Research on the Application of Generative Artificial Intelligence to Evaluate Responses Related to Questions About COVID-19 in Terms of Their Accuracy and Readability

Zongjing Liang , Yun Kuang [*] , Xiaobo Liang , Gongcheng Liang , Zhijie Li

Posted Date: 16 May 2025

doi: 10.20944/preprints202505.1319.v1

Keywords: COVID-19; Generative Artificial Intelligence; Infectious Disease Prevention and Con-trol; Performance Comparison; Public Health Knowledge Dissemination

Article

Research on the Application of Generative Artificial Intelligence to Evaluate Responses Related to Questions About COVID-19 in Terms of Their Accuracy and Readability

Zongjing Liang [1], Yun Kuang [2,*], Xiaobo Liang [3], Gongcheng Liang [3] and Zhijie Li [1]

[1] Guangxi Normal University, School of Economics Management
[2] Guilin Normal University, Library
[3] Guilin Normal University, Network and Education Technology Center
* Correspondence: kyun@mail.glnc.edu.cn; Tel.: +(086)-18007879626

Abstract: *Objective:* This study aims to compare the accuracy and readability of COVID-19 infectious disease prevention and control knowledge generated by four major generative artificial intelligence models—two international models (ChatGPT and Gemini) and two domestic models (Kimi and Ernie Bot)—to evaluate the performance characteristics of domestic and international models. *Methods:* The knowledge Q&A from the COVID-19 prevention guidelines issued by the U.S. Centers for Disease Control and Prevention (CDC) was used as the evaluation standard. The texts generated by the four models were compared with the standard in terms of accuracy, readability, and understandability. Then, a neural network model based on intelligent algorithms was used to extract the factors influencing the readability of the generated texts. Finally, text analysis was applied to explore the medical topics in the generated texts. *Results:* Text accuracy.Domestic models showed higher accuracy in generated texts, while international models demonstrated better reliability. Text readability.Domestic models produced fluent language and a style suitable for public reading; international models exhibited better stability and tended to generate formal documentation. Text understandability.Domestic models had better readability; international models had more stable output. Readability influencing factors.The sentence length indicator (AWPS) of texts generated by both domestic and international models was the most important factor affecting readability. *Topic analysis*: ChatGPT focused more on epidemiological knowledge; Gemini on the healthcare field; Kimi on multidisciplinary information; and Ernie Bot on clinical medical topics. *Conclusion:* Texts generated by domestic models are easy to understand and more suitable for public reading, and are better suited for clinical testing, health consultation, and similar applications. Texts generated by international models have higher accuracy and professionalism, focusing more on epidemiological analysis, disease severity assessment, and related fields. Based on the findings, it is recommended that infectious disease prevention knowledge systems—such as those for COVID-19—should pay more attention to the public's knowledge base and comprehension level, achieving an organic integration of professionalism and accessibility in AI-generated knowledge, thereby providing objective reference materials for future major infectious disease outbreaks.

Keywords: COVID-19; generative artificial intelligence; infectious disease prevention and control; performance comparison; public health knowledge dissemination

1. Introduction

The introduction of the new epidemic, which originated in the end of 2019, has caused great damage to global society, economy, and culture in the form of its strong infectiveness and infectious speed [1]. The World Health Organization announced on January 30, 2020, that the new outbreak constituted an international public health event of international concern because of the severity of the epidemic [2]. This is the highest level of warnings the World Health Organization has released under international health regulations. On May 5, 2023, the World Health Organization announced that COVID-19 was no longer an international public health event of great concern [3], marking the official end of the global emergency phase of COVID-19.

Although the World Health Organization declared that the outbreak was not in an emergency phase in 2023, this does not mean that infectious diseases are not a threat to humans. In contrast, currently, infectious diseases continue to present a global threat [4]. The existing epidemic of the disease is still spreading [5]. The COVID-19 virus remains intact, and there are reports from around the world of disease, hospitalization, and death caused by COVID-19. The COVID-19 virus is a respiratory epidemic, and other respiratory infections continue to cause significant public health burdens. For example, the recent influenza pandemic poses a new threat to public health security. Furthermore, there are new infectious diseases [6]. According to our research, there are still potential outbreaks of new infections in the world, and other infectious diseases from the coronavirus family may be re-erupting. Furthermore, there may be new mutations in the virus. Based on this, although the emergency state of the COVID-19 outbreak is over, a new outbreak of infectious diseases is still possible. Therefore, it is critical to be aware of the emergence of new infectious diseases and the emergence of effective early warnings of infectious diseases.

In order to effectively cope with the major infectious diseases that may occur again in the future, it is equally important to enhance the awareness of self-protection in public health management. There are many effective ways to promote the dissemination of knowledge about infectious diseases. The information network has become the main source for the prevention and response of common people, and search engines are traditionally people's source of information. In recent years, with the development of artificial intelligence, especially in November 2022, generative artificial intelligence technology, which is represented by ChatGPT, has been increasingly used to obtain information and is gradually replacing the traditional search engine [7]. Compared with the traditional search engine, generative artificial intelligence has the following advantages [8]: it can be automatically generated, it can answer questions accurately, and it can produce a personalized experience.

Generative artificial intelligence technology, with its unique response mode, has revolutionized the knowledge of the population, and the generative artificial intelligence model has become a new channel for the dissemination of knowledge about infectious diseases [9]. Generative artificial intelligence technologies have been applied in many areas, including in the field of medicine [10]. Although generative artificial intelligence has been widely used, problems remain in its application, including ethics, accuracy, and readability, especially regarding the accuracy and readability of the answers it provides regarding medical questions, which has become one of the focal issues of the current generation of artificial intelligence applications [11].

At present, at home and abroad, the study of the technical knowledge of artificial intelligence technology, in terms of medical knowledge, is focused on evaluating the correctness of ChatGPT in medical examinations and when answering questions [12–14]. ChatGPT used in basic and clinical medicine has been tested [15]. The performance of ChatGPT-3.5 in Polish medical exams was evaluated for a study [16]. ChatGPT's theoretical knowledge and specification accuracy regarding bacterial infection has been evaluated [17], as has the reliability and readability of ChatGPT-4's assessment of hypothyroidism in pregnancy [18]. Furthermore, ChatGPT's practical analysis of myopia has been evaluated [19], and a comparison of the performance of ChatGPT and the Google Bard language model, in terms of generating text, has been carried out [20]. Research has compared the ability of ChatGPT and Internet searches to answer patients' questions [21].

After summarizing, we find that there is little research on the accuracy and readability of the knowledge of infectious diseases, and there are few research results that affect the ability of text readability. China now has more than 1 billion Internet users, with the number of users of generative artificial intelligence who query infectious diseases through the platform reaching 230 million each day. China is a big country, and its population still needs to be aware of the occurrence and dissemination of infectious diseases. Therefore, this paper studies the accuracy and readability of generative text on infectious diseases on four major generative artificial intelligence platforms (including Chinese models). This study has important theoretical and practical value. Through this study, we can provide medical information for medical professionals, science and technology enterprises, and the general public and help them choose the most suitable generative AI Q&A tools to improve the efficiency of health consultations. At the same time, through a comparison of the performance of Chinese and foreign generative artificial intelligence platforms, this paper reveals the relative advantages and disadvantages of the domestic and foreign models in the prevention and control of infectious diseases and provides a reference for improving the production of artificial intelligence technology.

2. Materials and Methods

This paper compares the generative artificial intelligence model of four mainstream platforms at home and abroad. In addition, the article compares the four models' accuracy, accessibility, comprehension, and readability influence factors and the production of text research topics. As research material, we selected the COVID-19 control guide Q&A text of the CDC as the standard answer, and the responses provided by the four generative artificial intelligence models of the domestic and foreign generation of artificial intelligence models were compared with the standard answer. The domestic generative AI models tested were Kimi and Ernie Bot, and the foreign models were ChatGPT and Gemini.

2.1. Research Materials

This article is related to infectious diseases. It is based in the U.S. CDC and control center, and the corresponding research text is the COVID-19 prevention and control guide of the CDC from August 2020 [22]. This can be downloaded at the following URL: https://stacks.CDC.gov/view/CDC/89817. This article contains 53 questions and answers. The 53 questions and the answers covered the following areas [22]: (1) COVID-19 infection risk and control (six questions). (2) COVID-19 spread and prevention (seven questions). (3) Detection and diagnosis (six questions). (4) The clinical management of COVID-19 (eight questions). (5) The management of the special population (seven questions). (6) The treatment of and prevention measures for COVID-19 (six questions). (7) Prevention and vaccination (six questions). (8) Other relevant questions (seven questions). The question and answer documents cover the categories of infection risk, communication prevention, detection diagnosis, clinical management, special population, prevention and treatment, vaccination, and other problems. The guide is aimed at both general medical staff and the scientific community at large in order to provide authoritative resources for the prevention and control of outbreaks.

2.2. Research Methods

In order to compare the generative text properties of four generative artificial intelligence models at home and abroad, this paper will study several relevant indicators. By comparing the generative text of the four models to the standard text provided by the CDC, we compared the accuracy, accessibility, comprehension, and readability impact indicators and the generative text topics. Through comprehensive evaluation of the domestic and international models of the analysis and response of infectious diseases, this study provides a reference guide for medical workers and people to obtain knowledge regarding infectious diseases.

The variable metrics used for model performance comparison are as follows: SimHash, Flesch–Kincaid grade level (FKGL), Flesch Reading Ease Score (FRES), reading level (RL), average words per sentence (AWPS), average syllables per word (ASPW), and sentences and words, where SimHash stands for text similarity, i.e., accuracy. The Flesch–Kincaid grade level (FKGL) indicates the U.S. school grade corresponding to the reading difficulty of the text. The Flesch Reading Ease Score (FRES) indicates how intelligible the text is. Reading level (RL) indicates the minimum education level, in terms of reading, required for a text. Average words per sentence (AWPS) indicates the average number of words per sentence. Average syllables per word (ASPW) indicates the average number of syllables per word. The "sentences" variable indicates the total number of sentences the text contains and is used to calculate the average sentence length. The "words" variable indicates the total number of words. The definitions and instructions for the use of each indicator are as follows.

1. Comparison of text accuracy;

There are many algorithms for comparing the accuracy of two texts in current research, and the SimHash index describes the text accuracy. The SimHash algorithm is a method of detecting file accuracy based on hash functions [23]. This method can calculate the accuracy of two texts to measure the similarity of the two texts. The SimHash algorithm is suitable for text comparison, data classification, etc. It functions on the principle of ensuring the validity and accuracy of the algorithm by filtering and optimizing the calculation strategy [24]. The similarity of two texts is represented by the numerical value of the SimHash algorithm. The greater this number is, the more that the texts are accurate and similar.

2. Textual legibility comparison;

Text accessibility is evaluated by the FRES test, which is measured in terms of the extent to which the text is easy to read [9]. The text readability index uses Flesch Reading Ease Score (FRES). The Flesch Reading Index is a measure of the readability of English text which was proposed by Rudolf Flesch in 1948 [25]. It scores texts on a range of 0 to 100. The higher the score, the easier it is to read the text, and the lower the score, the more difficult it is to read.

3. Textual comprehensibility comparison;

The FKRGL index is used for easily understanding text in our study. It is measured in terms of the degree of understanding of text [9]. FKGL stands for Flesch-Kincaid grade level. This indicator is one of the important indicators of text comprehension, especially in areas such as medical care and education. The index was first proposed by Rudolf Flesch in 1948 and then revised by J. Peter Kincaid. This score represents the level of reading in the education system [26].

4. The study of the effect of text readability;

The text's readability is also evaluated with the reading level (RL) indicator [27]. RL indicates the minimum grade of education required to read the text in question. In this study, we use the neural network model as one of the intelligent algorithms to analyze the key factors affecting text readability. According to the requirements of neural network model construction, the input layer indicators of the model are the AWPS indicator, ASW indicator, RL indicator, word indicator, and SimHash indicator, and the output layer indicator is the RL indicator. This paper discusses the key factors affecting text readability by multilayer perceptron mining in neural networks (multilayer perceptron is one of many neural network model algorithms).

5. The semantic comparison of text content.

This paper studies the word frequency and word frequency semantics of the model by text analysis [28]. This method is the basic and most common analysis process used in natural language comprehension, and it mainly consists of two aspects. The first is word frequency statistics, and the second is topic mining. Word frequency statistics are the most basic parts of text analysis used to identify the most common words in the text in order to understand the main content or keywords of the text. Topic mining is a combination of high-frequency words and context induction topics. This method is suitable for the analysis of words in text in our study.

2.3. Statistical Method

This paper is a quantitative analysis study, and different calculation methods are used in each index. This study calculates the text accuracy index of SimHash using the network computing platform for the solution: https://kiwirafe.pythonanywhere.com/app/xiangsi/. Text readability and language size metrics (FKGL, FRES, RL, AWPS, ASPW, sentences, word) use the network computing platform: https://goodcalculators.com/flesch-kincaid-calculator/?utm_source=chatgpt.com. This study used the neural model analysis module in SPSS27 to calculate the multilayer perceptron and used the software to carry out the statistical analysis of the full text data. We used Python to record the word frequency statistics of the text.

3. Results

3.1. Comparison of Text Accuracy

In this study, SimHash similarity data were used as an accuracy index. In order to compare the accuracy of the four models in answering the text of the 53 questions of the CDC, this study calculated the SimHash accuracy, displayed the calculated data as a box plot, and measured the consistency of the four models in answering the text of the same question by analyzing the box plot. The calculation results are shown in Figure 1.

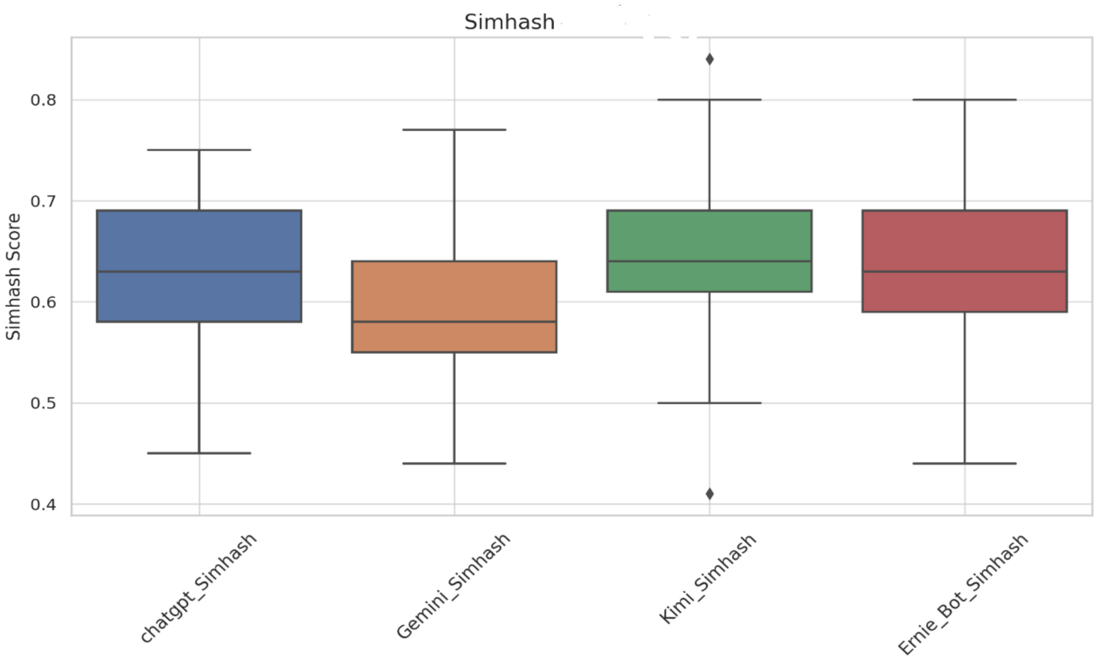


Figure 1. Box plots of ChatGPT, Gemini, Kimi, and Ernie Bot SimHash values.

(1) The consistency of the text

The consistency index is the standard for the SimHash similarity score. From the calculation results, Kimi scored the highest (0.646), indicating that Kimi was the most consistent in answering questions. ChatGPT's score was close to that of Ernie Bot, but its answers were poor. The Gemini model has the lowest average score, indicating that its answers were the worst among the four models.

(2) Stability comparison

The accuracy stability of the question is measured by standard deviation. The minimum standard deviation of ChatGPT is the lowest (0.074), indicating that it is the most stable and that its accuracy in answering different questions does not change. The standard deviation of Ernie Bot (0.082) indicates that the accuracy of the answer changes frequently, and the stability of the quality of the text is not high. The volatility of the Kimi and Gemini models is close (0.079 and 0.077, respectively), indicating that their responses demonstrated similar levels of accuracy stability.

(3) Extreme values

The extreme values are compared to the distribution of the group points. The minimum accuracy of the Kimi mode is very volatile in terms of the answer that it provides to questions, but in regard to answers, ChatGPT and Ernie Bot demonstrated large deviations in the box diagram, which shows that the standard deviation of the three models is relatively focused, while the observation value of the standard of the abnormal value is not, and the data may achieve a level of distribution that is close to normal or relatively uniform.

In summary, the response of the domestic model Kimi demonstrates the best accuracy in terms of the text of its answers text and the standard of its answer, but the answers were not as accurate as those of ChatGPT. Ernie Bot was the best of the four models, and, on the whole, ChatGPT and Kimi performed well. The stability of ChatGPT's answers were the best. The accuracy of Gemini's answers was the lowest, and the accuracy of its answers, overall, were worse than that of the other models. In general, Kimi has the best accuracy, but Kimi is less stable. ChatGPT was the second of the four models in terms of the stability of its answers and the reliability of its text.

3.2. Text Readability Comparison

The text readability index uses the Flesch Reading Ease Score. The FRES index of the four models is shown in Figure 2.

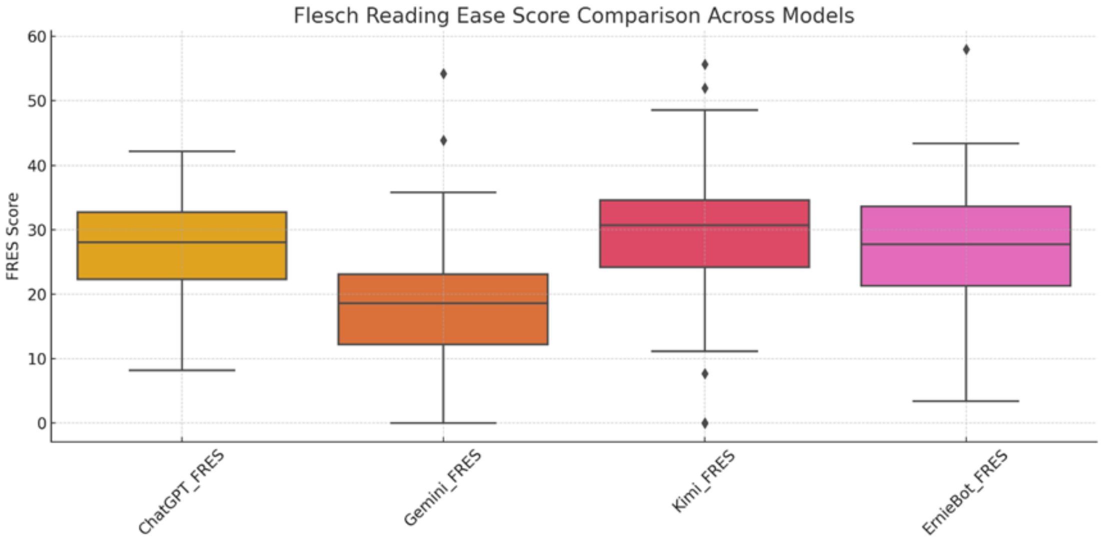


Figure 2. Flesch Reading Ease Score (i.e., FRES) index box diagram.

The following is a comparison of Chinese and foreign models in terms of text consistency, stability, and extreme values.

(1) Comparison of text consistency

The ChatGPT score distribution is more concentrated, indicating that the text demonstrates high consistency. The Gemini score distribution is also more concentrated, indicating that the text demonstrates less consistency than that of ChatGPT. The distribution of the domestic model Kimi is more diffuse, indicating that the text is not consistent with the previous two. The score distribution of Ernie Bot is also more fragmented, with a median of about 35, indicating that its text is the least consistent.

(2) Stability comparison

The ChatGPT box is less long, indicating that the score is more stable. The length of the Gemini box is a little greater than that of ChatGPT, indicating that Gemini is less stable than ChatGPT. The length of the box of the domestic model Kimi is greater, indicating that this fraction of the box is volatile and that its stability is poor. The length of Ernie Bot’s box is longer and more abnormal, indicating that its scores are the most volatile and that its stability is the worst.

(3) Extreme comparison

ChatGPT does not have obvious extreme values, which indicates that the distribution of the scores is more uniform. Gemini has some extreme values, but not many of them, indicating some anomalies in its score distribution. The domestic model Kimi has several extreme values, indicating that there are some anomalies in the distribution of the scores. Ernie Bot has multiple extreme values, indicating that there are many exceptions in the distribution of the scores.

To sum up, the performance of ChatGPT is the best in the three aspects of text consistency, stability, and extreme value, followed by Gemini, Kimi, and Ernie Bot.

3.3. Textual Comprehensibility Comparison

The text is easy to understand and uses the FKGL index. The Flesch–Kincaid grade level indicator is the FKGL index, which is one of the indicators of text readability. Text readability represents the readability of the text and the corresponding reading level in the United States. Figure 3 shows the FKGL index box diagram for the four large models.

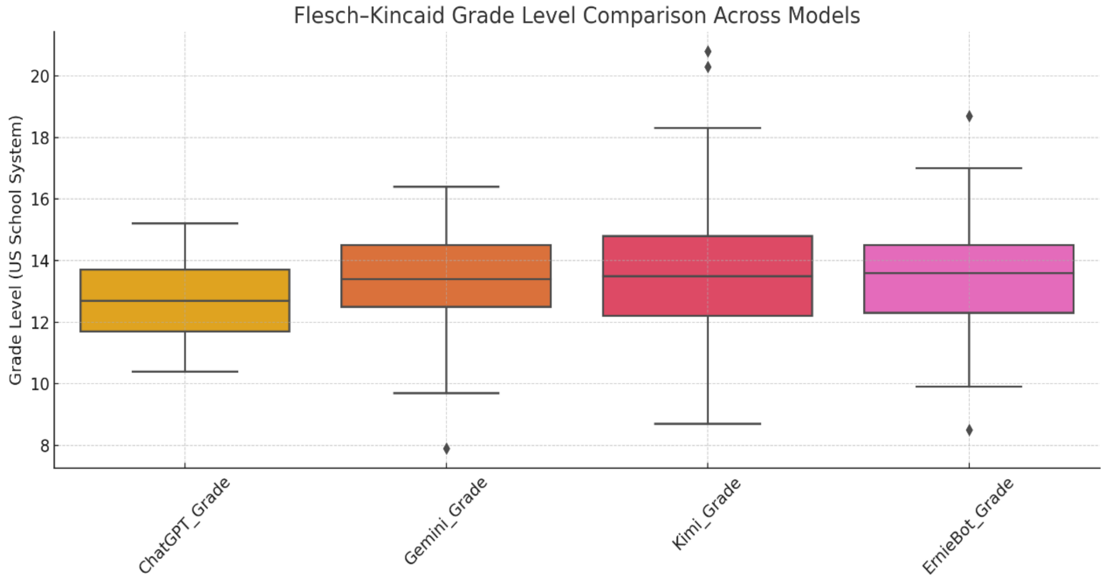


Figure 3. Flesch–Kincaid grade level (i.e., FKGL index).

This is based on the comparison of the four models in the above box, which is based on the output level, stability, and extreme value control.

(1) The overall comparison of the output level

In Figure 2, we can see the Flesch–Kincaid grade score, and the box diagram shows that the Chinese and foreign language models are similar in output complexity. The median output grades of the four models (ChatGPT, Gemini, Kimi, and Ernie Bot) were all between grades 12 and 14. In comparison, the output rank of the domestic models, Kimi and Ernie Bot, are slightly higher, and the language style is academic. The text output of ChatGPT is closer to the popular science class, and the output text of the foreign model is more readable.

(2) Analysis of output stability

In terms of stability, the foreign model ChatGPT is the most prominent. The output text is very volatile, and the output text is consistent. Gemini's text output is slightly undulating and stable. In comparison, the class span of the domestic models Kimi and Ernie Bot is obviously larger and shows higher linguistic diversity.

(3) Extreme comparison

In Figure 3, we can observe that the extreme values found with ChatGPT had almost no high-grade or ultra-low-grade text output, indicating that ChatGPT-generated content was more stable. By contrast, there are many extreme points in Kimi and Ernie Bot, with ultra-high-grade output (top 20 above). The occurrence of the display model is complex and readable.

In general, ChatGPT has the best output stability of the four models in Chinese, and the Gemini model is medium. The domestic model Kimi has good readability, but the stability is poor, and the Ernie Bot model has the most extreme value, indicating that the text that it outputs is more volatile than that of Kimi.

3.4. Text Readability Influence Factor Analysis

This section uses the neural network model as one of the intelligent algorithms to analyze the influence factors affecting text readability for the four models and to calculate the relative importance of each input variable. The above study is compared with the four models of text readability (that is, the text readability and textual comprehension), and a quantitative study is carried out on the text readability and the text. There are few reports of readability studies in terms of the words of the text. The basic unit of text and the structure of the basic unit has a direct effect on the readability of the text, and the study of words and of word structure is of practical significance for text to be easily readable. This study can be used to understand the intrinsic nature of the readable nature of the text from the internal organization of the text. Based on this, this section intends to use multilayer perceptron (multilayer perceptron is one of many neural network model algorithms) to generate the influence factors that affect text readability, and the four models are compared. In this paper, we use text computing software to generate text computing FRES and FKGL indicators. The same calculation can be obtained by determining the AWPS index of the other points [9]: AWPS indicator, ASPW indicator, RL indicator, word indicator, and sentence indicator.

3.4.1. Neural Network Algorithm Construction

This paper uses multilayer perceptron (multilayer perceptron is one of many neural network model algorithms) to make research tools for text readability influence factors. Multilayer perceptron (MLP) is a typical feedforward neural network [29] and consists mainly of the input layer, the hidden layers, and the output layer. It learns the feature representation of the data through a fully connected (FC) structure and nonlinear activation function, and MLP has the characteristics of full connection, nonlinear activation, error backpropagation, and multilayer feature extraction [30]. MLP can be applied to text classification.

According to the construction requirements of a neural network model, input layer and output layer variables are required. There are five input layer variables (X) used in our study: AWPS, ASPW, sentence, word and SimHash. The output layer variable (Y) is RL, which indicates that the reading level variable. Among these reading levels, RL 1 means the reading level is “College”, and a higher RL means the reading level is “College graduate”. The results of the four models of the neural network model are shown in Figure 4.

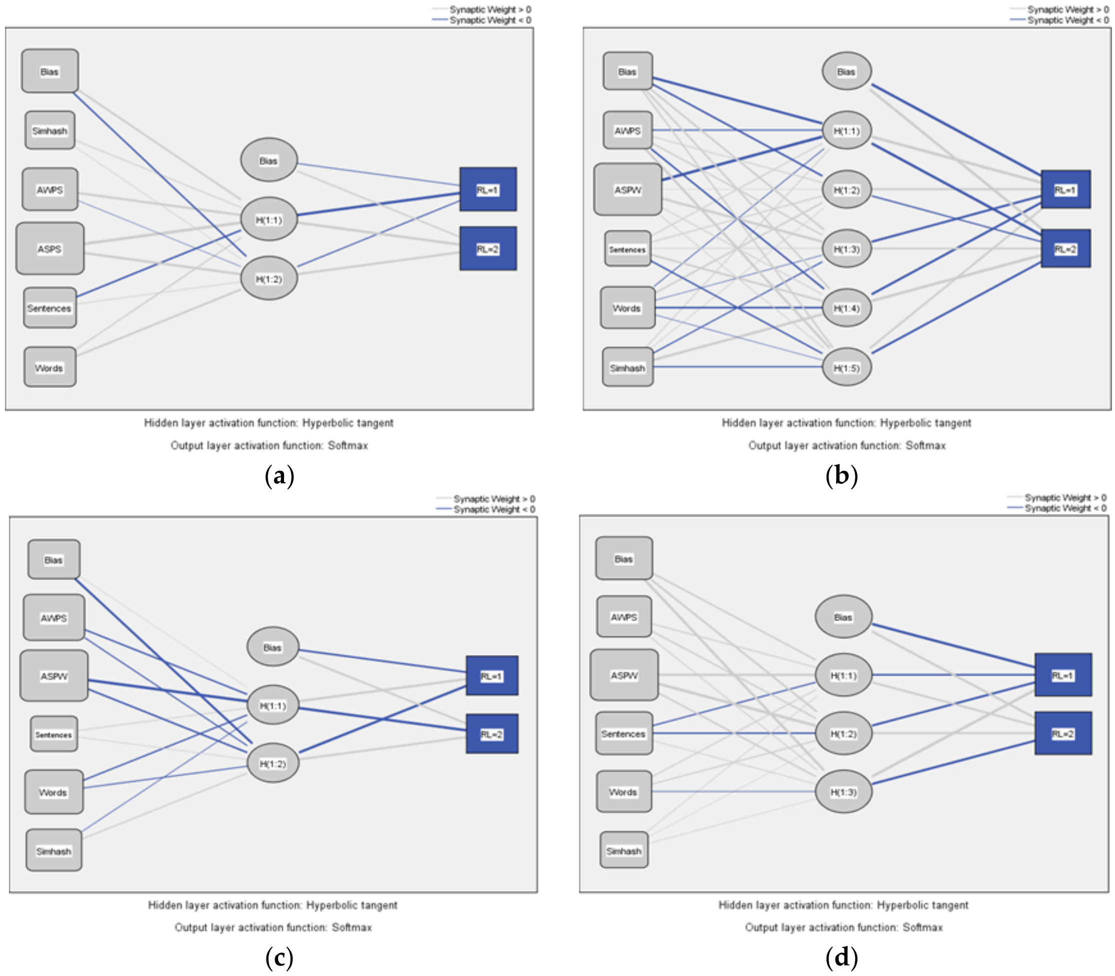
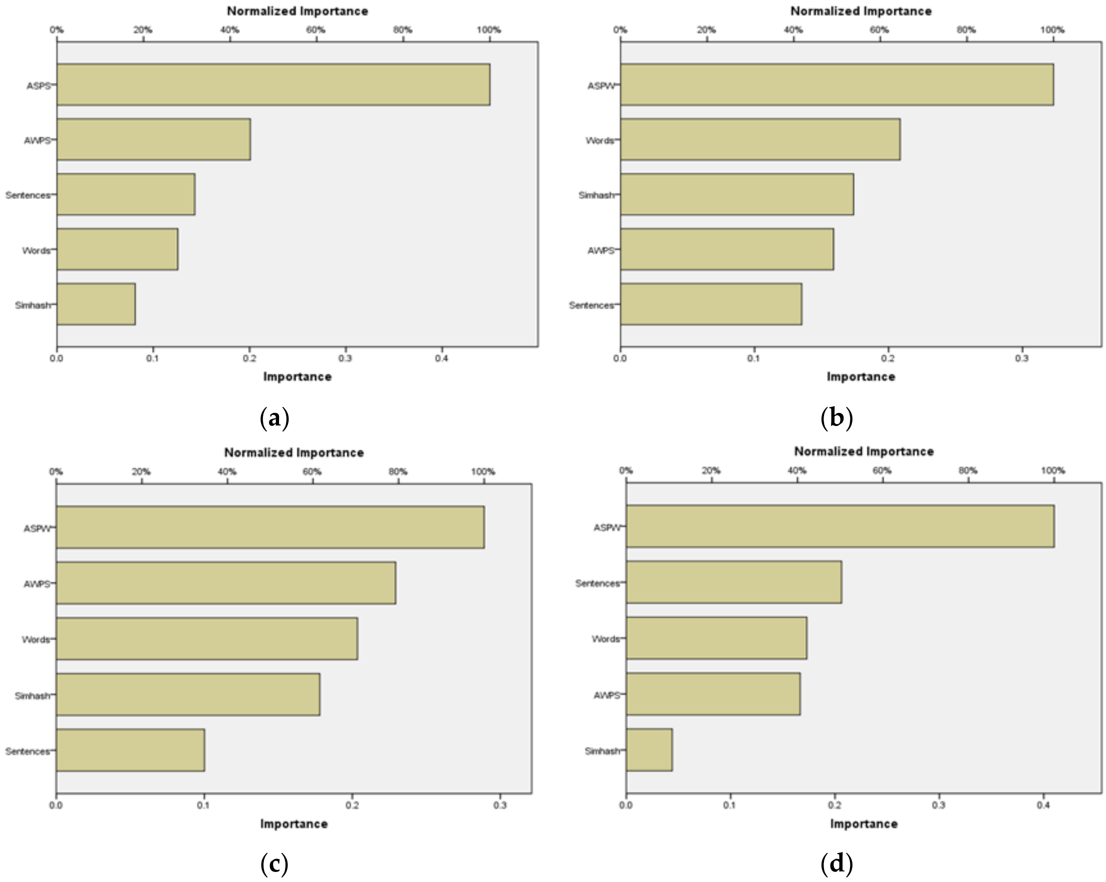


Figure 4. Diagram of the neural network structure of the four models. (**a**) ChatGPT neural network structure diagram; (**b**) Gemini neural network structure diagram; (**c**) Kimi neural network structure diagram; (**b**) Ernie Bot neural network structure diagram.

The neural network structure, which is built by four models, is composed of the input layer, hidden layer, and output. In this case, the hidden layer is the hyperbolic variable activation function, which is used to improve the function in the nonlinear relationship between the variables and helps to optimize the gradient propagation of the model. The output layer of the model is text-readable, and the four images of the four models are consistent. The influence of the input variables of each model is comparable to the influence of the model output variable.

3.4.2. Influence Factor Analysis

The study mentioned above, by building a model neural network training model that affects the output variable, obtains its influence path, analyzes the relative importance of the input variables of the output variable, and then provides the objective reference data for the quantitative comparison of the models [31]. In this study, we used SPSS software to conduct neural network calculation. The neural network model used in this article is a multilayer perceptron model (MLP model). After the model runs, the software can output the importance of the input variable directly, which indicates the relative importance of each input variable. The variable importance bar diagram corresponding to the model is shown in Figure 5.



(**a**) (**b**)

(**c**) (**d**)

Figure 5. Histogram of the importance of influencing variables. (**a**) ChatGPT affects the importance of variables histogram; (**b**) Gemini affects the importance of variables histogram; (**c**) Kimi affects the importance of variables histogram; (**b**) Ernie Bot affects the importance of variables histogram.

As can be seen from Figure 5, the readability of the text generated by the ChatGPT model is most affected by ASPS, followed by AWPS, while the variable SimHash has a lesser importance, indicating that the readability of the generated text is less affected by the accuracy of the text. The readability of the text generated by the Gemini model is most affected by the average number of syllables (ASPW), while the SimHash effect is slightly higher than that of ChatGPT, indicating that the text generated by Gemini is greatly affected by the accuracy of the text. For the domestic model, the readability of Kimi-generated text is most affected by ASPW, indicating that sentence length is still the main influencing factor. Sentence length (AWPS) is still the most important factor for the readability of Ernie Bot's generated text, followed by sentence factor, and it can be seen from the graph that the variable AWPS is much more important than the variable sentence. Other variables have a lesser impact.

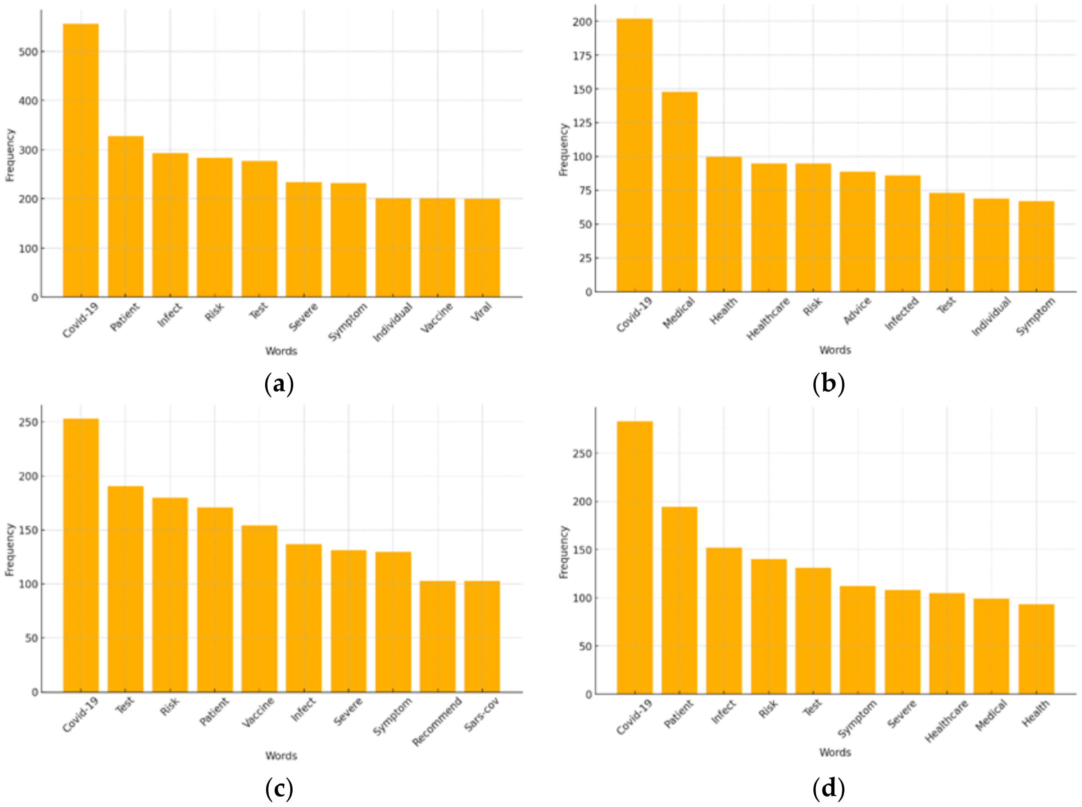
In conclusion, the influencing factor of sentence length (AWPS) is the most important factor affecting the readability of the generated text in all models. This conclusion shows that the average sentence length is the most critical factor in generating text readability. The influencing factor of SimHash has different degrees of influence in different models. The readability of the text generated by the foreign models ChatGPT and Gemini is less affected by SimHash, indicating that the foreign models have a high degree of innovation in generating text, and the generated text has little correlation with the CDC standard text. However, the domestic models Kimi and Ernie Bot are more sensitive to SimHash, indicating that the text generated by the domestic models is more similar to the CDC standard text, which also proves that the text accuracy affects the readability of the text.

3.5. Text Content Semantic Comparison

Text analysis is the process of extracting useful information from the text by means of mathematical statistics or related algorithms [32]. According to the purpose of this paper and the characteristics of the text of the answers, lexical analysis is proposed to extract the key words. Then, according to the high-frequency lexical induction theme of each text, the core thought of each text is understood. In the present study, there are several kinds of text theme-mining algorithms which require a small amount of text; thus, we use the inductive summary method to adapt to the research text feature [33].

3.5.1. Lexical Frequency Statistics

Based on the four large models and the lexical frequency data of COVID-19 generative text, the lexical frequency data statistics are obtained. The word frequency bar chart (10) and the word frequency ranking chart are shown in Figure 6. This paper analyzes the following aspects.



(**a**) (**b**)

(**c**) (**d**)

Figure 6. Generative histogram of text word frequency. (**a**) ChatGPT generative text word frequency diagram; (**b**) Gemini generative text word frequency diagram; (**c**) Kimi generative text word frequency diagram; (**b**) Ernie Bot generative text word frequency diagram.

"COVID-19" was the most frequently used word in the four big models, indicating that the models are highly focused on COVID-19. ChatGPT had the highest frequency of using "COVID-19" (556 times) and Gemini had the lowest (202 times), suggesting that ChatGPT used the term most intensively, while Gemini may have more frequently used alternative expressions. "Risk" and "test" have a high ranking in the model, showing that risk assessment and testing are key topics in the model's response. The frequency distribution of the word frequency distribution is higher than other models, which may indicate that the output text is longer or that the output text is more likely to be reused for a specific term.

3.5.2. Topic Mining

In this paper, the text of the various models is generated by statistical induction of the various models. A frequency table of the text generated by the four models is shown in Table 1.

Table 1. The models' generative word frequency table (the frequency number is the top 10).

	ChatGPT		Gemini		Kimi		Ernie Bot	
Sort	Word	Frequency	Word	Frequency	Word	Frequency	Word	Frequency
1	COVID-19	556	COVID-19	202	COVID-19	253	COVID-19	283
2	Patient	328	Medical	148	Test	191	Patient	194
3	Infect	293	Health	100	Risk	180	Infect	152
4	Risk	284	Healthcare	95	Patient	171	Risk	140
5	Test	277	Risk	95	Vaccine	154	Test	131
6	Severe	233	Advice	89	Infect	137	Symptom	112
7	Symptom	231	Infected	86	Severe	131	Severe	108
8	Individual	201	Test	73	Symptom	130	Healthcare	105
9	Vaccine	201	Individual	69	Recommend	103	Medical	99
10	Viral	200	Symptom	67	Sars-cov	103	Health	93

Statistical induction yields four model research themes. Through the analysis of Table 1, the following conclusions can be drawn: among the top 10 word frequency rankings, the words "Patient", "Infect", and "Vaccine" appear more in the text generated by ChatGPT, indicating that the model pays more attention to epidemiological knowledge. The words "Medical", "Healthcare", and "Advice" appear more frequently in the text generated by the model Gemini, indicating that the field of medical care is more emphasized in the text generated by the model. The words "Risk", "Symptom", and "Recommend" appear more often in the domestic model Kimi, indicating that the text generated by the model pays more attention to multidisciplinary background information or early risk prevention. The words "Patient", "Infect", and "Healthcare" appear more often in the text generated by the model Ernie Bot, indicating that the model pays more attention to medical clinical topics. In conclusion, the domestic models (Kimi, Ernie Bot) are more suitable for clinical testing, medical system research, health consultation, etc. The foreign models (ChatGPT, Gemini) were more focused on epidemiological analysis, vaccine research, and disease severity assessment. Generally speaking, the domestic model is more applicable, and the foreign model is more professional.

4. Results

This paper studies the application performance of generative artificial intelligence models in terms of providing knowledge and answers regarding infectious diseases. This study systematically compares the differences between the foreign models ChatGPT and Gemini, on the one hand, and the domestic models Kimi and Ernie Bot in the formation of text, the ease with which the text can be read and understood, and the semantic content. Furthermore, this study discusses the influence of generative artificial intelligence models in the dissemination of public health information. A concrete in-depth discussion is carried out below, and our findings are presented.

In this paper, the accuracy of artificial intelligence models is compared. The results show that the text of the foreign models ChatGPT and Gemini is similar to the standard answer provided by the CDC, which shows that these models have better accuracy and that the foreign model has a significant advantage for data training in the public health sector. In contrast, the text generated by the domestic models (Kimi and Ernie Bot) is more volatile and less stable than the CDC standard text. This may occur due to the number of training data in the domestic model, which limits the accuracy of the domestic model in the formation of English text. By comparing the accuracy of the text, we need to provide sufficient data to ensure that the semantic accuracy and professionalism of the text are unified.

The models used in this study generate an indicator of the readability of text, which is the Flesch Reading Ease Score (FRES index). The FRES index reveals the difference between the Chinese and foreign models in terms of the readability of the text that they produce. The empirical results show that the FRES index of the domestic models (Kimi and Ernie Bot) is higher, indicating that the domestic models produce text that can be read easily. The domestic model generates text that is easier to understand and language choreography that is more popular. In comparison, the FRES values of the foreign models (ChatGPT and Gemini) are lower, indicating that the generated text may be more professional and include words that are less frequently used by most people. However, the foreign models (ChatGPT and Gemini) generate text that is logical, rigorous, and suitable for professional reading. In the examination of the ease with which text can be read, we should focus on the organic combination of professionalism and universality in the dissemination of public health information.

The model generates text intelligibility using another FKGL metric in the readability metric. The empirical results show that the text of the foreign models (ChatGPT and Gemini) is above 12, and the level of understanding may correspond to the higher education level of the United States, which shows that the foreign model is suitable for people with higher levels of education. The FKGL value of the domestic model is between 810 and the average level of an individual with a background of secondary education background. The results of this study show that the training of generative artificial intelligence in the future needs to be understood in order to provide timely, accurate, and easy-to-understand professional text for the prevention and control of major infectious diseases.

In addition to the accuracy, readability, and comprehensibility of the four models, this work also used the human work neural model (multilayer perceptron) to study the effect of text readability. The empirical results show that the foreign models' generative text is more focused on the complexity of the language structure, especially the number of lexical syllables (the ASPW) and the length of the sentence (the length of the sentence AWPS), which affects the readability of the language. The domestic model also emphasizes the number of text sentences, reflecting stronger localization language. The linguistic differences of the text in the domestic and international models embody the different characteristics of Chinese and English language.

The results of topic mining show that the topic content of the foreign models (ChatGPT and Gemini) focuses on "vaccination", "virus mutation", and "protection suggestions", while domestic models (Kimi and Ernie Bot) mainly focus on practical knowledge such as "symptom recognition" and "test suggestions". In general, the textual subject content of the Chinese and English models, at home and abroad, reflects the cultural and traditional differences in the training data behind them.

5. Conclusions

The central aim of this research was to systematically evaluate the differences in performance between four mainstream generative AI models (ChatGPT, Gemini, Kimi, and Ernie Bot), at home and abroad, based on the Q&A content of COVID-19 prevention and control provided by the CDC as a text standard. The evaluation of their performance includes four aspects: the ability to generate text, the production of text that is easy to understand, the production of text that is readable, and text semantic topics. The empirical results show that the accuracy and professional performance of ChatGPT and Gemini (foreign models) are better, but the complexity of the formation language is high. The text of the domestic models (Kimi and Ernie Bot) is more popular, suitable for the health-related questions of ordinary people, but text specialization needs to be strengthened. In addition, domestic and international models have obvious differences in language generation strategy, audience suitability, and semantic coverage, and different models vary in terms of their language style and propagation preferences. Therefore, according to the results of the research in this study, it is suggested that in the future dissemination of knowledge regarding infectious diseases, we should pay more attention to the knowledge foundation and level of understanding of the general public and emphasize the organic unity of professionalism and universality in the training of artificial intelligence data, thus improving the effectiveness and accuracy of public health knowledge propagation.

In this paper, we studied the application of generative artificial intelligence in terms of its general knowledge of and ability to provide answers about the prevention and control of infectious diseases. In addition, we implemented three aspects of innovation: (1) theoretical innovation. Through the introduction of a generative artificial intelligence model analysis tool, the research boundary of public health information dissemination is expanded. Traditional public health communication is concentrated in the media and government propaganda, and generative artificial intelligence models are rarely applied to the study of the knowledge of infectious diseases. This study marks the first time that generative artificial intelligence is incorporated into the knowledge and answer framework of infectious diseases, and it has enriched the theoretical map of the spread of infectious diseases. The theoretical innovation of this study is also reflected in the selection of multiple indexes that generate text performance. This study selects the accuracy, readability, comprehension, readability factor, and five indexes of the text and overcomes the limitations of the traditional evaluation of individual indexes. (2) Innovative methods. In this paper, innovation is embodied in the multidimensional analysis framework of text accuracy and readability index, and, by introducing the neural network model, the ability to identify the capacity to generate text is realized, and a mechanism for comparing the quality of texts is produced. This study enhances the interpretative nature of text through text topics. (3) Application innovation. The application innovation of this study is mainly embodied in the introduction of major model performance comparison perspectives. In this study, we compare the output performance of foreign models and domestic models under the same task and reveal the differences between Chinese and foreign models in terms of semantic accuracy, popularity, and professionalism, which provides a valuable economic reference for the development of artificial intelligence.

In general, the results of this study can serve directly in the healthcare system and in the prevention and control department. In the use of or supervised generation of AI to conduct health knowledge dissemination, the evaluation of the quantity of the information is based on technical advice, thereby promoting the generation of artificial intelligence models to better serve the people and contribute to improving the level of control of major infectious diseases in the world.

6. Study Limitations

Although this study has carried out an effective exploration of the framework design of model evaluation, empirical data comparison, and multidimensional index extraction, it still contains a limitation which needs to be further improved in future studies.

The uniqueness of the standard answer limits the overall performance of the model. In this study, the answers to 53 COVID-19-related questions provided by the CDC were used as the standard answers, and the responses of the artificial intelligence models were similar. However, the answers of generative artificial intelligence models are diverse, and the uniqueness of the answer and the diversity of the generated text cannot be measured by the actual professional level of the model.

The model resolves the paradox of the dynamic nature of the release of the version. The research objects—that is, the models used (ChatGPT, Gemini, Kimi, Ernie Bot)—are all iteratively updated AI products, and the answer results may fluctuate with time and with different words revealed by users.

The topic analysis does not conduct emotional classification of the text. Although this study carried out text theme analysis, it did not carry out emotional analysis of the text, and in the process of actual dissemination of medical information, the emotionality of text and language is often associated with the efficiency of the transmission.

The quantity and limitation of the knowledge domain is a further shortcoming of this study. This study was based on the 53 COVID-19-related problems of the CDC, which have certain limitations. The knowledge sector is small and does not cover the wider range of infectious diseases. In future research, we will attempt to apply information about a wider range of diseases in order to improve the prevention and control of major infectious diseases.

Author Contributions: Conceptualization, Z.L. and Y.K.; methodology, Z.L., Y.K., G.L. and Z.L.; formal Analysis, G.L.; preparation of original drafts, Z.L., Z.L. and Y.K.; review and final approval, Z.L., Y.K., X.L., Z.L. and G.L. All authors have read and agreed to the published version of the manuscript.

Funding: National Social Science Foundation of China (No. 20XTJ004): " Spatio-temporal Modeling, Monitoring, Prevention and Control of the Spread of Major Infectious Diseases Based on Big Data".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used and analyzed in this study are available from the corresponding authors upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. CHAKRABORTY I, MAITY P. COVID-19 outbreak: Migration, effects on society, global environment and prevention. *Science of the total environment*, **2020**, 728(1), 138882.
2. JEE Y. WHO international health regulations emergency committee for the COVID-19 outbreak. *Epidemiology health Communication*, **2020**, 42(1), e2020013.
3. SARKER R, ROKNUZZAMAN A, HOSSAIN M J, et al. The WHO declares COVID-19 is no longer a public health emergency of international concern: benefits, challenges, and necessary precautions to come back to normal life. *International Journal of Surgery*, **2023**, 109(9), 2851-2852.
4. EVANS A, ALSHURMAN B A, SEHAR H, et al. Monkeypox: a mini-review on the globally emerging Orthopoxvirus. *International Journal of Environmental Research Public Health*, **2022**, 19(23), 15684.
5. LI G, HILGENFELD R, WHITLEY R, et al. Therapeutic strategies for COVID-19: progress and lessons learned. *Nature Reviews Drug Discovery*, **2023**, 22(6), 449-475.
6. BAKER R E, MAHMUD A S, MILLER I F, et al. Infectious disease in an era of global change. *Nature reviews microbiology*, **2022**, 20(4), 193-205.
7. ZHOU T, LI S. Understanding user switch of information seeking: From search engines to generative AI. *Journal of Librarianship Information Sciences*, **2024**, 09610006241244800.
8. BHARTI I, CHAUHAN K, AGGARWAL P. Generative AI: Next Frontier for Competitive Advantage. *Enhancing Communication and Decision-Making With AI. IGI Global*. **2025**, 1-36.
9. ÖZTüRK Z, BAL C, ÇELIKKAYA B N. Evaluation of Information Provided by ChatGPT Versions on Traumatic Dental Injuries for Dental Students and Professionals. *Dental Traumatology*, **2025**, 1-10.
10. SIEBIELEC J, ORDAK M, OSKROBA A, et al. Assessment Study of ChatGPT-3.5's performance on the final Polish Medical examination: Accuracy in answering 980 questions. *Healthcare*, **2024**, 12(16), 1637.
11. WANG G, GAO K, LIU Q, et al. Potential and limitations of ChatGPT 3.5 and 4.0 as a source of COVID-19 information: comprehensive comparative analysis of generative and authoritative information. *Journal of Medical Internet Research*, **2023**, 25(1), e49771.
12. ZONG H, LI J, WU E, et al. Performance of ChatGPT on Chinese national medical licensing examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. **2024**, 24(1), 1-9.
13. YANAGITA Y, YOKOKAWA D, UCHIDA S, et al. Accuracy of ChatGPT on medical questions in the national medical licensing examination in Japan: evaluation study. *JMIR Formative Research*, **2023**, 7(1), e48023.
14. SADEQ M A, GHORAB R M F, ASHRY M H, et al. AI chatbots show promise but limitations on UK medical exam questions: a comparative performance study. *Scientific Reports*, **2024**, 14(1), 1-11.
15. MEO S A, AL-MASRI A A, ALOTAIBI M, et al. ChatGPT knowledge evaluation in basic and clinical medical sciences: multiple choice question examination-based performance. *Healthcare*, **2023**, 11(14), 2046.
16. SUMBAL A, SUMBAL R, AMIR A. Can ChatGPT-3.5 pass a medical exam? A systematic review of ChatGPT's performance in academic testing. *Journal of medical education curricular development*, **2024**, 11(1), 23821205241238641.
17. DE VITO A, GEREMIA N, MARINO A, et al. Assessing ChatGPT's theoretical knowledge and prescriptive accuracy in bacterial infections: a comparative study with infectious diseases residents and specialists. *Infection, disease health Communication*, **2024**, 1-9.
18. ONDER C E, KOC G, GOKBULUT P, et al. Evaluation of the reliability and readability of ChatGPT-4 responses regarding hypothyroidism during pregnancy. *Scientific Reports*, **2024**, 14(1), 1-8.
19. BISWAS S, LOGAN N S, DAVIES L N, et al. Assessing the utility of ChatGPT as an artificial intelligence-based large language model for information to answer questions on myopia. **2023**, 43(6), 1562-1570.

20. AHMED W M, AZHARI A A, ALFARAJ A, et al. The Quality of AI-Generated Dental Caries Multiple Choice Questions: A Comparative Analysis of ChatGPT and Google Bard Language Models. *Heliyon*, **2024**, 10(7), e28198.
21. SHEN S A, PEREZ-HEYDRICH C A, XIE D X, et al. ChatGPT vs. web search for patient questions: what does ChatGPT do better?. **2024**, 281(6), 3219-3225.
22. CONTROL C F D, PREVENTION, CONTROL C F D, et al. Clinical questions about COVID-19: questions and answers [EB/OL].(2020-8-4).
23. TOPRAK A, TURAN M. Automated thematic dictionary creation using the web based on WordNet, Spacy, and Simhash. *Data Information Management*, **2024**, 100088.
24. SADOWSKI C, LEVIN G. Simhash: Hash-based similarity detection[EB/OL].(2007-12-13) .
25. ELEYAN D, OTHMAN A, ELEYAN A. Enhancing software comments readability using flesch reading ease score. *Information Development*, **2020**, 11(9), 430.
26. GRABEEL K L, RUSSOMANNO J, OELSCHLEGEL S, et al. Computerized versus hand-scored health literacy tools: a comparison of Simple Measure of Gobbledygook (SMOG) and Flesch-Kincaid in printed patient education materials. **2018**, 106(1): 38.
27. KARNAN N, FRANCIS J, VIJAYVARGIYA I, et al. Analyzing the Effectiveness of AI-Generated Patient Education Materials: A Comparative Study of ChatGPT and Google Gemini. *Cureus*, **2024**, 16(11), e74398-e74398.
28. ONAN A, KORUKOGLU S, BULUT H. LDA-based topic modelling in text sentiment classification: An empirical analysis. *Int J Comput Linguistics Appl*, **2016**, 7(1), 101-119.
29. DESAI M, SHAH M. An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN). *Clinical eHealth*, **2021**, 4(1), 1-11.
30. HE X, CHEN Y. Modifications of the multi-layer perceptron for hyperspectral image classification. *Remote Sensing*, **2021**, 13(17), 3547.
31. YILMAZ I, KAYNAR O. Multiple regression, ANN (RBF, MLP) and ANFIS models for prediction of swell potential of clayey soils. *Expert systems with applications*, **2011**, 38(5), 5958-5966.
32. PANG X, WAN B, LI H, et al. MR-LDA: an efficient topic model for classification of short text in big social data. *International Journal of Grid High Performance Computing*, **2016**, 8(4), 100-113.
33. DEBORTOLI S, MüLLER O, JUNGLAS I, et al. Text mining for information systems researchers: An annotated topic modeling tutorial . *Communications of the Association for Information Systems*, **2016**, 39(1), 7.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.