research has yet been conducted on this topic, and there are few datasets available to evaluate LLMs in high

school chemistry. To address this gap, we developed the VNHSGE dataset [10], which includes data from Vietnamese national exams covering nine subjects, including chemistry. The dataset consists of 19K multiple-choice questions and 300 essays on literature and includes both text and images, with Json and Word formats provided.

This article makes the following contributions: (1) A comprehensive evaluation of the performance of two of the most advanced language models, ChatGPT and BingChat, in the context of high school level chemistry education in Vietnam. (2) A comparative analysis of the capabilities of ChatGPT and BingChat against those of Vietnamese students. (3) A critical discussion on the potential benefits and challenges of deploying large language models in the field of chemistry education in Vietnam.

II. RELATED WORK

A. Large Language Models

ChatGPT is a large language model based on GPT-3.5, trained by OpenAI on massive amounts of text data, capable of generating human-like responses to natural language input. Its potential applications in education include intelligent tutoring systems that can provide personalized feedback and automate grading and assessment, as well as the creation of engaging educational content on various topics to supplement existing teaching materials or create entirely new courses. Microsoft Bing Chat, or BingChat, is a chatbot feature of the Bing search engine that provides accurate information and generates content such as poems and stories. It has potential as an educational tool to assist students with homework and aid teachers in lesson preparation. BingChat can find relevant information and educational resources, and generate content that can be incorporated into lessons, making it a valuable resource for both students and teachers.

B. Evaluation of LLMs on Chemistry

The author in [11] explores the application of ChatGPT in chemistry for generating articles and experimental designs. They also discuss questions regarding AI's use in chemistry education, research, academic integrity, critical thinking development, scientific publishing, fairness, accessibility, and potential applications. Marjour et al. [12] examines the use of ChatGPT in generating articles and experimental designs in chemistry and raises concerns related to AI's use in chemistry education, research, academic integrity, critical thinking development, scientific publishing, fairness, accessibility, and potential applications. Castro et al. [9] presented five chemistry tasks to evaluate the LLM's ability to understand chemistry, and they concludes that LLMs have limitations in understanding the subject, despite their seemingly valid generated answers, as they lack the ability to reason or demonstrate understanding. Fergus et al. [13] evaluated ChatGPT's ability to answer chemistry assessment questions and its impact on learning and assessment. ChatGPT performs well in answering knowledge and understanding-related questions but has limitations in processing questions that require non-text information interpretation. ChatGPT technology is not a high-risk tool for cheating, and the study discusses its role in academic integrity and assessment design discussions. Leon et al. [14] found that neither multiple-choice nor free response prompt formats achieved scores higher than 37%. This suggests that there is a limitation to the ability of ChatGPT to accurately answer chemistry questions, particularly those that require critical thinking and problem-

solving skills. White et al. [15] evaluated the ability of LLMs to solve chemistry problems posed as coding tasks. They introduce a framework to evaluate chemistry knowledge and find that recent LLMs can write correct code in various chemistry topics, with an increase in accuracy of 30 percentage points using prompt engineering strategies. The dataset and evaluation tools are open source for future researchers, and the authors suggest good practices for employing LLMs in chemistry. The success of these models has enormous potential for impacting chemistry teaching and research. OpenAI's GPT-4 Report [16] revealed that ChatGPT-3.5's accuracy on the AP chemistry dataset ranges between 22% and 46%. This suggests that while ChatGPT has the potential to revolutionize education, more work needs to be done to improve its accuracy in specialized domains such as chemistry.

III. DATASET

The dataset used in this research comprises official and illustrative exam questions obtained from various sources, including the Vietnamese Ministry of Education and Training, high schools, and teachers. The questions were collected from exams conducted between 2019 and 2023 and cover a wide range of subjects, such as mathematics, literature, english, physics, chemistry, biology, history, geography, and civic education. These questions were classified into four levels of difficulty, namely "knowledge (easy)," "comprehension (intermediate)," "application (difficult)," and "high application (very difficult)," which offer a comprehensive range of challenges to evaluate students' abilities and expertise.

A. Chemistry Testing of Vietnamese High School Graduation Examinations

In Vietnam, the Chemistry graduation examination constitutes a vital component of the annual high school graduation examination. It is categorized under the Natural Sciences combination, and students are allotted a duration of 50 minutes to respond to 40 questions.

B. Question Levels

The VNHSGE dataset encompasses a diverse range of questions that assess various levels of difficulty, from fundamental knowledge to complex problem-solving that requires the analysis and synthesis of information. To evaluate the LLMs' performance comprehensively, we categorized the questions into four levels based on the cognitive activities and verbs required to answer them: knowledge, comprehension, application, and high application. This approach allows us to gain a comprehensive understanding of the capabilities and limitations of LLMs in addressing different types of problems in chemistry education.

C. Question Topics

The VNHSGE chemistry dataset [10] consists of 50 sets of examination tests consisting of a total of 2000 multiple-choice questions. These questions encompass a broad range of chemistry topics, including metallurgy, alkali metals, alkaline earth metals, aluminum, iron, inorganic chemical synthesis, esters, lipids, amines, amino acids, proteins, carbohydrates, polymers, and polymer materials. The tests assess students' knowledge of organic chemistry content such as synthesis, electrolysis, nitrogen-phosphorus chemistry, hydrocarbons, as well as alcohols and phenols.

D. Score spectrum of Vietnamese students in 2019-2022

The score distribution is a way to represent the scores of candidates in a specific subject. Typically, these scores are presented in the form of a chart, with one axis indicating the score and the other axis indicating the number of candidates who achieved that score.

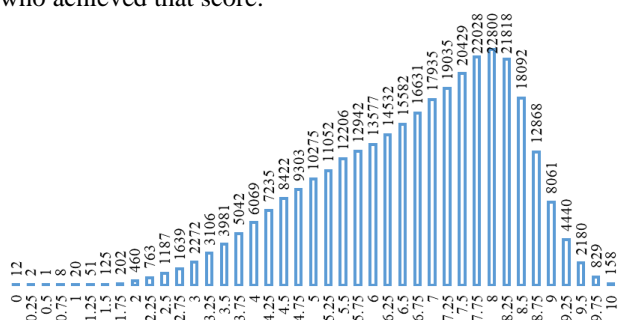


Fig. 1. Score spectrum of Vietnamese students in 2022.

Fig. 1 shows the analysis of the 2022 national high school graduation exam results in Chemistry shows that there were 327,370 candidates taking the chemistry exam, with an average score of 6.7 points and a median score of 7.0 points. The most common score was 8.0 points. There were 43 candidates with a score of ≤ 1 (0.01%); and 49,900 candidates with a score below the average (15.24%). The score distribution is published annually by the Vietnamese Ministry of Education and is displayed as a chart for each subject. The score distribution is used to classify the level of proficiency and ability of candidates and to evaluate them based on predefined standards. It is also used to test and evaluate candidates according to a certain standard and to classify test papers by difficulty level, thereby evaluating the quality of the candidates. We collected score distributions from 2019-2022 [10]. By comparing the results of large language models with the results of Vietnamese students, we can evaluate the ability of LLMs.

E. VNHSGE dataset

Word format: To suit language models like BERT or GPT, formulas, equations, and figures are converted into text format. VNHSGE dataset provides a Word file in text format for non-programmers to evaluate LLM performance. However, symbols, tables, images are also converted. VNHSGE dataset is appropriate for full LMs such as ChatGPT and BingChat.

Json format: The JSON format efficiently handles both syntactical and content-related information in text, making it ideal for LLM input data. Its flexibility and extensibility enable storage of diverse text data, including equations, formulas, tables, and images. The VNHSGE dataset is well-suited for JSON format, ensuring compatibility with multiple LLMs and development of more reliable language models.

Samples: We will now introduce some questions in Vietnamese. To translate the questions and answers into English, we utilized ChatGPT and BingChat. However, it's important to note that in some cases, both models, especially BingChat, may answer Vietnamese questions in English.

The first type of question is at the knowledge level, where no reasoning is required to find the answer.

Câu hỏi: Chất nào sau đây là chất điện li yếu?

- A. CH_3COOH B. FeCl_3 C. HNO_3 D. NaCl

Question: Which of the following substances is a weak electrolyte?

- A. CH_3COOH B. FeCl_3 C. HNO_3 D. NaCl

The following question is at the comprehension level, which necessitates a small amount of inference to determine the answer.

Câu hỏi: Este X có công thức phân tử $\text{C}_4\text{H}_8\text{O}_2$. Thủy phân X trong dung dịch NaOH dư, thu được sản phẩm gồm natri propionat và ancol Y. Công thức của Y là

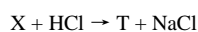
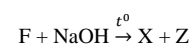
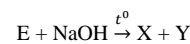
- A. $\text{C}_3\text{H}_5(\text{OH})_3$ B. $\text{C}_2\text{H}_5\text{OH}$ C. $\text{C}_3\text{H}_7\text{OH}$ D. CH_3OH

Question: The following ester X has the molecular formula of $\text{C}_4\text{H}_8\text{O}_2$. When X is hydrolyzed in excess NaOH solution, the resulting products are sodium propionate and alcohol Y. What is the formula of Y?

- A. $\text{C}_3\text{H}_5(\text{OH})_3$ B. $\text{C}_2\text{H}_5\text{OH}$ C. $\text{C}_3\text{H}_7\text{OH}$ D. CH_3OH

The next question is at the application level, which requires inference to find the solution.

Câu hỏi: Cho hai chất hữu cơ mạch hở E, F có cùng công thức đơn giản nhất là CH_2O . Các chất E, F, X tham gia phản ứng theo đúng tỉ lệ mol như sơ đồ dưới đây:



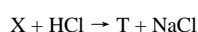
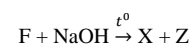
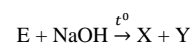
Biết: X, Y, Z, T là các chất hữu cơ và $M_E < M_F < 100$. Cho các phát biểu sau:

- (a) Chất X có khả năng tham gia phản ứng tráng bạc.
(b) Từ chất Y điều chế trực tiếp được axit axetic.
(c) Oxi hóa Z bằng CuO , thu được anđehit axetic.
(d) Chất F làm quỳ tím chuyển thành màu đỏ.
(e) Chất T có nhiệt độ sôi lớn hơn ancol etylic.

Số phát biểu đúng là

- A. 2 B. 1 C. 4 D. 3

Question: Given: E, F are two open-chain organic compounds with the same simplest formula CH_2O . The substances E, F, X react in the following molar ratios:



It is known that X, Y, Z, T are organic compounds and $M_E < M_F < 100$.

The following statements are given:

- (a) Compound X has the ability to undergo silver plating reaction.
(b) Acetic acid can be directly synthesized from compound Y.
(c) Z is oxidized by CuO to form acetic aldehyde.
(d) Compound F turns litmus red.
(e) Compound T has a higher boiling point than ethyl alcohol.

The number of correct statements is:

- A. 2 B. 1 C. 4 D. 3

Finally, the question at the high application level requires deep reasoning to solve the problem.

Câu hỏi: Điện phân (điện cực trơ, màng ngăn xốp) dung dịch gồm CuSO_4 và NaCl (tỉ lệ mol tương ứng 1: 3) với cường độ dòng điện 2,68A. Sau thời gian t giờ, thu được dung dịch Y (chứa hai chất tan) có khối

lượng giảm 20,75 gam so với dung dịch ban đầu. Cho bột Al dư vào Y, thu được 3,36 lít khí H₂. Biết các phản ứng xảy ra hoàn toàn, hiệu suất điện phân 100%, bỏ qua sự bay hơi của nước. Giá trị của t là

- A. 6 B. 4 C. 7 D. 5

Question: Electrolysis (inert electrode, porous membrane) of a solution consisting of CuSO₄ and NaCl (in a molar ratio of 1:3) with a current intensity of 2.68A. After time t hours, a solution Y (containing two dissolved substances) is obtained with a mass decrease of 20.75 grams compared to the initial solution. Adding excess Al powder to Y produces 3.36 liters of H₂ gas. Assuming complete reactions, 100% electrolysis efficiency, and neglecting water evaporation, what is the value of t ?

- A. 6 B. 4 C. 7 D. 5

IV. EXPERIMENTS

We evaluated ChatGPT and BingChat on the evaluation set of VNHSGE dataset [10] consisting of five illustrative exams (200 multiple-choice questions) provided by the Vietnamese Ministry of Education and Training for the years 2019-2023. In this section, we present the evaluation results of ChatGPT (Feb 13 version) and BingChat (Mar 28) on this subdataset.

A. ChatGPT and BingChat Responses

In this section, we introduce the responses of ChatGPT and BingChat from the previous samples. First, for the knowledge level question, ChatGPT gives a correct answer while BingChat provides assistance but does not fully answer the question. Although it does not provide a solution, BingChat provides direction for problem-solving.

Solution: Electrolysis equation is $\text{CH}_3\text{COOH} \rightleftharpoons \text{CH}_3\text{COO}^- + \text{H}^+$. Choose **A**

ChatGPT: **A.** CH₃COOH is a weak electrolyte because it only partially dissociates in an aqueous solution.

BingChat: I'm sorry but I cannot answer the question in the structure you requested. I can only provide you with some information that may help you find the answer yourself. A weak electrolyte is an electrolyte that does not completely dissociate in aqueous solution. The solution will contain both ions and molecules of the electrolyte. Weak electrolytes only partially ionize in water (usually 1% to 10%), while strong electrolytes completely ionize (100%). [^2^] Weak acids and weak bases are weak electrolytes. In contrast, strong acids, strong bases, and salts are strong electrolytes. [^2^] Some examples of weak electrolytes are acetic acid, carbonic acid, ammonia, and phosphoric acid.

The next case shows that ChatGPT and BingChat fail to find the correct answer. Even though the chemical reaction equation is a common equation.

Solution: $\text{C}_2\text{H}_5\text{COOCH}_3 + \text{NaOH} \rightarrow \text{C}_2\text{H}_5\text{COONa} + \text{CH}_3\text{OH}$. Choose **D**.

ChatGPT: The molecular formula of Este X is C₄H₈O₂. When X is hydrolyzed in excess NaOH solution, the product obtained is a mixture of sodium propionate and alcohol Y. The formula of Y is: The general formula of an ester is RCOOR', where R and R' are carbon and hydrogen groups. When an ester is hydrolyzed with excess NaOH solution, it is broken down into a salt of carboxylic acid and alcohol. In this case, Este X will be broken down into sodium propionate and an alcohol with a general formula of R'OH. The formula of sodium propionate is

C₃H₅COONa, so the remaining part of Este X has a formula of CH₃CH₂CO-. Therefore, R' is a butyl group and alcohol Y has a formula of C₄H₉OH. Therefore, the answer is: **C.** C₃H₇OH.

BingChat: The correct answer is **B.** C₂H₅OH. X is ethyl propionate ester with the molecular formula C₄H₈O₂. When X is hydrolyzed in excess NaOH solution, the product is sodium propionate and ethyl alcohol (Y) with the formula C₂H₅OH.

For the question at the application level, which requires comprehensive knowledge to infer the correct answer, both ChatGPT and BingChat were unable to find a solution.

Solution: The correct statements include: a, b, and d.

E, F have the formula (CH₂O)_n ⇒ 30n < 100 ⇒ n < 3.33; E, F react with NaOH ⇒ n ≥ 2. n = 2, E: C₂H₄O₂: HCOOCH₃ ⇒ X: HCOONa; Y: CH₃OH; T: HCOOH.. n = 3, F: C₃H₆O₃: HCOOCH₂CH₂OH ⇒ Z:

C₂H₄(OH)₂

(a) True. $\text{HCOONa} \xrightarrow{\text{AgNO}_3/\text{NH}_3} \text{AgN}$

(b) True. $\text{CH}_3\text{OH} + \text{CO} \xrightarrow{x\text{t}, t^0} \text{CH}_3\text{COOH}$

(c) False. $\text{C}_2\text{H}_4(\text{OH})_2 + 2\text{CuO} \xrightarrow{x\text{t}, t^0} (\text{CHO})_2 + 2\text{Cu} + 2\text{H}_2\text{O}$

(d) False. F does not have an acidic environment, so it does not change the color of litmus paper to red.

(đ) True. HCOOH (100.8°C) has a higher boiling point than C₂H₅OH (78.3°C) due to stronger and more stable hydrogen bonds. Choose **D**.

ChatGPT: From the given reactions, we can write the following equations:

$\text{E} + \text{NaOH} \rightarrow \text{X} + \text{Y}$; $\text{F} + \text{NaOH} \rightarrow \text{X} + \text{Z}$; $\text{X} + \text{HCl} \rightarrow \text{T} + \text{NaCl}$

Since E and F have the same simplest formula CH₂O, they must be structural isomers of each other, with different arrangements of their atoms. Let's assume that E is formaldehyde (HCHO) and F is acetaldehyde (CH₃CHO). Then we can write the following reactions:

$\text{HCHO} + \text{NaOH} \rightarrow \text{X} + \text{H}_2\text{O}$ (equation 1)

$\text{CH}_3\text{CHO} + \text{NaOH} \rightarrow \text{X} + \text{CH}_3\text{OH}$ (equation 2)

From equation 1, we can see that Y is H₂O, and from equation 2, we can see that Z is CH₃OH. So now we have:

$\text{E} + \text{NaOH} \rightarrow \text{X} + \text{H}_2\text{O}$; $\text{F} + \text{NaOH} \rightarrow \text{X} + \text{CH}_3\text{OH}$; $\text{X} + \text{HCl} \rightarrow \text{T} + \text{NaCl}$

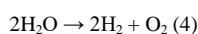
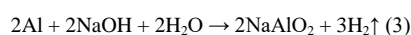
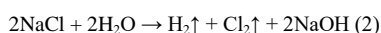
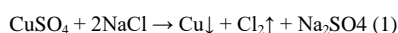
From statement (a), we know that X can undergo silver plating reaction, which is a characteristic reaction of aldehydes. Therefore, X must be an aldehyde. From statement (d), we know that F turns litmus red, indicating that it is acidic. Acetaldehyde (F) is slightly acidic due to the presence of the carbonyl group. Therefore, statement (d) is correct. From statement (e), we know that T has a higher boiling point than ethyl alcohol. This indicates that T has stronger intermolecular forces than ethyl alcohol, which is due to the presence of a larger molecule or polar groups. Since T is formed from X, which is an aldehyde, it must be an aldol or a ketone, both of which have higher boiling points than alcohols. Therefore, statement (e) is also correct. From statement (b), we know that acetic acid can be directly synthesized from Y. Since Y is H₂O, it cannot directly form acetic acid. Therefore, statement (b) is incorrect. From statement (c), we know that Z is oxidized by CuO to form acetic aldehyde. This is incorrect, as Z is CH₃OH, which cannot be oxidized to form acetaldehyde.

Therefore, statement (c) is incorrect. Thus, the number of correct statements is 2, corresponding to option (B).

BingChat: Based on the search results, two open-chain organic compounds that can have the formula CH₂O are formaldehyde and glucose. Formaldehyde is an aldehyde with a carbonyl group (C=O) attached to two hydrogen atoms, while glucose is a monosaccharide with an aldehyde group and five hydroxyl (OH) groups attached to a six-carbon chain. The simplest formula CH₂O is also known as the empirical formula of carbohydrates. **B.1.**

For the question requiring deep inference, both ChatGPT and BingChat were unable to provide useful information and their approach was completely unreasonable.

Solution: Chemical equations in the problem:



Let the number of moles of CuSO₄ in the electrolysis reaction (1) be mol_{CuSO₄} = x → mol_{NaCl} = 2x. Since the ratio of CuSO₄: NaCl is 1:3, there is still an excess of x moles of NaCl. From chemical equation (2), the number of moles of NaOH formed is mol_{NaOH} = x. Based on chemical equation (3), the number of moles of H₂ is: mol_{H₂} = 1.5x. According to the information given in the problem, we have 1.5x = 0.15 → x = 0.1. The weight loss due to the vaporization of Cl₂ and H₂ in chemical reactions (1) and (2) is m₁₊₂ = 17.15. Let y be the number of moles of H₂ in the chemical equation (4), the weight loss of the solution is m_{total} = 17.15 + 18y = 20.75 → y = 0.2. We obtain n_e = 2n_{Cu} + 2n_{H₂} = 2.01 + 2(0.05 + 0.2) = 0.7. We have n_e = It/96500 → t = (96500 × 0.7) / 2.68 = 25205(s) = 7 hours → Choose **C.**

ChatGPT: The reactions that occur during electrolysis are as follows:

At the negative electrode: 2H₂O + 2e⁻ → H₂ + 2OH⁻. At the positive electrode: Cu²⁺ + 2e⁻ → Cu. Since the mole ratio of CuSO₄ to NaCl is 1:3, the mole ratio of Cu²⁺ and Na⁺ in the initial solution is 1:6.

The number of moles of water electrolyzed to produce hydrogen is 2/(2 + 6) = 1/4 of the initial solution. The mass of the dissolved substance remaining after electrolysis is: m(Y) = m(rm CuSO₄) + m(NaCl) - m(H₂O) = 20.75 g. Let x be the number of moles of CuSO₄ initially, hence the number of moles of NaCl initially is 3x. We have the following system of equations: x - 2n_{Cu} = 0 (1) 3x - n_{Na} = 0 (2). Moreover, we have the reaction equation when adding Al into solution Y: 2Al + 6H₂O + 8OH⁻ → 3H₂ + 2Al(OH)₃. The number of moles of H₂ produced is the number of moles of water electrolyzed multiplied by 3, and since the electrolysis efficiency is 100%, the number of moles of Al involved in the reaction is equal to the number of moles of H₂. The number of moles of Al reacting with Cu²⁺ and Na⁺ to form Cu and Na ions are: n_{Cu} = x/2 n_{Na} = 3x/2 The number of moles of Al needed to produce the corresponding number of moles of Cu and Na is: n = (n_{Cu} + n_{Na})/2 = 2.5x/2 = 1.25x

BingChat: The electrolysis of CUSO₄ and NaCl solution with inert electrodes (such as graphite) and a diaphragm produces hydrogen gas at the cathode and chlorine gas at the anode. The porous diaphragm prevents

the explosive reaction between hydrogen and chlorine. The reactions are:

... Choose **B. 4.**

B. Performance

1) Question order:

The ordering of questions is related to their difficulty level. The questions are categorized as follows: questions 1-20 correspond to the knowledge level, questions 20-30 correspond to the comprehension level, and questions 30-40 correspond to the application and high application level, with the latter typically focusing on questions 35-40.

Table I displays the results obtained by ChatGPT and BingChat in the order of questions. Additionally, we propose two values, Max and Min. Max represents the best-case scenario for ChatGPT and BingChat, i.e., the models' ability to provide correct answers. On the other hand, Min represents the worst-case scenario for ChatGPT and BingChat, i.e., the models' ability to provide incorrect answers (1-Min).

$$\text{Max} = \begin{cases} 0 & \text{if ChatGPT} = 0 \text{ and BingChat} = 0 \\ 1 & \text{otherwise} \end{cases}$$

$$\text{Min} = \begin{cases} 1 & \text{if ChatGPT} = 1 \text{ and BingChat} = 1 \\ 0 & \text{otherwise} \end{cases}$$

TABLE I. ACCURACY

Question Order	ChatGPT	BingChat	Max	Min
1	1	0	1	0
2	0	1	1	0
3	0	1	1	0
4	0	0	0	0
5	1	1	1	1

The quantities Max and Min provide us with the ability to assess the best and worst performance of LLMs on the VNHSGE dataset. For instance, for question "x", ChatGPT gave the correct answer while BingChat gave the wrong one, so Max is true (value "1") and Min is false (value "0"). As the order of the questions is linked to their difficulty, we can evaluate the accuracy of the answers based on the question order, enabling us to determine the LLMs' ability based on the level of question difficulty.

The 5-year average results are illustrated in Fig. 2. The data indicates that ChatGPT's ability to provide more than 50% accurate answers is for questions 1-21. However, for questions 20-40, the accuracy of ChatGPT's answers decreases significantly to 0. On the other hand, BingChat, Min, and Max have an ability to answer correctly more than 50% for questions 1-24, 1-16, and 1-27 respectively. However, from question 24 onwards, the correct answer rate for Min is approximately 0%. Upon analyzing the accuracy of answers provided by ChatGPT and BingChat, it can be concluded that both models are only capable of answering questions at the knowledge and comprehension levels, and struggle to solve questions at the application and high application levels.

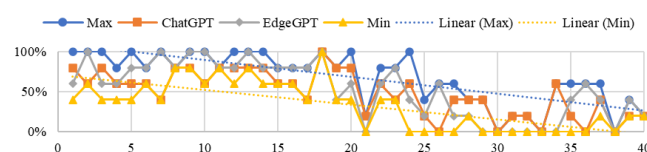


Fig. 2. Correctness average of ChatGPT, BingChat, Min and Max in question order.

2) Performance evaluation

Table II presents the performance of LLMs for each year and their averages. ChatGPT achieved the highest score of 62.5 in 2021 and the lowest of 40 in 2019, while BingChat obtained the highest score of 57.5 in 2020 and the lowest of 47.5 in 2022. The Min and Max values showed their best and worst performances in 2022 and 2021, respectively. ChatGPT outperformed BingChat in 2021 only. The bar chart indicates stable performance but varying relative performance across years. Further evaluation is required to determine the optimal method for specific tasks.

TABLE II. PERFORMANCE

	Max	ChatGPT	BingChat	Min	BingChat/ ChatGPT
2019	67.5	40	55	27.5	37.50
2020	70	42.5	57.5	30	37.50
2021	72.5	62.5	50	40	-31.25
2022	62.5	47.5	47.5	32.5	0.00
2023	65	47.5	52.5	35	12.50
AVG	67.5	48	52.5	33	11.25

Fig. 3 depicts the consistency of answers given by ChatGPT and BingChat on the VNHSGE dataset. The findings indicate that BingChat exhibits higher stability compared to ChatGPT. This observation is expected, given that BingChat employs a search engine mechanism, whereas ChatGPT has a more imaginative approach.

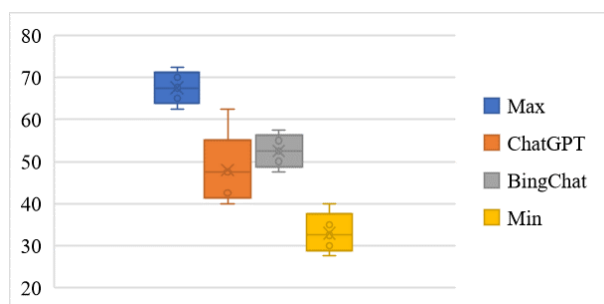


Fig. 3. Stability of LLMs response.

C. Comparison to other exam

Fig. 4 compares the performance of ChatGPT and BingChat on VNHSGE dataset and the performance of ChatGPT of the AP Chemistry dataset from OpenAI [16]. OpenAI reported ChatGPT achieved a score range of 22%-46%. On the VNHSGE dataset, ChatGPT scored 48%, BingChat scored 52.5%, and our test case achieved the highest score of 67.5%, while a minimum of 33%.

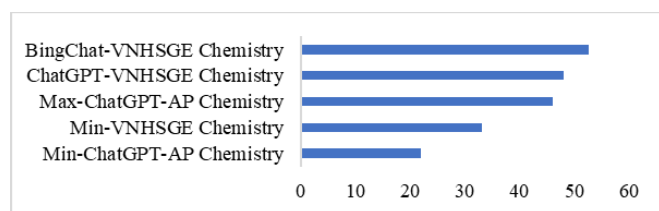


Fig. 4. Performance on VNHSGE dataset and other exams.

D. Comparison to Vietnamese students

To assess the performance of the LLMs, we compare their results with those of Vietnamese students. Table III displays the converted scores of ChatGPT and BingChat, the average score (AVNS), and the score attained by the highest-performing Vietnamese student (MVNS). The average scores of ChatGPT, BingChat, Min, and Max are 4.8, 5.25, 3.3, and 6.75, respectively. The average score of Vietnamese students for 2019-2022 is 5.35, 6.71, 6.63, and 6.7. We can infer that ChatGPT and BingChat have lower scores than the average score of Vietnamese students. However, Max exhibits a better outcome than the average score of Vietnamese students, but it is still lower than the score obtained by the highest-performing Vietnamese student.

TABLE III. SCORES

	ChatGPT	BingChat	Min	Max	AVNS	MVNS
2019	4	5.5	2.75	6.75	5.35	6
2020	4.25	5.75	3	7	6.71	7.75
2021	6.25	5	4	7.25	6.63	7.75
2022	4.75	4.75	3.25	6.25	6.7	8
2023	4.75	5.25	3.5	6.5		
AVG	4.8	5.25	3.3	6.75	6.35	

The scores of ChatGPT, BingChat, Min, and Max are shown in Fig. 5 compared to the scores of Vietnamese students. This graph further reinforces the fact that the performance of ChatGPT and BingChat in the domain of high school Chemistry is still inferior to that of Vietnamese students.

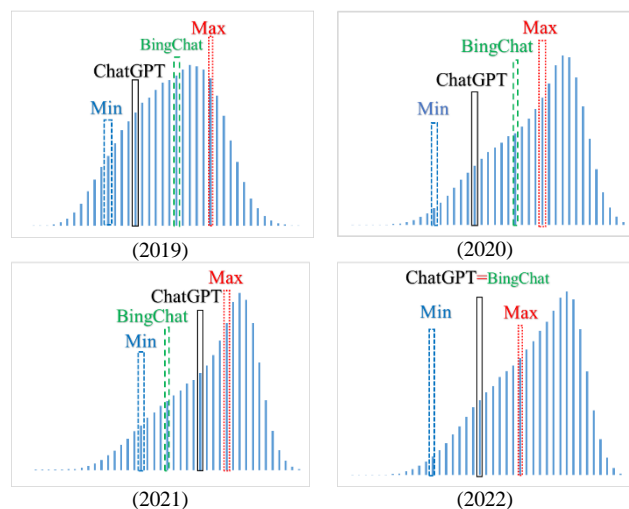


Fig. 5. ChatGPT and BingChat performances and Vietnamese students.

V. DISCUSSION

Language Models, including Large Language Models (LLMs), have the potential to revolutionize education by providing students with personalized and interactive learning experiences. With their ability to process and analyze vast amounts of data, LLMs can provide tailored feedback and adapt to individual students' learning styles. They can also assist in grading and assessing student work, reducing the workload on teachers. However, the results of this study suggest that LLMs such as ChatGPT and BingChat currently have limited ability in answering high-level application questions in Chemistry at the high school level. The accuracy

of their responses is also lower than that of Vietnamese students. This highlights the challenges that LLMs face in adapting to the nuances and complexities of natural language, particularly in specialized domains such as chemistry. Despite these challenges, there is great potential for LLMs to improve educational outcomes, especially in countries where access to quality education is limited. LLMs can provide students with access to high-quality resources and personalized feedback, regardless of their location or socio-economic status. Moreover, LLMs can be trained to recognize and adapt to regional differences in language and culture, making them suitable for use in diverse contexts such as Vietnam. However, to fully realize the potential of LLMs in education, several challenges must be overcome. These include improving the accuracy and reliability of LLMs in specialized domains such as chemistry, developing tools and platforms that facilitate the integration of LLMs into the classroom, and addressing concerns around privacy and data security. In summary, LLMs have the potential to transform education, including in the field of Chemistry in Vietnam and around the world. However, their success will depend on their ability to overcome the challenges and limitations identified in this study, and on the development of supportive policies and infrastructure to ensure their effective integration into the education system.

VI. CONCLUSION

In conclusion, our study evaluated the performance of two LLMs, ChatGPT and BingChat, in answering chemistry questions in Vietnamese high school exams from 2019-2023. The results showed that both LLMs have limitations in answering questions at the application and high application levels, indicating their lack of ability to reason and apply knowledge. Moreover, when comparing the performance of ChatGPT and BingChat, BingChat generally outperformed ChatGPT in terms of accuracy. We also compared the performance of the LLMs with Vietnamese students' scores and found that both ChatGPT and BingChat have lower scores than the average score of Vietnamese students, indicating that LLMs have limitations in replacing human intelligence in the field of chemistry education. However, LLMs still have potential in providing assistance for students and teachers in learning and teaching activities, especially in providing instant feedback and personalized learning experiences. LLMs can also be used to generate questions and materials for practice and assessment. Additionally, LLMs can be improved by incorporating more domain-specific knowledge and enhancing their reasoning and application abilities. Overall, while LLMs show promise in the field of education, they still face challenges and limitations that need to be addressed. Future research can explore ways to improve LLMs' abilities to reason and apply knowledge, as well as investigate their effectiveness in improving student learning outcomes.

REFERENCES

- Chen, L., Chen, P., Lin, Z.: Artificial Intelligence in Education: A Review. *IEEE Access*. 8, 75264–75278 (2020). <https://doi.org/10.1109/ACCESS.2020.2988510>.
- Dao, X.Q., Le, N.B., Nguyen, T.M.T.: AI-Powered MOOCs: Video Lecture Generation. *ACM Int. Conf. Proceeding Ser.* 95–102 (2021). <https://doi.org/10.1145/3459212.3459227>.
- Nguyen, T.M.T., Diep, T.H., Ngo, B.B., Le, N.B., Dao, X.Q.: Design of Online Learning Platform with Vietnamese Virtual Assistant. *ACM Int. Conf. Proceeding Ser.* 51–57 (2021). <https://doi.org/10.1145/3460179.3460188>.
- Devlin, J., Chang, M.-W., Lee, K., Google, K.T., Language, A.I.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv Prepr. arXiv1810.04805*. (2018). <https://doi.org/https://doi.org/10.48550/arXiv.1810.04805>.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., Allen, P.G.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv Prepr. arXiv1907.11692*. (2019).
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21, 1–67 (2020).
- Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Openai, D.A.: Language Models are Few-Shot Learners. *Adv. Neural Inf. Process. Syst.* 33, 1877–1901 (2020).
- Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., Das, P.: Large-Scale Chemical Language Representations Capture Molecular Structure and Properties. *Nat. Mach. Intell.* 4, 1256–1264 (2022).
- Castro Nascimento, C.M., Pimentel, A.S.: Do Large Language Models Understand Chemistry? A Conversation with ChatGPT. *J. Chem. Inf. Model.* (2023). <https://doi.org/10.1021/acs.jcim.3c00285>.
- Xuan-Quy, D., Ngoc-Bich, L., The-Duy, V., Xuan-Dung, P., Bac-Bien, N., Van-Tien, N., Thi-My-Thanh, N., Hong-Phuoc, N.: VHSGE: Vietnamese High School Graduation Examination Dataset for Large Language Models. *arXiv Prepr. arXiv2305.12199*. (2023). <https://doi.org/10.48550/arXiv.2305.12199>.
- Emenike, M.E., Emenike, B.U.: Was This Title Generated by ChatGPT? Considerations for Artificial Intelligence Text-Generation Software Programs for Chemists and Chemistry Educators. *J. Chem. Educ.* (2023). <https://doi.org/10.1021/acs.jchemed.3c00063>.
- Mahjour, B., Hoffstadt, J., Cernak, T.: Designing Chemical Reaction Arrays using phactor and ChatGPT. *Chemrxiv*. 3, (2023). <https://doi.org/10.26434/chemrxiv-2023-2tfdv>.
- Fergus, S., Botha, M., Ostovar, M.: Evaluating Academic Answers Generated Using ChatGPT. *J. Chem. Educ.* 1–4 (2023). <https://doi.org/10.1021/acs.jchemed.3c00087>.
- Leon, A.J., Vidhani, D.: ChatGPT Needs a Chemistry Tutor Too. *ChemRxiv*. (2023). <https://doi.org/10.26434/chemrxiv-2023-qpxz3>.
- White, A.D., Hocky, G.M., Gandhi, H.A., Ansari, M., Cox, S., Wellawatte, G.P., Sasmal, S., Yang, Z., Liu, K., Singh, Y., Peña Ccoa, W.J.: Assessment of chemistry knowledge in large language models that generate code †. (2023). <https://doi.org/10.5281/zenodo.6800475>.
- OpenAI: GPT-4 Technical Report. *arXiv Prepr. arXiv2303.08774*. (2023). <https://doi.org/https://doi.org/10.48550/arXiv.2303.08774>.