

# LLMs Performance on Vietnamese High School Biology Examination

**Xuan-Quy Dao**

School of Engineering, Eastern International University, Binh Duong, Vietnam

E-mail: [quy.dao@eiu.edu.vn](mailto:quy.dao@eiu.edu.vn)

ORCID iD: <https://orcid.org/0000-0001-5415-7538>

**Ngoc-Bich Le\***

School of Biomedical Engineering, International University, Vietnam National University HCM City, HCM City, Vietnam

E-mail: [lnbich@hcmiu.edu.vn](mailto:lnbich@hcmiu.edu.vn)

ORCID iD: <https://orcid.org/0000-0001-7431-0157>

\*Corresponding Author

Received: 17 May, 2023; Revised: 25 June, 2023; Accepted: 28 July, 2023; Published: 08 December, 2023

**Abstract:** Large Language Models (LLMs) have received significant attention due to their potential to transform the field of education and assessment through the provision of automated responses to a diverse range of inquiries. The objective of this research is to examine the efficacy of three LLMs - ChatGPT, BingChat, and Bard - in relation to their performance on the Vietnamese High School Biology Examination dataset. This dataset consists of a wide range of biology questions that vary in difficulty and context. By conducting a thorough analysis, we are able to reveal the merits and drawbacks of each LLM, thereby providing valuable insights for their successful incorporation into educational platforms. This study examines the proficiency of LLMs in various levels of questioning, namely Knowledge, Comprehension, Application, and High Application. The findings of the study reveal complex and subtle patterns in performance. The versatility of ChatGPT is evident as it showcases potential across multiple levels. Nevertheless, it encounters difficulties in maintaining consistency and effectively addressing complex application queries. BingChat and Bard demonstrate strong performance in tasks related to factual recall, comprehension, and interpretation, indicating their effectiveness in facilitating fundamental learning. Additional investigation encompasses educational environments. The analysis indicates that the utilization of BingChat and Bard has the potential to augment factual and comprehension learning experiences. However, it is crucial to acknowledge the indispensable significance of human expertise in tackling complex application inquiries. The research conducted emphasizes the importance of adopting a well-rounded approach to the integration of LLMs, taking into account their capabilities while also recognizing their limitations. The refinement of LLM capabilities and the resolution of challenges in addressing advanced application scenarios can be achieved through collaboration among educators, developers, and AI researchers.

**Index Terms:** ChatGPT, Microsoft Bing Chat, Google Bard, large language models, biology education.

## 1. Introduction

Several studies have investigated the use of artificial intelligence (AI) in education. Chen et al. [1] focused on the application of AI to help teachers manage administrative activities more effectively and adapt lesson plans to the unique demands of each student, resulting in encouraging improvements in education's overall quality. Dao et al. [2] integrated text-to-speech and speech-driven-face technology to automatically create video lectures using the instructor's voice and facial expressions, lowering workload and boosting learner engagement in online learning settings. Nguyen et al. [3] proposed an online learning platform that included a Vietnamese virtual assistant to help instructors with lesson delivery and student assessment, combining presentation slides with the instructor's face, a synthesized voice, and simple editing tools without the need for video capture.

Large Language Models (LLMs), a type of AI system, are trained on massive amounts of text data to understand natural language and generate human-like responses. These models have become a potent tool for building chatbots with applications in a variety of fields, including education. LLMs have the potential to revolutionize the way we interact with technology and access information. The potential of LLMs in jobs like content creation, language

translation, and instructional support is enormous. Notable developments in LLMs include the release of BERT [4] by Google in 2018, its extension RoBERTa [5] by Facebook in 2019, and T5 [6] by Google researchers in 2019. The 2020 release of OpenAI's GPT-3 [7] received recognition for its outstanding results in a variety of NLP tasks, even with a small amount of training data.

The potential for LLMs to revolutionize biology education is a topic of interest. However, a comprehensive understanding of their competence and limitations across different question types and difficulty levels poses a significant challenge. Given that biology education involves the acquisition of factual knowledge, understanding of context, and the ability to apply critical thinking, it is crucial to assess the proficiency of LLMs in addressing these complexities. Given the aforementioned difficulties, the objective of this study is to conduct a thorough evaluation of ChatGPT<sup>1</sup>, Microsoft's Bing Chat<sup>2</sup> (BingChat), and Google Bard<sup>3</sup> using the VNHSGE biology dataset [8]. Our objective is to provide insights that facilitate the effective integration of different models into educational platforms and assessment designs by highlighting their strengths and limitations. The primary objective is to make a valuable contribution toward the advancement of customized LLM programs that are specifically designed to address the complexities of biology education. This will ultimately result in the promotion of a more enhanced and enriched learning experience for students.

The topic of this article focuses on assessing LLMs' ability in relation to biology tests. The study provides various contributions: (1) a thorough analysis of three cutting-edge LLMs, ChatGPT, BingChat, and Bard in the context of high school biology education in Vietnam; (2) a comparison of the performance of ChatGPT, BingChat, and Bard with Vietnamese students; and (3) a thorough investigation of the advantages and drawbacks of using LLMs in the field of biology education in Vietnam. Through these contributions, this research highlights the issues that must be resolved for proper integration while illuminating the potential of LLMs to improve biology education.

## 2. Related Work

### 2.1. Large Language Models

ChatGPT, a language model created by OpenAI and based on the GPT-3.5 architecture, has the ability to provide human-like responses in natural language. It has the potential to be used in intelligent tutoring programs that enable assessment activities, offer tailored feedback, and automate grading procedures. BingChat, based on the GPT-4 architecture<sup>4</sup>, acts as a chatbot function within the Bing search engine and has a lot of potential as a teaching tool because it can deliver factual information. Google Bard is an experimental, conversational AI chat service developed by Google that was initially powered by Google's Language Model for Dialogue Applications (LaMDA)<sup>5</sup> and later upgraded to use Google's most advanced LLM, PaLM. These AI-powered solutions have the potential to revolutionize education by providing trustworthy and accessible educational support to both teachers and students.

### 2.2. Biology datasets for evaluation of LLMs

Due to their thorough training with enormous volumes of data, LLMs have outstanding natural language comprehension abilities. They are therefore excellent candidates for upholding academic and professional standards, especially in the area of biology education. For LLMs working in the field of biology, existing datasets like BioASQ [9] and SciQ [10] provide challenges. When it addresses tasks like question-answering and biomedical semantic indexing, BioASQ generally concentrates on the medical sector as opposed to the biological one. The SciQ dataset, on the other hand, utilizes multiple-choice questions to assess LLMs' understanding of and rationale for biological science principles. The dataset presents a considerable challenge for LLMs because it contains 13,679 science exam questions from disciplines including biology, chemistry, and physics as well as supporting data for the correct responses. The biology dataset offered in MMLU dataset [11] also includes material from high school and college biology courses, covering a variety of topics like natural selection, heredity, the cell cycle, the Krebs cycle, cellular structure, molecular biology, and ecology. This dataset is used to evaluate the precision of LLMs in certain particular biological domains. A multimodal technique has been introduced in ScienceQA dataset [12], with a focus on molecular and cellular biology. With matching lectures, explanations, and photos captioned, it has approximately 21,000 multiple-choice queries. In comparison to earlier datasets, this dataset offers a wide variety of research topics and annotations, which helps to create a broader domain diversity. Despite LLMs' promise to revolutionize education, they still need to do better in certain areas, like biology. On the AP Biology dataset, ChatGPT-3.5 achieves an accuracy range of 62% to 85%, according to the GPT-4 Report by OpenAI [13]. This shows that although LLMs like ChatGPT have a promising future in education, more work needs to be done to improve their accuracy and performance in biologically-related activities. VNHSGE biology dataset [8] encompasses information derived from the Vietnamese National High School Graduation Examination. This comprehensive dataset encompasses a wide array of biological topics, including genetic principles, population genetics, genetic applications, human genetics, evolution, ecology, plant organismal biology, and numerous

<sup>1</sup> <https://chat.openai.com/>

<sup>2</sup> <https://www.bing.com/chat>

<sup>3</sup> <https://bard.google.com/>

<sup>4</sup> Confirmed: the new Bing runs on OpenAI's GPT-4 | Bing Search Blog

<sup>5</sup> Google AI updates: Bard and new AI features in Search (blog.google)

other facets of the field. It is essential to note that this dataset poses a formidable challenge to large language models due to the intricate nature of inference tasks and the complexity of the Vietnamese language.

### 2.3. Evaluation of LLMs on biology

The study conducted by Agathokleous et al. [14] examines the effects and investigates the practical uses and consequences of AI LLMs in the fields of biology and environmental science. The article examines the potential advantages of LLMs in diverse domains, including education, research, publishing, outreach, and societal translation. The paper extends its analysis by presenting illustrative instances via interactive sessions with LLM, thereby showcasing its capacity to streamline intricate undertakings. In a separate scholarly investigation, Tong et al. [15] examine the ramifications of ChatGPT within the domain of Synthetic Biology. This study investigates the potential of utilizing ChatGPT for the purpose of comprehending research patterns and enhancing research endeavors within this particular field. The paper acknowledges the dual nature of the sentiments surrounding the impact of LLMs on society and research, encompassing both enthusiasm and apprehension. In an alternative perspective, Kumar et al. [16] analyze the usefulness and advantages of ChatGPT in the realm of academic writing within the field of biomedical sciences. The evaluation encompasses the measurement of various aspects, including the speed of response, the caliber of content, the dependability, and the level of novelty exhibited in the text generated by ChatGPT. The research acknowledges the various strengths and limitations associated with the use of ChatGPT in the context of academic writing. Finally, Shue et al. [17] propose an iterative model that aims to refine instructions for guiding a chatbot, specifically ChatGPT, in the generation of code for bioinformatics data analysis tasks. This study evaluates the viability of implementing this model in various bioinformatics domains and examines the practical considerations and constraints associated with utilizing such a model for chatbot-supported bioinformatics instruction. The papers under review collectively highlight the significant influence of LLMs on different fields such as biology, environmental science, synthetic biology, academic writing in biomedical sciences, and bioinformatics. The authors emphasize the possible advantages and difficulties linked to the incorporation of LLMs within these domains. Through the execution of our research, our objective is to make a scholarly contribution to the ongoing discussion pertaining to the responsible and efficient incorporation of LLMs, namely ChatGPT, Bing Chat, and Bard, within the realms of biology education.

## 3. Methods

Fig.1 illustrates the evaluation process of LLMs on the VNHSGE biology dataset. Firstly, the Vietnamese National High School Graduation Examination biology test is converted into a text format so that LLMs can understand it, with mathematical formulas and symbols being converted into LaTeX format. Next, the questions in text format are sent to the LLMs API via the zero-shot method. Finally, the LLMs' responses are scored, analyzed, and evaluated.

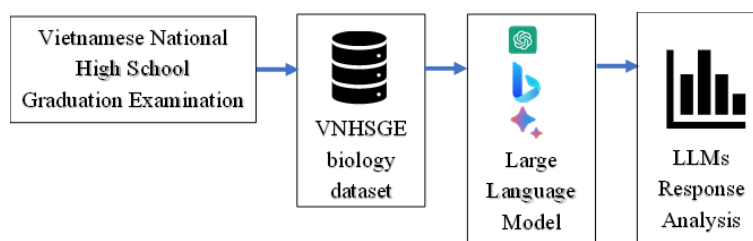


Fig. 1. Evaluation of LLMs on VNHSGE biology dataset

### 3.1 Dataset

The biology questions in the VNHSGE dataset [8] are exceedingly challenging and intricate, necessitating a thorough understanding of biology in order to provide adequate answers. The VNHSGE dataset tests students' biology knowledge and abilities with a variety of questions of varying complexity. The dataset's questions are categorized as "knowledge" (easy), "comprehension" (intermediate), "application" (difficult), and "high application" (very difficult). The dataset is intended to be made up of 75% theoretical questions and 25% exercises, with a focus on higher-order thinking abilities and applications making up 30% of the questions and knowledge and comprehension levels being assessed by the remaining 70%. The dataset includes a range of questions, with an emphasis on the capacity for calculation and inference. While comprehension-level questions only call for one to three steps of deductive reasoning, knowledge-level questions necessitate a thorough understanding of biology. Application-level questions concentrate on issues including genetic principles, population genetics, inheritance mechanisms, and mutation, requiring the synthesis of knowledge. High application-level questions provide another level of complexity because they require on extensive analysis and problem-solving abilities. Due to the complex nature of the VNHSGE biology dataset, LLMs encounter substantial challenges in comprehending and answering questions that cover a range of biology topics correctly.

### 3.2 Prompt

In this investigation, zero-shot evaluation was employed to gauge the effectiveness of LLMs in the evaluation of the VNHSGE dataset denoted as **D**. This dataset comprises a collection of question-answer pairs, designated as **Q** and **S**, respectively, where **S** represents the ground truth solutions. The contextual information represented by the term **P** is defined. LLMs' answer **A** is computed according to the formula presented below:

$$A = f(P, Q) \quad (1)$$

Here, the function **f** is embodied by LLMs and takes into consideration both the context **P** and the question **Q**. The context **P**, in this instance, is a specific structure that guides the response of LLMs. It instructs LLMs to generate 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 adhering to this format, LLMs generates its response **A**, which can subsequently be assessed and compared against the ground truth solution **S**. In the case of multiple-choice questions sourced from the VNHSGE dataset, the questions are adapted to align with the expected answer format and are then sent to the LLMs' API.

### 3.3 Grading

To assess the performance of LLMs in responding to questions, we conducted an evaluation by comparing LLMs' responses to the ground truth solution. The evaluation process was executed through a binary grading system, whereby the response provided by LLMs was categorized as either correct or incorrect. The ground truth solution **S** for each question **Q** was determined by a human expert. The response **A** generated by LLMs was subsequently compared to the ground truth solution using the subsequent equation:

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

## 4. Experimental Results and Discussion

### 4.1 Performance evaluation

Fig.2a illustrates LLMs performance on VNHSGE biology dataset while Fig.2b shows their consistent responses. It is observed that Bard distinguishes itself with a mean performance score of 69.5% on the VNHSGE dataset, suggesting a tendency to produce a greater number of accurate responses to biology questions compared to other models. This observation implies that Bard possesses a comprehensive understanding of concepts related to biology and is capable of delivering precise and reliable responses within this field. Meanwhile, the performance of BingChat is notable, as it exhibits a robust accuracy rate of 69.0%. It is worth mentioning that BingChat consistently demonstrates a commendable performance in addressing inquiries related to biology, as evidenced by its relatively limited range of performance, spanning from 65.0% to 72.5%. The attribute of consistency is highly advantageous for an educational tool. Another finding is that the performance of ChatGPT, despite exhibiting the lowest mean performance of 58%, managed to attain a commendable degree of accuracy. It is noteworthy to mention that ChatGPT exhibited a marginally superior level of consistency in comparison to BingChat. This implies that ChatGPT could potentially serve as a dependable choice for educational applications, despite exhibiting a comparatively lower average accuracy. Furthermore, one intriguing observation pertains to the notable variability observed in Bard's responses, which span a range from 60.0% to 77.5%. Although Bard attained the highest average accuracy, the presence of variability suggests that its responses may occasionally lack consistency. The observed variability in LLM responses could potentially be an aspect that warrants further attention, especially in relation to the dependability of these responses within educational contexts.

In general, the findings of this study suggest that the three LLMs, namely Bard, BingChat, and ChatGPT, exhibit promise for educational utilization within the domain of high school biology. Bard and BingChat exhibit commendable performance with elevated mean accuracies, whereas ChatGPT, while marginally less accurate on average, showcases commendable consistency. Nevertheless, it is imperative to take into account the balance between precision and uniformity when choosing an LLM for educational objectives, as both aspects play a role in the dependability of these models in supporting students and instructors in the field of biology education. Additional refinement and optimization efforts have the potential to improve the performance and reliability of these LLMs in the context of educational applications.

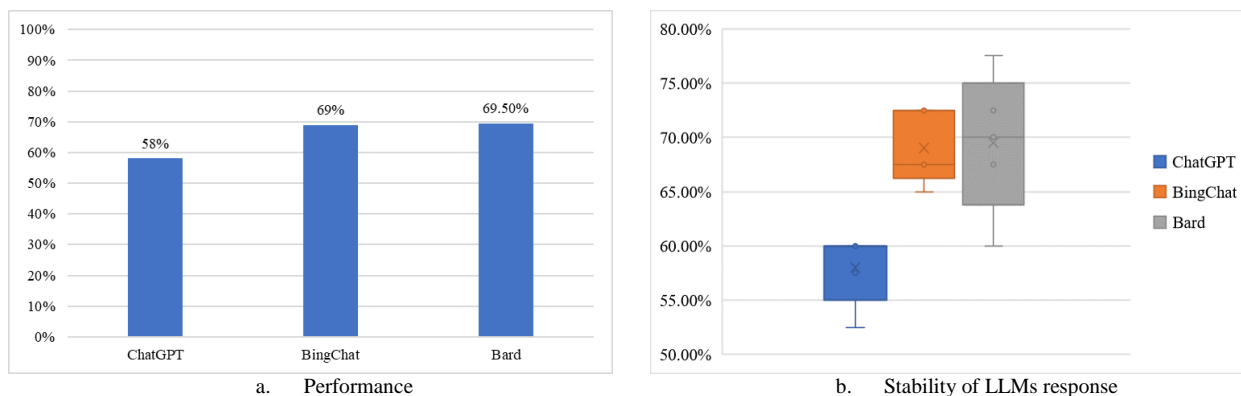


Fig. 2. Performance of LLMs on VNHSGE biology dataset

#### 4.2 LLMs capabilities in different question levels

An essential component of this study involved examining the efficacy of various LLMs in addressing challenging-level inquiries. The objective of the analysis was to evaluate the efficacy of ChatGPT, BingChat, and Bard in addressing various categories of challenging inquiries. The findings of the analysis are succinctly presented in Fig.3b, which illustrates the proportion of accurate responses across different categories of challenging questions for ChatGPT, BingChat, and Bard. The taxonomy of inquiries encompasses Knowledge-based inquiries, Comprehension-based inquiries, Application-based inquiries, and High Application-based inquiries (refer to Fig.3a). The following are the principal discoveries:

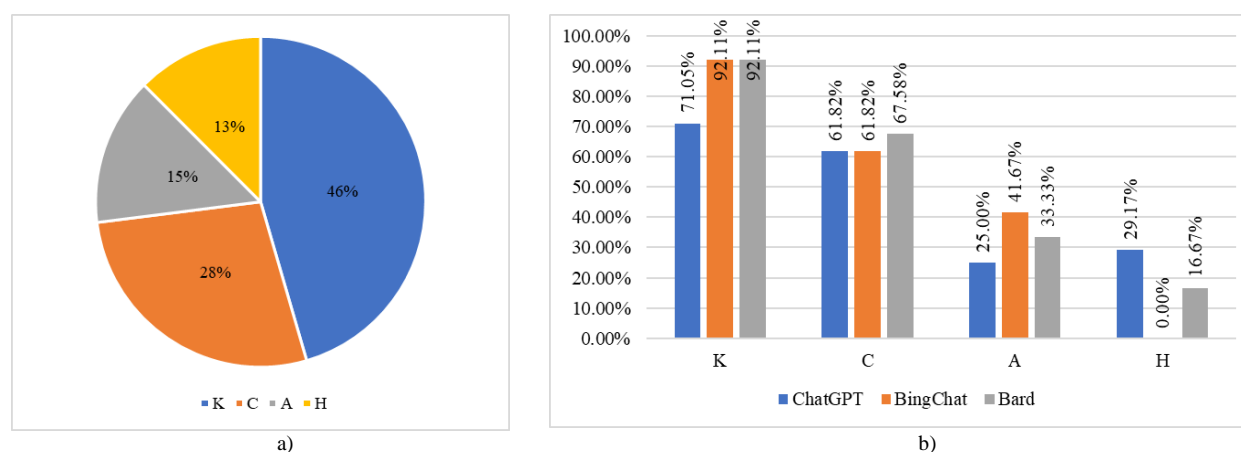


Fig. 3. a) Question levels proportion, and b) Performance of LLMs on question levels

First, knowledge questions are commonly associated with the retrieval of factual information. Both BingChat and Bard exhibited a significantly higher level of accuracy (92.11%) in responding to these inquiries when compared to ChatGPT (71%). The reason behind this is that BingChat and Bard use extensive training data and painstaking fine-tuning to provide accurate and reliable information. The models may have been exposed to a variety of real facts, improving their capacity to retrieve reliable information. Students seeking succinct explanations and definitions can use BingChat and Bard to quickly and accurately answer knowledge-based questions. However, these sources may not always provide complete explanations or contextual information, limiting their use in complex dialogue.

Second, comprehension questions pertain to the process of comprehending and interpreting provided information. All three models demonstrated comparable levels of accuracy in answering these questions. ChatGPT and Bard achieved an accuracy rate of approximately 61.82%, while BingChat slightly outperformed them in this regard. The similar efficacy of all three models in facilitating students' comprehension and interpretation of biology-related information indicates their overall capacity to grasp and elucidate concepts. This suggests that students have the option to refer to any of these models as a resource for understanding the content of a biology text, diagram, or concept. Educators have the opportunity to integrate these models into their instructional materials, thereby offering students supplementary explanations and alternative viewpoints, thereby augmenting their comprehension and fostering their critical thinking abilities.

Third, the application questions, which necessitate the utilization of acquired knowledge in specific scenarios, were found to be more difficult. Among the evaluated systems, BingChat demonstrated the most notable accuracy within this specific category, achieving a percentage of 41.67%. Subsequently, Bard attained a lower accuracy rate of 33.33%, while ChatGPT exhibited the lowest accuracy of 25.0%. This implies that although BingChat and Bard exhibit a greater capacity for the application of knowledge in practical situations, all models encounter challenges in this regard.



Finally, high-application questions necessitate the utilization of advanced analytical skills and the capacity to integrate knowledge in intricate situations. In this category, Bard demonstrated the highest level of accuracy, achieving a rate of 16.67%. Following closely behind was ChatGPT with a rate of 29.1%. Conversely, BingChat did not produce any correct responses within this category. This underscores the notable difficulties that the models encounter when confronted with inquiries that necessitate advanced cognitive abilities and the synthesis of ideas. It is understandable that AI models face inherent difficulties when confronted with high-application questions that necessitate intricate reasoning and the synthesis of ideas. The comparatively lower rates of accuracy observed within this particular category suggest that the models under consideration face challenges when confronted with tasks that necessitate higher-order cognitive skills, such as critical analysis, hypothesis formulation and testing, and the integration of knowledge from multiple disciplines. In order to enhance performance in this area, it is recommended that future advancements prioritize the augmentation of the models' reasoning capabilities, comprehension of context, and capacity to establish coherent connections between disparate pieces of information.

In summary, it is imperative to comprehend the merits and limitations of BingChat, Bard, and ChatGPT in relation to their ability to effectively address various categories of biology inquiries, as this knowledge is essential for their optimal utilization within educational environments. BingChat and Bard demonstrate proficiency in delivering accurate information and facilitating practical implementation. Conversely, ChatGPT distinguishes itself through its adaptability and reliability. Nevertheless, it is important to acknowledge that all models encounter difficulties when presented with challenging inquiries, underscoring the imperative for continuous advancements in artificial intelligence and the comprehension of natural language in order to effectively address the requirements of intricate educational assignments.

#### 4.3 Comparison to other exam

The evaluation of ChatGPT, BingChat, and Bard on the VNHSGE dataset not only yields valuable insights into their respective capabilities within the VNHSGE context but also presents an opportunity to compare their performance with ChatGPT's performance on the AP biology dataset provided by OpenAI [17]. Fig.4 presents a comprehensive comparison of the outcomes obtained by the models on the VNHSGE dataset, along with the performance of ChatGPT on the AP biology dataset. The score range previously determined by OpenAI for GPT-3.5/GPT-4 on the AP biology dataset is situated within the interval of 62% to 85%.

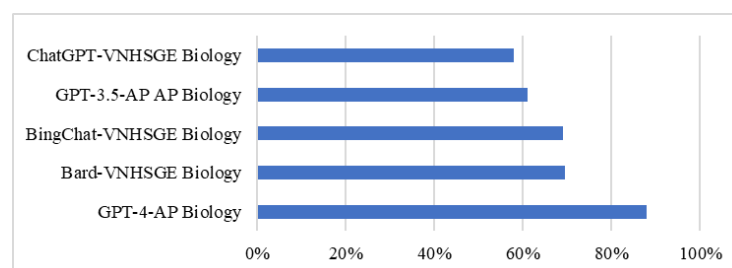
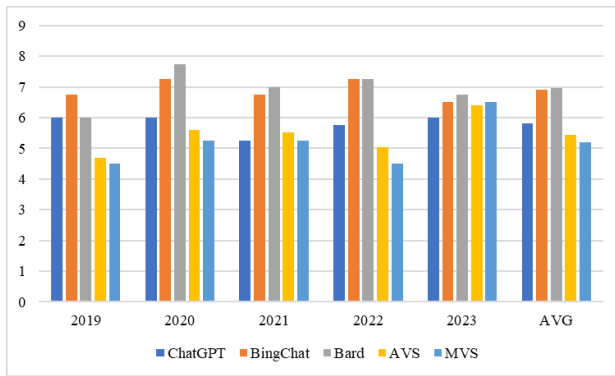


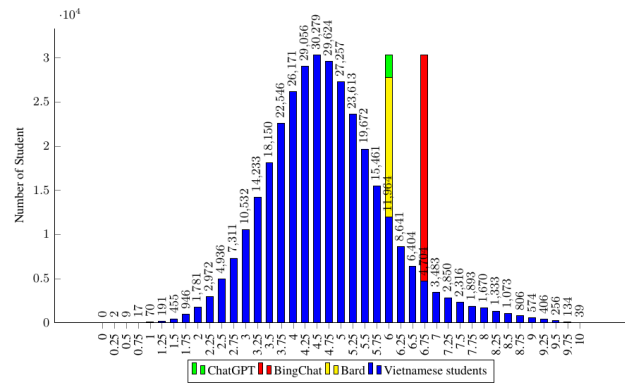
Fig. 4. Performance of LLMs on VNHSGE dataset and other exams.

#### 4.4 Comparison to Vietnamese students

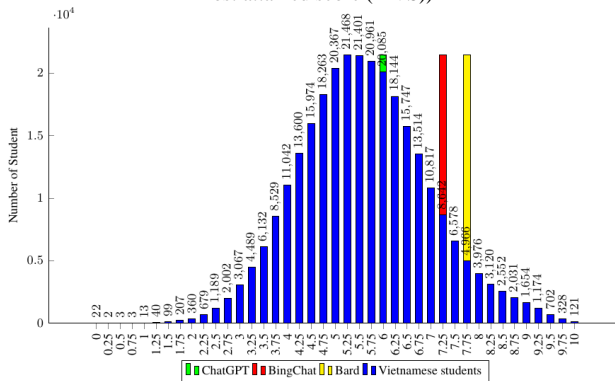
A comparison between the LLMs' scores and those of Vietnamese students was conducted in order to evaluate how well the LLMs performed in the setting of high school biology education in Vietnam. Fig.5 analyzes and displays the converted scores from ChatGPT, BingChat, and Bard together with the average score (AVS) and the most attained score (MVS). The findings show that ChatGPT performed better than both AVS and MVS overall. This demonstrates that ChatGPT performs better than the typical scores obtained by Vietnamese students. But it is remarkable that across all the years considered, BingChat and Bard continuously surpassed AVS and MVS. This demonstrates the potential application of LLMs, particularly BingChat and Bard, in Vietnam's high school biology education. The results highlight the positive potential for incorporating cutting-edge language models into the educational system. Teachers can improve their students' learning experiences by utilizing the features of BingChat and Bard, giving them access to a variety of information and tools. These LLMs could help students develop their critical thinking, problem-solving, and comprehension abilities in the area of biology.



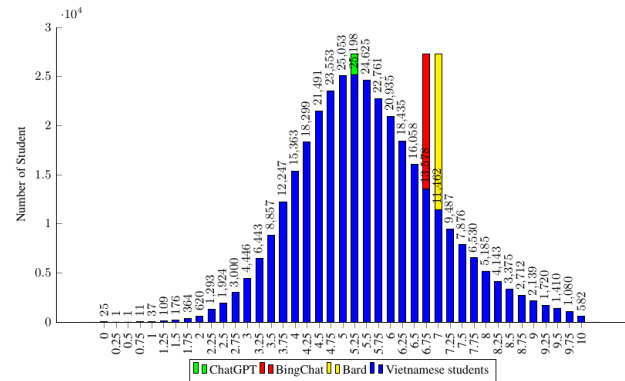
a. LLMs and Vietnamese students (the average score (AVS) and the most attained score (MVS))



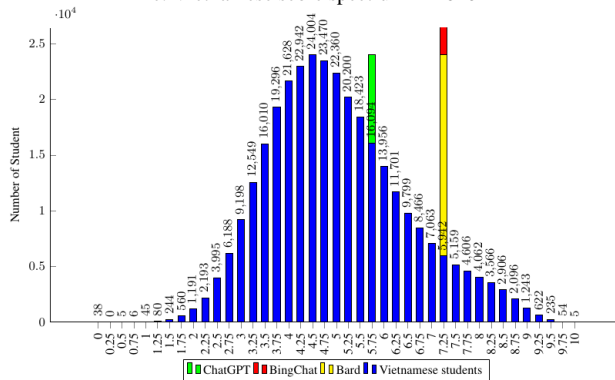
b. Vietnamese score spectrum in 2019



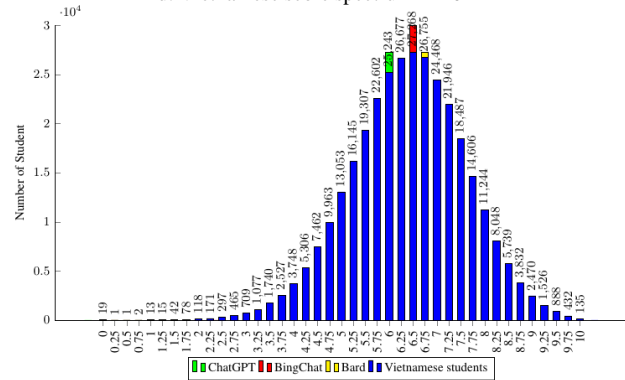
c. Vietnamese score spectrum in 2020



d. Vietnamese score spectrum in 2021



e. Vietnamese score spectrum in 2022



f. Vietnamese score spectrum in 2023

Fig. 5. Performance comparison of LLMs and Vietnamese students

#### 4.5 Suggestion on using the right LLM for a certain circumstance

Table 1 shows the evaluation of the academic performance of Vietnamese students and different LLMs - ChatGPT, BingChat, and Bard - across varying levels of questioning, offering insights into their proficiency and limitations. These results are valuable for users when querying information from LLMs. BingChat and Bard are the best choices for the Knowledge level, while Bard is a reasonable choice at the Comprehension level. However, none of the LLMs are reasonable choices for questions at the Application and High Application levels. The selection of LLMs should be aligned with the levels of questions when designing educational platforms. The limitations of LLMs in addressing intricate application questions should be acknowledged, emphasizing the role of human guidance. Therefore, it is not possible for LLMs to replace the role of teachers in the near future.

Table 1. Performance of Vietnamese student, ChatGPT, Bing Chat and Bard on different question levels of VNHSGE dataset.

Question level	Best performance ← Worst performance		
	BingChat ~ Bard	ChatGPT	Vietnamese students
Knowledge			
Comprehension	Bard	ChatGPT~BingChat	Vietnamese students
Application	Vietnamese students	Bing	ChatGPT
High application	Vietnamese students	ChatGPT	Bard

#### 4.6 Improving the Efficiency and Quality of LLMs

The strategic utilization of specialized biology datasets for fine-tuning LLMs is an effective method for improving their performance. The utilization of these datasets can facilitate the familiarization of LLMs with biology-specific terminology, context, and nuances, which are essential for generating precise and accurate responses. Specialized datasets not only enhance the precision of factual information but also facilitate the comprehension of biology inquiries by LLMs, thereby resulting in more pertinent and enlightening responses for students.

The involvement of biology experts and educators in the refinement of LLMs is highly valuable. Professionals possess specialized knowledge and expertise, enabling them to offer valuable insights that are specific to a particular field. This ensures that the models created are in accordance with the latest biology curricula and teaching goals. Professionals with expertise in the field of biology can contribute to the development of well-crafted training examples that are both of high quality and contextually rich. This, in turn, facilitates a deeper comprehension of intricate biology concepts and their practical implications for LLMs.

The integration of multi-modal learning represents a progressive and forward-looking educational strategy. The field of biology education frequently utilizes visual aids, including but not limited to images, diagrams, videos, and interactive simulations, to enhance the learning experience. The integration of these modalities with text-based learning has the potential to offer students a holistic and all-encompassing educational experience. Multi-modal learning can additionally facilitate the comprehension and interpretation of visual content for LLMs, thereby enhancing their adaptability and appropriateness for a wide range of educational resources.

Meta-learning techniques enable LLMs to quickly adapt to novel tasks and domains, thereby enhancing their overall versatility. The significance of this aspect is particularly pronounced within the realm of biology education, given the diverse and dynamic nature of the subject matter. Meta-learning enables LLMs to effectively apply previously acquired knowledge from one domain to another, thereby enhancing their proficiency in addressing diverse biological inquiries and obstacles.

It is of utmost importance to regularly update LLMs with the most recent advancements and insights in the field of biology. This practice is essential in order to maintain the currency and pertinence of the information they offer to students. The field of biology is characterized by its rapid evolution, necessitating the continuous updating of learning materials and resources to ensure that students are provided with accurate and current information. This practice is crucial in cultivating a dynamic learning environment.

Developing feedback loops with educators can play a crucial role in enhancing LLM responses. Educational professionals possess the ability to identify and communicate any inaccuracies, ambiguities, or potential areas for enhancement within content generated by LLM technology. The utilization of an iterative feedback process serves to improve the caliber and pertinence of LLM responses, thereby establishing a stronger alignment with the educational objectives and expectations of both educators and students.

In summary, the successful incorporation of LLMs into biology education necessitates a comprehensive strategy that encompasses technological advancements. The responsible and effective utilization of LLMs in the field of biology education is facilitated by various factors. These include the utilization of specialized datasets, fostering collaboration with experts, incorporating multi-modal learning approaches, employing meta-learning techniques, ensuring regular updates to the models, and implementing feedback mechanisms. The integration of these elements collectively enhances the overall learning experience for students.

## 5. Conclusion

The integration of LLMs has presented new opportunities for the field of biology education within the dynamic and constantly changing educational landscape. The objective of this research endeavor was to investigate the efficacy of three prominent LLMs - ChatGPT, BingChat, and Bard - within the specific context of the Vietnamese National High School Biology Examination. After conducting thorough evaluation and analysis, a number of significant insights and recommendations have been identified, which offer valuable guidance for enhancing the efficacy and ethical utilization of LLMs in the field of biology education. The results of our study have unveiled the distinct advantages and obstacles associated with each LLM in relation to a wide range of biological inquiries. BingChat and Bard have exhibited remarkable competence in delivering accurate information and assisting students in the practical application of their knowledge, respectively. Conversely, ChatGPT exhibits consistency; however, it faces difficulties when applied in practical scenarios and when engaging in complex cognitive endeavors. These findings emphasize the significance of comprehending the distinct capabilities of each LLM in order to optimize their efficacy as educational instruments. In order to fully exploit the capabilities of LLMs in the field of biology education, we suggest adopting a comprehensive strategy. The process of fine-tuning these models using biology datasets that are specifically tailored to the domain, and engaging in collaborations with experts in the field of biology, can significantly improve their ability to comprehend and interpret context and knowledge within that domain. By incorporating multi-modal learning, utilizing meta-learning techniques, and ensuring consistent updates, the versatility and adaptability of individuals can be expanded. Moreover, the notable efficacy of LLMs, specifically BingChat and Bard, when compared to the average academic achievements of students on the Vietnam National High School Graduation Examination, indicates a potentially fruitful approach for improving the quality of biology education in Vietnam. These models possess the potential to be highly beneficial



resources for educators, as they provide students with a wide range of information and tools that facilitate the development of critical thinking, problem-solving, and comprehension abilities.

In summary, this research offers significant contributions to our understanding of the capabilities and limitations of LLMs in the field of biology education. Additionally, it presents a comprehensive plan for the successful incorporation of LLMs into the educational journey. As we explore the continuously expanding realm of educational technology, the conscientious and knowledgeable utilization of LLMs has the potential to enhance the biology learning process, enabling students to engage in a voyage of exploration and comprehension within this crucial scientific domain.

## References

- [1] L. Chen, P. Chen, and Z. Lin, "Artificial Intelligence in Education: A Review," *IEEE Access*, vol. 8, pp. 75264–75278, 2020, doi: 10.1109/ACCESS.2020.2988510.
- [2] X. Q. Dao, N. B. Le, and T. M. T. Nguyen, "AI-Powered MOOCs: Video Lecture Generation," *ACM Int. Conf. Proceeding Ser.*, pp. 95–102, Mar. 2021, doi: 10.1145/3459212.3459227.
- [3] T. M. T. Nguyen, T. H. Diep, B. B. Ngo, N. B. Le, and X. Q. Dao, "Design of Online Learning Platform with Vietnamese Virtual Assistant," in *ACM International Conference Proceeding Series*, Feb. 2021, pp. 51–57, doi: 10.1145/3460179.3460188.
- [4] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv Prepr. arXiv1810.04805*, 2018, doi: <https://doi.org/10.48550/arXiv.1810.04805>.
- [5] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv Prepr. arXiv1907.11692*, 2019.
- [6] C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 1–67, 2020.
- [7] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 1877–1901, 2020.
- [8] X.-Q. Dao *et al.*, "VNHSGE: VietNameese High School Graduation Examination Dataset for Large Language Models," *arXiv Prepr. arXiv2305.12199*, May 2023, doi: 10.48550/arXiv.2305.12199.
- [9] G. Tsatsaronis *et al.*, "An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition," *BMC Bioinformatics*, vol. 16, no. 1, pp. 1–28, 2015, doi: 10.1186/s12859-015-0564-6.
- [10] J. Welbl, N. F. Liu, and M. Gardner, "Crowdsourcing Multiple Choice Science Questions," *arXiv Prepr. arXiv1707.06209*, pp. 94–106, 2017, doi: 10.18653/v1/w17-4413.
- [11] D. Hendrycks *et al.*, "Measuring Massive Multitask Language Understanding," *arXiv Prepr. arXiv2009.03300*, 2020, [Online]. Available: <http://arxiv.org/abs/2009.03300>.
- [12] P. Lu *et al.*, "Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 2507–2521, Sep. 2022.
- [13] OpenAI, "GPT-4 Technical Report," *arXiv Prepr. arXiv2303.08774*, 2023, doi: <https://doi.org/10.48550/arXiv.2303.08774>.
- [14] E. Agathokleous, C. J. Saitanis, C. Fang, and Z. Yu, "Use of ChatGPT: What does it mean for biology and environmental science?," *Sci. Total Environ.*, vol. 888, p. 164154, 2023, doi: <https://doi.org/10.1016/j.scitotenv.2023.164154>.
- [15] Y. Tong and L. Zhang, "Discovering the next decade's synthetic biology research trends with ChatGPT," *Synth. Syst. Biotechnol.*, vol. 8, no. 2, pp. 220–223, 2023, doi: 10.1016/j.synbio.2023.02.004.
- [16] A. HS Kumar, "Analysis of ChatGPT Tool to Assess the Potential of its Utility for Academic Writing in Biomedical Domain," *Biol. Eng. Med. Sci. Reports*, vol. 9, no. 1, pp. 24–30, 2023, doi: 10.5530/bems.9.1.5.
- [17] E. Shue, L. Liu, B. Li, Z. Feng, X. Li, and G. Hu, "Empowering Beginners in Bioinformatics with ChatGPT," *Quantitative Biology*, Vol. 11 (2), pp. 105–108, doi: 10.15302/J-QB-023-0327.

## Authors' Profiles



**Xuan-Quy Dao** obtained the Ph.D. degree in signal processing and telecommunication at the French National Institute for Research in Computer Science and Automation in 2014. He received both the M.S. and B.S. degrees in automation engineering from the Grenoble Institute of Technology in France in 2010.

He is currently a lecturer at the Eastern International University in Vietnam. From 2015 to 2017, he was a lecturer at Quang Binh University in Dong Hoi, Vietnam. His research interests include the development Online Learning Technology, Teaching-based Artificial Intelligence, Voice Cloning, and Speech-driven Face.



**Ngoc-Bich Le** received his undergraduate degrees at Bach Khoa University, Vietnam, his Master and Ph.D. in Mechatronics Science from Southern Taiwan University of Science and Technology – Taiwan in 2004, 2007, 2010, respectively.

He published a number of papers in preferred Journals such as J. Biomedical Microdevices, J. Microfluidics and Nanofluidics, J. Sensors and Actuators, and many Vietnamese Engineering Books in automation, CAD, and mold design. He also presented various academic as well as research-based papers at several national and international conferences. His articles focus on Medical devices, MEMs, Microfluidics, Robotics, and AI.

## Appendix

### A. LLMs Responses

This section shows samples of LLMs' responses, including BingChat, which may on occasion respond to Vietnamese inquiries in English. LLMs demonstrate the ability to give precise and well-supported answers to inquiries involving knowledge-level information retrieval in the subject of biology, utilizing their extensive training data and access to online resources. The first type of question can be answered without using any kind of reasoning, while a small degree of inference is needed for the second type of comprehension-level question. Inference is used for the third category of application-level questions, while in-depth reasoning is required for the fourth type of high application-level inquiry.

#### A.1 Knowledge level question

LLMs, by leveraging their trained data and external resources, can deliver precise responses at the knowledge level without the need for any reasoning. Their aptitude for handling information retrieval tasks makes them useful tools in the field of biology education. They can provide exact answers and comprehensive rationales for knowledge-level questions, benefiting students, instructors, and researchers. Through their retrieval and synthesis abilities, LLMs have the potential to enhance educational experiences, assist in studying complex biological topics, and promote biology teaching. The subsequent illustration serves as a representative instance:

<b>Câu hỏi:</b> Trong hệ sinh thái, nhóm sinh vật nào sau đây là sinh vật tự dưỡng? A. Động vật ăn thực vật.      B. Động vật kí sinh. C. Động vật ăn động vật.      D. Thực vật	<b>Question:</b> In an ecosystem, which of the following groups of organisms is autotrophic? A. Herbivorous animals.      B. Parasitic animals. C. Carnivorous animals.      D. Plants.
<b>Đáp án:</b> Trong hệ sinh thái, thực vật là sinh vật tự dưỡng vì chúng có khả năng quang hợp. Chọn <b>D</b> .	<b>Solution:</b> In an ecosystem, plants are autotrophic organisms because they have the ability to photosynthesize. Choose <b>D</b> .
<b>ChatGPT:</b> D. Thực vật là loại sinh vật tự dưỡng, tức là chúng có thể tự sản xuất thức ăn bằng cách sử dụng ánh sáng mặt trời, khí cacbonic và nước để chuyển hóa thành chất hữu cơ và oxy. Trong khi đó, các loại sinh vật khác như động vật ăn thực vật (tạp ăn), động vật ăn động vật (săn mồi) hoặc động vật kí sinh (sống trên hoặc bên trong cơ thể của sinh vật khác để lấy đồ ăn) đều phải dựa vào thực phẩm từ các nguồn khác để sống.	<b>ChatGPT:</b> D. Plants are autotrophic organisms, meaning they can produce their own food through photosynthesis using sunlight, carbon dioxide, and water to convert into organic matter and oxygen. On the other hand, other organisms such as herbivorous animals (eating plants), carnivorous animals (preying on other animals), or parasitic animals (living on or inside the bodies of other organisms to obtain food) all rely on food from other sources to survive.
<b>BingChat:</b> Trong hệ sinh thái, nhóm sinh vật tự dưỡng là Thực vật. Đó là đáp án <b>D</b> .	<b>BingChat:</b> In the ecosystem, the group of autotrophic organisms is Plants. That is answer <b>D</b> .
<b>Bard:</b> Đáp án đúng là: D. Thực vật. Sinh vật tự dưỡng là những sinh vật có khả năng tự tổng hợp chất hữu cơ từ các chất vô cơ đơn giản, sử dụng năng lượng từ ánh sáng mặt trời hoặc quá trình hóa học. Thực vật là nhóm sinh vật tự dưỡng điển hình nhất. Thực vật có thể sử dụng năng lượng từ ánh sáng mặt trời để tổng hợp chất hữu cơ từ nước và khí cacbon dioxide. Các sinh vật tự dưỡng là nguồn thức ăn chính cho các sinh vật dị dưỡng, bao gồm động vật và vi sinh vật.	<b>Bard:</b> The correct answer is: <b>D. Plants</b> . Autotrophs are organisms that are able to synthesize organic compounds from simple inorganic compounds, using energy from sunlight or chemical processes. Plants are the most typical group of autotrophs. Plants can use the energy from sunlight to synthesize organic compounds from water and carbon dioxide. Autotrophs are the main source of food for heterotrophs, including animals and microorganisms.
<b>All LLMs provide correct responses.</b>	

#### A.2 Comprehension level question

The limitations of LLM models in biological reasoning were highlighted in their responses to comprehension-level queries. The sample shows that language models can describe the steps required to make recombinant DNA but have trouble determining the exact solution. This highlights their dependence on their training data and limitations in understanding complex information. LLMs have a wide range of knowledge but have trouble inferring complex biological data. The effectiveness of LLMs in comprehension-level questions emphasizes the need for more development in their reasoning and inference abilities, particularly in biology. To successfully incorporate LLM models into biology education, it is imperative to address their difficulties in understanding complicated biological reasoning. By resolving these drawbacks, LLM models could become even more beneficial resources for biology teachers and students looking to better understand and analyze biological concepts. The subsequent text serves as an illustrative instance:

<p><b>Câu hỏi:</b> Cà chua lưỡng bội có 12 nhóm gen liên kết. Theo lý thuyết, thể một thuộc loài này có số lượng nhiễm sắc thể (NST) trong tế bào sinh dưỡng là</p> <p>A. 11. B. 23. C. 12. D. 24.</p>	<p>The polyploid tomato has 12 linked gene groups. <b>According to theory, the number of chromosomes in the somatic cells of an individual of this species is</b></p> <p>A. 11. B. 23. C. 12. D. 24.</p>
<p><b>Đáp án:</b> Số nhóm gen liên kết bằng số NST có trong bộ đơn bội của loài <math>\Rightarrow</math> Có 12 nhóm gen liên kết <math>\Rightarrow 2n = 24</math>. Thể một <math>có 2n - 1 = 23</math> NST. Chọn đáp án B.</p>	<p><b>Solution:</b> The number of linked gene groups is equal to the number of chromosomes in the haploid set of the species. In this case, since there are 12 linked gene groups, the number of chromosomes in the haploid set would be <math>2n = 24</math>. Therefore, the somatic cells of an individual would have <math>2n - 1 = 23</math> chromosomes. Choose B.</p>
<p><b>ChatGPT:</b> Trong các tế bào sinh dưỡng của thực vật, số lượng NST bằng số lượng bộ NST được di truyền từ bố và mẹ cộng thêm bộ NST độc lập xuất hiện mới do quá trình tái sắp xếp ngẫu nhiên. Trong trường hợp cà chua lưỡng bội, mỗi bộ NST bao gồm 12 nhóm gen liên kết, vì vậy tổng số NST trong thể sẽ là 2 lần 12 bộ NST, tức là 24 NST. Do đó, đáp án là D. 24.</p>	<p><b>ChatGPT:</b> In the somatic cells of plants, the number of chromosomes is equal to the number of sets of chromosomes inherited from both parents, plus any additional independent sets of chromosomes resulting from random rearrangement processes. In the case of a tetraploid tomato, each set of chromosomes consists of 12 linked gene groups. Therefore, the total number of chromosomes in the organism would be twice the number of sets, which is 2 times 12, resulting in 24 chromosomes. Therefore, the answer is D. 24.</p>
<p><b>BingChat:</b> Đáp án đúng là D. 24. Cà chua lưỡng bội có 12 nhóm gen liên kết, do đó thể một thuộc loài này có số lượng nhiễm sắc thể trong tế bào sinh dưỡng là 24 (12 cặp nhiễm sắc thể). Bạn có cần tôi giải thích thêm về câu hỏi này không?</p>	<p><b>BingChat:</b> The correct answer is D. 24. A tomato tetraploid has 12 linkage groups, so an individual of this species has 24 chromosomes in its somatic cell (12 pairs of chromosomes). Do you need me to explain more about this question?</p>
<p><b>Bard:</b> Đáp án là: C. 11. Lý giải: Cà chua lưỡng bội <math>có 2n = 12</math>. Thể một thuộc loài này có số lượng nhiễm sắc thể là <math>n - 1 = 12 - 1 = 11</math>. (3)</p>	<p><b>Bard:</b> The answer is: C. 11. Explanation: Tomatoes have a diploid number of chromosomes, <math>2n = 12</math>. Monosomic plants of this species have a chromosome number of <math>n - 1 = 12 - 1 = 11</math>. (3) Therefore, the answer is C.</p>
All LLMs provide incorrect responses.	

### A.3 Application level question

Providing precise answers to biology queries that call for extensive inference and image processing is difficult for LLMs, particularly when it comes to issues regarding human genetics and the interpretation of pedigree charts. These inquiries require a thorough grasp of biology and the ability to decipher complex data and visual cues. LLMs struggle with complex inference and image processing, especially in the setting of biology where subject-specific knowledge is essential. These models are constrained by their incapacity to decipher visual data or draw complex conclusions, emphasizing the value of human skill in genetics. Complex biological queries are difficult for LLMs to answer, highlighting the need for more development in image processing, inference tools, and domain-specific knowledge integration. Creating models that can efficiently process visual input and reason with complex biological concepts would be an enormous milestone toward improving their performance in biology education. The following is two typical examples:

<p><b>Câu hỏi:</b> Ở loài chuột, tính trạng màu lông được quy định bởi 1 gen có 4 alen trên nhiễm sắc thể thường, các alen trội là trội hoàn toàn. Cho 6 con P (khiêu là a, b, c, d, e, f) thuộc loài này thực hiện các phép lai sau:</p> <p>- Phép lai 1 (P): ♂ lông đen (a) <math>\times</math> ♀ lông đen (b) tạo ra F1 có 75% con lông đen : 25% con lông nâu.</p> <p>- Phép lai 2 (P): ♂ lông vàng (c) <math>\times</math> ♀ lông trắng (d) tạo ra F1 có 100% con lông vàng.</p> <p>- Phép lai 3 (P): ♂ lông nâu (e) <math>\times</math> ♀ lông vàng (f) tạo ra F1 có 25% con lông trắng : 25% con lông vàng : 50% con lông nâu.</p> <p><b>Theo lý thuyết, có bao nhiêu phát biểu sau đây đúng?</b></p> <p>I. Có 2 kiểu gen quy định kiểu hình lông nâu.</p> <p>II. Nếu chỉ dựa vào tỉ lệ kiểu hình F1 tạo ra từ các phép lai trên, xác định chính xác kiểu gen của 4 trong 6 con P.</p> <p>III. Cho con (d) giao phối với con (e) tạo ra đời con có tỉ lệ kiểu gen giống tỉ lệ kiểu hình.</p> <p>IV. Cho con (a) giao phối với con (f), đời con có thể có tỉ lệ 50% con lông đen : 50% con lông vàng.</p> <p>A. 2 B. 1 C. 3 D. 4</p>	<p><b>Question:</b> In mice, fur color is determined by a gene with 4 alleles on a regular chromosome, where the dominant alleles are completely dominant. Consider 6 individuals (denoted as a, b, c, d, e, f) of this species and the following crosses:</p> <p>Cross 1 (P): ♂ black fur (a) <math>\times</math> ♀ black fur (b) producing F1 with 75% black fur: 25% brown fur.</p> <p>Cross 2 (P): ♂ yellow fur (c) <math>\times</math> ♀ white fur (d) producing F1 with 100% yellow fur.</p> <p>Cross 3 (P): ♂ brown fur (e) <math>\times</math> ♀ yellow fur (f) producing F1 with 25% white fur: 25% yellow fur: 50% brown fur.</p> <p><b>According to theory, how many of the following statements are true?</b></p> <p>I. There are 2 types of genes that determine brown fur phenotype.</p> <p>II. Based solely on the ratio of fur types in F1 from the given crosses, the genetic types of 4 out of 6 individuals can be accurately determined.</p> <p>III. If individual (d) mates with individual (e), the offspring will have a genetic ratio identical to the phenotype ratio.</p> <p>IV. If individual (a) mates with individual (f), the offspring can have a ratio of 50% black fur: 50% yellow fur.</p> <p>A. 2 B. 1 C. 3 D. 4</p>
--	---

<p><b>Đáp án:</b>  Xác định trội lặn:  Từ phép lai 1, ta có P: lông đen x lông đen; F<sub>1</sub>: 3 lông đen : lông nâu  ⇒ Lông đen là trội so với lông nâu.  Từ phép lai 2, ta có lông vàng là trội so với lông trắng.  Từ phép lai 3, ta có lông nâu là trội so với lông vàng.  ⇒ Lông đen &gt; lông nâu &gt; lông vàng &gt; lông trắng.  Quy ước gen:  A<sub>1</sub> quy định lông đen; A<sub>2</sub> quy định lông nâu; A<sub>3</sub> quy định lông vàng;  A<sub>4</sub> quy định lông trắng.  I sai. Vì kiểu hình lông nâu do 3 kiểu gen quy định là A<sub>2</sub>A<sub>2</sub>; A<sub>2</sub>A<sub>3</sub>; A<sub>2</sub>A<sub>4</sub>.  II đúng. Dựa vào kiểu hình của F<sub>1</sub> ta chỉ xác định được kiểu gen của 4 con: c, d, e, f. Vì  Phép lai 2: F<sub>1</sub> tạo ra 100% lông vàng (A<sub>3</sub>A<sub>4</sub>) ⇒ P: A<sub>3</sub>A<sub>3</sub> (c) x A<sub>4</sub>A<sub>4</sub> (d).  Phép lai 3: F<sub>1</sub> tạo ra ¼ lông trắng (A<sub>4</sub>A<sub>4</sub>) ⇒ P: A<sub>2</sub>A<sub>4</sub> (e) x A<sub>3</sub>A<sub>4</sub> (f).  Ở phép lai 1:  F<sub>1</sub>: ¾ A<sub>1</sub>- : ¼ A<sub>2</sub>-  ⇒ P có thể là A<sub>1</sub>A<sub>2</sub> x A<sub>1</sub>A<sub>2</sub> hoặc A<sub>1</sub>A<sub>2</sub> x A<sub>1</sub>A<sub>3</sub> hoặc A<sub>1</sub>A<sub>2</sub> x A<sub>1</sub>A<sub>4</sub>.  ⇒ Không thể xác định được kiểu gen của 2 con a và b.  III đúng. P: A<sub>4</sub>A<sub>4</sub> (d) x A<sub>2</sub>A<sub>4</sub> (e)  F<sub>1</sub>: 1 A<sub>2</sub>A<sub>4</sub> : 1 A<sub>4</sub>A<sub>4</sub> (1 lông nâu : 1 lông trắng)  IV đúng.  Nếu a có kiểu gen A<sub>1</sub>A<sub>3</sub> thì P: A<sub>1</sub>A<sub>3</sub> (a) x A<sub>3</sub>A<sub>4</sub> (f)  F<sub>1</sub>: 50% lông đen : 50% lông vàng.  Chọn C.</p>	<p><b>Solution:</b> Determination of dominance and recessiveness:  From cross 1, we have P: black fur x black fur; F<sub>1</sub>: 3 black fur: 1 brown fur ⇒ Black fur is dominant over brown fur.  From cross 2, we have yellow fur is dominant over white fur.  From cross 3, we have brown fur is dominant over yellow fur.  ⇒ Black fur &gt; brown fur &gt; yellow fur &gt; white fur.  Gene notation:  A<sub>1</sub> determines black fur; A<sub>2</sub> determines brown fur; A<sub>3</sub> determines yellow fur; A<sub>4</sub> determines white fur.  Statement I is false. The brown fur phenotype is determined by 3 different genotypes: A<sub>2</sub>A<sub>2</sub>, A<sub>2</sub>A<sub>3</sub>, A<sub>2</sub>A<sub>4</sub>.  Statement II is true. Based on the phenotypes of F<sub>1</sub>, we can determine the genotypes of 4 individuals: c, d, e, f. Because:  Cross 2: F<sub>1</sub> produces 100% yellow fur (A<sub>3</sub>A<sub>4</sub>) ⇒ P: A<sub>3</sub>A<sub>3</sub> (c) x A<sub>4</sub>A<sub>4</sub> (d).  Cross 3: F<sub>1</sub> produces 1/4 white fur (A<sub>4</sub>A<sub>4</sub>) ⇒ P: A<sub>2</sub>A<sub>4</sub> (e) x A<sub>3</sub>A<sub>4</sub> (f).  In cross 1:  F<sub>1</sub>: 3/4 A<sub>1</sub>- : 1/4 A<sub>2</sub>-  ⇒ P could be: A<sub>1</sub>A<sub>2</sub> x A<sub>1</sub>A<sub>2</sub> or A<sub>1</sub>A<sub>2</sub> x A<sub>1</sub>A<sub>3</sub> or A<sub>1</sub>A<sub>2</sub> x A<sub>1</sub>A<sub>4</sub>. ⇒ We cannot determine the genotypes of individuals a and b.  Statement III is true. P: A<sub>4</sub>A<sub>4</sub> (d) x A<sub>2</sub>A<sub>4</sub> (e)  F<sub>1</sub>: 1 A<sub>2</sub>A<sub>4</sub>: 1 A<sub>4</sub>A<sub>4</sub> (1 brown fur: 1 white fur)  Statement IV is true.  If a has the genotype A<sub>1</sub>A<sub>3</sub>, then P: A<sub>1</sub>A<sub>3</sub> (a) x A<sub>3</sub>A<sub>4</sub> (f)  F<sub>1</sub>: 50% black fur: 50% yellow fur.  Choose C.</p>
<p><b>ChatGPT:</b> Trong loài chuột này, tính trạng màu lông được quy định bởi 1 gen có 4 alen trên NST thường và các alen trội là trội hoàn toàn.  Phép lai 1 (P): ♂ lông đen (a) × ♀ lông đen (b) tạo ra F<sub>1</sub> có 75% con lông đen: 25% con lông nâu.  • Điều này cho thấy alen a và b cùng quy định kiểu hình lông đen, vì con F<sub>1</sub> có 75% lông đen, tức là đa số mang alen a hoặc b.  Phép lai 2 (P): ♂ lông vàng (c) × ♀ lông trắng (d) tạo ra F<sub>1</sub> có 100% con lông vàng.  • Điều này cho thấy alen c là alen trội hoàn toàn so với d, vì con F<sub>1</sub> đều mang alen c.  Phép lai 3 (P): ♂ lông nâu (e) × ♀ lông vàng (f) tạo ra F<sub>1</sub> có 25% con lông trắng: 25% con lông vàng: 50% con lông nâu.  • Điều này cho thấy alen e và f là khác nhau và không trội lên nhau, vì con F<sub>1</sub> có tỉ lệ kiểu hình khác nhau.  Dựa trên các kết quả trên, chúng ta có thể suy ra:  I. Sai. Bởi vì chỉ có 1 gen quy định tính trạng màu lông và có 4 alen trên NST thường. Do đó, chỉ có tối đa 4 kiểu gen khác nhau.  II. Sai. Chỉ dựa vào tỉ lệ kiểu hình F<sub>1</sub>, chúng ta không thể xác định chính xác kiểu gen của các con chuột P. Ví dụ: Alen a và b đều có thể tạo ra lông đen, vậy con chuột P nào mang alen a, con nào mang alen b chúng ta không thể biết được.  III. Sai. Điều này không thể xảy ra vì tính trạng màu lông được quy định bởi 1 gen và không liên quan đến việc giao phối giữa các con chuột.  IV. Đúng. Alen a là trội so với alen f, và alen f là trội so với alen e. Do đó, khi con chuột a giao phối với con chuột f, đời con có thể có tỉ lệ 50% lông đen và 50% lông vàng.  Vậy đáp án là <b>B. 1</b>.</p>	<p><b>ChatGPT:</b> In this mouse species, the fur color is determined by a gene with 4 alleles on the normal chromosome, and the dominant alleles are completely dominant.  Cross 1 (P): ♂ black fur (a) × ♀ black fur (b) produces F<sub>1</sub> with 75% black fur and 25% brown fur.  • This indicates that alleles a and b both determine the black fur phenotype because the F<sub>1</sub> offspring have 75% black fur, meaning that the majority carry alleles a or b.  Cross 2 (P): ♂ yellow fur (c) × ♀ white fur (d) produces F<sub>1</sub> with 100% yellow fur.  • This shows that allele c is completely dominant over d because all F<sub>1</sub> offspring carry allele c.  Cross 3 (P): ♂ brown fur (e) × ♀ yellow fur (f) produces F<sub>1</sub> with 25% white fur, 25% yellow fur, and 50% brown fur.  • This indicates that alleles e and f are different and not dominant over each other because the F<sub>1</sub> offspring have different phenotypic ratios.  Based on the above results, we can infer:  I. False. Since there is only 1 gene determining the fur color and 4 alleles on the normal chromosome, there can be a maximum of 4 different genotypes.  II. False. Based solely on the F<sub>1</sub> phenotype ratios, we cannot accurately determine the genotypes of the P mice. For example, both alleles a and b can produce black fur, so we cannot determine which P mouse carries allele a and which carries allele b.  III. False. This cannot happen because the fur color trait is determined by a single gene and is unrelated to the mating between mice.  IV. True. Allele a is dominant over allele f, and allele f is dominant over allele e. Therefore, when mouse a mates with mouse f, the offspring can have a ratio of 50% black fur and 50% yellow fur.  Therefore, the correct answer is <b>B. 1</b>.</p>
	<p><b>BingChat:</b> The number of correct statements is <b>A. 2</b>. Statements I and II are correct. Statement I is correct because the 3:1 phenotypic ratio in the F<sub>1</sub> generation of cross 1 suggests that the brown fur phenotype is determined by a homozygous recessive genotype. Statement II is correct because based on the phenotypic ratios in the F<sub>1</sub> generation of the crosses, we can determine that individuals a and b are heterozygous (Aa), individual c is homozygous dominant (AA), and individual d is homozygous recessive (aa). Statements III and IV are incorrect.</p>



<p><b>Bard:</b> Đáp án: C. 3</p> <p>Giải thích:</p> <p>* Cho con (d) giao phối với con (e) tạo ra đời con có tỉ lệ kiểu gen giống tỉ lệ kiểu hình.</p> <p>- Kiểu gen của con (d): A_a</p> <p>- Kiểu gen của con (e): a_a</p> <p>- Tỉ lệ kiểu gen đời con: 1/2 A_a: 1/2 a_a</p> <p>- Tỉ lệ kiểu hình đời con: 1/2 lông trắng: 1/2 lông nâu</p> <p>* Cho con (a) giao phối với con (f), đời con có thể có tỉ lệ 50% con lông đen: 50% con lông vàng.</p> <p>- Kiểu gen của con (a): A_A</p> <p>- Kiểu gen của con (f): a_a</p> <p>- Tỉ lệ kiểu gen đời con: 1/2 A_A: 1/2 a_a</p> <p>- Tỉ lệ kiểu hình đời con: 1/2 lông đen: 1/2 lông vàng</p> <p>Vậy, có 3 phát biểu đúng.</p>	<p><b>Bard:</b> Answer: C. 3</p> <p>Explanation:</p> <p>F1 offspring from (d) x (e) have a genotype ratio equal to the phenotype ratio.</p> <p>Genotype of (d): A_a</p> <p>Genotype of (e): a_a</p> <p>Genotype ratio in F1 offspring: 1/2 A_a: 1/2 a_a</p> <p>Phenotype ratio in F1 offspring: 1/2 white coat: 1/2 brown coat</p> <p>F1 offspring from (a) x (f) can have a 50% black coat: 50% yellow coat ratio.</p> <p>Genotype of (a): A_A</p> <p>Genotype of (f): a_a</p> <p>Genotype ratio in F1 offspring: 1/2 A_A: 1/2 a_a</p> <p>Phenotype ratio in F1 offspring: 1/2 black coat: 1/2 yellow coat</p> <p>Therefore, there are 3 correct statements.</p>
In this instance, only Bard provides the correct answer, while ChatGPT and BingChat provide incorrect responses.	

<p><b>Câu hỏi:</b> Một loài thực vật lưỡng bội, xét 3 cặp gen: A, a; B, b và D, d trên 3 cặp NST cùng quy định màu hoa; kiểu gen có alen A, alen B và alen D quy định hoa tím; kiểu gen chỉ có alen A và alen B quy định hoa đỏ; các kiểu gen còn lại quy định hoa trắng. Theo lý thuyết, phát biểu nào sau đây đúng?</p> <p>A. Phép lai P: Cây hoa trắng dị hợp 2 cặp gen × Cây hoa trắng đồng hợp 2 cặp gen, tạo ra F1 có tỉ lệ 3 cây hoa đỏ: 5 cây hoa trắng.</p> <p>B. Phép lai P: Cây hoa tím dị hợp 1 cặp gen × Cây hoa đỏ dị hợp 2 cặp gen, tạo ra F1 không thể có 3 loại kiểu hình.</p> <p>C. Phép lai P: Cây hoa đỏ dị hợp 2 cặp gen × Cây hoa đỏ dị hợp 1 cặp gen, tạo ra F1 có tỉ lệ 1 cây hoa đỏ: 3 cây hoa trắng.</p> <p>D. Phép lai P: Cây hoa đỏ dị hợp 1 cặp gen × Cây hoa trắng dị hợp 2 cặp gen, tạo ra F1 có 9 loại kiểu gen.</p>	<p><b>Question:</b> A tetraploid plant species, considering 3 pairs of genes: A, a; B, b and D, d on 3 pairs of chromosomes that regulate flower color; genotype with allele A, allele B and allele D regulates purple flowers; genotype with only allele A and allele B regulates red flowers; the remaining genotypes regulate white flowers. According to theory, which of the following statements is correct?</p> <p>A. Crossing: Heterozygous white-flowered plant (2 gene pairs) × Homozygous white-flowered plant (2 gene pairs), producing F1 with a ratio of 3 red flowers: 5 white flowers.</p> <p>B. Crossing: Heterozygous purple-flowered plant (1 gene pair) × Heterozygous red-flowered plant (2 gene pairs), resulting in F1 that cannot have 3 types of phenotypes.</p> <p>C. Crossing: Heterozygous red-flowered plant (2 gene pairs) × Heterozygous red-flowered plant (1 gene pair), producing F1 with a ratio of 1 red flower: 3 white flowers.</p> <p>D. Crossing: Heterozygous red-flowered plant (1 gene pair) × Heterozygous white-flowered plant (2 gene pairs), resulting in F1 with 9 different genotypes</p>																														
<p><b>Đáp án:</b> A-B-D-: tím; A-B-dd: đỏ; Còn lại: trắng</p> <p>Xét các phát biểu:</p> <p>A sai. P: Cây hoa trắng dị hợp 2 cặp gen x Cây hoa trắng đồng hợp 2 cặp gen.</p> <p>VD: AabbDd x AAbbdd → Không tạo được hoa đỏ.</p> <p><b>B đúng.</b> Phép lai P: Cây hoa tím dị hợp 1 cặp gen x Cây hoa đỏ dị hợp 2 cặp gen.</p> <p>VD: AABbDd x AaBbdd → chỉ tạo được kiểu hình hoa đỏ và hoa tím.</p> <p>C sai. Phép lai P: Cây hoa đỏ dị hợp 2 cặp gen x Cây hoa đỏ dị hợp 1 cặp gen</p> <p>VD: AaBbdd x AaBbdd → (3A-:1aa)B-dd → 3 đỏ: 1 trắng.</p> <p>D sai. Phép lai P: Cây hoa đỏ dị hợp 1 cặp gen x Cây hoa trắng dị hợp 2 cặp gen</p> <p>VD: AaBBdd x AaBbDd → số kiểu gen: 4.</p>	<p><b>Solution:</b> A-B-D-: purple. A-B-dd: red. The rest: white.</p> <p>Considering the statements</p> <p>A is wrong. Crossing: Heterozygous white-flowered plant (2 gene pairs) × Homozygous white-flowered plant (2 gene pairs). For example: AabbDd x AAbbdd → Cannot produce red flowers.</p> <p><b>B is right.</b> Crossing: Heterozygous purple-flowered plant (1 gene pair) × Heterozygous red-flowered plant (2 gene pairs). For example: AABbDd x AaBbdd → can only produce red and purple flower phenotypes.</p> <p>C is wrong. Crossing: Heterozygous red-flowered plant (2 gene pairs) × Heterozygous red-flowered plant (1 gene pair). For example: AaBbdd x AaBBdd → (3A-: 1aa) B-dd → 3 red: 1 white</p> <p>D is wrong. Crossing: Heterozygous red-flowered plant (1 gene pair) × Heterozygous white-flowered plant (2 gene pairs). For example: AaBBdd x AaBbDd → number of genotypes: 4.</p>																														
<p><b>ChatGPT:</b> Đáp án đúng là B.</p> <p>Giải thích: Với phép lai P: Cây hoa tím dị hợp 1 cặp gen × Cây hoa đỏ dị hợp 2 cặp gen, ta có các kiểu gen là ABdd x AAbb, và ta có bảng Punnett sau đây:</p> <p>Bảng 1. Bảng Punnett</p> <table><tr><td></td><td>Ab</td><td>AB</td></tr><tr><td>dd</td><td>hoa trắng</td><td>hoa đỏ</td></tr><tr><td>db</td><td>hoa trắng</td><td>hoa đỏ</td></tr><tr><td>Dd</td><td>hoa trắng</td><td>hoa trắng</td></tr><tr><td>Db</td><td>hoa trắng</td><td>hoa đỏ</td></tr></table> <p>Từ bảng Punnett trên, ta thấy được F1 không thể có 3 loại kiểu hình, do đó phát biểu B đúng.</p> <p>Các phát biểu A, C, D đều sai:</p> <p>Phép lai P: Cây hoa trắng dị hợp 2 cặp gen × Cây hoa trắng đồng hợp 2 cặp gen tạo ra F1 với tỉ lệ kiểu hình là 1:1, do đó phát biểu A sai.</p> <p>Phép lai P: Cây hoa đỏ dị hợp 2 cặp gen × Cây hoa đỏ dị hợp 1 cặp gen tạo ra F1 với tỉ lệ kiểu hình là 1:1, do đó phát biểu C sai.</p> <p>Phép lai P: Cây hoa đỏ dị hợp 1 cặp gen × Cây hoa trắng dị hợp 2 cặp gen, tạo ra F1 có tỉ lệ kiểu hình là 1:1:1:1, tức là có 4 loại kiểu hình khác nhau, do đó phát biểu D sai.</p>		Ab	AB	dd	hoa trắng	hoa đỏ	db	hoa trắng	hoa đỏ	Dd	hoa trắng	hoa trắng	Db	hoa trắng	hoa đỏ	<p><b>ChatGPT:</b> The correct answer is B.</p> <p>Explanation: With the cross P: Heterozygous purple-flowered plant (1 gene pair) × Heterozygous red-flowered plant (2 gene pairs), we have the genotypes ABdd x AAbb, and the following Punnett square Table (Table 2):</p> <p>Table 2. Punnett square Table</p> <table><tr><td></td><td>Ab</td><td>AB</td></tr><tr><td>dd</td><td>white</td><td>red</td></tr><tr><td>db</td><td>white</td><td>red</td></tr><tr><td>Dd</td><td>white</td><td>white</td></tr><tr><td>Db</td><td>white</td><td>red</td></tr></table> <p>From the Punnett table, we can see that F1 cannot have 3 different types of phenotypes, therefore statement B is correct.</p> <p>Statements A, C, and D are incorrect:</p> <p>Cross P: Heterozygous white-flowered plant (2 gene pairs) × Homozygous white-flowered plant (2 gene pairs) produces F1 with a 1:1 phenotype ratio, thus statement A is incorrect.</p> <p>Cross P: Heterozygous red-flowered plant (2 gene pairs) × Heterozygous red-flowered plant (1 gene pair) produces F1 with a 1:1 phenotype ratio, thus statement C is incorrect.</p> <p>Cross P: Heterozygous red-flowered plant (1 gene pair) × Heterozygous white-flowered plant (2 gene pairs) produces F1 with a 1:1:1:1 phenotype ratio, meaning there are 4 different phenotypes, thus statement D is incorrect.</p>		Ab	AB	dd	white	red	db	white	red	Dd	white	white	Db	white	red
	Ab	AB																													
dd	hoa trắng	hoa đỏ																													
db	hoa trắng	hoa đỏ																													
Dd	hoa trắng	hoa trắng																													
Db	hoa trắng	hoa đỏ																													
	Ab	AB																													
dd	white	red																													
db	white	red																													
Dd	white	white																													
Db	white	red																													



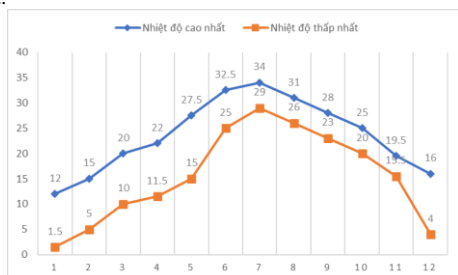
<p><b>BingChat:</b> Đáp án đúng là <b>B</b>. Phép lai P: Cây hoa tím dị hợp 1 cặp gen <math>\times</math> Cây hoa đỏ dị hợp 2 cặp gen, tạo ra F<sub>1</sub> không thể có 3 loại kiểu hình. Điều này đúng vì F<sub>1</sub> chỉ có thể có 2 loại kiểu hình: hoa tím (có alen A, alen B và alen D) và hoa đỏ (chỉ có alen A và alen B). Bạn có cần tôi giải thích thêm về câu hỏi này không?</p>	<p><b>BingChat:</b> The correct answer is <b>B</b>. Crossing: Heterozygous purple-flowered plant (1 gene pair) <math>\times</math> Heterozygous red-flowered plant (2 gene pairs), resulting in F<sub>1</sub> that cannot have 3 types of phenotypes. This is correct because F<sub>1</sub> can only have 2 types of phenotypes: purple flowers (with allele A, allele B and allele D) and red flowers (with only allele A and allele B). Do you need me to explain more about this question?</p>
<p><b>Bard:</b> Đáp án: <b>**D</b>. Phép lai P: Cây hoa đỏ dị hợp 1 cặp gen <math>\times</math> Cây hoa trắng dị hợp 2 cặp gen, tạo ra F<sub>1</sub> có 9 loại kiểu gen. <b>**</b></p> <p>* Kiểu gen của cây hoa đỏ dị hợp 1 cặp gen là AaBbDd.</p> <p>* Kiểu gen của cây hoa trắng dị hợp 2 cặp gen là aabbdd.</p> <p>* Các loại giao tử tạo ra từ cây hoa đỏ dị hợp 1 cặp gen là ABd, AbD, aBd, adD.</p> <p>* Các loại giao tử tạo ra từ cây hoa trắng dị hợp 2 cặp gen là abdd.</p> <p>* Tỷ lệ các loại kiểu gen ở F<sub>1</sub> là</p> $(ABd + AbD + aBd + adD) \times (abdd)$ $= (ABdd + ABdd + Abdd + aBdd + abdd + abdd + abdd + abdd) \quad (4)$ $= 9 \text{ loại kiểu gen.}$ <p>Vậy, đáp án đúng là <b>D</b>.</p>	<p><b>Bard:</b> Answer: <b>**D</b>. <b>**</b>The cross P: A heterozygous red flower plant <math>\times</math> A heterozygous white flower plant with 2 pairs of genes, produces 9 types of genotypes.</p> <p>The genotype of the heterozygous red flower plant is AaBbDd.</p> <p>The genotype of the heterozygous white flower plant with 2 pairs of genes is aabbdd.</p> <p>The types of gametes produced from the heterozygous red flower plant are: ABd, AbD, aBd, adD.</p> <p>The types of gametes produced from the heterozygous white flower plant with 2 pairs of genes are: abdd.</p> <p>The ratio of the types of genotypes in F<sub>1</sub> is: <math>(ABd + AbD + aBd + adD) \times (abdd) = (ABdd + ABdd + Abdd + aBdd + abdd + abdd + abdd + abdd) = 9</math> types of genotypes. Therefore, the correct answer is D.</p>

In contrast to the previous example, both ChatGPT and BingChat provide the correct answer, while Bard provides an incorrect response.

#### A.4 High application level question

Biology queries with a high applicability level is challenging for LLMs to deduce the relevant information to correctly respond to. However, these language models can provide accurate responses in some circumstances due to their exposure to comparable queries in their training data or their capacity to find pertinent material online. The internet is a useful tool for LLMs, enabling them to look up pertinent material and obtain additional viewpoints and insights. However, relying on existing knowledge and internet research may have its drawbacks, particularly in difficult and subtle high application-level biology topics. Breakthroughs must be made in training techniques, data curation, and reasoning skills for LLMs to perform better when answering high application-level queries. Their capacity to accurately infer and respond to complicated biology problems can be improved by including more varied and difficult question sets, improving inference algorithms, and expanding domain-specific expertise. Here are two illustrative instances:

**Câu hỏi:** Hình dưới đây (Hình 6) biểu thị sự biến động về nhiệt độ giữa ngày cao nhất và thấp nhất theo tháng ở một vùng. Thời gian sinh trưởng từ khi bắt đầu nuôi trong môi trường tự nhiên đến khi xuất chuồng của các giống vật nuôi A, B, C và D tối thiểu là 160 ngày. Bảng dưới đây (Bảng 3) cho biết giới hạn sinh thái về nhiệt độ của bốn giống vật nuôi A, B, C và D. Giả sử các điều kiện sinh thái khác của môi trường không ảnh hưởng đến sức sống của các giống vật nuôi đang nghiên cứu. Khi nhiệt độ môi trường thấp hơn giới hạn dưới hoặc cao hơn giới hạn trên của mỗi giống vật nuôi thì chúng sẽ bị chết.



Hình 6. Sự biến động về nhiệt độ giữa ngày cao nhất và thấp nhất theo tháng ở một vùng

Bảng 3. Giới hạn sinh thái về nhiệt độ

Giống vật nuôi	Giới hạn sinh thái
Giống A	12-32
Giống B	8-26
Giống C	14-40
Giống D	5-25

Dựa vào thông tin trong hình và bảng, có bao nhiêu nhận định sau đây đúng để lựa chọn các giống vật nuôi A, B, C và D chăn thả tại vùng này cho phù hợp?

- Giống A phù hợp để chăn thả ở vùng này.
- Có thể nuôi giống D từ tháng hai để đảm bảo năng suất khi xuất chuồng là cao nhất.
- Để đảm bảo đủ thời gian xuất chuồng, giống C là phù hợp nhất chăn thả ở vùng này.
- Không thể nuôi được giống B trong 160 ngày để xuất chuồng ở vùng này.

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

**Question:** The diagram below (Fig.6) represents the variation in maximum and minimum temperatures by month in a region. The minimum time for the growth of animal breeds A, B, C, and D from the start of rearing in the natural environment to the time of release is at least 160 days. The table below (Table 3) shows the temperature ecological limits for these four animal breeds. Assuming that the other ecological conditions of each breed do not affect their viability, when the environmental temperature is below the lower limit or above the upper limit for each breed, they will die.

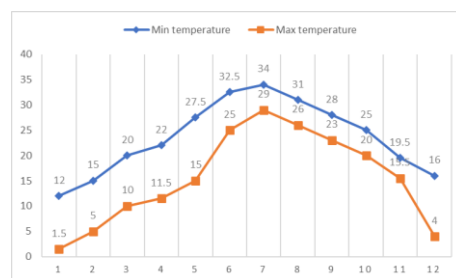


Fig. 6. The variation in maximum and minimum temperatures by month in a region

Table 3. The temperature ecological limits

Animal Breed	Ecological Limits
Breed A	12-32
Breed B	8-26
Breed C	14-40
Breed D	5-25

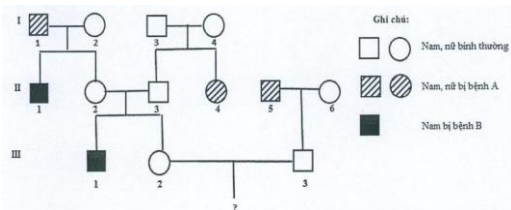
Based on the information in the diagram and table, how many of the following statements are correct in selecting suitable animal breeds A, B, C, and D for free-range grazing in this region?

- Breed A is suitable for free-range grazing in this region.
- Breed D can be raised from the second month to ensure the highest productivity at the time of release.
- To ensure sufficient time for release, breed C is the most suitable for free-range grazing in this region.
- Breed B cannot be raised in 160 days for release in this region.

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

<p><b>Đáp án:</b>  <b>I – Đúng.</b> Dựa vào khoảng nhiệt độ từng loài và khả năng nhiệt độ giao động trong tháng, năm của vùng này ta thấy: Giống A phù hợp để chăn thả ở vùng này.  <b>II – Sai.</b> Nếu nuôi từ tháng 2 giống D bị chênh lệch về nhiệt độ giới hạn sinh thái → Phát triển kém.  <b>III – Sai.</b> Giới hạn sinh thái nhiệt của C chênh lệch cao so với mức nhiệt độ của vùng.  <b>IV – Đúng.</b> So giới hạn nhiệt của B với mỗi 6 tháng liên tục của vùng ta thấy mức nhiệt độ không được đảm bảo.  <b>Chọn A.</b></p>	<p><b>Solution</b>  <b>I - Correct.</b> Based on the temperature range of each breed and the temperature fluctuations throughout the months and years in this area, it can be observed that Breed A is suitable for free-ranging in this area.  <b>II - False.</b> If Breed D is raised from February, it will experience temperature fluctuations outside its ecological limits, resulting in poor development.  <b>III - False.</b> The temperature ecological limit of Breed C deviates significantly from the temperature range of this area.  <b>IV - Correct.</b> Comparing the temperature limits of Breed B with the continuous six-month period in this area, it is evident that the temperature requirements are not met.  <b>Therefore, the correct answer is A.</b></p>
<p><b>ChatGPT:</b> Nhìn vào biểu đồ, giới hạn sinh thái về nhiệt độ của từng giống vật nuôi và yêu cầu về thời gian sinh trưởng, ta có thể suy ra các nhận định sau đây:  <b>I.</b> Giống A không phù hợp để chăn thả ở vùng này, vì giới hạn sinh thái của giống A có thể không đảm bảo sức sống của chúng trong mùa đông (tháng 12-2) khi nhiệt độ xuống thấp.  <b>II.</b> Không thể nuôi giống D từ tháng hai để đảm bảo năng suất khi xuất chuồng là cao nhất, vì giới hạn sinh thái của giống D vào khoảng nhiệt độ 18-29 °C và thời gian sinh trưởng tối thiểu là 160 ngày, do đó không có thời gian đủ để nuôi giống D từ tháng hai đến tháng chín để đạt được năng suất cao nhất khi xuất chuồng.  <b>III.</b> Giống C là phù hợp nhất để chăn thả ở vùng này, vì giới hạn sinh thái của giống C rộng và phù hợp với môi trường, thời gian sinh trưởng tối thiểu của giống C là 160 ngày, phù hợp với yêu cầu về thời gian xuất chuồng.  <b>IV.</b> Giống B có giới hạn sinh thái khá rộng và có thể chăn thả ở vùng này, tuy nhiên, yêu cầu thời gian sinh trưởng tối thiểu là 200 ngày, vượt quá thời gian tối thiểu là 160 ngày để đảm bảo đủ thời gian xuất chuồng, do đó không thể nuôi được giống B trong 160 ngày để xuất chuồng ở vùng này.  <b>V</b> vậy, có 3 nhận định đúng: II, III, IV, đáp án là <b>B. 3</b></p>	<p><b>ChatGPT:</b> Based on the temperature ecological limits of each animal breed and the growth time requirements, we can infer the following statements:  <b>I.</b> Breed A is not suitable for free-ranging in this area because its ecological limits may not ensure their survival during winter (December to February) when temperatures drop low.  <b>II.</b> It is not possible to raise Breed D starting from the second month to ensure the highest productivity at the time of release, as its ecological limits fall within the temperature range of 18-29 °C, and the minimum growth time required is 160 days. Therefore, there is not enough time to raise Breed D from February to September to achieve the highest productivity at the time of release.  <b>III.</b> Breed C is the most suitable for free-ranging in this area because its ecological limits are broad and compatible with the environment, and the minimum growth time for Breed C is 160 days, which aligns with the time required for release.  <b>IV.</b> Breed B has a relatively wide ecological limit and can be raised in this area. However, the minimum growth time required is 200 days, exceeding the minimum of 160 days needed to ensure sufficient time for release. Therefore, it is not possible to raise Breed B within 160 days for release in this area.  <b>Therefore, there are 3 correct statements: II, III, IV, and the answer is B. 3.</b></p>
<p><b>BingChat:</b> Based on the information in Table 1 and Table 2, we can determine which animal breeds are suitable for rearing in this region. Breed A has an ecological temperature range of 12-32 °C. According to Table 1, the minimum temperature in this region is 1.5 °C in January and the maximum temperature is 34 °C in July. This means that breed A can survive in this region for most of the year, but may not survive during the coldest and hottest months. Therefore, statement I is incorrect.  Breed D has an ecological temperature range of 5-25 °C. If breed D is reared starting in February, the minimum temperature would be 5 °C and the maximum temperature would be 15 °C. Over the next 160 days (about 5 months), the maximum temperature would reach 32.5 °C in June, which is above the upper limit of breed D's ecological temperature range. Therefore, statement II is incorrect.  Breed C has an ecological temperature range of 14-40 °C. According to Table 1, the minimum temperature in this region is 1.5 °C in January and the maximum temperature is 34 °C in July. This means that breed C can survive in this region for most of the year, but may not survive during the coldest months. Therefore, statement III is incorrect.  Breed B has an ecological temperature range of 8-26 °C. According to Table 1, the minimum temperature in this region is 1.5 °C in January and the maximum temperature is 34 °C in July. This means that breed B cannot survive in this region for a full 160 days due to both low and high temperatures outside its ecological temperature range. Therefore, statement IV is correct.  <b>In conclusion, only one of the statements (IV) is correct. The correct answer is D. 1.</b></p>	<p><b>Bard:</b> The correct answer is (A) 2.  Statement I is correct because the temperature range of breed A is within the average temperature range of the region.  Statement II is incorrect because the average temperature of the region from February to May is below the temperature range of breed B. Therefore, if breed B is raised from February, the yield at the time of sale will not be the highest.  Statement III is incorrect because the temperature range of breed C is within the average temperature range of the region. Therefore, breed C is not the most suitable breed for grazing in this region.  Statement IV is correct because the average temperature of the region from January to November is above the temperature range of breed B. Therefore, it is not possible to raise breed B for 160 days to sell in this region.  <b>Therefore, there are 2 correct statements among the given statements.</b></p>
<p><b>Bard:</b> Đáp án đúng là (A) 2.  Nhận định I đúng vì giống A có giới hạn nhiệt độ thích hợp nằm trong khoảng nhiệt độ trung bình của vùng này.  Nhận định II sai vì nhiệt độ trung bình của vùng này từ tháng 2 đến tháng 5 là dưới giới hạn nhiệt độ thích hợp của giống B. Do đó, nếu bắt đầu nuôi giống B từ tháng 2 thì năng suất khi xuất chuồng sẽ không cao nhất.  Nhận định III sai vì giống C có giới hạn nhiệt độ thích hợp nằm trong khoảng nhiệt độ trung bình của vùng này. Do đó, giống C không phải là giống phù hợp nhất để chăn thả ở vùng này.  Nhận định IV đúng vì nhiệt độ trung bình của vùng này từ tháng 1 đến tháng 11 là trên giới hạn nhiệt độ thích hợp của giống B. Do đó, không thể nuôi giống B trong 160 ngày để xuất chuồng ở vùng này.  <b>Vậy, có 2 nhận định đúng trong số các nhận định đã cho.</b></p>	<p><b>In this particular case, Bard alone furnishes the accurate solution, whereas ChatGPT and BingChat supply incorrect answers.</b></p>

**Câu hỏi:** Phả hệ sau đây (Hình 7) mô tả sự di truyền bệnh A và bệnh B. Biết rằng: mỗi bệnh đều do 1 trong 2 alen của 1 gen quy định; các gen phân li độc lập; alen trội là trội hoàn toàn; người I.1 không mang alen gây bệnh B.



Hình 7. Sự di truyền bệnh A và bệnh B

Theo lý thuyết, xác suất sinh con đầu lòng là con trai không bị bệnh A và không bị bệnh B của cặp vợ chồng III.2 và III.3 là

- A. 51/160      B. 119/320      C. 3/40      D. 17/80

**Đáp án:** Xét bệnh A: bố mẹ bình thường sinh con gái bị bệnh → Gen gây bệnh là gen lặn trên NST thường  
Xét bệnh B: Bố mẹ bình thường sinh con trai bị bệnh, người I.1 không mang gen gây bệnh B mà con trai II.1 bị bệnh → Gen gây bệnh là gen lặn trên NST X.

Quy ước:

A- không bị bệnh A; a- bị bệnh A.

B- không bị bệnh B; b- bị bệnh

B. Người III.3: Có bố II.5 bị bệnh A → người III.3: AaXBY.

Người III.2

+ Bệnh A:

Người II.2: Aa; người II.3 có em gái II.4 bị bệnh → II.3: 1AA:2Aa

Xét cặp

$$II.2 \times III.3: Aa \times (1AA:2Aa) \leftrightarrow (1A:1a) \times (2A:1a) \quad (5)$$

→ người III.2:

$$2AA:3Aa \quad II.2 \times III.3: Aa \times (1AA:2Aa) \leftrightarrow (1A:1a) \times (2A:1a) \quad (6)$$

+ Bệnh B: Có em trai bị bệnh → Bố mẹ:  $X^BX^b \times X^BY$

Người III.2:  $(1X^BX^b:1X^BX^B)$

Xét cặp vợ chồng:

$$III.2 \times III.3: (2AA:3Aa)(X^BX^b:1X^BX^B) \times AaX^BY \leftrightarrow (7A:3a)(3X^B:1X^b) \times (1A:1a)(1X^B:1Y) \quad (7)$$

→ xác suất sinh con đầu lòng là con trai không bị bệnh A và B:

$$A - X^BY = (1 - aa)X^BY = \left(1 - \frac{3}{10}a \times \frac{1}{2}a\right) \times \left(\frac{3}{4}X^B \times 1Y\right) = \frac{51}{160} \quad (8)$$

Chọn A

**ChatGPT:** Để tính xác suất sinh con đầu lòng là con trai không bị bệnh A và không bị bệnh B của cặp vợ chồng III.2 và III.3, ta cần phải xác định các kiểu gen của cặp vợ chồng này.

Theo phả hệ, ta có:

- Người II.1 mang 1 alen gây bệnh A và 1 alen gây bệnh B, do đó kiểu gen của người II.1 là AaBb.
- Người II.2 mang 2 alen trội về bệnh A và không mang alen gây bệnh B, do đó kiểu gen của người II.2 là AAbb.
- Người III.1 là con trai của người II.1, có thể mang kiểu gen là AB, Ab, aB hoặc ab.
- Người III.2 là con gái của người II.1, có thể mang kiểu gen là AB, Ab, aB hoặc ab.
- Người III.3 là con trai của người II.2, có kiểu gen là AAbb.

Do cặp vợ chồng III.2 và III.3 đều không mang alen gây bệnh B, ta chỉ cần quan tâm đến bệnh A. Xét xác suất của các kiểu gen có thể có của cặp vợ chồng III.2 và III.3 như sau:

- Kiểu gen của III.2: AB, Ab, aB hoặc ab, với xác suất đồng nhất là 1/4.
- Kiểu gen của III.3: AAbb.

Để sinh con trai không bị bệnh A, ta cần phải chọn gamete từ III.2 mang alen a, và gamete từ III.3 mang alen A, tức là chọn gamete Ab và gamete Ab. Xác suất của việc này là  $1/4 \times 1/2 = 1/8$ . Do đó, xác suất sinh con đầu lòng là con trai không bị bệnh A và không bị bệnh B của cặp vợ chồng III.2 và III.3 là

xác suất = xác suất sinh con trai không bị bệnh A × xác suất sinh con trai không bị bệnh B (vì cả cha và mẹ đều không mang alen gây bệnh B) =  $1/8 \times 1 = 1/8$

Vậy đáp án đúng là A. 51/160.

**Question:** The following pedigree (Fig.7) describes the inheritance of disease A and disease B. Knowing that: each disease is caused by 1 of 2 alleles of a gene; the genes segregate independently; the dominant allele is completely dominant; person I.1 does not carry the allele causing disease B.

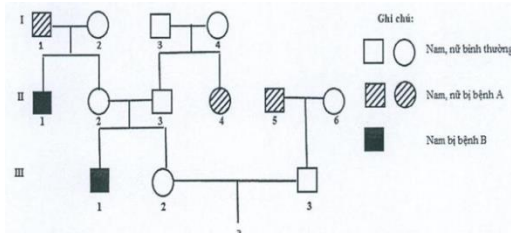


Fig. 7. The inheritance of disease A and disease B

According to theory, the probability of the first-born child being a boy who does not have disease A and does not have disease B of the couple III.2 and III.3 is

- A. 51/160      B. 119/320      C. 3/40      D. 17/80

**Solution:** Consider disease A: If normal parents have a daughter with the disease, the disease-causing gene is recessive on an autosomal chromosome.

Consider disease B: If normal parents have a son with the disease, and individual I.1 does not carry the disease-causing gene B but the son II.1 has the disease, the disease-causing gene is recessive on the X chromosome.

Notations: A- unaffected by disease A; a- affected by disease A. B- unaffected by disease B; b- affected by disease B.

Individual III.3: Has father II.5 affected by disease A ⇒ individual III.3: AaX<sup>BY</sup>.

Individual III.2:

+ Disease A: Father II.2: Aa; sibling II.3 has sister II.4 affected by the disease ⇒ II.3: 1AA:2Aa. Considering the pair II.2 × II.3: Aa × (1AA:2Aa) ⇒ (1A:1a) × (2A:1a) (5)

$$\Rightarrow \text{Individual person III.2: } 2AA:3Aa \quad II.2 \times III.3: Aa \times (1AA:2Aa) \leftrightarrow (1A:1a) \times (2A:1a) \quad (6)$$

+ Disease B: Having an affected brother ⇒ Parents: X<sup>BX</sup>b × X<sup>BY</sup> ⇒ Individual person III.2: (1X<sup>BX</sup>b:1X<sup>BX</sup>B:1X<sup>BX</sup>b). Considering the couple: III.2 × III.3: (2AA:3Aa)(X<sup>BX</sup>b:1X<sup>BX</sup>B) × AaX<sup>BY</sup> (7)

⇒ (7A:3a)(3X<sup>B</sup>:1X<sup>b</sup>) × (1A:1a)(1X<sup>B</sup>:1Y) ⇒ Probability of having a first-born son unaffected by diseases A and B:

$$A - X^BY = (1 - aa)X^BY = \left(1 - \frac{3}{10}a \times \frac{1}{2}a\right) \times \left(\frac{3}{4}X^B \times 1Y\right) = \frac{51}{160} \quad (8)$$

Choose A.

**ChatGPT:** To calculate the probability of having a first-born son who is not affected by disease A and disease B in the couple III.2 and III.3, we need to determine their genotypes.

According to the pedigree:

Individual II.1 carries 1 allele for disease A and 1 allele for disease B, so the genotype of individual II.1 is AaBb.

Individual II.2 carries 2 dominant alleles for disease A and does not carry the allele for disease B, so the genotype of individual II.2 is AAbb.

Individual III.1 is the son of individual II.1 and can have the genotype AB, Ab, aB, or ab.

Individual III.2 is the daughter of individual II.1 and can have the genotype AB, Ab, aB, or ab.

Individual III.3 is the son of individual II.2 and has the genotype AAbb.

Since both individuals III.2 and III.3 do not carry the allele for disease B, we only need to consider disease A. Let's examine the probabilities of the possible genotypes for the couple III.2 and III.3:

Genotype of III.2: AB, Ab, aB, or ab, with equal probabilities of 1/4 each.

Genotype of III.3: AAbb.

To have a first-born son who is not affected by disease A, we need to select a gamete from III.2 carrying the allele a and a gamete from III.3 carrying the allele A, which means selecting the gametes Ab and Ab. The probability of this is  $1/4 \times 1/2 = 1/8$ . Therefore, the probability of having a first-born son who is not affected by disease A and disease B in the couple III.2 and III.3 is:

Probability = Probability of having a first-born son without disease A \* Probability of having a first-born son without disease B (both parents do not carry the allele for disease B) =  $1/8 \times 1 = 1/8$ .

Thus, the correct answer is A. 51/160.

BingChat	BingChat: no answer
<p><b>Bard:</b> Đáp án: C.</p> <p>Xác suất sinh con trai không bị bệnh và không bị bệnh của cặp vợ chồng III. 2 và III.3 là</p> <ul style="list-style-type: none"> <li>* Xác suất cặp vợ chồng III. 2 và III.3 sinh con trai là <math>1/2</math>.</li> <li>* Xác suất con trai không bị bệnh là <math>1 - 1/4 = 3/4</math>.</li> <li>* Xác suất con trai không bị bệnh là <math>1 - 1/4 = 3/4</math>.</li> </ul> <p>Vậy xác suất chung là <math>1/2 \times 3/4 \times 3/4 = 9/32 = 3/40</math>.</p> <p>Vậy đáp án đúng là C.</p>	<p><b>Bard:</b> Answer: C.</p> <p>The probability of a son of couple III.2 and III.3 being born without disease and without disease is:</p> <p>The probability of couple III.2 and III.3 having a son is: <math>1/2</math>.</p> <p>The probability of a son not being affected by disease is: <math>1 - 1/4 = 3/4</math>.</p> <p>The probability of a son not being affected by disease is: <math>1 - 1/4 = 3/4</math>.</p> <p>Therefore, the overall probability is: <math>1/2 \times 3/4 \times 3/4 = 9/32 = 3/40</math>.</p> <p>Therefore, the correct answer is C.</p>
<p><b>In this instance, ChatGPT provides the correct answer by demonstrating relatively good mathematical ability, while Bard provides an incorrect response and BingChat does not provide an answer.</b></p>	

**How to cite this paper:** Xuan-Quy Dao, Ngoc-Bich Le, "LLMs Performance on Vietnamese High School Biology Examination", International Journal of Modern Education and Computer Science(IJMECS), Vol.15, No.6, pp. 14-30, 2023. DOI:10.5815/ijmecs.2023.06.02