Performance Comparison of Large Language Models on VNHSGE English Dataset: OpenAI ChatGPT, Microsoft Bing Chat, and Google Bard

Xuan-Quy DAO

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

Abstract

This paper presents a performance comparison of three large language models (LLMs), namely OpenAI ChatGPT, Microsoft Bing Chat (BingChat), and Google Bard, on the VNHSGE English dataset. The performance of BingChat, Bard, and ChatGPT (GPT-3.5) is 92.4%, 86%, and 79.2%, respectively. The results show that BingChat is better than ChatGPT and Bard. Therefore, BingChat and Bard can replace ChatGPT while ChatGPT is not yet officially available in Vietnam. The results also indicate that BingChat, Bard and ChatGPT outperform Vietnamese students in English language proficiency. The findings of this study contribute to the understanding of the potential of LLMs in English language education. The remarkable performance of ChatGPT, BingChat, and Bard demonstrates their potential as effective tools for teaching and learning English at the high school level.

1 Introduction

Artificial Intelligence (AI) has revolutionized the field of education, by transforming the ways of learning and teaching. One of the most prominent advancements in this domain is the development of large language models (LLMs), such as ChatGPT[1], BingChat[2], and Bard[3](developed by OpenAI, Microsoft, and Google, respectively). The development and continuous improvement of these LLMs have paved the way for AI to be integrated into various educational applications and domains (OpenAI, 2023). These models have advanced conversational abilities, closely resembling human-like interactions. This capability holds great promise for educational settings, including the utilization of virtual assistants, chatbots, and online learning support systems (Kasneci et al., 2023).

Although these models offer tremendous potential for personalized and interactive learning experiences, creating new opportunities for educational innovation and transforming the way we teach and learn, the development and deployment of AI in education require careful consideration of the effectiveness of these AI-powered educational tools in different educational and social contexts (Kung et al., 2023), (Thorp, 2023).

Similar to other parts of the world, LLMs hold great potential in the field of education in Vietnam. However, it is essential to conduct thorough research and evaluation to comprehensively understand their capabilities and limitations in the specific context of Vietnamese education. ChatGPT scored well on the Vietnamese High School Graduation Examination (VNHSGE) (Dao et al., 2023a) and obtained a good score on an English test case (Dao et al., 2023b). However, it is worth noting that ChatGPT has not been officially deployed in Vietnam. Meanwhile, Vietnamese students have access to BingChat and Bard, two other LLMs. Thus, it is crucial to evaluate the capabilities of BingChat and Bard within the context of Vietnamese education.

We focus on evaluating the performance of BingChat and Bard on the VNHSGE English dataset, and comparing them to ChatGPT. The objective of this evaluation is to determine whether BingChat and Bard can serve as potential alternatives to ChatGPT in English education at the high school level. Thus, we ask the following research questions:

- Research Question 1 (RS1): What is the performance of ChatGPT, BingChat, and Bard on the VNHSGE English dataset at the high school level in Vietnam?

- Research Question 2 (RS2): How do the LLMs perform in comparison to Vietnamese students in English language proficiency at

---

[1]https://chat.openai.com
[2]https://www.bing.com/chat
[3]https://bard.google.com

the high school level in Vietnam?

- Research Question 3 (RS3): What potential do LLMs hold for English language teaching and learning at the high school level in the context of Vietnam?

The main contribution of this paper provides a comprehensive evaluation of the performance of LLMs, comparing them to human performance on the same tasks, and identifying the potential applications of LLMs in the context of English language education at the high school level in Vietnam. These findings can inform educators, researchers, and policymakers in making informed decisions regarding the integration of LLMs into the English language curriculum and instructional practices.

2 Related works

2.1 Large Language Models

AI is now capable of understanding and engaging in human-like communication, thanks to recent advancements in LLMs. These breakthroughs have opened up new opportunities for their application in the field of education. LLMs have demonstrated immense potential in areas such as education, content development, and language translation. The two primary architectures of LLMs are BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer). In 2018, Google introduced BERT (Devlin et al., 2018), which has excelled in various natural language processing (NLP) tasks. Developed by OpenAI (Alec et al., 2018), the GPT algorithm was trained on extensive unlabeled text datasets. Building on Google's research, Facebook's RoBERTa (Liu et al., 2019) was introduced, and in 2019, Google released T5 (Raffel et al., 2020). The year 2020 witnessed the creation of GPT-3 by OpenAI (Brown et al., 2020), showcasing exceptional performance in a wide range of NLP tasks. Recently, OpenAI developed GPT-4 (OpenAI, 2023), a text-to-text machine learning system capable of processing both text and image inputs. GPT-4 has demonstrated human-level performance in many professional and academic criteria, although its performance may not match that of humans in all contexts. These advancements highlight the progress made in LLMs and their potential impact in various domains.

2.2 Evaluation of LLMs on English

ChatGPT scored an average of 7.18 on the national high school exam in the Netherlands, which is equivalent to the average score of all students participating in the exams in the Netherlands (de Winter, 2023). In other research, ChatGPT outperformed Vietnamese students by scoring 7.92 in the national high school graduation exam in Vietnam (Dao et al., 2023b). Additionally, ChatGPT/GPT-4 has been shown to be as good as commercial translation products at translation (Jiao et al., 2023), as well as having the ability to perform multilingual tasks (Bang et al., 2023).

3 Methods

3.1 Dataset

The VNHSGE English dataset (Dao et al., 2023c) consists of 2500 multiple-choice questions and comprises a diverse range of exam questions extracted from high school examinations, covering various topics and assessing different linguistic abilities. The dataset includes questions related to pronunciation and stress, vocabulary and grammar, communication skills, reading fill-in-the-blank, and reading comprehension. The evaluation dataset comprises 250 multiple-choice questions corresponding to 5 exams from the academic years 2019, 2020, 2021, 2022, and 2023.

3.2 Prompt

In this study, zero-shot learning was employed, where LLMs were tasked with answering questions directly without any prior data or example questions. For the VNHSGE English dataset D, let {Q, S} denote pairs where Q is the question and S is the ground truth solution. Additionally, let P denote the context of words. The LLM's answer A is determined by

$$A = f(P, Q) \tag{1}$$

where f is the LLM, which takes into account the context P and the question Q. The context P in this case is a specific structure that guides the LLM's response. It instructs the LLM to provide the answer in the following format: { Choice: "A" or "B" or "C" or "D"; Explanation: Explain the answer; The question is: [the actual question] }. By following this structure, the LLM generates its answer A, which can be evaluated and compared to the ground truth solution S.

Figure 1 illustrates the process of prompting LLMs and retrieving the results. In the case of multiple-choice questions from the VNHSGE English dataset, the questions are formatted to align with the expected answer format. The questions are then sent to the LLMs' API.

3.3 Grading

To evaluate the performance of LLMs in answering questions, we assessed the LLM's response by comparing it to the ground truth. Given the question Q, the corresponding ground truth solution S, and the LLM's answer A, we assessed the LLM's response by comparing it to the ground truth solution S:

$$G = g(Q, S, A) \qquad (2)$$

The evaluation process resulted in a binary grading system, where the LLM's answer was classified as correct or incorrect.

The answers of LLMs may be different because they are trained on different datasets. To capture the best and worst cases among the answers of LLMs, we introduced two variables: $\mathrm{LLM_B}$ represents the case with the highest graded answer among the LLMs, while $\mathrm{LLM_W}$ represents the case with the lowest graded answer among the LLMs

$$\begin{cases} \mathrm{LLM_B} = \max(G_i) \\ \mathrm{LLM_W} = \min(G_i) \end{cases} \qquad (3)$$

where G_i represents the grading corresponding to ChatGPT, BingChat, and Bard. These variables allow us to analyze the upper and lower bounds of the performance of LLMs in terms of grading their answers.

4 Results

4.1 Answers of LLMs

In this section, we present the answers provided by ChatGPT, BingChat, and Bard on the VNHSGE English dataset. We provide illustrative examples for different types of questions, including pronunciation and stress, grammar and vocabulary, communication, reading fill-in-the-blank, and reading comprehension. For detailed information about the answers provided by the LLMs, please refer to Section A.

4.2 Performance of LLMs

To answer RS 1, we evaluate the performance of ChatGPT, BingChat, and Bard on the VNHSGE English dataset. The findings provide insights into how accurate and effective these LLMs are in answering questions from the dataset. The contribution of this analysis is a comparative assessment of the performance of the three models, which can guide educators and researchers in selecting the most suitable LLM for English language tasks in the context of high school education.

4.2.1 Question order

Figure 2 illustrates the accuracy of the answers of LLMs according to the order of the questions. The results show that the LLMs models do not provide accurate answers for the first four questions, which are related to phonetics. This may be because models like ChatGPT, BingChat, and Bard were not sufficiently trained in the field of phonetics. LLMs perform well on the remaining portions of the exam, proving that LLMs have received adequate training in English vocabulary, grammar, and reading comprehension.

4.2.2 Performance evaluation

Table 1 presents the performance of LLMs. BingChat demonstrates better results compared to ChatGPT and Bard. The accuracy of $\mathrm{LLM_B}$ is 97.2%, demonstrating the significant potential of LLMs in English education at the high school level.

Figure 3 illustrates the stability of LLMs across the years 2019-2023. The performance indices $\mathrm{LLM_B}$, ChatGPT, and Bard demonstrate relatively stable results, while BingChat and $\mathrm{LLM_W}$ exhibit variations between academic years. The consistent performance of LLMs over the years indicates their ability to maintain stable performance. This demonstrates that LLMs are trained on a large amount of data and can be valuable tools in education.

Table 1: Performance (%)

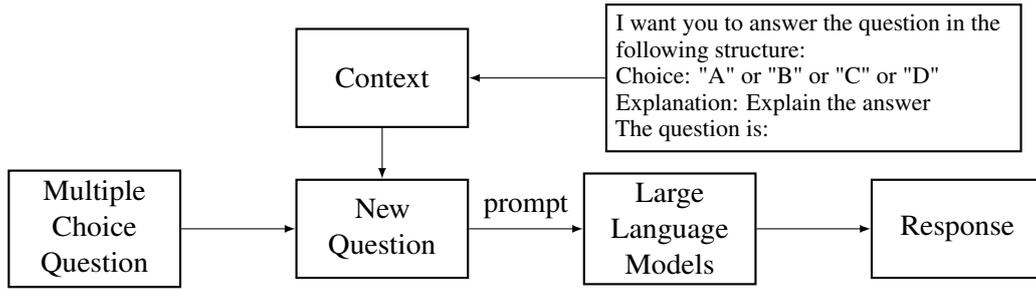|	2019	2020	2021	2022	2023	AVG
ChatGPT	76	86	76	80	78	79.2
BingChat	92	96	86	94	94	92.4
Bard	82	94	82	86	86	86
$\mathrm{LLM_W}$	66	82	68	74	70	72
$\mathrm{LLM_B}$	96	100	94	96	100	97.2

Figure 1: Formatted question and LLMs response.

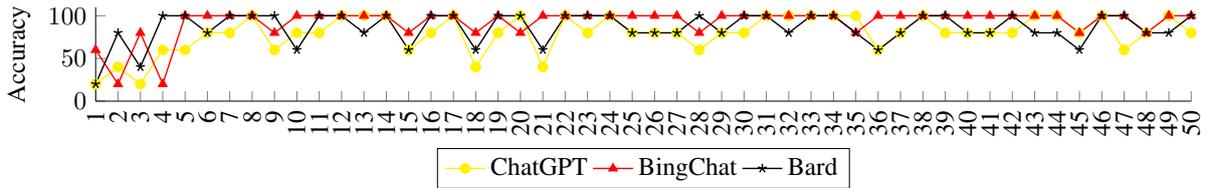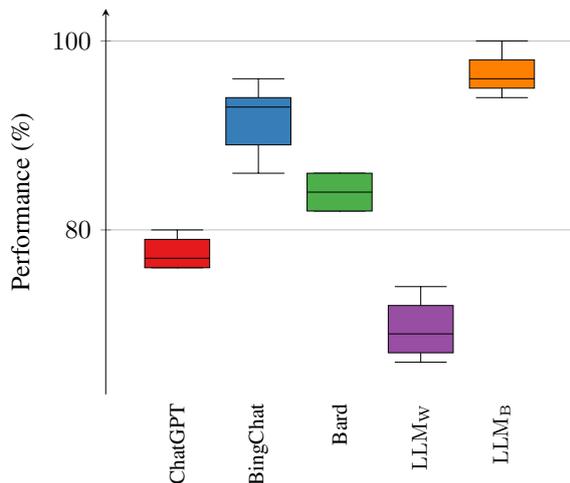Figure 2: Correctness average of ChatGPT, BingChat, and Bard in question order.

Figure 3: Stability of LLMs performance.

4.3 LLMs and Vietnamese students

4.3.1 LLMs perform better in English than Vietnamese students

To answer RS 2, we examine the performance of the LLMs in comparison to Vietnamese students' English language skills. Our aim is to determine whether LLMs possess abilities comparable to those of humans, although this comparison is challenging due to the dissimilar settings. By conducting this comparison, we can evaluate whether LLMs can serve as effective study aids for Vietnamese students in various subject areas.

The score distribution of the students' exams serves as an indicator for evaluating their perfor-

mance. The Vietnam Ministry of Education and Training releases score distribution charts for each subject every year, which help assess the candidates' competence and determine the difficulty level of the exams. By gathering the score distributions (2019, 2020, 2021, 2022), we can compare the performance of LLMs with that of Vietnamese students. To facilitate this comparison, we compare the score of LLMs to the average score (AVS) and the most reached score (MVS) by Vietnamese students. For instance, in 2023, the AVS and MVS were 5.84 and 3.8, respectively (see Appendix section B for comparison in Vietnamese score spectrum). Table 2 shows the score of LLMs on a 10-point scale. Figure 4 shows the effectiveness of LLMs with Vietnamese students.

Table 2: Scores

	2019	2020	2021	2022	2023	AVG
ChatGPT	7.6	8.6	7.6	8.0	7.8	7.92
BingChat	9.2	9.6	8.6	9.4	9.4	9.24
Bard	8.2	9.4	8.2	8.6	8.6	8.6

The results indicate that LLMs exhibit a higher level of English proficiency compared to Vietnamese students. One key advantage of LLMs is their ability to access vast amounts of information and data, including extensive language resources and pre-training on large corpora. This enables them to have a broader knowledge base and a deeper understanding of the English language.
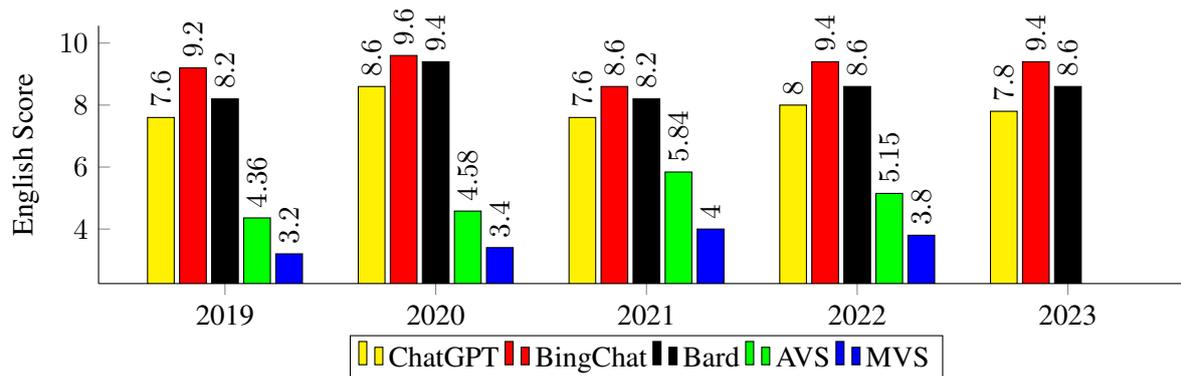
Figure 4: Comparison of the performance of LLMs and Vietnamese students.

Consequently, LLMs demonstrate superior performance in answering questions and providing accurate responses compared to Vietnamese students.

4.3.2 Potential of LLMs in helping Vietnamese students learn English

To answer RS 3, we explore the potential of LLMs for English language teaching and learning at the high school level in Vietnam. By evaluating the performance of the LLMs on the VNHSGE English dataset, the research highlights the capabilities of these models in assisting with language instruction and learning. The findings contribute to identifying the specific areas where LLMs can be leveraged, such as vocabulary acquisition, grammar comprehension, and reading comprehension. The findings of the research can inform the development of educational strategies and materials that incorporate LLMs to enhance English language education in Vietnam.

LLMs can indeed provide valuable assistance to Vietnamese students in learning English. Here are some ways LLMs can be helpful:

- Language practice: LLMs can act as conversational partners, allowing Vietnamese students to practice their English communication skills. Students can engage in dialogue with the model, asking questions, discussing topics, and receiving responses that simulate real conversations.

- Personalized learning: LLMs have the potential to adapt to individual student's needs and provide personalized learning experiences. By analyzing students' performance and identifying areas of improvement, LLMs can generate customized exercises, quizzes, or study materials tailored to each student's specific requirements.

- Language comprehension: LLMs can process and understand English text, including articles, books, and educational resources. Vietnamese students can use LLMs to improve their reading comprehension by getting instant definitions, explanations, and examples of unfamiliar words or phrases.

- Writing support: LLMs can assist students in improving their English writing skills. Students can use LLMs to generate suggestions for sentence structure, grammar, and vocabulary, and even receive feedback on their writing. This can help students enhance their written expression and produce more accurate and coherent English texts.

- Language resources: LLMs can serve as a vast repository of information, providing access to various English learning materials, such as grammar rules, idiomatic expressions, vocabulary lists, and sample essays. Students can leverage these resources to enhance their understanding of English language concepts and improve their overall language proficiency.

5 Conclusion

This study compared the performance of three LLMs (OpenAI ChatGPT, Microsoft Bing Chat, and Google Bard) on the VNHSGE English dataset. The results showed that these LLMs have the potential to be used in education, as they demonstrated high accuracy in answering multiple-choice questions and were consistent in their performance across different years of the dataset.

Additionally, LLMs are significantly superior to Vietnamese students in English language proficiency suggesting that they have the potential to be valuable tools for facilitating English language learning among Vietnamese students. By leveraging the strengths of language models, educators, and students can benefit from personalized and adaptive learning experiences that meet individual needs and enhance overall language proficiency. This highlights the potential of language models to revolutionize English language education and empower Vietnamese students to excel in their language-learning endeavors.

References

Radford Alec, Narasimhan Karthik, Salimans Tim, and Sutskever Ilya. 2018. Improving language understanding with unsupervised learning. *Citado*, 17:1–12.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Xuan-Quy Dao, Ngoc-Bich Le, Xuan-Dung Phan, and Bac-Bien Ngo. 2023a. Can chatgpt pass the vietnamese national high school graduation examination? *arXiv preprint arXiv:2306.09170*.

Xuan-Quy Dao, Ngoc-Bich Le, Xuan-Dung Phan, and Bac-Bien Ngo. 2023b. An evaluation of chatgpt's proficiency in english language testing of the vietnamese national high school graduation examination. *Available at SSRN 4473369*.

Xuan-Quy Dao, Ngoc-Bich Le, The-Duy Vo, Xuan-Dung Phan, Bac-Bien Ngo, Van-Tien Nguyen, Thi-My-Thanh Nguyen, and Hong-Phuoc Nguyen. 2023c. Vnhsge: Vietnamese high school graduation examination dataset for large language models. *arXiv preprint arXiv:2305.12199*.

Joost CF de Winter. 2023. Can chatgpt pass high school exams on english language comprehension. *Researchgate. Preprint*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

H Holden Thorp. 2023. Chatgpt is fun, but not an author.

A Answes of LLMs

This section illustrates the answers of LLMs on the VNHSGE English dataset.

A.1 Pronunciation and stress question

Question: Mark the letter A, B, C, or D on your answer sheet to indicate the word whose underlined part differs from the other three in pronunciation in each of the following questions.

    A. maintain<u>ed</u>

    B. promot<u>ed</u>

    C. requir<u>ed</u>

    D. argu<u>ed</u>

  Solution: Pronounce the ending "ed" Explain: A. maintained /mein'teind/ B. promoted /prə'məutid/ C. required /rI'kwaiə(r)d/

D. say /ˈɑːgjuːd/ How to pronounce the ending "ed": + The ending "ed" is pronounced /t/ when the verb ends in /t∫/, /θ/, /∫/, /s/, /k/, /p/, /f/+ "ed" is pronounced /Id/ when the verb ends in /t/, /d/ + The ending "ed" is pronounced/d/ for the rest of the cases Option B, the underlined part is pronounced /Id/, the rest is pronounced /d/. Choose **B**.

ChatGPT: Question 1: **D**. argued

BingChat: For question 1, the word whose underlined part differs from the other three in pronunciation is **D**. argued.

Bard: The answer is ****A. maintain****.

The other three words have the "ed" pronounced as /d/, but "maintain" has the "ed" pronounced as /t/.

A.2 Grammar and vocabulary questions

Question: Mark the letter A, B, C, or D on your answer sheet to indicate the correct answer to each of the following questions.

Question: He's American,....?

A. won't he

B. didn't he

C. doesn't he

D. isn't he

Solution: Tag question. Explain: The verb in the main clause is "is", the subject is "He". => Negative question tag "isn't he". => He's American, isn't he? Choose **D**.

ChatGPT: **D**. isn't he

BingChat: For question 5, the correct answer is **D**. isn't he.

Bard: The correct answer is **D**. The question is a tag question, which is a type of question that is used to confirm information. The tag question "isn't he?" is used to confirm that the person being referred to is American. The other options are not correct because they do not make sense in the context of the question.

Question: Mark the letter A, B, C, or D on your answer sheet to indicate the underlined part that needs correction in each of the following questions.

Question: It is <u>essential</u> to provide high school students with adequate ca-

reer <u>orientation</u>, so they can make more <u>informative</u> decisions about their future <u>major</u>.

A. essential

B. orientation

C. informative

D. major

Solution: Advanced Vocabulary. Explain: informative (adj): provides a lot of information, has the effect of improving knowledge informed (adj): to have a lot of knowledge or information about something Edit: "informative" => (informed) It is essential to provide high school students with adequate career guidance, so they can make more informed decisions about their future major. Choose **C**.

ChatGPT: It is essential to provide high school students with adequate career orientation, so they can make more informed decisions about their future major. No correction needed.

BingChat: **C**. informative

Bard: The correct answer is (**D**). The word major should be capitalized, as it is referring to a specific academic field of study. The other parts of the sentence are correct.

A.3 Communication question

Question: Mark the letter A, B, C, or D on your answer sheet to indicate the option that best completes each of the following exchanges.

Question 21: Jack and David are talking about taking a gap year. Jack: "I think taking a gap year is a waste of time." David: "....... It gives gap-year takers a lot of valuable experiences."

A. I agree with you

B. I don't quite agree

C. It's right

D. My pleasure

Solution: Explanation: A. I agree with you => not suitable meaning B. I don't quite agree C. That's right => doesn't match the meaning D. My pleasure => inappropriate meaning Choose **B**.

ChatGPT: Jack and David are talking about taking a gap year. Jack: "I think taking a gap year is a waste of time." David: "I don't quite agree. It gives gap-year takers a lot of valuable experiences." **B**. "I don't quite agree.

BingChat: the option that best completes the exchange is **B**. I don't quite agree.

Bard: The correct answer is ****B**. I don't quite agree**. David disagrees with Jack's opinion that taking a gap year is a waste of time. He believes that gap-year takers can gain a lot of valuable experiences, such as travel, work experience, and personal growth.

A.4 Reading Fill-in-the-Blank question

Question: Read the following passage and mark the letter A, B, C, or D on your answer sheet to indicate the correct word or phrase that bestfits each of the numbered blanks from 26 to 30.

THE BEST STUDENTS

According to most professors, mature students are ideal students because they are hardworking and become actively involved in all aspects of the learning process. The majority of mature students have a poor educational (26).....but they manage to do exceptionally well at tertiary level.

Question 26:

    A. development

    B. background

    C. basement

    D. institution

Solution: The majority of mature students have a poor educational (26) background but they manage to do exceptionally well at tertiary level. Choose **B**.

ChatGPT: **B**. background

BingChat: the correct word that best fits the blank is **B**. background.

Bard: The answer must be something that describes the educational history of mature students. Of the above choices, only background describes the educational history of mature students. So the answer is (**B**). Here are the other choices and why they are not the answer: * development: This does not describe the educational history of mature students. * basement: This is not a word that is typically used to describe educational history. * institution: This is a place where education takes place, but it does not describe the educational history of mature students.

A.5 Reading comprehension question

Question: Read the following passage and mark the letter A, B, C, or D on your answer sheet to indicate the correct answer to each of the questions.

Marcel Bich, a French manufacturer of traditional ink pens, was the man turning the ballpoint pen into an item that today almost anyone can afford. Bich was shocked at the poor quality of the ballpoint pens that were available, and also at their high cost. However, he recognised that the ballpoint was a firmly established invention, and he decided to design a cheap pen that worked well and would be commercially successful.

Bich went to the Biro brothers and asked them if he could use the design of their original invention in one of his own pens. In return, he offered to pay them every time he sold a pen. Then, for two years, Bich studied the detailed construction of every ballpoint pen that was being sold, often working with a microscope.

By 1950, he was ready to introduce his new wonder: a plastic pen with a clear barrel that wrote smoothly, did not leak and only cost a few cents. He called it the 'Bic Cristal'. The ballpoint pen had finally become a practical writing instrument. The public liked it immediately, and today it is as common as the pencil. In Britain, they are still called Biros, and many Bic models also say 'Biro' on the side of the pen, to remind people of their original inventors. Bich became extremely wealthy thanks to his invention, which had worldwide appeal. Over the next 60 years his company, Societe Bic, opened factories all over the world and expanded its range of inexpensive products. Today, Bic is as famous for its lighters and razors as it is for its pens, and you can even buy a Bic mobile phone.

Question: The word "which" in paragraph 4 refers to.......

A. factories

B. company

C. invention

D. range

Solution: Explain: The word "which" in paragraph 4 refers to ..... => which replaces the noun that comes before it, "invention". Information: Bich became extremely wealthy thanks to his invention, which had worldwide appeal.. Choose **C**.

ChatGPT: **C**. invention.

BingChat: the word "which" in paragraph 4 refers to **C**. invention .

Bard: The word "which" in paragraph 4 refers to ****C**.invention**.

In paragraph 4, it says:

Bich became extremely wealthy thanks to his invention, which had worldwide appeal. Over the next 60 years his company, Societe Bic, opened factories all over the world and expanded its range of inexpensive products.

The word "which" refers to the invention that Bich created. This invention was the ballpoint pen, which became a practical writing instrument that was affordable and widely used. .

B   Vietnamese score spectrum

Figures 5-9 show the English score spectrum of Vietnamese students in 2019-2023.

Figure 5 (English score spectrum of Vietnamese students in 2019):

Number of Student ($\cdot 10^4$)

Legend: ChatGPT, BingChat, Bard, Vietnamese students

Score	Number of Student
0	0
0.2	0
0.4	5
0.6	33
0.8	123
1	469
1.2	1,324
1.4	2,864
1.6	5,952
1.8	10,310
2	16,722
2.2	23,685
2.4	31,481
2.6	37,599
2.8	42,348
3	45,297
3.2	45,755
3.4	44,476
3.6	41,861
3.8	39,542
4	36,385
4.2	33,410
4.4	30,588
4.6	27,458
4.8	24,979
5	22,630
5.2	20,989
5.4	18,710
5.6	17,283
5.8	15,464
6	14,288
6.2	13,145
6.4	12,173
6.6	11,343
6.8	10,405
7	9,834
7.2	9,274
7.4	8,552
7.6	7,990 (ChatGPT)
7.8	7,612
8	7,108
8.2	6,979 (Bard)
8.4	6,416
8.6	6,045
8.8	5,378
9	4,845
9.2	3,968 (BingChat)
9.4	3,133
9.6	1,976
9.8	939
10	299

Figure 5: English score spectrum of Vietnamese students in 2019.

Figure 6 (English score spectrum of Vietnamese students in 2020):

Number of Student ($\cdot 10^4$)

Legend: ChatGPT, BingChat, Vietnamese students

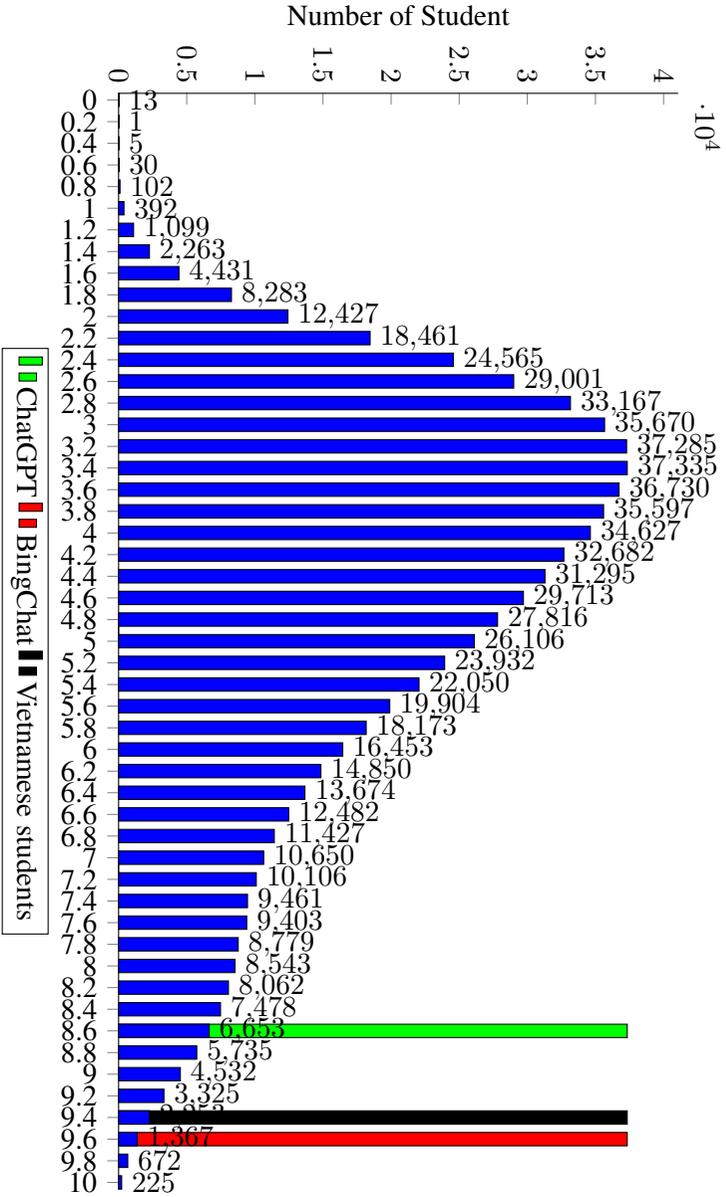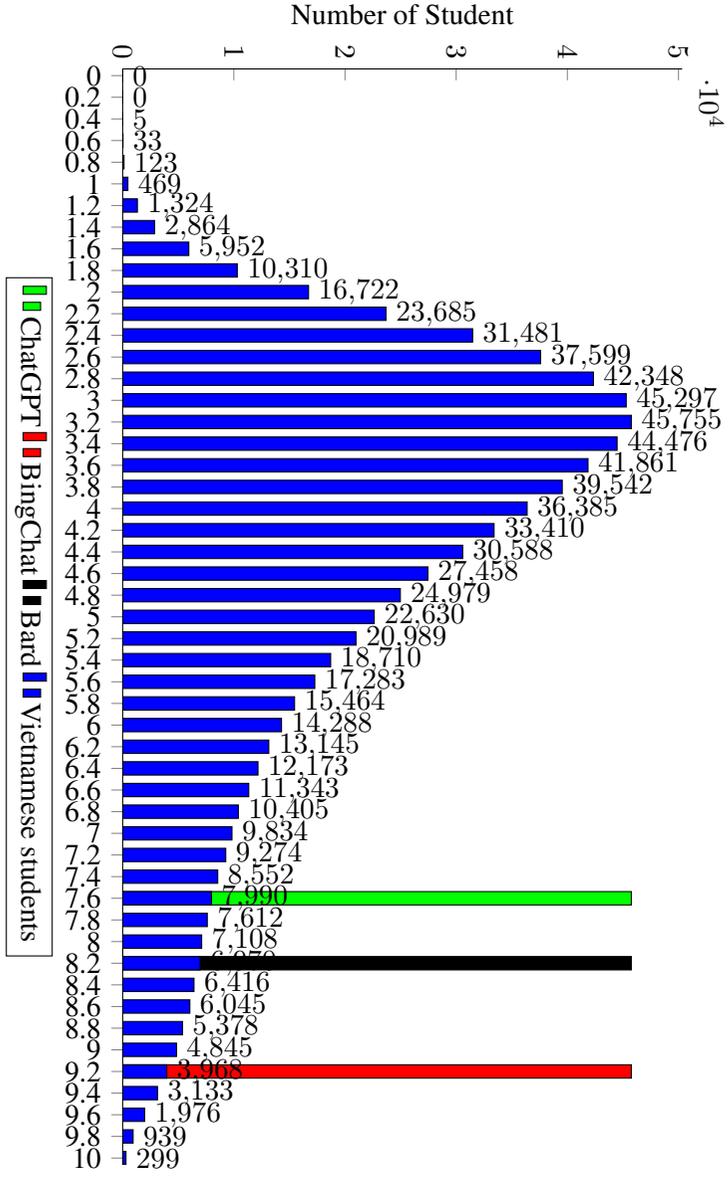| Score	Number of Student
0	13
0.2	1
0.4	5
0.6	30
0.8	102
1	392
1.2	1,099
1.4	2,263
1.6	4,431
1.8	8,283
2	12,427
2.2	18,461
2.4	24,565
2.6	29,001
2.8	33,167
3	35,670
3.2	37,285
3.4	37,335
3.6	36,730
3.8	35,597
4	34,627
4.2	32,682
4.4	31,295
4.6	29,713
4.8	27,816
5	26,106
5.2	23,932
5.4	22,050
5.6	19,904
5.8	18,173
6	16,453
6.2	14,850
6.4	13,674
6.6	12,482
6.8	11,427
7	10,650
7.2	10,106
7.4	9,461
7.6	9,403
7.8	8,779
8	8,543
8.2	8,062
8.4	7,478
8.6	6,653 (ChatGPT)
8.8	5,735
9	4,532
9.2	3,325
9.4	2,853 (BingChat)
9.6	1,367
9.8	672
10	225

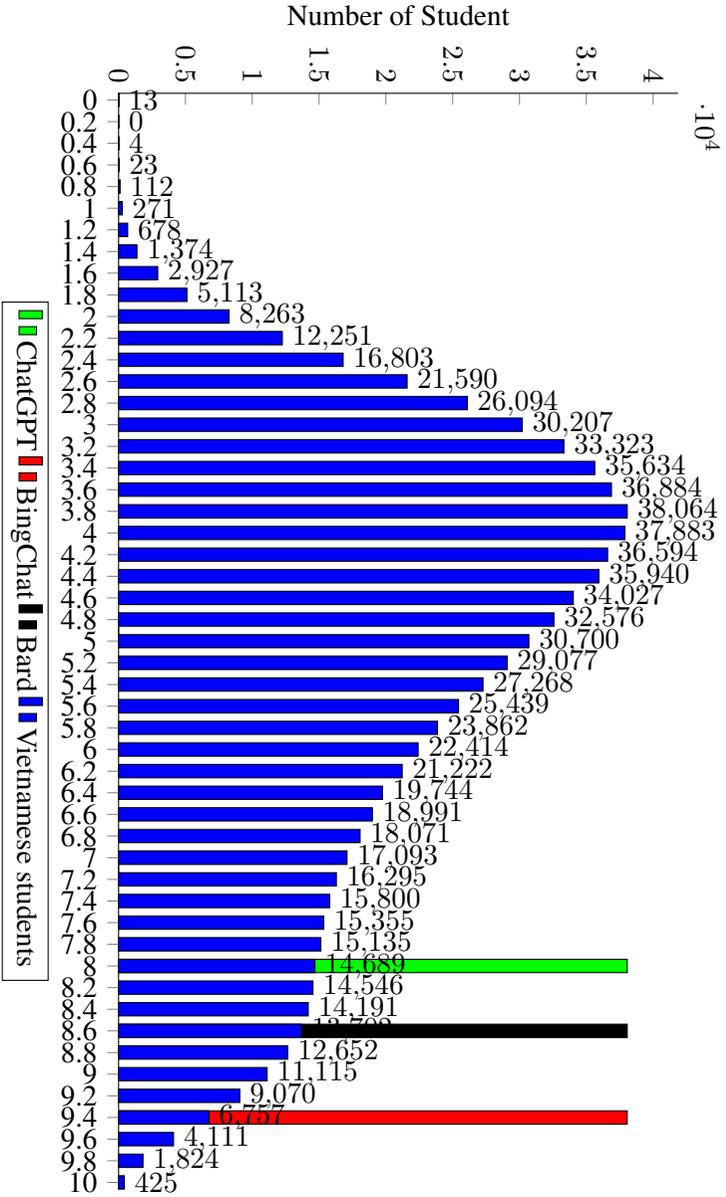Figure 6: English score spectrum of Vietnamese students in 2020.

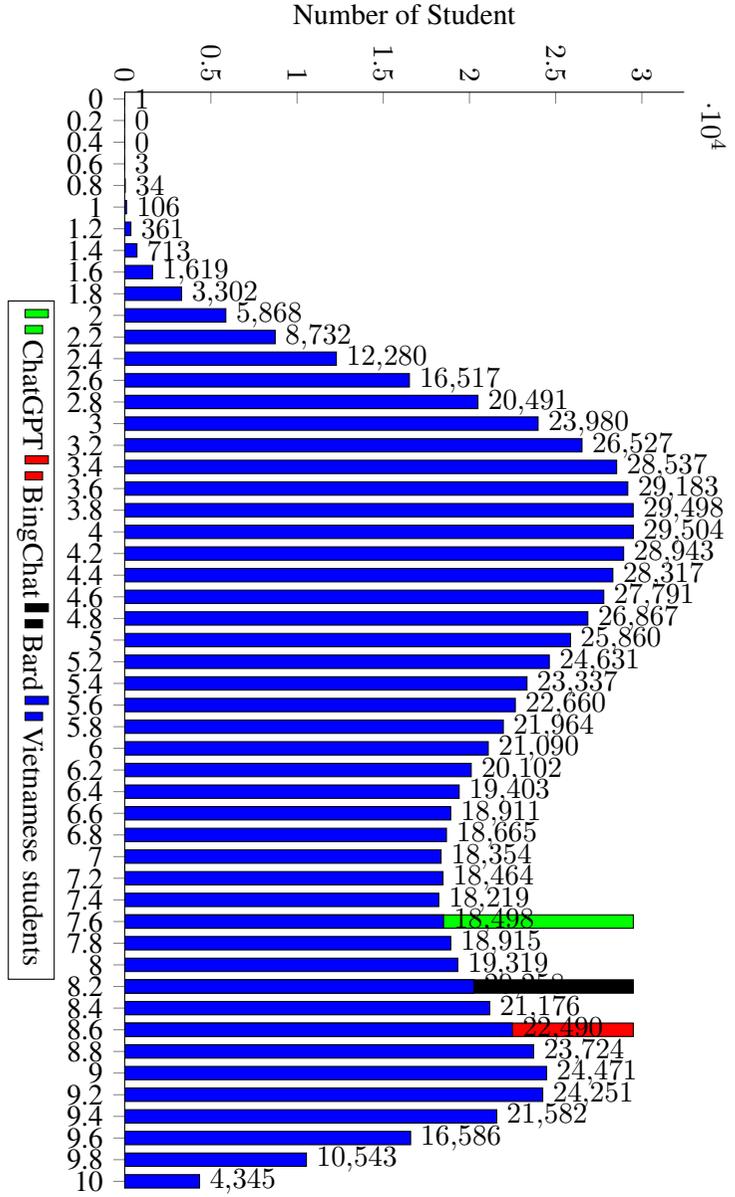Figure 8: English score spectrum of Vietnamese students in 2022.

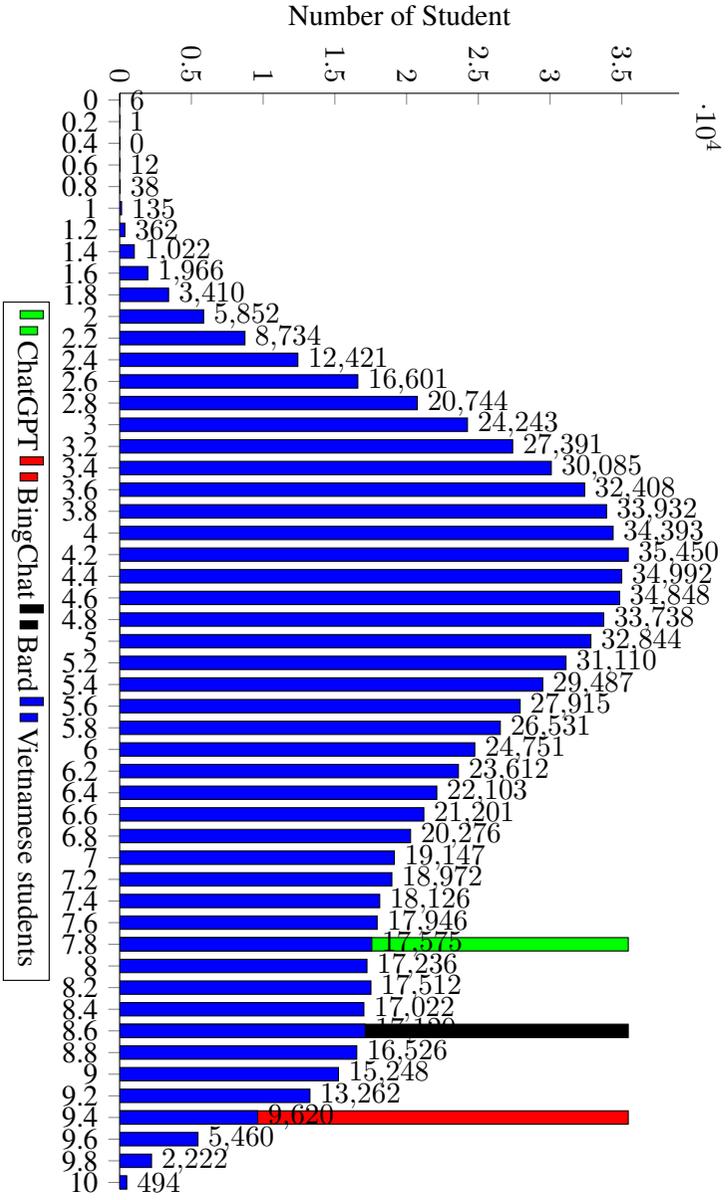Figure 7: English score spectrum of Vietnamese students in 2021.

Figure 9: English score spectrum of Vietnamese students in 2023.