REVIEW

Rick Rejeleene, PhD
Special Fellow for Artificial Intelligence, Cleveland Clinic, Cleveland, OH

Neil B. Mehta, MBBS, MS
Professor, Department of Medicine, and Associate Dean for Curricular Affairs, Cleveland Clinic Lerner College of Medicine of Case Western Reserve University, Cleveland, OH; Director, Center for Technology-Enhanced Knowledge and Instruction, Cleveland Clinic, Cleveland, OH; Co-Chair, AI Taskforce, Case Western Reserve University School of Medicine, Cleveland, OH

Artificial intelligence in medicine: How it works, how it fails

ABSTRACT

Artificial intelligence (AI) is transforming healthcare, with large language models emerging as important tools for clinical practice, education, and research. To use it safely and effectively, healthcare professionals need to understand how it works, and how it fails. Using practical clinical examples, the authors explain the subset of AI called large language models, highlighting their capabilities and their limitations.

KEY POINTS

AI is trained on vast amounts of data, which can itself be biased, leading to biased results.

AI essentially predicts the most probable sequence of words to complete a sentence. If the training data are ambiguous or incomplete, or if the query is outside the model's core knowledge, it might guess or infer information based on weak statistical signals, leading it to create plausible but untrue statements (hallucinations).

The way a question or prompt is phrased can significantly affect the response. Clinicians should thus triangulate answers by asking the same clinical question in 2 different ways to see if the model's reasoning remains consistent.

While large language models can enhance efficiency and clinical decision-making, they must be integrated with a human in the loop to ensure safe, ethical practice.

doi:10.3949/ccjm.93a.25089

Artificial intelligence (AI) is increasingly being integrated into education, clinical practice, and research. Clinicians today may encounter it through general-purpose tools like ChatGPT, automated documentation assistants like Ambience, or clinical decision support tools like Open Evidence.

We believe that AI can help us in our jobs as doctors—but with important caveats. We don't all have to be computer scientists, but we do need to know a little bit about AI to effectively and safely integrate it into practice, discerning when to rely on what it says, when to be skeptical, and when to strategically leverage its capabilities.[1]

Here, we try to explain AI, how it has evolved, how it works,[2] what it is good for, and critically, the specific risks, such as hallucination and bias, that can lead to clinical errors. We will use some clinical examples, such as cases of community-acquired pneumonia, to ground our discussion.

HOW ARTIFICIAL INTELLIGENCE HAS EVOLVED

Over the years, AI has progressed from rule-based systems, to machine learning, to deep learning, to large language models (**Figure 1**). Let's consider how each of these systems might help a clinician choose the best antibiotic for a patient with community-acquired pneumonia.

Rule-based systems are like very detailed flowcharts. We would program the computer with specific "if-then" instructions based on established clinical guidelines and local resistance patterns. For example, the system might

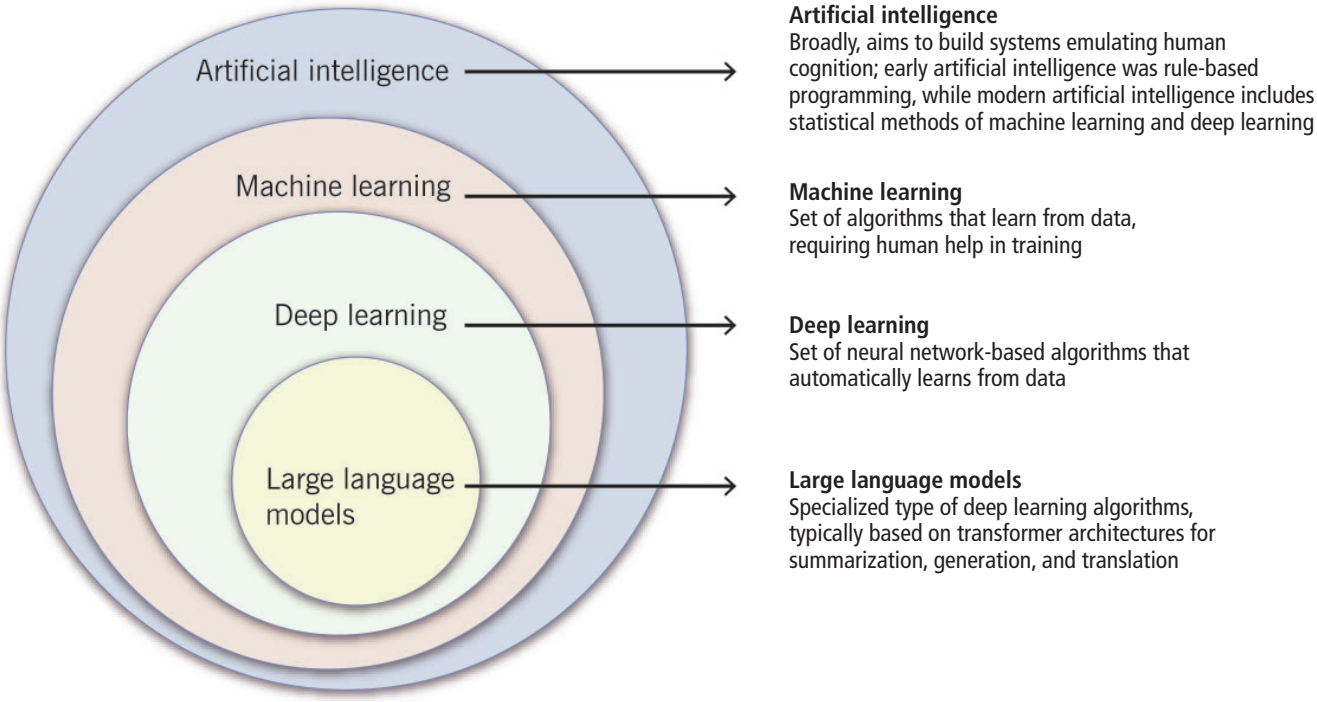


Figure 1. Different types of artificial intelligence.

be told: "IF the patient is over 65 AND has kidney problems, THEN suggest antibiotic X."

The major downside of these systems is their inflexibility: they can only handle situations they've been explicitly programmed for. If a new guideline comes out, or a patient presents with a slightly different set of symptoms not covered by a rule, or new antibiotics or bacterial resistance patterns emerge, then manual reprogramming is required.

Traditional machine learning systems came next.[3] Instead of being given rigid rules, these systems learn by analyzing vast amounts of past patient data.

To continue with our example, it's like showing the computer thousands of charts from patients with community-acquired pneumonia, with expert-selected variables that might influence the choice of correct antibiotics. We'd feed it data such as patient age, specific symptoms, and other health conditions, along with the successful antibiotic prescribed. The computer then finds its own patterns in that data to develop a model that provides antibiotic recommendations, learning which combinations of these factors tend to lead to a good outcome. This process of humans selecting the input or predictor variables to build a model is called *feature engineering*.

The advantage here is that the system can uncover connections that might not be obvious to humans and can continue to update itself over time.[4] However, the drawback is that feature engineering is labor-intensive and susceptible to "unknown unknowns." If a rare condition (eg, a specific genetic contraindication) was not included as a feature during training, the model will not account for it in future predictions.

Deep learning systems are the most advanced, often thought of as having a more brain-like structure.[5] What makes them powerful is their ability to learn directly from raw, messy data without the need for feature engineering. For our community-acquired pneumonia example, a deep learning system could analyze entire patient files, including the actual chest radiography images,[6] the free-text notes written by doctors and nurses,[7] and all the laboratory numbers and then figure out *on its own* which parts of this information are most important.

The huge advantage of deep learning systems is their ability to find complex and subtle patterns in vast and varied datasets, leading to highly accurate suggestions.[8] However, their primary drawbacks are that they require enormous amounts of data and processing power to learn effectively, and sometimes they act like a "black box"—it can be hard to understand exactly *why* they made a particular recommendation, which can be a concern in critical healthcare decisions. Large language models are a specialized subset of deep learning.

FEATURES IN, LABELS OUT

Note that when discussing AI we use the words "features" and "labels" in unconventional ways.

Features (input or predictive or independent variables) are like all the pieces of clinical information one gathers about a patient that help you choose an antibiotic. For a patient with suspected community-acquired pneumonia, these would be things like their age, their comorbidities or allergies, and their severity of illness.

The label (outcome or dependent variable) is the outcome you're trying to predict for that patient. It's the diagnosis or the most effective treatment that was determined for similar past patients, and what you want the AI model to learn to identify. So, for our community-acquired pneumonia example, the label for a past patient might be "antibiotic X was the most effective treatment for this specific case." The AI model learns from thousands of these features and their corresponding labels to predict the right label for a new patient.

Critical caveat: labels often reflect historical clinician behavior rather than objective truth. If the training data reflect that physicians historically prescribed antibiotic X most frequently, even if it wasn't the optimal evidence-based choice, the AI system will learn to mimic this habit rather than the best clinical practice.

MACHINE LEARNING: SUPERVISED, UNSUPERVISED, AND REINFORCED

Supervised learning trains the model on a dataset that provides both features and labels, eg, past cases of community-acquired pneumonia for which the correct antibiotic choice is known.[9] This model can then be used to predict the best antibiotic for future patients.

Unsupervised learning provides the model with features but not the labels. The model attempts to find patterns or groupings in the data, such as identifying clusters of infection types without knowing the correct antibiotic. This could reveal previously unrecognized subgroups of patients with community-acquired pneumonia that respond similarly or uniquely to certain interventions, potentially guiding new ways to categorize and treat patients more precisely.

Reinforcement learning trains an AI model to make decisions by learning from trial and error and continuous feedback on past actions, much like suggesting a series of moves in a game of chess as opposed to just the first move. For instance, in treating pneumonia, the AI model might suggest antibiotics in a simulated setting, then receive feedback on success or adverse events. This allows it to learn a dynamic strategy, providing evolving guidance throughout a patient's illness to optimize long-term outcomes, not just an initial recommendation.

Each learning type offers unique strengths, with supervised learning being widely used in clinical prediction, unsupervised learning helping discover unknown patterns, and reinforcement learning holding promise for adaptive decision-making in dynamic clinical environments.

NEURAL NETWORKS: THE ARCHITECTURES OF DEEP LEARNING

Each machine learning approach or learning framework uses layers of artificial neural networks arranged in specific designs and organizations, termed *architectures*, that mimic aspects of human learning. *Transformers* are one such architecture, and large language models are based on them. **Table 1** lists common machine learning architectures and their applications in healthcare; while most of these architectures are beyond the scope of this article, it is important for clinicians to be familiar with the breadth of AI applications in healthcare.

Transformers and attention

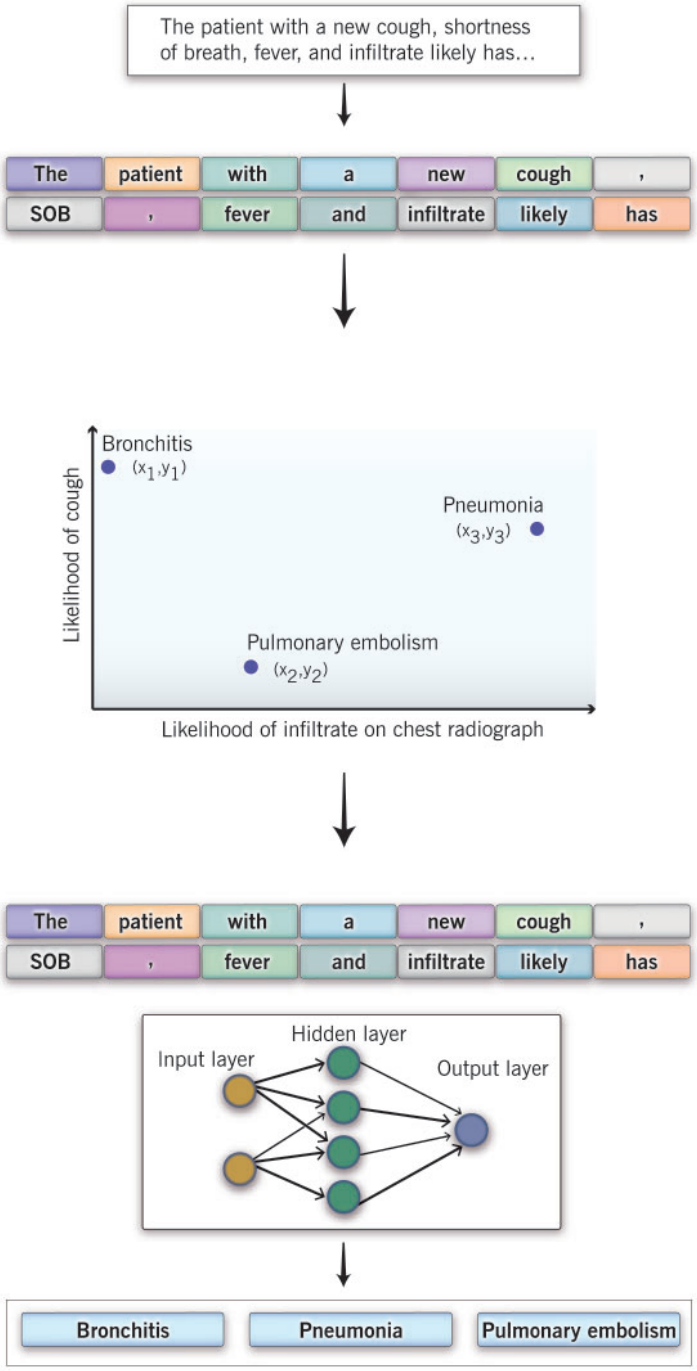
Transformers are designed to understand information that comes in a sequence, like words in a sentence or events over time.[2] They do this by figuring out how important each part of the input is to every other part—called the "attention mechanism"—allowing them to grasp the full meaning, even when pieces of information are far apart. The process goes through several steps—tokens, embedding, transformers, and feedback (**Figure 2**).

Imagine an expert clinician listening to a medical student presenting a new admission to the hospital. As they listen, they can identify and prioritize the key features that will help determine the diagnosis. This ability to weigh the importance of different pieces of information and focus on the most relevant parts of the input is like the attention mechanism of the transformer architecture.[10]

For example, when a large language model processes a sentence like, "The patient presented with cough and new infiltrate on chest x-ray, suggesting possible community-acquired pneumonia that may require antibiotic treatment," the attention mechanisms in transformers allow it to weigh the importance of each word. It recognizes that "new infiltrate on chest x-ray" is highly relevant to "cough" in the context of potential community-acquired pneumonia and establishes connections across the sentence to generate clinically appropriate interpretations.

TABLE 1
Architectures of neural networks in machine learning: What can they do?

Architectures	Uses and limitations in healthcare
Supervised	
Feedforward neural networks or multilayer perceptrons Data move in 1 direction through layers of "neurons," each fully connected to the next; the network learns by adjusting the strength of connections	Can be trained on structured patient data (eg, age, comorbidities, vaccination status) to predict a patient's individual risk of developing community-acquired pneumonia or experiencing a severe outcome, guiding preventive measures or early intervention While powerful, are less specialized for image or sequence data than other architectures
Convolutional neural networks Use filters that scan across the image to detect features like edges, shapes, and textures, the way your eye recognizes parts of a picture	Can be trained to quickly spot signs of pneumonia on a chest radiograph or computed tomography image, determine the extent of lung involvement, or even help assess the severity of the infection by identifying specific patterns, aiding in faster diagnosis and treatment planning Excellent at finding patterns in data that look like a grid, such as images Training data need to include normal studies and other thoracic conditions that mimic pneumonia so that the system can learn to distinguish pneumonia from look-alike abnormalities
Recurrent neural networks Designed to understand sequences of information where order matters Have a memory that allows them to consider past events when processing current information, making them ideal for data that unfold over time	Can monitor a patient's vital signs (eg, heart rate, respiratory rate), laboratory values, and symptom progression over hours or days Can learn patterns that predict worsening community-acquired pneumonia, like an increasing risk of sepsis or the need for intensive care, enabling earlier intervention
Transformers Excel at understanding complex relationships within long sequences of data (like text), even when important pieces of information are far apart Can process entire sequences at once, efficiently grasping the context and significance of each part	Can read and understand lengthy, unstructured physician's notes, nursing observations, and discharge summaries related to community-acquired pneumonia and then summarize key findings, extract all comorbidities, or identify subtle mentions of medication side effects, providing a comprehensive patient overview or flagging relevant clinical details for review
Unsupervised	
Autoencoders or variational autoencoders Learn to compress complex data into a simplified, meaningful "code" and then reconstruct similar data from that code Are particularly good at finding and representing underlying patterns in data; are often used for data compression, anomaly detection, and generating new data with learned characteristics	Can analyze vast clinical datasets (symptoms, laboratory results, demographics, treatment responses) for patients with community-acquired pneumonia without prior labels Could uncover hidden patient groups (eg, those with a specific inflammatory response that predicts poorer outcomes) that might benefit from targeted therapies, guiding personalized medicine approaches
Generative adversarial networks Composed of 2 competing artificial intelligence networks, one that creates new, realistic data (the "generator") and another that tries to tell whether the data are real or fake (the "discriminator"); through this competition, they learn to generate highly convincing, novel data	Can create realistic but entirely synthetic computed tomography images showing various stages of pneumonia, which can be invaluable for training other artificial intelligence models when real patient data are scarce or for creating diverse educational materials without compromising patient privacy Could also generate anonymized patient records for research
Supervised or unsupervised	
Graph neural networks Designed to understand and learn from data that are structured as a network or graph, where elements (nodes) are connected by relationships (edges); can analyze how information flows and interacts within these complex connections	Could model patient flow within a hospital, identifying high-risk infection transmission pathways for community-acquired pneumonia More broadly, can analyze complex networks of patient conditions, medications, and clinical events, helping to understand how different comorbidities influence community-acquired pneumonia outcomes or how specific antibiotics interact within a patient's overall drug regimen



Tokens are individual words, syllables, or punctuations that a body of text gets split into so a computer can work with it.

Embedding turns tokens into numbers that capture their meaning so that a computer can compare and work with them. Here, the words *bronchitis*, *pulmonary embolism*, and *pneumonia* might hypothetically be assigned numbers.

A transformer learns to pay attention to the important parts of a sentence so that it can understand and generate language. Here, the model may generate 3 words (with different probabilities) that best fit for completing the sentence. With feedback, it would be trained to suggest that *pneumonia* has the highest probability as the next word in the sentence.

Figure 2. How artificial intelligence chooses the next word—and makes a diagnosis.

Caveat. While transformers are excellent at understanding relationships within text, they can sometimes struggle with nuanced language, especially negation. For instance, if a physician asks a large language model about a patient's presentation, saying, "The patient has a cough, but denies fever or shortness of breath," a transformer-based model might focus strongly on "cough" and the general context of "pneumonia symptoms." It could then incorrectly suggest a higher likelihood of community-acquired pneumonia by

failing to adequately process the denial of fever and shortness of breath. This occurs because the model's attention mechanism primarily learns statistical connections between words, rather than a human-like logical comprehension of the denial.

Transformers learn statistical correlations rather than symbolic logic. If a prompt says, "The patient is on high-dose corticosteroids and has a cough but no fever," a human knows steroids suppress fever. However, unless the model was explicitly trained on many examples linking steroids to afebrile infection, it might statistically associate infection with fever. It could then incorrectly lower the probability of pneumonia because the word "fever" is absent, failing to account for the physiologic effect of the drug.

HOW MACHINES LEARN

Pretraining

Large language models are initially trained on massive datasets of digital information from the Internet. This process, called *pretraining*, is analogous to a medical student going through years of medical school, reading textbooks, research papers, and clinical notes to acquire general medical knowledge before seeing patients.[11]

A large language model pretrained on a large corpus of medical literature will learn about various diseases, common symptoms, and treatments.[12] It learns that pneumonia often presents with cough, fever, and shortness of breath, and that antibiotics are a common treatment. It would learn about typical organisms causing community-acquired pneumonia and their clinical presentations.

But it can make mistakes. Pretraining exposes large language models to a wide range of medical information, but the data are not exhaustive. As a result, large language models may struggle with diseases that are newly emergent, rare, or have an atypical presentation.

For example, a large language model that is pretrained on data up to late 2024 may not know about cases of avian influenza in the United States. When asked about the differential diagnosis of a dairy farmer in California with influenza-like illness, it would not suggest an H5N1 infection.

For another example, *Chlamydia psittaci* pneumonia can present with a headache and rash and without classic features of community-acquired pneumonia. If the pretraining dataset has very few cases of this condition, the model may not learn robust associations for this. Unless the user specifically mentions exposure to a sick parrot, the model would likely not suggest *Chlamydia psittaci* as a differential for this presentation.

Applying weights and biases

During pretraining, the model adjusts internal numerical parameters called weights and biases, which determine the strength of connections between words and concepts.

For example, if a large language model is trained on many reports in which hypertension is associated with increased risk of stroke, the weight connecting these 2 concepts increases. The values of these weights and biases are crucial for the large language model's accuracy.[13] If the training data are biased or incomplete, the weights and biases may be adjusted in a way that leads to incorrect or suboptimal responses.

But the weights may be wrong. Consider a large language model trained to predict the likelihood that a patient has a pulmonary embolism based on various clinical factors. If the training data come primarily from tertiary care hospitals, the model might be biased toward severe diagnoses. For example, if "calf pain" in the training data is almost always associated with pulmonary embolism rather than muscle strain, the model might aggressively suggest pulmonary embolism even for a young, active patient with a clear history of strain. Here, the model's parameters have been skewed by the prevalence of a particular association in the training data, resulting in an erroneous prediction in a different clinical context.

Fine-tuning

After pretraining, large language models can be further trained on specific datasets to improve their performance on particular tasks. *Fine-tuning* adapts a general-purpose large language model to perform better in specific domains. It is analogous to a medical resident trained on all of internal medicine joining a cardiology fellowship and focusing on heart-related conditions.

Fine-tuning improves accuracy. For instance, a general large language model pretrained on a broad range of Internet text might be asked, "What is the most common cause of elevated troponin in a patient without chest pain?" The large language model might generate a list of common causes of elevated troponin, such as myocardial infarction, but may not rank them in the correct order of probability for a patient without chest pain. It might overemphasize myocardial infarction because, in general, that's a very common cause of troponin elevation.

The large language model could be fine-tuned on a large dataset of electronic health records from patients presenting with elevated troponin but without chest pain. This dataset would include cases of myocarditis, pulmonary embolism, renal failure, and other less com-

mon causes in this specific patient population. After fine-tuning, the large language model learns the subtle differences in troponin elevation in the absence of chest pain. It adjusts the weights and biases in its neural network to prioritize conditions like myocarditis and pulmonary embolism in this specific clinical context. When asked the same question, the fine-tuned large language model provides a more accurate, clinically relevant answer, prioritizing causes appropriately for the patient's presentation.

Retrieval augmented generation

Retrieval augmented generation enhances the foundational large language models by enabling access to external knowledge sources.[14] When a user asks a question, the large language model first retrieves relevant information from a database or knowledge base and then uses this information to generate more accurate, informed responses. This process is analogous to a clinician consulting a resource such as a textbook or guideline before making a clinical decision.

For example, a physician asks a large language model, "What is the best treatment for a patient with heart failure with reduced ejection fraction?" The large language model uses retrieval augmented generation to search a database of clinical guidelines, retrieves the latest recommendations, and then provides a response based on that information, rather than relying solely on pretrained knowledge.

Risks and limitations. While the retrieval augmented generation model uses appropriate resources such as clinical guidelines, it still must interpret the information, which is where it may make errors. Thus, in a patient with a history of "mild penicillin allergy" in the electronic health record, the guideline would suggest using a medication such as levofloxacin. The large language model may misinterpret "mild" or the "low" risk of cross-reactivity with another beta-lactam antibiotic and suggest using a cephalosporin with azithromycin. This is an issue of the large language model using statistical inference rather than prioritizing patient safety.

For another example, a patient developed bromism after consulting ChatGPT and following its medical advice to take bromide supplements.[15] This highlights the risk of relying on large language models for medical decisions and requires careful clinical evaluation of unusual symptoms. While large language models can enhance efficiency and clinical decision-making, they must be integrated with a human in the loop to ensure safe, ethical practice.[16]

POTENTIAL PITFALLS OF ARTIFICIAL INTELLIGENCE

Hallucinations: When AI makes stuff up

Large language models can generate information that is not present in the training data or any external knowledge source. This is often called *hallucination*.[17]

When a large language model generates a response, it is essentially predicting the most probable sequence of words to complete a sentence. If the training data are ambiguous or incomplete, or if the query is outside its core knowledge, the model might "guess" or infer information based on weak statistical signals, leading it to create plausible but untrue statements.[17]

For example, a clinician asks a large language model to list the common side effects of a new antibiotic to treat community-acquired pneumonia. The large language model provides a list that includes side effects not reported in any clinical trials or postmarketing surveillance.

It is thus necessary to double-check primary evidence before applying a large language model–generated response to clinical care. Explicitly prompting the model with, "*Answer using only information from professional society guidelines. If you do not know, state that you do not know. Do not fabricate references.*" can help reduce hallucinations but not completely eliminate them.

Prompt sensitivity: What you ask is what you get

The way a question or prompt is phrased can significantly affect the response.[18] Large language models learn statistical patterns from vast amounts of text data. They don't possess genuine understanding[19] or conscious reasoning. When you change a prompt, even slightly, it shifts the statistical probabilities that the large language model uses to generate its next words. This can nudge the model down a different path in its vast learned knowledge space, leading it to emphasize different information, omit crucial details, or even present a different conclusion.

For example, the following 2 prompts might give different responses.

- "What are the recommended empiric antibiotics for outpatient community-acquired pneumonia in an adult without comorbidities?"
- "What are the recommended empiric antibiotics for outpatient community-acquired pneumonia in an adult without comorbidities? *Please specifically address optimal choices, given the prevalence of macrolide resistance in many regions.*"

The second prompt might elicit a different response that deemphasizes azithromycin as empiric treatment. Clinicians should thus triangulate answers by asking the same clinical question in 2 different ways to see if the model's reasoning remains consistent.

WHAT ARE AGENTS?

Unlike large language models that give a single answer to a question or prompt, an *agent* can autonomously perceive its environment, reason about its observations, formulate plans, take actions (often by using various digital tools and resources), and continuously evaluate its progress toward a defined goal.[20]

Agents are very useful for complex tasks that might be run multiple times.[21] For example, take a patient with multiple complex problems including renal impairment, on multiple medications, admitted with community-acquired pneumonia. The agent would look up the clinical guideline for the most appropriate medication, then check the medical records for renal function and the medication list, then check an online drug-interaction checker, and then recommend a medication that is dosed appropriately for the patient's renal function.

EMPOWERING HEALTHCARE PROFESSIONALS

In this article, we have tried to use a clinical lens to demystify AI, particularly large language models, by illustrating their core mechanisms, learning paradigms, and pivotal concepts like attention, pretraining, weights and biases, and retrieval-augmented generation in the context of practical clinical examples. We also illuminated potential pitfalls such as hallucinations and prompt sensitivity that clinicians must carefully consider.

Armed with this understanding, we hope clinicians will be empowered to critically evaluate AI-generated information, judiciously integrate these powerful tools into their practice, and ultimately deliver care that is not only safer and more efficient but also more effective, truly augmenting human expertise for the benefit of patients.

DISCLOSURES

The authors report no relevant financial relationships which, in the context of their contributions, could be perceived as a potential conflict of interest.

REFERENCES

1. **Deng J, Heybati K, Park YJ, Zhou F, Bozzo A.** Artificial intelligence in clinical practice: a look at ChatGPT. Cleve Clin J Med 2024; 91(3):173–180. doi:10.3949/ccjm.91a.23070
2. **Vaswani A, Shazeer N, Parmar N, et al.** Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Red Hook, NY: Curran Associates Inc.; 2017:6000–6010.
3. **Reddy S.** Artificial intelligence and healthcare—why they need each other? J Hosp Manag Health Policy 2021; 5:9.
4. **El Naqa I, Murphy MJ.** What is machine learning? In: El Naqa I, Li R, Murphy MJ, eds. Machine Learning in Radiation Oncology. Cham, Switzerland: Springer International Publishing; 2015:3–11.
5. **Rajpurkar P, Chen E, Banerjee O, Topol EJ.** AI in health and medicine. Nat Med 2022; 28(1):31–38. doi:10.1038/s41591-021-01614-0
6. **Mitsuyama Y, Tatekawa H, Takita H, et al.** Comparative analysis of GPT-4-based ChatGPT's diagnostic performance with radiologists using real-world radiology reports of brain tumors. Eur Radiol 2025; 35(4):1938–1947. doi:10.1007/s00330-024-11032-8
7. **Yang X, Chen A, PourNejatian N, et al.** A large language model for electronic health records. NPJ Digit Med 2022; 5(1):194. doi:10.1038/s41746-022-00742-2
8. **Bubeck S, Chandrasekaran V, Eldan R, et al.** Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv:2303.12712v5. Last revised April 13, 2023.
9. **Jordan MI, Mitchell TM.** Machine learning: Trends, perspectives, and prospects. Science 2015; 349(6245):255–260. doi:10.1126/science.aa8415
10. **Zhao WX, Zhou H, Li J, et al.** A survey of large language models. arXiv:2303.18223v16. Last revised March 11, 2025.
11. **Singhal K, Azizi S, Tu T, et al.** Large language models encode clinical knowledge. Nature 2023; 620(7972):172–180. doi:10.1038/s41586-023-06291-2
12. **Kaczmarczyk R, Wilhelm TI, Martin R, Roos J.** Evaluating multimodal AI in medical diagnostics. NPJ Digit Med 2024; 7(1):205. doi:10.1038/s41746-024-01208-3
13. **Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW.** Large language models in medicine. Nat Med 2023; 29(8):1930–1940. doi:10.1038/s41591-023-02448-8
14. **Lewis P, Perez E, Piktus A, et al.** Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv:2005.11401v4. Last revised April 12, 2021.
15. **Eichenberger A, Thielke S, Van Buskirk A.** A case of bromism influenced by use of artificial intelligence. AIM Clinical Cases 2025; 4:e241260. Epub 5 August 2025. doi:10.7326/aimcc.2024.1260
16. **Bakken S.** AI in health: keeping the human in the loop. J Am Med Inform Assoc 2023; 30(7):1225–1226. doi:10.1093/jamia/ocad091
17. **Valmeekam K, Olmo A, Sreedharan S, Kambhampati S.** Large language models still can't plan (a benchmark for LLMs on planning and reasoning about change). OpenReview. openreview.net/pdf?id=wUU-7XTL5XO. Accessed January 16, 2026.
18. **Wei J, Wang X, Schuurmans D, et al.** Chain-of-thought prompting elicits reasoning in large language models. arXiv:2201.11903v6. Last revised January 10, 2023.
19. **Kambhampati S.** Can large language models reason and plan? Ann N Y Acad Sci 2024; 1534(1):15–18. doi:10.1111/nyas.15125
20. **Weng L.** LLM powered autonomous agents. Lil'Log. June 23, 2023. lilianweng.github.io/posts/2023-06-23-agent/. Accessed January 16, 2026.
21. **Tu T, Azizi S, Driess D, et al.** Towards generalist biomedical AI. arXiv:2307.14334v1. Submitted on July 26, 2023.

Address: *Rick Rejeleene, PhD, Special Fellow for Artificial Intelligence, Cleveland Clinic, 9500 Euclid Avenue, Cleveland, OH 44195; rick.rejeleene@gmail.com*