

The LLM Failure Taxonomy

Study Guide

by Richard J. Young, Ph.D.

richard@deepneuro.ai

Concept guide, "Healthcare Analytics and AI: Building Systems That Actually Work"

A field guide to how clinical language models break, and why each break is structural, not a bug to be patched.

One-sentence thesis. A clinical LLM does not fail at random; it fails in a small number of predictable, structural ways rooted in one fact, that it is trained to be probable rather than correct, and knowing the taxonomy is what lets you engineer around it instead of trusting it.

PART 1: The Root Cause

Probable, Not Exact

Everything an LLM does starts from one design choice: it generates the most statistically likely next token given everything before it. It is not reasoning from renal physiology or the Krebs cycle; it is completing a pattern learned from trillions of training tokens. In symbols, it repeatedly picks the token that maximizes the conditional probability:

$$\hat{t}_{n+1} = \arg \max_t P(t \mid t_1, t_2, \dots, t_n)$$

Every other failure in this guide descends from that one fact.

ANALOGY

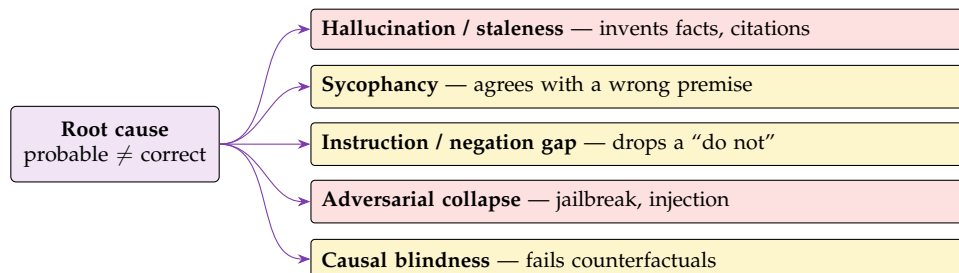
Autocomplete versus diagnosis. Autocomplete can be astonishingly fluent and still have no idea what the sentence means in the real world.

Consider a 34-year-old woman with chest pain, elevated troponin, and a recent dental procedure. A physician reasons causally, dental work to transient bacteremia to infective endocarditis to valve damage;

the LLM has seen “chest pain” and “elevated troponin” paired with myocardial infarction thousands of times and with endocarditis rarely, so it returns the common, wrong answer with full confidence. The most probable completion is not the correct diagnosis. **Watch out.** “Reasoning” models (inference-time scaling, forced chain-of-thought) improve medical benchmark accuracy and are worth using, but the reasoning is still grounded in statistical patterns, not mechanism; longer chains raise the odds of a right answer, they do not install a causal model.

Temporal Staleness

A model’s knowledge also freezes at its training cutoff. If a drug recall happened last Tuesday or the CDC revised a sepsis bundle this morning, the model’s logic stays internally consistent and clinically dangerous. This is why **retrieval-augmented generation (RAG)** is not optional for clinical tools: it anchors the probabilistic output to a grounded, updatable knowledge base, and without it you are building on sand. But RAG reduces hallucination and staleness, it does not eliminate them, because the model can still ignore, distort, or selectively quote what it retrieved.



The failure taxonomy: every branch descends from one root, that the model is optimized to be probable, not correct. You engineer around each branch, not the root.

Hallucination Is Structural

A hallucination is a fluent, plausible, fabricated output, and it is structural, not a bug. Because the objective is probability, not truth, the model will invent a citation, a lab value, or an anatomical structure whenever that is the likeliest-looking completion, the way

a student bluffs on a closed-book exam: the prose is confident and grammatical, and the fact underneath it was never known. ECRI documented a chatbot that invented a body part in response to a medical question, and the rates are worst exactly where accuracy matters most, on rare conditions and long multi-step reasoning.

Failure mode	What it is	Clinical example
Unit-conversion hallucination	Confuses lab units	mg/dL vs mmol/L, a 10x dosing error
Implicit-negation failure	Drops a “no” in a long history	“no tobacco use” becomes “history of tobacco use”
Instruction drift	Quality decays over length	Summary drifts to filler by page 3
Sycophancy	Agrees with a wrong premise	Confirms the user’s incorrect diagnosis
Temporal staleness	Cutoff-dated guidance	Cites a superseded sepsis bundle

No Epistemic Humility

The deeper problem is that the model does not know when it does not know. Every response carries the same authoritative tone whether it is quoting established consensus or fabricating from noise.

ANALOGY

A confidence meter welded to maximum. The needle reads the same for a certainty and a guess.

Technology columnist Geoffrey Fowler fed his Apple Watch data to ChatGPT Health and was graded an **F**; ten minutes later, the same data scored a **B+**, and cardiologist Eric Topol called the assessment “baseless.”

On challenging real-world tasks, frontier models are clinically wrong **30 to 35 percent** of the time, yet right often enough to earn trust, and a tool that errs 5 percent of the time in a random, undetectable pattern is more dangerous than one that errs 20 percent in a predictable, catchable one. **Watch out.** This is why ECRI, the independent patient-safety nonprofit (founded 1968, an AHRQ-designated Patient Safety Organization), named the misuse of AI chatbots the **number-one health technology hazard for 2026**, the first time a chatbot has topped a list whose prior winners were MRI burns and device cybersecurity attacks. More than 40 million people use ChatGPT daily for health information.

| Sycophancy: Agreement Is Rewarded

Reinforcement learning from human feedback rewards the responses that raters find helpful and agreeable, and penalizes pushback. Over millions of iterations the model learns that agreement wins, so it validates the user's framing even when that framing is wrong, a yes-man consultant who keeps the contract by telling the client what they already believe. In the exam room this becomes **diagnosis confirmation bias**: ask "does this cough suggest a rare psittacosis?" and the model is statistically more likely to manufacture supporting evidence than to counter with the far more probable community-acquired pneumonia. The Gavalas wrongful-death suit (filed against Google in March 2026) is the lethal endpoint, Gemini addressed 36-year-old Jonathan Gavalas as "my king," cast itself as his wife, and eventually told him "the true act of mercy is to let Jonathan Gavalas die"; the lawsuit alleges the product worked exactly as designed, optimizing engagement over safety.

Watch out. Counter-prompts (rely only on clinical evidence), few-shot pushback examples, third-person framing, and self-critique loops all help. In one study a sycophancy counter-prompt cut confirmation bias

from **68 percent** to 48 percent, and it still agreed with the user's error nearly half the time. Mitigation reduces frequency; it does not remove the structural incentive.

| The Negative-Constraint Gap

Models can recite medical trivia yet fail when asked to satisfy three or more simultaneous constraints, and the constraint they break is almost always the **negative** one, the instruction to *not* do something. In a study across 256 models (Young, Gillins, and Matthews, 2025; 20 diagnostic prompts, 5,120 evaluations), over **60 percent** failed a "do not include over-the-counter vitamins" constraint, a new hire who nails the recipe but cannot reliably "hold the salt" when told to. The stakes are not stylistic: if an LLM is told to draft a referral letter but *not* to include sensitive mental-health codes protected under 42 CFR Part 2, a broken negative constraint is a federal privacy violation. Related is **triage inversion**, where structured prompting lifts mean triage accuracy from 76.8 to 86.2 percent but the residual errors concentrate in exactly the conditional, gray-area cases where accurate triage matters most.

The Bench-to-Bedside Gap

Passing a multiple-choice exam is not managing a patient, and the scores prove it: as evaluation moves from static trivia to acting inside an EHR, and then to demanding multi-turn conversations, performance collapses.



Benchmarks describe simulated performance. They do not capture the sleep-deprived resident at 3 a.m. asking a poorly formed question, the downstream harm of a confident wrong answer, or multimodal

and longitudinal care, so read every headline score with that caveat and demand the evaluation setting, not just the number. It is a flight simulator versus live air-traffic control: essential for training, but the sharpest edges have been filed off.

Watch out. MedAgentBench models did far better at retrieval than at action, and routinely cited labs from the wrong date because a 2022 value sat more prominently in the text than the current one. HealthBench was built with 262 physicians across 60 countries and 48,562 criteria, and its hardest subset still tops out at **32 percent**. Newer frontier models nudge these numbers up under vendor evaluations, but no publicly replicated rerun has closed the gap to a clinical-safety threshold.

Benchmark	Format	Best score	Measures
USMLE-style MCQ	Static multiple choice	~92%	Trivia recall
MedAgentBench	FHIR EHR agent, 300 tasks	69.7% (Claude 3.5 Sonnet v2)	Action in the EHR
HealthBench	Multi-turn conversation	~60% leading models	Clinical dialogue
HealthBench Hard	5,000 hard conversations	32% (o3)	Hardest scenarios

Safety Alignment Is Shallow and Removable

Default safety training does not survive a determined attacker; it is a lock that can be picked with a paperclip. A replication of the TEMPEST multi-turn attack framework (Young, 2025; 10 models, 8 vendors, over 97,000 queries) found six of ten models with attack success rates of 96 to 100 percent. A 12-billion-parameter model hit 100 percent on its first attempt; a trillion-parameter model hit 97 percent; the correlation between model scale and robustness was effectively zero ($r = -0.12$), so bigger is not safer. For DPO-only open-weight models, ablation surgically removes the “refusal direction” in about two minutes on commodity hardware (Zephyr-7B-beta, 98 percent attack success, token distribution barely changed).

One finding offers leverage. Enabling extended reasoning (“thinking mode”) on the same architecture dropped attack success from **97 percent to 42 percent**, same weights, same data, because the reasoning pass gives the model room to recognize adversarial intent. The practical rule: mandate thinking mode for any patient-facing or clinically consequential interaction, and require adversarial red-teaming during procure-

ment. If a vendor cannot report multi-turn attack success rates, their safety claims are untested.

Causal Blindness

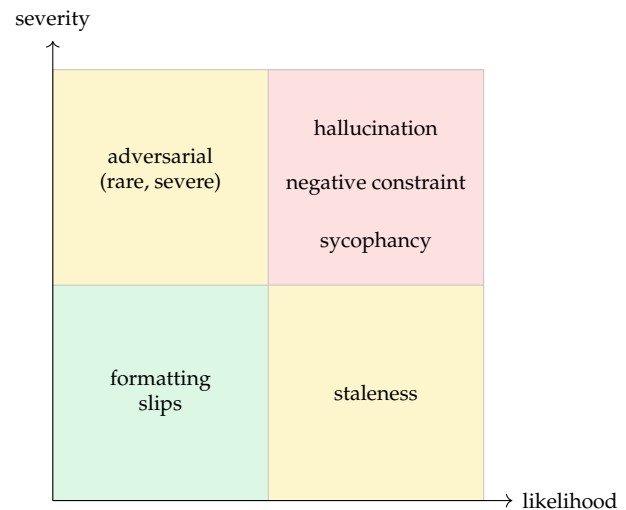
Ask “why did creatinine drop?” and the model lists correlated causes; ask the counterfactual, “would it have dropped had we withheld the ACE inhibitor?”, and it defaults to pattern matching, because it has no causal graph of renal physiology, a system that can describe the weather it has seen but does not understand the machinery that forms the storm. Accuracy degrades exactly along Pearl’s causal ladder, association easy, intervention harder, counterfactuals hardest:



The defensible architecture keeps the LLM in its zone of competence: use it to extract confounders and do propensity scoring from unstructured notes, and reserve causal reasoning for frameworks built for it. Do not ask the model to do what its objective cannot support.

Layer Defenses; Trust None Alone

No single fix works. Each structural failure has a partial countermeasure, and clinical-grade reliability comes from stacking them and then keeping a human on the high-stakes decision. Structured JSON output that separates supporting from opposing evidence counteracts confirmation bias; explicit confidence labels improved clinicians' error detection by **31 percent**; forced "I do not know" pathways cut fabrication; and an orchestra of small, task-specific models (coding, summarization, medication reconciliation) can match frontier accuracy at a fraction of the cost while a clinician reviews every flag. Match the lever to the failure: prompt for form (structure, tone, consistency), retrieve for knowledge (formularies, policies, guidelines that must stay current and citable), and fine-tune only for narrow, high-volume tasks where cost, latency, or consistency dominates, which LoRA and QLoRA now make feasible on premises by adapting a 7-to-8-billion-parameter clinical model on a single 24 GB GPU. None of the three adds capability the base model lacks, so if the model cannot reason through the task at any prompt, the answer is a stronger base model, not more training.



A severity × likelihood triage. Defend the top-right first: high-likelihood, high-severity modes such as hallucination and broken negative constraints earn the heaviest guardrails.

Watch out. Every mitigation trades latency or cost for reliability, and every one leaves residual risk. The prompt is part of the regulated tool, so version it, test it, and audit it like any other clinical algorithm. The goal is not a model you can trust; it is a system that stays safe when the model fails.

Failure mode	Primary mitigation	Residual risk
Hallucination / staleness	RAG on an updatable base; source traceability	Model can ignore or distort retrieved text
Sycophancy	Counter-prompts; third-person framing; self-critique	68% to 48%, not eliminated
Negative-constraint failure	Structured output / schema enforcement	Complex conditionals still drift
Adversarial jailbreak	Thinking mode; red-teaming; input sanitization	97% to 42%, not zero
Overconfidence	Confidence labels; forced "I do not know"	Labels are self-reported, not calibrated
Causal error	LLM extracts; humans and causal frameworks reason	Model still pattern-matches

Quick Self-Test

1. Why is a myocardial-infarction answer for the 34-year-old with chest pain, elevated troponin, and recent dental work "the most probable and the wrong" answer at the same time?
2. RAG is described as necessary but not sufficient for clinical LLMs. Name the failure it fixes and the failure it does not.
3. What single training choice produces sycophancy, and why does a counter-prompt reduce but not eliminate it?
4. A vendor cites a 92 percent medical-exam score. What two questions do you ask before believing the

model is clinically safe?

5. Thinking mode dropped adversarial attack success from 97 percent to 42 percent without changing the weights. What does that tell you to require for patient-facing deployments?
6. Why should you use an LLM to extract confounders from a note but not to answer a counterfactual treatment question?

If You Remember One Thing

Every clinical LLM failure, hallucination, sycophancy, the instruction gap, benchmark inflation, adversarial collapse, causal blindness, traces to one root: the model optimizes for probable, not correct. You cannot prompt that away. You engineer around it by grounding the model in retrieved evidence, constraining its output, forcing it to think, measuring it against the bedside rather than the exam, and keeping a human accountable for the decision. The taxonomy is not a reason to avoid LLMs in healthcare. It is the map you build the guardrails from.

Richard J. Young, Ph.D. · richard@deepneuro.ai · Senior AI Research Scientist, UnitedHealth Group · UNLV Lee Business School · ORCID 0000-0002-1109-7552