Original Paper

Detection of Medical Misinformation in Hemangioma Patient Education: Comparative Study of ChatGPT-4o and DeepSeek-R1 Large Language Models

Guoyong Wang[1], MD; Ye Zhang[1], MD, PHD; Weixin Wang[1], MD, PHD; Yingjie Zhu[1], MD; Wei Lu[1], MD; Chaonan Wang[1], MD, PHD; Hui Bi[2], MD, PHD; Xiaonan Yang[1], MD, PHD

[1]Department of Hemangioma and Vascular Malformation, Plastic Surgery Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[2]Department of Internal Medicine, Plastic Surgery Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

Corresponding Author:

Xiaonan Yang, MD, PHD
Department of Hemangioma and Vascular Malformation
Plastic Surgery Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College
33 Badachu Road, Shijingshan District
Beijing 100144
China
Phone: 86 18810601889
Fax: 86 01053968149
Email: yxnan@aliyun.com

Abstract

Background: This study examines the capability of large language models (LLMs) in detecting medical rumors, using hemangioma-related information as an example. It compares the performances of ChatGPT-4o and DeepSeek-R1.

Objective: This study aimed to evaluate and compare the accuracy, stability, and expert-rated reliability of 2 LLMs, ChatGPT-4o and DeepSeek-R1, in classifying medical information related to hemangiomas as either "rumors" or "accurate information."

Methods: We collected 82 publicly available texts from social media platforms, medical education websites, international guidelines, and journals. Of the 82 items, 47/82 (57%) were labeled as "rumors," and 35/82 (43%) were labeled as "accurate information." Three vascular anomaly specialists with extensive clinical experience independently annotated the texts in a double-blinded manner, and disagreements were resolved by arbitration to ensure labeling reliability. Subsequently, these texts were input into ChatGPT-4o and DeepSeek-R1, with each model generating 2 rounds of results under identical instructions. Output stability was assessed using bidirectional encoder representations from transformers–based semantic similarity scores. Classification accuracy, precision, recall, and F_1-score were calculated to evaluate the performance. Additionally, 2 medical experts independently rated the model outputs using a 5-point scale based on clinical guidelines. Statistical analyses included paired *t* tests, Wilcoxon signed-rank tests, and bootstrap resampling to compute confidence intervals.

Results: In terms of semantic stability, the similarity distributions for the 2 models largely overlapped, with no statistically significant difference observed (mean difference=–0.003, 95% CI –0.011 to 0.005; P=.30). Regarding classification performance, DeepSeek-R1 achieved higher accuracy (0.963) compared to ChatGPT-4o (0.910), and also performed better in terms of precision (0.978 vs 0.940), recall (0.957 vs 0.894), and F_1-score (0.967 vs 0.916). Expert evaluations revealed that DeepSeek-R1 significantly outperformed ChatGPT-4o on both "rumor" items (mean difference=0.431; P<.001; Cohen d_z=0.594) and "accurate information" items (mean difference=0.264; P=.045; Cohen d_z=0.352), with a particularly pronounced advantage in rumor detection.

Conclusions: DeepSeek-R1 demonstrated greater accuracy and rationale in detecting medical rumors compared with ChatGPT-4o. This study provides empirical support for the application of LLMs and recommends optimizing accuracy and incorporating real-time verification mechanisms to mitigate the harmful impact of misleading information on patient health.

JMIR AI2025;4:e76372; doi: 10.2196/76372

Keywords: medical rumors; large language models; hemangioma; semantic similarity; classification performance; artificial intelligence; AI

Introduction

In recent years, artificial intelligence (AI) has drawn considerable attention in detecting medical and health-related rumors [1,2]. Some studies have conducted systematic reviews on the application of AI technologies, such as text mining and machine learning, for the automatic identification of health misinformation [3]. Nonetheless, recognizing medical rumors remains a challenge due to the scarcity of high-quality specialized datasets and the extensive effort required by medical experts for annotation [4,5], making it difficult to train highly accurate rumor detection models. Moreover, as conversational AI assistants become increasingly integrated with and partially replace traditional search engine functionalities, more individuals are turning to chatbots for medical information [6,7]. However, current large language models (LLMs) lack robust verification mechanisms and often struggle to differentiate genuine from false medical information, frequently producing factually incorrect or imprecise answers—commonly known as "hallucinations" [6,8,9]. In the medical field, the risks posed by misinformation are particularly severe, as misleading content can undermine trust in health care systems, alter treatment decisions, and even lead patients to delay or reject scientifically validated therapies, opting instead for unsupported and potentially harmful treatments [10].

To ground our investigation concretely, we focused on vascular tumors and malformations—a field where rapidly evolving medical classifications often cause significant public confusion and misinformation [11]. The International Society for the Study of Vascular Anomalies classification is continuously updated, with the 2025 edition significantly revising its 2018 predecessor by introducing a new category, potentially unique vascular anomaly, incorporating multiple genetic syndromes into the classification framework, and implementing extensive terminology revisions. Such frequent updates complicate both clinical diagnosis and public comprehension [11,12]. A prominent example is the lesion previously termed "cavernous hemangioma," which has now been redefined as a subtype of "venous malformation." However, outdated terminology persists widely in patient forums and online sources, creating a gap between current medical standards and lay perceptions. This misinformation can lead directly to clinical risks, such as misdiagnosis, delayed treatments, or unnecessary interventions, highlighting the critical need to address inaccuracies and outdated information [13].

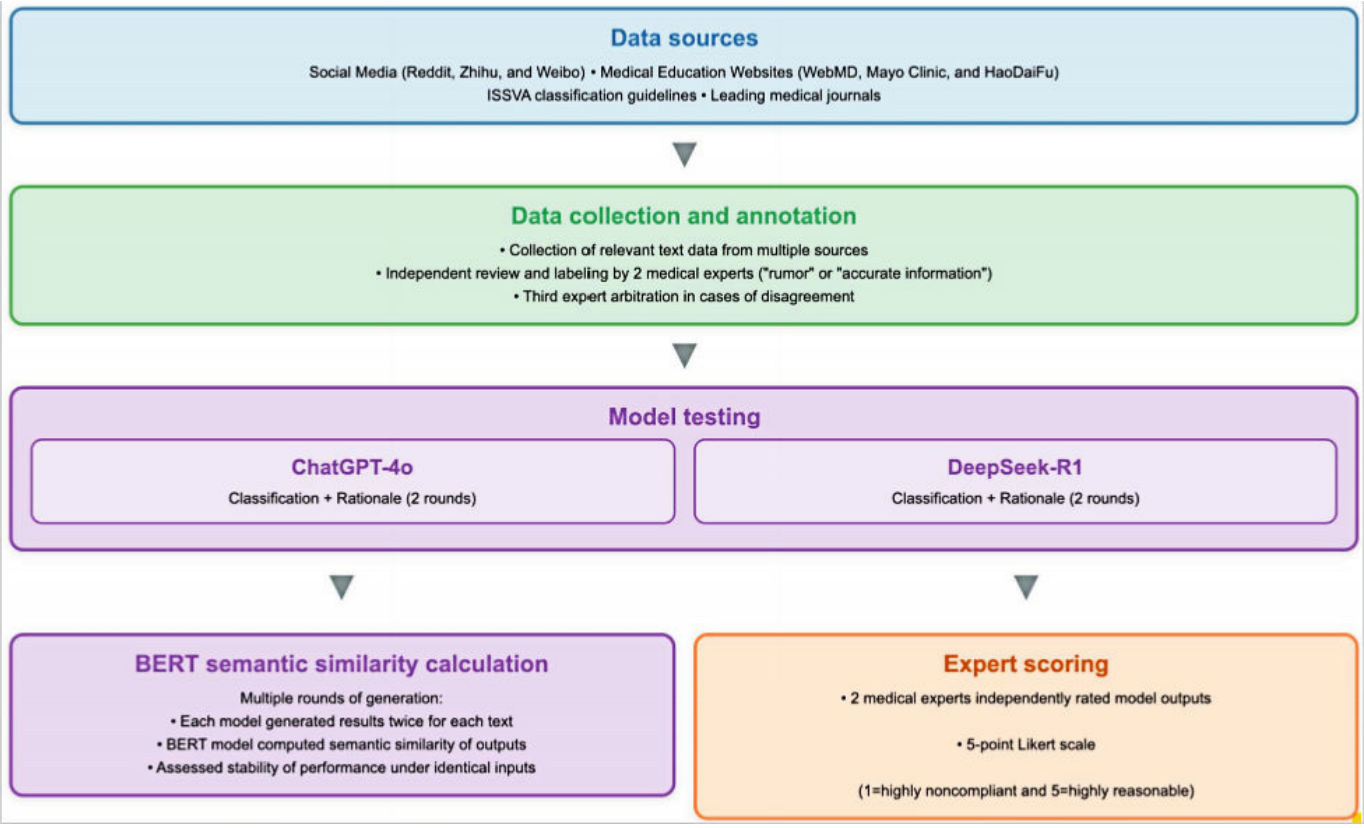
In this context, our study selected 2 widely adopted conversational AI models—OpenAI's ChatGPT-4o and the open-source DeepSeek-R1—as research subjects [14,15]. This combination not only represents the 2 primary development trajectories (closed-source versus open-source) of contemporary LLMs but also establishes a baseline task for subsequent benchmarking, allowing future studies to incorporate additional LLMs and facilitate longitudinal comparability. We conducted a classification evaluation of medical statements concerning hemangiomas and vascular malformations, focusing particularly on the models' ability to identify incorrect medical claims (rumors). By comparing the performance of these 2 models on relevant statements, our research aims to evaluate the current capabilities and limitations of AI models in verifying medical information and to provide insights for enhancing rumor-detection capabilities in medical AI systems in future work.

Methods

Study Design and Overview

Our study used publicly available texts from global social media platforms (eg, Reddit, Zhihu, and Weibo); medical education websites (eg, WebMD, Mayo Clinic, and HaoDF or HaoDaifu Online), the International Society for the Study of Vascular Anomalies classification resources, relevant guidelines, and medical journals (Multimedia Appendix 1). In total, 82 statements were collected, with 47 (57%) classified as "rumors" and 35 (43%) as "accurate information." These statements covered key educational aspects of patients with hemangiomas and vascular malformations, including (1) nomenclature and classification, (2) pathogenesis and natural history, (3) risk stratification and complications, (4) assessment and referral, (5) treatment and peritreatment issues, and (6) prognosis and follow-up. All texts collected were independently reviewed by medical experts and labeled as either "rumors" or "accurate information," based on guideline-supported factual accuracy. Figure 1 provides an overview of the study workflow.

Figure 1. Research methodology framework. BERT: bidirectional encoder representations from transformers; ISSVA: International Society for the Study of Vascular Anomalies.



Ethical Considerations

This study used only publicly available, nonidentifiable text data and did not involve clinical interventions, access to medical records, or collection of personal identifiers. In accordance with the Measures for the Ethical Review of Life Science and Medical Research Involving Humans, research using lawfully obtained public data or anonymized information may be exempt from ethics review (Article 32). Therefore, an ethics application was not required for this study [16]. Since the data were public and nonidentifiable, informed consent was not required. No compensation was provided to any individuals in relation to this study.

Data Collection and Annotation

Two medical experts specializing in vascular anomalies (with 5 and 10 y of clinical experience, respectively) independently reviewed and labeled each statement as either "rumor" or "accurate information." To minimize bias, all items were anonymized by removing source identifiers and engagement metrics prior to labeling, and annotators remained double-blinded to each other's decisions. In cases of disagreement, arbitration was conducted by a third medical expert with 15 years of clinical experience, resulting in a unified set of labels and ensuring labeling reliability. Potential biases were mitigated through independent dual review, third-party arbitration, and prespecified labeling guidelines.

Model Testing

After labeling, the texts were input into 2 LLMs—ChatGPT-4o and DeepSeek-R1—for testing. The process is presented in Textbox 1.

Textbox 1. Model testing process.

- Prompts and outputs: to minimize bias introduced by variations in prompting and to highlight baseline comparability, both models received the identical concise instruction: "evaluate the following statement for accuracy and reliability in the context of hemangioma and vascular malformation treatment." Each model classified the texts as either "rumor" or "accurate information," accompanied by a brief rationale (Multimedia Appendix 2).
- Multiple rounds of generation: to reduce the effects of random output, each model generated results twice for each text. A bidirectional encoder representations from transformers model was then used to compute the semantic similarity of these 2 outputs to assess the stability of the model's performance under identical inputs.

Expert Scoring

In addition to classification results, 2 medical experts independently assessed the compliance of each model's output with clinical guidelines. Evaluations were performed using a 5-point Likert scale (1= highly noncompliant, 5=highly reasonable). The medical experts remained blinded to both the model identities (ChatGPT-4o vs DeepSeek-R1) and each other's scores. Detailed scoring criteria are provided in Multimedia Appendix 3.

Statistical Analysis

Semantic Similarity and Stability

Semantic stability was assessed by calculating bidirectional encoder representations from transformers (BERT)–based similarity scores between 2 independently generated outputs for each statement (see Multimedia Appendix 4 for detailed code). Descriptive statistics, including means, SDs, medians, and IQRs, were reported. Differences between models were compared using paired Wilcoxon signed-rank tests (due to partially nonnormal distributions). Additionally, 95% bias-corrected and accelerated CIs for mean differences were computed via 10,000 bootstrap resamples to ensure robust interval estimation.

Classification Performance

Classification accuracy, precision, recall, and F_1-scores were calculated based on standard definitions, with error distributions visualized using confusion matrices. This approach allows comprehensive evaluation of global and class-specific performance and is particularly suitable for scenarios involving class imbalance.

Expert Ratings

Two clinical experts independently provided ratings on a 5-point Likert scale for each of the 82 statements (47 rumors and 35 accurate statements) in 2 separate rounds. The mean rating for each item was computed as the final score. For each model, descriptive statistics such as mean (SD) and 95% CIs were calculated, treating each statement as an independent unit. Between-model comparisons were performed using paired 2-tailed t tests (assuming normality of differences) supplemented by Wilcoxon signed-rank tests as a robust alternative, with Cohen d_z effect sizes reported. Within-model comparisons between "rumors" and "accurate information" were conducted using Welch t test to account for unequal sample sizes and potential variance heterogeneity. Reviewer agreement and reliability were assessed using Cronbach α and interclass correlation coefficients (ICCs), ICC(2,1)/ICC(2,k). All tests were 2-tailed, with statistical significance defined as $P<.05$.

Results

Overview

This study systematically compared the performance of ChatGPT-4o and DeepSeek-R1 in classifying statements related to hemangiomas and vascular malformations across three dimensions: (1) the stability of 2 independent outputs, assessed using BERT-based semantic similarity metrics; (2) classification performance, evaluated by accuracy, precision, recall, and F_1-score; and (3) clinical appropriateness of model outputs as rated by experts on a 5-point scale. For expert ratings, statistical inference was conducted using a paired design with Wilcoxon signed-rank tests, effect sizes (r), and 95% bias-corrected and accelerated CIs.

Semantic Similarity Analysis

To evaluate the semantic similarity between the model-generated responses, we used a BERT-based scoring approach (detailed in Multimedia Appendix 5). Multimedia Appendix 6 shows the distribution of the scores for ChatGPT-4o and DeepSeek-R1. Overall, the distributions for both models exhibited substantial overlap, with ChatGPT-4o displaying a slightly narrower distribution, while DeepSeek-R1 showed a marginally wider range. During paired comparisons, 1 pair with identical observations was excluded, resulting in 81 (99%) paired samples for analysis. The Wilcoxon signed-rank test indicated no significant difference in stability between the 2 models (W=1440.5; z=–1.036; P=.30), with a mean difference of only –0.003 (95% bootstrap CI –0.011 to 0.005, r=–0.115) as shown in Table 1. These findings suggest comparable semantic similarity and stability performance between the 2 models.

Table 1. Stability comparison between ChatGPT-4o and DeepSeek-R1 based on bidirectional encoder representations from transformers semantic similarity scores.[a]

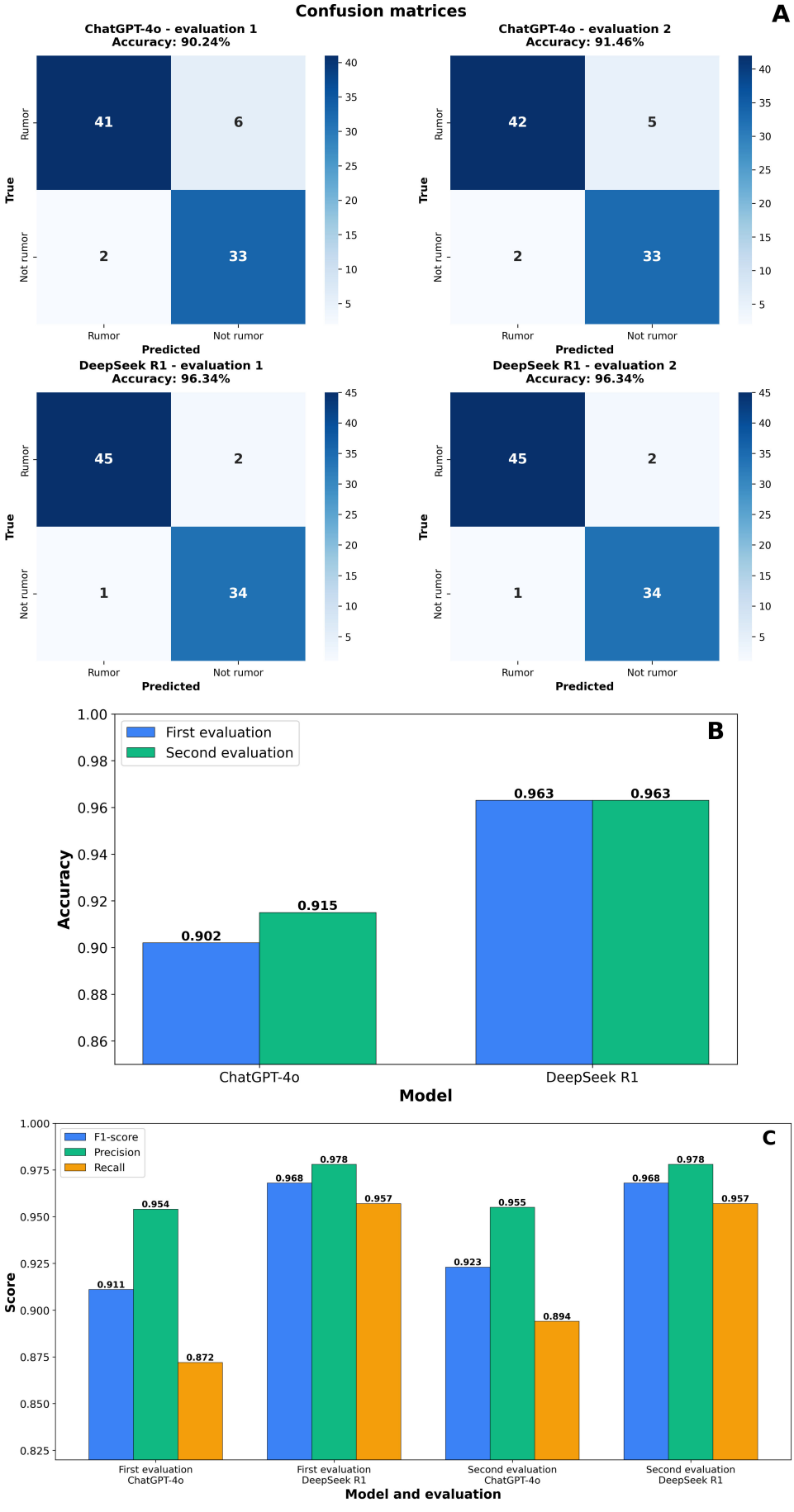
Model and comparison	Mean (SD)	Median (IQR)	Range
ChatGPT-4o (N=82)	0.9000 (0.0250)	0.9060 (0.8870-0.9180)	0.8250-0.9400
DeepSeek-R1 (N=82)	0.8970 (0.0320)	0.9010 (0.8850-0.9140)	0.7800-1.0000

[a]Paired difference (DeepSeek-R1 – ChatGPT-4o; n=81; of the original 82 pairs, 1 pair with identical values [tie] was excluded automatically during the Wilcoxon test, resulting in an effective sample size of 81): mean difference=–0.0030; 95% bias-corrected and accelerated CI –0.0110 to 0.005; Wilcoxon W=1440.5000; z=–1.0360; P=.30; r=–0.1150.

Classification Performance Evaluation

Classification performance for hemangioma and vascular malformation statements was evaluated by examining confusion matrices (Figure 2A) and key performance metrics. Confusion matrix analyses indicated no substantial differences in misclassification distribution between the 2 models, with overall good stability. In terms of the overall classification accuracy (Figure 2B), DeepSeek-R1 achieved 0.963, which was notably higher than ChatGPT-4o, which reached approximately 0.910. Additionally, DeepSeek-R1 surpassed ChatGPT-4o in terms of other metrics, including precision, recall, and F_1-score. Specifically, DeepSeek-R1 demonstrated a precision of approximately 0.978, recall of 0.957, and an F_1-score of 0.967, each marginally higher than the corresponding values for ChatGPT-4o (Figure 2C). These results highlight the superior classification accuracy of DeepSeek-R1.

Figure 2. (A) Confusion matrices for vascular lesion classification by ChatGPT-4o and DeepSeek-R1; (B) overall classification accuracy of ChatGPT-4o and DeepSeek-R1; (C) precision, recall, and F_1-scores of ChatGPT-4o and DeepSeek-R1.



Expert Rating Analysis

In qualitative assessments, both models demonstrated strong performance regarding the clinical appropriateness of their outputs, with subtle yet meaningful differences observed. Expert ratings (Multimedia Appendices 7 and 8) indicated that for statements classified as "rumors" (47/82, 57%), DeepSeek-R1 scored significantly higher with a mean (SD) of 4.39 (0.59) and 95% CI 4.21-4.56 compared to ChatGPT-4o with a mean of 3.96 (SD 0.81) and 95% CI of 3.72-4.20; the mean difference was 0.431 (95% CI 0.218-0.644); paired t_{46}=4.071; P<.001; Wilcoxon P<.001; and effect size Cohen d_z=0.594.

For statements labeled as "accurate information" (35/82, 43%), DeepSeek-R1 with a mean of 4.44 (SD 0.37) and 95% CI of 4.32-4.57 also significantly outperformed ChatGPT-4o with a mean of 4.18 (SD 0.69) and 95% CI of 3.94-4.41; the mean difference was 0.264 (95% CI 0.007-0.522); paired t_{34}=2.085; P=.045; Wilcoxon P=.046; and Cohen d_z=0.352.

These findings demonstrate significant superiority of DeepSeek-R1 over ChatGPT-4o in evaluating both "rumors" and "accurate information," with a particularly pronounced advantage in detecting "rumors."

DeepSeek-R1 performed slightly better than ChatGPT-4o across multiple evaluation dimensions, exhibiting higher output stability and classification accuracy. This finding suggests that DeepSeek-R1 holds greater potential for medical information classification tasks.

Discussion

This study compared ChatGPT-4o and DeepSeek-R1 in the task of identifying medical rumors, with hemangioma-related misinformation serving as the focal point [17,18]. Overall, both models demonstrated robust language comprehension capabilities but differed markedly in their approaches to recognizing inaccurate statements about hemangiomas. DeepSeek-R1 excelled at pinpointing erroneous claims and clearly categorizing them as rumors, showing its strength in explicit rumor detection and confident classification. In contrast, ChatGPT-4o demonstrated superior semantic similarity and exhibited more consistent stability in understanding nuanced languages, yet tended to approach rumor identification cautiously, often resorting to ambiguous wording rather than decisively refuting false information. Although these observed differences may stem from variations in training data, model architecture, and fine-tuning strategies, existing evidence from other studies suggests that specialized fine-tuning with medical information could further enhance the capability of LLMs in accurately and effectively detecting medical misinformation [19].

In our task, overly cautious responses—specifically, the failure to decisively refute rumors (false negatives)—may perpetuate harmful misconceptions, causing caregivers to delay specialist referrals or discontinue evidence-based treatments in favor of unproven remedies. Conversely, overconfidence—erroneously labeling accurate guidance as rumors (false positives)—may lead to unnecessary anxiety, undermine trust in clinicians, or impede appropriate interventions. In hemangioma treatment, such misclassification could negatively impact decisions regarding timely assessment (eg, ulceration and airway involvement), follow-up intervals, or continuation of guideline-adherent therapies. These risks support the use of conservative safety thresholds, verifiable citations, and escalation of human oversight when model confidence is low. One illustrative example is the claim that "sun exposure exacerbates hemangiomas," which lacks scientific support [20]. Authoritative sources indicate that sun exposure does not directly enlarge or worsen hemangiomas. While moderate sun protection can help safeguard the skin, it does not specifically address pathological changes in hemangiomas [21,22]. In this study, DeepSeek-R1 correctly identified this assertion as a rumor and provided a concise explanation consistent with medical consensus. ChatGPT-4o, in contrast, did not unequivocally refute the claim, instead offering a somewhat reserved answer that did not effectively dispel the misconception. Although both models possess extensive medical knowledge, DeepSeek-R1 displayed a stronger rumor-debunking ability when confronted with evidently incorrect statements, whereas the cautious approach of ChatGPT-40 diluted its capacity to correct misinformation.

As more users turn to AI assistants for medical information, traditional search engines are gradually being supplemented or even replaced by these systems [23,24]. Unlike search engines that merely provide links, AI chatbots often deliver comprehensive, single-point answers whose perceived authority may lead users to over-rely on them instead of consulting additional information sources [25,26]. Consequently, the adverse impact of inaccurate or ambiguous medical information disseminated by AI could be amplified, posing a considerable risk of misleading patients in their health care decisions. Therefore, ensuring higher accuracy in identifying medical rumors is both urgent and critical [27].

Recent research has proposed various methods for leveraging AI to detect medical rumors. For instance, studies comparing GPT-4 with other models trained specifically on health information have shown that specialized models tend to be more accurate in identifying and correcting misinformation [28,29]. These findings underscore that although LLMs have tremendous potential for conveying medical knowledge, they still exhibit shortcomings in fact-checking and real-time verification [30]. Incorporating real-time retrieval mechanisms and referencing authoritative data in responses represents a key direction for improving the accuracy of AI-generated medical information [28]. Notably, conclusions regarding model superiority depend heavily on the task design, dataset scope, and evaluation criteria. These factors help explain the inconsistencies observed in the existing literature and highlight the novelty of our research, which specifically addresses misinformation related to hemangiomas. The methodological workflow applied in this study—consisting of data annotation, multiround generation, BERT similarity assessment, and expert evaluation—not only validates the relative advantages of DeepSeekR1 in

our task but also underscores the insufficiency of any single metric for comprehensively assessing model performance. Multidimensional evaluations more effectively reveal nuanced differences between models in stability, accuracy, and clinical appropriateness, thereby offering valuable lessons and standardized protocols for the deployment and further study of large medical language models.

This study has several limitations. First, our data primarily address hemangiomas and vascular malformations, and the limited number and types of examples may not comprehensively encompass all medical rumors. Second, the labeling of rumors relies on expert judgment, introducing an element of subjectivity, and disagreements may arise when experts evaluate borderline cases. Additionally, discrepancies in the 2 AI models' training data and knowledge cutoff dates could affect their ability to capture the latest medical information. Finally, we did not evaluate aspects such as explanatory depth, response speed, and user-friendliness. For instance, we did not conduct a formal qualitative or user-centered analysis of explanation quality, which remains an important area for future investigation. For clinical decision support, patient-oriented education, or public health surveillance, LLM-generated outputs should be embedded within regulated workflows that include (1) retrieval-augmented validation from curated vascular anomaly sources, (2) human-in-the-loop review of high-risk recommendations, (3) audit trails and disclaimers clearly delineating accountability, (4) transparent rationales with explicit references to guidelines and clearly marked uncertainties, and (5) postdeployment monitoring for data drift and fairness. These safeguards are prerequisites for mitigating liabilities and improving interpretability and usability in practical applications.

In conclusion, this research highlights the performance differences between the 2 LLMs in detecting hemangioma-related medical rumors, stressing the urgency of maintaining accurate medical information as AI gradually supplants traditional search engines. DeepSeek-R1 showed higher accuracy and a more decisive approach to rumor detection, whereas the guarded stance of ChatGPT-4o sometimes led to less definitive answers. Future studies should optimize AI models' fact-checking capabilities, for example, by integrating real-time access to authoritative databases, enhancing domain-specific fine-tuning, and building human-machine collaborative monitoring systems. Continuous improvements in the accuracy and transparency of AI-driven medical communications will better protect patient health and reinforce public trust in evidence-based health care.

Acknowledgments

The authors would like to express their sincere gratitude to the Vascular Anomalies and Vascular Malformations Plastic Surgery Team at the Plastic Surgery Hospital of the Chinese Academy of Medical Sciences for their strong support of this research, the National Clinical Key Specialty Construction Project (23003), and the Plastic Medicine Research Fund of the Chinese Academy of Medical Sciences (2024-ZX-1-01). The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data Availability

All deidentified data that support the findings of this study are provided in Multimedia Appendix 1 (vascular anomaly information sources). Additional deidentified data and analysis materials that were generated during the study are available from the corresponding author on reasonable request for noncommercial purposes. Authors are prepared to provide the underlying (anonymized) data to the journal for inspection or verification upon request.

Authors' Contributions

GW wrote the main manuscript and designed the research methodology framework. Y Zhu and WW prepared Figures 1 and 2 and conducted data evaluation. GW, Y Zhang, and HB collected the data. GW and XY designed the study. All authors contributed to the statistical analysis and critically reviewed the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Vascular anomaly information sources.
[DOCX File (Microsoft Word File), 24 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Large language model (LLM) text classification prompts.
[DOCX File (Microsoft Word File), 14 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Likert scale for model output assessment.
[DOCX File (Microsoft Word File), 15 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Bidirectional encoder representations from transformers (BERT) semantic similarity code.
[DOCX File (Microsoft Word File), 16 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Bidirectional encoder representations from transformers (BERT) similarity scores model comparison.
[XLSX File (Microsoft Excel File), 16 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Bidirectional encoder representations from transformers (BERT) similarity ChatGPT-4o versus DeepSeek-R1.
[PNG File (Portable Network Graphics File), 148 KB-Multimedia Appendix 6]

Multimedia Appendix 7

YZ ratings for model reasonableness.
[XLSX File (Microsoft Excel File), 14 KB-Multimedia Appendix 7]

Multimedia Appendix 8

WW ratings for model reasonableness.
[XLSX File (Microsoft Excel File), 14 KB-Multimedia Appendix 8]

References

1. Fridman I, Boyles D, Chheda R, Baldwin-SoRelle C, Smith AB, Elston Lafata J. Identifying misinformation about unproven cancer treatments on social media using user-friendly linguistic characteristics: content analysis. JMIR Infodemiology. Feb 12, 2025;5:e62703. [doi: 10.2196/62703] [Medline: 39938078]
2. Wang J, Wang X, Yu A. Tackling misinformation in mobile social networks a BERT-LSTM approach for enhancing digital literacy. Sci Rep. Jan 7, 2025;15(1):1118. [doi: 10.1038/s41598-025-85308-4] [Medline: 39774143]
3. Schlicht IB, Fernandez E, Chulvi B, Rosso P. Automatic detection of health misinformation: a systematic review. J Ambient Intell Humaniz Comput. May 27, 2023;27:1-13. [doi: 10.1007/s12652-023-04619-4] [Medline: 37360776]
4. Rosenbacke R, Melhus Å, Stuckler D. False conflict and false confirmation errors are crucial components of AI accuracy in medical decision making. Nat Commun. Aug 13, 2024;15(1):6896. [doi: 10.1038/s41467-024-50952-3] [Medline: 39138179]
5. Lee Y, Ferber D, Rood JE, Regev A, Kather JN. How AI agents will change cancer research and oncology. Nat Cancer. Dec 2024;5(12):1765-1767. [doi: 10.1038/s43018-024-00861-7] [Medline: 39690222]
6. Menz BD, Modi ND, Abuhelwa AY, et al. Generative AI chatbots for reliable cancer information: evaluating web-search, multilingual, and reference capabilities of emerging large language models. Eur J Cancer. Mar 11, 2025;218:115274. [doi: 10.1016/j.ejca.2025.115274] [Medline: 39922126]
7. Huo B, Boyle A, Marfo N, et al. Large language models for chatbot health advice studies: a systematic review. JAMA Netw Open. Feb 3, 2025;8(2):e2457879. [doi: 10.1001/jamanetworkopen.2024.57879] [Medline: 39903463]
8. Maaz S, Palaganas JC, Palaganas G, Bajwa M. A guide to prompt design: foundations and applications for healthcare simulationists. Front Med (Lausanne). 2024;11:1504532. [doi: 10.3389/fmed.2024.1504532] [Medline: 39980724]
9. Meyrowitsch DW, Jensen AK, Sørensen JB, Varga TV. AI chatbots and (mis)information in public health: impact on vulnerable communities. Front Public Health. 2023;11:1226776. [doi: 10.3389/fpubh.2023.1226776] [Medline: 38026315]
10. Borges do Nascimento IJ, Pizarro AB, Almeida JM, et al. Infodemics and health misinformation: a systematic review of reviews. Bull World Health Organ. Sep 1, 2022;100(9):544-561. [doi: 10.2471/BLT.21.287654] [Medline: 36062247]
11. Classification of vascular anomalies. International Society for the Study of Vascular Anomalies. 2025. URL: https://www.issva.org/classification [Accessed 2025-03-26]
12. Sun Y, Su L, Zhang Y, et al. Dermatoscopic features differentiating among port wine stain, arteriovenous malformation, and capillary malformation-arteriovenous malformation syndrome: to detect potential fast-flow vascular malformations at an early stage. J Am Acad Dermatol. Dec 2022;87(6):1435-1437. [doi: 10.1016/j.jaad.2022.07.053] [Medline: 35952834]
13. Werner JA, Dünne AA, Lippert BM, Folz BJ. Optimal treatment of vascular birthmarks. Am J Clin Dermatol. 2003;4(11):745-756. [doi: 10.2165/00128071-200304110-00003] [Medline: 14572297]
14. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. Aug 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]

15. Normile D. Chinese firm's large language model makes a splash. Science. Jan 17, 2025;387(6731):238. [doi: 10.1126/science.adv9836] [Medline: 39818899]
16. National Health Commission, Ministry of Education, Ministry of Science and Technology, National Administration of Traditional Chinese Medicine. Measures for the Ethical Review of Life Science and Medical Research Involving Humans [Web page in Chinese]. The State Council of the People's Republic of China. URL: https://www.gov.cn/zhengce/zhengceku/2023-02/28/content_5743658.htm [Accessed 2025-11-13]
17. 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. Dec 14, 2023;25:e49771. [doi: 10.2196/49771] [Medline: 38096014]
18. Tan TF, Thirunavukarasu AJ, Campbell JP, et al. Generative artificial intelligence through ChatGPT and other large language models in ophthalmology: clinical applications and challenges. Ophthalmol Sci. Dec 2023;3(4):100394. [doi: 10.1016/j.xops.2023.100394] [Medline: 37885755]
19. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. Feb 15, 2020;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
20. Torrence D, Antonescu CR. The genetics of vascular tumours: an update. Histopathology. Jan 2022;80(1):19-32. [doi: 10.1111/his.14458] [Medline: 34958509]
21. Krowchuk DP, Frieden IJ, Mancini AJ, et al. Clinical practice guideline for the management of infantile hemangiomas. Pediatrics. Jan 2019;143(1):e20183475. [doi: 10.1542/peds.2018-3475] [Medline: 30584062]
22. Frenette C, Mendiratta-Lala M, Salgia R, Wong RJ, Sauer BG, Pillai A. ACG clinical guideline: focal liver lesions. Am J Gastroenterol. Jul 1, 2024;119(7):1235-1271. [doi: 10.14309/ajg.0000000000002857] [Medline: 38958301]
23. Fayos De Arizón L, Viera ER, Pilco M, et al. Artificial intelligence: a new field of knowledge for nephrologists? Clin Kidney J. Dec 2023;16(12):2314-2326. [doi: 10.1093/ckj/sfad182] [Medline: 38046016]
24. Kim TW. Application of artificial intelligence chatbots, including ChatGPT, in education, scholarly work, programming, and content generation and its prospects: a narrative review. J Educ Eval Health Prof. 2023;20:38. [doi: 10.3352/jeehp.2023.20.38] [Medline: 38148495]
25. Reddy S. Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. Implement Sci. Mar 15, 2024;19(1):27. [doi: 10.1186/s13012-024-01357-9] [Medline: 38491544]
26. Pugliese N, Wai-Sun Wong V, Schattenberg JM, et al. Accuracy, reliability, and comprehensibility of ChatGPT-generated medical responses for patients with nonalcoholic fatty liver disease. Clin Gastroenterol Hepatol. Apr 2024;22(4):886-889. [doi: 10.1016/j.cgh.2023.08.033] [Medline: 37716618]
27. Ismail N, Kbaier D, Farrell T, Kane A. The experience of health professionals with misinformation and its impact on their job practice: qualitative interview study. JMIR Form Res. Nov 2, 2022;6(11):e38794. [doi: 10.2196/38794] [Medline: 36252133]
28. Zakka C, Shad R, Chaurasia A, et al. Almanac - retrieval-augmented language models for clinical medicine. NEJM AI. Feb 2024;1(2). [doi: 10.1056/aioa2300068] [Medline: 38343631]
29. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature New Biol. Aug 2023;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
30. Alber DA, Yang Z, Alyakin A, et al. Medical large language models are vulnerable to data-poisoning attacks. Nat Med. Feb 2025;31(2):618-626. [doi: 10.1038/s41591-024-03445-1] [Medline: 39779928]

Abbreviations

AI: artificial intelligence
BERT: bidirectional encoder representations from transformers
ICC: interclass correlation coefficient
LLM: large language model

Edited by Yanshan Wang; peer-reviewed by Kaijun Zhang, Sandipan Biswas, Yunxuan Zhang; submitted 24.Apr.2025; final revised version received 31.Aug.2025; accepted 01.Oct.2025; published 18.Nov.2025

Please cite as:
Wang G, Zhang Y, Wang W, Zhu Y, Lu W, Wang C, Bi H, Yang X
Detection of Medical Misinformation in Hemangioma Patient Education: Comparative Study of ChatGPT-4o and DeepSeek-R1 Large Language Models
JMIR AI2025;4:e76372
URL: https://ai.jmir.org/2025/1/e76372

doi: 10.2196/76372