


# GP or ChatGPT? Ability of large language models (LLMs) to support general practitioners when prescribing antibiotics

Oanh Ngoc Nguyen<sup>1</sup>, Doaa Amin<sup>1</sup>, James Bennett<sup>2</sup>, Øystein Hetlevik<sup>3</sup>, Sara Malik<sup>4</sup>, Andrew Tout<sup>5</sup>, Heike Vornhagen<sup>6</sup> and Akke Vellinga <sup>1\*</sup>

<sup>1</sup>CARA Network, School of Public Health, Physiotherapy and Sports Science, University College Dublin, Dublin, Ireland; <sup>2</sup>NIHR In Practice Fellow, Hull York Medical School, University of Hull, Hull HU6 7RX, UK; <sup>3</sup>Department of Global Public Health and Primary Care, University of Bergen, Bergen, Norway; <sup>4</sup>Midleton Medi Center, Midleton, Co Cork, Ireland; <sup>5</sup>Division of General Internal Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA; <sup>6</sup>CARA Network, Insight Centre for Data Analytics, University of Galway, Galway, Ireland

\*Corresponding author. E-mail: Akke.vellinga@ucd.ie  
@Dr\_Akke

Received 11 November 2024; accepted 28 February 2025

**Introduction:** Large language models (LLMs) are becoming ubiquitous and widely implemented. LLMs could also be used for diagnosis and treatment. National antibiotic prescribing guidelines are customized and informed by local laboratory data on antimicrobial resistance.

**Methods:** Based on 24 vignettes with information on type of infection, gender, age group and comorbidities, GPs and LLMs were prompted to provide a treatment. Four countries (Ireland, UK, USA and Norway) were included and a GP from each country and six LLMs (ChatGPT, Gemini, Copilot, Mistral AI, Claude and Llama 3.1) were provided with the vignettes, including their location (country). Responses were compared with the country's national prescribing guidelines. In addition, limitations of LLMs such as hallucination, toxicity and data leakage were assessed.

**Results:** GPs' answers to the vignettes showed high accuracy in relation to diagnosis (96%–100%) and yes/no antibiotic prescribing (83%–92%). GPs referenced (100%) and prescribed (58%–92%) according to national guidelines, but dose/duration of treatment was less accurate (50%–75%). Overall, the GPs' accuracy had a mean of 74%. LLMs scored high in relation to diagnosis (92%–100%), antibiotic prescribing (88%–100%) and the choice of antibiotic (59%–100%) but correct referencing often failed (38%–96%), in particular for the Norwegian guidelines (0%–13%). Data leakage was shown to be an issue as personal information was repeated in the models' responses to the vignettes.

**Conclusions:** LLMs may be safe to guide antibiotic prescribing in general practice. However, to interpret vignettes, apply national guidelines and prescribe the right dose and duration, GPs remain best placed.

## Introduction

Inappropriate prescribing of antibiotics is a major cause of antimicrobial resistance, also referred to as the silent pandemic, as it has the potential to render many infections resistant to common antibiotics.<sup>1</sup> Up to 80% of antibiotics are prescribed in general practice and many interventions aim to address inappropriate prescribing.<sup>2</sup> Antibiotic prescribing in European general practice also shows wide variability in antibiotic prescribing practices. When evaluating antibiotic prescribing by indication and according to national guidelines, inconsistencies were observed in the adherence to recommended guidelines, with frequent overprescribing of broad-spectrum antibiotics or prescribing of antibiotics when

they are not indicated.<sup>3</sup> Similarly, research in the UK highlighted considerable variability in antibiotic prescribing across general practices, with changes in prescribing patterns not consistently aligning with updates to national guidelines.<sup>4</sup>

Large language models (LLMs) are artificial intelligence (AI) systems trained on extensive textual data such as books, articles and online content, with human-like capabilities such as providing summaries, translations and answers to questions. In 2018, the first LLM, generative pre-trained transformer 1 (GPT-1), was released by Open AI.<sup>5</sup> This has been followed by the release of many other LLMs by big tech companies, such as Meta's Llama and Google's Gemini.<sup>6,7</sup> LLMs can be improved through 'training' the models by providing questions and answers or other

**Table 1.** Overview of features of the six LLMs used

Feature	ChatGPT (GPT4o)	Gemini	Copilot	Mistral Large 2	Claude 3.5 Sonnet	Llama 3.1
Developer	OpenAI	Google DeepMind	Microsoft	Mistral AI	Anthropic	Meta
Country of origin	USA	USA	USA	France	USA	USA
Training data Sources	Web pages, books, articles and licensed datasets	Web pages, books, scientific papers and proprietary data	Code repositories (e.g. GitHub), documentation and open-source projects	Web pages, books and multilingual datasets	Web pages, books and curated datasets for ethical alignment	Web pages, books and open-source datasets
Primary use cases	General-purpose AI, chatbots, coding, Q&A	Multimodal tasks, research, creativity	Code generation, developer assistance	General-purpose AI, multilingual tasks	General-purpose AI, ethical AI, Q&A	Research, open-source applications
Strengths	High accuracy, large context window	Multimodal capabilities, strong reasoning	Excellent for coding, integrates with IDEs	Efficient, multilingual, lightweight	Ethical alignment, strong reasoning	Open-source, customizable, cost-effective
Open source	No	No	No	Partially	No	Yes
Performance	State-of-the-art for general tasks	Strong in multimodal and creative tasks	Best for coding and developer tasks	Efficient for multilingual tasks	Strong in ethical reasoning and Q&A	Good for research and customization

IDE, integrated development environment.

(updated) specific information.<sup>8</sup> LLMs have been increasingly employed to assist clinical diagnosis, clinical decision support, virtual medical assistants and health chatbots, as well as automated medical report synthesis.<sup>9</sup> LLMs have performed well on medical examinations and the selection of antidepressants, as well as supporting clinical decision-making and identification of drug-drug interactions<sup>10–13</sup>

A recent study showed the applicability of LLMs to support treatment suggestions for patients with resistant infections.<sup>14,15</sup> A correspondence to the *Lancet Infectious Diseases*, with one of the nine US-based scenarios not solely based within the hospital, presented overall good responses from ChatGPT but also identified deficits in situational awareness, inference and consistency.<sup>16</sup> No study based in general practice was identified. This study aimed to understand if LLMs provide accurate advice on antibiotic treatment and patient management in a general practice setting.

## Methods

Twenty-four vignettes were selected from the literature, which included upper respiratory tract infection, pneumonia, bronchitis, pharyngitis, urinary tract infection (UTI), COPD Gold II and III exacerbation, otitis media, cellulitis, sinusitis, acne, asymptomatic bacteriuria (ASB), wound, sore throat, varicella and dental infections.<sup>17–24</sup> The decision to prescribe an antibiotic was guided by the description given in the vignette. Four countries were identified: USA, UK and Ireland, each with English and online prescribing guidelines, and Norway, as a different language option but with easy-to-find and accessible online guidelines. For each vignette, a country-specific treatment plan (antibiotic yes/no, which one, duration,

other information) was based on the national guidelines for antibiotic prescribing: IDSA; NICE prescribing guidelines; the Irish antibiotic prescribing guidelines; and the Norwegian antibiotics prescribing guidelines in primary care.<sup>25–28</sup> Based on the vignettes, 12 out of 24 advised to prescribe an antibiotic.

In each country, a practising GP, without a specific interest in antibiotic prescribing, was approached and asked to look at each vignette and provide a treatment plan as they would in daily clinical practice.

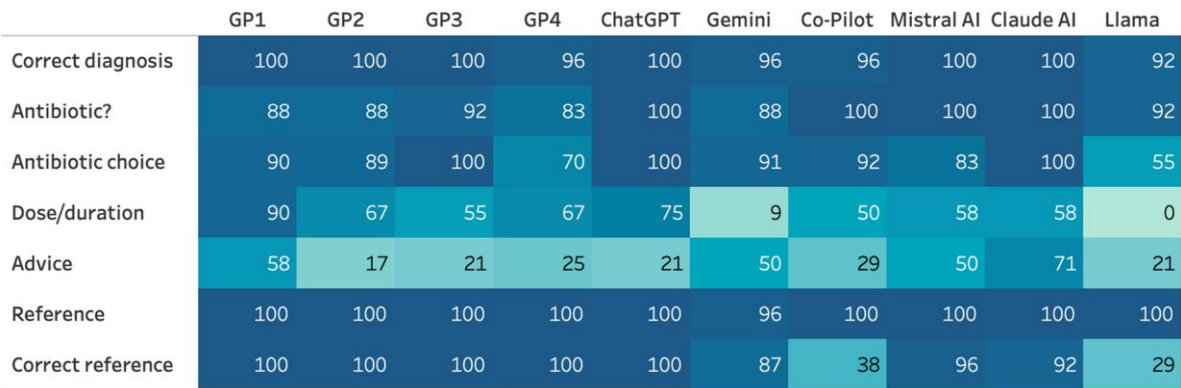
Six LLMs [ChatGPT (GPT-4o), Gemini, Copilot, Mistral Large 2, Claude 3.5 Sonnet and Llama 3.1 (8 billion parameter model)] were used (Table 1). ChatGPT is a chatbot, based on OpenAI's most recent generative pre-trained transformer and known for its advanced capabilities for smoother human-computer interaction and generation of output in a very short time.<sup>29</sup> Google's Gemini, previously known as Bard, is designed to integrate with Google's infrastructure but may be less flexible.<sup>7</sup> Mistral AI is designed to handle very complex tasks and more specialized use within smaller contexts but may not perform as well with more creative tasks. Anthropic's Claude is able to capture sophisticated instructions, write content with high quality and is focused on safer and more understandable AI outputs.<sup>30,31</sup> Llama 3.1 is Meta's LLM and works best within this environment. Llama is focused on flexibility and accessibility and has three versions: 8 billion (used in this study); 70 billion; and 405 billion.<sup>32</sup>

A standardized approach was taken to prompt the LLMs, mirroring the GPs but with inclusion of country. All questions were presented in English. The baseline comparison was the US guidelines (and the vignettes without country specification). Each LLM and GP was prompted with the same questions: diagnosis; yes/no antibiotic; if yes, antibiotic choice, dose and duration; any advice provided; and which guidelines were used (if any). The prompts were run on each LLM between 7 August and 16 August 2025, while the GPs provided their advice between 7 August and 10 August 2025.

**Table 2.** Overview of percentage correct answers for the 24 vignettes by GP and country

					USA					
	GP 1	GP 2	GP 3	GP 4	ChatGPT	Gemini	Copilot	Mistral AI	Claude	Llama
Correct diagnosis	100	100	100	96	100	96	96	100	100	92
Antibiotic?	88	88	92	83	100	88	100	100	100	92
Antibiotic choice	75	67	92	58	100	83	92	83	100	50
Dose/duration	75	50	50	50	75	8	50	58	50	0
Advice	58	0	21	25	21	50	29	50	71	21
Reference	100	100	100	100	100	96	100	100	100	100
Correct reference	100	100	100	100	100	83	38	96	92	29
Overall mean	83	67	76	69	83	68	68	81	85	49

GPs were compared with their own country’s guidelines.



**Figure 1.** Heatmap presenting differences between GPs and LLMs in response to vignettes. Darker colour represents higher accuracy.

The percentage of accurately provided prescribing/treatment options was assessed for the GPs and LLMs by comparing with the national guidelines.

Additional evaluations were for hallucination, which occurs when the LLM generates a response that appears to be correct, but is useless and incorrect, and toxicity, which is defined as the presence of disrespectful, rude, unreasonable or aggressive comments.<sup>33,34</sup> Hallucination was evaluated with the BERTScore, which evaluates the responses from LLMs to their prompts for relevance and out-of-context answers. Data leakage, which in this context can be defined as the ‘leakage of user’s input data’, was checked with Python (‘detect entities’ in Spacey library).<sup>35</sup>

### Results

A total of seven questions (diagnosis, antibiotic yes/no, choice of antibiotics, dose and duration, advice to patient, reference yes/no, which reference) prompted to six LLMs and four GPs were included.

When considering the GPs’ answers to the vignettes, their accuracy in relation to diagnosis (96%–100%) and yes/no antibiotic prescribing (83%–92%) was high (Table 2 and Figure 1). Once this decision to prescribe or not was made, between 70% and 100% identified the right antibiotic according to their national guidelines, with the dose/duration provided being less accurate (55%–90%). All GPs were aware of their country’s antibiotic

guidelines and could provide the right link for this. Advice to patients was not provided or was not added to the vignettes in 17%–58%. The overall mean score for the GPs was 79%.

In comparison with the GPs, the LLMs’ answers to the standard vignettes showed that each LLM scored highly in relation to diagnosis (92%–100%), antibiotic prescribing (88%–100%) and the choice of antibiotic (55%–100%), which showed similar variation as the GPs. However, the dose and duration provided by the LLMs varied from 0% to 75%. Between 21% and 71% of the LLMs provided advice; the best advice was provided by Claude (71%). References were provided, as asked in the prompts; however, the appropriateness of the references varied widely, from 29% to 100%.

Comparison of the LLMs using the country-specific vignettes, showed that for the standard (USA) the highest scores were observed for Claude (87%). ChatGPT scored similarly highly (83%), except for providing advice (21%). ChatGPT was the only model able to retrieve all the correct US references for its answers, followed by Mistral AI (96%) and Claude (92%). When comparing between the countries (Table 3 and Figure 2), diagnosis, antibiotic decision and choice were generally well provided; however, the application of the national guidelines, considering dose and duration, providing advice and identifying the appropriate reference were poor. The UK guidelines were better referenced than the Irish guidelines, in particular for ChatGPT (96% versus 9%),

**Table 3.** Overview of percentage correct answers for the 24 vignettes, prompted with country, by each LLM compared with the specific guidelines of each country: Ireland, UK and Norway

	Ireland						UK						Norway					
	ChatGPT	Gemini	Copilot	Mistral AI	Claude	Llama	ChatGPT	Gemini	Copilot	Mistral AI	Claude	Llama	ChatGPT	Gemini	Copilot	Mistral AI	Claude	Llama
Correct diagnosis	100	92	96	100	100	88	100	100	92	92	100	96	96	96	96	96	96	96
Antibiotic?	100	88	96	100	100	83	96	100	92	96	100	88	96	92	96	96	96	96
Antibiotic choice	83	92	83	83	100	58	75	75	50	75	100	50	58	50	58	67	50	50
Dose/duration	17	8	0	33	50	0	25	8	8	42	92	8	0	0	0	17	8	8
Advice	33	58	25	33	96	17	38	46	38	58	67	38	33	46	25	71	83	21
Reference	96	96	100	100	92	100	100	96	100	100	100	100	100	96	100	100	100	100
Correct reference	8	38	8	8	54	4	96	71	8	88	92	4	13	8	13	8	8	0
Overall mean	56	63	52	60	82	44	72	66	49	76	92	48	50	49	49	60	58	46

Claude (92% versus 59%) and Gemini (71% versus 38%). However, for the Norwegian guidelines, all LLMs performed very poorly (0%–13%). The overall mean of the LLMs, considering all vignettes and countries, was highest for Claude (80%), followed by Mistral AI (70%), ChatGPT (65%) and Gemini, Copilot and Llama (62%, 55% and 47%, respectively).

The BERTScore (hallucination) ranged between 0.68 and 0.82, reflecting moderate semantic similarity and indicating that the models are not hallucinating. However, human expertise was still needed to determine whether the antibiotic treatment proposed was in line with the national prescribing guideline. No toxicity (threats, insults or attacks on identity) was observed in any of the responses obtained from the LLMs. However, data leakage (any leakage of personal information) was observed in all the responses of the vignettes that included a patient’s name or their initials.

**Discussion**

Compared with GPs, LLMs evaluated against US guidelines, show high reliability in the decision to prescribe an antibiotic and the choice of antibiotic, as well as identifying the diagnosis described in the vignettes, whereas the dose/duration was less accurate. When comparing the performance of LLMs between countries, and against the country’s specific guidelines, LLMs often fail to provide accurate references to national guidelines and recommendations. GPs were reliable in diagnosis and the decision to prescribe, but importantly, more reliable in referencing and applying national guidelines.

The vignettes were written for antibiotic prescribing and for 12 out of 24 vignettes (50%) an antibiotic was indicated. For the vignettes where prescribing was indicated, accuracy in the choice of antibiotic was similar for LLMs and GPs, irrespective of the country, as the type of antibiotic would be relatively similar for different countries, due to their range and mechanism of action.<sup>36</sup>

The accuracy of LLMs reduces when more detailed (country-specific) information is required, which suggests that LLMs are largely trained on US information, as this was the most accurate model. However, information on the sources of data accessed by LLMs is not easily available, or has not been released.<sup>37</sup>

A strength of some LLMs was the inclusion of advice, in particular Claude, even though this has to be interpreted with caution. GPs may not have expanded on advice in their responses on the vignettes, which in daily clinical practice would be guided by the patient, and their questions. Also, the appropriateness of LLMs providing advice was challenged in a comparative vignette study with nurses, which showed that LLMs had a tendency towards over-triaging and too much information leading to indecisiveness.<sup>12</sup> Furthermore, the importance of human interpretation was clear as the LLM answers had to be interpreted in line with the national guidelines and checked individually. Human expertise regarding the application of the national guidelines may rely on the nuance and interpretation of experienced GPs; however, the answers regarding advice to patients did not provide enough evidence about this point. Assessment of advice, together with follow-up of the outcome for patients or, alternatively, the simultaneous application of LLMs in real-life consultations would be a logical next step.

	UK						Norway					
	ChatGPT	Gemini	Co-Pilot	Mistral AI	ClaudeAI	Llama	ChatGPT	Gemini	Co-Pilot	Mistral AI	ClaudeAI	Llama
Correct diagnosis	100	100	92	92	100	96	96	96	96	96	96	96
Antibiotic?	96	100	92	96	100	88	96	92	96	96	96	96
Antibiotic choice	82	75	50	82	100	58	58	50	58	73	55	50
Dose/duration	27	8	8	45	92	8	0	0	0	18	9	8
Advice	38	46	38	58	67	38	33	46	25	71	83	21
Reference	100	96	100	100	100	100	100	96	100	100	100	100
Correct reference	96	71	8	88	92	4	13	9	13	8	8	0

	US						Ireland					
	ChatGPT	Gemini	Co-Pilot	Mistral AI	ClaudeAI	Llama	ChatGPT	Gemini	Co-Pilot	Mistral AI	ClaudeAI	Llama
Correct diagnosis	100	96	96	100	100	92	100	92	96	100	100	88
Antibiotic?	100	88	100	100	100	92	100	88	96	100	100	83
Antibiotic choice	100	91	92	83	100	55	83	100	83	83	100	64
Dose/duration	75	9	50	58	58	0	17	9	0	33	50	0
Advice	21	50	29	50	71	21	33	58	25	33	96	17
Reference	100	96	100	100	100	100	96	96	100	100	92	100
Correct reference	100	87	38	96	92	29	9	38	8	8	59	4

Figure 2. Overview of accuracy of LLMs in response to vignettes—country comparison.

A wide variety was observed between different LLMs, reflecting the variety in the training data, as well as the contextual interpretation and model capabilities. More studies compare the quality of answers between LLMs, and between LLMs and human decision.<sup>10,38</sup> Depending on the context, training data and model capacity, LLMs performed well, though often suffering from a lack of interpretation, incorrect statements and a lack of references. ChatGPT has been shown to recognize clinically important factors when explicit information was provided but missed relevant issues in scenarios of increasing complexity.<sup>16</sup> In the presented comparison, however, the GPs, probably due to a more nuanced interpretation and inherent understanding of national guidelines, outperformed LLMs, except for Claude. LLMs can be improved in this area by more training data but also through algorithms that apply to country-specific guidelines and/or other factors such as the inclusion of different languages.

Of particular interest in relation to LLMs is the concerns in relation to data privacy and security, and a tailored approach to regulatory oversight has been suggested previously.<sup>37,39</sup> From our findings, some leakage occurred, when patient information was included in the vignettes. This should be a real concern for GPs and a warning never to include patient details when using LLMs.<sup>40</sup>

In a recent, but not yet reviewed, paper on how LLMs can influence medical decision-making, clinicians were asked to assess triage, risk and treatment before and after receiving advice

generated by ChatGPT.<sup>41</sup> Clinicians were willing to change their decisions based on the AI assistance, and improvements could be made. Considering our results, in particular the lack of nuance and application of national guidelines, the antibiotic prescribing decision should remain with the GP; however, LLMs can be used to provide advice.

The study showed the potential of LLMs to suggest when antibiotic treatment is appropriate, potential antibiotic agents and advice. However, GPs or other prescribers of antibiotics should interpret and use the information with caution and awareness of its limitations. Whereas the results have shown that LLMs can support antibiotic prescribing, the identification and implementation of national guidelines is suboptimal and there are potential concerns in relation to patient privacy. Improvements can be made by optimizing model training and the application of relevant algorithms.

---

### Funding

This work was funded by grant number RL-20200-03, Health Research Board, Ireland, Research Leader Award 2020. D.A. was funded by this grant and is a SPHeRE scholar.

---

### Transparency declarations

None to declare.



## References

- 1 Courtenay M, Castro-Sanchez E, Fitzpatrick M *et al*. Tackling antimicrobial resistance 2019–2024 – the UK's five-year national action plan. *J Hosp Infect* 2019; **101**: 426–7. <https://doi.org/10.1016/j.jhin.2019.02.019>
- 2 Llor C, Bjerrum L. Antimicrobial resistance: risk associated with antibiotic overuse and initiatives to reduce the problem. *Ther Adv Drug Saf* 2014; **5**: 229–41. <https://doi.org/10.1177/2042098614554919>
- 3 Vellinga A, Luke-Currier A, Garzon-Orjuela N *et al*. Disease-specific quality indicators for outpatient antibiotic prescribing for respiratory infections (ESAC quality indicators) applied to point prevalence audit surveys in general practices in 13 European countries. *Antibiotics (Basel)* 2023; **12**: 572. <https://doi.org/10.3390/antibiotics12030572>
- 4 Palin V, Mölter A, Belmonte M *et al*. Antibiotic prescribing for common infections in UK general practice: variability and drivers. *J Antimicrob Chemother* 2019; **74**: 2440–50. <https://doi.org/10.1093/jac/dkz163>
- 5 Thirunavukarasu AJ, Ting DSJ, Elangovan K *et al*. Large language models in medicine. *Nat Med* 2023; **29**: 1930–40. <https://doi.org/10.1038/s41591-023-02448-8>
- 6 Meta. Introducing LLaMA: a foundational, 65-billion-parameter large language model. 2023. <https://ai.meta.com/blog/large-language-model-llama-meta-ai/>.
- 7 Manyika J. An overview of the Gemini App. 2024. <https://gemini.google/overview-gemini-app.pdf>.
- 8 Clusmann J, Kolbinger FR, Muti HS *et al*. The future landscape of large language models in medicine. *Commun Med (Lond)* 2023; **3**: 141. <https://doi.org/10.1038/s43856-023-00370-1>
- 9 Nazi ZA, Peng W. Large language models in healthcare and medical domain: a review. *Informatics (MDPI)* 2024; **11**: 57. <https://doi.org/10.3390/informatics11030057>
- 10 Perlis RH. Research letter: Application of GPT-4 to select next-step antidepressant treatment in major depression. *medRxiv* 2023; 2023.04.14.23288595. <https://doi.org/10.1101/2023.04.14.23288595>
- 11 Perlis RH, Goldberg JF, Ostacher MJ *et al*. Clinical decision support for bipolar depression using large language models. *Neuropsychopharmacology* 2024; **49**: 1412–6. <https://doi.org/10.1038/s41386-024-01841-2>
- 12 Saban M, Dubovi I. A comparative vignette study: evaluating the potential role of a generative AI model in enhancing clinical decision-making in nursing. *J Adv Nurs* 2024; <https://doi.org/10.1111/jan.16101>
- 13 Alshehri BMJ, Kraiem N, Sakly H *et al*. Enhancing medication safety with large language models: advanced detection and prediction of drug-drug interactions. 2024 *IEEE 7th International Conference on Advanced Technologies, Signal and Image Processing (ATSIP), Sousse, Tunisia, July 2024*, 547–52. <https://doi.org/10.1109/ATSIP62566.2024.10638993>
- 14 Lv J, Deng S, Zhang L. A review of artificial intelligence applications for antimicrobial resistance. *Biosaf Health* 2021; **3**: 22–31. <https://doi.org/10.1016/j.bsheat.2020.08.003>
- 15 Chakraborty C, Pal S, Bhattacharya M *et al*. ChatGPT or LLMs can provide treatment suggestions for critical patients with antibiotic-resistant infections: a next-generation revolution for medical science? *Int J Surg* 2024; **110**: 1829–31. <https://doi.org/10.1097/JIS.0000000000000987>
- 16 Howard A, Hope W, Gerada A. ChatGPT and antimicrobial advice: the end of the consulting infection doctor? *Lancet Infect Dis* 2023; **23**: 405–6. [https://doi.org/10.1016/S1473-3099\(23\)00113-5](https://doi.org/10.1016/S1473-3099(23)00113-5)
- 17 Delory T, Maillard A, Tubach F *et al*. Appropriateness of intended antibiotic prescribing using clinical case vignettes in primary care, and related factors. *Eur J Gen Pract* 2024; **30**: 2351811. <https://doi.org/10.1080/13814788.2024.2351811>
- 18 Ung E, Czarniak P, Sunderland B *et al*. Assessing pharmacists' readiness to prescribe oral antibiotics for limited infections using a case-vignette technique. *Int J Clin Pharm* 2017; **39**: 61–9. <https://doi.org/10.1007/s11096-016-0396-0>
- 19 Tran J, Danchin M, Pirotta M *et al*. Management of sore throat in primary care. *Aust J Gen Pract* 2018; **47**: 485–9. <https://doi.org/10.31128/AJGP-11-17-4393>
- 20 Kistler CE, Beeber A, Becker-Dreps S *et al*. Nursing home nurses' and community-dwelling older adults' reported knowledge, attitudes, and behavior toward antibiotic use. *BMC Nurs* 2017; **16**: 12. <https://doi.org/10.1186/s12912-017-0203-9>
- 21 Fergie J, Pawaskar M, Veeranki P *et al*. Recognition & management of varicella infections and accuracy of antimicrobial recommendations: case vignettes study in the US. *PLoS One* 2022; **17**: e0269596. <https://doi.org/10.1371/journal.pone.0269596>
- 22 Schneider-Smith EG, Suda KJ, Lew D *et al*. How decisions are made: antibiotic stewardship in dentistry. *Infect Control Hosp Epidemiol* 2023; **44**: 1731–6. <https://doi.org/10.1017/ice.2023.173>
- 23 Taylor LN, Wilson BM, Singh M *et al*. Syndromic antibiograms and nursing home clinicians' antibiotic choices for urinary tract infections. *JAMA Netw Open* 2023; **6**: e2349544. <https://doi.org/10.1001/jamanetworkopen.2023.49544>
- 24 Gidengil CA, Linder JA, Beach S *et al*. Using clinical vignettes to assess quality of care for acute respiratory infections. *Inquiry* 2016; **53**: 0046958016636531. <https://doi.org/10.1177/0046958016636531>
- 25 IDSA. IDSA Practice Guideline. 2024. [https://www.idsociety.org/practice-guideline/practice-guidelines/#/+0/date\\_na\\_dt/desc/](https://www.idsociety.org/practice-guideline/practice-guidelines/#/+0/date_na_dt/desc/).
- 26 NICE. NICE Guidance—Conditions and Diseases. 2024. <https://www.nice.org.uk/guidance/conditions-and-diseases>.
- 27 Health Service Executive (HSE). Antibiotic Prescribing—Conditions and Treatments. 2024. <https://www.hse.ie/eng/services/list/2/gp/antibiotic-prescribing/conditions-and-treatments/list-of-conditions-and-treatments.html>.
- 28 Norwegian Directorate of Health. Antibiotika i primærhelsetjenesten. Antibiotika i primærhelsetjenesten. 2024. <https://www.helsedirektoratet.no/retningslinjer/antibiotika-i-primærhelsetjenesten>.
- 29 OpenAI. Hello GPT-4o. <https://openai.com/index/hello-gpt-4o/>.
- 30 Mistral. Frontier AI. In your hands. 2024. <https://mistral.ai/>.
- 31 Anthropic. Claude 3.5 Sonnet. 2024. <https://www.anthropic.com/news/claude-3-5-sonnet>.
- 32 Meta. Introducing Llama 3.1: our most capable models to date. 2024. <https://ai.meta.com/blog/meta-llama-3-1/>.
- 33 Farquhar S, Kossen J, Kuhn L *et al*. Detecting hallucinations in large language models using semantic entropy. *Nature* 2024; **630**: 625–30. <https://doi.org/10.1038/s41586-024-07421-0>
- 34 Taleb M, Hamza A, Zouitni M *et al*. Detection of toxicity in social media based on natural language processing methods. 2022 *International Conference on Intelligent Systems and Computer Vision (ISCV), Fez, Morocco, 18 May 2022*. <https://doi.org/10.1109/ISCV54655.2022.9806096>
- 35 Chen C, Wu Z, Lai Y *et al*. Challenges and remedies to privacy and security in AIGC: exploring the potential of privacy computing, blockchain, and beyond. *arXiv* 2023; 2306.00419. <https://doi.org/10.48550/arXiv.2306.00419>
- 36 Shaw L. Representing antibiotic relationships using measurements of efficacy against clinical isolates. *Wellcome Open Res* 2019; **4**: 86. <https://doi.org/10.12688/wellcomeopenres.15304.2>
- 37 Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med* 2023; **6**: 120. <https://doi.org/10.1038/s41746-023-00873-0>
- 38 He Z, Bhasuran B, Jin Q *et al*. Quality of answers of generative large language models versus peer users for interpreting laboratory test results

for lay patients: evaluation study. *J Med Internet Res* 2024; **26**: e56655. <https://doi.org/10.2196/56655>

**39** 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. <https://doi.org/10.1136/bmj-2023-078538>

**40** Rydzewski NR, Dinakaran D, Zhao SG et al. Comparative evaluation of LLMs in clinical oncology. *NEJM AI* 2024; **1**: 10.1056/aioa2300151. <https://doi.org/10.1056/aioa2300151>

**41** Goh E, Bunning B, Khoong E et al. ChatGPT influence on medical decision-making, bias, and equity: a randomized study of clinicians evaluating clinical vignettes. *medRxiv* 2023; <https://doi.org/10.1101/2023.11.24.23298844>