



OPEN ACCESS

EDITED BY

Ariel Soares Teles,
Federal Institute of Education, Science and
Technology of Maranhão, Brazil

REVIEWED BY

Shoukat Arain,
Alfaisal University, Saudi Arabia
Emre Emekli,
Eskişehir Osmangazi University, Türkiye

*CORRESPONDENCE

Chuan Wang
✉ wangchuan1308@163.com
Kaiyu Zhou
✉ Drkaiyuzhou@scu.edu.cn

[†]These authors have contributed equally to
this work

RECEIVED 17 February 2025

ACCEPTED 06 March 2025

PUBLISHED 31 March 2025

CITATION

Yan C, Li Z, Liang Y, Shao S, Ma F, Zhang N,
Li B, Wang C and Zhou K (2025) Assessing
large language models as assistive tools in
medical consultations for Kawasaki disease.
Front. Artif. Intell. 8:1571503.
doi: 10.3389/frai.2025.1571503

COPYRIGHT

© 2025 Yan, Li, Liang, Shao, Ma, Zhang, Li,
Wang and Zhou. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License
\(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication
in this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Assessing large language models as assistive tools in medical consultations for Kawasaki disease

Chunyi Yan^{1,2†}, Zexi Li^{3†}, Yongzhou Liang^{1,2}, Shuran Shao^{1,2},
Fan Ma^{1,2}, Nanjun Zhang^{1,2}, Bowen Li^{1,2}, Chuan Wang^{1,2*} and
Kaiyu Zhou^{1,2*}

¹Department of Pediatric Cardiology, West China Second University Hospital, Sichuan University, Chengdu, China, ²Key Laboratory of Birth Defects and Related Diseases of Women and Children (Sichuan University), Ministry of Education, Chengdu, China, ³Department of Cardiology, West China Hospital, Sichuan University, Chengdu, China

Background: Kawasaki disease (KD) presents complex clinical challenges in diagnosis, treatment, and long-term management, requiring a comprehensive understanding by both parents and healthcare providers. With advancements in artificial intelligence (AI), large language models (LLMs) have shown promise in supporting medical practice. This study aims to evaluate and compare the appropriateness and comprehensibility of different LLMs in answering clinically relevant questions about KD and assess the impact of different prompting strategies.

Methods: Twenty-five questions were formulated, incorporating three prompting strategies: No prompting (NO), Parent-friendly (PF), and Doctor-level (DL). These questions were input into three LLMs: ChatGPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro. Responses were evaluated based on appropriateness, educational quality, comprehensibility, cautionary statements, references, and potential misinformation, using Information Quality Grade, Global Quality Scale (GQS), Flesch Reading Ease (FRE) score, and word count.

Results: Significant differences were found among the LLMs in terms of response educational quality, accuracy, and comprehensibility ($p < 0.001$). Claude 3.5 provided the highest proportion of completely correct responses (51.1%) and achieved the highest median GQS score (5.0), outperforming GPT-4o (4.0) and Gemini 1.5 (3.0) significantly. Gemini 1.5 achieved the highest FRE score (31.5) and provided highest proportion of responses assessed as comprehensible (80.4%). Prompting strategies significantly affected LLM responses. Claude 3.5 Sonnet with DL prompting had the highest completely correct rate (81.3%), while PF prompting yielded the most acceptable responses (97.3%). Gemini 1.5 Pro showed minimal variation across prompts but excelled in comprehensibility (98.7% under PF prompting).

Conclusion: This study indicates that LLMs have great potential in providing information about KD, but their use requires caution due to quality inconsistencies and misinformation risks. Significant discrepancies existed across LLMs and prompting strategies. Claude 3.5 Sonnet offered the best response quality and accuracy, while Gemini 1.5 Pro excelled in comprehensibility. PF prompting with Claude 3.5 Sonnet is most recommended for parents seeking KD information. As AI evolves, expanding research and refining models is crucial to ensure reliable, high-quality information.

KEYWORDS

artificial intelligence, large language models, medical information, medical education, Kawasaki disease

Background

Kawasaki disease (KD), also known as mucocutaneous lymph node syndrome, is a serious pediatric disease that predominantly affects infants and young children younger than 5 years of age (Wang et al., 2005). Initially, KD manifests with high fever, mucocutaneous inflammation, and cervical lymphadenopathy, which may later progress to involve the coronary arteries and other cardiovascular structures (Kato et al., 1975). KD remains a significant challenge in pediatric healthcare, particularly among East Asian populations. In Japan, the annual incidence of KD reaches approximately 200 cases per 100,000 children, indicating that nearly 1% of children by the age of 4–5 may be affected (Elakabawi et al., 2020; Nakamura et al., 2012). Even in the United States, around 5,440 hospitalizations for KD were reported in 2016 (Iwata et al., 2024). Notably, KD has surpassed rheumatic fever as the primary etiology of acquired cardiac disorders in children across developed nations (Newburger et al., 2016). The severity and complexity of KD's cardiovascular implications underscore the importance of early detection and intervention. For instance, coronary artery aneurysms (CAA), a critical complication of KD, develop in 25% of untreated cases, but this rate reduces to 4% with timely intervention (McCrindle et al., 2017). Myocardial infarction or rupture of CAA caused by KD can lead to cardiogenic shock or even sudden death. Moreover, as KD patients grow older, the pre-existing cardiac complications often correlate with conditions such as arrhythmias, heart failure, and peripheral arterial occlusion (An et al., 2021). Additionally, even after the acute phase has been controlled, KD patients, particularly those who have previously developed CAA, necessitate lifelong cardiovascular management, as they remain at high risk for coronary thrombosis and stenosis (Kato et al., 1996). Therefore, it is imperative for both parents and physicians to cultivate a comprehensive understanding of Kawasaki disease, from initial diagnosis and treatment to long-term management after recovery.

Currently, the internet has become a dominant platform for accessing health-related information, with a steadily increasing number of individuals utilizing these resources (Daraz et al., 2019). And many patients turn to search engines and online medical content instead of consulting healthcare professionals (Popovac and Roomaney, 2022). In tandem with this digital shift, Artificial intelligence (AI) has established itself as an integral component in the field of medicine (Topol, 2019; Elemento et al., 2021). Recent advancement in AI have catalyzed the development of large language models (LLMs) such as Generative Pre-trained Transformer (GPT) (Sufi, 2024), which shows potential for enhancing medical practice assistance (Haug and Drazen, 2023). Both patients and healthcare providers can access medical information through online AI tools (Hopkins et al., 2023). An

increasing number of studies are investigating the potential prospects of LLMs in medicine. These models have been applied in various domains, including diagnostic examination image recognition (Günay et al., 2024; Dehdab et al., 2024), case diagnoses (Lechien et al., 2024) and medical tests (Massey et al., 2023; Ali et al., 2023). With the increasing accessibility and development of LLMs, their application in healthcare has expanded significantly. Prompt engineering, as an emerging field, focuses on the systematic development and refinement of prompts to optimize the performance of LLMs (Wang et al., 2024). The design and quality of prompts can significantly affect the accuracy, relevance, and usefulness of the information produced by these models (Kiyak and Emekli, 2024). More and more individuals may rely on LLMs to seek information regarding health conditions, medications, and medical procedures. Therefore, evaluating the quality and accuracy of LLM-generated content is essential for comprehending information being consumed by patients (Yang et al., 2024).

While existing literature has explored the accuracy and reliability of LLMs in various medical domains, there remains a notable gap in evaluating their performance regarding KD. Given the diverse and complex clinical manifestations of KD, it is imperative to assess the quality of information retrieved using LLMs on this condition. Hence, this research aimed to assess and compare the appropriateness and comprehensibility of different LLMs to clinically pertinent questions related to KD. To evaluate their performance, three distinct prompts were designed to compare the effectiveness of the LLMs under varying conditions.

Methods

This study was performed during October 2024. To evaluate patient-related information pertinent to KD, a total of 25 questions were carefully formulated through a multi-step process (Table 1). Initially, a preliminary list of 50 questions was generated based on a hospital-wide survey assessing parents' most frequent inquiries and concerns regarding their children's KD diagnosis and treatment, and informal interviews with KD patients' families during follow-up visits. Subsequently, a panel comprising three senior pediatricians, each with over 10 years of experience in general pediatrics and KD management, independently scored these 50 questions and selected 25 highest-scoring questions. The scoring criteria were based on frequency of occurrence and clinical significance. The panel then conducted a series of discussions to refine the question list, ensuring comprehensive coverage of clinically relevant topics. The detailed prompts are provided in Table 2. To minimize any potential bias on the responses, each question was entered in a separate dialogue box.

In this study, we utilized three advanced large language models (LLMs): ChatGPT-4 "Omni" [GPT-4o] (GPT-4, OpenAI, San Francisco, California, United States), Claude 3.5 Sonnet [Claude 3.5] (Claude 3, Anthropic, San Francisco, California, United States), and Gemini 1.5 Pro [Gemini 1.5] (LaMDA, Google, Mountain View, California, United States). The responses generated by these models

Abbreviations: KD, Kawasaki disease; AI, Artificial intelligence; LLMs, Large language models; PF, Parent-friendly; DL, Doctor-level; NO, No prompting; GQS, Global Quality Scale; FRE, Flesch Reading Ease; CAA, Coronary artery aneurysms; GPT, Generative Pre-trained Transformer; WoC, Word count; ICC, Intraclass correlation coefficient; HNSCC, Head and neck squamous cell carcinoma.

TABLE 1 Questions concerning Kawasaki disease.

Number	Questions
Q1	What is Kawasaki disease?
Q2	How common is Kawasaki disease in children?
Q3	What are the common symptoms of Kawasaki disease?
Q4	Is Kawasaki disease hereditary?
Q5	What are the types of Kawasaki disease?
Q6	How can I differentiate between typical and atypical Kawasaki disease?
Q7	How is Kawasaki disease diagnosed?
Q8	How is atypical Kawasaki disease diagnosed?
Q9	A child has a high fever. Should I be worried about Kawasaki disease?
Q10	What complications can arise from Kawasaki disease?
Q11	Why does Kawasaki disease cause damage to the heart's coronary arteries?
Q12	What tests will be used to diagnose Kawasaki disease?
Q13	When should echocardiography be done for Kawasaki patients?
Q14	What are the common treatment options for Kawasaki disease?
Q15	Why is immunoglobulin used to treat Kawasaki disease?
Q16	What role does aspirin play in the treatment of Kawasaki disease?
Q17	When would children need steroids to treat Kawasaki disease?
Q18	Are there any side effects from using steroids to treat Kawasaki disease?
Q19	When would a second dose of immunoglobulin be needed in the treatment of Kawasaki disease?
Q20	How can I help children prevent coronary artery problems caused by Kawasaki disease?
Q21	When can children with Kawasaki disease stop taking aspirin?
Q22	How does the doctor evaluate blood test results for Kawasaki disease?
Q23	Does children with Kawasaki disease need long-term management, and how should it be managed?
Q24	What should we pay special attention to during children's recovery from Kawasaki disease?
Q25	Will children have long-term effects from Kawasaki disease?

were systematically evaluated and compared. Each question was posed three times, and the responses were independently reviewed by two experienced physicians, each blinded to the other's assessments. In case of any discrepancies, a third reviewer was consulted to arbitrate between the two initial assessments. This process ensured the achievement of a unanimous result while maintaining the integrity of the dual-reviewer system. To assess responses on questions concerning KD comprehensively, both subjective ratings and objective evaluations were employed.

TABLE 2 Prompts applied for each form.

Form name	Prompt provided
No prompting (NO)	(Input the instruction directly)
Parent-friendly prompting (PF)	I am a parent looking to learn more about Kawasaki disease. Please provide answers to the question below in the language that would be appropriate for my understanding. Ensure that your responses are accurate and thorough.
Doctor-level prompting (DL)	I am a doctor attempting to learn the most up-to-date information on Kawasaki disease. Please provide answers to the question below in the language that would be appropriate for my expert-level understanding of medical concepts. Be as specific as possible in your answers.

Each prompt was input before the question.

Initially, evaluators employed an Information Quality Grade, with the following criteria: Grade 1: completely correct; Grade 2: correct but inadequate; Grade 3: partially correct; Grade 4: entirely incorrect or irrelevant. Then the Global Quality Scale (GQS) was utilized to provide a more detailed evaluation of the response quality (Table 3). Grades 1 and 2 could be collectively referred to as "acceptable responses." These two information quality assessment tools, previously utilized and validated in prior studies (Wang et al., 2024; Cakir et al., 2024; Dyckhoff-Shen et al., 2024; Onder et al., 2024; Ozgor et al., 2024), demonstrated efficacy in evaluating both the correctness and comprehensiveness of the provided information. Subsequently, responses rated at grade 3 or 4 were subjected to further analysis to categorize the types of errors. These were classified into five distinct categories, (detailed in Table 3). Additionally, the responses were evaluated for different aspects: (Wang et al., 2005) comprehensibility, (Kato et al., 1975) cautionary statements, reference, (Elakabawi et al., 2020) confabulation, as shown in Table 3.

To quantify readability objectively, we employed two metrics: word count (WoC) and the Flesch Reading Ease (FRE) score. The FRE score, a validated readability measure, assesses text complexity on a 100-point scale, with higher scores indicating greater ease of reading (Jindal and MacDermid, 2017; Hillmann et al., 2023). The interpretation of FRE score is detailed in Table 4. The FRE score is calculated based on two variables: average sentence length (based on the number of words) and average word length (based on the number of syllables) (Jindal and MacDermid, 2017). Both WoC and FRE score were computed by Microsoft Word 365, (Microsoft, Redmond, Washington, United States).

Statistical analyses were performed by SPSS 26.0 (IBM, New York, NY, United States). Categorical variables, including the Information Quality Grade, were calculated and represented as frequencies and percentages, while continuous variables were represented as median values with interquartile ranges due to their non-normal distribution. Reliability, referred to the repeatability of responses to identical questions, was evaluated using the intraclass correlation coefficient (ICC) of the GQS scores. ICC values were interpreted as follows: excellent reliability (> 0.90), good reliability (0.75–0.90), moderate reliability (0.50–0.75) and poor reliability (< 0.50) (Koo and Li, 2016). Comparative analyses between LLMs and prompt types were

TABLE 3 Procedures evaluating the response quality of LLMs.

Criterion	Description
Information Quality Grade	
Grade 1	Completely correct (Response is entirely accurate and up-to-date with current KD guidelines)
Grade 2	Correct but inadequate (Response is accurate but misses some important aspects of KD)
Grade 3	Partially correct (containing a mixture of accurate and inaccurate information about KD)
Grade 4	Entirely incorrect or irrelevant (Responses directly contradicts the guidelines or offers information irrelevant with KD)
Global Quality Scale (GQS)	
1	Poor quality, poor flow of the site, most information missing, not at all useful for patients
2	Generally poor quality and poor flow, some information listed but many important topics missing, of very limited use to patients
3	Moderate quality, suboptimal flow, some important information is adequately discussed but others poorly discussed, somewhat useful for patients
4	Good quality and generally good flow, most of the relevant information is listed, but some topics not covered, useful for patients
5	Excellent quality and excellent flow, very useful for patients
Error classifications	
A	Misunderstanding of medical terms or jargon
B	Incorrect usage of medical terms
C	Errors in diagnosis/treatment/management
D	Entirely irrelevant information
E	A combination of two or more error types among A–C
Comprehensibility	
Yes/no	Refers to the clarity and ease of understanding for patients
Cautionary statements	
Yes/no	Refers to the presence of a recommendation to consult licensed healthcare professionals or the provision of explicit disclaimers
Reference	
Yes/no	Refers to the presence of any referenced studies in the response
Confabulation	
Yes/no	Refers to the presence of fabricated or distorted information in the response

TABLE 4 Interpretation of FRE score.

Score	Readability
91–100	Very easy
81–90	Easy
71–80	Fairly easy
61–70	Standard
51–60	Fairly difficult
31–50	Difficult
0–30	Very difficult

FRE, Flesch Reading Ease.

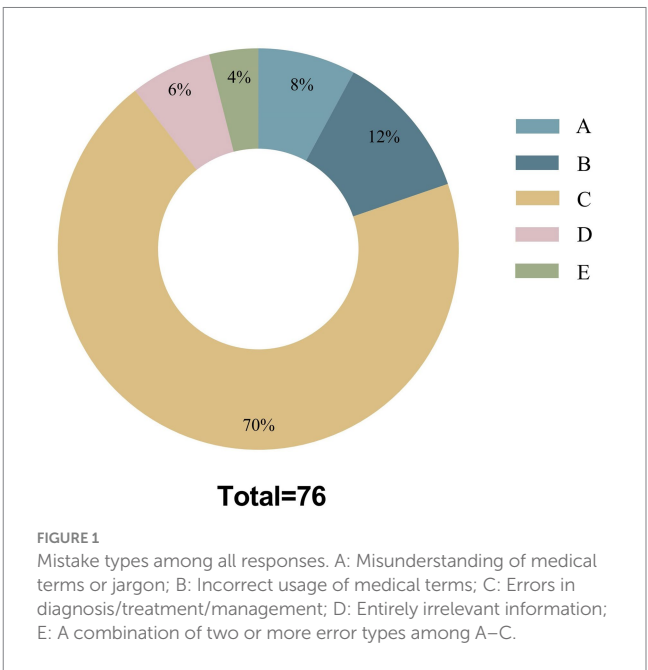
conducted using the Kruskal-Wallis test, while categorical data were analyzed via chi-square tests. The Bonferroni correction was applied for post-hoc analysis. All p values were two-tailed, with statistical significance set as $p < 0.05$.

Results

Evaluation of responses of different LLMs

There were significant differences in Information Quality Grade, GQS scores, WoC, FRE scores, comprehensibility and cautionary statements among different LLMs ($p < 0.001$). The likelihood of Claude 3.5 giving completely correct responses was 51.1%, which was significantly higher than 18.2% of GPT-4o and 11.1% of Gemini 1.5. Additionally, most of the responses provided by Gemini 1.5 were correct but inadequate (78.2%), significantly higher than GPT-4o (67.1%) and Claude 3.5 (40.4%), with all differences among the three models being statistically significant. There was no significant difference in the frequency of Information Quality Grade 3 or 4 across different LLMs, and it was found that the most prevalent errors were related to diagnosis, treatment, and management (70%) in these responses containing inaccurate or incorrect information (Figure 1). The median GQS values of different LLMs were as follows: Claude 3.5 at 5.0, GPT-4o at 4.0, and Gemini 1.5 at 3.0, with all inter-model comparisons demonstrating statistically significant differences (Table 5).

The WoC of responses from GPT-4o (231.0) was significantly lower compared to Claude 3.5 (281.0) and Gemini 1.5 (292.0). Furthermore, in calculating FRE scores, Gemini 1.5 exhibited the highest score (31.5), which was significantly higher than GPT-4o (16.0) and Claude 3.5 (19.5). Both the percentage of responses assessed as comprehensible and the percentage containing cautionary statements from Gemini 1.5 were significantly higher compared to those from GPT-4o and Claude 3.5, with detailed information shown in Table 5. Additionally, Gemini



1.5 consistently referenced studies or websites in all its responses, and the articles or websites cited were verifiable. Conversely, neither GPT-4o nor Claude 3.5 included references in any of their responses.

The ICC values of GQS scores and associated *p*-values for three LLMs responding to three prompts were shown in [Supplementary Table 1](#), where notably, the majority of *p*-values exceeded 0.05, indicating that the observed ICCs may not reflect statistically significant reliability. This outcome may be attributed to insufficient sample sizes within the subgroups. Upon analyzing the reliability of different LLMs, GPT-4o (ICC: 0.429, 95% CI: 0.289–0.565) and Gemini 1.5 (ICC: 0.442, 95% CI: 0.302–0.576) exhibited a poor to moderate reliability, while Claude 3.5 (ICC: 0.334, 95% CI: 0.192–0.480) exhibited a poor reliability ([Supplementary Table 2](#)).

Evaluation of information quality across different LLMs and prompts

The findings indicated that Claude 3.5 outperformed the other models in terms of GQS scores and the frequency of achieving Information Quality Grade 1 ([Table 5](#)). The proportions of acceptable responses across all models and prompts were considerable, exceeding 80% ([Figure 2](#)). The combination of Claude 3.5 and DL achieved the highest GQS score of 5.0 (5.0–5.0) and the highest proportion of completely correct responses at 81.3%, while the combination of Claude 3.5 and PF achieved the highest proportion of acceptable responses at 97.3%. Notably, the combination of Claude 3.5 and DL yielded the highest incidence of confabulation, reaching 12% ([Figure 3C](#)), and there was a significant difference observed between DL and NO prompting in Claude 3.5 ([Table 6](#)).

The proportions of completely correct responses under DL prompting were significantly higher than those under NO and PF prompting across all three models. Conversely, the proportion of acceptable responses under DL prompting was lower than that under PF prompting in these models, with a significant difference observed in Claude 3.5. Furthermore, no significant difference in GQS scores was observed between PF and DL prompting in Claude

3.5. For GPT-4o and Gemini 1.5, the GQS scores under DL prompting were significantly higher than those under PF and NO prompting. For GPT-4o and Claude 3.5, the GQS scores and proportions of completely correct responses under PF prompting were significantly higher than those under NO prompting ([Table 6](#)).

Analysis of the cautionary statements across different LLMs and prompts

The combination of Gemini 1.5 and PF achieved the highest proportion (77.3%) of reminders for users to consult a licensed healthcare professionals or provide of explicit disclaimers in its responses. The proportion of cautionary statements provided under DL prompting was significantly lower than that under NO and PF prompting across all three models, with values of 1.3, 6.7, and 20.0% for GPT-4o, Claude 3.5, and Gemini 1.5, respectively, indicating a notably low level. The proportion of cautionary statements provided under PF prompting was significantly higher than that under NO prompting in GPT-4o and Claude3.5, while in Gemini 1.5, there was no significant difference observed between NO (69.3%) and PF prompting ([Table 6](#)). Overall, the percentages of cautionary statements provided by all three models under PF prompting, as well as those from the combination of Gemini 1.5 and NO, were considerable, around 70%. ([Figure 3B](#)).

Evaluation of comprehensibility and readability across different LLMs and prompts

The percentages of responses assessed as comprehensible under PF prompting by all three models, as well as those from the combination of Gemini 1.5 and NO, were substantial, exceeding 90%. In contrast, the proportions of responses assessed as comprehensible under DL prompting were significantly lower than those under NO and PF prompting across all three models, with respective values of 25.3, 24.0, and 48.0% for GPT-4o, Claude 3.5, and Gemini 1.5,

TABLE 5 Assessment of responses given by three different LLMs.

	ChatGPT-4o	Claude 3.5 Sonnet	Gemini 1.5 Pro	<i>P</i>
Information Quality Grade, <i>n</i> (%)				
1	41 (18.2)	115 (51.1)*	25 (11.1) [‡]	< 0.001
2	151 (67.1)	91 (40.4)*	176 (78.2) ^{†‡}	
3	30 (13.3)	18 (8.0)	22 (9.8)	
4	3 (1.4)	1 (0.5)	2 (0.9)	
GQS score	4.0 (3.0–4.0)	5.0 (4.0–5.0)*	3.0 (3.0–4.0) ^{†‡}	< 0.001
WoC	231.0 (186.0–267.5)	281.0 (218.5–341.0)*	292.0 (222.5–369.5) [†]	< 0.001
FRE score	16.0 (0.0–31.3)	19.5 (9.0–31.2)	31.5 (14.2–44.6) ^{†‡}	< 0.001
Comprehensibility, <i>n</i> (%)	146 (64.9)	154 (68.4)	181 (80.4) ^{†‡}	< 0.001
Cautionary statements, <i>n</i> (%)	80 (35.6)	79 (35.1)	125 (55.6) ^{†‡}	< 0.001
Reference, <i>n</i> (%)	0 (0)	0 (0)	225 (100) ^{†‡}	< 0.001
Confabulation, <i>n</i> (%)	13 (5.8)	12 (5.3)	9 (4.0)	0.675

GQS, Global Quality Scale; FRE, Flesch Reading Ease; WoC, Word count.

*Claude vs. ChatGPT significantly (*P* < 0.05), [†]Gemini vs. ChatGPT significantly (*P* < 0.05), [‡]Gemini vs. Claude significantly (*P* < 0.05). Bold value: *P* < 0.05.

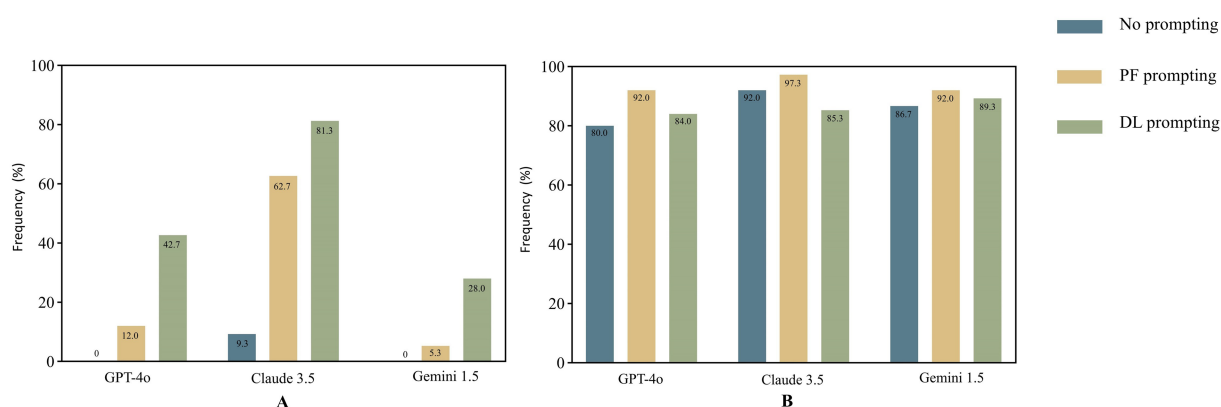


FIGURE 2
Frequency of Information Quality Grade 1 (A) and Grades 1&2 (B) across models and prompts.

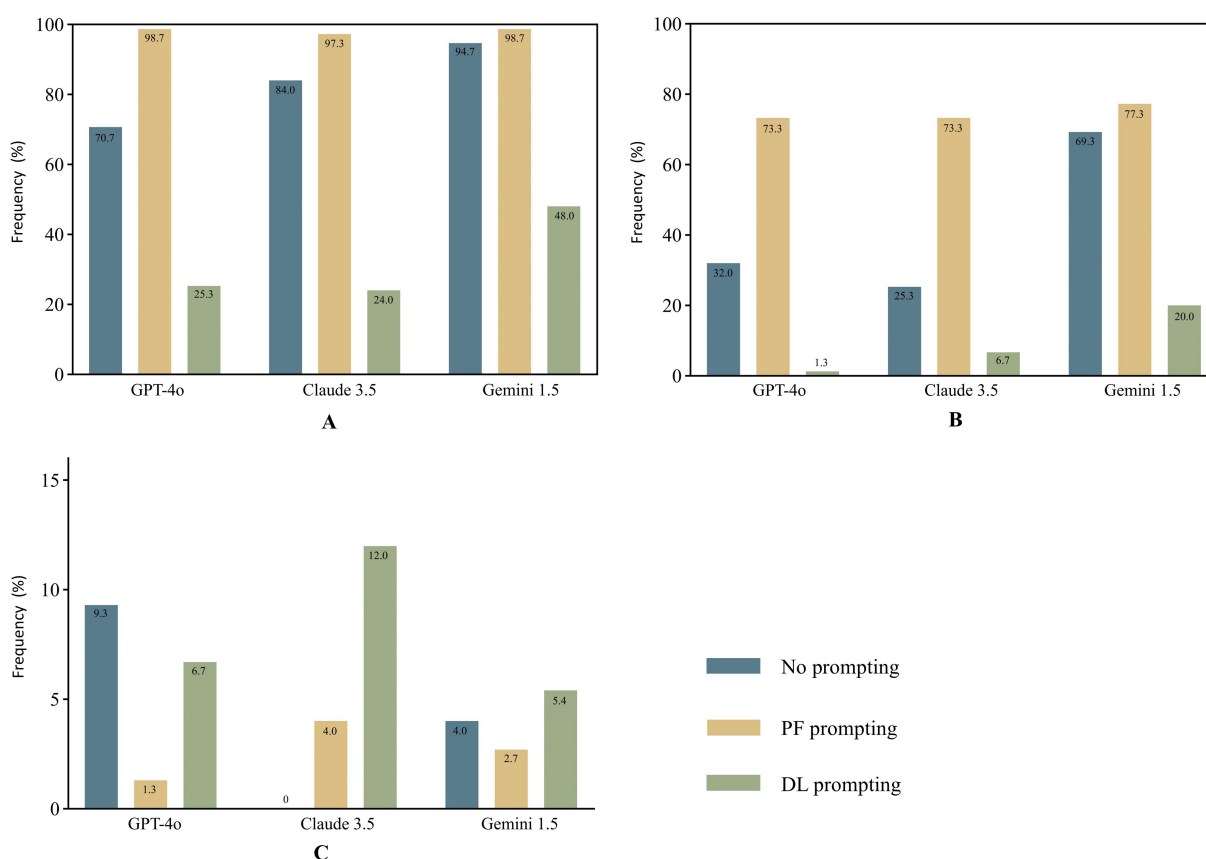


FIGURE 3
Assessment of comprehensibility (A), cautionary statements (B) and confabulation (C) across models and prompts.

respectively, suggesting a substantially inadequate level (Figure 3A). The combination of Gemini 1.5 and PF achieved the highest FRE score of 45.3 (40.4–50.8), while the combination of Gemini 1.5 and DL attained the highest WoC of 386 (292–430) (Figure 4). The FRE scores of responses under PF promoting were significantly higher than those under NO and PF prompting across all three model. Detailed analysis of WoC and FRE is presented in Table 6.

Discussion

Approximately 70–80% of internet users actively seek health-related information online, which plays a crucial role in shaping their initial understandings and perceptions of health issues (Finney Rutten et al., 2019). With the rising prevalence and application of LLMs, there might be an increasing number of individuals depending on these

TABLE 6 Assessment of responses across different prompts in different LLMs.

	No prompting	Parent-friendly prompting	Doctor-level prompting	<i>P</i>
ChatGPT-4o				
Information Quality Grade 1, <i>n</i> (%)	0 (0.0)	9 (12.0)*	43 (42.7) ^{†‡}	< 0.001
Information Quality Grades 1&2, <i>n</i> (%)	60 (80.0)	69 (92.0)	63 (84.0)	0.107
GQS score	3.0 (3.0–3.0)	4.0 (3.0–4.0)*	4.0 (4.0–5.0) ^{†‡}	< 0.001
WoC	183 (156–221)	224 (212–258)*	267 (227–309) ^{†‡}	< 0.001
FRE score	17.5 (8.5–27.0)	31.9 (26.9–38.8)*	0.0 (0.0–2.6) ^{†‡}	< 0.001
Comprehensibility, <i>n</i> (%)	53 (70.7)	74 (98.7)*	19 (25.3) ^{†‡}	< 0.001
Cautionary statements, <i>n</i> (%)	24 (32.0)	55 (73.3)*	1 (1.3) ^{†‡}	< 0.001
Confabulation, <i>n</i> (%)	7 (9.3)	1 (1.3)	5 (6.7)	0.089
Claude 3.5 Sonnet				
Information Quality Grade 1, <i>n</i> (%)	7 (9.3)	47 (62.7)*	61 (81.3) ^{†‡}	< 0.001
Information Quality Grades 1&2, <i>n</i> (%)	69 (92.0)	73 (97.3)	64 (85.3) [‡]	0.030
GQS score	4.0 (3.0–4.0)	5.0 (4.0–5.0)*	5.0 (5.0–5.0) [†]	< 0.001
WoC	196 (170–239)	281 (248–331)*	342 (308–386) ^{†‡}	< 0.001
FRE score	22.7 (15.7–29.3)	33.1 (27.9–38.9)*	3.8 (0.0–11.3) ^{†‡}	< 0.001
Comprehensibility, <i>n</i> (%)	63 (84.0)	73 (97.3)*	18 (24.0) ^{†‡}	< 0.001
Cautionary statements, <i>n</i> (%)	19 (25.3)	55 (73.3)*	5 (6.7) ^{†‡}	< 0.001
Confabulation, <i>n</i> (%)	0 (0.0)	3 (4.0)	9 (12.0) [†]	0.003
Gemini 1.5 Pro				
Information Quality Grade 1, <i>n</i> (%)	0 (0.0)	4 (5.3)	21 (28.0) ^{†‡}	< 0.001
Information Quality Grades 1&2, <i>n</i> (%)	65 (86.7)	69 (92.0)	67 (89.3)	0.571
GQS score	3.0 (3.0–3.0)	3.0 (3.0–3.0)	4.0 (3.0–5.0) ^{†‡}	< 0.001
WoC	271 (216–323)	252 (186–308)	386 (292–430) ^{†‡}	< 0.001
FRE score	31.5 (26.5–43.6)	45.3 (40.4–50.8)*	9.8 (3.7–16.1) ^{†‡}	< 0.001
Comprehensibility, <i>n</i> (%)	71 (94.7)	74 (98.7)	36 (48.0) ^{†‡}	< 0.001
Cautionary statements, <i>n</i> (%)	52 (69.3)	58 (77.3)	15 (20.0) ^{†‡}	< 0.001
Confabulation, <i>n</i> (%)	3 (4.0)	2 (2.7)	4 (5.4)	0.646

GQS, Global Quality Scale; FRE, Flesch Reading Ease; WoC, Word count.

*Parent-friendly prompting vs. No prompting significantly ($P < 0.05$), [†]Doctor-level prompting vs. No prompting significantly ($P < 0.05$), [‡]Doctor-level prompting vs. Parent-friendly prompting significantly ($P < 0.05$). Bold value: $P < 0.05$.

models to obtain details on health conditions, which necessitates healthcare professionals to engage proactively in developing standards and evaluating the quality of information provided by these models (Rajpurkar et al., 2022). Nonetheless, comprehensive data on the utilization of LLMs by healthcare professionals and parents, especially within pediatrics and specifically for KD, remain scarce. This study presents the first evaluation of LLMs' performance on addressing clinically pertinent questions of KD. Three recently published LLMs, namely Claude 3.5 Sonnet, Gemini 1.5 Pro, and ChatGPT-4o were applied and we also devised two prompts in the perspective of parents and doctors to identify the optimal combination of models and prompting strategy.

Through quantitative scoring approaches including GQS, FRE scores, and additional indices such as Information Quality Grade and

WoC, we assessed the responses with an emphasis on educational quality and comprehensibility, while also considering other dimensions. The current results reveals that the responses generated by LLMs: (1) vary in educational quality and proportions in completely correct answers, but generally meet acceptable standards; (2) are mostly easy to understand with varying readability; (3) were unsatisfactory in terms of consistent quality; (4) may exhibit inaccuracies and instances of confabulation occasionally; (5) exhibited notable differences across different LLMs and prompting strategies. Among the 76 inaccurate responses, it was discerned that the most frequent errors pertained to diagnosis, treatment, and management (70%). This prevalence of errors could be attributed to the fact that the majority of inquiries involved KD diagnostics and treatment. Moreover, the variability in treatment protocols across different

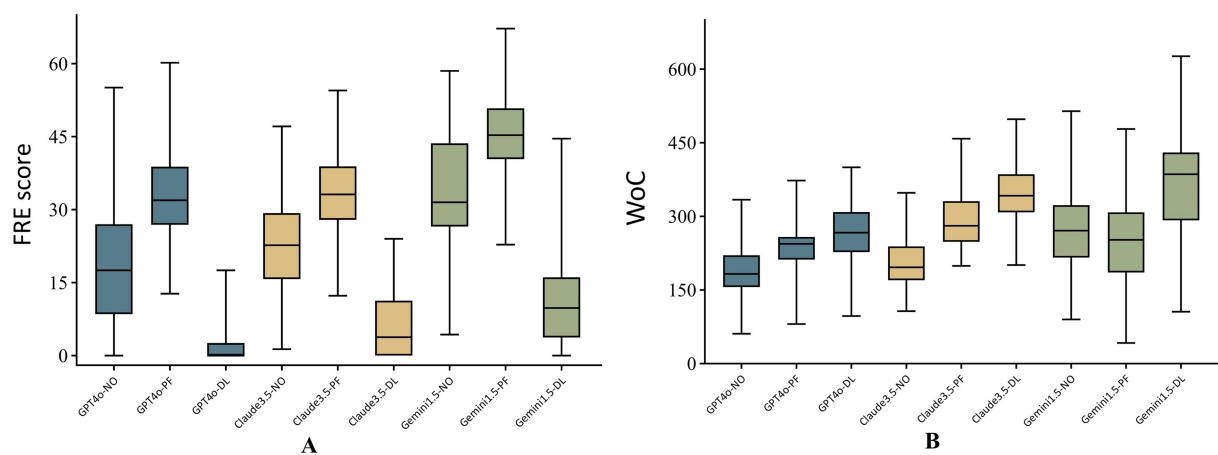


FIGURE 4
F1E scores (A) and word count (B) across models and prompts.

medical centers may result in discrepancies between evaluators' assessments and the standards employed by the LLMs.

In the comparative analysis, Claude 3.5 Sonnet demonstrated superior response quality and accuracy, with significantly higher GQS scores (5.0 vs. 4.0 for ChatGPT and 3.0 for Gemini) and a greater proportion of completely correct responses (51.1% vs. 18.2% for ChatGPT and 11.1% for Gemini). This finding aligns with some previous research in the field of LLM application. For instance, Schmidl et al. (2024) reported that Claude 3 Opus demonstrates superior performance in diagnosing head and neck squamous cell carcinoma (HNSCC) compared to ChatGPT 4.0. In contrast, certain studies have suggested a different finding. Meyer et al. (2024) analyzed LLM responses to questions encountered in rhinoplasty practice and found that ChatGPT surpassed both Claude and Gemini in terms of accuracy and overall quality. Additionally, Liu et al. (2024) found no significant difference between the Claude 3 Opus and ChatGPT 4.0 in dermoscopic image analysis for melanoma diagnosis. ChatGPT-4o, on the other hand, provided responses that tended to be more concise, characterized by a lower median word count (231 vs. 281 for Claude and 292 for Gemini). While such brevity can be advantageous in certain contexts, it may result in less comprehensive answers that could omit critical details necessary for fully understanding the topic. Nonetheless, GPT-4o still maintained a reasonable level of accuracy, with GQS scores significantly higher than those of Gemini 1.5 Pro. Meanwhile, in an analysis by Wang et al. (2024) concerning LLM responses on myopia prevention and control, ChatGPT-4.0 exhibited a significantly higher word count than Claude 2 and Gemini, while also achieving the highest scores in comprehensiveness, accuracy, and relevance. These discrepancies may be attributed to various factors, such as differences in studied fields, application scenarios and model versions.

Among three models, Gemini 1.5 Pro excelled in delivering comprehensible responses, as reflected in its higher F1E scores (31.5 vs. 16.0 for ChatGPT and 19.5 for Claude) and percentages of comprehensibility (80.4% vs. 64.9% for ChatGPT and 68.4% for Claude). However, it is important to note that the GQS scores of Gemini were significantly lower than those of the other two models, indicating a potential risk of not delivering high-quality information. This observation is consistent with previous studies (Tepe and Emekli, 2024; Gondode et al., 2024). Despite its lower accuracy, Gemini remains more

accessible to a broader patient audience. Notably, each response from Gemini in our study was accompanied by references and external websites that could be independently verified. In some instances, Gemini even cited images from external websites to bolster its textual explanations. These features enhanced the credibility of the information provided, making Gemini suitable for users who prioritize verifiable and easy-to-understand information, such as parents of children with KD. Previous studies had indicated that up to half of all citations presented by ChatGPT were fabricated (Bhattacharyya et al., 2023). And the phenomenon of falsified citations had been extensively documented in the literature (Walters and Wilder, 2023). This highlighted the technological advancements in Gemini, which had effectively mitigated this issue. Additionally, Gemini exhibited a higher proportion of responses containing cautionary statements, which recommended consulting professionals or included explicit disclaimers. While these statements could encourage parents to be more vigilant and nuanced in evaluating the information, they might also imply a lack of confidence in the medical judgments provided. Therefore, further investigations are warranted to assess the potential risks and benefits of this response mode (Wang et al., 2024).

An appropriate prompt could improve the accuracy and comprehensibility of responses to medical questions and different prompts had variable effects across diverse models (Wang et al., 2024). In our study, we found that DL prompting consistently outperformed NO prompting in terms of overall educational quality. However, from assessment of comprehensibility and readability, the responses generated under DL prompting were often more challenging for users to understand. Furthermore, within the Claude 3.5, there was a significantly higher incidence of confabulation with DL prompting compared to NO prompting, and the proportion of acceptable responses under DL prompting was significantly lower than that observed with PF prompting. These findings illustrated that while DL prompting could provide high-quality and comprehensive information, it may also carry a higher likelihood of inaccuracies in certain scenarios. Nonetheless, recommending the use of DL prompting for medical professionals might still prove beneficial, as it offered deeper insights, and medical professionals were more likely to critically evaluate the information provided. For ChatGPT-4o and Claude 3.5 Sonnet, the responses generated under PF prompting demonstrated significant improvements

in educational quality, accuracy, and comprehensibility compared to those generated using NO prompting, which indicated that PF prompting not only enhanced the quality, but also facilitated better user understanding. Moreover, the performance of Gemini 1.5 Pro did not exhibit significant differences in response quality and comprehensibility between PF and NO prompting. However, PF prompting within Gemini did result in superior FRE scores and interestingly, incorporated the use of metaphors in certain responses, which could aid readability and make the information more accessible for patients.

Among the three models, Claude 3.5 Sonnet demonstrated the highest overall educational quality. The quality achieved under PF prompting within Claude was comparable to that of DL prompting, both of which were statistically superior to NO prompting. Furthermore, the proportion of acceptable responses generated by the combination of Claude 3.5 and PF prompting approached nearly 100%, indicating a minimal likelihood of inaccuracies in the responses. This combination also exhibited considerable comprehensibility, with the percentage of comprehensible responses nearing 100%. Consequently, for parents seeking information on KD, Claude with PF prompting emerges as the optimal choice due to its outstanding educational quality, accuracy and clarity.

While this was the first study assessing the responses of LLMs concerning KD, it had several limitations. Firstly, since a LLM is trained until a specific cut-off date, the version utilized in this analysis may have been updated subsequently, potentially leading to variations in responses in future iterations. Secondly, although the questions were meticulously devised through a multi-step process, their limited number may not fully capture the comprehensive scope of the topic. Despite the implementation of a double-review process, the assessment remained subjective, causing unavoidable confounders in our evaluation. Additionally, as the evaluations of comprehensibility were performed by pediatric professionals rather than KD patients or parents of children with KD, the results may be subject to biases and not accurately reflect a patient's perspective. Furthermore, the investigation was conducted in English, which is the predominant global language. Future research could explore the responses of LLMs in similarly widely used languages, such as Chinese. Moreover, although cautionary statements promoted responsible and prudent application, the exploration of the potential advantages and drawbacks inherent in this approach remains necessary.

In future research, broadening the scope and depth of inquiry and assessment, would provide a more comprehensive understanding of LLMs' capabilities and limitations in pediatric and KD contexts. Considering clinical applications of LLMs, particularly from the parents' perspective, KD is a challenging disease that requires long-term management. Therefore, prompt engineering could play a crucial role in guiding patients to inquiry medical questions correctly, potentially improving patient education and effectively addressing their queries. Further research is warranted to refine prompt engineering across LLMs, tailoring them to specific medical inquiries and target audiences. Furthermore, it is essential to develop specialized LLMs with advanced medical expertise to assist physicians in the field such as KD diagnosis, cardiac ultrasound interpretation, and decision-making (Pan and Jiao, 2024). As AI technology evolves, continuous evaluations and updates would be necessary to ensure that these tools remain reliable and effective in clinical application.

Conclusion

The utilization of LLMs for patient education on KD holds substantial potential as a help resource. Generally, responses generated by different LLMs meet to acceptable standards and are mostly comprehensible, albeit with variations in readability and educational quality. However, the consistency in quality remains unsatisfactory and issues of misinformation and confabulation persist. Therefore, parents and physicians should be cautious when utilizing LLMs for medical information on KD. Notable discrepancies were observed across different LLMs and prompting strategies. Claude 3.5 Sonnet demonstrated superior response quality and accuracy, whereas the Gemini 1.5 Pro excelled in delivering comprehensible responses. The efficacy of different prompts varied across the various models, with Claude 3.5 Sonnet employing PF prompting being most recommended for parents seeking information on KD. As AI technology rapidly advances, it is crucial to broaden the scope and depth of inquiry, continuously evaluate and update models, and develop specialized LLMs with advanced medical expertise to assist physicians.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

CY: Conceptualization, Methodology, Writing – original draft, Writing – review & editing. ZL: Conceptualization, Methodology, Writing – original draft, Writing – review & editing. YL: Data curation, Formal analysis, Writing – review & editing. SS: Data curation, Formal analysis, Writing – review & editing. FM: Data curation, Formal analysis, Writing – review & editing. NZ: Data curation, Formal analysis, Writing – review & editing. BL: Formal analysis, Visualization, Writing – review & editing. CW: Funding acquisition, Supervision, Writing – review & editing. KZ: Funding acquisition, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported financially by grants from National Natural Science Foundation of China (No. 82370236) and National Key Research and Development Program of China (No. 2023YFC2706402).

Conflict 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.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations,

or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/frai.2025.1571503/full#supplementary-material>

References

- Ali, R., Tang, O. Y., Connolly, I. D., Zadnik Sullivan, P. L., Shin, J. H., Fridley, J. S., et al. (2023). Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. *Neurosurgery* 93, 1353–1365. doi: 10.1227/neu.0000000000002632
- An, H. S., Kim, G. B., Song, M. K., Lee, S. Y., Kwon, H. W., Lee, J. W., et al. (2021). The occurrence of coronary artery lesions in Kawasaki disease based on C-reactive protein levels: a retrospective cohort study. *Pediatr. Rheumatol. Online J.* 19:78. doi: 10.1186/s12969-021-00566-6
- Bhattacharyya, M., Miller, V. M., Bhattacharyya, D., and Miller, L. E. (2023). High rates of fabricated and inaccurate references in ChatGPT-generated medical content. *Cureus* 15:e39238. doi: 10.7759/cureus.39238
- Cakir, H., Caglar, U., Yildiz, O., Meric, A., Ayranci, A., and Ozgor, F. (2024). Evaluating the performance of ChatGPT in answering questions related to urolithiasis. *Int. Urol. Nephrol.* 56, 17–21. doi: 10.1007/s11255-023-03773-0
- Daraz, L., Morrow, A. S., Ponce, O. J., Beuschel, B., Farah, M. H., Katabi, A., et al. (2019). Can patients trust online health information? A meta-narrative systematic review addressing the quality of health information on the internet. *J. Gen. Intern. Med.* 34, 1884–1891. doi: 10.1007/s11606-019-05109-0
- Dehdab, R., Brendlin, A., Werner, S., Almansour, H., Gassenmaier, S., Brendel, J. M., et al. (2024). Evaluating ChatGPT-4V in chest CT diagnostics: a critical image interpretation assessment. *Jpn. J. Radiol.* 42, 1168–1177. doi: 10.1007/s11604-024-01606-3
- Dyckhoff-Shen, S., Koedel, U., Brouwer, M. C., Bodilsen, J., and Klein, M. (2024). ChatGPT fails challenging the recent ESCMID brain abscess guideline. *J. Neurol.* 271, 2086–2101. doi: 10.1007/s00415-023-12168-1
- Elakabawi, K., Lin, J., Jiao, F., Guo, N., and Yuan, Z. (2020). Kawasaki disease: global burden and genetic background. *Cardiol. Res.* 11, 9–14. doi: 10.14740/cr993
- Elemento, O., Leslie, C., Lundin, J., and Tourassi, G. (2021). Artificial intelligence in cancer research, diagnosis and therapy. *Nat. Rev. Cancer* 21, 747–752. doi: 10.1038/s41568-021-00399-1
- Finney Rutten, L. J., Blake, K. D., Greenberg-Worisek, A. J., Allen, S. V., Moser, R. P., and Hesse, B. W. (2019). Online health information seeking among US adults: measuring Progress toward a healthy people 2020 objective. *Public Health Rep. (Washington, DC: 1974)* 134, 617–625. doi: 10.1177/0033354919874074
- Gondode, P. G., Khanna, P., Sharma, P., Duggal, S., and Garg, N. (2024). End-of-life care patient information leaflets-a comparative evaluation of artificial intelligence-generated content for readability, sentiment, accuracy, completeness, and suitability: ChatGPT vs Google Gemini. *Indian J. Crit. Care Med.* 28, 561–568. doi: 10.5005/jp-journals-10071-24725
- Günay, S., Öztürk, A., Özerol, H., Yiğit, Y., and Erenler, A. K. (2024). Comparison of emergency medicine specialist, cardiologist, and chat-GPT in electrocardiography assessment. *Am. J. Emerg. Med.* 80, 51–60. doi: 10.1016/j.ajem.2024.03.017
- Haug, C. J., and Drazen, J. M. (2023). Artificial intelligence and machine learning in clinical medicine, 2023. *N. Engl. J. Med.* 388, 1201–1208. doi: 10.1056/NEJMra2302038
- Hillmann, H. A. K., Angelini, E., Karfoul, N., Feickert, S., Mueller-Leisse, J., and Duncker, D. (2023). Accuracy and comprehensibility of chat-based artificial intelligence for patient information on atrial fibrillation and cardiac implantable electronic devices. *Europace* 26. doi: 10.1093/europace/euad369
- Hopkins, A. M., Logan, J. M., Kichenadasse, G., and Sorich, M. J. (2023). Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. *JNCI Cancer Spectr.* 7. doi: 10.1093/jncics/pkad010
- Iwata, H., Kobayashi, S., Itoh, M., Itoh, S., Mesfin Ketema, R., Tamura, N., et al. (2024). The association between prenatal per- and polyfluoroalkyl substance levels and Kawasaki disease among children of up to 4 years of age: a prospective birth cohort of the Japan environment and Children's study. *Environ. Int.* 183:108321. doi: 10.1016/j.envint.2023.108321
- Jindal, P., and MacDermid, J. C. (2017). Assessing reading levels of health information: uses and limitations of Flesch formula. *Educ. Health (Abingdon)* 30, 84–88. doi: 10.4103/1357-6283.210517
- Kato, H., Koike, S., Yamamoto, M., Ito, Y., and Yano, E. (1975). Coronary aneurysms in infants and young children with acute febrile mucocutaneous lymph node syndrome. *J. Pediatr.* 86, 892–898. doi: 10.1016/S0022-3476(75)80220-4
- Kato, H., Sugimura, T., Akagi, T., Sato, N., Hashino, K., Maeno, Y., et al. (1996). Long-term consequences of Kawasaki disease. A 10- to 21-year follow-up study of 594 patients. *Circulation* 94, 1379–1385. doi: 10.1161/01.CIR.94.6.1379
- Kiyak, Y. S., and Emekli, E. (2024). ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: a literature review. *Postgrad. Med. J.* 100, 858–865. doi: 10.1093/postmj/qgae065
- Koo, T. K., and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* 15, 155–163. doi: 10.1016/j.jcjm.2016.02.012
- Lechien, J. R., Naunheim, M. R., Maniaci, A., Radulesco, T., Saibene, A. M., Chiesa-Estomba, C. M., et al. (2024). Performance and consistency of ChatGPT-4 versus otolaryngologists: a clinical case series. *Otolaryngology* 170, 1519–1526. doi: 10.1002/ohn.759
- Liu, X., Duan, C., Kim, M. K., Zhang, L., Jee, E., Maharjan, B., et al. (2024). Claude 3 opus and ChatGPT with GPT-4 in dermoscopic image analysis for melanoma diagnosis: comparative performance analysis. *JMIR Med. Inform.* 12:e59273. doi: 10.2196/59273
- Massey, P. A., Montgomery, C., and Zhang, A. S. (2023). Comparison of ChatGPT-3.5, ChatGPT-4, and orthopaedic resident performance on orthopaedic assessment examinations. *J. Am. Acad. Orthop. Surg.* 31, 1173–1179. doi: 10.5435/JAAOS-D-23-00396
- McCrindle, B. W., Rowley, A. H., Newburger, J. W., Burns, J. C., Bolger, A. F., Gewitz, M., et al. (2017). Diagnosis, treatment, and long-term Management of Kawasaki Disease: a scientific statement for health professionals from the American Heart Association. *Circulation* 135, e927–e999. doi: 10.1161/CIR.0000000000000484
- Meyer, M. K. R., Kandathil, C. K., Davis, S. J., Durairaj, K. K., Patel, P. N., Pepper, J. P., et al. (2024). Evaluation of rhinoplasty information from ChatGPT, Gemini, and Claude for readability and accuracy. *Aesth. Plast. Surg.* doi: 10.1007/s00266-024-04343-0
- Nakamura, Y., Yashiro, M., Uehara, R., Sadakane, A., Tsuboi, S., Aoyama, Y., et al. (2012). Epidemiologic features of Kawasaki disease in Japan: results of the 2009–2010 nationwide survey. *J. Epidemiol.* 22, 216–221. doi: 10.2188/jea.JE20110126
- Newburger, J. W., Takahashi, M., and Burns, J. C. (2016). Kawasaki Disease. *J. Am. Coll. Cardiol.* 67, 1738–1749. doi: 10.1016/j.jacc.2015.12.073
- Onder, C. E., Koc, G., Gokbulut, P., Taskaldiran, I., and Kuskonmaz, S. M. (2024). Evaluation of the reliability and readability of ChatGPT-4 responses regarding hypothyroidism during pregnancy. *Sci. Rep.* 14:243. doi: 10.1038/s41598-023-50884-w
- Ozgor, F., Caglar, U., Halis, A., Cakir, H., Aksu, U. C., Ayranci, A., et al. (2024). Urological cancers and ChatGPT: assessing the quality of information and possible risks for patients. *Clin. Genitourin. Cancer* 22, 454–7.e4. doi: 10.1016/j.clgc.2023.12.017
- Pan, Y., and Jiao, F. Y. (2024). Application of artificial intelligence in the diagnosis and treatment of Kawasaki disease. *World J. Clin. Cases* 12, 5304–5307. doi: 10.12998/wjcc.v12.i23.5304
- Popovac, M., and Rومانey, R. (2022). Measuring online health-seeking behaviour: construction and initial validation of a new scale. *Br. J. Health Psychol.* 27, 756–776. doi: 10.1111/bjhp.12571
- Rajpurkar, P., Chen, E., Banerjee, O., and Topol, E. J. (2022). AI in health and medicine. *Nat. Med.* 28, 31–38. doi: 10.1038/s41591-021-01614-0
- Schmidl, B., Hütten, T., Pigorsch, S., Stögbauer, F., Hoch, C. C., Hussain, T., et al. (2024). Assessing the use of the novel tool Claude 3 in comparison to ChatGPT 4.0 as an artificial intelligence tool in the diagnosis and therapy of primary head and neck cancer cases. *Eur. Arch. Otorhinolaryngol.* 281, 6099–6109. doi: 10.1007/s00405-024-08828-1
- Sufi, F. (2024). Generative pre-trained transformer (GPT) in research: a systematic review on data augmentation. *Information* 15:99. doi: 10.3390/info15020099

Tepe, M., and Emekli, E. (2024). Assessing the responses of large language models (ChatGPT-4, Gemini, and Microsoft copilot) to frequently asked questions in breast imaging: a study on readability and accuracy. *Cureus* 16:e59960. doi: 10.7759/cureus.59960

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* 25, 44–56. doi: 10.1038/s41591-018-0300-7

Walters, W. H., and Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Sci. Rep.* 13:14045. doi: 10.1038/s41598-023-41032-5

Wang, L., Chen, X., Deng, X., Wen, H., You, M. K., Liu, W. Z., et al. (2024). Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *NPJ Digit. Med.* 7:41. doi: 10.1038/s41746-024-01029-4

Wang, Y., Chen, Y., and Sheng, J. (2024). Assessing ChatGPT as a medical consultation assistant for chronic hepatitis B: cross-language study of English and Chinese. *JMIR Med. Inform.* 12:e56426. doi: 10.2196/56426

Wang, Y., Liang, L., Li, R., Wang, Y., and Hao, C. (2024). Comparison of the performance of ChatGPT, Claude and Bard in support of myopia prevention and control. *J. Multidiscip. Healthc.* 17, 3917–3929. doi: 10.2147/JMDH.S473680

Wang, C. L., Wu, Y. T., Liu, C. A., Kuo, H. C., and Yang, K. D. (2005). Kawasaki disease: infection, immunity and genetics. *Pediatr. Infect. Dis. J.* 24, 998–1004. doi: 10.1097/01.inf.0000183786.70519.fa

Yang, J., Ardavanis, K. S., Slack, K. E., Fernando, N. D., Della Valle, C. J., and Hernandez, N. M. (2024). Chat generative Pretrained transformer (ChatGPT) and bard: artificial intelligence does not yet provide clinically supported answers for hip and knee osteoarthritis. *J. Arthroplast.* 39, 1184–1190. doi: 10.1016/j.arth.2024.01.029