Large language models for antibiotic prescribing and antimicrobial stewardship

Daniele Roberto Giacobbe[1,2]*, MD, PhD; Cristina Marelli[2], BS; Bianca La Manna[3], BS; Donatella Padua[4], PhD; Alberto Malva[5], MD; Sabrina Guastavino[6], PhD; Alessio Signori[7], PhD; Sara Mora[8], PhD; Nicola Rosso[8], PhD; Cristina Campi[6,9], PhD; Michele Piana[6,9], PhD; Ylenia Murgia[3], BS; Mauro Giacomini[3], PhD; Matteo Bassetti[1,2], MD, PhD

[1] Department of Health Sciences (DISSAL), University of Genoa, Genoa, Italy

[2] UO Clinica Malattie Infettive, IRCCS Ospedale Policlinico San Martino, Genoa, Italy

[3] Department of Informatics, Bioengineering, Robotics and System Engineering (DIBRIS), University of Genoa, Genoa, Italy

[4] UniCamillus - International University of Health and Medical Science

[5] Italian Interdisciplinary Society for Primary Care

[6] Department of Mathematics (DIMA), University of Genoa, Genoa, Italy

[7] Section of Biostatistics, Department of Health Sciences (DISSAL), University of Genoa, Genoa, Italy

[8] UO Information and Communication Technologies, IRCCS Ospedale Policlinico San Martino, Genoa, Italy

[9] Life Science Computational Laboratory (LISCOMP), IRCCS Ospedale Policlinico San Martino, Genoa, Italy

*** Corresponding author:**

Daniele Roberto Giacobbe, MD, PhD

Department of Health Sciences (DISSAL)

University of Genoa

Via A. Pastore 1 – 16132 Genoa, Italy

Email address: danieleroberto.giacobbe@unige.it

Abstract

With the advent of large language models (LLMs) to support healthcare decisions, researchers are exploring their role in antibiotic prescribing. Currently, there is a lack of standardization in research on using LLMs for this purpose, necessitating more efforts to identify biases and misinformation in outputs from black box models. Additionally, antimicrobial prescribing requires balancing optimal treatment for individual patients with reducing antimicrobial resistance globally, aligning with antimicrobial stewardship core objectives. This challenge demands dedicated and standardized guidance to measure the appropriateness of LLMs' suggestions. Educating future medical professionals on these aspects is crucial for ensuring the proper use of LLM-based support in antibiotic prescribing, providing a deeper understanding of their strengths and limitations. Antimicrobial resistance is projected to cause up to 10 million deaths annually by 2050, surpassing deaths from other widespread diseases. Thus, achieving balanced and safe use of LLM support in antibiotic prescribing and antimicrobial stewardship initiatives is an opportunity not to be missed.

Key words: artificial intelligence; machine learning; neural networks; natural language processing; antimicrobial resistance; antimicrobial stewardship.

Introduction

Imagine you are a hospital-based infectious diseases specialist receiving a consultation request from another ward. When you first read the request for consultation on your computer screen, an intelligent artificial assistant, leveraging large language models (LLMs) technology, has already prepared a coherent summary of the patient's medical and microbiological history, relevant laboratory and instrumental test results, and the evolution of their acute phase conditions in the last few days.[1-3] This summary immediately provides you with an initial idea about what to do, without laboriously spending many minutes searching for information across clinical notes in the patient's clinical chart.

Subsequently, you go to the other ward to visit the patient and gather additional information from the patient and their treating physicians. During the consultation, your intelligent artificial assistant can: (*i*) directly register and summarize the additional information provided by the patient and their treating physicians; (*ii*) suggest additional relevant questions to be posed. After coherently merging the already known information from the patient's history and test results with the new information collected during the consultation, your artificial intelligent assistant can explicitly offer some suggestions for your revision (e.g., prescription of a given antibiotic at a certain dosage and for a certain duration), supported by reasonable, summarized explanations.

This is only a hypothetical example of how LLMs could aid physicians in the near future in prescribing antibiotics, likely not exhaustive of all potential applications of LLMs for this purpose.[3-8] Since the advantages (above all, dramatic reduction of repetitive tasks for clinicians, thereby making time for more sophisticated clinical reasoning) of introducing LLMs in daily clinical practice could be transformative in healthcare, and considering that profound implementation of LLMs within electronic health records has already been announced[9], a thorough understanding of both potential advantages and relevant limitations is essential for current and future clinicians who will very likely deal with this emerging technology in their daily clinical practice.[1,10]

In this perspective, we focus on the potential advantages and limitations of introducing LLMs to support the prescription of antibiotics, both in terms of improving the efficacy and safety of the therapeutic

approach to the single patient and in terms of appropriate use of antibiotics in line with antimicrobial stewardship principles (i.e., responsible and appropriate prescription of antibiotics at both patient and global levels, to ensure availability in the present and preservation of efficacy in future populations[11]). Notably, these are complex medical tasks, requiring dedicated medical expertise and involving a multi-component and dynamic clinical reasoning process (Box 1).[12-16]

Brief History of LLMs and How They Work

Natural language processing (NLP) studies how to elaborate and produce natural human language through a computer, including healthcare-related text.[17-19] NLP is considered part of the domain of artificial intelligence (AI) since it tries to reproduce tasks typically performed by humans. Consequently, the evolution of language models progresses alongside the implementation of dedicated AI algorithms. The first NLP models were rule-based systems, relying on pre-written rules defined by domain experts. These models performed well on specific simple tasks but poorly on unseen data.

This limitation was overcome with the application of neural networks (NNs) for this task. NNs are machine learning algorithms designed to emulate the biological architecture of the human brain, i.e., networks of interconnected 'nodes' capable of transferring information. NNs are considered "black box" models because the composition and computations of features within the initial (input) layer and the final (output) layer may be partly or sometimes totally unclear to data scientists building and testing the model, as well as to physicians assessing how a NN model arrived at a given output, e.g., suggesting a certain antibiotic prescription.[20,21] The need to improve understanding of how such models arrive at their outputs/predictions, fundamental in healthcare to reduce the risk of overlooking biases and misinformation possibly perpetuated by black box models, has led to the expansion of research on explainable AI.[22,23]

Regarding the task of human language recognition, recurrent NNs (RNNs), which are directed graphs that process sequential inputs, and long short-term memory (LSTM) NNs, which can store past information, improved prediction skills connected to text decoding. However, RNNs and LSTMs were proven unable to make accurate predictions over extended sequences of text.

In late 2017, Vaswani and colleagues introduced transformers, NNs with architectures able to handle long-range dependencies.[24] Both RNNs and transformers rely on attention mechanisms, methods which assign "weights" to each node to achieve better predictions and decisions by considering the importance of a word in context. However, while RNNs achieve this within a selected context window, transformers exploit self-attention and calculate the weights considering all the words in a sentence. With this solution, NLP models started to perform not only as decoders of textual information but as encoders too, like BERT, which is capable of producing natural language answers according to user input.[25]

Transformers show better generalization and prediction ability than previous NLP models but are limited by the lack of large-scale datasets and adequate computational resources. They laid the basis for the advent of LLMs, along with the introduction of graphic processing units that increased the performance of mathematical calculations, allowing the processing of huge quantities of data. Public awareness of LLMs was maximized after the release of OpenAI's GPT-3 in 2020.[26] Technically, LLMs are AI algorithms that can understand, extract, summarize, predict, and generate human-like text based on patterns and relationships learned from vast amounts of data. LLMs are considered "large" because they are trained on massive amounts of data and comprise a huge number of learnable parameters, with popular LLMs reaching hundreds of billions of parameters. This allows them to produce more accurate responses than previously proposed NLP models. Currently, some main companies releasing LLMs are OpenAI, NVIDIA/Microsoft, Meta, Google, Cohere, Anthropic (public benefit company), and EleutherAI (non-profit company). For interested readers, more details on the components and functioning architecture of LLMs are available in Box 2.[27-30]

Current Literature on the Use of LLMs for Supporting Antibiotic Prescribing

There are already some examples in the scientific literature on the use of general-purpose or domain-specific LLMs, or of chatbots based on LLMs, for supporting antibiotic prescriptions. For example, the performance of generative pretrained transformer (GPT) was recently assessed by Maillard and colleagues using a GPT-4-based chatbot (ChatGPT-4) to provide appropriate antimicrobial therapy recommendations in 44 retrospective cases of bloodstream infection (BSI).[31] In this study, ChatGPT-4 was provided with all the (anonymized) information available to clinicians who performed the consultation (without the aid of LLMs as

per standard clinical practice), and the chatbot's performance in terms of appropriateness was classified (appropriate vs. inappropriate) by infectious diseases specialists not involved in the care of that given patient. Standardized prompts were provided (once for each case) to ChatGPT-4, contextualizing the need for a comprehensive response regarding the management of a specific case of bloodstream infection in a French hospital, to be provided as if ChatGPT was the infectious diseases specialist consulting on that given patient. Appropriateness was measured according to local and international guidelines. Furthermore, recommendations provided by ChatGPT-4 were also classified in terms of their harmfulness (potentially harmful for patients vs. not harmful). Overall, appropriateness of suggestions for empirical and targeted therapy was 64% and 36%, respectively, whereas 2% and 5% of empirical and targeted prescriptions, respectively, were considered potentially harmful. For example, a potentially harmful suggestion for empirical therapy was narrowing the spectrum of antibiotic therapy to a regimen not covering Gram-negative bacteria in a patient with febrile neutropenia while waiting for culture results, whereas for targeted therapy a potentially harmful suggestion was de-escalating from cefepime and vancomycin to cloxacillin in a neutropenic patient with a non-bacteremic infection by *Staphylococcus aureus* and concomitant ongoing sepsis of suspected unrelated origin.[31]

In another study, Fisch and colleagues evaluated LLMs' adherence to good clinical practice principles and guidelines from the Infectious Diseases Society of America (IDSA) and the European Society of Clinical Microbiology and Infectious Diseases (ESCMID) when providing management indications for a clinical case (hypothetical) of pneumococcal meningitis originating from mastoiditis.[32] No definite diagnosis was provided to LLMs. Several LLMs (Llama, Bard, Claude-2, PaLM, Bing, GPT-3.5, GPT-4) were presented with the same case thrice, and, besides appropriateness of recommendations, the heterogeneity of the suggested management provided by the same LLM across the three different sessions was evaluated. Regarding prompting, LLMs were asked to act as expert medical assistants to suggest a junior doctor how to manage a 52-year-old patient with headache and confusion, with subsequent conversation on the case with a more specific illustration of signs and symptoms. Among questions inherent to antibiotic prescribing, LLMs were evaluated based on: (*i*) whether or not prescription of antibiotics was necessary; (*ii*) whether, if antibiotic administration was suggested, the type and dosages of suggested empirical antibiotics were in line with IDSA

and ESCMID guidelines. Overall, a total of 21 responses were collected for each question, from the three different sessions for each of the seven evaluated LLMs. The need for rapid antibiotic administration was correctly recognized in 81% of cases. The correct type of empirical antibiotics (in line with IDSA and ESCMID guidelines) was suggested in 38% of cases, with correct dosages (whenever correct antibiotics were suggested) being suggested in almost 90% of cases. Some misleading statements were also identified. For example, hallucinations included the presence of Kernig's sign and a stiff neck (not depicted in the presented case), and misleading interpretations included recognizing herpes ophthalmicus instead of bacterial meningitis. Heterogeneity was observed for all models during the three different sessions, impacting the rate of adherence to guidelines. Among evaluated LLMs, ChatGPT-4 provided the most consistent responses across the three sessions.[32]

In another study, the performance of the LLM-based chatbots ChatGPT-3.5 and ChatGPT-4 in replying to different questions regarding antibiotic prophylaxis in patients undergoing spine surgery was evaluated against the North American Spine Society (NASS) guidelines, which served as the reference standard for evaluating the accuracy of responses.[33] Prompts were formulated exactly as 16 original questions of the NASS guidelines (with the addition of a reference to spine surgery whenever not included in the questions, to provide the necessary context present in the guidelines but not in isolated prompts). The accuracy of responses was 63% (10/16) and 81% (13/16) for ChatGPT-4 and ChatGPT-3.5, respectively. ChatGPT-3.5 showed a tendency towards overconfident but potentially erroneous or contradictory responses, whereas ChatGPT-4 showed an increased tendency to support its statements with references, including the NASS guidelines.[33]

Recently, Lai and colleagues assessed the accuracy and repeatability of responses provided by ChatGPT-3.5 to queries about *Helicobacter pylori*, including those regarding treatment of *H. pylori* infections (queries regarding treatment were six out of a total of 22 queries).[34] Regarding repeatability, the same question (prompt) was presented to ChatGPT 3.5 two weeks after the first session. Responses provided by ChatGPT 3.5 were independently assessed by two expert gastroenterologists using the following scoring system: (*i*) comprehensive (four points); (*ii*) correct but inadequate (three points); (*iii*) mixed correct/incorrect or outdated (two points); (*iv*) completely incorrect (one point). Confirmation vs. rejection

of repeatability between the two responses to the same query provided two weeks apart was also based on the independent judgment of two expert gastroenterologists (another expert with >20 years of experience in *H. pylori* infection was involved for the final decision in case of disagreement). Notably, responses regarding treatment showed the lowest score (mean 3.25, standard deviation ±0.48). Over 80% of these responses were rated as comprehensive (four points) or correct but inadequate (three points), but 16.6% were rated as mixed correct/incorrect or outdated (two points). Regarding repeatability, ChatGPT-3.5 provided similar responses between the two sessions in 95.2% of cases.[34]

In another recent paper, Chakraborty and colleagues asked two questions to ChatGPT (version not provided) regarding the management of antibiotic-resistant infections.[35] For the first question, ChatGPT was provided with susceptibility test results for several antibiotics without clinical context or bacterial genus and species. While ChatGPT appropriately suggested thorough evaluation of the patient's condition and consultation with an infectious diseases specialist, it also recommended meropenem without sufficient context, which could be inappropriate without more information. The second question was similar, with resistance to carbapenems included. Again, ChatGPT emphasized the need for more context and specialist consultation but recommended colistin, not aligning with recent guidelines for managing carbapenem-resistant Gram-negative infections, which no longer include colistin as a first-line therapy.[36,37] No other sessions were performed to assess response consistency to the same prompt.[35]

Finally, De Vito and colleagues recently evaluated ChatGPT-4's performance in responding to true/false and open-ended questions regarding clinical cases of bacterial infections, with susceptibility test results available, totaling 96 questions.[38] Experts in antibiotic prescribing formulated the questions, that were also posed to four senior residents and four infectious diseases specialists. Responses from humans and ChatGPT-4 were assessed by the experts (blinded to whether responses were from humans or ChatGPT-4) for accuracy and completeness. ChatGPT-4 showed similar accuracy to humans in true/false questions (approximately 70% correctness) and provided more complete and accurate responses to open-ended questions than human participants. However, ChatGPT-4 struggled with recognizing resistance mechanisms and tended not to prescribe recently approved antibiotics for multidrug-resistant Gram-negative infections,

favoring older, more toxic antibiotics such as polymyxins. ChatGPT-4 also tended to suggest longer-than-necessary antibiotic treatment durations compared to human participants.[38]

Discussion and Perspective

With the advent of LLMs to support healthcare decisions, researchers have begun exploring their role in supporting antibiotic prescribing, a crucial step in evaluating and improving the use for this purpose of LLM-based tools already accessible through the web or apps (Box 3).[31-35,38-41] Several fundamental points should be considered, based on the initial literature on this topic.

The first point is the lack of standardization in research on the use of LLMs to support antibiotic prescribing. Standardization is likely required in building prompts, the number of sessions in which the same prompt should be presented to a given LLM or LLM-based chatbot, how subsequent questions should be prepared and posed, and how to measure accuracy and consistency of responses. The term used to describe the comparison of responses to the same prompt varies across studies (e.g., consistency vs. repeatability). This heterogeneity is not unique to evaluating LLMs for antibiotic prescribing but also affects research on LLMs supporting healthcare decisions more generally. Initiatives such as the Chatbot Assessment Reporting Tool (CHART) project aim to improve the standardization of research methodology on using LLMs to support healthcare decisions, which could be fundamental in improving the generalizability and comparability of research findings on LLMs supporting antibiotic prescribing.[42]

The second point is the need to improve human ability to identify biases or misinformation in confident and convincing outputs from black box models, which may theoretically mislead even expert antibiotic prescribers when subtle errors or biases are perpetuated. Some authors advocate for relying on interpretable models only, arguing that the assumption of increased accuracy of black box models over interpretable models should not be taken for granted.[20] Nonetheless, while interpretable models should be preferred when their accuracy is similar to that of black box models, this may not always be the case for machine learning models working on unstructured data like LLMs.[21-23] This raises the issue of explainability of LLMs, or more importantly, of the degree of explainability in terms of accuracy/correctness of explanations that can be accepted for healthcare decisions. Defining and identifying a similar threshold would require

openness about datasets used for training and model architectures, and recognition potential biases in training data not to be perpetuated. To this aim, pre-processing of data must also ensure privacy preservation, grammatical correction of errors, and proper recognition of medical terms and abbreviations. Expert human evaluation of responses provided by LLMs during development and before/during implementation in clinical practice would also be crucial. Against this background, initiatives like the Translational Evaluation of Healthcare AI (TEHAI) have been taken with the aim to develop and standardize a comprehensive and multi-stage evaluation of AI models, including LLMs.[43,44] Finally, in our opinion, exploring dedicated design and metrics for randomized studies to assess LLMs' performance in clinical practice may also prove essential for evaluating, with the highest certainty of evidence, the efficacy and safety of LLMs or LLM-based chatbots in improving appropriate antibiotic prescriptions.

A third point specific to antimicrobial prescribing involves balancing the best possible treatment for individual patients with reducing antimicrobial resistance at a more global level, in line with antimicrobial stewardship core objectives. This peculiar challenge will require dedicated and standardized guidance for measuring the appropriateness of LLMs' suggestions for antimicrobial prescribing. Overall, all the above considerations necessitate a multidisciplinary approach to LLM development, approval, and clinical use for antibiotic prescribing and antimicrobial stewardship. Collaborations between clinicians and data scientists should be supported by structured governance, regulatory, and ethical frameworks that can keep pace with the rapid development of LLMs and their application in healthcare. Transparency in LLM models (data openness, data quality, and model explainability) and clear accountability for LLM-supported decisions are crucial from a regulatory standpoint, and intended across all phases from model development to the evaluation of the trustworthiness of responses/suggestions provided by fully developed models. Educating future medical professionals on these aspects will also play a fundamental role in improving the proper use of LLM-based support for antibiotic prescribing by ensuring a deeper understanding of their strengths and limitations (Box 4).[45-70]

Antimicrobial resistance is projected to cause up to 10 million deaths annually by 2050, surpassing deaths from other widespread diseases.[71,72] Achieving a balanced and safe use of LLM support in antibiotic prescribing and antimicrobial stewardship initiatives is thus an opportunity not to be missed.

Acknowledgments

An LLM-based chatbot (ChatGPT-4o) was utilized to enhance the readability and conciseness of this manuscript. Although several of the authors have over 10 years of experience in writing scientific articles, English is not our primary language. Therefore, we typically review the article multiple times after the initial draft to improve readability and conciseness. For this specific paper on LLMs, the initial draft (written by the authors without the support of LLMs, available as supplementary file S1, and including this acknowledgment section) was submitted to ChatGPT-4o on 20 July 2024, with the following prompt: "*Could you improve (in terms of readability and conciseness) the text of the following scientific article (without adding information, removing information, or changing the meaning of the sentences)? The total length of the text should remain around 3000 words*". The text produced by ChatGPT-4o was then thoroughly reviewed by all authors to ensure the preservation of content and meaning.

Authors contribution

D.R.G. originated the idea for this paper. D.R.G., C.M., B.L.M, D.P., and A.M. prepared the first draft of this paper. N.R., C.C., M.P., M.G. A.S., and M.B. supervised all aspects of the research and provided inputs on early drafts of the paper. S.G., Y.M., and S.M. revised the paper and discussed its contents with the other authors. All authors reviewed and agreed to the final version of this paper.

Competing interest

Outside the submitted work, Matteo Bassetti has received funding for scientific advisory boards, travel, and speaker honoraria from Angelini, Astellas, BioMérieux, Cidara, Gilead, Menarini, MSD, Pfizer, Shionogi, Tetraphase, Nabriva. Outside the submitted work, Daniele Roberto Giacobbe reports investigator-initiated grants from Pfizer, Shionogi, BioMérieux, and Gilead Italia, and speaker/advisor fees from Pfizer, Menarini, and Tillotts Pharma. The other authors have no conflicts of interests to disclose.

Additional information

No funding was received for the present work. Correspondence and requests for materials should be addressed to Daniele Roberto Giacobbe.

References

1 Omiye, J. A., Gui, H., Rezaei, S. J., Zou, J. & Daneshjou, R. Large Language Models in Medicine: The Potentials and Pitfalls : A Narrative Review. *Ann Intern Med* **177**, 210-220 (2024). https://doi.org/10.7326/M23-2772

2 Nassiri, K. & Akhloufi, M. A. Recent Advances in Large Language Models for Healthcare. *BioMedInformatics* **4**, 1097-1143 (2024).

3 Clusmann, J. *et al.* The future landscape of large language models in medicine. *Commun Med (Lond)* **3**, 141 (2023). https://doi.org/10.1038/s43856-023-00370-1

4 Thirunavukarasu, A. J. *et al.* Large language models in medicine. *Nat Med* **29**, 1930-1940 (2023). https://doi.org/10.1038/s41591-023-02448-8

5 Shah, N. H., Entwistle, D. & Pfeffer, M. A. Creation and Adoption of Large Language Models in Medicine. *JAMA* **330**, 866-869 (2023). https://doi.org/10.1001/jama.2023.14217

6 Park, Y. J. *et al.* Assessing the research landscape and clinical utility of large language models: a scoping review. *BMC Med Inform Decis Mak* **24**, 72 (2024). https://doi.org/10.1186/s12911-024-02459-6

7 Meng, X. *et al.* The application of large language models in medicine: A scoping review. *iScience* **27**, 109713 (2024). https://doi.org/10.1016/j.isci.2024.109713

8 Cascella, M. *et al.* The Breakthrough of Large Language Models Release for Medical Applications: 1-Year Timeline and Perspectives. *J Med Syst* **48**, 22 (2024). https://doi.org/10.1007/s10916-024-02045-3

9 Eddy, N. Epic, Microsoft partner to use generative AI for better EHRs. Healthcare IT News. 18 April 2023 [last accessed 20 Jul 2024]. Available at: www.healthcareitnews.com/news/epic-microsoft-partner-use-generative-aibetter-ehrs.

10 Giacobbe, D. R., Zhang, Y. & de la Fuente, J. Explainable artificial intelligence and machine learning: novel approaches to face infectious diseases challenges. *Ann Med* **55**, 2286336 (2023). https://doi.org/10.1080/07853890.2023.2286336

11 Hibbard, R. *et al.* Antimicrobial stewardship: a definition with a One Health perspective. *npj Antimicrobials and Resistance* **2**, 15 (2024). https://doi.org/10.1038/s44259-024-00031-w

12 Society for Healthcare Epidemiology of, A., Infectious Diseases Society of, A. & Pediatric Infectious Diseases, S. Policy statement on antimicrobial stewardship by the Society for Healthcare Epidemiology of America (SHEA), the Infectious Diseases Society of America (IDSA), and the Pediatric Infectious Diseases Society (PIDS). *Infect Control Hosp Epidemiol* **33**, 322-327 (2012). https://doi.org/10.1086/665010

13 Dyar, O. J., Huttner, B., Schouten, J., Pulcini, C. & Esgap. What is antimicrobial stewardship? *Clin Microbiol Infect* **23**, 793-798 (2017). https://doi.org/10.1016/j.cmi.2017.08.026

14 Deresinski, S. Principles of antibiotic therapy in severe infections: optimizing the therapeutic approach by use of laboratory and clinical data. *Clin Infect Dis* **45 Suppl 3**, S177-183 (2007). https://doi.org/10.1086/519472

15 Bassetti, M., Giacobbe, D. R., Vena, A. & Brink, A. Challenges and research priorities to progress the impact of antimicrobial stewardship. *Drugs Context* **8**, 212600 (2019). https://doi.org/10.7573/dic.212600

16 Barlam, T. F. *et al.* Implementing an Antibiotic Stewardship Program: Guidelines by the Infectious Diseases Society of America and the Society for Healthcare Epidemiology of America. *Clin Infect Dis* **62**, e51-77 (2016). https://doi.org/10.1093/cid/ciw118

17 Mora, S. *et al.* Towards the automatic calculation of the EQUAL Candida Score: Extraction of CVC-related information from EMRs of critically ill patients with candidemia in Intensive Care Units. *J Biomed Inform* **156**, 104667 (2024). https://doi.org/10.1016/j.jbi.2024.104667

18 Datta, S., Bernstam, E. V. & Roberts, K. A frame semantic overview of NLP-based information extraction for cancer-related EHR notes. *Journal of Biomedical Informatics* **100**, 103301 (2019). https://doi.org/https://doi.org/10.1016/j.jbi.2019.103301

19 Rabhi, S., Jakubowicz, J. & Metzger, M.-H. Deep Learning versus Conventional Machine Learning for Detection of Healthcare-Associated Infections in French Clinical Narratives. *Methods Inf Med* **58**, 031-041 (2019). https://doi.org/10.1055/s-0039-1677692

20 Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1**, 206-215 (2019). https://doi.org/10.1038/s42256-019-0048-x

21 Giacobbe, D. R. *et al.* Explainable and Interpretable Machine Learning for Antimicrobial Stewardship: Opportunities and Challenges. *Clin Ther* (2024). https://doi.org/10.1016/j.clinthera.2024.02.010

22 Amann, J. *et al.* To explain or not to explain?-Artificial intelligence explainability in clinical decision support systems. *PLOS Digit Health* **1**, e0000016 (2022). https://doi.org/10.1371/journal.pdig.0000016

23 Ali, S. *et al.* The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review. *Computers in Biology and Medicine* **166**, 107555 (2023). https://doi.org/https://doi.org/10.1016/j.compbiomed.2023.107555

24 Vaswani, A. *et al.* in *Proceedings of the 31st International Conference on Neural Information Processing Systems* 6000–6010 (Curran Associates Inc., Long Beach, California, USA, 2017).

25 Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018).

26 Brown, T. B. *et al.* Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

27 Singhal, K. *et al.* Towards Expert-Level Medical Question Answering with Large Language Models. *ArXiv* **abs/2305.09617** (2023).

28 Naveed, H. *et al.* A Comprehensive Overview of Large Language Models. *ArXiv* **abs/2307.06435** (2023).

29 Liu, P. *et al.* Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys* **55**, 1 - 35 (2021).

30 Chockalingam, A., Patel, A., Verma, S. & Yeung, T. NVIDIA. A Beginner's Guide to Large Language Models. Part 1. https://resources.nvidia.com/en-us-large-language-model-ebooks/llm-ebook-part1 (2023).

31 Maillard, A. *et al.* Can Chatbot Artificial Intelligence Replace Infectious Diseases Physicians in the Management of Bloodstream Infections? A Prospective Cohort Study. *Clin Infect Dis* **78**, 825-832 (2024). https://doi.org/10.1093/cid/ciad632

32 Fisch, U., Kliem, P., Grzonka, P. & Sutter, R. Performance of large language models on advocating the management of meningitis: a comparative qualitative study. *BMJ Health Care Inform* **31** (2024). https://doi.org/10.1136/bmjhci-2023-100978

33 Zaidat, B. *et al.* Performance of a Large Language Model in the Generation of Clinical Guidelines for Antibiotic Prophylaxis in Spine Surgery. *Neurospine* **21**, 128-146 (2024). https://doi.org/10.14245/ns.2347310.655

34 Lai, Y. *et al.* Exploring the capacities of ChatGPT: A comprehensive evaluation of its accuracy and repeatability in addressing helicobacter pylori-related queries. *Helicobacter* **29**, e13078 (2024). https://doi.org/10.1111/hel.13078

35 Chakraborty, C., Pal, S., Bhattacharya, M. & Islam, M. A. ChatGPT or LLMs can provide treatment suggestions for critical patients with antibiotic-resistant infections: a next-generation revolution for medical science? *Int J Surg* **110**, 1829-1831 (2024). https://doi.org/10.1097/JS9.0000000000000987

36 Tamma, P. D. *et al.* Infectious Diseases Society of America 2022 Guidance on the Treatment of Extended-Spectrum beta-lactamase Producing Enterobacterales (ESBL-E), Carbapenem-Resistant Enterobacterales (CRE), and Pseudomonas aeruginosa with Difficult-to-Treat Resistance (DTR-P. aeruginosa). *Clin Infect Dis* **75**, 187-212 (2022). https://doi.org/10.1093/cid/ciac268

37 Paul, M. *et al.* European Society of Clinical Microbiology and Infectious Diseases (ESCMID) guidelines for the treatment of infections caused by multidrug-resistant Gram-negative bacilli (endorsed by European society of intensive care medicine). *Clin Microbiol Infect* **28**, 521-547 (2022). https://doi.org/10.1016/j.cmi.2021.11.025

38 De Vito, A. *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

39 Schwartz, I. S., Link, K. E., Daneshjou, R. & Cortes-Penfield, N. Black Box Warning: Large Language Models and the Future of Infectious Diseases Consultation. *Clin Infect Dis* **78**, 860-866 (2024). https://doi.org/10.1093/cid/ciad633

40 Yuan, K. *et al.* Leveraging transformers and large language models with antimicrobial prescribing data to predict sources of infection for electronic health record studies. *medRxiv*, 2024.2004.2017.24305966 (2024). https://doi.org/10.1101/2024.04.17.24305966

41 Chang, Y. *et al.* A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.* **15**, Article 39 (2024). https://doi.org/10.1145/3641289

42 Protocol for the development of the Chatbot Assessment Reporting Tool (CHART) for clinical advice. *BMJ Open* **14**, e081155 (2024). https://doi.org/10.1136/bmjopen-2023-081155

43 Reddy, S. *et al.* Evaluation framework to guide implementation of AI systems into healthcare settings. *BMJ health & care informatics* **28** (2021).

44 Reddy, S. Evaluating large language models for use in healthcare: A framework for translational value assessment. *Informatics in Medicine Unlocked* **41**, 101304 (2023). https://doi.org/https://doi.org/10.1016/j.imu.2023.101304

45 Wysocka, M., Wysocki, O., Delmas, M., Mutel, V. & Freitas, A. Large language models, scientific knowledge and factuality: A systematic analysis in antibiotic discovery. *arXiv preprint arXiv:2305.17819* (2023).

46 Williamson, S. M. & Prybutok, V. Balancing privacy and progress: a review of privacy challenges, systemic oversight, and patient perceptions in AI-driven healthcare. *Applied Sciences* **14**, 675 (2024).

47 Wang, X. & Wang, Y. Analysis of trust factors for AI-assisted diagnosis in intelligent Healthcare: Personalized management strategies in chronic disease management. *Expert Systems with Applications*, 124499 (2024).

48 Wang, L. *et al.* Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *npj Digital Medicine* **7**, 41 (2024).
49 Vellido, A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural computing and applications* **32**, 18069-18083 (2020).
50 Tang, X. *et al.* Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537* (2023).
51 Sezgin, E. Artificial intelligence in healthcare: Complementing, not replacing, doctors and healthcare providers. *Digital health* **9**, 20552076231186520 (2023).
52 Sahni, N. R. & Carrus, B. Artificial intelligence in US health care delivery. *New England Journal of Medicine* **389**, 348-358 (2023).
53 Ravi, A., Neinstein, A. & Murray, S. G. Large language models and medical education: Preparing for a rapid transformation in how trainees will learn to be doctors. *ATS scholar* **4**, 282-292 (2023).
54 Pirson, M. & Malhotra, D. K. Unconventional insights for managing stakeholder trust. *Harvard Business School NOM Working Paper* (2008).
55 Park, S. H., Do, K.-H., Kim, S., Park, J. H. & Lim, Y.-S. What should medical students know about artificial intelligence in medicine? *Journal of educational evaluation for health professions* **16** (2019).
56 Padua, D. *Trust, social relations and engagement: Understanding customer behaviour on the web*. (Springer, 2012).
57 Ochodo, E. A. *et al.* Overinterpretation and misreporting of diagnostic accuracy studies: evidence of "spin". *Radiology* **267**, 581-588 (2013).
58 McCoy, L. G. *et al.* What do medical students actually need to know about artificial intelligence? *npj Digital Medicine* **3**, 86 (2020). https://doi.org/10.1038/s41746-020-0294-7
59 Mayer, R. C., Davis, J. H. & Schoorman, F. D. An integrative model of organizational trust. *Academy of management review* **20**, 709-734 (1995).
60 Li, S. S. *et al.* MEDIQ: Question-Asking LLMs for Adaptive and Reliable Medical Reasoning. *arXiv preprint arXiv:2406.00922* (2024).
61 Johri, S. *et al.* Guidelines For Rigorous Evaluation of Clinical LLMs For Conversational Reasoning. *medRxiv*, 2023.2009. 2012.23295399 (2023).
62 Jigyasu, D., Kumar, S., Shekhawat, R. S. & Vats, S. in *Healthcare Solutions Using Machine Learning and Informatics* 1-24 (Auerbach Publications, 2022).
63 Göndöcs, D. & Dörfler, V. AI in medical diagnosis: AI prediction & human judgment. *Artificial Intelligence in Medicine* **149**, 102769 (2024).
64 Gautam, P. & Sharma, R. Legal And Ethical Concerns In AI Driven Healthcare-A Study Of Legal Approaches. *Educational Administration: Theory and Practice* **30**, 11781-11788 (2024).
65 Driesnack, S. *et al.* A practice-based approach to teaching antimicrobial therapy using artificial intelligence and gamified learning. *JAC Antimicrob Resist* **6**, dlae099 (2024). https://doi.org/10.1093/jacamr/dlae099
66 Bornstein, B. H. & Emler, A. C. Rationality in medical decision making: a review of the literature on doctors' decision-making biases. *Journal of evaluation in clinical practice* **7**, 97-107 (2001).
67 Bommareddy, S., Khan, J. A. & Anand, R. A review on healthcare data privacy and security. *Networking Technologies in Smart Healthcare*, 165-187 (2022).

68 Atluri, H. & Thummisetti, B. ENHANCING ANTIBIOTIC PRESCRIBING IN URGENT CARE BY LEVERAGING LARGE LANGUAGE MODELS FOR OPTIMIZED CLINICAL DECISION SUPPORT. (2024).

69 A Compact Guide to Retrieval Augmented Generation (RAG). Definitions, components and basics for practitioners, E-Book, DataBricks. RAG "offers the greatest potential control over the model's expressiveness" of LLMs. Downloaded with permission from Databricks (3 July 2024).

70 Schwartz, S., Yaeli, A. & Shlomov, S. Enhancing trust in LLM-based AI automation agents: New considerations and future challenges. *arXiv preprint arXiv:2308.05391* (2023).

71 Antimicrobial Resistance, C. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* **399**, 629-655 (2022). https://doi.org/10.1016/S0140-6736(21)02724-0

72 O'Neill J. Tackling drug-resistant infections globally: final report and recommendations. London: The Review on Antimicrobial Resistance, 2016.

Box 1. The complexity of antibiotic prescribing and antimicrobial stewardship[11-16]

Clinicians prescribing antibiotics typically undergo complex clinical reasoning involving several key considerations:

1. ***Assessing the Cause of the Clinical Picture***
    - Determining whether the patient has a bacterial infection or if the symptoms could be due to another infectious agent (e.g., a virus) or a non-infectious disease.
2. **Severity and Necessity of Empirical Antibiotic Therapy**
    - Deciding if the clinical presentation is severe enough to warrant empirical antibiotic therapy while awaiting a definitive diagnosis, to prevent any perilous delay in treatment that could adversely impact the prognosis.
3. **Choice of Antibiotic Treatment**
    - If empirical antibiotic treatment is necessary, deciding whether to prescribe a single antibiotic or a combination, and selecting the appropriate antibiotic(s) based on the infection site, infection severity, expected causative agents, and the risk of antibiotic-resistant infections considering the patient's medical and microbiological history and the local microbiological epidemiology.

As microbiological test results become available at different times, the clinical reasoning evolves:

1. **Rapid Molecular Tests**
    - If rapid molecular tests (which provide information on the presence or absence of certain etiological agents and resistance determinants within hours) are available, deciding whether to discontinue, escalate, or de-escalate antibiotic therapy based on the results. This requires peculiar expertise also because a negative result does not exclude infections by organisms or resistance determinants not included in the test panel.
2. **Microscopy Results and Positive Cultures**
    - In the case of positive cultures, considering whether to escalate or de-escalate antibiotic therapy based on microscopy results (Gram-staining) while waiting for complete culture results.
3. **Complete Culture Results**
    - Adjusting antibiotic therapy based on complete culture results providing identification of the etiological agents.
4. **Addressing possible contamination or colonization**
    - Determining whether the identified organism is a true etiological agent or a contaminant (e.g., coagulative-negative staphylococci in a single positive blood culture) or colonization (in case of cultures from non-sterile sites), and if the clinical picture in similar cases might be due to another unidentified organism or a non-infectious cause.
5. **Phenotypic Antibiotic Susceptibility Tests**
    - Typically, within 24 hours after identifying the etiological agents, phenotypic antibiotic susceptibility test results also become available. This prompts further clinical reasoning to choose the best antibiotic for targeted therapy based on proven susceptibility.

The evolution of the clinical picture and laboratory results indicating organ function and inflammatory status necessitate continuous evaluation:

1. **Worsening Clinical Picture**
    - If the patient's condition worsens while waiting for an etiological diagnosis, considering whether the worsening is due to insufficient antibiotic efficacy requiring dosage adjustments (e.g., in case of augmented renal clearance in critically ill patients for antibiotics with renal excretion) or the consideration of other etiological agents (e.g., fungal infections in high-risk patients).
2. **Evaluating Therapy Based on Clinical and Laboratory Evolution**
    - In the absence of an etiological diagnosis (e.g., negative cultures), continuously assessing the clinical picture and laboratory results (e.g., inflammatory markers trends) to decide on escalation, de-escalation, or discontinuation of antibiotic therapy.
3. **Discontinuation of Antibiotics**
    - Favorable clinical and laboratory evolution, alongside guideline recommendations, should prompt decisions on discontinuing antibiotics after completing the appropriate course to avoid excessive therapy durations, which can increase risks like invasive fungal infections or *Clostridioides difficile* infections.

An essential aspect of clinical reasoning for antibiotic prescription is adherence to antimicrobial stewardship objectives, which aim to ensure:

1. **Reduction of Antimicrobial-Resistant Bacteria Emergence**
    - Minimizing selective antibiotic pressure on bacterial populations to prevent the emergence of antimicrobial-resistant bacteria.
2. **Cost Reduction**
    - Reducing excessive healthcare costs associated with suboptimal antimicrobial use.
3. **Responsible Use of Antimicrobials**
    - Promoting responsible antimicrobial use at national and global levels across human health, animal health, and the environment.
    - Ensuring sustainable access to effective antibiotic therapy for all those in need.
    - Preserving the future efficacy of antimicrobials.

Effective support for antibiotic prescribing through LLM-based tools must enhance accuracy across all these tasks. It should not merely suggest an antibiotic but consider the dynamic phases of antibiotic prescription and management, ensuring comprehensive information is used to maximize treatment efficacy and safety for the patient and adhere to global antimicrobial stewardship goals.

Box 2. Main structural components of large language models (LLMs)[27-30]

A LLM is any language model trained on a vast amount of data, which can subsequently be fine-tuned for specific tasks or domains in natural language processing (NLP). Most LLMs are built through three core stages:

1. **Pre-training**: The model is trained in a self-supervised manner on large corpora to extract parameters. The quantity and quality of the data are essential for building a successful model.
2. **Fine-tuning**: The pre-trained model is further trained on a new, specific dataset to adapt it better to a particular task.
3. **Prompting**: This involves querying the trained model to generate responses.

The latest LLMs follow the "pre-training, prompt, predict" paradigm, meaning they can predict the most suitable text according to a human prompt, similar to text autofill.

Components and Construction of LLMs

Several necessary components and steps are involved in constructing an LLM:

1. **Data Pre-processing**: Ensuring data quality through techniques like data cleaning and deduplication is crucial. These steps allow the computer to read the data correctly and enable the algorithm to perform the task. Pre-processing also includes techniques to help the machine encode implicit information, such as tokenization, which parses text into tokens (characters, words, or symbols). This process makes the concept of 'words' understandable to a computer. Additionally, precautionary measures like privacy reduction (removing private information such as names and phone numbers) are taken to prevent undesired outcomes or incorrect data usage.
2. **Attention**: Assigning weights to input tokens according to their relevance within the text is essential for understanding context.
3. **Encoding Positions**: Positional embedding vectors are added to encode the sequential relationship between words in a sentence, as this information is not included in the attention module.
4. **Activation Function**: This mathematical function determines the importance of a node (token) in the network based on its input parameters and weight, calculating the node's output to the subsequent layer to provide the best data representation and maximize the algorithm's prediction ability.
5. **Layer Normalization**: Normalizing the inputs reduces the impact of different scales and value ranges, leading to faster model convergence.
6. **Parallelization**: Distributing computational tasks over multiple processors efficiently handles and accelerates the significant computational demands of training and inference on available hardware.

The architecture of an LLM is determined by applying attention together with connecting transformer blocks. While an LLM can be built from scratch, it is often customized from an existing pre-trained model (PLM). Using PLMs allows for fine-tuning an existing language model, requiring less computational power since the model has already been trained.

Building an LLM Using a PLM

The steps to build an LLM using a PLM are:

1. **Finding a Well-suited PLM**: Consider the task, dataset structure, and model size.
2. **Fine-tuning the Model**: Adjust the parameters according to the specific use case.
3. **Optimizing the Model (Model Alignment)**: Use appropriate techniques, such as Reinforcement Learning from Human Feedback.

The final step in building an LLM, even when starting from a PLM, is evaluating its performance. Since these models are usually trained with unsupervised learning algorithms on unlabeled data to infer patterns, evaluation is crucial to ensure the LLM can perform the required tasks correctly. Standardized datasets and evaluation metrics (benchmarks) are available for this purpose, facilitating objective comparisons across different models and methods and identifying LLMs' strengths and weaknesses.

Challenges and Performance

LLMs perform probabilistic computations without making the data's nature and categorizations explicit. This process reveals underlying patterns in the text structure. However, the unsupervised nature of the algorithms, combined with the large data volume, makes it challenging to trace the type and underlying reasons for textual predictions. Consequently, LLMs can lack interpretability, which can be problematic, especially in healthcare and financial contexts.

On the positive side, the performance of the latest general-purpose LLMs has improved significantly due to the large quantity of high-quality data used for training. Recent domain-specific models have shown that fine-tuning a general-purpose LLM for a specific field can lead to excellent performance. For instance, Med-PaLM 2, released in 2023, passed the US Medical License Examination (USMLE) with a score of 86.5% on the MedQA dataset, a substantial improvement from the first version's 67.2%.

In summary, LLMs represent a significant advancement in NLP, driven by extensive data and sophisticated training techniques. While challenges such as interpretability remain, the potential for fine-tuning these models for specific tasks and domains promises continued improvements in their application, particularly in critical fields like healthcare.

Box 3. Current state of large language models (LLMs)-based tools for assisting antibiotic prescription in real life clinical practice[31-35,38-41]

LLMs and LLM-based tools are increasingly being explored for their potential to assist clinicians in providing antibiotic treatment suggestions. These tools can generate tailored antibiotic prescriptions based on patient data and the latest medical literature. Specifically, LLMs could:

- Assist clinicians in diagnosing infectious cases and identifying appropriate empirical and targeted antibiotic treatments.
- Provide treatment suggestions with first-line, second-line, and third-line options and alternatives.
- Serve as platforms for continuous learning and staying updated with the latest scientific literature on antimicrobial resistance epidemiology and treatment guidelines.

Although not specifically designed for antibiotic prescription, some LLM-based tools available to the public can provide treatment suggestions based on user queries (supplementary table S1). However, the use of generative AI for antibiotic treatment requires careful validation and oversight to ensure accuracy and safety. While promising, LLM-based tools have limitations and challenges. Given their nature of generating content based on probabilistic models, they can produce convincing yet non-factual outputs, known as "hallucinations." The reliability of generated content is also intertwined with issues related to big data and artificial intelligence, such as intellectual property, the validity of training data, biases influencing treatment suggestions, privacy, and accountability for the generated content.

Despite these challenges, the transformative potential of these tools in healthcare is evident. They are being rapidly adopted across various applications due to their ability to generate text, audio, music, programming code, images, and videos with a quality often indistinguishable from human-produced content. While algorithmically generated products have numerous applications in consumer markets, the healthcare sector demands dedicated reflection due to the specific requirements for accuracy and governance of medical content.

Language models like GPT-4 are trained using textual information commonly available on the web. This approach enables the model to learn and reproduce human language effectively but also means that the quality of the information reflects the general and

often non-specialized nature of web sources. Consequently, the informational quality of the generated responses is generally suitable for the general public but not necessarily reliable or precise enough to aid professionals seeking up-to-date and highly technical information in their fields.

Challenges and Opportunities in Healthcare

In healthcare, there is a need for professional tools specialized in the medical domain. The challenge lies in merging the communicative capabilities of language models with reliable medical data repositories. Medical data is often unstructured, inconsistent, fragmented, and continuously evolving, adding to its complexity. However, the opportunities are vast. For example, since the public launch of ChatGPT (a chatbot tool based on the GPT LLM) in November 2022, over 25,000 articles have been published on PubMed with the keyword "Generative Artificial Intelligence," highlighting the rapid growth and interest in this technology.

Generative AI represents one of the most promising innovations in the medical field, with applications ranging from diagnostics to pharmaceutical research. In the near future, generative AI could safely enhance medical diagnosis and the appropriate choice of treatments, including antibiotic prescriptions, by providing integrated analyses of clinical data and immediate responses based on medical literature. However, this requires the standardization of regulatory, ethical, and governance frameworks to ensure transparency, accuracy, safety, and accountability for treatment decisions supported by LLM-based tools.

Current Use and Future Considerations

LLM-based tools are already available, easily accessible, and used in real life for seeking advice. This accessibility is not necessarily a disadvantage, as it means their aid can be exploited now. However, this further emphasizes the need for users to be fully aware of the current evolving regulatory frameworks and the need for standardization in the development and use of LLMs for decision support in healthcare. Users must also be mindful of crucial limitations, such as the black box nature of models, the possibility of hallucinations, and the lack of updated training on the most recent literature and guidelines on antibiotic prescriptions in both community and hospital settings.

In summary, while LLMs and LLM-based tools hold immense potential for transforming antibiotic treatment suggestions and broader healthcare applications, their integration must be approached with careful consideration of validation, oversight, and continuous updates. The ongoing evolution of regulatory and ethical frameworks will be critical in ensuring these tools provide reliable and safe support for medical professionals.

Box 4. How can education foster trust in large language models (LLMs) for antibiotic prescribing and antimicrobial stewardship?[45-70]

The integration of LLMs into antibiotic prescribing and stewardship practices presents both opportunities and significant educational challenges. Beyond current educational programs that focus on the fundamentals of clinical decision-making based on LLMs, the main challenge for the medical education system is to support physicians in changing their cultural approach towards responsibly adopting LLMs in clinical decision-making. This complex transition requires innovative reasoning skills and critical thinking, necessitating tailor-made educational programs.

Currently, training offerings related to the integration of LLMs in antibiotic prescribing and antimicrobial stewardship appear limited or not specifically focused on these tasks. Most rely on simulation-based learning, integration with electronic health records, data verification, and hands-on practice. There is a clear lack of specialized training programs on critical thinking, which is essential for antibiotic prescribing. The term 'critical' originates from the Greek word krínō, meaning 'I judge' or 'I make distinctions.'

A key component of trust is judgment, and critical thinking underpins trustworthiness and reliability. Studies show that physicians' propensity to use LLMs depends on their ability to adopt critical thinking. This mindset is crucial for transitioning to new approaches in clinical decision-making. Additionally, personal factors, technological considerations, and environmental influences affect physicians' initial confidence in adopting technology.

Building Trust in LLMs

To build a trusting relationship between physicians and LLMs for integrating these models into antibiotic prescribing practices, several specific educational contents must be addressed:

- The accuracy and reliability of the recommendations generated by the model
- Potential bias in the data
- Issues related to the interpretability and transparency of model results
- Patient privacy and data security in managing sensitive health information

From a sociological perspective, these areas are linked to the five Trust Beliefs: Competence, Benevolence, Integrity, Identification, and Transparency. Research confirms that trust is a complex sociological construct with its own rules but can be managed by acting on these five beliefs. This theoretical approach supports the feasibility of designing an innovative educational program focused on strengthening the trustworthiness of LLMs in antibiotic stewardship.

In this section, we apply the Trust Beliefs to the relationship between physicians (trustors) and AI-based LLMs (trustees) to design the content and objectives of a proposed course. A brief description of trusteeship follows:

- **Competence**: The ability of an LLM to achieve goals effectively and efficiently, providing support to the physician beyond their competence and capacity.
- **Benevolence**: In the context of antibiotic prescribing and antimicrobial stewardship, benevolence is the extent to which an LLM is considered to enable the physician to fulfil their primary mission: providing healthcare while respecting antibiotic prescription stewardship.
- **Integrity**: The trustor perceives the LLM as adhering to acceptable principles, including honesty, fair treatment, and avoiding hypocrisy. This belief pertains to the ethical aspect of LLMs.
- I**dentification**: Also called "value congruence," it expresses the integration or sharing of values. In this framework, the relationship between an LLM and a physician should be complementary, not one of replacement, as the physician makes the final decision. This Trust Belief is excluded from our proposal.
- **Transparency**: The possibility for the trustor to acquire information about the trustee's integrity, emphasizing the importance of education and training.

Supplementary Table S2 illustrates how teaching modules may cover all Trust Beliefs according to their specific objectives and build trust in LLMs, focusing on critical thinking. Each proposed teaching module covers a single Trust Belief, analyzed based on interdisciplinary scientific literature. The outline of the proposed course is tentative and does not specify learning objectives or topics. Notably, the teaching modules on "accuracy and reliability" and "addressing bias" comply with all Trust Beliefs and require specific attention in module design, particularly in methods for cross-checking AI recommendations and identifying biases in medical healthcare.

In summary, integrating LLMs into antibiotic prescribing and stewardship practices demands specialized educational programs that emphasize critical thinking and trust-building. By addressing key areas such as accuracy, reliability, bias, and transparency, these programs can help physicians adopt LLMs responsibly, ultimately improving patient health and health system economics.

Supplementary table S1. Some examples of generative artificial intelligence tools developed with the aim of supporting evidence-based medical decisions

Tool	Website	Description	Access
Glass AI	https://glass.health/	Medical Q&A	Free limited account with premium option
MediSearch	https://medisearch.io/	Analyzed healthcare Q&A with NLP.	Free limited account with premium option
Medical Chat	https://medical.chat-data.com/	Medical Q&A	Free limited account with premium option
MedQuestio	https://medquestio.it	Medical Q&A	Free
Glass AI	https://glass.health/	Medical Q&A	Free limited account with premium option
Kahun	https://www.kahun.com	Clinical Decision Support	Premium

Supplementary table S2. A proposal of course design on critical thinking to build trust in large language models (LLMs) to support antibiotic prescribing and antimicrobial stewardship

Teaching Modules	Learning Objectives	Trust Beliefs coverage			
		Competence	*Benevolence*	*Integrity*	*Transparency*
Introduction to Artificial Intelligence in Healthcare	Provide an overview of AI technologies, including LLMs, and their applications in healthcare	X			
Ensuring Accuracy and Reliability of LLMs in Clinical Practice	Train physicians to evaluate and ensure the accuracy and reliability of AI-generated recommendations	X	X	X	X
Addressing Bias in AI and LLMs	Educate physicians on the potential biases in data and algorithms and how to mitigate them	X	X	X	X
Interpretability and Transparency of AI Models	Equip physicians with skills to understand and interpret AI model outputs and ensure transparency in AI-assisted decision-making	X		X	X
Patient Privacy and Data Security in AI Integration	Ensure physicians understand the importance of safeguarding patient privacy and data security in the use of AI technologies		X	X	X
Ethical and Legal Considerations in AI-Driven Healthcare	Provide comprehensive knowledge on the ethical and legal aspects of integrating AI in healthcare		X	X	X

Supplementary file S1. Initial draft

An LLM-based chatbot (ChatGPT-4o) was utilized to enhance the readability and conciseness of the final version of this manuscript. Although several of the authors have over 10 years of experience in writing scientific articles, English is not our primary language. Therefore, we typically review the article multiple times after the initial draft to improve readability and conciseness. For this specific paper on LLMs, the initial draft (written by the authors without the support of LLMs and available in this supplementary file S1 starting from next page) was submitted to ChatGPT-4o on 20 July 2024, with the following prompt: "*Could you improve (in terms of readability and conciseness) the text of the following scientific article (without adding information, removing information, or changing the meaning of the sentences)? The total length of the text should remain around 3000 words*". The text produced by ChatGPT-4o was then thoroughly reviewed by all authors to ensure the preservation of content and meaning.

Large language models for antibiotic prescribing and antimicrobial stewardship

Daniele Roberto Giacobbe[1,2]*, Cristina Marelli[2], Bianca La Manna[3], Donatella Padua[4], Alberto Malva[5], Sabrina Guastavino[6], Alessio Signori[7], Sara Mora[8], Nicola Rosso[8], Cristina Campi[6,9], Michele Piana[6,9], Ylenia Murgia[3], Mauro Giacomini[3], Matteo Bassetti[1,2]

[1] Department of Health Sciences (DISSAL), University of Genoa, Genoa, Italy

[2] UO Clinica Malattie Infettive, IRCCS Ospedale Policlinico San Martino, Genoa, Italy

[3] Department of Informatics, Bioengineering, Robotics and System Engineering (DIBRIS), University of Genoa, Genoa, Italy

[4] UniCamillus - International University of Health and Medical Science

[5] Italian Interdisciplinary Society for Primary Care

[6] Department of Mathematics (DIMA), University of Genoa, Genoa, Italy

[7] Section of Biostatistics, Department of Health Sciences (DISSAL), University of Genoa, Genoa, Italy

[8] UO Information and Communication Technologies, IRCCS Ospedale Policlinico San Martino, Genoa, Italy

[9] Life Science Computational Laboratory (LISCOMP), IRCCS Ospedale Policlinico San Martino, Genoa, Italy

*** Corresponding author:**

Daniele Roberto Giacobbe, MD, PhD

Department of Health Sciences (DISSAL)

University of Genoa

Via A. Pastore 1 – 16132 Genoa, Italy

Email address: danieleroberto.giacobbe@unige.it

Abstract

With the advent of large language models (LLMs) to support healthcare decisions, researchers have begun exploring their role in supporting antibiotic prescribing. Overall, there is currently a lack of standardization in research on the use of LLMs to support antibiotic prescribing, and more efforts are needed to improve our ability to identify biases or misinformation in confident and convincing outputs from black box models. Furthermore, a point specific to antimicrobial prescribing involves the need for balancing the best possible treatment for individual patients with reducing antimicrobial resistance at a global level, in line with antimicrobial stewardship core objectives. This peculiar challenge requires dedicated and standardized guidance for measuring the appropriateness of LLMs' suggestions. Educating future medical professionals on these aspects will also play a fundamental role in improving the proper use of LLM-based support for antibiotic prescribing by ensuring a deeper understanding of their strengths and limitations. Antimicrobial resistance is projected to cause up to 10 million deaths annually by 2050, surpassing deaths from other widespread diseases. Achieving a balanced and safe use of LLM support in antibiotic prescribing and antimicrobial stewardship initiatives is thus an opportunity not to be missed.

Key words: artificial intelligence; machine learning; neural networks; natural language processing; antimicrobial resistance; antimicrobial stewardship.

Introduction

Imagine you are an hospital-based infectious diseases specialist receiving a request for consultation from another ward. When you first read the request for consultation on your computer screen, an intelligent artificial assistant, exploiting large language models (LLMs) technology, has already prepared for you a coherent summary of the patient's medical and microbiological history, relevant results of laboratory and instrumental tests, and evolution of their acute phase conditions in the last few days.[1-3] This immediately provides you with a first idea about what to do, without laboriously spending many minutes searching for information across clinical notes in the patient's clinical chart. Then, you go in the other ward to visit the patient and gather additional information from the patients and their treating physicians. During the consultation, your intelligent artificial assistant can: (*i*) directly register and summarize the additional information provided by the patient and their treating physicians; (*ii*) suggest additional relevant questions to be posed. Subsequently, after coherently merging the already knew information from the patient's history and tests results with the new information collected during the consultation, your artificial intelligent assistant can explicit some suggestions for your revision (e.g., prescription of a given antibiotic at a certain dosage and for a certain duration), supported by reasonable, summarized explanations. This is only an hypothetical example of how LLMs could aid physicians in the near future in prescribing antibiotics, likely not exhaustive of all potential applications of LLMs for this purpose.[3-8] Since the advantages (above all, dramatic reduction of repetitive tasks for clinicians, thereby making time for the more sophisticated clinical reasoning) of the introduction of LLMs in daily clinical practice could be transformative in healthcare and that profound implementation of LLMs within electronic health records (EHRs) has been already announced[9], a thorough understanding of both potential advantages and relevant limitations may be essential for current and future clinicians that will very likely deal with this emerging technology in their daily clinical practice.[1,10] In the present perspective, we focus on potential advantages and limitations of the introduction of LLMs for supporting prescription of antibiotics, both in terms of improving efficacy and safety of the therapeutic approach to the single patient and in terms of appropriate use of antibiotics in line with antimicrobial stewardship principles (i.e., responsible and appropriate prescription of antibiotics antimicrobials at both patient and global levels, to ensure availability in the present and preservation of efficacy in future

populations[11]). Notably, these are complex medical tasks, requiring dedicated medical expertise and involving a multi-component and dynamic clinical reasoning process (Box 1).[12-16]

Brief history of LLMs and how they work

Natural Language Processing (NLP) studies how to elaborate and produce natural human Language through a computer, including healthcare-related text.[17-19] NLP is considered under the domain of artificial Intelligence (AI) since it tries to reproduce tasks which are typical of human beings. For this reason, the evolution of language models goes at the same pace with the implementation of AI algorithms. The first NLP models were rule-based, i.e. systems based on pre-written rules defined by domain experts. These models performed well on specific simple tasks but poorly on unseen data.

This limit was overcome with the application of neural networks (NNs) for this task. NNs are machine learning algorithms, designed to emulate the biological architecture of human brain, i.e., networks of interconnected 'nodes' capable of transferring information. NNs are considered among "black box" models, i.e., the Indeed, the composition of features of "composite" features within the initial (input) layer and the final (output) layer, as well as their computations leading to the output of interest may be partly, or sometimes totally, unclear to data scientists building and testing the model and physicians assessing how a NN arrived to provide a given output, e.g., suggestion for a certain antibiotic prescription [20,21]. The need for improving the understanding of how similar models arrive to their outputs/predictions, which is fundamental in healthcare for reducing the risk of overlooking biases and misinformation possibly perpetuated by black box models, has led to the expansion on research on so called explainable AI.[22,23]

With regard to the task of human language recognition, recurrent NNs (RNNs), which are directed graphs that process sequential inputs, and long short-term memories (LSTMs) NNs, which are able to store past information, improved prediction skills connected to the task of text decoding. However, RNNs and LSTMs are unable to make accurate predictions over extended sequences of text. In late 2017, Vaswani and colleagues introduced transformers, i.e., NNs with architectures able to handle long-range dependencies.[24] Both RNNs and transformers rely on attention mechanisms, i.e., methods which assign "weights" to each node, in order to achieve better prediction and decisions by taking into account the importance of a word in

context. However, while RNNs are able to achieve the above on a selected context window, transformers exploit self-attention, and calculate the weights considering all the words in a sentence. With this solution, NLP models started to perform not only as a decoder of textual information, but as an encoder too, like BERT, which is capable of producing natural language answers according to the user input.[25] Similar transformers show better generalization and prediction ability than previous NLP models, but are limited by the lack of large-scale datasets and adequate computational resources.

They laid the basis for the advent of LLMs, together with the introduction of graphic processing units (GPUs), i.e., processors that increased the performance of mathematical calculations, allowing to process huge quantities of data. Public awareness of the advent of LLMs was maximized after the release of OpenAI's GPT-3 in 2020.[26] Technically, LLMs are AI algorithms that can understand, extract, summarize, predict, and generate human-like text based on patterns and relationships learnt from vast amounts of data. LLMs are considered "large" because they are trained on massive amounts of data and comprise a huge number of learnable parameters, with popular LLMs reaching hundreds of billions of parameters. This allows them to produce more accurate responses than previously proposed NLP models. Currently, some main companies releasing LLMs are OpenAI , NVIDIA/Microsoft, Meta, Google , Cohere , Anthropic (public benefit company), and EleutherAI (non-profit company). For interested readers, more details on the components and functioning architecture of LLMs are available in Box 2.[27-29]

Current literature on the use of LLMs for supporting antibiotic prescribing

There are already some examples in the scientific literature on the use of general or domain-specific LLMs, or of chatbot based on LLMs, for supporting antibiotic prescriptions. For example, the performance of generative pretrained transformer (GPT) 4.0 was recently assessed by Maillard and colleagues, through using a GPT 4.0-based chatbot (ChatGPT 4.0) in terms of its ability to provide the appropriate antimicrobial therapy recommendation in 44 retrospective cases of bloodstream infection (BSI).[30] In this study, ChatGPT 4.0 was provided with all the (anonymized) information available to clinicians who actually performed the consultation (without the aid of LLMs as per standard clinical practice), and the chatbot performance in terms of appropriateness was classified (appropriate vs. inappropriate) by infectious diseases specialists who were

not involved in the care of that given patient. Standardized prompts were provided (once for each case) to ChatGPT 4.0, contextualizing the need for a comprehensive response regarding the management of a specific case of bloodstream infection in a French hospital, to be provided as if ChatGPT was the infectious diseases specialist consulting on that given patients. Appropriateness was measured according to local and international guidelines). Furthermore, recommendations provided by ChatGPT 4.0 were also classified in terms of their harmfulness (potentially harmful for patients vs. not harmful). Overall, appropriateness of suggestions for empirical and targeted therapy was 64% and 36%, respectively, whereas 2% and 5% of empirical and targeted prescriptions, respectively, were considered as potentially harmful. For example, among potentially harmful suggestion for empirical therapy was that of a narrowing the spectrum of antibiotic therapy to a regimen not covering Gram-negative bacteria in a patient with febrile neutropenia while waiting for culture results, whereas for targeted therapy a potentially harmful suggestion was that of de-escalating from cefepime and vancomycin to cloxacillin in a neutropenic patient with a nonbacteremic infection by *Staphylococcus aureus* and a concomitant ongoing sepsis of suspected unrelated origin.[30]

In another study, Fisch and colleagued evaluated LLMs adherence to good clinical practice principles and guidelines from the Infectious Diseases Society of America (IDSA) and the European Society of Clinical Microbiology and Infectious Diseases (ESCMID) when providing management indications for a clinical case (hypothetical) of pneumococcal meningitis originating from mastoiditis.[31] No definite diagnosis was provided to LLMs. Overall, several LLMs (Llama, Bard, Claude-2, PaLM, Bing, GPT 3.5, GPT 4.0) were presented with the same case thrice, and, besides appropriateness of recommendations, also heterogeneity of the suggested management provided by the same LLM across the three different sessions was evaluated. Regarding prompting, LLMs were asked to act as expert medical assistant to suggest a junior doctor about how to manage a 52-year old patient with headache and confusion, with subsequent conversation on the case with more specific illustration of signs and symptoms. Finally, among questions inherent to antibiotic prescribing, LLMs were evaluated based on the following: (*i*) whether or not prescription of antibiotics was necessary; (*ii*) whether, if antibiotic administration was suggested, the type and dosages of suggested empirical antibiotics were in line with IDSA and ESCMID guidelines. Overall, a total of 21 responses were collected for each question, from the three different sessions for each of the seven evaluated LLMs. The need for rapid

antibiotic administration was correctly recognized in 81% of cases. The correct type of empirical antibiotics (in line with IDSA and ESCMID guidelines) was suggested in 38% of cases, with correct dosages (whenever correct antibiotics were suggested) being suggested in almost 90% of cases. Of note, some misleading statements were also identified. For example, among registered hallucinations were presence of Kernig's sign and of stiff neck (not depicted in the presented case), whereas among misleading interpretations was recognition of herpes ophthalmicus instead of bacterial meningitis based on the presented cases. With regard to variability of response during the three different session (for each LLM and with standardized prompts), heterogeneity was observed for all models, impacting the rate of adherence to guidelines across the three different sessions. Among evaluated LLMs, the one providing most consistent responses (not change in the adherence to guidelines across the three different session) was ChatGPT 4.0.[31]

In another study, the performance of the LLM-based chatbots ChatGPT 3.6 and ChatGPT 4.0 in replying to different questions regarding antibiotic prophylaxis in patients subjected to spine surgery was evaluated against the North American Spine Society (NASS) guidelines (reference standard for evaluating accuracy of responses).[32] Overall, prompts were formulated exactly as 16 questions of NASS guidelines (with the addition of reference to spine surgery whenever not included in the questions, in order to provide the necessary context present in the guidelines but not in isolated prompts). No information was provided as to whether different sessions of the same prompt were conducted with the tested chatbots. Questions included in the guidelines regards efficacy of antibiotic prophylaxis in spine surgery, selection and timing of antibiotic prophylaxis, redosing of prophylactic antibiotics, discontinuation of prophylactic antibiotics, effect of comorbidities, and possible complications of prophylaxis. Furthermore, it is worth noting that 6 questions require binary responses, whereas 10 require a more complex and articulated response. Eventually, the accuracy of response was 63% (10/16) and 81% (13/16) for ChatGPT 4.0 and ChatGPT 3.5, respectively. Regarding the difference in accuracy between ChatGPT 3.5 and ChatGPT 4.0, it is worth noting that a tendence was observed for ChatGPT towards providing overconfident (but possibly erroneous or contradictory) responses, whereas an increased (positive) tendency of ChatGPT 4.0 was that of supporting its statements with references, including the NASS guidelines.[32]

Recently, Lai and colleagues assessed the accuracy and repeatability of responses provided by ChatGPT 3.5 to queries about *Helicobacter pylori*, including those regarding treatment of *H. pylori* infections (queries regarding treatment were 6 out of a total of 22 queries).[33] Regarding repeatability, the same question (prompt) was presented to ChatGPT 3.5 at 2 weeks after the first session. Responses provided by ChatGPT 3.5 were independently assessed by two expert gastroenterologists through the following scoring system: (i) comprehensive (four points); (*ii*) correct but inadequate (3 points); (iii) mixed correct/incorrect or outdated (2 points); (iv) completely incorrect (1 point). Confirmation vs. rejection of repeatability between the two responses to the same query provided two weeks apart was also based on the independent judgment of two expert gastroenterologists (another experts with >20 years of experience in *H. pylori* infection was involved for final decision in the case of disagreement). Notably, regarding accuracy of responses, those regarding treatment showed the lowest score (mean 3.25, standard deviation ±0.48). Of note, >80% of them being rated as comprehensive (four points) or correct but inadequate (three points), but as many of 16.6% were rated as mixed correct/incorrect or outdated (2 points). Regarding repeatability, ChatGPT 3.5 was reported to provide similar responses between the two sessions in 95.2% of cases.[33]

In another recent paper, Chakraborty and colleagues asked two question to ChatGPT (version not provided) regarding the management of antibiotic-resistant infections.[34] For the first question, they provided ChatGPT with susceptibility test results for several antibiotics (but without providing ChatGPT with the clinical picture and even with the genus and name of the bacterium/a causing infection). While the authors deemed the response as appropriate (ChatGPT correctly suggested the importance of thorough evaluation of the patient's condition, the specific bacteria causing the infection, and local resistance patterns, as well as the fact the decisions regarding treatment should be made by an healthcare professional and that an infectious diseases specialist consultation should be required), it is of note that ChatGPT also provided meropenem as a suggested option (which of course, could be eventually an appropriate choice in some situations, but, in our opinion, more context and information should be necessary before suggesting a specific antibiotic). The second question was very similar, with only susceptibility results being presented to ChatGPT, this time with resistance also to carbapenems. While some important concepts were retained (need for more context and suggestion of infectious diseases specialist consultation), an antibiotic (colistin) was suggested, again without

enough context and also not in line with recent guidelines for the management of infections by carbapenem-resistant Gram-negative bacteria, which no longer include colistin among first line therapies.[35,36] No other sessions were performed to assess consistency of responses to the same prompt.[34]

Finally, De Vito and colleagues recently evaluated the performance of ChatGPT 4.0 in replaying to true/false questions and open-ended questions regarding clinical cases of bacterial infections (with availability of susceptibility test results) for a total of 96 questions.[37] Questions were formulated by experts in antibiotic prescribing, and were also asked to 4 senior residents and 4 specialists in infectious diseases. Responses (from both humans and ChatGPT 4.0) were assessed by the experts (blinded to whether responses were from humans or from ChatGPT 4.0) in terms of their accuracy and completeness. Overall, they noticed that ChatGPT 4.0 showed similar accuracy to humans in its responses to true/false questions (70% approximate correctness), while it provided more complete and accurate responses to open-ended question than human participants. Regarding evaluation of clinical cases, ChatGPT 4.0 showed lower accuracy in recognizing the potential resistance mechanism underlying resistance than human participants, and also tended not to prescribe recently approved antibiotics for multidrug-resistant Gram-negative infections, favoring "older" and more toxic antibiotics such as polymyxins. Notably, ChatGPT 4.0 also tended to suggest longer than necessary antibiotic treatment duration than human participants.[37]

Discussion and perspective

With the advent of LLMs for support of healthcare decisions, researchers have started to explore their role also for supporting antibiotic prescribing, which represent a crucial step in evaluating and improving the use of some LLMs-based tools that are already easily accessible through the web or apps (Box 3).[30-34,37-40] In this regard, several fundamental points should be considered, in our opinion, based on the initial literature on this topic discussed in the previous sections. The first point, perhaps more technical, is the lack of standardization in research conduct on the use of LLMs for supporting antibiotic prescribing. Indeed, standardization would likely be required, for example, in the building of prompts, in the number of sessions in which the same prompt should be presented to a given LLM or LLM-based chatbot, in how subsequent questions should be prepared and posed to the given LLM, and in how to measure accuracy (e.g., in the

literature above both a dichotomic measurement and a grading system were employed by different researchers) and consistency across response. In this latter regard, it is of note that event the term with which the comparison of responses to the same prompt was labeled was different across studies (consistency vs. repeatability). Notably, this heterogeneity does not belong specifically to the evaluation of the use of LLMs for supporting antibiotic prescribing, but also for supporting healthcare decisions more in general. In this regard, important initiatives such as the Chatbot Assessment Reporting Tool (CHART) project have been recently taken aiming to improve standardization of research methodology on the use of LLMs in supporting healthcare decisions in clinical practice, that we think could be fundamental if we were to improve generalizability and comparability of research findings regarding the use of LLMs also for supporting antibiotic prescribing.[41] The second point to be considered, again not specific to antibiotic prescribing but also involving other healthcare fields, is more profoundly connected to the human ability to spot either biases or hallucinations perpetuating misinformation in provided outputs of black box models, that, even when providing wrong or suboptimal suggestions, may theoretically mislead even human experts in antibiotic prescribing. In this regards, some authors advocates reliance on interpretable models only, stressing that the assumptions of increased accuracy of black box models over white box models should not be taken as granted [20]. Nonetheless, while we support that an interpretable model should be preferred over an explainable (i.e., not fully interpretable) ones whenever their accuracy/performance is similar, this could not be the rule for machine learning models working on unstructured data like LLMs.[21-23] This open door to the issue of explainability of LLMs or, more importantly, which degree of explainability (in term of accuracy/correctness of explanations) can be accepted for healthcare decisions. Defining and identifying such a threshold would also require openness regarding datasets employed for training (e.g., resources used for training some of the most recent version of private LLMs have not been fully disclosed) and model architectures, as well as careful assessment of any possible pre-existing potential for biases in data employed for training. Furthermore, pre-processing of data would also require privacy preservation (both for general LLMs trained on texts publicly available on the web and for domain-specific LLMs if integrated within electronic medical charts to be trained on texts from clinical notes), grammatical corrections of errors, proper recognition of medical terms and their abbreviations. Finally, expert human evaluation of responses provided by LLMs during development and also

before/during implementation in clinical practice has been recognized as a crucial step. In this regard, initiative like the Translational Evaluation of Healthcare AI (TEHAI) have been taken in the attempt to develop and standardize a comprehensive and multi-stage evaluation of artificial intelligence models such as LLMs.[42,43] Furthermore, development of proper design and metrics of randomized studies to assess the performance of LLMs for antibiotic prescribing in clinical practice are, in our opinion, a further crucial step to evaluated for the highest degree of certainty of evidence the efficacy in terms of improvement of appropriate prescriptions and safety of using LLMs or LLM-based chatbot for supporting medical decisions about antibiotic prescribing. Finally, a third point to be considered, specific to antimicrobial prescribing, lies in the peculiarity of antimicrobial stewardship, with the need at the same time to provide the best possible antibiotic treatment to the single patient and to relieve pressure for emergence and dissemination of antimicrobial resistance. This likely represent an additional challenge also for measuring appropriateness of suggestion of LLMs for antimicrobial prescribing and will require dedicated and standardized guidance.

Overall, the above considerations cannot be separated from a multidisciplinary approach to LLMs development, approval, and use in clinical practice for antibiotic prescribing and within antibiotic stewardship initiatives. Collaborations between clinicians and data scientist developing LLM models for healthcare should be supported by structured and dedicated governance, regulatory, and ethical frameworks, that in the future should be able to keep pace with the rapid development of LLMs and their application in healthcare. This would require transparency of LLMs models (in terms of data openness, data quality, and model explainability) and accountability of LLM-supported decision to be clearly defined from a regulatory standpoint from the first steps of model development to the evaluation of trustworthiness of both accuracy and safety of responses/suggestions provided by fully developed models. Last, but not least in terms of crucial importance, proper education of future medical professionals about all these aspect will play a fundamental role in improving proper use of LLMs-based support for antibiotic prescribing by guaranteeing a more profound baseline knowledge of their strengths and limitations (Box 4).[39,44-68]

It has been estimated that antimicrobial resistance could result in up to 10 million deaths per year by 2050, more than those caused by any of other widespread infectious and non-infectious diseases.[69,70] Should the above described complex equilibrium be reached to guarantee a fair and safe use of LLMs support,

exploiting their transformative potential to further improve antibiotic prescription and within antimicrobial stewardship initiatives is, in our opinion, an opportunity not to be missed.

Acknowledgments

An LLM-based chatbot (ChatGPT-4o) was employed for improving readability and conciseness of the text. Despite several of the authors have more than 10 years of experience in writing scientific articles, English is not our primary language, therefore we usually review the article several times after the initial draft is produced to improve readability and conciseness. For this specific paper on LLM, the initial draft (written by the authors without the support of LLMs, available as supplementary file S1, and also including the present acknowledgment section) was presented to ChatGPT on 20 July 2024, using the following prompt: "*Could you improve (in terms of readability and conciseness) the text of the following scientific article (without adding information and without removing information, and without changing the meaning of the sentences)? The total length of the text should remain around 3000 words*". The produced text was then revised by all authors for guaranteeing preservation of contents and meaning.

Authors contribution

D.R.G. originated the idea for this paper. D.R.G., C.M., B.L.M, D.P., and A.M. prepared the first draft of this paper. N.R., C.C., M.P., M.G. A.S., and M.B. supervised all aspects of the research and provided inputs on early drafts of the paper. S.G., Y.M., and S.M. revised the paper and discussed its contents with the other authors. All authors reviewed and agreed to the final version of this paper.

Competing interest

Outside the submitted work, Matteo Bassetti has received funding for scientific advisory boards, travel, and speaker honoraria from Angelini, Astellas, BioMérieux, Cidara, Gilead, Menarini, MSD, Pfizer, Shionogi, Tetraphase, Nabriva. Outside the submitted work, Daniele Roberto Giacobbe reports investigator-initiated grants from Pfizer, Shionogi, BioMérieux, and Gilead Italia, and speaker/advisor fees from Pfizer, Menarini, and Tillotts Pharma. The other authors have no conflicts of interests to disclose.

Additional information

No funding was received for the present work. Correspondence and requests for materials should be addressed to Daniele Roberto Giacobbe.

References

1 Omiye, J. A., Gui, H., Rezaei, S. J., Zou, J. & Daneshjou, R. Large Language Models in Medicine: The Potentials and Pitfalls : A Narrative Review. *Ann Intern Med* **177**, 210-220 (2024). https://doi.org/10.7326/M23-2772

2 Nassiri, K. & Akhloufi, M. A. Recent Advances in Large Language Models for Healthcare. *BioMedInformatics* **4**, 1097-1143 (2024).

3 Clusmann, J. *et al.* The future landscape of large language models in medicine. *Commun Med (Lond)* **3**, 141 (2023). https://doi.org/10.1038/s43856-023-00370-1

4 Thirunavukarasu, A. J. *et al.* Large language models in medicine. *Nat Med* **29**, 1930-1940 (2023). https://doi.org/10.1038/s41591-023-02448-8

5 Shah, N. H., Entwistle, D. & Pfeffer, M. A. Creation and Adoption of Large Language Models in Medicine. *JAMA* **330**, 866-869 (2023). https://doi.org/10.1001/jama.2023.14217

6 Park, Y. J. *et al.* Assessing the research landscape and clinical utility of large language models: a scoping review. *BMC Med Inform Decis Mak* **24**, 72 (2024). https://doi.org/10.1186/s12911-024-02459-6

7 Meng, X. *et al.* The application of large language models in medicine: A scoping review. *iScience* **27**, 109713 (2024). https://doi.org/10.1016/j.isci.2024.109713

8 Cascella, M. *et al.* The Breakthrough of Large Language Models Release for Medical Applications: 1-Year Timeline and Perspectives. *J Med Syst* **48**, 22 (2024). https://doi.org/10.1007/s10916-024-02045-3

9 Eddy, N. Epic, Microsoft partner to use generative AI for better EHRs. Healthcare IT News. 18 April 2023 [last accessed 20 Jul 2024]. Available at: www.healthcareitnews.com/news/epic-microsoft-partner-use-generative-aibetter-ehrs.

10 Giacobbe, D. R., Zhang, Y. & de la Fuente, J. Explainable artificial intelligence and machine learning: novel approaches to face infectious diseases challenges. *Ann Med* **55**, 2286336 (2023). https://doi.org/10.1080/07853890.2023.2286336

11 Hibbard, R. *et al.* Antimicrobial stewardship: a definition with a One Health perspective. *npj Antimicrobials and Resistance* **2**, 15 (2024). https://doi.org/10.1038/s44259-024-00031-w

12 Society for Healthcare Epidemiology of, A., Infectious Diseases Society of, A. & Pediatric Infectious Diseases, S. Policy statement on antimicrobial stewardship by the Society for Healthcare Epidemiology of America (SHEA), the Infectious Diseases Society of America (IDSA), and the Pediatric Infectious Diseases Society (PIDS). *Infect Control Hosp Epidemiol* **33**, 322-327 (2012). https://doi.org/10.1086/665010

13 Dyar, O. J., Huttner, B., Schouten, J., Pulcini, C. & Esgap. What is antimicrobial stewardship? *Clin Microbiol Infect* **23**, 793-798 (2017). https://doi.org/10.1016/j.cmi.2017.08.026

14 Deresinski, S. Principles of antibiotic therapy in severe infections: optimizing the therapeutic approach by use of laboratory and clinical data. *Clin Infect Dis* **45 Suppl 3**, S177-183 (2007). https://doi.org/10.1086/519472

15 Bassetti, M., Giacobbe, D. R., Vena, A. & Brink, A. Challenges and research priorities to progress the impact of antimicrobial stewardship. *Drugs Context* **8**, 212600 (2019). https://doi.org/10.7573/dic.212600

16 Barlam, T. F. *et al.* Implementing an Antibiotic Stewardship Program: Guidelines by the Infectious Diseases Society of America and the Society for Healthcare Epidemiology of America. *Clin Infect Dis* **62**, e51-77 (2016). https://doi.org/10.1093/cid/ciw118

17 Rabhi, S., Jakubowicz, J. & Metzger, M.-H. Deep Learning versus Conventional Machine Learning for Detection of Healthcare-Associated Infections in French Clinical Narratives. *Methods Inf Med* **58**, 031-041 (2019). https://doi.org/10.1055/s-0039-1677692

18 Mora, S. *et al.* Towards the automatic calculation of the EQUAL Candida Score: Extraction of CVC-related information from EMRs of critically ill patients with candidemia in Intensive Care Units. *J Biomed Inform* **156**, 104667 (2024). https://doi.org/10.1016/j.jbi.2024.104667

19 Datta, S., Bernstam, E. V. & Roberts, K. A frame semantic overview of NLP-based information extraction for cancer-related EHR notes. *Journal of Biomedical Informatics* **100**, 103301 (2019). https://doi.org/https://doi.org/10.1016/j.jbi.2019.103301

20 Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1**, 206-215 (2019). https://doi.org/10.1038/s42256-019-0048-x

21 Giacobbe, D. R. *et al.* Explainable and Interpretable Machine Learning for Antimicrobial Stewardship: Opportunities and Challenges. *Clin Ther* (2024). https://doi.org/10.1016/j.clinthera.2024.02.010

22 Amann, J. *et al.* To explain or not to explain?-Artificial intelligence explainability in clinical decision support systems. *PLOS Digit Health* **1**, e0000016 (2022). https://doi.org/10.1371/journal.pdig.0000016

23 Ali, S. *et al.* The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review. *Computers in Biology and Medicine* **166**, 107555 (2023). https://doi.org/https://doi.org/10.1016/j.compbiomed.2023.107555

24 Vaswani, A. *et al.* in *Proceedings of the 31st International Conference on Neural Information Processing Systems* 6000–6010 (Curran Associates Inc., Long Beach, California, USA, 2017).

25 Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018).

26 Brown, T. B. *et al.* Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

27 Singhal, K. *et al.* Towards Expert-Level Medical Question Answering with Large Language Models. *ArXiv* **abs/2305.09617** (2023).

28 Naveed, H. *et al.* A Comprehensive Overview of Large Language Models. *ArXiv* **abs/2307.06435** (2023).

29 Liu, P. *et al.* Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys* **55**, 1 - 35 (2021).

30 Maillard, A. *et al.* Can Chatbot Artificial Intelligence Replace Infectious Diseases Physicians in the Management of Bloodstream Infections? A Prospective Cohort Study. *Clin Infect Dis* **78**, 825-832 (2024). https://doi.org/10.1093/cid/ciad632

31 Fisch, U., Kliem, P., Grzonka, P. & Sutter, R. Performance of large language models on advocating the management of meningitis: a comparative qualitative study. *BMJ Health Care Inform* **31** (2024). https://doi.org/10.1136/bmjhci-2023-100978

32 Zaidat, B. *et al.* Performance of a Large Language Model in the Generation of Clinical Guidelines for Antibiotic Prophylaxis in Spine Surgery. *Neurospine* **21**, 128-146 (2024). https://doi.org/10.14245/ns.2347310.655

33 Lai, Y. *et al.* Exploring the capacities of ChatGPT: A comprehensive evaluation of its accuracy and repeatability in addressing helicobacter pylori-related queries. *Helicobacter* **29**, e13078 (2024). https://doi.org/10.1111/hel.13078

34 Chakraborty, C., Pal, S., Bhattacharya, M. & Islam, M. A. ChatGPT or LLMs can provide treatment suggestions for critical patients with antibiotic-resistant infections: a next-generation revolution for medical science? *Int J Surg* **110**, 1829-1831 (2024). https://doi.org/10.1097/JS9.0000000000000987

35 Paul, M. *et al.* European Society of Clinical Microbiology and Infectious Diseases (ESCMID) guidelines for the treatment of infections caused by multidrug-resistant Gram-negative bacilli (endorsed by European society of intensive care medicine). *Clin Microbiol Infect* **28**, 521-547 (2022). https://doi.org/10.1016/j.cmi.2021.11.025

36 Tamma, P. D. *et al.* Infectious Diseases Society of America 2022 Guidance on the Treatment of Extended-Spectrum beta-lactamase Producing Enterobacterales (ESBL-E), Carbapenem-Resistant Enterobacterales (CRE), and Pseudomonas aeruginosa with Difficult-to-Treat Resistance (DTR-P. aeruginosa). *Clin Infect Dis* **75**, 187-212 (2022). https://doi.org/10.1093/cid/ciac268

37 De Vito, A. *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

38 Yuan, K. *et al.* Leveraging transformers and large language models with antimicrobial prescribing data to predict sources of infection for electronic health record studies. *medRxiv*, 2024.2004.2017.24305966 (2024). https://doi.org/10.1101/2024.04.17.24305966

39 Schwartz, I. S., Link, K. E., Daneshjou, R. & Cortes-Penfield, N. Black Box Warning: Large Language Models and the Future of Infectious Diseases Consultation. *Clin Infect Dis* **78**, 860-866 (2024). https://doi.org/10.1093/cid/ciad633

40 Chang, Y. *et al.* A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.* **15**, Article 39 (2024). https://doi.org/10.1145/3641289

41 Protocol for the development of the Chatbot Assessment Reporting Tool (CHART) for clinical advice. *BMJ Open* **14**, e081155 (2024). https://doi.org/10.1136/bmjopen-2023-081155

42 Reddy, S. *et al.* Evaluation framework to guide implementation of AI systems into healthcare settings. *BMJ health & care informatics* **28** (2021).

43 Reddy, S. Evaluating large language models for use in healthcare: A framework for translational value assessment. *Informatics in Medicine Unlocked* **41**, 101304 (2023). https://doi.org/https://doi.org/10.1016/j.imu.2023.101304

44 Wysocka, M., Wysocki, O., Delmas, M., Mutel, V. & Freitas, A. Large language models, scientific knowledge and factuality: A systematic analysis in antibiotic discovery. *arXiv preprint arXiv:2305.17819* (2023).

45 Williamson, S. M. & Prybutok, V. Balancing privacy and progress: a review of privacy challenges, systemic oversight, and patient perceptions in AI-driven healthcare. *Applied Sciences* **14**, 675 (2024).

46 Wang, X. & Wang, Y. Analysis of trust factors for AI-assisted diagnosis in intelligent Healthcare: Personalized management strategies in chronic disease management. *Expert Systems with Applications*, 124499 (2024).

47 Wang, L. *et al.* Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *npj Digital Medicine* **7**, 41 (2024).

48 Vellido, A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural computing and applications* **32**, 18069-18083 (2020).
49 Tang, X. *et al.* Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537* (2023).
50 Sezgin, E. Artificial intelligence in healthcare: Complementing, not replacing, doctors and healthcare providers. *Digital health* **9**, 20552076231186520 (2023).
51 Sahni, N. R. & Carrus, B. Artificial intelligence in US health care delivery. *New England Journal of Medicine* **389**, 348-358 (2023).
52 Ravi, A., Neinstein, A. & Murray, S. G. Large language models and medical education: Preparing for a rapid transformation in how trainees will learn to be doctors. *ATS scholar* **4**, 282-292 (2023).
53 Pirson, M. & Malhotra, D. K. Unconventional insights for managing stakeholder trust. *Harvard Business School NOM Working Paper* (2008).
54 Park, S. H., Do, K.-H., Kim, S., Park, J. H. & Lim, Y.-S. What should medical students know about artificial intelligence in medicine? *Journal of educational evaluation for health professions* **16** (2019).
55 Padua, D. *Trust, social relations and engagement: Understanding customer behaviour on the web*. (Springer, 2012).
56 Ochodo, E. A. *et al.* Overinterpretation and misreporting of diagnostic accuracy studies: evidence of "spin". *Radiology* **267**, 581-588 (2013).
57 McCoy, L. G. *et al.* What do medical students actually need to know about artificial intelligence? *npj Digital Medicine* **3**, 86 (2020). https://doi.org/10.1038/s41746-020-0294-7
58 Mayer, R. C., Davis, J. H. & Schoorman, F. D. An integrative model of organizational trust. *Academy of management review* **20**, 709-734 (1995).
59 Li, S. S. *et al.* MEDIQ: Question-Asking LLMs for Adaptive and Reliable Medical Reasoning. *arXiv preprint arXiv:2406.00922* (2024).
60 Johri, S. *et al.* Guidelines For Rigorous Evaluation of Clinical LLMs For Conversational Reasoning. *medRxiv*, 2023.2009. 2012.23295399 (2023).
61 Jigyasu, D., Kumar, S., Shekhawat, R. S. & Vats, S. in *Healthcare Solutions Using Machine Learning and Informatics* 1-24 (Auerbach Publications, 2022).
62 Göndöcs, D. & Dörfler, V. AI in medical diagnosis: AI prediction & human judgment. *Artificial Intelligence in Medicine* **149**, 102769 (2024).
63 Gautam, P. & Sharma, R. Legal And Ethical Concerns In AI Driven Healthcare-A Study Of Legal Approaches. *Educational Administration: Theory and Practice* **30**, 11781-11788 (2024).
64 Driesnack, S. *et al.* A practice-based approach to teaching antimicrobial therapy using artificial intelligence and gamified learning. *JAC Antimicrob Resist* **6**, dlae099 (2024). https://doi.org/10.1093/jacamr/dlae099
65 Bornstein, B. H. & Emler, A. C. Rationality in medical decision making: a review of the literature on doctors' decision-making biases. *Journal of evaluation in clinical practice* **7**, 97-107 (2001).
66 Bommareddy, S., Khan, J. A. & Anand, R. A review on healthcare data privacy and security. *Networking Technologies in Smart Healthcare*, 165-187 (2022).
67 Atluri, H. & Thummisetti, B. ENHANCING ANTIBIOTIC PRESCRIBING IN URGENT CARE BY LEVERAGING LARGE LANGUAGE MODELS FOR OPTIMIZED CLINICAL DECISION SUPPORT. (2024).

68 A Compact Guide to Retrieval Augmented Generation (RAG). Definitions, components and basics for practitioners, E-Book, DataBricks. RAG "offers the greatest potential control over the model's expressiveness" of LLMs. Downloaded with permission from Databricks (3 July 2024).

69 Antimicrobial Resistance, C. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* **399**, 629-655 (2022). https://doi.org/10.1016/S0140-6736(21)02724-0

70 O'Neill J. Tackling drug-resistant infections globally: final report and recommendations. London: The Review on Antimicrobial Resistance, 2016.

71 Chockalingam, A., Patel, A., Verma, S. & Yeung, T. NVIDIA. A Beginner's Guide to Large Language Models. Part 1. https://resources.nvidia.com/en-us-large-language-model-ebooks/llm-ebook-part1 (2023).

72 Schwartz, S., Yaeli, A. & Shlomov, S. Enhancing trust in LLM-based AI automation agents: New considerations and future challenges. *arXiv preprint arXiv:2308.05391* (2023).

Box 1. The complexity of antibiotic prescribing and antimicrobial stewardship[11-16]

Clinicians prescribing antibiotics usually undergo a complex clinical reasoning, involving at least the following considerations:

- Whether the patient could have a bacterial disease or the clinical picture may be due to another infectious agent (e.g., a virus) or to a non-infectious disease
- Independent of the eventually infectious or non-infectious nature of the clinical picture, whether the clinical presentation is enough severe to require administration of an empirical antibiotic therapy while waiting for definitive diagnosis, in order not result in perilous delay in antibiotic treatment unfavourably impacting prognosis
- If empirical antibiotic treatment is administered, whether a single antibiotic or a combination of antibiotic should be prescribed, and which antibiotic/s to prescribe, based on the site of infection, severity of the infection, expected causative agents, and expected risk of antibiotic-resistant infections according to the patient medical and microbiological history and the local microbiological epidemiology

If antibiotic/s are administered, the clinical reasoning evolve according to the increasing availability of microbiological tests result at different points in time (dynamic evolution of antibiotic prescribing):

- If rapid molecular tests (able to provide some information about the presence or absence of some possible etiological agents and resistance determinants within a few hours) are provided, whether to discontinue, escalate, or de-escalate antibiotic therapy based on rapid tests results; furthermore, it should be noted that clinical reasoning regarding decision based on the results of rapid molecular tests requires dedicated expertise, since antibiotic prescribing experts also need to consider the possibility of etiological agents/resistance determinants not included in the panel of the employed test and thus not intercepted (i.e., a negative result does not exclude infection by such organisms or presence of resistance determinants other than those included in the given panel)
- In case of positive cultures, whether to escalate or de-escalate antibiotic therapy according to microscopy results (Gram-staining) while waiting for complete culture results
- Whether to escalate or de-escalate antibiotic therapy based on complete culture results (identification of the etiological agent/s)
- Whether the type of identified organism is a true etiological agent or may represent contamination (e.g., in the case of a single positive blood culture yielding coagulative-negative staphylococci) or colonization (in the case of cultures for normally non-sterile sites) and thus the underlying clinical picture could be due to another unidentified organisms or to non-infectious causes
- Usually within 24 hours after identification of the etiological agent/s in culture, also phenotypic antibiotic susceptibility test results become available, prompting dedicated clinical reasoning about which antibiotic represent the best suited option for targeted therapy according to proven susceptibility

The evolution of the clinical picture and of laboratory results informing about organ function and inflammatory status also bring dynamic considerations into the clinical reasoning of antibiotic prescribing:

- If the clinical picture is worsening in the absence of while waiting for etiological diagnosis, whether the worsening could be due to insufficient efficacy requiring adjustment of antibiotic dosages (e.g., required adjustment of dosages based on variations in organ function such as augmented renal clearance) or considerations of other etiological agents (e.g. in presence of risk factors for fungal infections)
- In the absence of etiological diagnosis (e.g., in case of negative cultures), the evolution of the clinical picture (either improving or worsening) and of laboratory results (e.g., either favorable or unfavourable trend in the value of inflammatory markers) need to be considered for evaluating escalation, de-escalation, or discontinuation on antibiotic therapy
- Finally, favorable evolution of the clinical picture and of laboratory test results, together with recommendations by guidelines, should prompt decisions about when to discontinue antibiotics due to completion of the proper course of antibiotic therapy, in order to avoid excessive durations of therapy with no additional gain in efficacy and increasing risks connected to prolonged antibiotic use (e.g., development of invasive fungal infections or *Clostridioides difficile* infections)

A peculiar aspect of clinical reasoning regarding antibiotic prescription is also that of respecting the objectives of antimicrobial stewardship, that, besides guaranteeing the best suited treatment for any given infected patient, are also the following at ward/hospital/national/international levels:

- Reduce selective antibiotic pressure on bacterial populations driving the emergence of antimicrobial-resistant bacteria. Antimicrobial stewardship
- Reduction of excessive costs that can be attributed to suboptimal antimicrobial use
- Promotion the responsible use of antimicrobials at the national and global level, and across human health, animal health and the environment
- Promotion the responsible use of antimicrobials to ensure sustainable
- access to efficacious antibiotic therapy for all those needing it
- Promotion the responsible use of antimicrobials to preserve their future efficacy

It appears that evident that an efficient support of antibiotic prescribing by LLMs-based tools should guarantee improvement of accuracy in all the above tasks, and does not reduce to the suggestion of a given antibiotic without proper considerations of the dynamic phases of antibiotic prescription and management and of the complete context of information necessary to pursue and maximize both efficacy and safety in the treatment of the infected patient and adherence to antimicrobial stewardship aims at a more global level.

Box 2. Main structural components of large language models (LLMs)[27-29,71]

A LLM is any language model trained on a huge amount of data, which can subsequently be fine-tuned on a specific task or domain in NLP.

Most of the LLMs are built according to these three core adaptation stages:

1. *Pre-training*: the process of training the model in a self-supervised on a corpora, in order to extract the parameters. In this step the quantity and the quality of the data is essential for building a successful model.
2. *Fine-tuning*: the process of taking the pre-trained model and further training it on a new, specific dataset to better adapt it to a particular task.
3. *Prompting*: the querying process of the trained model for generating responses.

The latest LLMs follow the paradigm “pre-training, prompt, predict” meaning that they are capable of predicting the most suitable text according to a human prompt, e.g. autofill of the text.

Necessary components and steps for the construction of a LLM are:

- *Data pre-processing:* in order to ensure data’s quality, some pre-processing techniques are required, such as data *cleaning* and data *deduplication*. These steps are required for the computer to read the data correctly and for the algorithm to perform the task. The pre-processing step also includes techniques that help the machine to encode information that is not explicit for a computer, such as the *tokenization* that parses the text into *tokens*, i.e. non-decomposing units such as characters, words, or symbols, allowing to make understandable the concept of ‘word’ for a computer.

Furthermore, there are some precautionary measures that are included in this phase, like *privacy reduction*, i.e., cutting off all private information, such as names and phone numbers, that are aimed to prevent undesired outcomes or incorrect use of the data.

- *Attention*: the process of assigning weights to input tokens according to their relevance within the text.
- *Encoding positions*: the positional information, i.e. the sequential relation between the words in a sentence, is not included in the attention module. For this reason, *positional embedding vectors* are added, encoding this information in the tokens.
- *Activation function*: a mathematical function which establishes the importance of a node (which, in this case, represents a token) in the network, according to the input parameters and its weight, and calculates the node's output to the subsequent layer, in order to provide the best representation of the data and maximise the prediction ability of the algorithm.
- *Layer normalisation*: the normalisation transformation of the inputs, reducing the impact of different scales and ranges of values, to faster convergence of the model.
- *Parallelization*: the technique of distributing computational tasks over multiple processors to make feasible, to handle efficiently, and accelerate the considerable computational demands of training and inference on available hardware.

The architecture of the LLM is determined by the application of the attention together with the connection of transformer blocks.

A LLM can be built from scratch, but more often it is customized from an existing pre-trained model (PLM).

With PLMs it is possible to fine-tune an existent language model, and since the model has already been trained, PLMs require less computational power. The steps to build an LLM using an PLM can be summed as follows:

1. Finding a well-suited PLM, considering the task, the structure of the dataset, and the model size.
2. Fine-tuning the model, adjusting the parameters according to the specific use case.
3. Optimising the model (model alignment) using appropriate techniques (e.g., Reinforcement Learning from Human Feedback).

The last step when building an LLM, even when starting from a PLM, is the evaluation of the performance. Since these models are usually trained with unsupervised learning algorithms, i.e. algorithms trained on unlabelled data from which they infer patterns, the evaluation of a model is a crucial step for knowing if the LLM is able to perform the required tasks correctly. Different benchmarks are usually available for this purpose, i.e., standardised datasets and evaluation metrics for a specific language-related task. Benchmarks facilitate objective comparisons across different models and methods, helping to identify the strengths and weaknesses of LLMs.

LLMs perform probabilistic computations without making explicit the data's nature and the categorizations extracted by the algorithm. This process brings out the underlying patterns of the entire text structure. This unsupervised nature of the algorithms, combined with the large volume of data, makes it difficult, but not impossible, to trace the type and underlying reasons for textual predictions. Hence, LLMs can lack interpretability, i.e., the ability to interpret models' predictions. This can be harmful especially in healthcare and financial contexts.

On the other hand, the performance of the latest general-purpose LLMs, due to the large quantity of quality data from which they are trained, has improved in the medical domain too. Recent domain-specific models have shown how taking a general-purpose LLMs and fine-tuning it on a specific field can lead to good performances. As an example, Med-PaLM 2, released in 2023, was able to pass the US Medical Licence Examination (USMLE) with a score of 86.5% on the MedQA dataset, a big improvement from the first version, which obtained 67.2%.

Box 3. Current state of large language models (LLMs)-based tools for assisting antibiotic prescription in real life clinical practice[30-34,37-40]

LLMs and LLMs-based tools have started to be explored for their potential to assist clinicians in providing antibiotic treatment suggestions by generating tailored antibiotic prescriptions based on patient data and latest medical literature. For this purpose, they could, for example:

- Assist clinicians in diagnosing infectious cases and identifying appropriate empirical and targeted antibiotic treatment choices
- Provide treatment suggestions with first line, second line, and third line options and alternatives

- Serve as a platform for continuous learning and for staying updated with the latest scientific literature regarding antimicrobial resistance epidemiology and treatment guidelines

While not specifically designed for antibiotic prescription, some LLM-based tools are already available to the public that can reply by providing treatment suggestions based on queries from users (supplementary table S1). However, the use of generative AI for antibiotic treatment must be approached carefully, with proper validation and oversight to ensure accuracy and safety. Indeed, while promising, LLMs-based tools are not without limitations and challenges. Given their intrinsic nature of generating content based on probabilistic models, they can produce convincing yet non-factual outputs, a phenomenon known as "hallucinations". The reliability of generated content is intertwined with issues commonly associated with big data and artificial intelligence, such as intellectual property, the validity of training data, biases influencing treatment suggestions, privacy, and accountability for the generated content. Despite these challenges, the immense and transformative potential of these tools in healthcare is evident as they are rapidly being adopted across various applications, leveraging their ability to generate text, audio, music, programming code, images, and videos with a quality often indistinguishable from human-produced content. While algorithmically generated products find numerous applications in consumer markets, the healthcare sector demands a dedicated reflection due to the specific requirements for accuracy and governance of medical content. Language models like generative pre-trained transformer (GPT) 4.0 are trained using textual information commonly available on the web. This approach enables the model to learn and reproduce human language effectively but also means that the quality of the information reflects the general and often non-specialized nature of web sources. Consequently, the informational quality of the generated responses is generally suitable for the general public but not necessarily reliable or precise enough to aid professionals seeking up-to-date and highly technical information in their fields. In the healthcare sector, there is a need for professional tools specialized in the medical domain. The challenge lies in merging the communicative capabilities of language models with reliable medical data repositories. Medical data itself poses a challenge as it is often unstructured, inconsistent, fragmented, and continuously evolving, inherently complex in content. However, the opportunities are vast, as evidenced by the 25,000 articles published on PubMed in the last 18 months, since the public launch of ChatGPT (a chatbot tool based on the GPT LLM) in November 2022, with the keyword "Generative Artificial Intelligence." This is in line with the fact that generative AI represents one of the most promising innovations in the medical field, with applications ranging from diagnostics to pharmaceutical research. In the (perhaps near) future, generative AI could safely enhance medical diagnosis and appropriate choose of treatment including antibiotic prescriptions by providing integrated analyses of clinical data and immediate responses based on medical literature, provided standardization of the regulatory, ethical, and governance frameworks are aligned on guaranteeing transparency, accuracy, safety and accountability for treatment decisions supported by LLMs-based tools. In the meantime, the fact that LLMs-based tools are already available and easily accessible and used in real life for seeking advice cannot be ignored. This is not necessarily a disadvantage, since it means their aid can be already exploited. Nonetheless, this further stresses the need for the user to be fully aware of the current evolving regulatory frameworks and need for standardization in development and use of LLMs for decision support in healthcare, as well as potentially crucial limitations such as the black box nature of models, the possibility of hallucinations, and that of lack of updated training on most recent literature and guidelines on antibiotic prescriptions in either community or medical setting.

Box 4. How can education foster trust in large language models (LLMs) for antibiotic prescribing and antimicrobial stewardship?[44-68,72]

The integration of LLMs into antibiotic prescribing and stewardship practices poses several opportunities but also decisive educational challenges. Beyond current educational programmes based on knowledge of the fundamentals of clinical decision-making based on LLMs, the main challenge of the medical education system is to support physicians to change their cultural approach for a responsible adoption of LLMs in clinical decision-making. This complex mindset transition needs innovative reasoning skills and critical thinking that require tailor-made educational programmes.

From this perspective, current training offerings regarding the integration of LLMs in antibiotic prescribing and antimicrobial stewardship, appear limited or not specifically focused on such tasks, and rely mostly on simulation-based learning, integration with electronic health records (EHR), data verification and hands-on practice. Based on this evidence, there is a clear lack of specialised training programmes on critical thinking, which is essential for antibiotic prescribing. Etymologically, the term

‘critical’ is derived from the Greek krínō, meaning ‘I judge’, ‘I make distinctions’. As described below, a pillar of trust is judgement. Therefore, ‘critical thinking’ underlies the concept of trustworthiness and reliability. Studies show that the propensity of physicians to use LLMs depends on their ability to adopt critical thinking. This attitude plays a key role in opening the door to the transition to the new mindset. Furthermore, elements influencing physicians' initial confidence in adopting technology include personal factors, technological considerations and environmental influences. Against this backdrop, the hypothesis developed in this box is, therefore, whether it is possible to design a specialised course aimed at physicians or students with the objective to strengthen the accountability of LLMs and to promote the adoption of LLMs in the field of antibiotic prescribing antibiotic stewardship, with an impact on patient health and health system economics.

Building Trust in LLMs

In order to build a trusting relationship between physicians and LLMs, aimed at the integration of LLMs into the clinical practice of antibiotic prescribing, it is of paramount importance to address some specific educational content, such as: the accuracy and reliability of the recommendations generated by the model; potential bias in the data; issues related to the interpretability and transparency of model results; patient privacy and data security in the management of sensitive health information. From a sociological perspective, all these areas appear to be linked to the five Trust Beliefs: *Competence, Benevolence, Integrity, Identification, Transparency*. Research confirms that trust is a very complex sociological construct, has its own rules, but can be managed by acting on the five trust beliefs. This theoretical approach confirms the feasibility hypothesis to design an innovative educational programme focused on strengthening the trustworthiness of LLMs in antibiotic stewardship towards physicians.

To this end, in this section, we originally apply the Trust Beliefs to the relationship between physicians (trustor) and AI-based LLMs (trustee) to design contents and objectives of the proposed course. Before explaining the course outline, a brief description of trusteeship follows:

- *Competence* refers to the ability of an LLM to achieve the goal effectively and efficiently, i.e. to provide support to the physician in a distinctive way, that is, beyond his or her competence and capacity.
- The application of *Benevolence* to the context of antibiotic prescribing and antimicrobial stewardship has to be interpreted as the extent to which an LLM (trustee) is considered to be doing good to the physician (trustor), i.e. enabling the physician to fulfil his or her primary mission: providing health care while respecting antibiotic prescription stewardship.
- *Integrity* implies that the trustor perceives the LLM (trustee) as adhering to a set of principles (integrity) considered acceptable by the medical (trustor), including honesty, fair treatment, and the avoidance of hypocrisy . This belief refers to the ethical aspect of LLMs.
- *Identification* is also called “value congruence”, a sociological concept expressing integration or sharing of values. In the context of the framework provided in this box, the relationship between an LLM and a physician has to be of complementarity and not of replacement as it is the medical to take the final decision. For this reason, this Trust Belief is excluded from our proposal.
- *Transparency* may be interpreted as the possibility of the trustor to acquire information about the trustee’s integrity. In this case, education and training are crucial.

Supplementary table S2 illustrates how teaching modules may cover all Trust Beliefs according to their specific and related objectives and build trust in LLMs, focusing on critical thinking. The coverage of a single Trust Belief for each proposed teaching module is attributed through the analysis of course-specific modules and content, based on the interdisciplinary scientific literature. Importantly, here we have presented a tentative, synthetic outline of a proposed course. Therefore, it does not illustrate specific learning objectives nor topics. Finally, it should be noted that the two proposed teaching modules covering “accuracy and reliability” and “addressing bias” comply with all Trust Beliefs. These two areas require specific attention in terms of specific module design, in particular, in terms of methods for cross-checking AIs and recommendations and key methods and strategies for identifying biases in medical healthcare.