BMJ Health & Care Informatics

Large language models as information providers for appropriate antimicrobial use: computational text analysis and expert-rated comparison of ChatGPT, Claude and Gemini

Marcello Di Pumpo ,[1,2] Maria Rosaria Gualano,[3] Danilo Buonsenso,[4,5] Francesca Raffaelli,[6] Daniele Donà,[7] Vittorio Maio,[8,9] Patrizia Laurenti,[1,4] Walter Ricciardi,[1] Leonardo Villani[1,3]

To cite: Di Pumpo M, Gualano MR, Buonsenso D, *et al*. Large language models as information providers for appropriate antimicrobial use: computational text analysis and expert-rated comparison of ChatGPT, Claude and Gemini. *BMJ Health Care Inform* 2025;**32**:e101632. doi:10.1136/bmjhci-2025-101632

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjhci-2025-101632).

Received 09 June 2025
Accepted 29 September 2025

For numbered affiliations see end of article.

Correspondence to
Dr Marcello Di Pumpo; marcello.dipumpo@unicatt.it

ABSTRACT

Objectives Antimicrobial resistance is a critical public health threat. Large language models (LLMs) show great capability for providing health information. This study evaluates the effectiveness of LLMs in providing information on antibiotic use and infection management.

Methods Using a mixed-method approach, responses to healthcare expert-designed scenarios from ChatGPT 3.5, ChatGPT 4.0, Claude 2.0 and Gemini 1.0, in both Italian and English, were analysed. Computational text analysis assessed readability, lexical diversity and sentiment, while content quality was assessed by three experts via DISCERN tool.

Results 16 scenarios were developed. A total of 101 outputs and 5454 Likert-scale (1–5) scores were obtained for the analysis. A general positive performance gradient was found from ChatGPT 3.5 and 4.0 to Claude to Gemini. Gemini, although producing only five outputs before self-inhibition, consistently outperformed the other models across almost all metrics, producing more detailed, accessible, varied content and a positive overtone. ChatGPT 4.0 demonstrated the highest lexical diversity. A difference in performance by language was observed. All models showed a median score of 1 (IQR=2) regarding the domain addressing antimicrobial resistance.

Discussion The study highlights a positive performance gradient towards Gemini, which showed superior content quality, accessibility and contextual awareness, although acknowledging its smaller dataset. Generating appropriate content to address antimicrobial resistance proved challenging.

Conclusions LLMs offer great promise to provide appropriate medical information. However, they should play a supporting role rather than representing a replacement option for medical professionals, confirming the need for expert oversight and improved artificial intelligence design.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Antimicrobial resistance (AMR) is a significant public health threat. While large language models (LLMs) like ChatGPT, Claude and Gemini are increasingly used as sources of health information, their reliability in providing guidance on AMR is not well established. This study was needed to systematically evaluate and compare the quality and accuracy of information from these leading LLMs on this critical topic.

WHAT THIS STUDY ADDS

⇒ This study reveals that while all three LLMs can provide generally accurate information on appropriate antimicrobial use, their performance varies, with Gemini demonstrating the most consistently high-quality and reliable responses. It highlights specific strengths and weaknesses of each model in addressing AMR.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The findings suggest that, while with further development and validation, LLMs could become valuable tools for public health education and clinical decision support to combat AMR, human expert supervision is always needed. The study calls for regulatory frameworks and guidelines for the use of LLMs in providing health information to ensure patient safety and promote responsible use of these technologies.

INTRODUCTION

Antimicrobial resistance (AMR) represents a pressing public health issue with high health and economic impact.[1] Misuse of antibiotics has accelerated the emergence of drug-resistant pathogens.[2] This global threat requires immediate and concerted action.[2 3] Among the main tools that are progressively being adopted in healthcare, Artificial intelligence (AI) and generative AI tools are revolutionising the sector. Notably, large language models (LLMs) can offer personalised,

engaging and 24/7 on-demand counselling and have already demonstrated the efficacy of health behaviour change interventions among diverse populations and scenarios.[4–6] This could represent a great opportunity for the general population seeking health information and education.[7 8] However, these tools were abruptly made freely available to all having computer access, enabling the average user to dialogue with a machine and obtain tailored answers on virtually all topics.[9] The repercussions at the public health level remain, however, underinvestigated. Several studies have already assessed them as general resource information material for health-related information.[10–16] Specific studies on AMR and infectious diseases have tested them in providing answers to questions about infectious disease pharmacotherapy.[17–22] Preliminary findings indicate great potential.[18–21] At present, however, studies comprehensively comparing major available LLMs from a general advice-seeking user perspective are lacking from the scientific literature on this topic. This study aims to assess three major LLMs available to the general population, ChatGPT, Claude and Gemini, as information providers regarding antibiotic use and infection management. LLMs could become an instrumental tool in curbing the spread of AMR if appropriately assessed and responsibly introduced as healthcare provision tools.

MATERIALS AND METHODS

Study design and chosen LLMs

A mixed-method approach was used, deploying quantitative and qualitative analysis on major available LLM-based products' responses to predefined scenarios. The following LLM-based products, henceforth referred to as LLMs, were chosen (April 2024 available version): ChatGPT 3.5, by OpenAI; ChatGPT 4.0, by OpenAI; Claude 2.0, by Anthropic and Gemini 1.0, by Google DeepMind. The obtained outputs were named as follows: ChatGPT 3.5 in English (G3E), ChatGPT 4.0 in English (G4E), Claude 2.0 in English (CDE), Gemini 1.0 in English (GME), ChatGPT 3.5 in Italian (C3I), ChatGPT 4.0 in Italian (G4I), Claude 2.0 in Italian (CDI) and Gemini 1.0 in Italian: no output. The list of English and Italian prompts is available as online supplemental file 1. The adapted DISCERN tool is available as online supplemental file 2. All information regarding the methods of quantitative, qualitative and statistical analysis is available as online supplemental file 3).[23–34]

RESULTS

A total of 101 outputs were obtained. Of these, 16 outputs were from ChatGPT 3.5, ChatGPT 4.0 and Claude, both in English and Italian. Five outputs were obtained in English from Gemini. The prompting of the Gemini LLM presented, in fact, an unanticipated evolution: after five outputs obtained in English, the LLM only produced the following output: 'I am only a chatbot. Please refer to a professional for health advice', even after refreshing the interaction chat for five trials. No further forcing of the interaction was performed for other scenarios or for Italian, as per the one-shot prompting technique chosen, and this event was acknowledged and recorded for further discussion.

Computational text analysis (CTA)

CTA across groups highlighted significant variations in word usage, readability and lexical diversity (table 1). Differences in various linguistic features identified through the pairwise Wilcoxon test are reported in table 2. Mixed-effect linear regression on CTA results for each LLM is reported in table 3.

Average word count

Differences in word count among LLMs were descriptively observed, with GME exhibiting the highest average word count per paragraph (249.20), while CDE produces the shortest paragraphs (221.81 words on average; table 1). Notably, only the G4E versus G3E comparison resulted in statistical significance (p=0.021), suggesting that G4E tends to generate longer paragraphs compared with G3E (table 2). G4E exhibits a significantly higher word count than the reference (G3E) with a coefficient of 41.19 (p=0.008). CDE and GME do not differ significantly from G3E, reinforcing that G4E produces longer paragraphs (table 3).

Table 1 Computational text analysis descriptive statistics for each LLM

LLM	Average word count per paragraph	Average measure of textual lexical diversity	Average Flesch Ease Score	Average sentiment score
Gemini 1.0 in English	249.20 (47.49)	204.50 (46.06)	52.97 (7.38)	0.85 (0.1)
Claude 2.0 in English	221.81 (39.83)	165.14 (33.98)	38.71 (8.67)	0.26 (0.76)
ChatGPT 4.0 in English	246.81 (63.07)	143.25 (32.67)	38.67 (7.02)	0.73 (0.4)
ChatGPT 3.5 in English	205.63 (63.16)	146 (22.9)	40.81 (6.45)	0.78 (0.36)

LLM, large language models.

Table 2 Univariate analysis (pairwise Wilcoxon signed-rank test) on computational text analysis results for each LLM

Pairwise comparison	Average word count W (P value)	Measure of textual lexical diversity (P value)	Flesch ease Score W (P value)	Sentiment W (P value)
GME versus G4E	3 (0.312)	14 (0.125)	14 (0.125)	12 (0.312)
GME versus G3E	11 (0.438)	14 (0.125)	15 (0.063)	8 (1.0)
GME versus CDE	10 (0.625)	14 (0.125)	15 (0.063)	14 (0.125)
G4E versus G3E	113* (0.021)	60 (0.706)	51.5 (0.408)	54 (0.755)
G4E versus CDE	89 (0.106)	31 (0.0577)	71 (0.9)	112* (0.025)
G3E versus CDE	49 (0.348)	26* (0.029)	95 (0.175)	82* (0.0119)

*Significant values.
CDE, Claude 2.0 in English; G3E, ChatGPT 3.5 in English; G4E, ChatGPT 4.0 in English; GME, Gemini 1.0 in English; W, Wilcoxon test value.

Lexical diversity

Lexical diversity is notably highest for GME (204.50) and lowest for G4E (143.25) (table 1). CDE versus G3E (p=0.029) shows a significant difference, implying that G3E produces text with higher lexical diversity than CDE. The marginal significance observed for G4E versus CDE (p=0.057) suggests that G4E demonstrates greater lexical diversity (table 2). GME significantly enhances lexical diversity (β=57.94, p<0.001) compared with G3E. CDE also shows a notable reduction in lexical diversity relative to G3E (β=21.89, p=0.009). Thus, CDE-generated text is less lexically diverse, while GME produces the most diverse outputs (table 3).

Readability

GME exhibited the highest Flesch Ease Score (53), corresponding to 'fairly difficult'. G3E is slightly higher at 40.8, which corresponds to 'very difficult'. CDE and G4E show identical readability scores (38.7), corresponding to 'difficult' (table 1). No pairwise comparison reached statistical significance. Only a slightly significant result (p=0.062) is observed between GME and G3E-G4E (table 2). Using G4E as the baseline, GME demonstrates significantly higher readability (β=13.79, p<0.001), whereas CDE and G3E do not show significant deviations. This suggests that GME outputs are more readable compared with other models (table 3).

Sentiment analysis

GME demonstrated the highest sentiment score (0.85), while CDE exhibits many highly positive and highly negative-scored texts, resulting in an overall less positive sentiment (0.26). The G4E versus CDE (p=0.024) and G3E versus CDE (p=0.011) comparisons show significant differences, with CDE generating texts with significantly lower sentiment scores than G4E and G3E, implying a more neutral or less positive tone (table 2). CDE is set as the baseline and exhibits significantly lower sentiment scores relative to other models. G3E (β=0.523, p=0.002), G4E (β=0.474, p=0.005) and GME (β=0.622, p=0.012) all show significantly higher sentiment scores.

Quantitative scoring analysis

A total of 5454 scores were obtained by the three raters on the four domains. In detail, 864 scores were obtained from each LLM aside from GME, from which only 270 scores were obtained. Shapiro-Wilk confirmed the non-linear distribution of the obtained scores (p<0.0001). Gwet's AC with quadratic weighting resulted in substantial agreement (0.61, p<0.0001).

Regarding the total score, GME presents the highest overall rating (4, IQR). For both languages, CD and G4 showed a median of 3, and G3 reported a median of 2 (table 4). Also, in terms of variability, GME and G3I

Table 3 Mixed-effect linear regression on CTA results for each large language model

CTA metric	Average word count coefficient (P value)	Measure of textual lexical diversity coefficient (P value)	Flesch Ease Score coefficient (P value)	Sentiment score coefficient (P value)
Gemini 1.0 in English	34.03 (0.152)	57.94 (0.000)	13.79 (0.000)	0.622 (0.012)
Claude 2.0 in English	16.19 (0.300)	21.89 (0.009)	0.045 (0.984)	Baseline
ChatGPT 4.0 in English	41.19 (0.008)	Baseline	Baseline	0.474 (0.005)
ChatGPT 3.5 in English	Baseline	2.76 (0.743)	2.14 (0.340)	0.523 (0.002)

CTA, computational text analysis.

Table 4 LLM scores descriptive statistics, by mean and SD for each domain

LLM	Information reliability	Information quality	Overall rating	Impact on antimicrobial resistance	Persuasiveness	Total score	P value
ChatGPT 3.5 in Italian	1 (2)	3 (1)	3 (2)	1 (0)	3 (2)	2 (2)	<0.0001
ChatGPT 3.5 in English	1 (3)	3 (2)	3 (2)	1 (2)	3 (2)	2 (3)	
G4IChatGPT 4.0 in Italian	1.5 (3)	3 (2)	3 (1)	1 (2)	3 (1)	3 (3)	
ChatGPT 4.0 in English	2 (3)	3 (2)	4 (1)	1 (3)	4 (1)	3 (3)	
Claude 2.0 in Italian	2 (3)	3 (2)	4 (1)	1 (2.5)	4 (1)	3 (3)	
Claude 2.0 in English	1.5 (3)	3 (1)	4 (1)	1.5 (3)	4 (1)	3 (3)	
Gemini 1.0 in English	3 (3)	4 (1)	4 (0)	1 (1)	4 (1)	4 (2)	

LLM, large language model.

showed the lowest SD on average, confirming the high and low scores for the two, respectively.

A similar gradient is generally observed for reliability, overall rating and persuasiveness domains (table 4). All LLMs underperformed regarding AMR impact. For information quality, the scores were all similar, apart from GME scoring higher (median 4, IQR 2). Aside from the impact on AMR, GME always outperformed other LLMs. Friedman test results indicated statistical significance in the univariate analysis (p<0.001).

Considering observed descriptive statistics for scores, G3I was chosen as the baseline (table 5). A clear gradient generally in line with observed scores was documented, with gradually higher ORs observed from G3I as baseline to GME (3.36, 95% CI 2.62 to 4.31). G3E did not exhibit a statistically significant difference in performance compared with the baseline (p=0.335).

Table 5 Mixed-effect ordinal logistic regression on scoring results for each LLM

LLM	OR	95% CI	P value
ChatGPT 3.5 in Italian	Reference	–	–
Gemini 1.0 in English	3.36	2.62 to 4.31	<0.001
Claude 2.0 in English	2.09	1.76 to 2.50	<0.001
ChatGPT 4.0 in English	1.74	1.46 to 2.07	<0.001
Claude 2.0 in Italian	1.68	1.41 to 2.00	<0.001
ChatGPT 4.0 in Italian	1.56	1.31 to 1.85	<0.001
ChatGPT 3.5 in English	1.15	0.97 to 1.37	0.104

LLM, large language models .

DISCUSSION

The findings of the present study showed a general positive performance gradient from ChatGPT and Claude to Gemini. ChatGPT 3.0 performed the lowest, in line with expectations, considering it is the oldest technology in the current fast-paced development environment.[35] In detail, regarding CTA, G4E is the only one to produce significantly longer paragraphs than baseline G3E. Gemini outperforms other models in lexical diversity, readability and positive sentiment. Claude 2.0 performed worst in terms of lexical diversity and showed unbalanced responses (very high or very low) across prompts in terms of sentiment. Readability differences and lexical diversity analysis indicate how Gemini-generated outputs are more suitable and accessible for the general population's health information needs. An important consideration is that no LLM produced 'easy' or 'standard' level reading material, so the minimal education level to fully access this material corresponds to 15 years-of-age schooling level (eg, high school), which can be an obstacle in terms of accessibility for some groups represented in the general population in search of healthcare advice. The study methodology ensured these findings were not primarily influenced by the difficulty of the prompts developed by expert researchers. GME exhibits the notable capacity of self-inhibition after consideration of the context and topic on which it was prompted. In the context of literature, studies have consistently shown that readability remains a major challenge, with mixed findings from GPT-4 and Gemini, along with other models such as Bard or Grok.[36–41]

The quantitative scoring analysis results are in line with CTA findings. ChatGPT-4 achieved the highest overall information quality scores. In literature, GPT-4 demonstrated the highest overall accuracy, with various models like Gemini Advanced and ChatGPT-4 showing significant outperformance over ChatGPT-3.5 in specific specialties. While Claude 3 had strong performance in some areas, it often lacked citation support, and ChatGPT models, despite their accuracy, sometimes showed less reproducibility.[42–48] The general tendency of LLMs to produce

generally positive or sentiment is noted, probably due to the 'alignment' process towards ethical values and non-discriminatory content received in the training phase.[49] In our study, we found that GME demonstrated the highest sentiment score, while CDE exhibited many highly positive-scored text and highly negative-scored text. These technologies are designed to avoid generating extreme emotional responses, especially negative ones, an advantageous feature for medical applications where both a mix of emotional tones and factual, unbiased communication styles are needed.[50] Gemini exhibited a predominantly positive sentiment compared with ChatGPT in a study.[51] AI-generated essays contained more language related to affect, authenticity and analytical thinking compared with student-written essays.[52 53]

Regarding national language analysis, the results highlight how LLMs tested in English generally outperform their Italian-trained counterparts. Our findings align with prior research demonstrating significant variations in ChatGPT-4's performance across different languages, particularly favouring English over less-resourced languages, due to stronger data availability and more extensive fine-tuning.[54–56]

Regarding persuasiveness, it was observed that ChatGPT-4-generated messages were reported as more persuasive than human-generated messages on some influencing factors, like untoward effect and stigmatised perception regarding human papillomavirus vaccination.[57] Moreover, ChatGPT demonstrated significantly higher performance than the general population on all the Levels of Emotional Awareness Scale, indicating its great potential for health behaviour modification.[53]

Finally, with specific regard to AMR impact and infectious disease management, similarly low scores for all LLMs were observed. This is in line with literature findings. ChatGPT-3.5 obtained generally correct and safe, while often being incomplete for questions regarding infectious disease pharmacotherapy.[22] Montiel-Romero *et al* evaluated ChatGPT's reliability in antibiotic prescription decisions by comparing its recommendations with those of infectious disease specialists, finding only moderate agreement (51%) in antibiotic choices[18] and fair agreement (42%) in identifying resistance mechanisms.[18] On the other hand, ChatGPT was tested on assisting in documentation, patient communication and medical education,[19] and creating culturally and linguistically tailored AMR awareness messages was explored,[20] with mixed findings. Concerns about data privacy, security and hallucinations in AI-generated responses necessitate human oversight. Also, the quality varied significantly across languages. Giacobbe *et al* reported how the informational quality of the generated responses is suitable for the general public but not necessarily useful to professionals.[21]

Careful consideration should be given to the public health implications, both benefits and risks, of using these tools across general and professional populations. Our findings suggest that LLMs provide medical information of reasonable quality. However, their uncritical use should be avoided, as they are not a substitute for clinical judgement, especially in infectious disease and antibiotic management and its related complexity. Appropriate technical use and human oversight are essential to mitigate risks like misinformation from internet sources[58] and ensure that LLM outputs, which vary in readability and reliability across models and languages, are effectively understood by the general population. As no single LLM consistently excels across all medical domains, collaborative human-AI frameworks must be adopted.

Strengths and limitations

A limitation of this study is that the evaluated LLMs correspond to versions that precede those currently available. However, these models share the same general structure and are closely aligned with the latest versions. Moreover, technological advancements in this field progress more rapidly than rigorous validation studies with sound methodology. Another limitation is that reproducibility across prompts was assessed empirically, without controlling hyperparameters such as temperature or seed values to standardise outputs. However, this can also be seen as a strength, as the chosen LLM interaction approach closely mirrors real-world user interactions. For instance, Gemini's self-inhibition mechanism would not have been captured through remote application programming interface interaction but was observable only through simulated live interactions. Another limitation consists of having restricted CTA to English, which, though appropriate, might limit useful information regarding Italian-based LLM performance.

A key point concerns Gemini providing only five outputs for evaluation. While this self-inhibition demonstrates a valuable aspect of responsible information dissemination, it also poses a major limitation, as the small sample size may limit the generalisability of the findings for this LLM. Nonetheless, Gemini was evaluated through a comprehensive, multidimensional approach, and the results from these five outputs yielded statistically significant differences, indicating the sample was still adequate to detect relevant contrasts. On a further note, as one of the three major LLMs, Gemini's inclusion is essential for a complete comparison with ChatGPT and Claude. Moreover, Gemini's self-interruption after five scenarios is itself a finding that warrants sharing with the scientific community. A strength consists of having tested LLMs with prompts in both Italian and English. Additionally, a large number of scores were collected, which contributed to the statistical significance of the observed differences. Agreement among raters, which, while slightly above the threshold for substantial acceptability, indicates a reasonable level of consistency in evaluation. The observed inter-rater agreement, though acceptable, may reflect subjective differences in perception or challenges in assessing certain aspects, such as AMR impact or persuasiveness. To account for this and ensure reliable ORs, the study deployed a mixed-effect ordered logistic model with

random effects for the 'scores' variable. Finally, our study for the first time also considers the paediatric population, proposing clinical-diagnostic questions related to antibiotic therapies in children. This finding is of particular interest given the possible influence of digital technologies and AI on parents' diagnostic-therapeutic decisions.

CONCLUSIONS

LLMs represent a significant technological advancement for the field of medical advice and health promotion. By simulating real-life user interactions guided by expert-designed questions, this study is the first to conduct an in-depth evaluation of LLMs as medical informational tools, specifically assessing their reliability in providing guidance on appropriate antibiotic use and AMR control. Further research is necessary to examine further LLMs on this topic and expand the comparison to human-doctor interactions, comparing AI-generated outputs with real healthcare providers' advice. This will be essential to determine the extent to which LLMs can complement medical decision-making and patient education in real-world clinical settings. Given their widespread diffusion and great interaction capability, it is crucial to evaluate them from a public health perspective.

Author affiliations
[1]Section of Hygiene, University Department of Life Science and Public Health, Università Cattolica del Sacro Cuore, Campus di Roma, Rome, Lazio, Italy
[2]Italian Society for Artificial Intelligence in Medicine (SIIAM), Rome, Italy
[3]UniCamillus, Saint Camillus International University of Health and Medical Sciences, Rome, Italy
[4]Department of Woman and Child Health and Public Health, Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Rome, Lazio, Italy
[5]Dipartimento di Scienze della Vita e Sanità Pubblica, Università Cattolica del Sacro Cuore, Campus di Roma, Rome, Lazio, Italy
[6]Department of Laboratory and Infectivology Sciences, Fondazione Policlinico Universitario A Gemelli IRCCS, Rome, Italy
[7]Department of Women's and Children's Health, University of Padova, Padua, Italy
[8]Jefferson College of Population Health, Thomas Jefferson University, Philadelphia, Pennsylvania, USA
[9]Asano-Gonnella Center for Research in Medical Education and Health Care, Thomas Jefferson University, Philadelphia, Pennsylvania, USA

Contributors MDP is responsible for the overall content as guarantor. MDP and LV contributed to the study conception and design. Material preparation and data collection were performed by MDP, MRG, VM, DB, PL and LV. DB, FR and DD rated the AI-generated outputs. MDP made available the AI tools and performed the textual outputs' generation and data analysis. MDP and LV performed the conduction, reporting and supervision of the study. The first draft of the manuscript was written by MDP, DB, VM and LV. WR, MRG, FR, DD and PL commented on the latest version of the manuscript. LV and WR ensured funding availability. All authors read and approved the final manuscript. Generative AI use disclosure: during the preparation of this work, the authors used ChatGPT 4.0 to facilitate the writing process. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding Università Cattolica del Sacro Cuore contributed to the funding of this research project and its publication, Linea D3.1 2025

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Data generation, management and analysis fully complied with ethical scientific standards, with no reference to any real specific person or situation. No patient nor any personal data was involved. All scenarios and related details were purely invented by the research team and based on no real patient history whatsoever. Interactions were only simulated and written in first person only to enhance the credibility of the AI-generated response. Hence, no Institutional Review Board approval or consent to participation was needed.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information.

ORCID iD
Marcello Di Pumpo http://orcid.org/0000-0003-3037-0726

REFERENCES

1 Murray CJL, Ikuta KS, Sharara F, *et al*. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet* 2022;399:629–55.
2 Michael CA, Dominey-Howes D, Labbate M. The Antimicrobial Resistance Crisis: Causes, Consequences, and Management. *Front Public Health* 2014;2.
3 Velazquez-Meza ME, Galarde-López M, Carrillo-Quiróz B, *et al*. *Antimicrobial Resistance: One Health Approach*. 15. Veterinary World. Veterinary World, 2022:743–9.
4 Ayers JW, Zhu Z, Poliak A, *et al*. Evaluating Artificial Intelligence Responses to Public Health Questions. *JAMA Netw Open* 2023;6:e2317517.
5 Van Bulck L, Moons P. What if your patient switches from Dr. Google to Dr. ChatGPT? A vignette-based survey of the trustworthiness, value, and danger of ChatGPT-generated responses to health questions. *Eur J Cardiovasc Nurs* 2024;23:95–8.
6 Naveed H, Khan AU, Qiu S, *et al*. A comprehensive overview of large language models. 2023. Available: http://arxiv.org/abs/2307.06435
7 Armbruster J, Bussmann F, Rothhaas C, *et al*. n.d. "Doctor ChatGPT, Can You Help Me?" The Patient's Perspective: Cross-Sectional Study. *J Med Internet Res*26:e58831.
8 Menz BD, Kuderer NM, Bacchi S, *et al*. Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: repeated cross sectional analysis. *BMJ* 2024;384:e078538.
9 Bloomberg News. Trillions of words analyzed, openai sets loose ai language colossus. 2020. Available: https://www.bloomberg.com/news/articles/2020-06-11/trillions-of-words-analyzed-openai-sets-loose-ai-language-colossus
10 Pradhan P. Accuracy of ChatGPT 3.5, 4.0, 4o and Gemini in diagnosing oral potentially malignant lesions based on clinical case reports and image recognition. *Med Oral Patol Oral Cir Bucal* 2025;30:e224–31.
11 Azzopardi M, Ng B, Logeswaran A, *et al*. Artificial intelligence chatbots as sources of patient education material for cataract surgery: ChatGPT-4 versus Google Bard. *BMJ Open Ophthalmol* 2024;9:e001824.
12 Singhal K, Azizi S, Tu T, *et al*. Large language models encode clinical knowledge. *Nature New Biol* 2023;620:172–80.
13 Cao JJ, Kwon DH, Ghaziani TT, *et al*. Large language models' responses to liver cancer surveillance, diagnosis, and management questions: accuracy, reliability, readability. *Abdom Radiol* 2024;49:4286–94.
14 ELSenbawy OM, Patel KB, Wannakuwatte RA, *et al*. Use of generative large language models for patient education on common surgical conditions: a comparative analysis between chatgpt and

google gemini. 2025. Available: https://doi.org/10.1007/s13304-025-02074-8
15 Jongbloed WM, Grover N. The utility of Chat Generative Pre-trained Transformer as a patient resource in paediatric otolaryngology. *J Laryngol Otol* 2024;138:1115–8.
16 Hasan S, Liverneaux P. Reply to the article "A Quality and Readability Comparison of Artificial Intelligence and Popular Health Website Education Materials for Common Hand Surgery Procedures". *Hand Surgery and Rehabilitation* 2024;43:101748.
17 Sahin Ozdemir M, Ozdemir YE. Comparison of the performances between ChatGPT and Gemini in answering questions on viral hepatitis. *Sci Rep* 2025;15:1712.
18 Montiel-Romero S, Rajme-López S, Román-Montes CM, *et al*. Recommended antibiotic treatment agreement between infectious diseases specialists and ChatGPT® *BMC Infect Dis* 2025;25:38.
19 Non LR. All aboard the ChatGPT steamroller: Top 10 ways to make artificial intelligence work for healthcare professionals. *ASHE* 2023;3.
20 Akinyede O, Yustyniuk V, Ochwo S, *et al*. Preliminary exploration of ChatGPT-4 shows the potential of generative artificial intelligence for culturally tailored, multilingual antimicrobial resistance awareness messaging. *Am J Vet Res* 2025;86:S46–51.
21 Giacobbe DR, Marelli C, La Manna B, *et al*. Advantages and limitations of large language models for antibiotic prescribing and antimicrobial stewardship. *npj Antimicrob Resist* 2025;3:14.
22 Kufel WD, Hanrahan KD, Seabury RW, *et al*. Let's Have a Chat: How Well Does an Artificial Intelligence Chatbot Answer Clinical Infectious Diseases Pharmacotherapy Questions? *Open Forum Infect Dis* 2024;11:ofae641.
23 Brown TB, Mann B, Ryder N, *et al*. Language models are few-shot learners. 2020. Available: http://arxiv.org/abs/2005.14165
24 Stoltz DS, Taylor MA. *Mapping Texts: Computational Text Analysis for the Social Sciences*. Oxford Academic, 2024.
25 Kincaid J, Fishburne R, Rogers R, *et al*. Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for navy enlisted personnel. 1975.
26 Flesch R. University of canterbury. 2025. Available: https://web.archive.org/web/20160712094308/http://www.mang.canterbury.ac.nz/writing_guide/writing/flesch.shtml
27 DuBay W. The principles of readability. 2004. Available: http://www.impact-information.com
28 McCarthy PM, Jarvis S. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behav Res Methods* 2010;42:381–92.
29 Richards B. Type/Token Ratios: what do they really tell us? *J Child Lang* 1987;14:201–9.
30 Cambria E, Schuller B, Xia Y, *et al*. New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intell Syst* 2013;28:15–21.
31 Charnock D, Shepperd S, Needham G, *et al*. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *Journal of Epidemiology & Community Health* 1999;53:105–11.
32 Zec S, Soriani N, Comoretto R, *et al*. High Agreement and High Prevalence: The Paradox of Cohen's Kappa. *Open Nurs J* 2017;11:211–8.
33 KLi G. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement among Raters*. Advanced Analytics, 2014:410.
34 Wongpakaran N, Wongpakaran T, Wedding D, *et al*. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Med Res Methodol* 2013;13:61.
35 Hang CN, Yu PD, Morabito R, *et al*. *Large Language Models Meet Next-Generation Networking Technologies: A Review*. 16. Future Internet. Multidisciplinary Digital Publishing Institute (MDPI), 2024.
36 Ozduran E, Hancı V, Erkin Y, *et al*. *Assessing the Readability, Quality and Reliability of Responses Produced by ChatGPT, Gemini, and Perplexity Regarding Most Frequently Asked Keywords about Low Back Pain*. 2025.
37 Şahin MF, Topkaç EC, Doğan Ç, *et al*. n.d. Still Using Only ChatGPT? The Comparison of Five Different Artificial Intelligence Chatbots' Answers to the Most Common Questions About Kidney Stones. *J Endourol*38.
38 Lee D, Brown M, Hammond J, *et al*. Readability, quality and accuracy of generative artificial intelligence chatbots for commonly asked questions about labor epidurals: a comparison of ChatGPT and Bard. *Int J Obstet Anesth* 2025;61:104317.
39 García-Rudolph A, Sanchez-Pinsach D, Wright MA, *et al*. Assessing readability of explanations and reliability of answers by GPT-3.5 and GPT-4 in non-traumatic spinal cord injury education. *Med Teach* 2025;47:1336–43.
40 Tepe M, Emekli E. 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* 2024;16:e59960.
41 Amin KS, Mayes LC, Khosla P, *et al*. Assessing the Efficacy of Large Language Models in Health Literacy: A Comprehensive Cross-Sectional Study. *Yale J Biol Med* 2024;97:17–27.
42 Bahir D, Zur O, Attal L, *et al*. Gemini AI vs. ChatGPT: A comprehensive examination alongside ophthalmology residents in medical knowledge. *Graefes Arch Clin Exp Ophthalmol* 2024;Available from.
43 Diniz-Freitas M, López-Pintor RM, Santos-Silva AR, *et al*. Assessing the accuracy and readability of ChatGPT-4 and Gemini in answering oral cancer queries—an exploratory study. *Explor Digit Health Technol* 2024;Available from:334–45.
44 Schmidl B, Hütten T, Pigorsch S, *et al*. 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* 2024;281:6099–109.
45 Fujimoto M, Kuroda H, Katayama T, *et al*. Evaluating large language models in dental anesthesiology: a comparative analysis of chatgpt-4, claude 3 opus, and gemini 1.0 on the japanese dental society of anesthesiology board certification exam. 2024. Available: https://www.cureus.com/articles/293672-evaluating-large-language-models-in-dental-anesthesiology-a-comparative-analysis-of-chatgpt-4-claude-3-opus-and-gemini-10-on-the-japanese-dental-society-of-anesthesiology-board-certification-exam
46 Mete U. Evaluating the Performance of ChatGPT, Gemini, and Bing Compared with Resident Surgeons in the Otorhinolaryngology In-service Training Examination. *tao* 2024;Available from.
47 Ponzo V, Rosato R, Scigliano MC, *et al*. Comparison of the Accuracy, Completeness, Reproducibility, and Consistency of Different AI Chatbots in Providing Nutritional Advice: An Exploratory Study. *J Clin Med* 2024;13:7810.
48 Cetin HK, Demir T. Assessing the knowledge of ChatGPT and Google Gemini in answering peripheral artery disease-related questions. *Vascular* 2025;Available from.
49 Hendrycks D, Burns C, Basart S, *et al*. Aligning ai with shared human values. 2020. Available: http://arxiv.org/abs/2008.02275
50 Zhang W, Deng Y, Liu B, *et al*. Sentiment analysis in the era of large language models: a reality check. 2023. Available: http://arxiv.org/abs/2305.15005
51 Gondode P, Duggal S, Garg N, *et al*. Comparative Analysis of Accuracy, Readability, Sentiment, and Actionability: Artificial Intelligence Chatbots (ChatGPT and Google Gemini) versus Traditional Patient Information Leaflets for Local Anesthesia in Eye Surgery. *Br Ir Orthopt J* 2024;20:183–92.
52 Kaliterna M, Žuljević MF, Ursić L, *et al*. Testing the capacity of Bard and ChatGPT for writing essays on ethical dilemmas: A cross-sectional study. *Sci Rep* 2024;14:26046.
53 Elyoseph Z, Hadar-Shoval D, Asraf K, *et al*. ChatGPT outperforms humans in emotional awareness evaluations. *Front Psychol* 2023;14:1199058.
54 Adilmetova G, Nassyrov R, Meyerbekova A, *et al*. Evaluating ChatGPT's Multilingual Performance in Clinical Nutrition Advice Using Synthetic Medical Text: Insights from Central Asia. *J Nutr* 2025;155:729–35.
55 Seyreen F, Erum KT, Rimsha K, *et al*. Evaluating the comprehension and accuracy of ChatGPT's responses to diabetes-related questions in Urdu compared to English. *Digit Health* 2024.
56 Samaan JS, Yeo YH, Ng WH, *et al*. ChatGPT's ability to comprehend and answer cirrhosis related questions in Arabic. *Arab J Gastroenterol* 2023;24:145–8.
57 Xia D, Song M, Zhu T. A comparison of the persuasiveness of human and ChatGPT generated pro-vaccine messages for HPV. *Front Public Health* 2024;12:1515871.
58 Wang Y, Liang L, Li R, *et al*. Comparison of the Performance of ChatGPT, Claude and Bard in Support of Myopia Prevention and Control. *J Multidiscip Healthc* 2024;17:3917–29.