

Identification of Online Health Information Using Large Pre-trained Language Models: An Effectiveness Assessment

Dongmei Tan, Yi Huang, Ming Liu, Ziyu Li, Xiaoqian Wu, Cheng Huang

Submitted to: Journal of Medical Internet Research
on: December 31, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript.....	5
Supplementary Files.....	35
Figures	36
Figure 1.....	37
Figure 2.....	38
Figure 3.....	39
Figure 4.....	40
Figure 5.....	41
Figure 6.....	42
Figure 7.....	43
Multimedia Appendixes	44
Multimedia Appendix 1.....	45

Identification of Online Health Information Using Large Pre-trained Language Models: An Effectiveness Assessment

Dongmei Tan¹ MS; Yi Huang¹ MS; Ming Liu¹ MS; Ziyu Li² MS; Xiaoqian Wu³ MS; Cheng Huang¹ MD

¹ College of Medical Informatics Chongqing Medical University Chongqing CN

² Human Resources Department Army Medical Center Army Medical University (The Third Military Medical University) Chongqing CN

³ Department of Quality Management Army Medical Center Army Medical University (The Third Military Medical University) Chongqing CN

Corresponding Author:

Cheng Huang MD

College of Medical Informatics
Chongqing Medical University
1 Yixueyuan Road Yuzhong District
Chongqing
CN

Abstract

Background: Online health information is widely available, but a significant portion of it is inaccurate or misleading, including exaggerated, incomplete, or unverified claims. Such misinformation can significantly influence public health decisions and pose serious challenges to healthcare systems. With advances in artificial intelligence and natural language processing, large pre-trained language models (LLMs) have shown promise in identifying and distinguishing misleading health information, though their effectiveness in this area remains under explored.

Objective: This paper aims to evaluate the performance of four mainstream LLMs (ChatGPT-3.5, ChatGPT-4.0, ERNIE Bot, and iFLYTEK Spark) in the identification of online health information, providing empirical evidence for their practical application in this field.

Methods: Web scraping was used to collect data from online rumor-refuting websites, resulting in 2,708 samples of online health information, including both true and false claims. The four LLMs' APIs were employed for authenticity verification, with expert results as benchmarks. Model performance was evaluated using semantic similarity, accuracy, recall, F1 score, and content analysis.

Results: The study found that the four models performed well in identifying online health information. Among them, ChatGPT-4.0 achieved the highest accuracy at 87.27%, followed by ERNIE Bot at 87.25%, iFLYTEK Spark at 87.00%, and ChatGPT-3.5 at 81.82%. Furthermore, text length and semantic similarity analysis showed that ERNIE Bot had the highest similarity to expert texts, while ChatGPT-4.0 showed good overall consistency in its explanations. Further analysis suggested that the primary areas of misjudgment included child health, drug nutrition, disease prevention, lifestyle, food safety, and diet. Overall, the research suggests that LLMs have potential in online health information identification; however, their understanding of certain specialized health topics may require further improvement.

Conclusions: The results demonstrate that while these models show potential in providing assistance, their performance varies significantly in terms of accuracy, semantic understanding, and cultural adaptability. Key findings highlight the models' ability to generate accessible and context-aware explanations; however, they fall short in areas requiring specialized medical knowledge or updated data, particularly for emerging health issues and context-sensitive scenarios. Significant discrepancies were observed in the models' ability to distinguish scientifically verified knowledge from popular misconceptions and in their stability when processing complex linguistic and cultural contexts. These challenges reveal the importance of refining training methodologies to improve the models' reliability and adaptability. Future research should focus on enhancing the models' capability to manage nuanced health topics and diverse cultural and linguistic nuances, thereby facilitating their broader adoption as reliable tools for online health information identification.

(JMIR Preprints 31/12/2024:70733)

DOI: <https://doi.org/10.2196/preprints.70733>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

Please make my preprint PDF available to anyone at any time (recommended).

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

✓ **No, I do not wish to publish my submitted manuscript as a preprint.**

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in [http://www.jmir.org/](#)

No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Identification of Online Health Information Using Large Pre-trained Language Models: An Effectiveness Assessment

Abstract

Background: Online health information is widely available, but a significant portion of it is inaccurate or misleading, including exaggerated, incomplete, or unverified claims. Such misinformation can significantly influence public health decisions and pose serious challenges to healthcare systems. With advances in artificial intelligence and natural language processing, large pre-trained language models (LLMs) have shown promise in identifying and distinguishing misleading health information, though their effectiveness in this area remains under-explored.

Objective: This paper aims to evaluate the performance of four mainstream LLMs (ChatGPT-3.5, ChatGPT-4.0, ERNIE Bot, and iFLYTEK Spark) in the identification of online health information, providing empirical evidence for their practical application in this field.

Methods: Web scraping was used to collect data from online rumor-refuting websites, resulting in 2,708 samples of online health information, including both true and false claims. The four LLMs' APIs were employed for authenticity verification, with expert results as benchmarks. Model performance was evaluated using semantic similarity, accuracy, recall, F1 score, content analysis, and credibility.

Results: The study found that the four models performed well in identifying online health information. Among them, ChatGPT-4.0 achieved the highest accuracy at 87.27%, followed by ERNIE Bot at 87.25%, iFLYTEK Spark at 87.00%, and ChatGPT-3.5 at 81.82%. Furthermore, text length and semantic similarity analysis showed that ERNIE Bot had the highest similarity to expert texts, while ChatGPT-4.0 showed good overall consistency in its explanations. Additionally, the credibility assessment results indicated that ChatGPT-4.0 provided the most reliable evaluations. Further analysis suggested that the highest misjudgment probabilities in the LLMs occurred within the topics of Food and Maternal-Infant Nutrition Management and Nutritional Science and Food Controversies. Overall, the research suggests that LLMs have potential in online health information identification; however, their understanding of certain specialized health topics may require further improvement.

Conclusion: The results demonstrate that while these models show potential in providing assistance, their performance varies significantly in terms of accuracy, semantic understanding, and cultural adaptability. The principal findings highlight the models' ability to generate accessible and

context-aware explanations; however, they fall short in areas requiring specialized medical knowledge or updated data, particularly for emerging health issues and context-sensitive scenarios. Significant discrepancies were observed in the models' ability to distinguish scientifically verified knowledge from popular misconceptions and in their stability when processing complex linguistic and cultural contexts. These challenges reveal the importance of refining training methodologies to improve the models' reliability and adaptability. Future research should focus on enhancing the models' capability to manage nuanced health topics and diverse cultural and linguistic nuances, thereby facilitating their broader adoption as reliable tools for online health information identification.

Keywords: Large pre-trained language models; LLMs; online health information; information identification; text similarity analysis; performance evaluation; ChatGPT; text generation; LDA

Introduction

The widespread global usage of Internet media has had a tremendous influence on social interactions between individuals, communities, and societies [1]. Kepios analysis shows that active social media user identities had grown to 4.95 billion by the start of October 2023, equating to 61.4 percent of the world's total population [2]. Recent years have witnessed a dramatic increase in consumer online health information seeking [3]. As access to information has become more convenient, however, the issue of false health information online has become increasingly prominent [4], posing significant risks to public health. Although there is no conclusive evidence that health misinformation spreads more broadly than scientific information, health misinformation reliably leads to misperceptions on health issues, which can have serious consequences for individuals' health decisions [5]. In some cases, the spread of false health information has even led to uncontrollable consequences [6]. Furthermore, with the rise of fake news [7], internet users are facing growing challenges in discerning the authenticity of information [8]. False health information, often emotionally charged and highly shareable, spreads rapidly and has far-reaching negative effects. While public demand for understanding and accessing health information has increased, this has inadvertently contributed to the widespread dissemination of false health claims [9,10]. Therefore, scientifically evaluating online health information has become a critical issue that requires urgent attention.

The task of online information identification involves the automated classification and identification of various types and sources of information within vast amounts of online data, with the goal of effectively identifying and utilizing valuable information [11]. The identification of online health information is a subtask within this broader framework, primarily focused on classifying and discerning health-related content through textual analysis and feature extraction. This task heavily relies on natural language processing techniques for in-depth analysis, enabling the evaluation of the authenticity of online health information. This task not only requires models to have strong text classification capabilities but also necessitates their ability to identify key elements in online health information, uncover potential misleading content, and comprehensively assess the accuracy and reliability of the information. Compared to information in other fields such as entertainment, education, finance, and technology, online health information carries higher social sensitivity and risk due to its direct impact on public life and health [12]. Furthermore, as the internet becomes the primary source of health information for the public, the accuracy and reliability of online health information have significant implications for public health [13]. Therefore, this study evaluates the practical performance of various pre-trained language models in online health information identification, aiming to better understand their effectiveness in addressing false health information and contributing to further research and practical applications.

On November 30, 2022, OpenAI launched ChatGPT. This large-scale pre-trained language model quickly garnered widespread attention due to its advanced conversational abilities [14]. ChatGPT has shown impressive performance in areas such as search engine optimization and code development, and its extensive pre-training dataset also opens possibilities for application in other specialized fields [15]. A review of previous studies reveals the evolution of text classification techniques, from traditional machine learning algorithms to deep learning models [16,17], and more recently, to the emergence of pre-trained models, which have become crucial tools for identifying online health information [18]. Models like ChatGPT have also demonstrated relatively strong natural language processing and text generation capabilities in the medical field [19,20]. Reports indicate that ChatGPT, without specialized medical training, managed to pass the United States Medical Licensing Examination with a "qualified" score, highlighting its potential in understanding and applying fundamental medical knowledge [21]. However, despite the advancements made by models such as ChatGPT across various domains, their effectiveness in the task of online health information identification still requires further research and evaluation.

Recently, research on the identification of false health information has received increasing attention. False health information has been defined as health-related information that lacks scientific evidence [22], which provides an important theoretical foundation for subsequent studies. It was found that false news spreads faster, farther, and to a larger audience on Twitter (now rebranded as X), with its wider dissemination primarily driven by the novelty of the information and emotional responses, especially in political contexts [23]. Besides, a social network analysis of Twitter data regarding COVID-19 and the 5G conspiracy theory revealed the characteristics of misinformation dissemination on social media platforms [24]. Another study tracked social media discourse during the COVID-19 pandemic and developed a public Twitter dataset on coronavirus to study the patterns of misinformation spread [25]. It has also been noted that people's insecurity about institutions led to many conspiracy theories and low-quality information during the pandemic, which deepened distrust in science and institutions and amplified the impact of these views and ideologies [26].

With the rapid advancement of AI-generated content technologies, particularly the widespread use of LLMs like ChatGPT, the focus has shifted from traditional machine learning to evaluating the authenticity and reliability of AI-generated content. One study pointed out the valuable role that LLMs could play in healthcare, particularly in supporting clinical work, research, and education, while also noting concerns about data privacy and dependence on these models [27]. Further analysis highlighted that ChatGPT-4.0 carries a higher risk of generating false information, with misinformation generation rates exceeding 80% [28]. This shift not only demonstrates the potential of AI technology in false health information identification but also introduces new risks and challenges [29].

In conclusion, the unique emotional characteristics and significant impact on the dissemination of false health information highlight the urgent need for continuous optimization and improvement of current identification methods. Therefore, this study investigates the feasibility of applying four mainstream LLMs (ChatGPT-3.5, ChatGPT-4.0, ERNIE Bot, and iFLYTEK Spark) to identify online health information. The study is structured into three phases: data collection and preprocessing, model application and evaluation, and performance summary, as illustrated in Figure 1. In the first phase, data were gathered from domestic debunking platforms, followed by preprocessing to create a dataset containing 2,708 pieces of online health information. The second phase focused on verifying this information using the four LLMs. The results from these models were then analyzed

through data evaluation. Finally, the third phase involved analyzing the models' identification performance, and assessing their strengths, limitations, and practical application potential. Through empirical analysis, this study systematically evaluates the identification capabilities and performance of these models in real-world application scenarios. The findings of this study offer insights into the potential of LLMs for identifying false health information, which could inform future improvements in health misinformation detection methods.

Methods

Data Collection and Preprocessing

The dataset used in this study comes from Tencent Fact Check, a Chinese platform that verifies misinformation by cross-referencing authoritative official sources and gathering debunking statements from various platforms. Verified sources include the World Health Organization (WHO), the National Drug Administration (NMPA), and others. To ensure the clarity and completeness of the data sources, the appendix provides the website and specific verification sources. The platform conducts professional reviews to assess the authenticity of the information, ensuring the accuracy and reliability of the data labels. A Python-based web crawler collected a total of 3,720 pieces of online health information, each annotated with authenticity labels and accompanied by expert or official explanations. During data preprocessing, the data underwent cleaning, label standardization, and format unification. The research team manually reviewed and removed redundant or semantically similar information, resulting in 2,708 pieces of health information, which were then stored in an Excel file. Table 1 presents a portion of the online health information in the dataset, including expert assessments and explanatory texts. Of these, 25.11% were labeled as true, while 74.89% were labeled as false. This distribution reflects the relatively high proportion of false information on the platform, while also including a certain proportion of authentic information.

Table 1. Example of Network Online Health Information Data (Partial).

API Calls

In this
study, a

Information Content	Expert Assessment	Expert or Institution Explanatory Text
Cancer Cannot Be Prevented	FALSE	<i>"Some cancers can be prevented, while others do not have effective prevention methods. However, even for preventable cancers, taking certain measures does not guarantee a 100% certainty of not developing cancer; rather, it reduces the risk and probability of occurrence. Many cancers lack specific preventive measures, but for certain preventable cancers, such as lung cancer, there are numerous associated factors, both known and unknown. What we can do is to stay as far away as possible from known risk factors to minimize risk. Additionally, there are many cancers for which risk can be reduced through specific actions. For example, preventing hepatitis B virus infection can lower the risk of liver cancer; receiving the HPV vaccine and practicing safe hygiene can reduce the risk of cervical cancer; and changing poor dietary habits can decrease the risk of esophageal and gastric cancers, among others."</i>
Receiving the COVID-19 Vaccine Increases Breast Size	FALSE	<i>"Recently, some women who received the COVID-19 vaccine have noticed signs of breast enlargement. This phenomenon is actually a result of the body's immune response, causing "swelling of the axillary lymph nodes," which is not unusual and is quite normal. This adverse reaction typically subsides on its own after a few weeks. The actual change occurs in the lymph nodes, not in the breast tissue, and receiving the COVID-19 vaccine does not have a breast enhancement effect. Lymph node swelling is not limited to the COVID-19 vaccine; it can be observed with many other vaccinations as well, so there is no need for excessive concern. However, lymph node swelling can also occur when the body is experiencing a bacterial or viral infection, in which case it is important to identify the specific cause and consider appropriate treatment."</i>
Over 90% of Hypertension Patients Have Primary Hypertension	TRUE	<i>"Hypertension is classified into two main categories: primary hypertension and secondary hypertension. The majority of hypertension cases in the population are primary hypertension, accounting for over 90%. Secondary hypertension constitutes a small portion, approximately 5-10%. The occurrence of primary hypertension is often due to multiple factors rather than a single cause. The increase in blood pressure is related to the patient's living and working environment, dietary and exercise habits, and is typically the result of a combination of various influences. To prevent the rise in blood pressure, it is essential to address risk factors such as obesity, smoking, and alcohol consumption, adopt healthy dietary and exercise habits, and maintain a positive emotional state."</i>
Children Cannot Get Diabetes	FALSE	<i>"The occurrence of diabetes is the result of the interaction between genetic and environmental factors. The genes associated with type 1 diabetes come from both parents, and three environmental factors—climate, viral infections, and infant feeding practices—can also influence the development of type 1 diabetes. If one or both parents have type 1 diabetes, the likelihood of their children developing the condition is higher. Type 2 diabetes has a stronger familial hereditary tendency, which may be related to a family history of obesity and a high degree of similarity in unhealthy dietary and lifestyle habits within families. A healthy lifestyle can help delay or prevent the onset of type 2 diabetes."</i>

combination of API calls and prompt design was employed to obtain classification labels and explanatory texts generated by four LLMs for online health information, thereby facilitating subsequent evaluation. Specifically, the API calls served as the method of interaction with each model, while the prompt design ensured that these models produced consistent response formats.

Table 2 summarizes the key steps involved, along with the corresponding code functions, providing an overview of the entire process from environment setup to model interaction.

(1) Data Loading and Preprocessing: Health-related information was initially loaded from an Excel file. Each record was individually processed and sent to the respective pre-trained language models for verification to ensure data integrity and accuracy.

(2) API Calls: This study utilized specific functions to interact with various pre-trained language models. The `open35_response(query)` and `open4_response(query)` functions were employed to communicate with ChatGPT-3.5 and ChatGPT-4.0, respectively. The `spark35_response(query)` function enabled real-time interaction with Spark AI, while the `wenxin_response(query)` function was used to send queries to the Wenxin AI service. These functions transmitted user queries and received corresponding results to evaluate the authenticity of online health information, returning classification labels and their explanation texts.

(3) Access Token Retrieval: To establish secure connections with the Wenxin API, the `get_access_token()` function was used to generate an access token. The function authenticated using the API key and API Secret, and the obtained token was used in subsequent API requests to ensure secure and accurate data transmission.

(4) Overall Process Management: In the main function, the WebSocket connection for the Spark API was initialized, managing the overall query workflow. After retrieving data from the various models, the results were stored and formatted to facilitate the generation of a comprehensive view of online health information evaluations.

For effective evaluation, a uniform prompt structure was employed across all models to ensure adherence to the same classification criteria and response format. The prompt was designed to request the model to classify online health information as either true or false and provide a brief explanation for the classification decision. This approach was based on preliminary research findings, which indicated that most users interact with LLMs by posing simple, straightforward queries, rather than engaging in more complex tasks like fine-tuning or model pretraining. The prompt design aims to mimic typical user queries, making it both user-friendly and effective. Table 3 presents the unified prompt structure used for evaluating the authenticity of online health information, which includes the content, verdict, explanation, and an example. This structure allows all LLMs to generate consistent classification labels and brief explanations, ensuring clarity and

consistency in the evaluation results. This uniform structure ensures that all models generate explanations in the same format, maintaining clarity and consistency in the evaluation results.

Table 2. Overview of API Calls and Result Processing Workflow.

Step	Description	Code Snippet/Function
1	Set Up Environment	os.environ to set proxies
2	Define API Functions	open35_response, open4_response, wenxin_response, spark35_response
3	Read Data from Excel	read_xlsx(filename)
4	Prepare Data Structure	Convert Excel rows to a list of dictionaries
5	Initialize Sample Data	Add keys for model labels and reasons
6	Create Prompt for AI Models	Define a prompt template for the models
7	Read Results from JSONL Files	read_jsonl(file_path)
8	Extract and Assign Results	Use re_patten(input_string) to parse results
9	Process Results from Each Model	Loop through samples and assign labels and reasons
10	Output Results	Print or save processed results

Table 3. Prompt Structure for Evaluating the Authenticity of Online Health Information.

Step	Prompt Component	Description
1	Information Content	The health-related information to be judged (e.g., "Drinking too much soda can cause leukemia").
2	Verdict	The judgment result (e.g., 'False').
3	Explanation	A brief explanation of why the judgment was made (e.g., "Sweeteners are legal and safe food additives. The causes of leukemia are still not fully understood, and there is no scientific basis for the claim that sweeteners cause childhood leukemia. The claim that drinking soda leads to leukemia is false.").
4	Example	The full prompt including the information content, verdict, and explanation. Example: "Please judge the following content based on your knowledge, assess whether it is a rumor, return {true or false}, and provide a brief explanation. For example:" {"Information content": 'Drinking too much soda can cause leukemia', 'Verdict': 'False', 'Explanation': 'Sweeteners are legal and safe food additives. The causes of leukemia are still not fully understood, and there is no scientific basis for the claim that sweeteners cause childhood leukemia. The claim that drinking soda leads to leukemia is false.'}"
5	Prompt Structure	The complete prompt structure for input into all models: <i>prompt = "Please judge the following content based on your knowledge, assess whether it is a rumor, return {true or false}, and provide a brief explanation. For example, see below:"</i> <i>{'Information content': 'Drinking too much soda can cause leukemia', 'Verdict': 'False', 'Explanation': 'Sweeteners are legal and safe food additives. The causes of leukemia are still not fully understood, and there is no scientific basis for the claim that sweeteners cause childhood leukemia. The claim that drinking soda leads to leukemia is false.'}</i> <i>Please assess the following content {'Information content':</i>

Performance Evaluation

Evaluation Metrics

The dataset contains 2,708 health-related information samples. It covers a variety of online health information types, which is sufficient for evaluating model performance.

To construct a reasonable evaluation framework, this study incorporates Precision (P), Recall (R), and F1 score as the core evaluation metrics. The formulas for calculating these metrics are as follows:

Precision (P): This metric measures the proportion of true positive samples among all samples classified as "positive" by the model. The formula is:

$$\text{Precision } P = \frac{TP}{TP+FP}$$

Where TP (True Positives) refers to the number of samples correctly classified as "positive" by the model, and FP (False Positives) refers to the number of "negative" samples incorrectly classified as "positive."

Recall (R): This metric evaluates the proportion of actual positive samples that the model successfully retrieves. The formula is:

$$\text{Recall } R = \frac{TP}{TP+FN}$$

Where FN (False Negatives) represents the number of positive samples that were incorrectly classified as "negative."

F1 Score: The F1 score is the harmonic mean of Precision and Recall, providing a balance between the two metrics. The formula is:

$$F1\text{ Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

In processing 28 samples for which the model outputs were labeled as "uncertain," "undetermined," "unknown," "to be verified," or "partly true partly false," a conservative approach was applied. These outputs were considered model errors to ensure comprehensive coverage of potential true information. During the calculation, these samples were counted as FP (if they were actually "negative") or excluded from affecting TP (if they were actually "positive"). In this study, although two of the 28 samples were labeled as "true" by experts, all 28 samples were treated as errors under the principle of "unknown equals wrong," primarily impacting the FP calculation.

Model Evaluation and Analysis

In this study, we used four methods to evaluate and analyze the performance of the four LLMs: text length analysis, semantic similarity analysis, credibility assessment, and Principal Component Analysis (PCA).

Text Length Analysis

Text length analysis is a method for statistically comparing the length of one or more texts to assess the characteristics of text output. By calculating the number of characters (average length, median, standard deviation, minimum, and maximum) in the texts generated by medical experts and the four LLMs, this analysis provides a quantitative basis for the length features of the generated text. It reveals the diversity and complexity of text generation across different models. Although text length does not directly reflect the accuracy of the health information generated by the models, its distribution reveals the consistency and stability of the models in generating health information and highlights the differences in text length characteristics across the models.

Semantic Similarity Analysis

Semantic textual similarity calculation is a key task in natural language processing, aiming to assess the semantic consistency between two texts and quantify the degree of their similarity. This study used semantic similarity analysis to evaluate the semantic consistency between the text generated by the models and the explanations provided by medical experts. This approach quantifies the differences in semantic understanding between the LLMs and expert explanations in the context of health information. This study used the jina-embeddings-v2-base-zh model developed by Jina AI. This model converts input text into fixed-length vectors (embeddings), ensuring that semantically similar texts are positioned closer together in the vector space. To quantify the similarity between model-generated text and expert explanations, feature data for both expert and model-generated texts were loaded from predefined data folders. The cosine similarity between expert features and model features is computed using the `F.cosine_similarity` function in PyTorch, reflecting the semantic consistency between the two. The similarity scores for each model were calculated, and the mean, 75th percentile, and 25th percentile values were printed. Finally, the similarity distributions were visualized using violin plots, with Matplotlib used to illustrate the similarity range and median for each model, providing a basis for assessing the degree of semantic matching between the model-generated content and expert explanations.

Credibility Assessment

To evaluate the ability of four LLMs to assess the credibility of online health information, this study selected 52 online health information samples from various sources. The data sources included PubMed, Tencent Fact Check, Chunyu Doctor, Microblog, Bilibili, and REDnote, covering a wide range of information types and levels of evidence, ranging from authoritative medical journals to online health information disseminated on social media. The selected samples encompassed both

true and false health information and included various types of evidence, such as anecdotal evidence, observational studies, and randomized controlled trials.

In modern medical practice, Evidence-Based Medicine (EBM) emphasizes the use of systematic evidence to assess the rationale behind clinical decision-making. To enhance the scientific rigor and precision of the evaluation, particularly in the assessment of online health information credibility, this study adopted the standard tool of EBM—the GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) framework as the evaluation tool. The GRADE framework is widely applied in the medical field and internationally recognized for its systematic approach to evaluating the quality of evidence and the strength of recommendations [30]. Moreover, the evidence quality is classified into four levels: very low, low, moderate, and high. High-quality evidence, often from well-conducted randomized controlled trials (RCTs), suggests that further research is unlikely to change the results; Moderate-quality evidence, typically from improved observational studies or RCTs with limitations, implies that more research could affect the findings; Low-quality evidence, usually from flawed observational studies, suggests that further studies are likely to alter the conclusions; Very low-quality evidence, often from observational studies with numerous downgrading factors, indicates significant uncertainty in the results [31].

In previous related studies, methods such as semantic similarity have often been used, which focus solely on the surface similarity between texts. However, these methods fail to fully consider multidimensional factors, such as the quality of the information, its source, and its clinical applicability. Therefore, the introduction of the GRADE framework for credibility assessment effectively addresses the limitations of semantic similarity-based evaluations, offering a more comprehensive and objective perspective. The specific steps are as follows: First, two independent researchers separately rated each sample according to the GRADE criteria, which included study design, risk of bias, consistency of results, directness and applicability, precision and confidence intervals, and publication bias. Their ratings were then compared, and any discrepancies were identified. In cases where disagreements occurred, the researchers engaged in discussions to resolve inconsistencies and reach a consensus on the final ratings. These finalized ratings were then used as the benchmark for evaluating the credibility of the online health information. Following this, the raw text of each online health information sample, along with relevant metadata (author information, source platform, and research type), was input into the four LLMs. When designing the prompts for the models, the GRADE rating criteria were explicitly provided, instructing the models

to assess the credibility of the information based on its source and author qualifications and to assign a credibility rating. Finally, inter-rater agreement between the ratings generated by the LLMs and those assigned by the research team was evaluated using the Kappa coefficient, and a heatmap was generated to visualize the consistency of the credibility assessments across models. The process provides a systematic approach for evaluating and comparing the performance of different models in assessing the credibility of online health information.

Principal Component Analysis

PCA is a multivariate technique that reduces the dimensionality of data by transforming it into a set of new, mutually independent principal components, thereby extracting and visualizing key information in the data [32]. This study evaluates the similarity between the four LLMs and expert explanations by extracting text embeddings, reducing dimensionality using PCA, and performing visualization analysis. First, the API of each model encoded the health-related information explanation texts and extracted their respective feature embeddings. The same process encoded and extracted features from the expert's explanation texts. Next, PCA was employed to reduce the high-dimensional text features to two principal components for visualization in a two-dimensional space. Specifically, the expert's text embeddings combined with those generated by each model, and PCA reduced the feature representations. Finally, this study visualized the relative distributions of the expert's and each model's text embeddings in the same space. Feature distribution plots were generated for each model to compare the embeddings with those of the expert, offering an intuitive comparison of the semantic similarity between the model-generated explanations and the expert's interpretations.

LDA Topic Modeling Analysis

To identify the latent topics within the 2,708 online health information samples, the Latent Dirichlet Allocation (LDA) model was applied. LDA is an unsupervised learning algorithm used to infer latent topics in a set of documents, thereby revealing the deep structure of the text [33]. In this study, the LDA analysis was conducted as a post-hoc analysis, applied to the expert explanation texts after data collection and initial processing to ensure the extracted topics' accuracy and representativeness, and to analyze the categories of misjudged instances.

The process was as follows: (1) Data Preprocessing: During data preprocessing, the dataset consisting of 2,708 online health information samples was cleaned by removing missing values and irrelevant data. Tokenization was then performed to break the text into words or phrases. To

construct the feature vocabulary, the TF-IDF method was used to encode words in each document and extract the most representative features. (2) Topic Number Determination: The optimal number of topics was determined by evaluating multiple LDA models using perplexity and coherence scores. Based on these metrics, seven topics were selected as the optimal balance between interpretability and model effectiveness. (3) Topic Modeling: After determining seven topics as optimal, the LDA model was applied to the dataset with this parameter setting. The model automatically identified the latent topics within the texts and produced the topic distribution for each document, along with the corresponding main topics. Analyzing the topic distributions uncovered latent semantic relationships between different online health information texts, providing valuable insights for subsequent accuracy assessment. The seven topics extracted by the LDA model covered key areas of online health information, each exhibiting distinct semantic characteristics, providing a solid foundation for further analysis and model validation.

Results

Identification Results and Analysis

The four models demonstrated a high level of detail and fluency in generating explanatory texts, as illustrated in Table 4. The first row of the table presents the results of the four LLMs in evaluating the claim "Diabetes is caused by eating too much sugar." All models correctly identified this claim as false, consistent with the expert's judgment. The second row shows the models' evaluation results for the claim "Rice washing water can be used for beauty." In this case, all models incorrectly classified the claim as true. Although the models made an incorrect judgment regarding the claim "Rice washing water can be used for beauty," the explanatory texts they generated still exhibited high fluency. Additionally, Table 5 provides a comparison of the identification results for the four LLMs, showing the number of correct, incorrect, and uncertain classifications, along with the evaluation metrics of accuracy, precision, recall, and F1 score. Among them, ChatGPT-4.0 achieved the best performance, with an accuracy of 87.27%, precision of 87.56%, and an F1 score of 87.46%, indicating strong reliability. ERNIE Bot and iFLYTEK Spark had accuracies of 87.25% and 87.00%, respectively, demonstrating stable performance. In contrast, ChatGPT-3.5 had an accuracy of 81.82%, which was lower than the other models, suggesting that there is still considerable room for improvement in its ability to accurately identify online health information.

Analysis of model performance shows that all four models performed well in distinguishing between true and false information. The true positive (TP) and true negative (TN) values for each model were relatively high, while the false positive (FP) and false negative (FN) values were relatively low, indicating that the models have a certain degree of reliability in filtering incorrect information and identifying accurate online health information. Figure 2 presents the confusion matrices for the four models, where the ratio of TN to FP reflects the models' ability to filter out incorrect information, while the ratio of TP to FN reveals their performance in identifying true health information.

Table 4. Identification Results of Pre-trained Language Models (Partial).

Information	ChatGPT-4.0	ChatGPT-3.5	iFLYTEK Spark	ERNIE Bot
Diabetes is caused by eating too much sugar	<i>"The main cause of diabetes is the inability of pancreatic β-cells to secrete insulin properly, or the body's tissue cells becoming insulin resistant, leading to high blood sugar. Although eating too much sugar may increase the risk of diabetes, it is not the direct cause."</i>	<i>"Diabetes is a chronic metabolic disease primarily caused by insufficient insulin secretion or impaired insulin response in cells. While excessive sugar intake may increase the risk of diabetes, the onset and progression of diabetes are influenced by multiple factors and cannot be simply attributed to eating too much sugar. Therefore, diabetes is not solely caused by eating too much sugar."</i>	<i>"Diabetes is related to various factors, including genetics, environment, and lifestyle. While excessive sugar intake may increase the risk of the disease, it should not be simply concluded that diabetes is caused by eating too much sugar. The pathogenesis of diabetes is complex, involving insulin secretion and action abnormalities. Therefore, diabetes should not be simply attributed to eating too much sugar."</i>	<i>"The cause of diabetes is not just eating too much sugar. While a high-sugar diet may increase the risk of diabetes, the real cause of diabetes usually involves multiple factors, including genetics, lifestyle, obesity, and lack of exercise. Additionally, diabetes is categorized into different types, such as Type 1 and Type 2 diabetes, which have different causes and treatments. Therefore, attributing diabetes solely to eating too much sugar is inaccurate."</i>
Rice washing water can improve skin	<i>"Rice washing water contains some vitamins and minerals, which have certain moisturizing and whitening effects on the skin. Therefore, washing or applying rice washing water to the face may have certain cosmetic benefits. However, since each person's skin type and condition vary, the effectiveness of rice washing water in beauty care may differ from person to person."</i>	<i>"Rice washing water contains vitamins B and C, as well as minerals, making it a natural skincare product. Regular use of rice washing water for cleansing can help alleviate skin dryness, brighten the complexion, and provide other cosmetic benefits."</i>	<i>"Rice washing water contains a large amount of vitamins B and minerals, which have cosmetic and skincare benefits. It can be used to wash the face or soak feet, contributing to skin whitening and smoothness. However, it is important to note that rice washing water should ideally be freshly made and not stored for too long, as bacteria may grow, which could be harmful to the skin."</i>	<i>"Rice washing water contains starch, vitamins, and minerals, which have certain moisturizing and whitening effects on the skin. However, the effectiveness may vary depending on individual differences, and rice washing water should not be solely relied upon for skincare."</i>

Table 5. Comparison of Identification Results of LLMs.

Model	Correct Count	Incorrect Count	Uncertain Count	Accuracy (%)	Precision (%)	Recall(%)	F1 Score (%)
ERNIE Bot	2361	339	8	87.25	87.47	87.20	87.34
ChatGPT-4.0	2371	327	10	87.27	87.56	87.37	87.46
ChatGTP-3.5	2221	481	6	81.82	82.13	81.75	81.94
iFLYTEK Spark	2353	351	4	87.00	87.02	86.89	86.95

Text Analysis

Text Length Analysis

Figure 3 presents the length similarity violin plot of explanation texts generated by the expert, ChatGPT-4.0, ChatGPT-3.5, iFLYTEK Spark, and ERNIE Bot. According to the plot, the text lengths of ChatGPT-4.0 and ERNIE Bot are closest to those of the expert's explanation texts, showing more consistent text generation characteristics. In contrast, ChatGPT-3.5 generates the shortest texts, with its text length distribution significantly lower than that of the other models, indicating greater variability. Table 6 summarizes and compares the statistical analysis of text lengths for the expert and the four LLMs, providing detailed information on mean, median, standard deviation, and coefficient of variation. According to the statistical analysis in Table 6, the mean text length for the expert is 172.83 (with a coefficient of variation of 43.68%), for ChatGPT-4.0 it is 128.70 (with a coefficient of variation of 29.16%), for ChatGPT-3.5 it is 81.73 (with a coefficient of variation of 23.43%), for iFLYTEK Spark it is 112.97 (with a coefficient of variation of 29.85%), and for ERNIE Bot it is 136.05 (with a coefficient of variation of 38.99%). The statistical results indicate that ERNIE Bot's text length is most similar to that of the expert. Although its coefficient of variation is marginally higher than those of ChatGPT-4.0 and iFLYTEK Spark, it still demonstrates notable stability and consistency. On the other hand, the text length generated by iFLYTEK Spark is relatively short, with less variability. In contrast, ChatGPT-4.0's text length and degree of dispersion are closer to those of the medical expert, highlighting its relative strength in generating explanation texts.

Table 6. Statistical analysis results of text length.

Model	Mean Length	Median Length	Std Dev Length	Minimum Length	25% Length	50% Length	75% Length	Maximum Length	Coefficient of Variation
Expert	172.83	170	75.5	18	111.75	170	237	355	43.68%
ChatGPT-4.0	128.7	127	37.52	19	102	127	153	326	29.16%
ChatGPT-3.5	81.73	80	19.15	26	68	80	94	275	23.43%
ERNIE Bot	136.05	127	53.05	38	105	127	153	797	38.99%
iFLYTEK Spark	112.97	108	33.72	13	90	108	132	390	29.85%

Semantic Similarity Analysis

Through semantic similarity analysis, we found that the texts generated by the four pre-trained language models exhibit relatively high semantic similarity to the expert's texts. ERNIE Bot shows an advantage, demonstrating its ability to understand and generate content with semantics closely aligned with the experts' explanations. Figure 4 presents the violin plot of semantic similarity, illustrating the distribution of similarity scores between the model-generated texts and the experts' explanations. Table 7 further presents the detailed statistical data of the semantic similarity between the model-generated texts and the experts' explanations. The table includes statistical metrics for each model, such as mean similarity, standard deviation, coefficient of variation, 25th

and 75th percentile similarities, maximum, minimum, and median similarity. The mean semantic similarity scores for ChatGPT-4.0, ChatGPT-3.5, ERNIE Bot, and iFLYTEK Spark are 0.7622, 0.7592, 0.7656, and 0.7588, respectively. The corresponding median scores are similarly high: 0.7836 for ChatGPT-4.0, 0.7825 for ChatGPT-3.5, 0.7869 for ERNIE Bot, and 0.7783 for iFLYTEK Spark. These findings suggest that, in most cases, the models are effective in capturing the semantic nuances of online health information and generating texts that are comparable to expert explanations. Notably, ERNIE Bot achieved the highest mean and median similarity scores, suggesting its generated texts are the closest to the expert's explanations.

Additionally, PCA was used to reduce the dimensionality of the features of the texts generated by each model and the expert's explanations and visualized them in two-dimensional space. Figure 5 illustrates that the projection points of the explanation texts generated by each model, when compared to the expert's explanation text, generally cluster in adjacent areas. Some of these points are very close to the expert's projection points, indicating a high degree of similarity in semantic content between the model-generated texts and the expert's explanations. However, there are a few projection points that are quite far from the expert's projection points, indicating that some models' generated texts still show significant differences from the expert's explanations.

In conclusion, the four models demonstrate relatively high semantic similarity to the expert's explanation texts and exhibit consistency, but there are still noticeable differences in understanding certain online health information.

Table 7. Statistical Analysis Results of Semantic Similarity.

Model	Mean Similarity	Standard Deviation	Coefficient of Variation	25th Percentile Similarity	75th Percentile Similarity	Avg. Similarity	Max. Similarity	Min. Similarity	Median Similarity	Mean Length
ChatGP-T4.0	0.7622	0.0652	0.0856	0.7116	0.8394	0.7622	0.9705	0.0501	0.7836	128.0642
ChatGP-3.5	0.7592	0.07	0.0922	0.7027	0.8405	0.7592	0.9686	0.009	0.7825	81.4597
ERNIE Bot	0.7656	0.0657	0.0859	0.7124	0.8419	0.7656	0.9498	0.0874	0.7869	135.825
iFLYTEK Spark	0.7588	0.0657	0.0866	0.7062	0.8349	0.7588	0.9658	0.0369	0.7783	112.484

Text Content Analysis

To further evaluate the performance of LLMs in understanding online health information, representative samples were selected for content analysis and summarization across different models. As shown in Table 8 and Table 9, the models displayed varying performances in semantic similarity analysis. Table 8 presents examples where the models achieved higher semantic similarity to expert explanations, while Table 9 highlights examples with lower semantic similarity. These

variations reveal differences in the models' ability to match semantics and the depth of understanding when processing online health information.

The data from the table show that, for some online health information, the texts generated by the models exhibit high semantic consistency with expert recommendations. For example, in response to the question "Are lumps and nodules always cancer?", the response generated by ChatGPT-3.5 shows a high degree of semantic similarity to the expert's explanation, with a similarity score of 0.9686. This indicates that, in certain contexts, pre-trained language models can accurately understand and restate complex health knowledge. However, when addressing the false information that "Vitamin C can boost immunity and prevent colds," ChatGPT-3.5's generated text significantly diverges from the expert's opinion, with a semantic similarity score of 0.3684. Specifically, the expert's explanation highlighted the limitations of Vitamin C in preventing colds and offered a comprehensive set of prevention strategies, including regular exercise, a regular sleep schedule, hand washing, and maintaining a clean environment. In contrast, ChatGPT-3.5 mentioned that Vitamin C may help enhance immunity but failed to adequately address its limitations and lacked detailed prevention advice. This explanation may result in users misunderstanding the information. This comparison underscores the limitations of pre-trained language models when processing complex online health information and highlights the importance of combining expert opinions.

iFLYTEK Spark demonstrated a semantic similarity of 0.9658 with the expert's explanation when addressing the misleading claim "Cosmetic allergy causes vitiligo." However, when responding to the question, "Does applying facial masks provide hydration for the skin?", the generated text showed lower similarity to the expert's opinion. The expert explained that the skin's ability to absorb moisture is limited and that the main function of a face mask is to promote the absorption of active ingredients and provide physical cooling, not to hydrate. The expert also emphasized the importance of moisturizing and warned that frequent wet applications could damage the skin barrier. In contrast, iFLYTEK Spark focused on the moisturizing effects of face masks, mentioning ingredients such as hyaluronic acid and glycerin, but it lacked a detailed scientific explanation and risk warnings, which could lead to misunderstandings. This highlights the model's limitations in addressing certain health topics, suggesting a need for further training to improve its explanatory capabilities.

As shown in Figure 6, the semantic similarity between the texts generated by the four LLMs and the expert's explanations, after misjudging online health information, is further illustrated using a violin plot, highlighting the differences in performance among these models. This figure clearly shows that the semantic similarity of the model-generated explanation texts after misjudging online health information is lower than that of the overall sample. This indicates that when the model fails to accurately process online health information, the semantic distance between the generated texts and the expert's explanations increases. Moreover, the figure indicates that ChatGPT-4.0 performs better than the other three models in maintaining the minimum semantic similarity. Even in cases of misjudgment, its generated explanation texts still maintain a relatively high degree of semantic similarity. This suggests that ChatGPT-4.0 retains relatively strong language comprehension and text generation abilities when handling complex or controversial online health information.

Table 8. Illustration of Examples with High Semantic Similarity between LLMs and Expert Explanations.

Model	Similarity Score	Text content	Expert or Institution Explanatory Text	Model-Generated Comment
ChatGPT-3.5	Max: 0.9686	Sugar increases the risk of cancer (false)	<i>"A lump or nodule is not necessarily cancer; it can be non-tumorous (such as an inflammatory mass) or a benign tumor. However, it should not be taken lightly, and further examination is recommended for a definitive diagnosis."</i>	<i>"A lump or nodule is not necessarily cancer; it could be caused by non-malignant conditions such as inflammation or cysts, but it could also be a malignant tumor. Therefore, the appearance of a lump or nodule requires prompt medical attention for further examination to determine the cause."</i>
iFLYTEK Spark	Max: 0.9658	Cosmetic allergies easily lead to vitiligo (false)	<i>"Vitiligo is a pigment disorder, and its exact cause remains unclear. It is currently believed that vitiligo has a genetic predisposition and results from a combination of internal and external factors. Cosmetic allergies are a type of allergic skin condition and are different from vitiligo. There is no definitive connection between cosmetic allergies and vitiligo, making the statement "cosmetic allergies easily lead to vitiligo" scientifically unfounded. However, cosmetics can potentially trigger vitiligo, but the likelihood is very low, and it can easily be confused with chemical leukoderma."</i>	<i>"Cosmetic allergies and vitiligo are two distinct conditions with different causes and mechanisms. Cosmetic allergies are primarily due to the skin's reaction to certain ingredients in cosmetic products, while vitiligo is a pigment loss disorder with an unclear cause, potentially related to genetic, immune, neurological, and environmental factors. Therefore, cosmetic allergies do not cause vitiligo."</i>
Ernie Bot	Max: 0.9498	The calcium supplementation effect of kidney beans is stronger than that of milk (false)	<i>"Kidney beans are rich in dietary calcium, and their calcium content per unit weight even exceeds that of milk. However, according to the Dietary Guidelines for Chinese Residents, the recommended intake of legumes for children is much lower than that of dairy products. Therefore, when considering calcium intake over a specific period, kidney beans provide less calcium than milk. The bioavailability of calcium in kidney beans is lower than that in"</i>	<i>"Although kidney beans do contain some calcium, it is not accurate to say that their calcium supplementation effect is stronger than that of milk. Milk is a high-quality calcium source with both high calcium content and absorption rate. In contrast, while kidney beans also contain calcium, their calcium content is relatively lower, and factors such as cooking methods may affect the absorption rate of calcium from kidney beans, which may not be as efficient as milk. Therefore, from a calcium"</i>

ChatGPT-4.0	Max: 0.9705	Myopia, hyperopia, and astigmatism are all refractive errors (true).	milk, as the calcium in milk and the effects of lactic acid fermentation make it more efficient for calcium absorption. So, in terms of calcium absorption, kidney beans are inferior to milk. That being said, relying solely on milk for calcium supplementation is insufficient; it is also necessary to obtain calcium from other sources such as algae, soy products, and shrimp.” “Myopia, hyperopia, and astigmatism are all types of refractive errors. Myopia occurs when the curvature of the cornea and lens is too steep, or the eyeball is too long, causing light to focus in front of the retina. Hyperopia happens when the curvature of the cornea and lens is too flat, or the eyeball is too short, causing light to focus behind the retina. Astigmatism occurs when the shape of the eyeball is irregular, causing light to refract in such a way that it does not form a single focal point on the retina.”	supplementation perspective, milk remains a more ideal choice.” “Myopia, hyperopia, and astigmatism are all abnormalities of the refractive state, known as refractive errors. Myopia occurs due to an excessively long eyeball axis or an overly strong refractive system, causing distant objects to focus in front of the retina. Hyperopia happens when the eyeball axis is too short or the refractive system is too weak, resulting in distant objects being focused behind the retina. Astigmatism occurs when the curvature of the cornea or lens is uneven, causing light entering the eye to fail to focus uniformly, leading to blurred vision. Therefore, myopia, hyperopia, and astigmatism all fall under the category of refractive errors.”
-------------	----------------	--	---	---

Table 9. Illustration of Examples with Low Semantic Similarity between LLMs and Expert Explanations.

Model	Similarity Score	Text content	Expert Text	Model-Generated Comment
ChatGPT-3.5	Min: 0.3684	Taking vitamin C can boost immunity and prevent colds (false)	“Supplementing with extra vitamin C cannot prevent colds; this is similar to taking dietary supplements. The only effective methods for preventing colds include regular exercise, adequate sleep, avoiding overexertion and exposure to cold, frequent handwashing, maintaining a clean and well-ventilated home environment, limiting activities in crowded places, avoiding contact with individuals with respiratory infections, and getting vaccinated against the flu.”	“Vitamin C can contribute to boosting the immune system and may help in preventing colds to some extent. However, it cannot completely prevent colds. It is still important to maintain a balanced diet and healthy lifestyle habits.”
iFLYTEK Spark	Min: 0.3980	Applying a face mask can hydrate the skin (false)	“The skin has limited water absorption; only a small amount of water is absorbed by the intact epidermis. When the external humidity is significantly higher than the moisture content within the epidermis, the skin can absorb some water, but after about ten minutes, the skin stops absorbing and becomes saturated. This is why there is no need to leave a face mask on for more than 10 minutes. After removing the mask, the small amount of water absorbed will evaporate. Face masks are not useless, but their benefits can be summarized into two categories: promoting the absorption of active ingredients and providing physical cooling. Furthermore, “hydration” is a	“Face masks are a common skincare product, and one of their main functions is to provide moisture and nutrients to the skin. The various ingredients found in face masks, such as hyaluronic acid, glycerin, etc., have excellent moisturizing effects and can effectively replenish the skin with moisture. Of course, the hydration effect may vary between different face mask products, and the skin's ability to absorb these ingredients also differs, so the results after using a face mask can vary from person to person.”

			<i>misleading concept. In fact, all methods that truly improve epidermal hydration ultimately work by reducing water loss through the skin, which should be called "moisturization." The main product for daily moisturizing is a moisturizer. Lastly, wet compresses can damage the skin's lipid layer, and frequent wet compresses are harmful to the skin barrier. Therefore, the use of face masks should be limited to 2-3 times per week and should not be excessive."</i>
Ernie Bot	Min: 0.2769	Taking vitamin C can boost the immune system and help prevent colds (false)	<i>"Supplementing with extra vitamin C cannot prevent colds, just like taking dietary supplements. The real ways to prevent colds are regular exercise, getting enough sleep, avoiding overexertion, staying warm, washing hands frequently, keeping the home environment clean and well-ventilated, minimizing visits to crowded places, avoiding contact with individuals with respiratory infections, and getting the flu vaccine to prevent the flu."</i> <i>"Vitamin C plays an important role in the human immune system and helps enhance the body's resistance to diseases. Some studies suggest that adequate supplementation of vitamin C may help prevent common illnesses like colds. However, for the general population, a balanced diet can typically meet the body's vitamin C requirements, and excessive intake of vitamin C may lead to side effects. Therefore, it is advisable to consult a doctor or nutritionist before taking vitamin C supplements."</i>
ChatGPT-4.0	Min: 0.4545	Shampoo opens the hair cuticle, while conditioner closes it (false)	<i>"The hair cuticle is made up of the outermost layer of dead cells, also known as the cuticle scales. They function like fish scales, protecting the inner structure of the hair. When you wash your hair, it absorbs water and swells, causing the cuticle scales to lift (the cuticle scales only lift, not open). The lifting of the cuticle scales does not mean that the "nutrients" in the shampoo can be absorbed by the hair, as these chemicals cannot penetrate the hair structure. After washing, the cuticle scales do not return to their original position, so they no longer provide protection to the hair. Conditioners contain moisturizing ingredients that, while they cannot close the lifted cuticle scales, can form a lubricating layer around them, helping to smooth and protect the hair."</i> <i>"To some extent, this statement is scientifically grounded. Shampoos contain surfactants, which help remove oils and dirt from the hair. This process can cause the outer layer of the hair, the cuticle scales, to lift. On the other hand, conditioners contain softening agents and oils, which help close the lifted cuticle scales, making the hair appear smoother and softer. Therefore, this claim is reasonable based on the fundamental principles of hair cleaning and care."</i>

Credibility Assessment

To quantify the consistency between the credibility assessment results of the four LLMs and the ratings provided by the research members, we calculated the Kappa coefficient and generated a heatmap. The Kappa coefficient measures the level of agreement between the models' credibility evaluations and the research members' ratings. Table 10 presents the Kappa coefficients for the four LLMs in relation to the research members' ratings. Specifically, the Kappa coefficient for ChatGPT-4.0 was 0.58, indicating a moderate level of agreement. ERNIE Bot achieved a Kappa coefficient of 0.52, reflecting a certain degree of consistency. ChatGPT-3.5 had a Kappa coefficient of 0.42, suggesting a lower level of agreement, while iFlytek Spark exhibited the lowest agreement, with a Kappa coefficient of 0.31. Additionally, the heatmap (Figure 7) further visualizes the performance of the four LLMs in credibility assessment. The results indicate that the ratings generated by ChatGPT-4.0 were the most aligned with those of the research members, followed by ERNIE Bot.

Table 10. Kappa Coefficients Between Models and Evidence Level.

Model	Kappa Value with Evidence Level
ChatGPT-4.0	0.5833
ChatGPT-3.5	0.4198
Ernie Bot	0.5201
iFlytek Spark	0.3131

Misjudgment Text Feature Analysis

To analyze the expert-explained texts of 2,708 online health information samples, we employed LDA topic modeling. Table 11, based on the results of LDA topic analysis, presents the sample size and proportion of each topic in the total dataset, along with the number and proportion of consistently misjudged online health information across the four LLMs within each topic. To be more precise, experiments determined that the optimal number of topics was seven. These topics include: Food and Maternal-Infant Nutrition Management, Chronic Disease and Women's Health Management, Children's Skin Health and Vision Issues, Nutritional Science and Food Controversies, Cancer Prevention and Vaccine Immunization, Weight Control and Metabolic Health, and Environmental Hazards Affecting Reproductive Health.

Further analysis of the online health information that was misjudged by all four models revealed a total of 66 instances (2.44%) that were incorrectly classified by all four LLMs. The distribution of these misjudged instances across the seven topics is shown in Table 11. While some topics have a higher overall proportion in the dataset, the misjudgment rates vary. The Food and Maternal-Infant Nutrition Management topic had the most misjudged instances (18), while Environmental Hazards Affecting Reproductive Health had the least (4). However, in terms of misjudgment ratios within each topic, the highest misjudgment rate was observed in Food and Maternal-Infant Nutrition Management, at 3.28%, followed by Nutritional Science and Food Controversies, with a misjudgment rate of 2.94%.

Table 11. Distribution and Proportions of Topics and Misjudged Information across Health Data.

Topic	Topic Name	Total Sample Size (n)	Proportion (%) Relative to 2,708 Total	Misjudged Sample Size (n)	Proportion (%) Relative to Topic's Total
Topic 2	Food and Maternal-Infant Nutrition Management	548	20.24%	18	3.28%
Topic 4	Chronic Disease and Women's Health Management	469	17.32%	7	1.49%
Topic 0	Children's Skin Health and Vision Issues	433	15.99%	12	2.77%
Topic 1	Nutritional Science and Food Controversies	374	13.81%	11	2.94%
Topic 6	Cancer Prevention and Vaccine Immunization	363	13.40%	8	2.21%
Topic 5	Weight Control and Metabolic Health	291	10.75%	6	2.07%
Topic 3	Environmental Hazards Affecting Reproductive Health	230	8.49%	4	1.74%

Discussion

Principal Findings

In the task of online health information identification, this study evaluated the performance of four LLMs: ChatGPT-4.0, ChatGPT-3.5, iFLYTEK Spark, and Ernie Bot. Semantic similarity analysis was used to assess the similarity between the texts generated by these models and expert explanations, aiming to better understand the differences between the models. While these models are capable of performing this task in most cases, they cannot fully replace the judgment of human experts [34]. Specifically, health misinformation is prevalent in certain topics, such as smoking products, opioids, and marijuana, with misinformation reaching 87% in some cases. Misinformation about vaccines, especially the human papillomavirus vaccine, and noncommunicable diseases like cancer also remain common, with rates of 43% and 40%, respectively [35].

ChatGPT can be divided into ChatGPT-3.5 and ChatGPT-4.0, which exhibit differences in their performance in online health information identification tasks. ChatGPT-4.0 demonstrates slight improvements in accuracy and precision, showcasing its enhanced ability to deliver detailed explanations for intricate health-related queries [36]. In contrast, while ChatGPT-3.5 possesses certain natural language processing capabilities and logical reasoning skills, the texts it generates are generally shorter. Research has shown that ChatGPT-4.0 outperforms ChatGPT-3.5 and Google Bard in terms of accuracy and self-correction ability when addressing myopia-related queries [37]. Therefore, this study suggests that ChatGPT-3.5 demonstrates comparatively weaker performance in managing specific details and ensuring accuracy than ChatGPT-4.0. Furthermore, ChatGPT-4.0 has improved logical coherence, enabling it to better integrate professional terminology and provide more compelling explanations. By comparison, Ernie Bot and iFLYTEK Spark exhibit notable strengths in Chinese language processing and cultural comprehension. These two models are likely to align more closely with the thinking patterns and cultural contexts of Chinese users when generating explanatory texts. In particular, Ernie Bot has the potential to provide relatively personalized health recommendations based on user needs and contextual information. In conclusion, the four LLMs exhibit distinct characteristics in the context of online health information identification. Each model provides varying degrees of explanatory text support, assisting users in evaluating the reliability of online health information.

Semantic similarity analysis revealed that, although the overall similarity was relatively high, there were notable variations between the models in specific cases. For instance, ChatGPT-3.5 recorded the lowest semantic similarity score of 0.009, markedly lower than other models, whereas Ernie Bot demonstrated a comparatively higher minimum similarity score of 0.0874, reflecting greater

stability. This discrepancy may be attributed to the diverse strategies employed by these models in processing online health information. Furthermore, an analysis of the coefficient of variation revealed the consistency of semantic similarity across different models. ChatGPT-4.0, ChatGPT-3.5, Ernie Bot, and iFLYTEK Spark exhibited coefficients of variation of 0.0856, 0.0922, 0.0859, and 0.0866, respectively. Lower coefficients of variation indicate greater stability and consistency in the semantic similarity between the model-generated and expert-generated texts. From this perspective, ChatGPT-4.0 and Ernie Bot displayed greater semantic stability. Analysis of the content revealed that model-generated explanatory texts differed from expert-generated texts in aspects like focus and comprehensiveness. Expert explanations tend to emphasize the application of medical knowledge and risk evaluation, providing in-depth analysis while frequently incorporating professional terminology and clinical context. In contrast, while model-generated texts are more accessible and convey information succinctly, they often lack depth and contextual support, which may lead to omissions of key information.

When addressing specific professional domains or emerging health issues, AI models may exhibit misunderstandings or provide insufficient information, which is related to issues of data availability, quality, and privacy [38]. For example, on COVID-19-related topics such as "COVID-19 is a self-limiting disease," judgments require up-to-date research and data [39]. However, models may struggle to make accurate assessments due to outdated training data. The KFF Health Misinformation Tracking Pilot Poll indicates that a significant proportion of the U.S. population is exposed to and holds inaccurate beliefs about health topics, such as COVID-19, gun violence, and reproductive health. This underscores the urgent need for more reliable information sources and greater trust in health media. In light of this, both the general public and large language models are susceptible to misinformation on complex and evolving health issues like COVID-19 [40]. Given the challenges both face in accessing accurate and up-to-date information, it is increasingly crucial to verify health-related content through trusted expert sources. Similarly, for emerging health concerns such as "Vitamin D supplementation can prevent uterine fibroids" or "Children wearing N95 masks may suffer irreversible harm," models are prone to misjudgments. These issues often lack sufficient research or scientific consensus, complicating accurate identification. Additionally, discussions surrounding nutritional products (e.g., "Fish oil can protect the heart") and emerging health technologies (e.g., "Laser surgery can cure myopia") may result in inaccuracies due to the

models' limited medical knowledge. Therefore, users are advised to exercise caution and cross-verify model-generated health information with other reliable sources.

When analyzing online health information in which LLMs and expert judgments disagree, certain recurring patterns and potential causes were identified. Firstly, in cases where experts have deemed information false, model misjudgments often stem from their adoption of popular yet scientifically unverified health concepts. This highlights the challenges models face in distinguishing verified medical knowledge from unsubstantiated popular beliefs during training, particularly when evaluating vague or exaggerated claims. Such ambiguity in these claims not only increases the difficulty of model interpretation but also reveals their limitations in recognizing accurate information. Secondly, when models misjudge expert-verified true information as false, it reflects the limitations they encounter in handling complex biomedical knowledge. Insufficient training data and limitations in knowledge bases may hinder the models' ability to accurately recognize information specific to professional domains. Additionally, their limited ability to understand and apply context often results in higher error rates for context-specific analysis. Collectively, these findings reveal the models' shortcomings in handling popular information, professional knowledge, and context-dependent data [41]. This provides direction for future improvements in model training methods and optimization strategies, particularly in enhancing their ability to understand scientific knowledge and contextual information.

In this study, all four LLMs misjudged 66 online health information samples, with none providing correct classifications. Although these misjudgments account for only 2.44% of the dataset, their potential implications in medical informatics applications should not be overlooked. Misjudged online health information may contribute to suboptimal health decisions. For instance, misjudgments in the Food and Maternal-Infant Nutrition Management domain may lead to inappropriate dietary choices during pregnancy, potentially affecting both maternal and infant health. Similarly, misjudgments in Nutritional Science and Food Controversies may cause individuals to disregard evidence-based dietary recommendations, increasing the risk of chronic diseases. In collaborative AI-driven environments, such errors may propagate across models, amplifying their impact. Misjudgments in critical areas such as Cancer Prevention and Vaccine Immunization could lead to confusion or vaccine hesitancy, thereby undermining public health initiatives. At the system level, misjudged information in Chronic Disease Management and Women's Health Management could influence healthcare policies, potentially resulting in suboptimal health interventions and

delays in patient care. On a broader scale, misjudged health data could disrupt international public health efforts, particularly during global health crises, by influencing policy decisions and international collaborations. Despite the demonstrated accuracy of these models, minimizing misjudgments remains a priority. Future research should prioritize model optimization, incorporating expert feedback and multi-tier validation to enhance reliability in health information processing.

Limitations and Future Work

While this study conducted a comprehensive analysis of LLMs in online health information identification, several limitations remain to be addressed. First, the quality and diversity of the training data may constrain the models' comprehension of domain-specific medical knowledge, particularly in handling recent advancements and intricate biomedical concepts. Although trained on large-scale corpora, these models are not specifically optimized for the health information domain, nor do they comprehensively cover the latest medical advancements and complex biomedical concepts. As a result, the models may exhibit errors when handling cutting-edge research or highly specialized medical content. Second, the study primarily drew on data made publicly available by rumor-debunking websites. However, these sources may not encompass the full spectrum of health-related information, particularly content originating from clinical settings or platforms with limited access. Since most of the data was gathered from Chinese websites, there is a possibility of regional or systemic bias in how information is categorized as "true" or "false." The reliance on a specific dataset raises concerns about the generalizability and external validity of the research findings. This issue is even more important when LLMs are expected to handle false health information in global or culturally diverse settings. As a result, the experimental dataset may not fully represent real-world clinical information, which could lead to differences in the model's performance in practical applications.

Additionally, this study primarily evaluates the effectiveness of LLMs in identifying true and false health information; however, it does not fully account for the diversity of user interactions, particularly the influence of health literacy on query formulation. Recent research has shown that patients and consumers differ significantly in their access to digital technology, ability to assess information accuracy, and trust in digital health tools, all of which affect their interactions with AI-based health information systems [42]. It is worth noting that lower health literacy not only hinders communication in healthcare settings but also impacts individuals' ability to search for and

comprehend health information in online environments [43]. In real-world applications, users with varying health literacy levels may interact with AI models differently. Individuals with higher health literacy are likely to pose queries with appropriate medical terminology and more precise expressions, while those with lower health literacy may struggle to accurately formulate questions or may use simplified language. This disparity in query formulation can affect how models interpret and respond to health information, thereby impacting the effectiveness and practicality of the models. For individuals with lower health literacy, model outputs may be too complex, lack necessary context, or fail to provide adequate clarification, leading to misunderstandings or an inability to access meaningful information. Therefore, future research should explore how to optimize LLMs to better accommodate users with different levels of health literacy. Specifically, models could assess the complexity of a user's query (such as the use of medical terms) and adjust the response accordingly. For users with lower health literacy, models could provide simplified language and clearer explanations, while for users with higher health literacy, more detailed and professional responses could be offered. Additionally, providing multiple versions of answers based on users' comprehension levels could enhance user experience and improve the overall effectiveness of health information retrieval through AI models.

Furthermore, this study primarily evaluated semantic similarity between model-generated texts and expert explanations. While this approach provides an objective measure of output quality, it may overlook critical nuances in online health information, such as contextual dependencies and the precision of medical terminology. Semantic similarity metrics often rely on general text comprehension, potentially resulting in a one-dimensional assessment. Future research should incorporate more comprehensive evaluation frameworks that account for the complexity of medical contexts. Finally, while multiple pre-trained language models were compared, the evaluation was limited to their default architectures and parameter settings without further exploration or optimization of the models' structures. In future research, we will explore fine-tuning models using domain-specific databases and developing customized architectures to enhance their capability in identifying and interpreting online health information.

Conclusion

Based on the results of this study, we found that the performance of the four LLMs in identifying online health information varied. Specifically, ChatGPT-4.0 demonstrated higher accuracy, with both accuracy and recall rates outperforming the other models, indicating strong abilities in

understanding and assessing online health information. The Kappa coefficient analysis also indicated that ChatGPT-4.0 demonstrated the highest level of consistency with the ratings provided by the research members, further confirming its superiority in credibility assessment. Notably, Ernie Bot showed a higher similarity to expert explanatory texts in both semantic similarity and text length analyses, indicating that Ernie Bot may hold greater potential in Chinese contexts, particularly in handling complex or specialized online health information. Additionally, in the analysis of misjudgments by the four models, 2.44% of the information did not align with expert judgment. The misjudgment rates were highest within the themes of "Food and Maternal-Infant Nutrition Management" and "Nutritional Science and Food Controversies." These findings suggest that future research should focus on these areas to enhance the models' performance in identifying online health information, improving the accuracy of their judgments. Although this study provides valuable insights into the evaluation of model efficacy, it has certain limitations, such as the potential impact of sample selection on the generalizability of the results and the lack of consideration for the models' differentiated handling of different types of online health information. Therefore, future research should expand the sample size to include a broader range of health topics and explore the applicability of the models in real-time data processing across diverse information sources. In summary, this study offers empirical support for the use of LLMs in online health information identification, highlighting their potential to improve public health literacy. As technology continues to advance, further exploration and refinement of these models' applications will be an important focus for future research, aiming for more accurate online health information identification and promoting more effective public health interventions.

Acknowledgments

The authors wish to thank the authors of the false health information debunking platforms for providing the data used in this study. The authors wish to extend their sincere gratitude to the developers of ChatGPT-3.5, ChatGPT-4.0, ERNIE Bot, and iFLYTEK Spark for making their APIs available for research. The authors wish to acknowledge the support of the School of Medical Information, Chongqing Medical University, and express their appreciation for the valuable feedback and suggestions from peers and reviewers, which greatly contributed to improving this paper.

Data Availability

The data sets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

Dongmei Tan was responsible for the conceptualization, data curation, formal analysis, investigation, methodology, visualization, and writing of the original draft. Yi Huang focused on formal analysis, writing-review & editing, and validation. Ming Liu managed software, project administration, resources, writing-review & editing, and validation. Ziyu Li managed project administration, resources, and supervision. Xiaoqian Wu played a key role in validation and resources. Cheng Huang handled funding acquisition, supervision, and writing-review & editing.

Abbreviations

LLMs: large language models

PCA: Principal Component Analysis

EBM: Evidence-Based Medicine

GRADE: Grading of Recommendations, Assessment, Development, and Evaluations

RCTs: Randomized Controlled Trials

References

References:

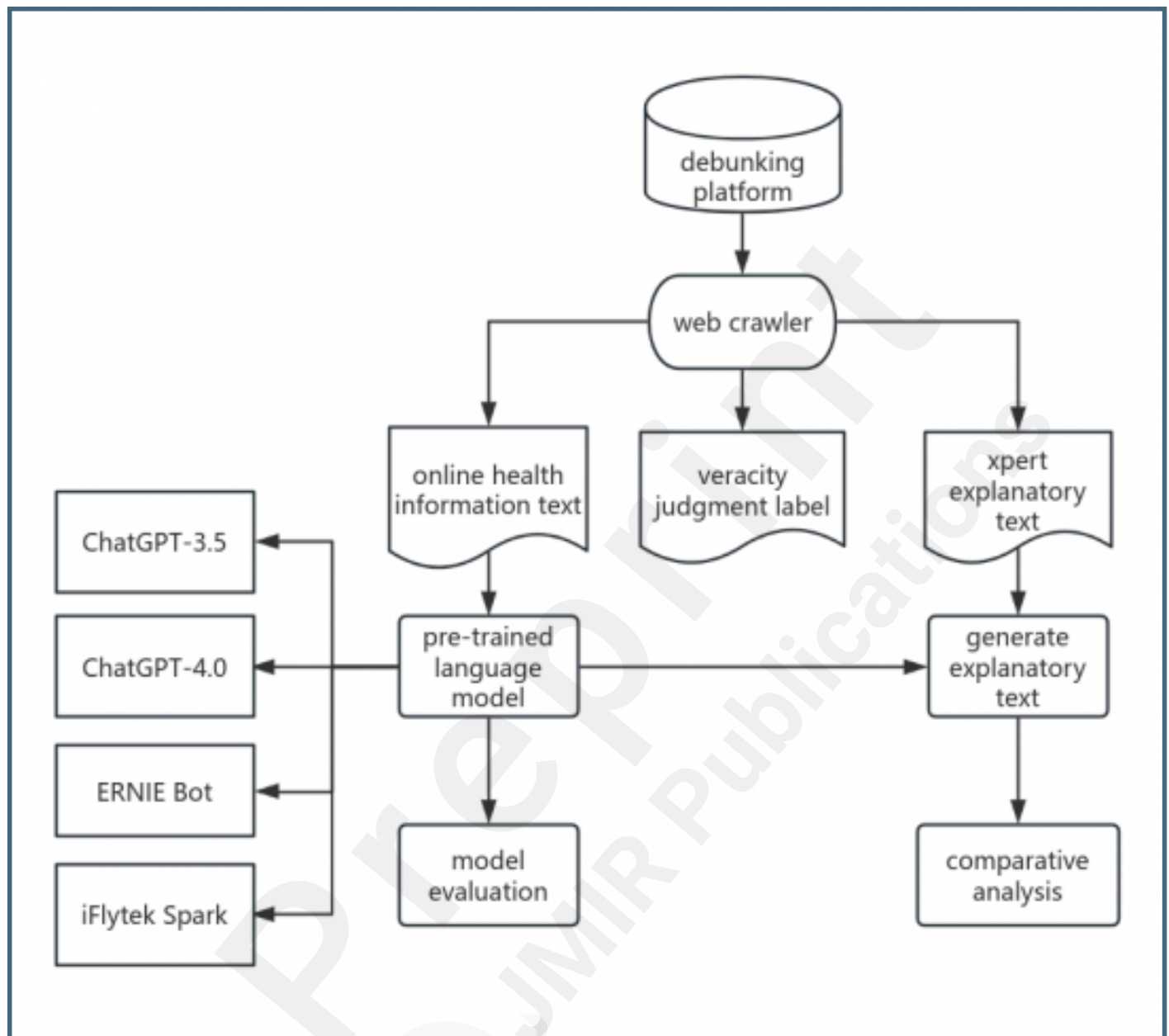
1. Zolkepli IA, Kamarulzaman Y. Social media adoption: The role of media needs and innovation characteristics. *Comput Human Behav.* 2015;43:189-209
2. Digital 2023 October Global Statshot Report — DataReportal – Global Digital Insights. 2023;2025(2025/3/2)
3. Diviani N, van den Putte B, Giani S, van Weert JC. Low Health Literacy and Evaluation of Online Health Information: A Systematic Review of the Literature. *J Med Internet Res.* 2015;17(5):e112
4. Sachdev MN, Ashfaq DR. Uncovering the Impact of social media on Disseminated Misinformation on Public Health: A Study of the Effects of False and Misleading Information on social media on Consumer Health Behaviors and Decision Making. *Journal Healthcare Treatment Development (JHTD).* 2023;3(04):43-55
5. Nan X, Wang Y, Thier K. Health misinformation. 2021:318-332
6. Stieglitz S, Dang-Xuan L. Emotions and information diffusion in social media —sentiment of microblogs and sharing behavior. *J Manag Inf Syst.* 2013;29(4):217-248
7. Lazer DM, Baum MA, Benkler Y et al. The science of fake news. *Science.* 2018;359(6380):1094-1096
8. Allcott H, Gentzkow M. Social Media and Fake News in the 2016 Election. *J Econ Perspect.* 2017;31(2):211-36
9. Chao W, Xianghui K. Application of Large-scale Pre-Training Language Model in Network Health Information Identification. *Journal of Library & Information Science in Agriculture.* 2023;35(6)
10. Cao YN, Ke Q. Why people show differences in their susceptibility to health misinformation? —research based on information processing configurations. *Journal of modern information.* 2023;43(1):40-54
11. Chai KEK, Anthony S, Coiera E, Magrabi F. Using statistical text classification to identify health information technology incidents. *J Am Med Inform Assoc.* 2013;20(5):980-985
12. Benigeri M, Pluye P. Shortcomings of health information on the Internet. *Health Promot Int.* 2003;18(4):381-386
13. Sillence E, Briggs P, Harris PR, Fishwick L. How do patients evaluate and make use of online health information? *Soc Sci Med.* 2007;64(9):1853-1862
14. Wu T, He S, Liu J et al. A brief overview of ChatGPT: The history, status quo and potential future development.

- IEEE/CAA Journal of Automatica Sinica*. 2023;10(5):1122-1136
15. Biswas SS. Role of chat gpt in public health. *Ann Biomed Eng*. 2023;51(5):868-869
 16. Kowsari K, Jafari Meimandi K, Heidarysafa M et al. Text Classification Algorithms: A Survey. *Information (Basel)*. 2019;10(4)
 17. Gasparetto A, Marcuzzo M, Zangari A, Albarelli A. A survey on text classification algorithms: From text to predictions. *Information (Basel)*. 2022;13(2):83
 18. Qiu X, Sun T, Xu Y et al. Pre-trained models for natural language processing: A survey. *Science China technological sciences*. 2020;63(10):1872-1897
 19. Gobira M, Nakayama LF, Moreira R et al. Performance of ChatGPT-4 in answering questions from the Brazilian national examination for medical degree revalidation. *Revista da Associação Médica Brasileira*. 2023;69:e20230848
 20. Favara G, Barchitta M, Maugeri A, Magnano San Lio R, Agodi A. The Research Interest in ChatGPT and Other Natural Language Processing Tools from a Public Health Perspective: A Bibliometric Analysis. *Informatics*. 2024;11(2)
 21. Bedi S, Liu Y, Orr-Ewing L et al. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA*. 2024;
 22. Chou WS, Oh A, Klein WM. Addressing health-related misinformation on social media. *JAMA*. 2018;320(23):2417-2418
 23. Vosoughi S, Roy D, Aral S. The spread of true and false news online. *Science*. 2018;359(6380):1146-1151
 24. Ahmed W, Vidal-Alaball J, Downing J, Seguí FL. COVID-19 and the 5G conspiracy theory: social network analysis of Twitter data. *J Med Internet Res*. 2020;22(5):e19458
 25. Chen E, Lerman K, Ferrara E. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health Surveill*. 2020;6(2):e19273
 26. Scardigno R, Mininni G. The rhetoric side of fake news: a new weapon for anti-politics? *World futures*. 2020;76(2):81-101
 27. Denecke K, May R, Rivera Romero O. Potential of Large Language Models in Health Care: Delphi Study. *J Med Internet Res*. 2024;26:e52399
 28. Arvanitis L, Sadeghi M, Brewster J. Despite OpenAI's promises, the company's new AI tool produces misinformation more frequently, and more persuasively, than its predecessor. *NewsGuard Technologies, Inc*. 2023;
 29. Wang B, Xie Q, Pei J et al. Pre-trained language models in biomedical domain: A systematic survey. *ACM Comput Surv*. 2023;56(3):1-52
 30. Guyatt GH, Oxman AD, Vist GE et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 2008;336(7650):924-926
 31. Atkins D, Best D, Briss PA et al. Grading quality of evidence and strength of recommendations. *BMJ*. 2004;328(7454):1490
 32. Abdi H, Williams LJ. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*. 2010;2(4):433-459
 33. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *J Mach Learn Res*. 2003;3(Jan):993-1022
 34. Javaid M, Haleem A, Singh RP. ChatGPT for healthcare services: An emerging stage for an innovative perspective. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*. 2023;3(1):100105
 35. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. *J Med Internet Res*. 2021;23(1):e17187
 36. 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. *J Med Internet Res*. 2023;25:e49771
 37. Lim ZW, Pushpanathan K, Yew SME et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *EBioMedicine*. 2023;95
 38. Aung YY, Wong DC, Ting DS. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. *Br Med Bull*. 2021;139(1):4-15
 39. Zhu C, Zhu J. The effect of self-limiting on the prevention and control of the diffuse COVID-19 epidemic with delayed and temporal-spatial heterogeneous. *BMC Infect Dis*. 2021;21:1-22
 40. Suran M, Bucher K. False health claims abound, but physicians are still the most trusted source for health information. *JAMA*. 2024;331(19):1612-1613
 41. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? *Patterns (N Y)*. 2024;5(3):100943
 42. Nutbeam D. Artificial intelligence and health literacy—proceed with caution. *Health Literacy and Communication Open*. 2023;1(1):2263355
 43. McCray AT. Promoting Health Literacy. *J Am Med Inform Assoc*. 2005;12(2):152-163

Supplementary Files

Figures

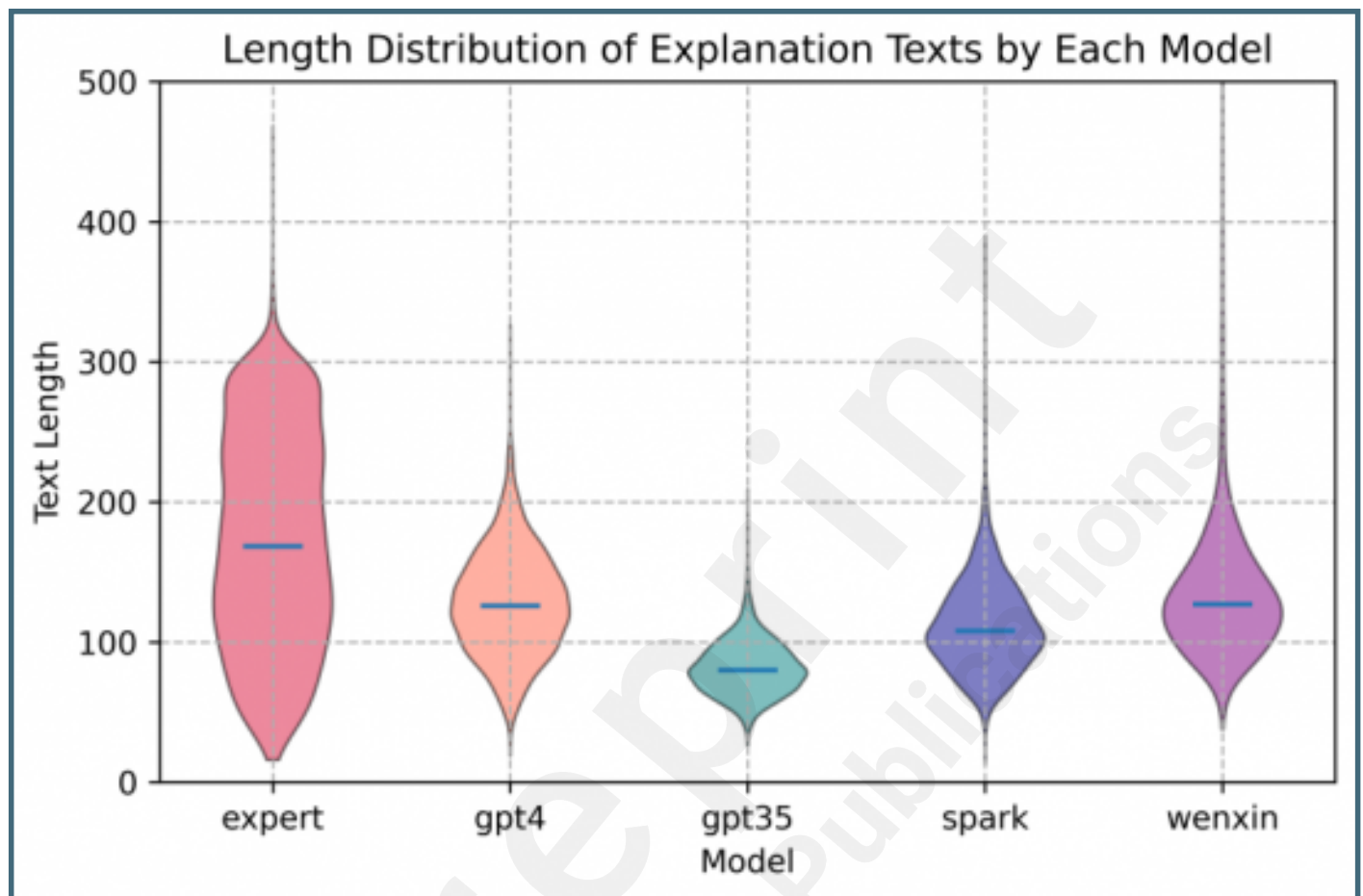
Framework Flowchart.



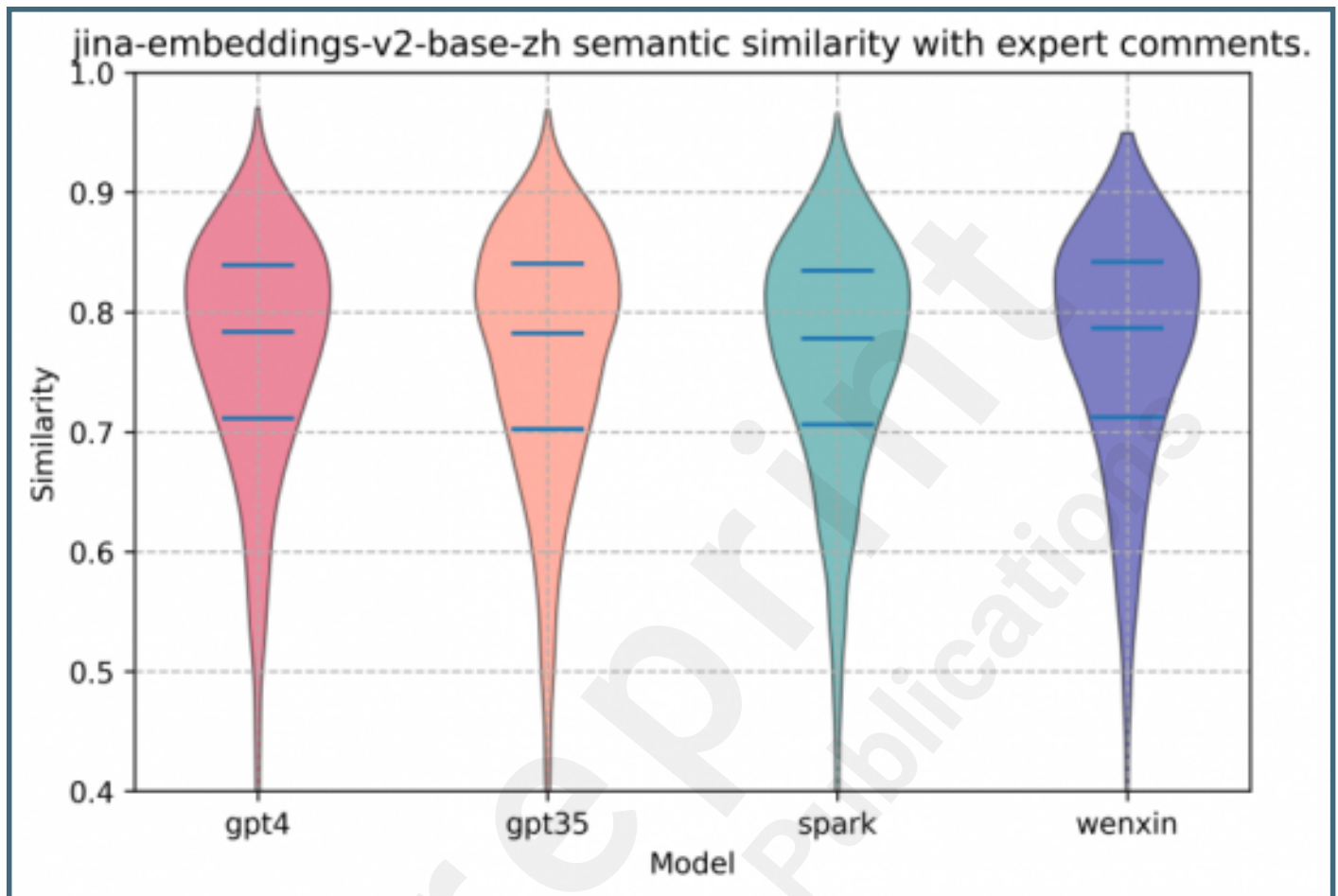
Confusion Matrix.



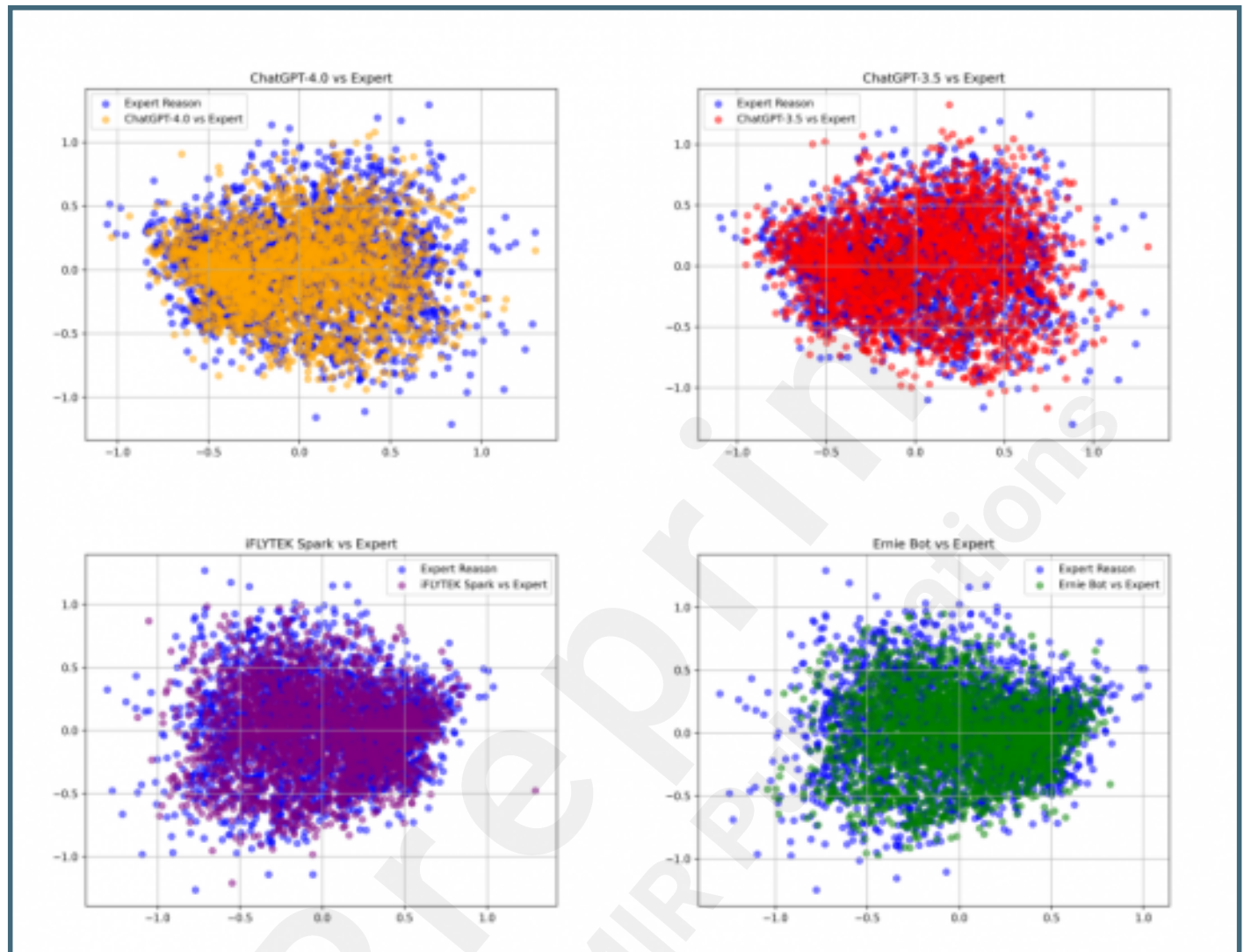
Text length distribution of explanations from experts, ChatGPT4.0, ChatGPT3.5, iFLYTEK Spark, and ERNIE Bot. ("gpt4" refers to ChatGPT-4.0, "gpt35" refers to ChatGPT-3.5, "sprk" refers to iFLYTEK Spark, and "wenxin" refers to ERNIE Bot.).



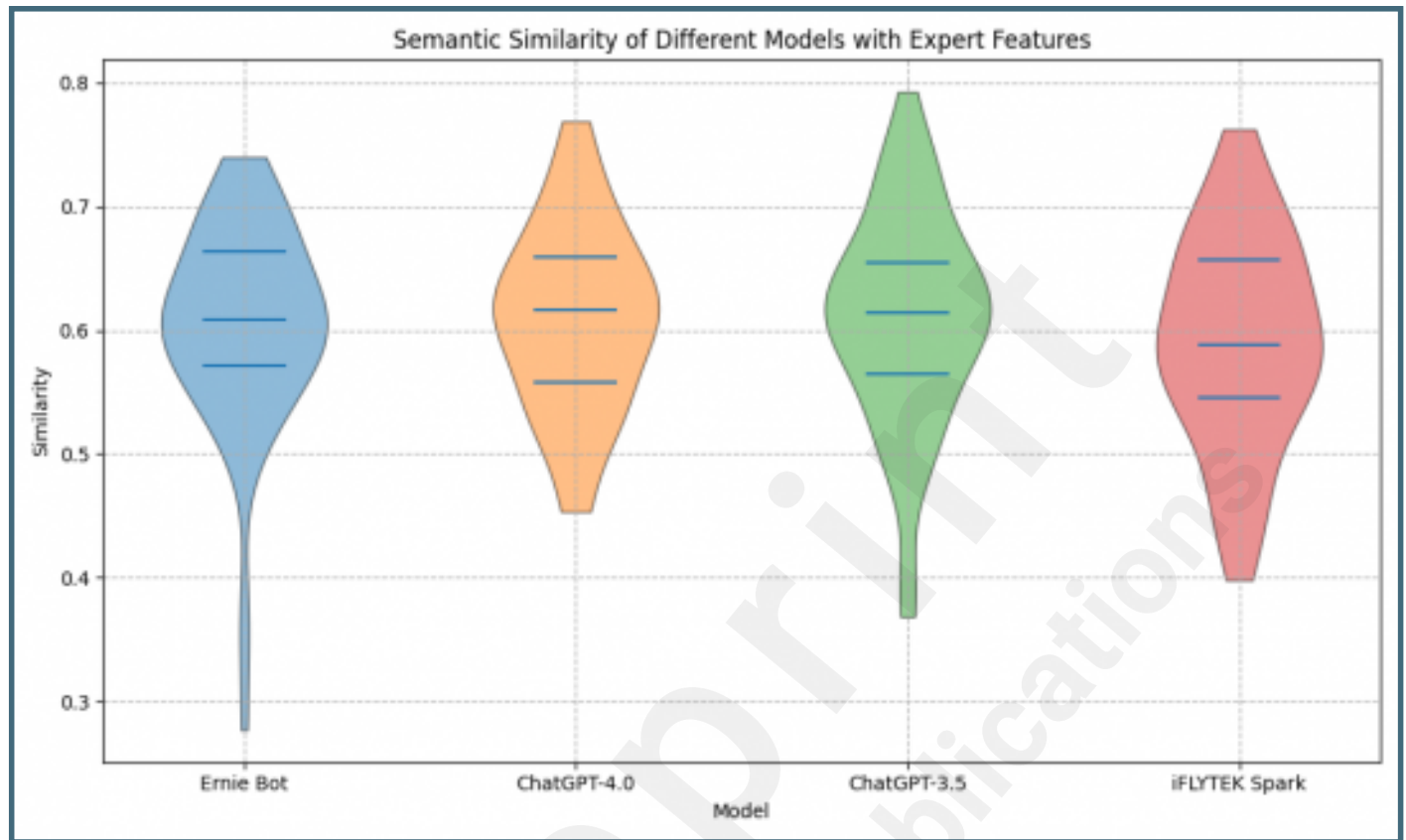
Distribution of Semantic Similarity between Texts Explained by ChatGPT-4.0, ChatGPT-3.5, iFLYTEK Spark, ERNIE Bot, and Experts. ("gpt4" refers to ChatGPT-4.0, "gpt35" refers to ChatGPT-3.5, "sprk" refers to iFLYTEK Spark, and "wenxin" refers to ERNIE Bot.).



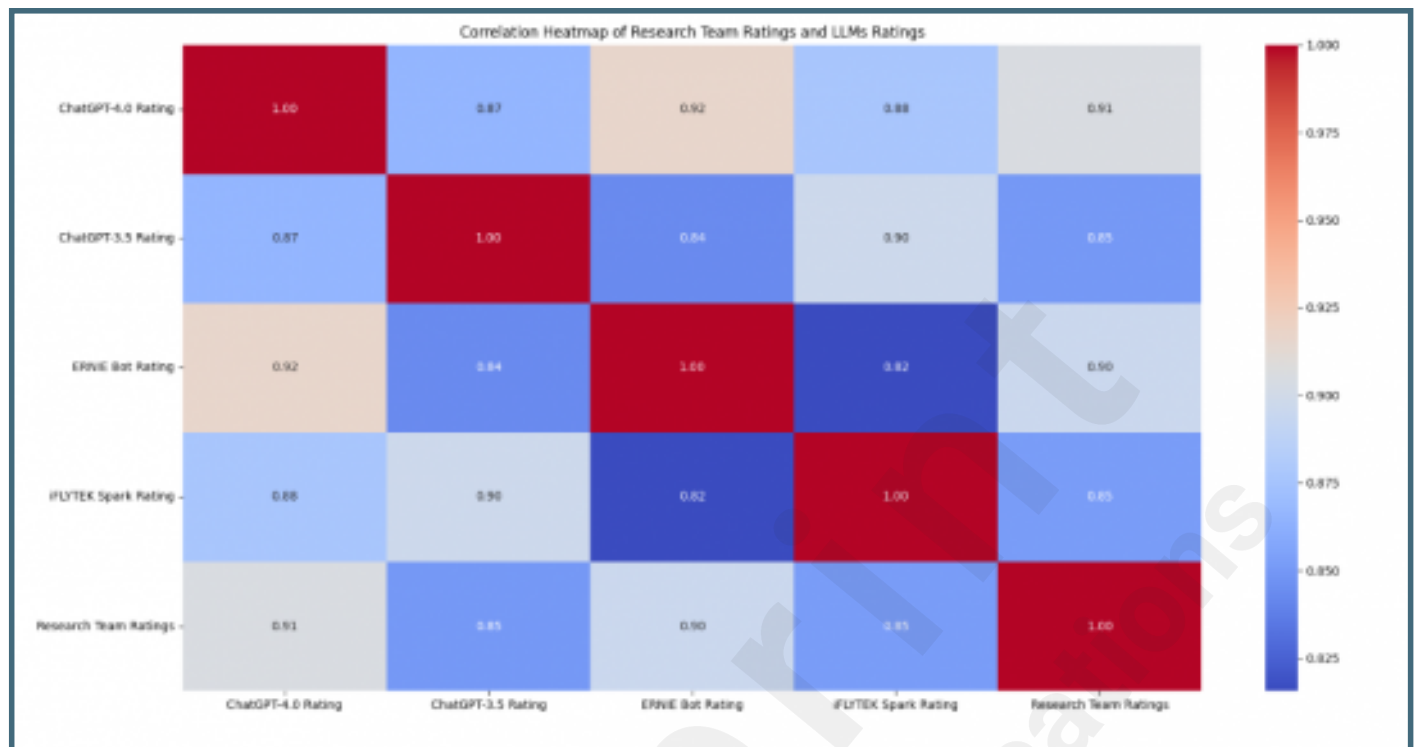
PCA Dimensionality Reduction Visualization of Texts Generated by Pre-trained Language Models and Expert Explanations.



Comparison of Semantic Similarity between Model Explanations and Expert Explanations for Misclassified Online Health Information.



Correlation Heatmap of Research Team Ratings and LLMs Ratings.



Multimedia Appendixes

Debunking platforms and their websites.

URL: <http://asset.jmir.pub/assets/67a683d21d0aa69912c597aea57c7ab4.docx>

