Clinical Microbiology and Infection xxx (xxxx) xxx

Contents lists available at ScienceDirect

Clinical Microbiology and Infection

journal homepage: www.clinicalmicrobiologyandinfection.com

Original article

Comparing large language models for antibiotic prescribing in different clinical scenarios: which performs better?

Andrea De Vito [1,*], Nicholas Geremia [2,3], Davide Fiore Bavaro [4,5], Susan K. Seo [6], Justin Laracy [6], Maria Mazzitelli [7], Andrea Marino [8], Alberto Enrico Maraolo [9], Antonio Russo [10], Agnese Colpani [1], Michele Bartoletti [4,5], Anna Maria Cattelan [7], Cristina Mussini [11], Saverio Giuseppe Parisi [12], Luigi Angelo Vaira [13], Giuseppe Nunnari [8], Giordano Madeddu [1]

[1] *Unit of Infectious Diseases, Department of Medicine, Surgery and Pharmacy, Sassari, Italy*
[2] *Unit of Infectious Diseases, Department of Clinical Medicine, Ospedale dell'Angelo, Venice, Italy*
[3] *Unit of Infectious Diseases, Department of Clinical Medicine, Ospedale Civile S.S. Giovanni e Paolo, Venice, Italy*
[4] *Department of Biomedical Sciences, Humanitas University, Pieve Emanuele, Milan, Italy*
[5] *Infectious Diseases Unit - Department of Biomedical Sciences - Istituti di Ricovero e Cura a Carattere Scientifico (IRCCS) Humanitas Research Hospital, Rozzano, Milan, Italy*
[6] *Infectious Diseases Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA*
[7] *Infectious and Tropical Diseases Unit, Department of Molecular Medicine, Padua University Hospital, Padua, Italy*
[8] *Unit of Infectious Diseases, Department of Clinical and Experimental Medicine, Azienda Ospedaliera di Rilievo Nazionale e di Alta Specializzazione (ARNAS), Garibaldi Hospital, University of Catania, Catania, Italy*
[9] *Section of Infectious Diseases, Department of Clinical Medicine and Surgery, University of Naples "Federico II," Naples, Italy*
[10] *Department of Mental Health and Public Medicine-Infectious Diseases Unit, University of Campania Luigi Vanvitelli, Naples, Italy*
[11] *Infectious Diseases Unit, Department of Surgical, Medical, Dental and Morphological Sciences, Azienda Ospedaliera-Universitaria of Modena, University of Modena and Reggio Emilia, Modena, Italy*
[12] *Department of Molecular Medicine, University of Padua, Padua, Italy*
[13] *Maxillofacial Surgery Operative Unit, Department of Medicine, Surgery and Pharmacy, University of Sassari, Sassari, Italy*

ARTICLE INFO

Article history:
Received 4 January 2025
Received in revised form
6 March 2025
Accepted 11 March 2025
Available online xxx

Editor: Andre Kalil

Keywords:
Antibiotic treatment
Antimicrobial susceptibility testing
ChatGPT-o1
Difficult-to-treat infection
Large language models
LLMs

ABSTRACT

Objectives: Large language models (LLMs) show promise in clinical decision-making, but comparative evaluations of their antibiotic prescribing accuracy are limited. This study assesses the performance of various LLMs in recommending antibiotic treatments across diverse clinical scenarios.
Methods: Fourteen LLMs, including standard and premium versions of ChatGPT, Claude, Copilot, Gemini, Le Chat, Grok, Perplexity, and Pi.ai, were evaluated using 60 clinical cases with antibiograms covering 10 infection types. A standardized prompt was used for antibiotic recommendations focusing on drug choice, dosage, and treatment duration. Responses were anonymized and reviewed by a blinded expert panel assessing antibiotic appropriateness, dosage correctness, and duration adequacy.
Results: A total of 840 responses were collected and analysed. ChatGPT-o1 demonstrated the highest accuracy in antibiotic prescriptions, with 71.7% (43/60) of its recommendations classified as correct and only one (1.7%) incorrect. Gemini and Claude 3 Opus had the lowest accuracy. Dosage correctness was highest for ChatGPT-o1 (96.7%, 58/60), followed by Perplexity Pro (90.0%, 54/60) and Claude 3.5 Sonnet (91.7%, 55/60). In treatment duration, Gemini provided the most appropriate recommendations (75.0%, 45/60), whereas Claude 3.5 Sonnet tended to over-prescribe duration. Performance declined with increasing case complexity, particularly for difficult-to-treat microorganisms.
Discussion: : There is significant variability among LLMs in prescribing appropriate antibiotics, dosages, and treatment durations. ChatGPT-o1 outperformed other models, indicating the potential of advanced LLMs as decision-support tools in antibiotic prescribing. However, decreased accuracy in complex cases

DOI of original article: https://doi.org/10.1016/j.cmi.2025.03.022.
* Corresponding author. Andrea De Vito, Unit of Infectious Diseases, Department of Medicine, Surgery and Pharmacy, Sassari, Italy.
E-mail address: andreadevitoaho@gmail.com (A. De Vito).

https://doi.org/10.1016/j.cmi.2025.03.002

and inconsistencies among models highlight the need for careful validation before clinical utilization.
Andrea De Vito, Clin Microbiol Infect 2025;▪:1

Introduction

The appropriate selection of antibiotics is a cornerstone of effective patient care. It is widely recognized that delayed or incorrect therapy can lead to severe complications, increased mortality, and the development of antimicrobial resistance (AMR) [1–3]. Antimicrobial stewardship (AMS) programmes emphasize the necessity of accurate, timely, and tailored prescriptions to optimize patient outcomes and reduce the risk of resistance [4,5]. However, the process of selecting the right antibiotic, the correct dose, and the appropriate duration is nuanced, especially in settings where multidrug-resistant organisms (MDROs) are prevalent and in which an infectious disease (ID) consultation is absent. Although guidelines for the management of MDROs are available, clinical adjudication is crucial for the management of complex clinical cases.

In hospitals with dedicated ID teams, specialists are often consulted to guide these complex decisions. ID physicians bring invaluable experience in interpreting antibiotic susceptibility testing (AST), understanding local resistance patterns, and navigating the evolving landscape of antimicrobial therapy [6–8]. However, access to ID specialists is often limited in smaller hospitals or community settings increasing the risk of treatment failure, adverse reactions, and both the emergence and spreading of AMR [9].

These challenges underscore the importance of enhancing antibiotic decision-making support in all healthcare settings. In this context, artificial intelligence systems and large language models (LLMs) can potentially revolutionize clinical practice. The application of LLMs in medicine has rapidly gained attention, with LLMs like ChatGPT and other advanced systems demonstrating potential in various clinical domains. The ability of LLMs-driven tools to support diagnostics, decision-making, and medical education has been studied across multiple specialties [10–12]. Regarding the use of LLMs for bacterial infections, few studies have been conducted, but no comparison between different LLMs has been performed [13–15]. Only a few studies evaluated the performance of different LLMs [16], but none have specifically examined the ability of ChatBots in the context of antibiotic prescribing. Therefore, this paper aims to address this gap by conducting a comprehensive evaluation of leading LLMs and comparing their performance in providing antibiotic recommendations.

Methods

We conducted a comparative study to assess the ability of various LLMs to evaluate clinical scenarios and AST and prescribe the appropriate antibiotic treatment. Three specialists in ID formulated 60 clinical cases with AST focusing on 10 different topics: (a) acute bacterial skin and skin structure infections; (b) bloodstream infection (BSI); (c) bone and joint infection, (d) central nervous system (CNS) infections; (e) ear and eye infections; (f) endocarditis; (g) intrabdominal infections (IAI); (h) pneumonia; (i) transplant-related infections; (j) urinary tract infections. The list of clinical cases is available in Supplementary material.

ChatBot selection

We investigated the performance of 8 LLMs, both standard and premium version, for a total of 14 LLMs (Table S1).

For each clinical case, queries were entered manually between September 1 and 30, 2024, and responses were directly collected from the interface. To prevent memorization from influencing results, we created a new task for each scenario, and the same prompt was used for all LLMs and cases (Table S2).

Blind review and evaluation

All responses were anonymized and reviewed by a blind panel of experts who evaluated the answers between October 1 and November 10, 2024. They were asked to evaluate three aspects: appropriateness, correctness of suggested dosage, and adequacy of treatment duration (Table S3).

Each clinical case was independently evaluated by a panel of nine ID specialists with at least 5 years of experience in AMS and clinical decision-making (Table S4). Each expert reviewed the cases individually and assigned a score based on predefined criteria. In cases where discrepancies arose among reviewers, a secondary adjudication panel consisting of senior ID professors re-evaluated the responses and reached a consensus. The experts did not adhere to a single set of guidelines but instead based their assessments on the most recent and robust evidence available. When both European and U.S. guidelines were available for a specific infection, U.S. experts primarily followed U.S. guidelines, whereas European experts used European guidelines.

For clinical conditions without specific international guidelines, such as ocular and ear infections, the panel relied on their clinical expertise and the latest peer-reviewed literature to evaluate antibiotic appropriateness, dosage, and treatment duration.

Microorganisms were categorized based on their difficulty to treat (DTR) according to AST. We divided them into four groups: (a) DTR microorganisms (*Acinetobacter baumannii, Stenotrophomonas maltophilia* resistant to trimethoprim/sulfamethoxazole, DTR *Pseudomonas aeruginosa* [17]), carbapenem-resistant *Enterobacterales*, and vancomycin-resistant *Enterococcus*; (b) microorganisms with limited treatment options (methicillin-resistant staphylococci, penicillin-resistant streptococci, and Gram-negative bacteria resistant to third-generation cephalosporins [extended-spectrum beta-lactamases or AmpC beta-lactamases]); (c) wild-type microorganisms; (d) microorganisms without an AST. The list of microorganisms and the AST distribution by Gram stain and resistance categories are available in Tables S5 and S6.

Statistical analysis

Statistical methods were conducted to compare the performance across different ChatBots. Inter-rater reliability was assessed using Fleiss' Kappa to measure agreement among the reviewers. Kappa values were interpreted based on standard thresholds. Chi-squared test was used to assess the presence of differences between groups. Binomial logistic regression was performed to estimate ORs with 95% CIs, adjusting for predefined covariates and those that were statistically significant in univariable analysis. Statistical significance was set at p values of less than 0.05, and data analysis was carried out through STATA (Version 16.1 StataCorp, College Station, TX).

Please cite this article as: De Vito A et al., Comparing large language models for antibiotic prescribing in different clinical scenarios: which performs better?, Clinical Microbiology and Infection, https://doi.org/10.1016/j.cmi.2025.03.002

Ethical considerations

Given the nature of the study involving hypothetical clinical scenarios based on AST from real bacterial strains yet containing no real patient data or demographic characteristics, an ethical review exemption was sought and granted, aligning with institutional guidelines on human subject research.

Results

Overall, 840 answers were collected and evaluated by a blinded panel of experts in antibiotic treatments.

Accuracy of prescribed antibiotics

Regarding the antibiotic prescription, expert evaluations showed substantial agreement in antibiotic selection ($\kappa = 0.799$, $p < 0.001$).

Among the different LLMs, ChatGPT-o1 performed the best in terms of correct antibiotic prescriptions, with 71.7% of responses classified as correct, followed by ChatGPT 4o (53.3%) and Perplexity Pro (56.7%). In contrast, Gemini and Claude 3 Opus were the least accurate, with only 30.0% and 48.3% of responses deemed correct, respectively (Table 1). The proportion of incorrect answers was the lowest for ChatGPT-o1 and ChatGPT 4o, whereas Gemini and Claude 3 Opus had the highest rates of incorrect prescriptions.

Among the free-access LLMs, Perplexity had the highest percentage of correct answers but also the highest percentage of partially correct answers. When combining the correct answers with overtreatment answers, Copilot outperformed the other free-access LLMs and was only surpassed by ChatGPT-o1.

When categorizing by Gram stain (Tables 2 and 3), correct answers were significantly higher for Gram-positive microorganisms ($p < 0.001$). ChatGPT-o1 consistently outperformed others, whereas Gemini performed notably poorly, especially in Gram-negative scenarios (24.3% correct) (Table 3).

Microorganisms classified as DTR had the highest number of incorrect responses, with overtreatment rates significantly lower for DTR microorganisms (Table 4). ChatGPT-o1 excelled across all difficulty categories, whereas Gemini performed the worst (Tables S7–S12).

Regarding infection types, wrong answers were most frequent for IAI (22.6%), ear and eye infections (19.1%), CNS infections (13.1%), and BSI (10.7%). Overtreatment was most common in pneumonia (60.7%) and ear and eye infections (39.4%) (Table S12).

Dosage recommendations

Inter-rater agreement on dosage was moderate to substantial ($\kappa = 0.644$, $p < 0.001$). ChatGPT-o1 achieved the highest accuracy in dosage recommendations (96.7%), with no underdosing cases and minimal overdosage. Perplexity Pro and Claude 3.5 Sonnet also performed well (90.0% and 91.7%, respectively). Gemini had the poorest accuracy (60.0%) and the highest underdosing rate (26.7%). Only one instance of a fabricated dosage was identified, with Gemini suggesting an implausible ceftazidime/avibactam regimen for endocarditis caused by *Escherichia coli* (Fig. 1).

Table 1
Percentage of antibiotic choice answers evaluated as incorrect, partially correct, correct, and overtreatment, divided by different large language models

Large language model	Wrong, *n* (%)	Partially correct, *n* (%)	Correct, *n* (%)	Overtreatment, *n* (%)
ChatGPT	4 (6.7)	9 (15.0)	29 (48.3)	18 (30.0)
ChatGPT-o1	1 (1.7)	10 (16.7)	43 (71.7)	6 (10.0)
ChatGPT4o	2 (3.3)	11 (18.3)	32 (53.3)	15 (25.0)
Claude 3 Opus	9 (15.0)	13 (21.7)	29 (48.3)	9 (15.0)
Claude 3.5 Sonnet	6 (10.0)	10 (16.7)	29 (48.3)	15 (25.0)
Copilot	3 (5.0)	10 (16.7)	29 (48.3)	18 (30.0)
Copilot Pro	5 (8.3)	10 (16.7)	26 (43.3)	19 (31.7)
Gemini	14 (23.3)	8 (13.3)	19 (31.7)	19 (31.7)
Gemini Advance	7 (11.7)	12 (20.0)	24 (40.0)	17 (28.3)
Grok 2	5 (8.3)	12 (20.0)	24 (40.0)	19 (31.7)
Le Chat - Large 2	7 (11.7)	10 (16.7)	28 (46.7)	15 (25.0)
Perplexity	4 (6.7)	14 (23.3)	31 (51.7)	11 (18.3)
Perplexity Pro	4 (6.7)	11 (18.3)	34 (56.7)	11 (18.3)
pi.ai	4 (6.7)	12 (20.0)	26 (43.3)	18 (30.0)
Total	75 (8.9)	152 (18.1)	403 (48.0)	210 (25.0)

Table 2
Percentage of antibiotic choice answers about Gram-positive microorganisms evaluated as incorrect, partially correct, correct, and overtreatment, divided by different large language models

Large language model	Wrong, *n* (%)	Partially correct, *n* (%)	Correct, *n* (%)	Overtreatment, *n* (%)
ChatGPT	1 (4.3)	0	15 (65.2)	7 (30.4)
ChatGPT-o1	0	1 (4.3)	20 (87.0)	2 (8.7)
ChatGPT4o	1 (4.3)	1 (4.3)	18 (78.3)	3 (13.0)
Claude 3 Opus	3 (13.0)	4 (17.4)	15 (65.2)	1 (4.3)
Claude 3.5 Sonnet	2 (8.7)	1 (4.3)	16 (69.6)	4 (17.4)
Copilot	1 (4.4)	0	15 (65.2)	7 (30.4)
Copilot pro	1 (4.4)	1 (4.4)	14 (60.9)	7 (30.4)
Gemini	4 (17.4)	1 (4.4)	10 (43.5)	8 (34.8)
Gemini Advance	2 (8.7)	3 (13.0)	11 (47.8)	7 (30.4)
Grok 2	2 (8.7)	5 (21.7)	11 (47.8)	5 (21.7)
Le Chat	1 (4.4)	3 (13.0)	17 (73.9)	2 (8.7)
Perplexity	1 (4.4)	3 (13.0)	15 (65.2)	4 (17.4)
Perplexity pro	2 (8.7)	3 (13.0)	16 (69.6)	2 (8.7)
pi.ai	0	4 (17.4)	14 (60.9)	5 (21.7)
Total	21 (6.5)	30 (9.3)	207 (64.3)	64 (19.9)

Please cite this article as: De Vito A et al., Comparing large language models for antibiotic prescribing in different clinical scenarios: which performs better?, Clinical Microbiology and Infection, https://doi.org/10.1016/j.cmi.2025.03.002

Table 3
Percentage of antibiotic choice answers about Gram-negative microorganisms evaluated as incorrect, partially correct, correct, and overtreatment, divided by different large language models

Large language model	Wrong, *n* (%)	Partially correct, *n* (%)	Correct, *n* (%)	Overtreatment, *n* (%)
ChatGPT	3 (8.1)	9 (24.3)	14 (37.8)	11 (29.7)
ChatGPT-o1	1 (2.7)	9 (24.3)	23 (62.2)	4 (10.8)
ChatGPT4o	1 (2.7)	10 (27.0)	14 (37.8)	12 (32.4)
Claude 3 Opus	6 (16.2)	9 (24.3)	14 (37.8)	8 (21.6)
Claude 3.5 Sonnet	4 (10.8)	9 (24.3)	13 (35.1)	11 (29.7)
Copilot	2 (5.4)	10 (27.0)	14 (37.8)	11 (29.7)
Copilot pro	4 (10.8)	9 (24.3)	12 (32.4)	12 (32.4)
Gemini	10 (27.0)	7 (18.9)	9 (24.3)	11 (29.7)
Gemini Advance	5 (13.5)	9 (24.3)	13 (35.1)	10 (27.0)
Grok 2	3 (8.1)	7 (18.9)	13 (35.1)	14 (37.8)
Le Chat	6 (16.2)	7 (18.9)	11 (29.7)	13 (35.1)
Perplexity	3 (8.1)	11 (29.7)	16 (43.2)	7 (18.9)
Perplexity pro	2 (5.4)	8 (21.6)	18 (48.7)	9 (24.3)
pi.ai	4 (10.8)	8 (21.6)	12 (32.4)	13 (35.1)
Total	54 (10.4)	122 (23.6)	196 (37.8)	146 (28.2)

Table 4
Percentage of antibiotic choice answers about microorganisms difficult-to-treat evaluated as incorrect, partially correct, correct, and overtreatment, divided by different large language models

Large language model	Wrong, *n* (%)	Partially correct, *n* (%)	Correct, *n* (%)	Overtreatment, *n* (%)
ChatGPT	3 (14.3)	8 (38.1)	10 (47.6)	0
ChatGPT-o1	1 (4.8)	6 (28.6)	14 (66.7)	0
ChatGPT4o	1 (4.8)	10 (47.6)	9 (42.9)	1 (4.8)
Claude 3 Opus	7 (33.3)	6 (28.6)	8 (38.1)	0
Claude 3.5 Sonnet	3 (14.3)	8 (38.1)	10 (47.6)	0
Copilot	2 (9.5)	9 (42.9)	10 (47.6)	0
Copilot pro	4 (19.0)	8 (38.1)	8 (38.1)	1 (4.8)
Gemini	10 (47.6)	5 (23.8)	4 (19.0)	2 (9.5)
Gemini Advance	5 (23.8)	8 (38.1)	7 (33.3)	1 (4.8)
Grok 2	4 (19.0)	7 (33.3)	10 (47.6)	0
Le Chat	6 (28.6)	6 (28.6)	7 (33.3)	2 (9.5)
Perplexity	2 (9.5)	9 (42.9)	10 (47.6)	0
Perplexity pro	3 (14.3)	7 (33.3)	11 (52.4)	0
pi.ai	4 (19.0)	6 (28.6)	8 (38.1)	3 (14.3)
Total	55 (18.7)	103 (35.0)	126 (42.9)	10 (3.4)

Fig. 1. Percentage of antibiotic dosage answers evaluated as low dosage, correct dosage, and high dosage, divided by different large language models.

Please cite this article as: De Vito A et al., Comparing large language models for antibiotic prescribing in different clinical scenarios: which performs better?, Clinical Microbiology and Infection, https://doi.org/10.1016/j.cmi.2025.03.002

No significant differences in dosage accuracy were noted between Gram-positive and Gram-negative microorganisms. However, correct dosage recommendations were highest for microorganisms without AST and DTR microorganisms, whereas underdosing was most common in wild-type microorganisms and over-dosing in those with limited treatment options (Tables S13 and S14).

Treatment duration

Agreement on treatment duration was moderate ($\kappa = 0.653$, $p < 0.001$) and Gemini produced the highest proportion of correct recommendations, with 75.0% classified as appropriate, followed by Perplexity Pro at 73.3% and ChatGPT-o1 at 70.0%. Furthermore, Gemini had the highest rate of overly short treatment (18.3%), followed by Copilot Pro (13.3%). Regarding the recommendations for overly long treatment, Claude 3.5 Sonnet had the highest proportion (41.7%), followed by Grok 2 and ChatGPT-4o (Fig. 2).

Discussion

Our findings highlighted significant variability in LLMs' performance in prescribing accurate antibiotics and dosages. These results could have important implications for the potential integration of LLM tools in clinical practice.

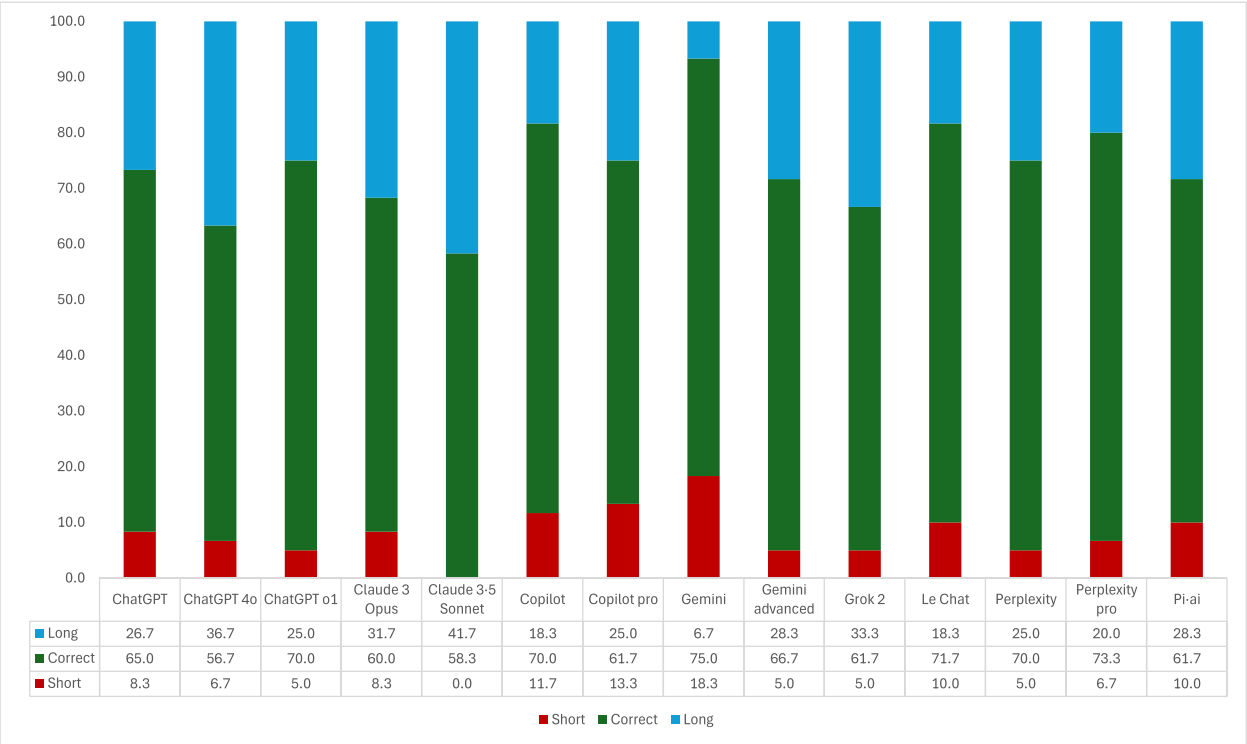
Our analysis revealed that ChatGPT-o1 achieved the highest accuracy in antibiotic prescription, with 70.0% of its responses classified as correct and only 1.7% deemed as incorrect. In contrast, models like Gemini and Claude 3 Opus showed considerably lower accuracy, with correct prescription rates of 30.0% and 48.3%, respectively. When considering dosage accuracy, ChatGPT-o1 led with 96.7% correct dosages, whereas Gemini was the poorest. Of note, incorrect prescriptions can have severe consequences, including potentially fatal outcomes for patients. Thus, we believe that although LLMs show promise as supportive tools, they are not yet ready for unsupervised use in clinical settings. It remains crucial that users possess substantial expertise in the relevant medical domain and utilize LLMs primarily as a source for alternative perspectives. Although LLMs may aid in decision-making, they should not be solely relied upon for critical clinical judgements.

The superior performance of ChatGPT-o1 compared with the other LLMs evaluated may be attributed to the implementation of chain-of-thought reasoning within its architecture, and to the advanced training data and fine-tuning techniques [18].

However, the variability observed among different LLMs underscores the necessity for caution. Although ChatGPT-o1 demonstrated high accuracy, other models like Gemini and Claude 3 Opus did not. The poor results from Gemini are in contrast with the literature. Pirkle et al. [19] found similar performances between ChatGPT and Gemini regarding appropriate recommendations for common paediatric orthopaedic conditions, and Tong et al. [20] found that it provides more concise and intuitive responses than ChatGPT-3.5 and ChatGPT-4. However, a key distinction between these studies and ours is that orthopaedic evaluations were based on theoretical questions, whereas our study utilized clinical scenarios with AST, requiring complex reasoning and individualized decision-making. LLMs may perform well in structured, guideline-based tasks, but nuanced antimicrobial prescribing involves factors such as host status, infection site, and resistance patterns, making it a more challenging domain.

Our study also found that LLMs accuracy decreased with increasing complexity, particularly for DTR microorganisms. Incorrect responses were more frequent in cases involving MDRO, whereas overtreatment was less common. These findings suggest that although LLMs may perform adequately in standard cases, their reliability diminishes in more complex scenarios.

These limitations observed in LLM performance align with well-documented challenges in antibiotic prescribing by non-ID



	ChatGPT	ChatGPT 4o	ChatGPT o1	Claude 3 Opus	Claude 3·5 Sonnet	Copilot	Copilot pro	Gemini	Gemini advanced	Grok 2	Le Chat	Perplexity	Perplexity pro	Pi·ai
Long	26.7	36.7	25.0	31.7	41.7	18.3	25.0	6.7	28.3	33.3	18.3	25.0	20.0	28.3
Correct	65.0	56.7	70.0	60.0	58.3	70.0	61.7	75.0	66.7	61.7	71.7	70.0	73.3	61.7
Short	8.3	6.7	5.0	8.3	0.0	11.7	13.3	18.3	5.0	5.0	10.0	5.0	6.7	10.0

Fig. 2. Percentage of treatment duration answers evaluated as low dosage, correct dosage, and high dosage, divided by different large language models.

Please cite this article as: De Vito A et al., Comparing large language models for antibiotic prescribing in different clinical scenarios: which performs better?, Clinical Microbiology and Infection, https://doi.org/10.1016/j.cmi.2025.03.002

specialists. Studies indicate that 30–50% of inpatient antibiotic prescriptions may be inappropriate because of suboptimal guideline adherence, broad-spectrum overuse, or AST misinterpretation [21–23]. In contrast, ID specialists and AMS teams improve prescribing accuracy, patient outcomes, and reduced AMR [24]. Although our study highlights the LLMs limitations in clinical decision-making, it reinforces a broader issue, even human prescribers struggle with complex antimicrobial decisions, particularly in the absence of ID consultation. This underscores the need for validated, expert-driven decision-support tools rather than reliance on off-the-shelf LLMs.

Our results align with the findings of Maillard et al. [14] who demonstrated that although ChatGPT-4 provided detailed and well-structured medical responses, its performance was far from perfect in more complex clinical situations. They found that although ChatGPT-4 could generate satisfactory management plans in some cases, significant errors arose in diagnosing, prescribing antibiotics, and handling source control measures.

In our study, the highest proportions of incorrect responses were observed in cases of IAI (22.6%), ear and eye infections (19.1%), CNS infections (13.1%), and BSI (10.7%). A high percentage of overtreatment was identified for pneumonia (60.7%) and ear and eye infections (39.4%). These findings reinforce that, despite strong performance in some areas, LLMs have clear limitations in complex cases, particularly those involving MDROs or intricate clinical decision-making.

Although LLMs may complement clinical practice, they cannot replace human expertise, especially in situations where standard guidelines may not fully address the complexity of real-world cases. Moreover, we must also consider the limits of updating LLMs with respect to published articles given the inability to access all the information available, especially those behind the paywall. This limitation means that LLMs may not incorporate a substantial part of the literature if not adequately trained by the researcher or clinicians, potentially leading to gaps in knowledge and less informed clinical guidance. In addition, LLMs require extensive computational resources and time to train on large datasets, including the latest clinical advances, guidelines and published studies, resulting in a significant delay between the publication of new medical research and its inclusion in the training data of an LLM.

This highlights the essential role of human intervention in interpreting guidelines and tailoring management to meet the specific needs of each patient, as clinical realities often surpass the straightforward application of protocols.

Several limitations of our study should be acknowledged. First, the clinical scenarios may not capture real-world cases' full complexity, such as allergy to antibiotics, and drug–drug interactions. Second, the LLMs were assessed at a single point in time, but these models are continually evolving, and their performance may change with updates. Third, we used a uniform prompt for all LLMs, which may not have been optimized for each model's strengths. Additionally, although our study evaluates optimal antibiotic treatments based on global standards, it is important to consider that dosing guidelines and drug availability can vary widely by region especially in low- and middle-income countries [25,26]. Consequently, treatments deemed suboptimal in our study categorized as ‘partially correct’ may represent the best or only available options in these settings. This disparity in access highlights a crucial limitation of our study: we did not fully consider the availability of antibiotics across different regions. Finally, we observed that LLMs only suggested treatments not tested in the AST in a few cases. This often resulted in overtreatment or answers categorized as ‘partially correct.’ Therefore, the introduction of newer antibiotics may present an additional challenge.

It will be important to account for these regional discrepancies in future studies, as this could shift the definition of ‘optimal treatment’ depending on the healthcare context. In addition, further research is needed to evaluate the integration of LLMs in real-world settings. Studies should also assess the models' performance over time, considering updates and improvements to the algorithms.

Conclusions

Our study highlights both the potential and significant limitations of current LLMs in antibiotic prescribing. Although models like ChatGPT-o1 demonstrated higher accuracy in some aspects, substantial variability across LLMs and inconsistent performance in complex cases underscore their unreliability in clinical decision-making. No model consistently met the standards required for safe antibiotic selection, dosage, or treatment duration.

These findings reinforce that off-the-shelf LLMs should not be relied upon for antibiotic prescribing in real-world practice. The risks of incorrect or inappropriate recommendations underscore the need for extreme caution. Although artificial intelligence-driven tools may hold promise for future integration into AMS programmes, they remain insufficient for standalone clinical use. Clinicians must avoid using LLMs for direct medical decisions, as this could compromise patient safety and contribute to AMR. Further research, model refinement, and validation through randomized clinical trials are essential before LLMs can be considered for clinical use. Collaboration among clinicians, developers, and researchers is crucial to harnessing LLMs' benefits while mitigating risks, ultimately improving patient care and supporting AMS.

Author contributions

A.D.V. and N.G. were responsible for conceptualization and visualization. A.D.V., N.G., and G.M. were responsible for methodology. A.D.V. was responsible for formal analysis. D.F.B., S.K.S., J.L., M.M., A.M., A.E.M., A.R., A.C., M.B., A.M.C., C.M., S.G.P., L.A.V., and G.N. were responsible for investigation. D.F.B., S.K.S., J.L., M.M., A.M., A.E.M., A.R., A.C., M.B., A.M.C., C.M., S.G.P., L.A.V., and G.N. were responsible for data curation. A.D.V., N.G., and A.C. were responsible for writing–original draft preparation. D.F.B., S.K.S., J.L., M.M., A.M., A.E.M., A.R., M.B., A.M.C., C.M., S.G.P., L.A.V., G.N., and G.M. were responsible for writing–review and editing. G.M. was responsible for supervision. A.D.V. and N.G. contributed equally to this work.

Transparency declaration

Potential conflict of interest

The authors declare that they have no conflicts of interest.

Financial report

No funding has been received for this manuscript.

Ethics declaration

The ethical review and approval requirement was waived because the study did not include any analysis of humans or animals.

Please cite this article as: De Vito A et al., Comparing large language models for antibiotic prescribing in different clinical scenarios: which performs better?, Clinical Microbiology and Infection, https://doi.org/10.1016/j.cmi.2025.03.002

Data availability

The data collected for this study will be made available to others, and they will be available upon publication of this article and can be accessed in the Supplementary Materials accompanying the manuscript. The data are freely available to anyone without any restrictions.

Acknowledgements

We want to thank Stefano Zugno Brunetti for his invaluable assistance in creating and refining the graphical representations for this study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cmi.2025.03.002.

References

[1] Worldwide Antimicrobial Resistance National/International Network Group (WARNING) Collaborators. Ten golden rules for optimal antibiotic use in hospital settings: the WARNING call to action. World J Emerg Surg 2023;18:50. https://doi.org/10.1186/S13017-023-00518-3.

[2] López Romo A, Quirós R. Appropriate use of antibiotics: an unmet need. Ther Adv Urol 2019;11:1756287219832174. https://doi.org/10.1177/1756287219832174.

[3] Patel K, Bunachita S, Agarwal AA, Bhamidipati A, Patel UK. A comprehensive overview of antibiotic selection and the factors affecting it. Cureus 2021;13:e13925. https://doi.org/10.7759/CUREUS.13925.

[4] Zay Ya K, Win PTN, Bielicki J, Lambiris M, Fink G. Association between antimicrobial stewardship programs and antibiotic use globally: a systematic review and meta-analysis. JAMA Netw Open 2023;6:e2253806. https://doi.org/10.1001/JAMANETWORKOPEN.2022.53806.

[5] Abdel Hadi H, Eltayeb F, Al Balushi S, Daghfal J, Ahmed F, Mateus C. Evaluation of hospital antimicrobial stewardship programs: implementation, process, impact, and outcomes, review of systematic reviews. Antibiotics (Basel) 2024;13:253. https://doi.org/10.3390/ANTIBIOTICS13030253.

[6] Hollingshead CM, Khazan AE, Franco JH, Ciricillo JA, Haddad MN, Berry JT, et al. A needs assessment for infectious diseases consultation in community hospitals. Infect Dis Ther 2023;12:1725. https://doi.org/10.1007/S40121-023-00810-4.

[7] Vaughn VM, Greene MT, Ratz D, Fowler KE, Krein SL, Flanders SA, et al. Antibiotic stewardship teams and *Clostridioides difficile* practices in United States hospitals: a national survey in the joint commission antibiotic stewardship standard era. Infect Control Hosp Epidemiol 2020;41:143–8. https://doi.org/10.1017/ICE.2019.313.

[8] Sunenshine RH, Liedtke LA, Jernigan DB, Strausbaugh LJ, Infectious Diseases Society of America Emerging Infections Network. Role of infectious diseases consultants in management of antimicrobial use in hospitals. Clin Infect Dis 2004;38:934–8. https://doi.org/10.1086/382358.

[9] Pulcini C, Botelho-Nevers E, Dyar OJ, Harbarth S. The impact of infectious disease specialists on antibiotic prescribing in hospitals. Clin Microbiol Infec 2014;20:963–72. https://doi.org/10.1111/1469-0691.12751.

[10] Lechien JR, Briganti G, Vaira LA. Accuracy of ChatGPT-3.5 and -4 in providing scientific references in otolaryngology-head and neck surgery. Eur Arch Otorhinolaryngol 2024;281:2159–65. https://doi.org/10.1007/S00405-023-08441-8.

[11] De Vito A, Colpani A, Moi G, Babudieri S, Calcagno A, Calvino V, et al. Assessing ChatGPT's potential in HIV prevention communication: a comprehensive evaluation of accuracy, completeness, and inclusivity. AIDS Behav 2024;28:2746–54. https://doi.org/10.1007/S10461-024-04391-2.

[12] Bahir D, Zur O, Attal L, Nujeidat Z, Knaanie A, Pikkel J, et al. Gemini AI vs. ChatGPT: a comprehensive examination alongside ophthalmology residents in medical knowledge. Graefes Arch Clin Exp Ophthalmol 2025;263:527–36. https://doi.org/10.1007/S00417-024-06625-4.

[13] De Vito A, Geremia N, Marino A, Bavaro DF, Caruana G, Meschiari M, et al. Assessing ChatGPT's theoretical knowledge and prescriptive accuracy in bacterial infections: a comparative study with infectious diseases residents and specialists. Infection 2024. https://doi.org/10.1007/s15010-024-02350-6.

[14] Maillard A, Micheli G, Lefevre L, Guyonnet C, Poyart C, Canouï E, et al. Can Chatbot artificial intelligence replace infectious diseases physicians in the management of bloodstream infections? A prospective cohort study. Clin Infect Dis 2024;78:825–32. https://doi.org/10.1093/CID/CIAD632.

[15] Kaneda Y. ChatGPT in infectious diseases: a practical evaluation and future considerations. New Microbe. New Infect 2023;54:101166. https://doi.org/10.1016/J.NMNI.2023.101166.

[16] Meyer A, Soleman A, Riese J, Streichert T. Comparison of ChatGPT, Gemini, and Le Chat with physician interpretations of medical laboratory questions from an online health forum. Clin Chem Lab Med 2024;62:2425–34. https://doi.org/10.1515/CCLM-2024-0246.

[17] Cosentino F, Viale P, Giannella M. MDR/XDR/PDR or DTR? Which definition best fits the resistance profile of *Pseudomonas aeruginosa*? Curr Opin Infect Dis 2023;36:564–71. https://doi.org/10.1097/QCO.0000000000000966.

[18] OpenAI. OpenAI O1 system card. https://openai.com/index/openai-o1-system-card/. [Accessed 1 November 2024].

[19] Pirkle S, Yang JW, Blumberg TJ. Do ChatGPT and Gemini provide appropriate recommendations for pediatric orthopaedic conditions? J Pediatr Orthop 2025;45:e66–71. https://doi.org/10.1097/BPO.0000000000002797.

[20] Tong L, Zhang C, Liu R, Yang J, Sun Z. Comparative performance analysis of large language models: ChatGPT-3.5, ChatGPT-4 and Google Gemini in glucocorticoid-induced osteoporosis. J Orthop Surg Res 2024;19:574. https://doi.org/10.1186/S13018-024-04996-2.

[21] Durkin MJ, Keller M, Butler AM, Kwon JH, Dubberke ER, Miller AC, et al. An assessment of inappropriate antibiotic use and guideline adherence for uncomplicated urinary tract infections. Open Forum Infect Dis 2018;5:ofy198. https://doi.org/10.1093/OFID/OFY198.

[22] Fleming-Dutra KE, Hersh AL, Shapiro DJ, Bartoces M, Enns EA, File TM, et al. Prevalence of inappropriate antibiotic prescriptions among US ambulatory care visits, 2010-2011. JAMA 2016;315:1864–73. https://doi.org/10.1001/JAMA.2016.4151.

[23] Vazquez Deida AA, Bizune DJ, Kim C, Sahrmann JM, Sanchez GV, Hersh AL, et al. Opportunities to improve antibiotic prescribing for adults with acute sinusitis, United States, 2016–2020. Open Forum Infect Dis 2024;11:ofae420. https://doi.org/10.1093/OFID/OFAE420.

[24] Davey P, Marwick CA, Scott CL, Charani E, Mcneil K, Brown E, et al. Interventions to improve antibiotic prescribing practices for hospital inpatients. Cochrane Database Syst Rev 2017;2:CD003543. https://doi.org/10.1002/14651858.CD003543.pub4.

[25] Ren M, So AD, Chandy SJ, Mpundu M, Peralta AQ, Åkerfeldt K, et al. Equitable access to antibiotics: a core element and shared global responsibility for pandemic preparedness and response. J Law Med Ethics 2022;50:34–9. https://doi.org/10.1017/JME.2022.77.

[26] Wasan H, Reeta KH, Gupta YK. Strategies to improve antibiotic access and a way forward for lower middle-income countries. J Antimicrob Chemother 2024;79:1–10. https://doi.org/10.1093/JAC/DKAD291.