

Chapter 12 / Capítulo 12

Primum non nocere: Software engineering
for medicine
ISBN: 978-9915-9876-8-2
DOI: 10.62486/978-9915-9876-8-2.Ch12
Pages: 198-210



© 2026 The authors. This is an open access
book chapter distributed under the terms of
the Creative Commons Attribution (CC BY)
4.0 License.

Chapter 12. Generative AI and Large Language Models

Annier Jesús Fajardo Quesada  

Eduardo Antonio Hernández González 

René Herrero Pacheco 

Rolando Javier Álvarez Pérez 

1 Granma University of Medical Sciences. Granma, Cuba.

2 University of Medical Sciences of Pinar del Río. Pinar del Río, Cuba.

“Generative artificial intelligence is not just another tool in the digital health arsenal; it is a paradigm shift as profound as the introduction of the electronic health record. For the first time, we have machines that not only classify or predict but also create, write, summarize, and converse. In addition, for the first time, the risk is not just that they will make mistakes but that they will invent things with such conviction that no one notices the error.”

Imagine a doctor who ends their day with twenty medical records to write, discharge reports to complete, and referral letters to send. They spend hours on this bureaucratic task—time they could be spending with their patients. A generative AI system could, in seconds, produce a draft of each document, extracting relevant information from scattered notes, structuring it appropriately, and drafting it in clear, professional language. The doctor would only have to review and sign. The promise is tempting: freeing up clinical time, reducing the administrative burden, and improving documentation. However, it also poses risks: what if the system invents data? What if it omits critical information? What if its language is inappropriate? Who is held accountable when an automatically generated report contains an error?

Generative artificial intelligence, particularly large language models such as the GPT, have burst into public scenes with overwhelming force. Their ability to produce coherent text, summarize documents, answer questions, and hold conversations has generated enormous expectations regarding their potential in medicine. However, it has also raised concerns about their limitations and risks. Unlike traditional predictive models, which produce a number or a classification, generative models produce text, and text is inherently more difficult to evaluate and validate. An error in a numerical prediction is easy to detect; a hallucination in a text can go unnoticed for months.

In this chapter, we explore the applications and risks of generative AI in the clinical setting. We will begin with one of the most promising applications: the automatic summarization of medical records, which can free up clinical time and improve documentation. We will continue with the automation of bureaucratic tasks, such as report writing and diagnostic coding. We address the most feared risk of these models: hallucinations—those moments when AI confidently invents information that does not exist. Finally, we discuss how to evaluate these models in healthcare settings, which is a first-order methodological challenge. This is not a technical guide for implementing these models but rather a reflection on what it means to introduce systems that create—rather than merely predict—at the heart of clinical practice.

12.1 AUTOMATIC SUMMARY OF MEDICAL RECORDS

There is a cruel paradox in modern medical practice. Professionals who entered the field driven by a desire to care for people spend increasing amounts of time caring for documents. A primary care physician in a developed healthcare system may spend two hours on every eight-hour shift typing notes, writing reports, filling out forms, and justifying decisions for future audits. After seeing thirty patients, a hospital specialist faces a mountain of paperwork before being able to leave. This administrative burden is not just a bureaucratic nuisance; it is a recognized factor in professional burnout, a source of errors by omission when fatigue overcomes attention, and an invisible barrier between the clinician and the patient. The promise that artificial intelligence systems can automatically generate summaries of medical records, freeing up hours for direct care, is one of the most enticing of the era of large language models. However, like any technological promise in medicine, its realization depends on navigating technical, clinical, and ethical complexities that initial enthusiasm tends to obscure.

The automatic summarization of medical records is not merely a matter of text compression. A contemporary medical record is an ecosystem of heterogeneous information: progress notes in natural language—sometimes telegraphic, sometimes narrative—laboratory results in structured tables, imaging reports with technical descriptions and conclusions, medication lists with dosages and routes of administration, coded and uncoded allergies, and family history scattered across different sections. Summarizing all of this coherently requires not only processing words but also understanding what is relevant to each clinical context, what can be safely omitted, how to structure the information for maximum utility, and in what tone to communicate it. Large language models trained on massive medical corpora have demonstrated surprising capabilities for these tasks, but surprisingly, they should not be confused with a guarantee.

12.1.1 THE DIALECTS OF SUMMARIZATION: ADAPTING TO CONTEXTS AND AUDIENCES

The use cases for automatic summarization are as diverse as the moments in clinical practice. A system can generate admission summaries for specialists receiving referred patients, highlighting active diagnoses, ongoing treatments, critical allergies, and pending tests that justify the consultation. Another can produce hospital discharge reports based on progress notes from admission, organizing the reason for admission, key findings, treatment received, follow-up recommendations, and warning signs for the patient. Another can create simplified versions for the patients themselves, translating technical jargon into understandable language without losing medical accuracy. Each of these summaries has a different audience, a different purpose, a different level of detail, and a different risk of error. A single, unadapted model serves all these needs poorly.

Customization based on the recipient is where generative artificial intelligence demonstrates its flexibility. Through appropriate prompting, the same base model can be trained to produce heart-centered summaries for the cardiologist, lung-centered summaries for the pulmonologist, or summaries focused on functional impact for the physical therapist. It can adjust the level of technical detail, the depth of past and remote history, and the emphasis on preventive versus curative aspects. This adaptability is advantageous over the rigid template systems that dominated previous clinical documentation generation. However, it is also a source of risk: if the prompt does not adequately specify the recipient's needs, the model may produce a summary that is elegantly written but clinically useless or dangerously incomplete.

Managing contradictory information is a particularly acute challenge. Real medical records contain inconsistencies: one doctor documents an allergy to penicillin, while another later prescribes amoxicillin without comment; one imaging report describes a stable lesion, while a subsequent report suggests progression; and the patient denies symptoms in one consultation that they described in detail in another. A naive summary that simply condenses all of this may convey confusion or, worse, take sides with one version without pointing out the contradiction. More sophisticated models must explicitly identify discrepancies, present them within their temporal context, and, when appropriate, indicate that they require clinical clarification before important decisions are made. This ability to handle uncertainty in the source text is a hallmark of mature systems.

12.1.2 QUALITY BEYOND FLUENCY: EVALUATING WHAT TRULY MATTERS

The evaluation of automatic medical summaries cannot be limited to traditional natural language metrics. BLEU or ROUGE scores, which compare surface-level similarity with reference summaries, are insufficient: a summary may use completely different phrvalues while better capturing essential information, or it may copy phrases from the original while critical points are omitted. Human evaluation by clinicians is the gold standard, but it is costly, subjective, and difficult to scale. Hybrid metrics that combine coverage of key information (extracted via structured language processing), verifiable factual accuracy against the original record, relevance to the requested context, and readability for the target audience provide a more comprehensive evaluation.

The omission of critical information is the most dangerous error and the most difficult to detect automatically. A summary that fails to mention an iodine allergy prior to a contrast study, that overlooks a history of long QT syndrome before prescribing certain antiarrhythmics, or that ignores polypharmacy in an elderly patient can cause direct harm. Summarization systems must incorporate context-specific checklists of critical elements, ensuring that certain information is always included when it is present in the source. This “hardcoding” of medical priorities is a necessary safeguard against model variability.

Hallucination—the generation of information not present in the source—is a characteristic risk of generative models. A summary that invents a diagnosis, attributes symptoms to the patient whom they never reported, or suggests treatments that were not administered is worse than useless; it is actively dangerous. Grounding techniques, which link each statement in the summary to its specific location in the source documentation, enable efficient human verification and automatic detection of unsupported claims. Models must be designed to prefer omission over invention when the source is ambiguous.

12.1.3 INTEGRATION INTO THE WORKFLOW: BETWEEN AUTOMATION AND SUPERVISION

The timing and method of summary generation determine its practical utility. Is it generated automatically upon opening a new patient's chart, presenting itself to the physician initiating the consultation? Is it produced on demand when the clinician requests a "summary for referral to cardiology"? Is it created automatically upon discharge, for review before signing? Each option has implications for workflow, the physician's perception of control, and the risk of excessive dependence. The most widely adopted systems tend to offer on-demand generation with configurable options, respecting professional autonomy.

Human review of generated summaries is an essential safeguard, but its nature evolves as the system matures. Initially, the review must be thorough, line by line, and compare the text to the sources. As evidence of reliability accumulates for certain types of summaries in certain contexts, the review can become more targeted: verification of critical elements, sampling of sections, and conditional trust in the history of previous corrections. Feedback from these reviews—what errors are detected, how often, and in what contexts—must inform the retraining or adjustment of the model. A system that does not learn from its detected errors is static in a dynamic world.

The relationship between efficiency and safety is a constant tension. A system that saves thirty minutes a day but introduces errors that require twenty minutes of correction and potentially cause harm is a bad deal. A system that saves time but reduces the physician's familiarity with the patient's complete history, creating a reliance on the summary—can degrade the quality of care in subtle ways. Impact assessment must include not only documented time but also perceived quality of care, patient satisfaction, and long-term clinical outcomes.

12.1.4 PRIVACY, CONSENT, AND THE GHOST OF MEMORIZATION

Large language models have a concerning ability to memorize information from their training and reproduce it in response to appropriate prompts. In medicine, where training data contain real—albeit anonymized—medical records, there is a risk that the model will generate summaries that include specific information about patients from the training dataset, violating privacy if these prompts are cleverly designed. Differential privacy training techniques, which add mathematically calibrated noise to ensure formal privacy, and techniques for verifying that the model does not memorize sensitive data, are areas of active research but are not yet mature enough for widespread clinical deployment.

Consent for the use of data in training generative models is evolving regulatory territory. If a patient consented to their medical record being used to "improve healthcare," does that include training AI systems that will generate summaries for other patients? What if the model—or partially trained on their medical history—produces a summary for another patient who contains phrases identical to those in their own documentation? Consent policies must be updated to specify these uses, and systems must be designed to minimize the risks of reidentification and retention.

For the developer, building medical summary systems requires hybrid skills. Not just language model engineering but also a deep understanding of clinical workflows, interface design that facilitates efficient review, feedback infrastructure, and continuous improvement. It requires close collaboration with clinical informaticians, quality professionals, and privacy specialists. It also requires humility in terms of limitations: generative models are powerful but imperfect tools that improve with human supervision rather than without it.

For the clinician-user, automatic summarization demands the development of new skills. The ability to critically read a generated summary, verifying it against their knowledge of the patient and the source documentation when in doubt. The discipline to systematically report errors, not just correct them individually. In addition, the judgment is to know when to trust the summary and when to dive into the full history. Technology does not eliminate professional responsibility; it transforms it.

Ultimately, the automated summary must serve a more humane medicine, not one that is more efficient but cold. If it frees up time for doctors to listen, explain, and accompany patients, it will have fulfilled its promise. If it simply accelerates an already frenetic pace, or worse, if it creates a distance between the clinician and the patient's full history, it will have betrayed the medical purpose. The direction depends on how we design, implement, and regulate these technologies. The choice is ours, and so is the responsibility.

12.2 AUTOMATION OF BUREAUCRATIC TASKS

There is a monster in modern medicine that does not appear in medical records or management reports but that devours hours, energy, and vocations with silent efficiency. It is called healthcare bureaucracy, and it is estimated to consume between one-third and one-half of doctors' working hours. It is not a single task but an ecosystem of small demands that accumulate: coding each diagnosis into standardized taxonomies, justifying each costly treatment to payers with shifting criteria, responding to requests from insurers who question decisions already made, filling out forms for quality indicators that no one seems to use, and drafting appointment letters and reminders that should be automated. Each of these tasks, examined in isolation, has its own administrative logic. Taken together, they constitute a burden that distracts professionals from what truly matters: caring for people who are suffering. The promise that generative artificial intelligence can take on part of this burden is, therefore, extraordinarily appealing. However, like any promise of power, it raises questions about accuracy, accountability, and the risk that technically gained efficiency will be lost in new forms of clinical detachment.

12.2.1 CODING AND ITS LABYRINTHS: WHEN DIAGNOSIS MEETS THE CODE

Diagnostic coding is perhaps the most burdensome and least rewarding bureaucratic task in medical practice. Every clinical encounter must be translated from the rich, narrative, sometimes ambiguous language of medical judgment into standardized codes: the ICD-10 for general diagnoses, the CIAP for primary care, the CPT for procedures, the SNOMED-CT for more sophisticated systems, and national variants that complicate the landscape. This translation is necessary for billing, public health records, epidemiological research, and quality management. However, it is tedious, prone to errors in selecting among similar codes, and consumes time because the physician does not spend listening to the patient.

A generative artificial intelligence system can, on the basis of clinical notes in natural language, suggest the most appropriate codes and even justify the selection. The model analyzes the case description, identifies medical entities, maps them to the target taxonomy, and presents suggestions ranked by confidence. The physician reviews, confirms, or corrects, and the system—learns from these corrections for future suggestions. In theory, this reduces coding time from minutes to seconds, improves consistency by eliminating variability among human coders, and increases specificity by suggesting codes that are more precise than those the physician would recall.

The reality is more complex. Medical coding systems are labyrinths where seemingly similar codes have very different implications. The difference between “type 2 diabetes mellitus with complications” and “type 2 diabetes mellitus without complications” is not academic; it affects billing, resource prioritization, and facility quality metrics. A model that suggests the most common code because it appears most frequently in its training, without capturing the subtleties of the particular case, introduces systematic biases. Human oversight is not optional; it is a guarantee that clinical judgment prevails over the model's statistics.

12.2.2 JUSTIFICATIONS, REQUIREMENTS, AND THE DANCE WITH PAYERS

Justifying treatments represents another frontier of bureaucratic automation. When a physician prescribes an expensive drug, a complex procedure, or a prolonged hospital stay, they must often document why it is necessary, especially when dealing with payers with budget constraints or specific eligibility criteria. This justification must be clinical, evidence-based, and tailored to the format and criteria of each payer: private insurers with their own forms, public systems with prior authorization protocols, and rare disease programs with specific requirements.

An artificial intelligence system can, by knowing the patient's history, the proposed treatment, and the relevant clinical guidelines, draft a justification that the physician reviews and submits. The model identifies elements in the clinical documentation that meet the payer's criteria, structures the argument clinically, and adapts the language to the required format. This promises to save hours of work and, potentially, improve the approval rate by ensuring that no relevant elements are omitted.

The challenge lies in the variability of requirements. Each payer has its own form, procedure codes, coverage criteria, and preferred documentation styles. A system that works perfectly for insurer A may fail completely for insurer B. Adaptability through prompting—natural language

instructions that specify the payer's context—is promising but not infallible. Models may “hallucinate” criteria that do not exist or omit specific requirements they do not recognize in their general training. Human verification that the draft exactly meets the specific payer's requirements remains mandatory.

The response to administrative requests follows a similar pattern. Insurers, accrediting bodies, and disease registries periodically request information that already exists in medical records but is scattered across different formats, requiring manual synthesis. An artificial intelligence system can automatically extract the relevant information, structure it according to the requested format, and generate draft responses. The potential efficiency is enormous, especially for routine and repetitive requests.

Exception handling is where automation reaches its limits. Not all cases fit standard patterns: patients with complex situations requiring unconventional reasoning, low-level treatments with emerging but not yet established evidence, and rare diagnoses where guidelines do not apply directly. In these cases, a model that always produces a response—even if it is inadequate—is more dangerous than one that recognizes its limitations and refers to the case of a human. Systems must be designed to detect when confidence is insufficient, when the case deviates from trained patterns, and when human intervention is imperative.

12.2.3 INDICATORS, ALIGNMENT, AND MEASURING WHAT MATTERS

Generating reports on quality indicators is a task that consumes substantial resources in hospitals and health centers. Dozens of indicators—wait times, readmission rates, vaccination coverage, adherence to guidelines—require data extraction from multiple sources, aggregation, and presentation in standardized formats for regulatory and accreditation bodies. An artificial intelligence system can automate much of this process, generating preliminary reports that quality managers review and finalize.

The risk here is the “dictatorship of the measurable.” Quality indicators are imperfect proxies for the true quality of care. A system that optimizes report generation may inadvertently prioritize the appearance of quality over actual quality. If the model learns that certain codes or formulations improve indicators without reflecting better care, it may systematically suggest them, distorting both documentation and practice. Human oversight must ensure that indicators reflect reality, not that reality is shaped to meet indicators.

Correspondence with patients—appointment letters, test reminders, test results, and postdischarge instructions—is an area where automation promises a dual benefit: efficiency and improved communication. A system that generates personalized drafts, tailored to the patient's level of understanding, their specific clinical situation, and the tone desired by the institution, can save time on drafting while improving clarity. Reminders can be more effective if personalized; postdischarge instructions are safer if they are adapted to the patient's specific polypharmacy.

Mass personalization, however, has ethical limits. Medical correspondence is not merely the transmission of information; it is also the building of a therapeutic relationship. A patient who receives letters clearly generated by a machine may feel depersonalized, perhaps reduced to a database entry. The balance between efficiency and humanization is delicate: automation should free up time for personal interaction and not replace all interactions with efficiency. The design of these systems must include “human touches,” options for physician customization, variations in tone, and spaces to add handwritten notes, all of which maintain the human connection.

12.2.4 INTEGRATION, ACCOUNTABILITY, AND CONTROL OF AUTOMATION

Integrating these capabilities into existing electronic health records systems is a technical and organizational challenge. Powerful models alone are not enough; connectivity is required with patient databases, clinical guideline repositories, up-to-date code catalogs, and constantly evolving payer systems. Interfaces are needed that enable efficient review: not only display the model's suggestion but also link it to the source in the clinical documentation, allowing for intuitive editing and capturing feedback on corrections. Workflows are required that integrate automation without disrupting practice, given that different physicians work at different paces.

The relationship between automation and legal liability is uncharted. If an incorrect diagnostic code generated by the system leads to fraudulent billing, who is liable? If a poorly drafted treatment justification causes a denial of coverage and harm to the patient, is it the

physician who signed off without thoroughly reviewing it? The hospital that implemented the system? The developer who trained the model? Case law is only just beginning to take shape, and in the meantime, prudence calls for defensive design: mandatory human oversight, full traceability of decisions, and clear documentation that medical liability remains with the professional.

Concern about loss of control is legitimate and profound. If the system automatically generates codes, justifications, and responses, how can we ensure that they reflect actual clinical judgment and not the statistical optimization of the model? Won't physicians, pressed for time, end up accepting suggestions without the critical review that safety demands? Aren't we silently delegating to the machine decisions that should be human? These concerns are not Luddism; they are an expression of the tension between efficiency and professional autonomy that characterizes all meaningful automation.

Careful design can mitigate these tensions. Systems must be “reviewable by design”: they should display not only the final suggestion but also the reasoning behind it, allow the physician to explore alternatives, and make it explicit when the model's confidence is low. They must be “ignorable without penalty”: a physician who prefers to code manually or draft their own justifications should be able to do so without friction. They must be “learners from their own corrections”: When the physician corrects a suggestion, the system must incorporate that feedback to improve future suggestions, creating a virtuous cycle of adaptation.

In terms of job satisfaction, the impact is a critical factor that is often overlooked. The automation of bureaucratic tasks can free up time, but it can also strip the work of aspects that some professionals value. Some physicians prefer to write their own reports because they retain narrative control or find intellectual satisfaction in precise coding. Imposing uniform automation can lead to resistance and alienation. Implementation must be flexible, allowing professionals to choose when and how to use these tools while respecting the diversity of work styles.

12.3 MEDICAL HALLUCINATIONS: THE RISK OF INVENTION

There is a phenomenon in large language models that has evolved from academic curiosity into a critical obstacle to their adoption in medicine. When these models encounter questions with no clear answer in their training data or when they encounter ambiguous information, they do not always respond with the honesty of an experienced physician: “I don't know for sure; I need to verify.” Instead, they frequently make things up. They do so with such fluency, with such an air of scientific authority, that even expert readers may be slow to detect deception. This phenomenon has been termed “hallucination,” a borrowed psychiatric term that captures its quality of perception without reality. In any other context, it would be an amusing anecdote. In medicine, where decisions are made in seconds on the basis of the information presented, it is a potentially lethal risk that demands a deep understanding and rigorous safeguards.

Hallucinations are not ordinary errors of knowledge or interpretation. An ordinary error is understandable and detectable: the model confuses two drugs with similar names, misinterprets a dose, or applies a guideline to a context where it does not apply. A hallucination is qualitatively different: the model generates information that exists nowhere in its training or in reality, citing studies that were never conducted and describing symptoms that do not appear in the literature, attributing to recognized authors statements they never made. It does so with the same expressive certainty with which it describes an established fact. For a physician consulting the model in the midst of an urgent clinical decision, distinguishing between what is real and what is invented may be impossible without external verification that time does not permit.

12.3.1 ARCHITECTURE OF INVENTION: WHY MODELS HALLUCINATE

The cause of hallucinations lies in the fundamental nature of large language models. Unlike traditional knowledge-based systems, which consult structured repositories of verified facts, these models operate through the statistical prediction of sequences. They have no internal representation of “truth” against which to test statements. What they possess is the ability to predict, given a sequence of preceding words, what the next most likely word is within the context of the trained language. When they generate text, they produce statistically plausible sequences, not necessarily true ones.

If certain combinations of concepts appeared frequently together in the training data, the model learned that they were “cooccurring” and reproduced them. If the requested combination

has no established true answer in the literature but the model has seen similar patterns, it can generate something that sounds convincing but is completely fictitious. Linguistic plausibility—correct grammar, appropriate terminology, and logical structure—is independent of factual accuracy. A model trained on medical texts learns to speak like a doctor, not necessarily to think like a rigorous doctor.

In medicine, hallucinations take on particularly dangerous forms because of the gravity of the consequences of each error. They can invent drug interactions that do not exist, leading doctors to avoid perfectly safe and effective therapeutic combinations. They can describe drug side effects that have never been documented, causing unnecessary anxiety in patients or leading to the unjustified discontinuation of necessary treatments. They can cite clinical studies with completely fictitious authors, dates, journals, and numerical results, giving the appearance of scientific evidence to recommendations that have no basis whatsoever.

12.3.2 DOCUMENTED CASES: WHEN FICTION RESEMBLES SCIENCE

A particularly instructive case involved a language model that was asked to generate a summary of evidence regarding an experimental treatment for a rare oncological condition. The model produced text that was flawless in form: a contextualized introduction, methods described with apparent precision, numerical results with confidence intervals, and a balanced discussion. The citations included titles from prestigious journals, names of authors recognized in the field, and recent publication years. Only when a researcher attempted to locate the cited articles did they discover that none existed. The model had invented everything: the titles, the authors, the journals, the methods, and the results. The presentation was so convincing that without access to bibliographic databases for verification, any user would have accepted the presented evidence as valid.⁽¹⁾

Hallucinations are particularly difficult to detect because the models have learned to perfectly mimic scientific language. They use jargon appropriate to the subspecialty, structure their responses as an expert in systematic reviews, maintain internal consistency that inspires confidence, and cite sources in the correct format. A reader who is not a specialist in a specific subject may have no way of knowing that they are reading fiction. Even specialists may have doubts, especially if the hallucination pertains to a very specific area where their own knowledge has gaps. The model's authoritative tone, confident manner, apparent detail, and absence of signs of doubt are counterindicators of reliability, but psychologically, they function as indicators of credibility.

12.3.3 MITIGATION STRATEGIES: BETWEEN PREVENTION AND DETECTION

For developers of systems based on generative artificial intelligence, hallucinations represent the most pressing technical problem. There is no complete solution, but various approaches can reduce their frequency and severity. Retrieval-augmented generation (RAG) is a promising strategy: instead of generating responses purely from the model's internal knowledge, the system first consists of reliable and verifiable databases, indexed medical literature, official clinical guidelines, and drug registries and then generates a response based strictly on the retrieved information. This drastically reduces hallucinations but requires that the database be comprehensive and up-to-date and that the model be capable of correctly extracting relevant information without distortion.

Reinforcement learning with human feedback (RLHF) can reduce the frequency of hallucinations by rewarding verifiable answers and penalizing in s. However, it does not eliminate the following fundamental problem: the model still lacks an internal truth mechanism; it merely learns to avoid answers that humans identified as having been diagnosed with hallucinations in the past. New situations can generate new forms of invention not seen in the feedback training.

Designing systems that, when in doubt, prefer to respond “I don't know” rather than risk hallucinating is a desirable but technically challenging goal. It requires the model to have a precise calibration of its own uncertainty—a capability that current models possess only in a limited and unreliable manner. A model that underestimates its ignorance will continue to hallucinate; one that overestimates its uncertainty will be useless because of excessive caution.

Human oversight remains the most effective barrier against hallucinations in clinical contexts. Any text generated by artificial intelligence that is intended to influence medical

decisions must be reviewed by a professional before use. This review cannot be superficial; it must include verification of key statements, especially those affecting diagnostic or therapeutic decisions. However, this oversight limits the utility of automation: if every response requires exhaustive verification, the time savings are drastically reduced, and the risk that oversight will become routine and less rigorous over time is real.

12.3.4 EDUCATION, SKEPTICISM, AND THE FUTURE OF TRUTHFULNESS

For clinicians, the existence of hallucinations demands a radically skeptical attitude toward any text generated by artificial intelligence. They cannot trust that something is true simply because it is wellwritten and expressed with authority. They must verify important claims, especially if they are going to base clinical decisions on them. They must be alert to potential signs of hallucination: vague references (“some studies suggest...”), the absence of verifiable citations where they would be expected, details that do not align with their knowledge of the subject, or suspiciously perfect data—which in real research always contain some noise.

Education on hallucinations must be a mandatory component of any artificial intelligence training program for healthcare professionals. The goal is not to instill paralyzing fear but to provide training for safe use. A physician who understands how the models work and who knows that they can hallucinate and through what mechanisms is better positioned to use these tools in a way that leverages their benefits while mitigating their risks. They are also better positioned to demand that developers implement safeguards and to actively participate in system evaluation prior to institutional adoption.

The future of generative artificial intelligence in medicine necessarily depends on resolving the problem of hallucinations or at least rigorously containing their risks. Architectures that explicitly incorporate the notion of truth, with mechanisms to verify claims against reliable sources in real time, are an active area of research. Automatic verification methods that detect internal inconsistencies or statistical improbabilities in generated text can provide early warning signals. The development of specialized “verification models,” trained specifically to detect hallucinations in the outputs of other models, is a defense-in-depth strategy.

In the meantime, prudence must guide all clinical applications. Hallucinations are neither a minor flaw nor a temporary bug; they are a manifestation of the fundamental nature of large language models. They were designed to generate plausible text, not to be accurate encyclopedias. Plausibility is not true, and in medicine, the difference between the two can have irreversible consequences. Generative artificial intelligence must be used with extreme caution, with mandatory human supervision, and with full awareness that every model output is potentially a mixture of trained wisdom and statistical invention. Only in this honesty about limitations lies the possibility of beneficial and nonharmful use.

12.4 EVALUATING LLMS IN HEALTHCARE SETTINGS

Evaluating a language model in medicine is not like evaluating a traditional predictive algorithm. When a model classifies images or predicts risk, we have a clear gold standard: a biopsy that confirms the diagnosis and an outcome that verifies the prediction. The metrics are numerical, comparable, and reproducible. However, when a model generates a medical record summary, answers a distressed patient’s question, or drafts a treatment justification, against what we compare its output? Who decides if the text is correct, complete, appropriate, and safe? The evaluation of generative models is a methodological challenge of the first order, and its imperfect resolution is a necessary—although not sufficient—condition for these systems to be safely integrated into clinical practice.

12.4.1 THE ABSENCE OF A SINGLE TRUTH: WHEN “CORRECT” IS PLURAL

The first profound difficulty is the absence of an unequivocal “fundamental truth” for many generative tasks. A summary of a patient’s medical history can be valid in multiple ways: focused on active problems versus comprehensive in background, detailed on medication versus centered on diagnoses, and structured chronologically versus by systems. Two expert physicians may produce summaries that are equally correct but substantially different. How, then, do we evaluate whether the model’s output is acceptable? The answer inevitably involves human judgment, but

human judgment is costly, variable across evaluators, difficult to scale, and potentially biased by expectations regarding the technology.

This plurality of valid answers does not mean that everything is valid. There are hard limits: omission of allergies is unacceptable, invention of diagnoses is unacceptable, and recommendations contrary to established guidelines are unacceptable. However, within safety limits, there is ample room for legitimate variation. Evaluation must capture this complexity: not only detecting serious errors but also rating the quality of acceptable responses, differentiating between excellent, adequate, and improvable.

12.4.2 THE DIMENSIONS OF QUALITY: BEYOND BINARY ACCURACY

The quality of generated medical text is inherently multidimensional. Factual accuracy—are the contained data correct?—is only the first dimension. Completeness—does it include all information relevant to the context?—is distinct: a summary may be factually correct in what it includes but dangerously incomplete in what it omits. Coherence—is it easy to follow logically?—affects practical utility. Conciseness—does it avoid irrelevant, distracting information?—determines whether the physician will actually read the summary. Tailoring to the audience—does it use language appropriate for the intended reader?—varies between a specialist, a general practitioner, an educated patient, or a patient with low health literacy. In addition, safety—does it contain statements that could lead to erroneous decisions if taken at face value?—is a critical dimension that goes beyond factual accuracy.

Each dimension requires different evaluation methods. Factual accuracy can be verified against structured medical records. Completeness requires context-specific checklists of critical elements. Coherence can be assessed using readability metrics and human judgment. Suitability for the target audience requires evaluators who represent that audience. In addition, safety requires adversarial testing designed to provoke problematic responses. A model may be excellent in terms of factual accuracy but deficient in safety or brilliant in coherence but dangerous in completeness. One-dimensional evaluation is misleading.

12.4.3 DETECTING THE NONEXISTENT: THE EVALUATION OF HALLUCINATIONS

Hallucinations—convincing fabrications of false information—pose a particularly acute evaluation challenge. They are subtly plausible: a fabricated study will have a credible title, authors with real names, and reasonable numerical results. Detecting that it is false requires verification against reliable sources, a task that no human can perform exhaustively at scale. Automated methods for detecting hallucinations—verifying citations against bibliographic databases, detecting internal inconsistencies, and analyzing sequence probabilities—have their own limitations and error rates.

Assessing hallucinations in medical contexts requires the deliberate design of “traps”: questions designed to elicit fabricated responses because the true answer does not exist or is extremely rare. Ambiguous clinical cases for which there is no established consensus exist. Requests for information about drugs that do not exist. Questions about clinical guidelines for recently discovered diseases where the literature is scarce. The frequency with which the model falls into these traps and the nature of its responses are critical safety metrics.

12.4.4 PRACTICAL APPROACHES: BETWEEN THE LAB AND THE CLINIC

Faced with these complexities, multiple evaluation approaches have emerged, none perfect, all partial. Task-based evaluation defines sets of test cases with ground-truth answers crafted by experts and compares the model's output to these ground-truths using similarity metrics or human judgment. It is the closest to traditional evaluation but requires immense effort to create cases and does not capture the legitimate variability of acceptable answers. It is useful for evaluating specific skills but is limited for general evaluation.

Clinical peer evaluation uses groups of expert clinicians who evaluate model outputs across multiple dimensions using standardized scales. It better captures the complexity of the task but is costly and slow and suffers from interrater variability that necessitates multiple evaluations per case and statistical aggregation methods. Interrater reliability is a critical metric: if experts cannot agree on what constitutes a good response, the model cannot be consistently optimized.

The use of language models as evaluators, known as LLM-as-a-judge, is an emerging and promising approach. A second model evaluates the quality of the first by following detailed instructions regarding dimensions and criteria. It is scalable and consistent, but it raises obvious questions: if models can hallucinate, can't they also hallucinate in their evaluations? Research suggests that with careful instructions and on well-defined tasks, this approach correlates reasonably well with human judgment, but it should not be the sole or final evaluation.

Safety evaluation through adversarial testing is a critical component. It includes cases designed to provoke failures: scenarios where incorrect information would have serious consequences, questions on topics where the model should admit ignorance, and requests that combine concepts in ways not seen during training. The frequency and nature of failures in these tests determine confidence limits for deployment.

Fairness evaluation examines whether the model generates responses of differing quality or content for patients from different demographic groups. It requires systematic evaluation across population subgroups and comparisons of multidimensional quality metrics across groups. It is complex but essential work: a model that performs well on average but systematically fails specific minorities perpetuates health inequalities.

12.4.5 REAL-WORLD VALIDATION: WHERE TESTING ENDS

Evaluation in real-world contexts is the final and indispensable step. A model may excel in laboratory tests and fail in real-world practice because of usability issues, integration challenges, changes in user behavior, or dataset shifts. Prospective studies, in which the model is used under real-world conditions with real patients and its impact on clinical processes and outcomes is evaluated, are the sole source of evidence regarding real-world utility. They are costly and complex and require specific ethical approval, but they are mandatory prior to widespread institutional adoption.

For developers, the evaluation of LLMs in medicine is a rapidly evolving field with no established standards. Prudence calls for multiple complementary methods, exhaustive documentation of limitations, transparency regarding uncertainties, and close collaboration with clinicians in the evaluation design to ensure the clinical relevance of metrics. There are no shortcuts: rigorous evaluation is a necessary investment for justified trust.

For clinicians, understanding how these models are evaluated empowers them to interpret performance claims. When a provider announces "95% accuracy," the following question should be answered: When is accuracy in what dimension? evaluated by whom? in which populations? with what types of cases? including adversarial testing? Without clear answers, the claim is marketing, not evidence. Clinicians must demand transparency and participate in validation before adoption.

The future of evaluation lies in the development of shared standards, similar to the CONSORT guidelines for clinical trials. Initiatives such as TRIPOD-LLM extend the TRIPOD guidelines to generative models. Public repositories of test cases will enable standardized comparisons between models. In addition, collaboration among novel Rs, clinicians, regulators, and patients will define what it truly means for a language model to be safe and effective. In this cocreation of standards lies the potential for generative artificial intelligence that is trustworthy in healthcare.

If we have learned to evaluate the quality and safety of generative models, how do we ensure that when these systems get it right, they also explain why, allowing clinicians not only to trust but also to understand, question, and, when necessary, correct?

REFERENCES

Emsley R. ChatGPT: these are not hallucinations – they're fabrications and falsifications. *Schizophrenia* [Internet]. December 1, 2023 [cited May 14, 2026];9(1):52. Available at: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10439949/>