npj | antimicrobials & resistance

Perspective

https://doi.org/10.1038/s44259-025-00084-5

Advantages and limitations of 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[7,8], Sara Mora[9], Nicola Rosso[9], Cristina Campi[6,10], Michele Piana[6,10], Ylenia Murgia[3], Mauro Giacomini[3] & Matteo Bassetti[1,2]

Antibiotic prescribing requires balancing optimal treatment for patients with reducing antimicrobial resistance. There is a lack of standardization in research on using large language models (LLMs) for supporting antibiotic prescribing, necessitating more efforts to identify biases and misinformation in their outputs. Educating future medical professionals on these aspects is crucial for ensuring the proper use of LLMs for supporting antibiotic prescribing, providing a deeper understanding of their strengths and limitations.

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 and (ii) suggest additional relevant questions to be posed. After coherently merging the already known information from the patient's history and tests results with the new information collected during the consultation, your artificial intelligent assistant can explicitly offer some suggestions for your revision (e.g., prescribing 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 antibiotic prescribing, both in terms of improving the efficacy and safety of the therapeutic approach to the single patient and in terms of the appropriate use of antibiotics in line with antimicrobial stewardship principles (i.e., responsible and appropriate antibiotic prescribing 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.

[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]Departmental Faculty of Medicine, UniCamillus - International University of Health and Medical Science, Rome, Italy. [5]Italian Interdisciplinary Society for Primary Care, Bari, Italy. [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]IRCCS Ospedale Policlinico San Martino, Genoa, Italy. [9]UO Information and Communication Technologies, IRCCS Ospedale Policlinico San Martino, Genoa, Italy. [10]Life Science Computational Laboratory (LISCOMP), IRCCS Ospedale Policlinico San Martino, Genoa, Italy. ✉e-mail: danieleroberto.giacobbe@unige.it

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. **Identification of the cause of the clinical picture**
    1. 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**
    1. 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**
    1. 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**
    1. 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**
    1. 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**
    1. Adjusting antibiotic therapy based on complete culture results providing identification of the etiological agents.

4. **Possible contamination or colonization**
    1. 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**
    1. Typically, within 24 h 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**
    1. 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. **Evaluation of therapy based on clinical and laboratory results**
    1. 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**
    1. 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 prescribing is adherence to antimicrobial stewardship objectives, which aim to ensure:

1. **Reduction of antimicrobial-resistant bacteria emergence**
    1. Minimizing selective antibiotic pressure on bacterial populations to prevent the emergence of antimicrobial-resistant bacteria.

2. **Cost reduction**
    1. Reducing excessive healthcare costs associated with suboptimal antibiotic use.

3. **Responsible use of antibiotics**
    1. Promoting responsible antibiotic use at national and global levels across human health, animal health, and the environment.
    2. Ensuring sustainable access to effective antibiotic therapy for all those in need.
    3. Preserving the future efficacy of antibiotics.

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 prescribing and management, ensuring comprehensive information is used to maximize treatment efficacy and safety for the patient and adhere to global antimicrobial stewardship goals.

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[20].

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[21]. 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[22,23]. 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[24,25].

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[20].

In late 2017, Vaswani and colleagues introduced transformers, deep NNs with architectures able to handle long-range dependencies[26]. Transformers rely on attention mechanisms, methods which assign "weights" to each token to achieve better predictions and decisions by determining the importance of a word in its context, a mechanism that has also been applied to RNNs. Transformers exploit self-attention and calculate the weights considering all the words in a sentence, while RNNs consider a selected context window. Thanks to this new mechanism, that allowed the modeling of dependencies between words, independently from their distance or sequence, NLP models started to perform both as decoders and encoders of textual information. Through transformers it was possible to design architectures capable of both taking in input and producing in output natural language texts, e.g., conversational transformers, a function that was anticipated by task-oriented dialog systems which relied on encoders such as bidirectional encoder representations from transformers (BERT)[27].

Transformers show better generalization and prediction ability than previous NLP models but were limited by the lack of large-scale datasets and adequate computational resources[26,28]. 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-3.5 in 2022. Technically, LLMs are AI algorithms that work by predicting the next tokens in a sentence and are able to extract, summarize, 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 improve outputs and generalization of 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[20,29–31].

Current literature on the use of LLMs for supporting antibiotic prescribing

The scientific literature already includes examples of using general-purpose or domain-specific LLMs, as well as chatbots powered by LLMs, to support antibiotic prescribing. For example, the performance of generative pre-trained 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)[32]. 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, the 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[32].

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[33]. 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 to 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 antibiotic prescribing 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[33].

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[34]. Prompts

Box 2 | Main structural components of large language models (LLMs)[20,29–31]

An 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 two 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 pretrained model is further trained on a new, specific dataset to adapt it better to a particular task.

After completing these building stages, the next step is the inference phase, which includes:

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 preprocessing**: 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. Preprocessing 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 by learning 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 pretrained language 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 self-supervised 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 self-supervised 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.

were formulated exactly as the 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[34].

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)[35]. Regarding repeatability, the same question (prompt) was presented to ChatGPT-3.5 2 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[35].

In another recent paper, Chakraborty and colleagues asked two questions to ChatGPT (version not provided) regarding the management of antibiotic-resistant infections[36]. 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 a 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[37,38]. No other sessions were performed to assess response consistency to the same prompt[36].

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[39]. 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[39].

Discussion and perspective

With the emergence of LLMs in healthcare decision-making, researchers have also started investigating their potential to support antibiotic prescribing (Box 3)[32–36,39–42]. 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 the 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[43].

The second point is the need to improve the 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[22]. 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[23–25]. This highlights the issue of LLM explainability, or more specifically, the level of explanation accuracy and correctness that can be deemed acceptable for healthcare decisions. Establishing such a standard would require transparency about the datasets used for training, clarity regarding model architectures, and acknowledgment of potential biases in the training data to avoid their perpetuation. However, complicating this picture is the fact that, in the lack of sufficient reliability of explanations, the need to consider explainability as an absolute requirement for AI models has also been questioned, with some authors suggesting focusing on rigorous internal and external validation of AI models instead[44,45]. Independent of the final trajectory (favoring either explanations or validation), preprocessing of data will need to ensure privacy preservation, grammatical correction of errors, and proper recognition of medical terms and abbreviations.

Other techniques besides, in alternative, or potentially synergistic to improving explainability have emerged to reduce hallucinations and biases in supporting healthcare decisions. Retrieval-augmented generation (RAG) combines the pretrained parametric memory of LLM with external non-parametric memory (e.g., we can imagine a link to the most updated guidelines on antibiotic prescribing for different infectious diseases), also in the form of knowledge graphs, as a fine-tuning approach leading to LLM responses more grounded in real factual knowledge[46,47]. Chain-of-thought prompting has been shown to elicit multi-step reasoning behavior in LLM, that could improve their performance independent of fine-tuning[48]. Dedicated evaluation frameworks to assess the reliability of LLMs as assistants in the biomedical domain have also been developed, prioritizing prompt robustness, high recall, and lack of hallucinations as necessary criteria for this use case[49], although it is likely that specific evaluation frameworks will be needed for antibiotic prescribing considering the peculiar need to adequately balance benefits for individual patients and global benefits in terms of reducing development and selection of antimicrobial resistance. In this light, while very high recall would likely be essential when supporting the selection of effective empirical antibiotics in patients with severe clinical presentation and reduced survival in case of delayed initiation of an effective therapy, the same may not always hold true in case of a less severe clinical presentation with no immediate prognostic repercussions, that could require a different balance between recall and other performance metrics from an antimicrobial stewardship perspective.

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[50,51]. Scaling human evaluation through crowdsourcing and the development of dedicated benchmarks to assess AI alignment to human preferences are also been explored as a way to expand and possibly improve the global evaluation of models' performances[52]. 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 prescribing.

A third point specific to antibiotic prescribing, as introduced in previous paragraphs, 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 antibiotic prescribing. Overall, all the above considerations necessitate a multidisciplinary approach to LLM development, approval, and clinical use for antibiotic prescribing and antimicrobial stewardship. Collaborations

Box 3 | Current state of large language models (LLMs)-based tools for assisting antibiotic prescribing in real-life clinical practice[32–36,39–42]

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 prescribing, some LLM-based tools available to the public can provide treatment suggestions based on user queries (Supplementary Table 1). 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 prescribing 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.

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. Furthermore, the complex and nuanced nature of antibiotic prescribing will likely require the use of complex platforms that connect and coordinate different models dedicated to the various dynamic phases of prescribing, which should act globally as moral agents balancing the needs of an individual with those of the broader society[53]. Notably, no LLM tools are currently approved by regulators for use in antibiotic-prescribing settings. Transparency in LLM models (data openness, data quality, and model explainability) and clear accountability for LLM-supported decisions could be 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)[54–79]. Despite all these current limitations and areas needing improvement, the theoretical potential advantages of LLM-based support of antibiotic prescribing are undeniable, for example: (i) reduced cognitive-load on automatic tasks in busy hospitals, with more time for clinicians to focus on more complex, high-value tasks related to antibiotic prescribing and management of complex infections; (ii) enhanced efficiency, with reduced time spent manually searching antibiotic guidelines; (iii) improve and rapid provision of contextual insights for clinical reasoning based on the elaboration and combination of medical records information with external knowledge based on the most updated research findings and guidelines; (iv) integration within healthcare records with automated monitoring for early identification of adverse events and clinical events or new laboratory results requiring changes in antimicrobial therapy, aiming both at improving patients' outcomes and at reducing selection and dissemination of antimicrobial resistance. Recent estimates indicate that more than 8 million deaths annually could be associated with antimicrobial resistance by 2050, surpassing deaths from other widespread diseases[80]. Achieving a balanced and safe use of LLM support in antibiotic

Box 4 | How can education foster trust in large language models (LLMs) for antibiotic prescribing and antimicrobial stewardship?[54–79]

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 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 fulfill their primary mission: providing healthcare while respecting antibiotic 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.
- **Identification**: 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 2 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 preliminary 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.

prescribing and antimicrobial stewardship initiatives is thus an opportunity not to be missed.

Data availability

No datasets were generated or analysed during the current study.

Received: 30 July 2024; Accepted: 6 February 2025;
Published online: 27 February 2025

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).
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).
4. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
5. Shah, N. H., Entwistle, D. & Pfeffer, M. A. Creation and adoption of large language models in medicine. *JAMA* **330**, 866–869 (2023).
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).
7. Meng, X. et al. The application of large language models in medicine: a scoping review. *iScience* **27**, 109713 (2024).
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).
9. Eddy, N. Epic, Microsoft partner to use generative AI for better EHRs. *Healthcare IT News* (18 April 2023).
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).

11. Hibbard, R. et al. Antimicrobial stewardship: a definition with a one health perspective. *NPJ Antimicrob. Resist.* **2**, 15 (2024).
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).
13. Dyar, O. J., Huttner, B., Schouten, J., Pulcini, C. & Esgap. What is antimicrobial stewardship? *Clin. Microbiol. Infect.* **23**, 793–798 (2017).
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**, S177–S183 (2007).
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).
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–e77 (2016).
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).
18. Datta, S., Bernstam, E. V. & Roberts, K. A frame semantic overview of NLP-based information extraction for cancer-related EHR notes. *J. Biomed. Inform.* **100**, 103301 (2019).
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).
20. 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).
21. Aggarwal, C. C. *Neural Networks and Deep Learning. A Textbook* (Springer, 2023).
22. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
23. Giacobbe, D. R. et al. Explainable and interpretable machine learning for antimicrobial stewardship: opportunities and challenges. *Clin. Ther.* **46**, 474–480 (2024).
24. Amann, J. et al. To explain or not to explain?-Artificial intelligence explainability in clinical decision support systems. *PLoS Digit. Health* **1**, e0000016 (2022).
25. Ali, S. et al. The enlightening role of explainable artificial intelligence in medical & healthcare domains: a systematic literature review. *Comput. Biol. Med.* **166**, 107555 (2023).
26. Vaswani, A. et al. Attention is all you need. In *Proc. 31st International Conference on Neural Information Processing Systems* (eds Von Luxburg, U. et al.) 6000–6010 (Curran Associates Inc, 2017).
27. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (eds Burstein, J. et al.) 4171–4186 (Association for Computational Linguistics, 2019).
28. Brown, T. B. et al. Language models are few-shot learners. In *NIPS'20: Proc. 34th International Conference on Neural Information Processing System* (eds Larochelle, H. et al.) 1877–1901 (Curran Associates Inc, 2020).
29. Singhal, K. et al. Towards expert-level medical question answering with large language models. *Nat. Med.* https://doi.org/10.1038/s41591-024-03423-7 (2025).
30. Naveed, H. et al. A comprehensive overview of large language models. Preprint at arXiv:2307.06435v10 (2023).
31. Liu, P. et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* **55**, 1–35 (2021).
32. 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).
33. 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**, e100978 (2024).
34. 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).
35. 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).
36. 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).
37. 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).
38. 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).
39. 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).
40. 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).
41. Yuan, K. et al. Leveraging transformers and large language models with antimicrobial prescribing data to predict sources of infection for electronic health record studies. Preprint at medRxiv, 2024.2004.2017.24305966 (2024).
42. Chang, Y. et al. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol*. **15**, 1–45 (2024).
43. Protocol for the development of the Chatbot assessment reporting tool (CHART) for clinical advice. *BMJ Open* **14**, e081155 (2024).
44. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit. Health* **3**, e745–e750 (2021).
45. Nagendran, M., Festor, P., Komorowski, M., Gordon, A. C. & Faisal, A. A. Quantifying the impact of AI recommendations with explanations on prescription decision making. *NPJ Digit. Med.* **6**, 206 (2023).
46. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NIPS'20: Proc. 34th International Conference on Neural Information Processing System* (eds Larochelle, H. et al.) 9459–9474 (Curran Associates Inc, 2020)
47. Pan, S. et al. Unifying large language models and knowledge graphs: a roadmap. *IEEE Trans. Knowl. Data Eng.* **36**, 3580–3599 (2024).
48. Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models. In *NIPS '22: Proc. 36th International Conference on Neural Information Processing Systems* (eds Koyejo, S. et al.) 24824–24837 (Curran Associates Inc, 2024).

49. Bolton, W. J., Poyiadzi, R., Morrell, E. R., van Bergen Gonzalez Bueno, G. & Goetz, L. RAmBLA: a framework for evaluating the reliability of LLMs as assistants in the biomedical domain. Preprint at arXiv:2403.14578 (2024).
50. Reddy, S. et al. Evaluation framework to guide implementation of AI systems into healthcare settings. *BMJ Health Care Inform.* **28**, e100444 (2021).
51. Reddy, S. Evaluating large language models for use in healthcare: a framework for translational value assessment. *Inf. Med. Unlocked* **41**, 101304 (2023).
52. Zheng, L. et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *NIPS '23: Proc. 37th International Conference on Neural Information Processing Systems* (eds Oh, A. et al.) 46595-46623 (Curran Associates Inc, 2023).
53. Bolton, W. J., Badea, C., Georgiou, P., Holmes, A. & Rawson, T. M. Developing moral AI to support decision-making about antimicrobial use. *Nat. Mach. Intell.* **4**, 912–915 (2022).
54. Wysocka, M., Wysocki, O., Delmas, M., Mutel, V. & Freitas, A. Large language models, scientific knowledge and factuality: A framework to streamline human expert evaluation. *J. Biomed. Inform.* **158**, 104724 (2024).
55. Williamson, S. M. & Prybutok, V. Balancing privacy and progress: a review of privacy challenges, systemic oversight, and patient perceptions in AI-driven healthcare. *Appl. Sci.* **14**, 675 (2024).
56. Wang, X. & Wang, Y. Analysis of trust factors for AI-assisted diagnosis in intelligent healthcare: personalized management strategies in chronic disease management. *Expert Syst. Appl.* **255**, 124499 (2024).
57. Wang, L. et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *NPJ Digit. Med.* **7**, 41 (2024).
58. Vellido, A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Comput. Appl*. **32**, 18069–18083 (2020).
59. Tang, X. et al. Medagents: large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024* (eds Ku, L.-W. et al.) 599-621 (Association for Computational Linguistics, 2024).
60. Sezgin, E. Artificial intelligence in healthcare: complementing, not replacing, doctors and healthcare providers. *Digit. Health* **9**, 20552076231186520 (2023).
61. Sahni, N. R. & Carrus, B. Artificial intelligence in US health care delivery. *N. Engl. J. Med.* **389**, 348–358 (2023).
62. 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 Sch* **4**, 282–292 (2023).
63. Pirson, M. & Malhotra, D. K. Unconventional insights for managing stakeholder trust. Harvard Business School NOM Working Paper No. 8-057 (2008).
64. Park, S. H., Do, K.-H., Kim, S., Park, J. H. & Lim, Y.-S. What should medical students know about artificial intelligence in medicine? *J. Educ. Eval. Health Prof.* **16**, 18 (2019).
65. Padua, D. *Trust, Social Relations and Engagement: Understanding Customer Behaviour on the Web* (Springer, 2012).
66. Ochodo, E. A. et al. Overinterpretation and misreporting of diagnostic accuracy studies: evidence of "spin". *Radiology* **267**, 581–588 (2013).
67. McCoy, L. G. et al. What do medical students actually need to know about artificial intelligence? *NPJ Digit. Med.* **3**, 86 (2020).
68. Mayer, R. C., Davis, J. H. & Schoorman, F. D. An integrative model of organizational trust. *Acad. Manag. Rev.* **20**, 709–734 (1995).
69. Li, S. S. et al. MEDIQ: question-asking LLMs for adaptive and reliable medical reasoning. Preprint at arXiv:2406.00922 (2024).
70. Johri, S. et al. Guidelines for rigorous evaluation of clinical LLMs for conversational reasoning. Preprint at *medRxiv* 2023.2009.2012.23295399 (2023).
71. Jigyasu, D., Kumar, S., Shekhawat, R. S. & Vats, S. in *Healthcare Solutions Using Machine Learning and Informatics* (eds Gupta, P. et al.) (Auerbach Publications, 2022).
72. Göndöcs, D. & Dörfler, V. AI in medical diagnosis: AI prediction & human judgment. *Artif. Intell. Med.* **149**, 102769 (2024).
73. Gautam, P. & Sharma, R. Legal and ethical concerns in AI driven healthcare-a study of legal approaches. *Educ. Adm. Theory Pract.* **30**, 11781–11788 (2024).
74. Driesnack, S. et al. A practice-based approach to teaching antimicrobial therapy using artificial intelligence and gamified learning. *JAC Antimicrob. Resist.* **6**, dlae099 (2024).
75. Bornstein, B. H. & Emler, A. C. Rationality in medical decision making: a review of the literature on doctors' decision-making biases. *J. Eval. Clin. Pract.* **7**, 97–107 (2001).
76. Bommareddy, S. et al. A review on healthcare data privacy and security. In *Networking Technologies in Smart Healthcare - Innovations and Analytical Approaches* (eds Singh, P. et al.) 165-187 (CRC Press, 2022).
77. Atluri, H. & Thummisetti, B. Enhancing antibiotic prescribing in urgent care by leveraging large language models for optimized clinical decision support. *Int. Res. J. Modern Eng. Technol. Sci.* https://doi.org/10.56726/IRJMETS48495 (2024).
78. *A Compact Guide to Retrieval Augmented Generation (RAG). Definitions, Components and Basics for Practitioners.* E-Book (DataBricks, 2024).
79. Schwartz, S., Yaeli, A. & Shlomov, S. Enhancing trust in LLM-based AI automation agents: new considerations and future challenges. Preprint at arXiv:2308.05391 (2023).
80. Naghavi, M. et al. Global burden of bacterial antimicrobial resistance 1990-2021: a systematic analysis with forecasts to 2050. *Lancet* **404**, 1199–1226 (2024).

Acknowledgements

No funding was received for the present work. 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 was submitted to ChatGPT-4o on 20 July 2024, with the following initial 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.

Author contributions

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. D.P. authored Box 4 and Supplementary Table 2. All authors reviewed and agreed to the final version of this paper.

Competing interests

Outside the submitted work, Matteo Bassetti has received funding for scientific advisory boards, travel, and speaker honoraria from Cidara, Gilead, Menarini, MSD, Mundipharma, Pfizer, and Shionogi. Outside the submitted work, Daniele Roberto Giacobbe reports investigator-initiated grants from Pfizer, Shionogi, BioMérieux, Menarini, Tillotts Pharma, and Gilead Italia, travel support from Pfizer, and speaker/advisor fees from Pfizer, Menarini, BioMérieux, Advanz Pharma, and Tillotts Pharma. Alberto Malva was previously an employee of GlaxoSmithKline and is the founder of MedQuestio,

an LLM-based support tool for Italian general practitioners. The other authors have no conflicts of interest to disclose.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44259-025-00084-5.

Correspondence and requests for materials should be addressed to Daniele Roberto Giacobbe.