Supplementary Material

Antibiotics and Artificial Intelligence (A&A):

Clinical Considerations on a Rapidly Evolving Landscape

Daniele Roberto Giacobbe[1,2,*], Sabrina Guastavino[3], Cristina Marelli[4,5], Ylenia Murgia[6], Sara Mora[7], Alessio Signori[8,9], Nicola Rosso[7], Mauro Giacomini[6], Cristina Campi[3,10], Michele Piana[3,10], Matteo Bassetti[1,2]

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

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

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

[4] Institut Curie - INSERM U1331 Team statistics applied to personalized medicine, Paris, France

[5] Gustave Roussy INSERM CESP team OncoStat - University Paris Saclay, Villejuif, France

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

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

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

[9] IRCCS Ospedale Policlinico San Martino, Genoa, Italy

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

*** Corresponding author:**

Daniele Roberto Giacobbe

Clinica Malattie Infettive, IRCCS Ospedale Policlinico San Martino

L.go R. Benzi, 10 – 16132 Genoa, Italy

Telephone: +390105554654

Email address: danieleroberto.giacobbe@unige.it

1. Responses of large language models to queries

Provided references have been deleted in order to prevent use of misleading/non-existent references.

Chat GPT 4o (OpenAI)

Prompt: Help me write a short introduction of what is artificial intelligence in general and then what are LLMs

Response: Here's a concise introduction to artificial intelligence and large language models (LLMs):

Artificial Intelligence (AI)
Artificial Intelligence refers to the simulation of human intelligence by machines, particularly computer systems. It encompasses a broad range of technologies designed to perform tasks typically requiring human cognition, such as problem-solving, decision-making, learning, and language understanding. AI systems rely on algorithms to process data, identify patterns, and make predictions or decisions. Subfields of AI include machine learning, where models improve their performance with experience, and deep learning, which uses neural networks inspired by the human brain to tackle complex tasks.

Large Language Models (LLMs)
Large Language Models are a subset of AI designed to understand and generate human-like text. These models are built using deep learning techniques, particularly transformer architectures, and are trained on massive datasets containing text from diverse sources. LLMs, such as OpenAI's GPT series, learn to predict and generate coherent sentences based on contextual input, making them capable of tasks like translation, summarization, content creation, and conversational interaction. Their ability to process and generate text in a contextually relevant way has revolutionized natural language processing, enabling applications in numerous fields, from healthcare to customer service.

If you'd like, we can expand or adapt this for your specific audience or use case!

Prompt: Please elaborate on how LLMs learn in a way that can be understood by doctors

Response: Here's an explanation of how large language models (LLMs) learn, tailored to the perspective of medical professionals:

How LLMs Learn: A Doctor-Friendly Analogy

Large Language Models (LLMs) are akin to highly trained specialists in language. Just as medical training involves absorbing vast amounts of knowledge and applying it to solve specific patient cases, LLMs undergo a two-step learning process: pretraining and fine-tuning.

1. Pretraining: The "Medical School" Phase

During pretraining, LLMs are exposed to enormous volumes of text data, much like medical students immerse themselves in textbooks, journal articles, and case studies. The model's objective is to learn patterns, relationships, and context within the text.

- How It Works:
  The model uses a structure called a transformer, which enables it to focus on relevant parts of the text (akin to prioritizing critical patient details). It learns to predict the next word in a sentence based on the words that came before it. For example, if the text is "The patient presents with chest pain and...," the model might predict "shortness of breath" as a plausible continuation.
- Outcome:
  By processing billions of such sentences, the model gains a general understanding of language—grammar, context, and associations between words. This is like a doctor building foundational knowledge in anatomy, physiology, and pathology.

2. Fine-Tuning: The "Residency Training" Phase

After pretraining, LLMs undergo fine-tuning to specialize in specific tasks, similar to how doctors enter residency programs to hone their skills in a particular field.

- How It Works:
  Fine-tuning involves training the model on smaller, curated datasets that are relevant to the desired task. For example, if the LLM is to assist in medical documentation, it might be fine-tuned on clinical notes and patient histories. This step adjusts the model's responses to be more precise and context-appropriate.
- Outcome:
  Fine-tuning ensures the model is better equipped to handle specific scenarios, just as a cardiologist develops expertise in interpreting echocardiograms and diagnosing heart conditions.

3. Continuous Learning: Staying Up-to-Date Like a Doctor

While most LLMs are static after fine-tuning, newer methods allow models to learn continuously, much like doctors who stay current through continuing medical education (CME). These models can integrate new data to remain relevant as language, guidelines, and knowledge evolve.

Technical Mechanisms: A Simplified View

- Attention Mechanism (like Diagnostic Prioritization):
  Transformers use "attention" to determine which parts of a sentence are most relevant, similar to how a doctor focuses on key symptoms and test results while forming a diagnosis.
- Neural Networks (like Neural Pathways):
  LLMs are powered by neural networks—layers of interconnected nodes that mimic the human brain's function. These networks analyze data inputs and adjust their connections (weights) to refine predictions, much like neurons strengthen pathways with repeated use.

Why This Matters in Medicine

Understanding how LLMs learn helps doctors appreciate their strengths and limitations:

- Strengths:
  They excel in recognizing patterns and generating text, making them valuable for summarizing patient histories, drafting reports, or even generating differential diagnoses based on symptoms.
- Limitations:
  LLMs can only provide responses based on patterns they've seen before. They lack true understanding, so outputs must always be critically evaluated, especially in clinical contexts where accuracy is paramount.

Would you like additional analogies or examples specific to clinical applications?

Prompt: Could you briefly explain hallucinations of LLMs to a medical audience?

Response: In the context of large language models (LLMs), "hallucinations" refer to instances where these AI systems generate plausible-sounding but factually incorrect or fabricated information. This phenomenon poses significant challenges in the medical field, where accuracy is paramount.

Understanding Hallucinations in LLMs

LLMs are trained on extensive datasets to predict and generate human-like text. However, they may produce outputs that are not grounded in reality, especially when faced with ambiguous or incomplete prompts. In medicine, such hallucinations can lead to the dissemination of incorrect medical advice, misinterpretation of patient data, or fabrication of nonexistent studies, potentially compromising patient care.

Implications for Healthcare

The consequences of LLM hallucinations in healthcare are profound:

- Patient Safety Risks: Inaccurate information can result in misdiagnoses or inappropriate treatment recommendations.

- Erosion of Trust: Healthcare professionals may lose confidence in AI tools that produce unreliable outputs.

- Legal and Ethical Concerns: Disseminating false medical information can lead to legal liabilities and ethical dilemmas.

Mitigation Strategies

To address hallucinations in LLMs within the medical domain, several approaches are being explored:

- Retrieval-Augmented Generation (RAG): This technique grounds LLM outputs in verified external data sources, reducing the likelihood of generating fabricated information.

- Self-Reflection Mechanisms: Incorporating self-assessment processes enables LLMs to evaluate and correct their outputs, enhancing factual accuracy.

- Human-in-the-Loop Systems: Engaging medical experts to review and validate AI-generated content ensures reliability before clinical application.

Conclusion

While LLMs oeer significant potential in healthcare, addressing the challenge of hallucinations is crucial to ensure their safe and eeective integration into medical practice. Ongoing research and the implementation of robust mitigation strategies are essential steps toward achieving this goal.

Claude 3.5 Sonnet (Anthropic)

Prompt: Help me write a short introduction of what is artificial intelligence in general and then what are LLMs

Response: Artificial Intelligence (AI) refers to computer systems designed to perform tasks that typically require human intelligence. These tasks include recognizing patterns, learning from experience, understanding language, solving problems, and making decisions. Unlike traditional computer programs that follow rigid, pre-defined rules, AI systems can adapt and improve their performance based on the data they process.

At its core, AI aims to create machines that can simulate aspects of human cognitive functions. This field encompasses various approaches, from rule-based systems to machine learning, where computers learn patterns from data without being explicitly programmed for each scenario.

Large Language Models (LLMs) are a specific type of AI that represent a significant breakthrough in natural language processing. They are neural networks trained on vast amounts of text data to understand and generate human-like text. Think of them as sophisticated pattern recognition systems that have learned the statistical relationships between words and concepts from millions or even billions of text examples.

LLMs work by predicting what words are likely to come next in a sequence, based on their training. This seemingly simple capability allows them to:

- Generate coherent and contextually appropriate text
- Answer questions by drawing on their training data
- Assist with writing and editing
- Translate between languages
- Help with coding and problem-solving
- Engage in conversation-like interactions

However, it's important to understand that LLMs don't truly "understand" language in the way humans do. They don't have consciousness or genuine comprehension – rather, they're extremely sophisticated statistical models that can recognize and reproduce patterns in text. While they can be remarkably capable, they can also make mistakes, exhibit biases present in their training data, and occasionally generate incorrect or nonsensical information.

Would you like me to elaborate on any particular aspect of AI or LLMs?

Prompt: Please elaborate on how LLMs learn in a way that can be understood by doctors

Response: Let me explain LLMs in a way that connects to medical concepts and training:

Think of an LLM's learning process as similar to how medical residents learn pattern recognition in clinical practice, but at a massive scale. Here's the breakdown:

The Training Process: Just as medical residents learn by seeing thousands of patient cases, LLMs learn by processing billions of text examples. However, instead of learning to recognize physical symptoms and lab values, they learn patterns in how words and concepts relate to each other.

Pattern Recognition System: Similar to how doctors learn to recognize patterns in symptoms that indicate specific conditions, LLMs learn patterns in text. For instance, just as you might recognize the pattern of fever + joint pain + rash that suggests certain diagnoses, an LLM learns patterns like "fever often appears in text near words like temperature, chills, and infection."

Prediction Mechanism: The core function is similar to clinical prediction - but instead of predicting disease progression, LLMs predict what words should come next in a sequence. For example, if given "The patient presents with chest pain that is __", the model has learned from its training data that words like "radiating," "sharp," "dull," or "crushing" are likely to follow, based on common medical descriptions.

Neural Network Structure: The architecture can be compared to a highly complex neural pathway:

- Input Layer: Like sensory neurons receiving initial information
- Hidden Layers: Similar to the processing that happens in interneurons, where the information is analyzed and patterns are detected
- Output Layer: Like motor neurons producing a response, this is where the model generates its text output

Training Data Quality: Just as medical education relies on peer-reviewed literature and verified case studies, the quality of an LLM's training data is crucial. Poor quality data (like unreliable medical information) leads to poor quality outputs.

Limitations: Like a medical student who has only read textbooks but never treated patients, LLMs have important limitations:

- They can recognize patterns but don't have real-world experience
- They can make confident-sounding but incorrect statements
- They don't actually understand causation, just correlation in text patterns
- They can't perform physical examinations or interpret real-time data

Would you like me to elaborate on any of these aspects or discuss specific applications of LLMs in medicine?

Prompt: Could you briefly explain hallucinations of LLMs to a medical audience?

Response: Let me explain LLM hallucinations using medical analogies:

LLM hallucinations can be compared to confabulation in neurological patients – where the brain creates false memories or makes incorrect connections while maintaining a confident narrative. Just as patients with Korsakoe's syndrome might fill in memory gaps with plausible but false information, LLMs can generate text that seems coherent and authoritative but is factually incorrect.

From a technical perspective, hallucinations occur when LLMs:
1. Fill in gaps in their knowledge with plausible-sounding but incorrect information
2. Make incorrect associations between concepts they've learned
3. Generate synthetic content that superficially matches patterns in their training data but lacks factual grounding

To use a medical analogy: imagine a medical student who has memorized thousands of case studies but occasionally mixes up details between cases or creates composite cases that never actually existed. The output sounds medically plausible but isn't real.

Would you like me to elaborate on any particular aspect of LLM hallucinations or their implications for medical practice?

Gemini 1.5 Flash (Google)

Prompt: Help me write a short introduction of what is artificial intelligence in general and then what are LLMs

Response: Artificial intelligence (AI) is a branch of computer science that aims to create intelligent agents, which are systems that can reason, learn, and act autonomously. AI encompasses a wide range of techniques and technologies, including machine learning, natural language processing, and computer vision.

Large language models (LLMs) are a type of AI model that are trained on massive amounts of text data. LLMs are able to generate text, translate languages, write different kinds of creative content, and answer your questions in an informative way. Some LLMs are trained on a massive scale and others are trained on a more limited scale.

Prompt: Please elaborate on how LLMs learn in a way that can be understood by doctors

Response: Large Language Models (LLMs) learn in a way that can be likened to how a child learns language. Imagine a child being exposed to a vast amount of text data, such as books, articles, and conversations. The child doesn't just memorize the words; they learn to understand the patterns, relationships, and context of the language.

Similarly, LLMs are trained on massive amounts of text data. They learn to recognize patterns in the data, such as grammar, syntax, and semantics. This allows them to generate text that is coherent and relevant to the input they receive.

However, unlike a child, LLMs don't have an innate understanding of the world. They learn through a process called machine learning, which involves training a model on a large dataset of text and labels. The model is then able to make predictions about new data, such as generating text or translating languages.

One of the key techniques used to train LLMs is called self-supervised learning. This involves training the model on a large dataset of unlabeled text. The model is tasked with predicting the next word in a sequence of words. This forces the model to learn the underlying patterns and relationships in the language.

Another important technique is called fine-tuning. This involves training the model on a smaller dataset of labeled text. This allows the model to specialize in a particular task, such as answering questions or writing different kinds of creative content.

In conclusion, LLMs learn in a way that is similar to how humans learn language. They are trained on massive amounts of text data and learn to recognize patterns and relationships in the language. This allows them to generate text that is coherent and relevant to the input they receive.

Prompt: Could you briefly explain hallucinations of LLMs to a medical audience?

Response: Large Language Models (LLMs) are powerful tools with the potential to revolutionize healthcare. However, they are prone to a phenomenon known as "hallucinations," where they generate factually incorrect or misleading information.

Hallucinations in LLMs occur when the model generates text that is not grounded in the provided information or real-world knowledge. This can lead to inaccurate diagnoses, inappropriate treatment recommendations, or misinterpretation of medical research.

It is crucial to be aware of the limitations of LLMs and to critically evaluate the information they generate, especially in healthcare settings. As the field of AI continues to advance, it is essential to develop methods to mitigate hallucinations and ensure the safe and responsible use of LLMs in healthcare.

2. 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 many years of experience in writing scientific articles, English is not our primary language. Therefore, we typically review the initial draft multiple to improve readability and conciseness. In this case, the initial draft (written by the authors without the support of LLMs and available below) was submitted to ChatGPT-4o on 23 December 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 that of the original one”. The text produced by ChatGPT-4o was then thoroughly reviewed by all authors to ensure the preservation of content and meaning.

Abstract

The growing interest in leveraging artificial intelligence (AI)-based tools for improving healthcare decisions has not spared the possibility of improving antibiotic prescriptions. Large language models (LLMs) are a type of AI model able to process and generate text in a contextually relevant way, trained on massive datasets containing text from diverse sources. The advantages are potentially massive in terms of improvement of patients’ health, but this does not mean that implementation and correct us of LLMs-based support is not complex, in particular for antibiotic prescription, that peculiarly require considering both improving the treatment of the given patient and reducing the risk of antimicrobial resistance selection in line with antimicrobial stewardship principles. There are many new grey areas pertaining both to the identification and the definition of an acceptable margin of risk, that depends on the inherent probabilistic nature of LLMs, the knowledge and quality of training data, and the recognition of hallucinations and omission, that are all fields of active research. Besides this, the training of current and future clinicians should consider education about how to maximize gains when interacting with LLMs, since it is crucial to develop a synergistic interaction that improves the performance compared to humans or LLMs alone rather than diminish it. There is an understandable excitement for a truly revolutionary technology, but also still a long road to go.

Keywords: artificial intelligence; large language models; deep learning; antimicrobial stewardship; antibiotics; antimicrobial resistance.

Introduction

The growing interest in leveraging artificial intelligence (AI)-based tools for improving healthcare decisions has not spared the possibility of improving antibiotic prescriptions [1-5]. This specific field is of particular interest, because, besides clinicians, it also faces artificial intelligence tools with a peculiar dual objective: (i) improving clinical outcomes of the treated patient; (ii) reducing selective pressure of antimicrobial resistance in the population in line with antimicrobial stewardship principles. Finding the correct balance between the two is a dutiful but not easy task for clinicians, thus the possible support of artificial intelligence in the attempt of promoting advances in the field merits particular consideration.

In this brief commentary, we provide our view on some crucial aspects of the adoption and role of artificial intelligence in supporting clinicians for antibiotic prescription, focusing in particular on the role of large language models (LLMs).

What are artificial intelligence and LLMs?

To succinctly write this introductory section and in order to anticipate one of the main themes of this commentary (the role of human dealing with AI) we wrote this brief introduction exploiting the help of LLMs (see the next section for the connected implications to the topic of this commentary).

AI refers to computer systems designed to perform tasks that typically require human intelligence [6-8]. It encompasses a broad range of technologies designed to perform tasks typically requiring human cognition, such as problem-solving, decision-making, learning, and language understanding. AI systems rely on algorithms to process data, identify patterns, and make predictions or decisions.

LLMs are a type of AI model able to generate text, translate languages, write dieerent kinds of creative content, and answer questions [9]. These models are built using deep learning techniques, and are trained on massive datasets containing text from diverse sources [10]. Their ability to process and generate text in a contextually relevant way has enabled potential applications in numerous fields, including healthcare. LLMs learn in a way that can somewhat be likened to how a medical student learn. Imagine a medical student being exposed to a vast amount of text data, such as books, articles, and conversations. The student does not just memorize the words; they learn to understand the patterns, relationships, and context of the medical language. Just as medical students learn by seeing many patient cases, LLMs learn by processing billions of text examples. However, instead of learning to recognize physical symptoms and laboratory values, they learn patterns in how words and concepts relate to each other.

What the above section means for antibiotic prescription?

It took only a few minutes to write the previous paragraph with the support of LLMs. More in detail, we asked three dieerent tools based on LLMs, namely ChatGPT 4o (OpenAI), Claude 3.5 Sonnet (Anthropic), and Gemini 1.5 Flash (Google) to reply to the following two questions and to provide the related scientific references: (i) “Help me write a short introduction of what is artificial intelligence in general and then what are LLMs”; (ii) “Please elaborate on how LLMs learn in a way that can be understood by doctors”. The original responses by the dieerent tools are available in the supplementary material. We then elaborated on the responses, verified the proposed references (discarding some that did not exist, or at least were not to be found on the web) and provided others of which we were aware based on our knowledge of the topic. The crucial point is that doing all of this took far less that what was expected if we had to write the paragraph from scratch, even if the structure of the paragraph, the topics to discuss, and the references to provide were already clear in our minds. Furthermore, LLMs provided additional useful suggestions that we had not consider in advance (e.g., the metaphor of LLMs as medical students), although an additional limitation is that LLMs could have based this suggestion on existing articles that we did not find, precluding to properly cite the original authors if and wherever necessary.

Now, translate all of this in the context of therapeutic suggestions for antibiotic prescription. The topic clearly deserves consideration, if only for the following: (i) potential reduction of the time needed for decision about whether to administer antibiotics and which antibiotic/s to administer; (ii) potential suggestions of eeective and safe alternative antibiotic

choices initially not considered by the clinician. Against this background, please also note that receiving support by LLMs does not equal to just report what guidelines dictates, since, based on what discussed in the previous paragraph, LLMs are able to elaborate contextually on the specific case.
However, several considerations, explored in the next paragraphs, should be made regarding current potentials and pitfalls of exploiting LLMs for decision support in antibiotic prescription. Notably, the present commentary focuses on clinical considerations, therefore privacy and legal implications of the use of LLMs in healthcare decisions are not discussed (some reviews and perspectives on these equally crucial aspects have been recently published [11, 12]).

Prescribing an antibiotic is not writing a commentary article

While certainly writing a commentary article is also a dieicult task to accomplish, it (usually) does not have (immediate) consequences for patients' health. Conversely, choosing the right antibiotic/s for a given patient, like many other healthcare decisions, has, intuitively, both direct and indirect repercussions on the patients' health. Thus, even if the advantages in terms of rapidity and precious additional suggestions of AI support for antibiotic prescription are shared with those of writing support by AI, there are additional considerations that should be made. By comparing the initial suggestions by LLMs-based tools and the final form of the previous section, it is clear that they are dieerent by a large extent, since we had to merge the dieerent suggestions, select pertinent contents, discard erroneous/misleading claims, verify proposed references, add other references we already selected based on our search/experience. Translated, again, in terms of AI support for antibiotic prescription, this means that expert judgment remains necessary to spot mistakes/incongruences/omissions in the proposed suggestion. Regarding this aspect, a well-known concept is that of hallucinations [13-15]. To describe hallucinations, we exploit again the help of LLMs, with the same methods and human revision/verification described above (prompt: "Could you briefly explain hallucinations of LLMs to a medical audience?"). Hallucinations can be compared (metaphorically) to a human condition in which the brain creates false memories or makes incorrect connections while maintaining a confident narrative. Consequently, LLMs can generate text that seems coherent and authoritative but is factually incorrect. Technically, hallucinations occur when LLMs fill in gaps in their knowledge with plausible-sounding but incorrect information, make incorrect associations between concepts they have learned, generate synthetic content that superficially matches patterns in their training data but lacks factual grounding. Consequently, it is also possible that LLMs fabricate non-existing references or cite non-pertinent references, thereby stressing for the need of references verification.

Imagine thus the risk of not recognizing such hallucinations when prescribing an antibiotic, or during other "high-risk" healthcare decisions (in terms of potential impact on patients). Inaccurate information can result in misdiagnoses or inappropriate treatment recommendations. Furthermore, due to the possibility of hallucinations, healthcare professionals may lose confidence in AI tools that produce unreliable outputs. Given all these reasons, why still possibly advice the future use of LLMs (provided the correct and shared ethical and legal frameworks are defined, since even unintended dissemination of false medical information created by hallucinations can lead to legal liabilities and ethical dilemmas) for supporting antibiotic prescription or other "high-risk" decisions? To articulate a response to this complex question we think it is useful to explore the concept of the risk of error, and then to discuss the expertise paradox.

The risk of error

On the clinical ground, the main reason to support the use of LLMs in healthcare is the potential for improving the risk of erroneous diagnoses and prescriptions in comparison with human doctors. Certainly, this seems in contrast with much of the exponentially increasing literature evaluating the accuracy of LLMs-based tools for healthcare decisions, mostly indicating still unacceptably high rates of potentially harmful suggestions, and the consequent inability of LLMs-based tools to "substitute" human experts in medical tasks, e.g., the prescription of antibiotics [16-20]. However, considering the rapid advances in the field, it is plausible to expect considerable improvements in the accuracy of LLMs suggestions in the near future. Furthermore, there is dedicated research aimed at reducing the risk of hallucinations [21]. In line with this vision, some works have started to appear suggesting improved performance of LLMs-based tools in comparison with humans. For example, a very recent article from the US (currently available as pre-print) aimed to assess the performance of ChatGPT o1-preview in comparison to historical human controls and benchmarks of previous LLMs with regard to five experiments including dieerential diagnosis generation, triage dieerential diagnosis, probabilistic reasoning, display of diagnostic reasoning, and management reasoning. The dieerent performance measures were adjudicated by physician experts through validated psychometrics. Overall, while no improvement over previous LLMs benchmarks were observed in terms of triage dieerential diagnosis and probabilistic reasoning (but with some gains in comparison with human baseline), the evaluate model showed over both human baseline and previous benchmark LLMs in terms of dieerential diagnosis generation and quality of diagnostic and management reasoning [22]. While this study did not specifically explore antibiotic prescription, we think the underlying implications are shared also with this latter task. Indeed, the crucial point, in our opinion, is that LLMs are becoming, on average, less prone to diagnostic/therapeutic errors than the human counterpart, which means that, if we accept that on average they may now improve diagnostic and therapeutic decisions, their use should be seriously considered as a mean to improve patients' health (which is the response to the first of our questions, i.e., why still possible advice the future use of LLMs for supporting antibiotic prescription or other "high-risk" decisions?). However, the use of LLMs is not exempt of risks, so who should use them? And how should they be used? To explore these aspects, we need first to introduce the concept of the expert paradox, in order to then discuss the risk of error just introduced above more in-depth.

The expertise paradox

Considering the above described ability of LLMs to provide rapid but comprehensive responses, potentially including also suggestions not initially considered by the asking human, it is consequent that those that may benefit the most from a given answer in a specific field (in terms of actionable information gain) are not the most experts in that field, but those with reduced expertise in that specific field. Regarding the specific topic of the present commentary, less experts may mean medical doctors without specialty in infectious diseases. Notably, we used the term "less expert" and not "not expert", since without knowledge and training in the medical field it could be very dieicult or impossible to spot hallucinations, incorrect information, or omission. For the reasons discussed in the previous paragraphs, we think this would be inacceptable for "high-risk" decisions such as antibiotic prescription. Contrarily, for a "less expert" it could be possible to recognize hallucinations and verify information for some responses, albeit not all. Indeed, for the most complex clinical

reasoning regarding antibiotic prescription in dieicult cases, expertise would be required to decipher subtle mistakes and hallucinations, considering the plausible-sounding responses of LLMs that could mislead less experts. This situation, that we may name the “expertise paradox” (more gains for less experts but more risks without experts) calls, again, for the need to find the correct balance. Indeed, while, in our opinion, it remains advisable that for healthcare decision LLMs need to be paired with human expertise, it is also desirable to have LLMs helping less experts whenever an expert is not available. However, an open question remains how to define which are the clinical situations/cases in which less experts could use LLMs safely (i.e., when the risk of missing hallucinations is so low to be considered acceptable). This is intimately connected to the “moving target” nature of the risk of error.

The risk of error is a moving target

The most dieicult thing for defining the correct balance described above is that the risk of error may vary greatly from response to response, also for very close clinical situations and sometime for the same question regarding the same patient. This because LLMs “works” by predicting the next “token” (that we may define as a word or part of a word) in the sentence based on a probability ground based on previous training, with also some “noise” be present as a regularization measure to preserve generalization of responses. Consequently, although we expect responses to the same question to be very similar in their contents, this is not an absolute certainty. With regard to the fluctuations of the magnitude of the risk of error between responses to dieerent questions, an additional variable contributing to the dieerence is risk is how solid the training of the LLM on that specific topic was both in terms of quantity of literature and quality of literature. This is inherently dieicult to measure, and potentially impossible for proprietary models for which training data is not fully disclosed. The third variable contributing to the “moving target” nature of the risk of error is, in our opinion, the how humans use LLMs. In a recent single-blind, randomized trial, 50 human participants (physicians with training in family medicine, internal medicine, or emergency medicine) were asked to reply to multiple-choice and open-ended medical reasoning questions and were randomized to: (i) having access to LLMs in addition to conventional resources; (ii) having access only to conventional resources [23]. The performance in replying to questions was measured by means of a standardized scale in percentage points. Notably, while the median score was slightly higher in the LLMs-supported arm than in the only conventional resources arm (76% vs. 74%), the maximum score was obtained by LLMs alone without human interaction (92%), i.e., 16 percentage points higher than the LLMs-supported arm (with 95% confidence interval for the dieerence from 2% to 30%, $p = 0.03$) [23]. The authors hypothesized that one of the possible drivers of this could be ineeicient prompt formulation (i.e. how the question is posed/written to LLMs), with lack of the proper details on the task, context, and instructions (that were conversely presented to the LLMs when asked to reply without the human participants). While it should be recognized that currently there is still not wide consensus on how to exactly build prompts (although there already are many prompt engineering courses on the web), it is indisputable that, if the quality and correctness of the responses by LLMs depend on how humans use them, this latter is an additional variable jeopardizing the moving target nature of the risk of error, making it even more dieicult to define “safe use” of LLMs for healthcare decisions.

Finally, antibiotic prescription requires and additional, peculiar consideration. Indeed, in addition to the above, when prescribing an antibiotic by means of a LLM-supported approach there would be to consider the balance between two margin of errors: (i) the margin of error for prescribing the correct antibiotic therapy to the patient; (ii) the margin of error for making

the best antibiotic selection in terms of antibiotic stewardship. Certainly, antibiotic experts already weigh these two needs (best for the patient in terms of cure and best for the community in terms of resistance selection) when prescribing antibiotics without the support of LLMs. However, the need for recognizing hallucinations and omission may also impact this dual balance in a dieerent but subtle way (e.g., prioritizing reduction of only one of the two errors), configuring a new aspect to be considered and recognized

Conclusions

In this commentary, we have provided our view on the evolving field of LLMs based support for “high-risk” healthcare decisions, focusing in particular on the peculiarity of antibiotic prescription. As discussed in the previous section, we are convinced that the advantages are potentially massive in terms of improvement of patients’ health, but this does not mean that implementation and correct us of LLMs-based support is not complex. There are many new grey areas pertaining both to the identification and the definition of an acceptable margin of risk, that depends on the inherent probabilistic nature of LLMs, the knowledge and quality of training data, and the recognition of hallucinations and omission, that are all fields of active research. Besides this, the training of current and future clinicians should consider education about how to maximize gains when interacting with LLMs, since it is crucial to develop a synergistic interaction that improves the performance compared to humans or LLMs alone rather than diminish it. There is an understandable excitement for a truly revolutionary technology, but also still a long road to go.

Funding

This work was not funded.

Acknowledgments

None

Conflicts of interest

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, and Gilead Italia, and speaker/advisor fees from Pfizer, Menarini, and Tillotts Pharma. The other authors have no conflicts of interests to disclose.

Author Contributions

Conceptualization: D.R.G.; writing—original draft preparation: D.R.G., S.G., C.M., Y.M., S.M.; writing—review and editing: D.R.G., S.G., C.M., Y.M., S.M., A.S., M.G., N.R, C.C., M.P., M.B.; supervision: D.R.G., A.S., N.R., M.G., C.C., M.P., M.B.

References

1. Nigo M, Rasmy L, Mao B, Kannadath BS, Xie Z, Zhi D. Deep learning model for personalized prediction of positive MRSA culture using time-series electronic health records. Nat Commun. 2024;15(1):2036. doi: 10.1038/s41467-024-46211-0.

2. Kanjilal S, Oberst M, Boominathan S, Zhou H, Hooper DC, Sontag D. A decision algorithm to promote outpatient antimicrobial stewardship for uncomplicated urinary tract infection. Sci Transl Med. 2020;12(568). doi: 10.1126/scitranslmed.aay5067.
3. Kim C, Choi YH, Choi JY, Choi HJ, Park RW, Rhie SJ. Translation of Machine Learning-Based Prediction Algorithms to Personalised Empiric Antibiotic Selection: A Population-Based Cohort Study. Int J Antimicrob Agents. 2023;62(5):106966. doi: 10.1016/j.ijantimicag.2023.106966.
4. Chang A, Chen JH. BSAC Vanguard Series: Artificial intelligence and antibiotic stewardship. J Antimicrob Chemother. 2022;77(5):1216-7. doi: 10.1093/jac/dkac096.
5. Giacobbe DR, Marelli C, Guastavino S, Signori A, Mora S, Rosso N, et al. Artificial intelligence and prescription of antibiotic therapy: present and future. Expert Rev Anti Infect Ther. 2024;22(10):819-33. doi: 10.1080/14787210.2024.2386669.
6. Duan Y, Edwards JS, Dwivedi YK. Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda. International Journal of Information Management. 2019;48:63-71. doi: https://doi.org/10.1016/j.ijinfomgt.2019.01.021.
7. Collins C, Dennehy D, Conboy K, Mikalef P. Artificial intelligence in information systems research: A systematic literature review and research agenda. International Journal of Information Management. 2021;60:102383. doi: https://doi.org/10.1016/j.ijinfomgt.2021.102383.
8. Jaboob A, Durrah O, Chakir A. Artificial Intelligence: An Overview. In: Chakir A, Andry JF, Ullah A, Bansal R, Ghazouani M, editors. Engineering Applications of Artificial Intelligence. Cham: Springer Nature Switzerland; 2024. p. 3-22.
9. Kumar P. Large language models (LLMs): survey, technical frameworks, and future challenges. Artificial Intelligence Review. 2024;57(10):260. doi: 10.1007/s10462-024-10888-y.
10. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nature Medicine. 2023;29(8):1930-40. doi: 10.1038/s41591-023-02448-8.
11. Maccaro A, Stokes K, Statham L, He L, Williams A, Pecchia L, et al. Clearing the Fog: A Scoping Literature Review on the Ethical Issues Surrounding Artificial Intelligence-Based Medical Devices. J Pers Med. 2024;14(5). doi: 10.3390/jpm14050443.
12. Meszaros J, Minari J, Huys I. The future regulation of artificial intelligence systems in healthcare services and medical research in the European Union. Front Genet. 2022;13:927721. doi: 10.3389/fgene.2022.927721.
13. Belisle-Pipon JC. Why we need to be careful with LLMs in medicine. Front Med (Lausanne). 2024;11:1495582. doi: 10.3389/fmed.2024.1495582.
14. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Trans Inf Syst. 2024. doi: 10.1145/3703155.
15. Shah SV. Accuracy, Consistency, and Hallucination of Large Language Models When Analyzing Unstructured Clinical Notes in Electronic Medical Records. JAMA Network Open. 2024;7(8):e2425953-e. doi: 10.1001/jamanetworkopen.2024.25953.
16. Giacobbe DR, Marelli C, Guastavino S, Mora S, Rosso N, Signori A, et al. Explainable and Interpretable Machine Learning for Antimicrobial Stewardship: Opportunities and Challenges. Clin Ther. 2024;46(6):474-80. doi: 10.1016/j.clinthera.2024.02.010.
17. Zaidat B, Shrestha N, Rosenberg AM, Ahmed W, Rajjoub R, Hoang T, et al. Performance of a Large Language Model in the Generation of Clinical Guidelines for Antibiotic Prophylaxis in Spine Surgery. Neurospine. 2024;21(1):128-46. doi: 10.14245/ns.2347310.655.

18. 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. 2024;31(1). doi: 10.1136/bmjhci-2023-100978.
19. Maillard A, Micheli G, Lefevre L, Guyonnet C, Poyart C, Canoui E, et al. Can Chatbot Artificial Intelligence Replace Infectious Diseases Physicians in the Management of Bloodstream Infections? A Prospective Cohort Study. Clin Infect Dis. 2024;78(4):825-32. doi: 10.1093/cid/ciad632.
20. De Vito A, Geremia N, Marino A, Bavaro DF, Caruana G, Meschiari M, et al. Assessing ChatGPT's theoretical knowledge and prescriptive accuracy in bacterial infections: a comparative study with infectious diseases residents and specialists. Infection. 2024. doi: 10.1007/s15010-024-02350-6.
21. Kwong JCC, Wang SCY, Nickel GC, Cacciamani GE, Kvedar JC. The long but necessary road to responsible use of large language models in healthcare research. NPJ Digit Med. 2024;7(1):177. doi: 10.1038/s41746-024-01180-y.
22. Brodeur PG, Buckley TA, Kanjee Z, Goh E, Bin Ling E, Jain P, et al. Superhuman performance of a large language model on the reasoning tasks of a physician. arxiv 2024. https://arxiv.org/abs/2412.10849.
23. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. JAMA Network Open. 2024;7(10):e2440969-e. doi: 10.1001/jamanetworkopen.2024.40969.