THE REMOTE AI DOC

AI Literacy for Clinicians

What Every Physician Needs to Know Before Trusting the Algorithm

Dr. Javier Rosas

Ship Physician · Clinical AI Specialist

EDITION 1.0

AI Literacy for Clinicians: What Every Physician Needs to Know Before Trusting the Algorithm

What Every Physician Needs to Know Before Trusting the Algorithm

Javier Rosas

This book is available at https://leanpub.com/AiforDoctors

This version was published on 2026-02-15

This is a Leanpub book. Leanpub empowers authors and publishers with the Lean Publishing process. Lean Publishing is the act of publishing an in-progress ebook using lightweight tools and many iterations to get reader feedback, pivot until you have the right book and build traction once you do.

© 2026 Javier Rosas

To my daughter Valery, who reminds me every day why getting this right matters.

Contents

What Happens When You Talk to AI

> *It is 2:47 AM. You are the only physician on a cruise ship carrying 4,800 passengers somewhere in the South Pacific. A 62-year-old man presents with acute-onset chest pain, diaphoresis, and a blood pressure of 88/54. Your nearest cardiac catheterization lab is 14 hours away by helicopter, if one is even available. You open a general-purpose AI chatbot on your laptop, not a regulated clinical decision support tool, but the same large language model millions of people use every day, and type: "Management of suspected STEMI in resource-limited setting, no PCI capability, limited formulary."*

Within seconds, the AI generates a detailed, confident response. It reads like a well-written clinical reference. But here is the question most clinicians never think to ask: what just happened inside that system between the moment you pressed enter and the moment those words appeared on your screen?

This chapter answers that question. Not with computer science jargon, but with the clinical clarity you would expect from a colleague explaining a diagnostic tool you are about to rely on. Because if you are going to use AI in your practice, and many of you already are, you need to understand what it is actually doing. Not to become an engineer, but to become a better judge of when to trust it, when to question it, and when to override it entirely.

A Note on Terminology

Throughout this book, it is important to distinguish between two categories of AI that clinicians encounter. **General-purpose large language models**, such as ChatGPT, Claude, and Gemini, are designed for broad conversational use. They have no FDA clearance, no validated clinical performance data, and no intended medical use. When a clinician uses one for a clinical question, that is off-label use, and the clinician assumes full liability for any decision that follows. **Clinical decision support (CDS) tools**, by contrast, may be built on similar underlying technology but are designed for specific medical tasks, may carry FDA 510(k) clearance, and come with documented performance characteristics and intended use cases. The distinction matters for liability, for trust calibration, and for understanding when the output you are reading has been validated against clinical outcomes and when it has not. Most of the scenarios in this book involve general-purpose LLMs, because that is what most clinicians are actually using today.

* * *

The Most Important Thing to Understand

Large language models, the technology behind ChatGPT, Claude, Gemini, and most clinical AI tools, do not think in the way clinicians mean when they describe medical reasoning. They do not reason through differential diagnoses. They do not understand your patient. What they do, at the most fundamental level, is predict the next word.

That sentence is worth pausing on, because it contradicts almost everything the output suggests. When an AI generates a paragraph about managing acute coronary syndrome, it reads like someone who understands cardiology. The formatting is clean. The drug names are correct. The dosing often checks out. But the mechanism that produced those words is not medical reasoning. It is statistical pattern completion operating at extraordinary scale.

Think of it this way. You have used the predictive text feature on your phone. When you type “I’ll be there in” your phone suggests “5 minutes” or “an hour”

because those completions are statistically common in the text it has seen. A large language model does exactly the same thing, but instead of drawing from your message history, it draws from a training dataset that includes billions of pages of text: medical textbooks, journal articles, clinical guidelines, online forums, Wikipedia, and vast quantities of general internet content.

Or consider another analogy. A search engine is like a librarian who knows where every book is shelved and can point you to the exact page. An LLM is like a colleague who has read every book in the library, returned them all, and is now reconstructing an answer from memory. The knowledge is real. The recall is often impressive. But the retrieval mechanism is entirely different, and that distinction has profound implications for how much you should trust the output.

When you ask it about STEMI management, it is not consulting Harrison's or recalling a case it managed. It is calculating, word by word, what text is most likely to follow your question based on patterns it absorbed during training. The output looks like medical knowledge because it was trained on medical knowledge. But the process is fundamentally different from how you arrive at a clinical decision.

Clinical Parallel

When you recall the management of diabetic ketoacidosis, you are not reading a textbook in your mind. You are reconstructing knowledge from training, experience, and pattern recognition. LLMs do something structurally similar, but without the clinical judgment to know when their reconstruction has gone wrong.

* * *

Why This Matters Clinically

This distinction is not academic. It has direct implications for how you should interpret AI-generated clinical content. A system that predicts the next word can produce text that is fluent, well-structured, and entirely wrong. It can

recommend a drug that does not exist. It can cite a study that was never published. It can generate a treatment plan that sounds textbook-perfect but contradicts your patient's specific clinical context. It does all of this with the same confident tone, because confidence is not a reflection of accuracy. It is a reflection of how common that phrasing pattern was in the training data.

It is worth distinguishing between three layers of confidence in AI-generated output. The first is **surface confidence**: the fluent, authoritative tone that is always present regardless of whether the content is accurate. The model does not hedge because it is certain. It does not hedge because hedging was not the most probable next-token pattern for that particular sequence. The second is **expressed uncertainty**: when prompted specifically, such as "how confident are you in this recommendation," some models can generate statements of uncertainty. This can be a useful technique, and clinicians should know it exists. The third is **actual calibration**, meaning whether the model's stated confidence matches its real accuracy. Here the evidence is concerning. Studies consistently show that LLMs are poorly calibrated on medical questions, meaning they express high confidence on questions they answer incorrectly at roughly the same rate as questions they answer correctly. The tone is not a signal. It is noise.

In my practice, I have tested clinical AI tools with ship-specific scenarios: acute presentations where the recommended intervention requires resources I do not have, drugs that are not on my formulary, or specialist referrals that are physically impossible a thousand miles from shore. The AI does not flag these constraints. It generates the standard answer because the standard answer is what appears most frequently in its training data. The gap between what the AI recommends and what I can actually do is where the risk lives, and understanding why that gap exists starts with understanding this fundamental mechanism.

* * *

How the Model Learned What It Knows

Before a large language model can predict anything, it has to be trained. Training is the process by which the model absorbs patterns from text. The simplest way to think about it is this: the model reads an enormous amount of

text with words randomly hidden, and it practices guessing what the hidden words should be. Over billions of repetitions, it develops an internal statistical map of how language works, which words tend to follow which other words, how medical sentences are typically structured, what drug names appear alongside which conditions.

This is where the analogy to clinical training is useful, but also where it breaks down in important ways.

A medical resident builds knowledge through a combination of didactic learning, clinical observation, supervised practice, and reflective feedback. When you encounter a patient with chest pain, you are drawing on pathophysiology you studied, ECGs you have read, patients you have managed, and attending feedback that corrected your reasoning in real time. Your knowledge is grounded in causal understanding. You know *why* nitroglycerin reduces preload, not just that it is associated with chest pain management in text.

An LLM has no causal understanding. It has statistical associations. It knows that the words "nitroglycerin," "sublingual," "chest pain," and "preload reduction" tend to appear near each other in medical texts. That correlation is often sufficient to produce correct output. But it also means the model has no mechanism to distinguish between a well-established treatment and an outdated one, between a guideline-concordant recommendation and a confidently stated error, unless the patterns in its training data happen to make that distinction clear.

* * *

The Training Data Problem

This leads to a critical point about training data. The quality of what an LLM produces is bounded by the quality of what it was trained on. Most frontier models are trained on a mixture of curated and uncurated sources: peer-reviewed literature alongside blog posts, clinical guidelines alongside patient forums, FDA labels alongside promotional pharmaceutical content.

It is fair to note that model developers have made significant efforts to address this. Techniques like reinforcement learning from human feedback

(RLHF), constitutional AI methods, and active data curation attempt to up-weight reliable sources and reduce harmful output during training and fine-tuning. These efforts are real, and they have measurably improved model behavior. However, they are imperfect, largely opaque to end users, and do not eliminate the fundamental problem: the model has no mechanism to verify claims against reality, only to reproduce patterns associated with perceived authority. A model that has been trained to sound more like a peer-reviewed journal is not the same as a model that has been validated against clinical outcomes.

The scale of training data is difficult to grasp intuitively. Models like GPT-4, Claude, and Gemini were trained on datasets measured in trillions of tokens (a token is roughly a word fragment, so this represents hundreds of billions of words of text). To put that in perspective, the entire contents of PubMed, every abstract ever indexed, would represent a small fraction of a percent of a modern training corpus. The model has, in a statistical sense, processed more medical literature than any physician could consume in a hundred lifetimes. But volume is not the same as curation, and this is where clinicians need to pay close attention.

For clinical content, this creates a specific risk. If a drug interaction that is well-documented in pharmacology references but rarely discussed in general medical text, the model may underweight it. If a common but incorrect practice appears frequently in online discussions, the model may reproduce it confidently. The training data is not a curated medical library. It is a reflection of what exists on the internet, with all the noise, bias, and inconsistency that implies.

There is an additional nuance that matters for clinicians specifically. Most LLMs have a knowledge cutoff date, a point after which they have not seen new information. If a major guideline update was published after that date, the model does not know about it. It will answer based on the prior version with the same confidence. It will not flag that its knowledge may be out of date. It will not suggest you check for recent updates. It will simply give you the best answer it can based on the data it has, which may no longer be the right answer.

Clinical Parallel

Imagine a medical student who studied by reading every document in the hospital's archive without anyone telling them which ones were the current evidence-based protocols and which were the outdated notes from a retired physician's personal filing cabinet. They would accumulate enormous knowledge. They would also accumulate an unpredictable number of confident misconceptions. And they would have no reliable way to distinguish between the two.

* * *

How AI Reads Your Question

When you type a clinical question into an AI tool, the first thing that happens is tokenization. The model does not read words the way you do. It breaks your input into smaller units called tokens, which are roughly equivalent to word fragments. The word "nitroglycerin" might be split into "nitro," "glyc," and "erin." Common words like "the" or "is" are typically single tokens. Medical terminology, because it is less common in general text, often gets split into multiple tokens.

This matters more than it might seem. The model processes each token in relation to every other token in the input using a mechanism called attention. In simple terms, the model calculates how much each part of your question relates to every other part, building a contextual map of what you are asking. When you write "STEMI management in resource-limited setting," the model assigns attention weights that connect "STEMI" to "management," but also to "resource-limited," which should modify the type of management it suggests.

Whether it actually modifies its response based on that constraint depends on how strongly that pattern exists in the training data. If most of the text the model saw about STEMI management was written in the context of hospitals with full PCI capability, then "resource-limited" may not shift the output as much as you need it to. The model hears the words but does not necessarily know what to do with them the way an experienced clinician would.

* * *

Context Windows: What the AI Can See

Every AI model has a limit on how much text it can process at once, known as the context window. Think of it as the model's working memory. Current models range from about 8,000 tokens to over 200,000 tokens, with newer models pushing these boundaries further. For context, a typical page of medical text contains roughly 500 tokens, so a model with a 100,000-token context window can process the equivalent of about 200 pages of text at once.

This has direct clinical implications. If you paste an 83-page patient chart into a clinical AI tool, the model can technically process it, but its ability to attend to specific details degrades as the input grows longer. Critical information buried on page 47 of a discharge summary may receive less attention weight than information at the beginning or end of the document. This is not a hypothetical concern. Studies have demonstrated that LLM performance degrades on tasks requiring retrieval of information from the middle of long inputs, a phenomenon sometimes called the "lost in the middle" effect (Liu et al., 2023).

For clinicians, the practical takeaway is this: do not assume that because a model can accept a large input, it will process all of that input with equal fidelity. If you need the AI to reason about a specific lab value or medication change, provide that information prominently rather than buried in a long document paste.

Clinical Parallel

In my practice as a ship physician, this constraint takes on a specific dimension. I do not have access to a passenger's complete medical history. They arrive with medication lists that are sometimes incomplete, sometimes handwritten, sometimes in a language I do not speak. If I paste what I have into an AI tool, the model will generate recommendations based on that partial picture and present them with full confidence. The AI does not know what it is missing. In maritime medicine, neither do I. The difference is that I have been trained to recognize the limits of my information. The model has not.

* * *

How the Response Is Built

Once the model has processed your input, it generates a response one token at a time. This is worth emphasizing because it contradicts the intuitive sense that the AI "thinks about" your question and then delivers an answer. There is no planning phase. There is no moment where the model constructs a complete treatment plan and then writes it out. Each word is generated based on the input plus all the words generated so far, and each word choice influences what comes next.

This sequential generation process is where many clinical AI errors originate. If the model generates an incorrect drug name early in a treatment plan, subsequent text will build on that error rather than correcting it. The model does not step back and review its output for internal consistency the way a physician would review a note before signing it. It moves forward, always forward, one token at a time.

There is also a parameter called **temperature** that controls how much randomness the model introduces during generation. At low temperature, the model consistently selects the most probable next token, producing predictable and repetitive output. At higher temperature, it samples from a wider range of possibilities, introducing variability. This is why you can ask the same clinical question twice and receive different answers. It is not that the model changed its mind. It is that the random sampling led it down a different sequence of word choices, which cascaded into a different response.

You generally cannot control or even know the temperature setting of commercial AI tools, whether general-purpose or clinical. This opacity is itself a limitation. The same tool may produce more variable output during software updates, A/B testing, or backend configuration changes without disclosing this to users. When a clinical recommendation changes between Tuesday and Thursday, the clinician has no way to determine whether the model was updated, the temperature was adjusted, or the random seed simply fell differently.

For clinical applications, this variability is a significant concern. Reproducibility is not guaranteed and can vary significantly based on system design, prompting technique, and timing. At low temperature settings and without system updates, outputs are often consistent. But clinicians cannot verify these conditions, and the unpredictability itself violates clinical norms of

reliability. If you rely on an AI tool for a clinical question, you should not assume the same query will produce the same output tomorrow.

Clinical Parallel

If you ask an LLM about drug dosing for a pediatric patient, and the model gives you a number, that number emerged from a probability calculation, not from a pharmacokinetic database. If you ask the same question again and get a slightly different number, neither answer is more "certain" than the other. The model cannot tell you which one is correct. It cannot tell you why they differ. Probability is not the same thing as accuracy.

* * *

What This Means for Your Practice

Understanding these mechanics does not require you to become a machine learning expert. But it should fundamentally change how you interpret AI-generated clinical content. Here are the practical principles that follow from what we have covered:

The output is not evidence. An AI-generated treatment recommendation is not equivalent to a guideline, a systematic review, or even an expert opinion. It is a statistical reconstruction of text patterns. Treat it as a starting point for your own clinical reasoning, not as a conclusion.

Confidence is cosmetic. The model generates text in the same fluent, authoritative tone whether the content is accurate or fabricated. You cannot assess reliability from how the response reads. You can only assess it by verifying the content against known sources. Asking the model to express its level of confidence can sometimes surface useful hedging, but do not mistake expressed uncertainty for calibrated accuracy.

Specificity matters in your prompts. The more precisely you describe your clinical context, including constraints, available resources, patient-specific factors, the better the model can attend to those details. A vague question produces a generic answer. A specific question at least gives the model the opportunity to generate a relevant one.

Longer inputs are not better inputs. Pasting entire charts is less effective than providing curated, relevant clinical information. The model's attention is finite. Help it focus on what matters.

Reproducibility is not guaranteed. If you rely on an AI tool for a clinical decision, document what you asked, what it generated, and what you decided. The same query may produce different output tomorrow.

There is one additional principle that does not concern the model itself but concerns *you*. Clinical AI failures do not always originate inside the algorithm. They frequently occur at the interface between the clinician and the tool: in how the prompt was constructed, in what the clinician expected to hear, in the confirmation bias that makes a fluent, well-formatted response feel more trustworthy than it is. If you ask a vague question, you will get a generic answer, and you may not notice the gap between what you needed and what you received. If the AI confirms your preliminary assessment, you may stop questioning it sooner than you would question a colleague. The model is a mirror of statistical patterns. The interpretation is yours.

When to Verify AI Output Externally

The recommendation involves a drug, dose, or interaction you do not recognize.

The response does not acknowledge constraints you specified in your prompt.

The output cites specific studies, guidelines, or authors. Verify that these exist.

The same question, rephrased, produces substantively different answers.

The recommendation feels too clean for a complex, ambiguous clinical scenario.

The AI suggests a specialist referral, imaging study, or intervention that is not available in your setting.

* * *

Before You Turn the Page

Everything we have discussed in this chapter describes a system that works as designed. The next-word prediction, the training on internet-scale data, the token-by-token generation, the temperature-driven variability: none of these are bugs. They are features of the architecture. The system does exactly what it was built to do.

The problem is that what it was built to do is not what most clinicians assume it does. Most clinicians assume it reasons. It does not. Most assume its confidence reflects its accuracy. It does not. Most assume that a well-formatted, medically fluent response is more likely to be correct than a poorly formatted one. There is no such correlation.

In the next chapter, we will examine what happens when this system fails, specifically, why AI can sound completely confident while being completely wrong. We will look at hallucination and confabulation, terms that are themselves debated in the field, and the specific failure modes that matter most

in clinical practice. We will also revisit the ship scenario from this chapter's opening, because the answer that AI generated for our 2:47 AM STEMI patient had a problem. A significant one. The AI did not hallucinate a drug. It did not invent a medication that does not exist. It recommended a real, guideline-concordant medication for STEMI management. The problem was simpler and, in some ways, more dangerous: it had no way to know that the drug was not in a ship's formulary of roughly forty medications, because that constraint does not exist in the training data of a general-purpose model.

The AI recommended a drug that was not on board.

Why AI Sounds Confident But Gets It Wrong

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

The Terminology Problem

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

The Anatomy of a Confident Error

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

Why “Just Verify It” Is Harder Than It Sounds

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

The Five Failure Modes That Matter Most

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

1. Formulary-Blind Recommendations

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

2. Citation Fabrication

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

3. Contraindication Blindness

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

4. The “Too Clean” Answer

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

5. The Stale Guideline Problem

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

What the Research Shows

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

The Ship at 2:47 AM, Continued

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

Toward a Different Kind of Verification

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

Before You Turn the Page

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

Where the Training Data Came From (And Why That Matters)

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

The Corpus: What the Model Read

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

What RLHF Actually Does (And What It Cannot)

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

Bias: What the Data Does Not Contain

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

The Forty-Drug Formulary Problem

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

The Knowledge Cutoff and the Guideline Lag

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

Why the Model Treats UpToDate and Reddit the Same

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

Before You Turn the Page

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

The Six-Platform Study — What I Found When I Tested Clinical AI

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

Why I Built This Study

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

The Six Platforms

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

The Rubric: 20 Items, 100 Points, One Standard

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

The Results

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

The Final Scores

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

What the Clinical AI Tools Got Right

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

Where the General LLMs Failed

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

GPT-5 (Score: 34 / 100)

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

Claude Opus 4 (Score: 35 / 100)

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

Gemini 2.5 Pro (Score: 29 / 100)

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

Llama 4 (Score: 21 / 100)

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

What the 2.4x Gap Actually Means

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

The Benchmark Paradox

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

What No Rubric Captures

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

Why the Gap Exists: Architecture, Not Intelligence

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

What This Means for Your Practice

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

Study Limitations

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.

Before You Turn the Page

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/AiforDoctors.