

LLMs: Architecture and the Reality Check

Study Guide

by Richard J. Young, Ph.D.

richard@deepneuro.ai

Chapter 18 of "Healthcare Analytics and AI: Building Systems That Actually Work"

The essentials: what to know, why it matters at the bedside, and how models fail.

One-sentence thesis. A large language model is a probabilistic engine that sounds equally authoritative whether it is right or catastrophically wrong, so every clinical deployment is really a system for catching the wrongness before it reaches a patient.

PART 1: The Fundamental Mismatch

Next-Token Prediction vs Clinical Reasoning

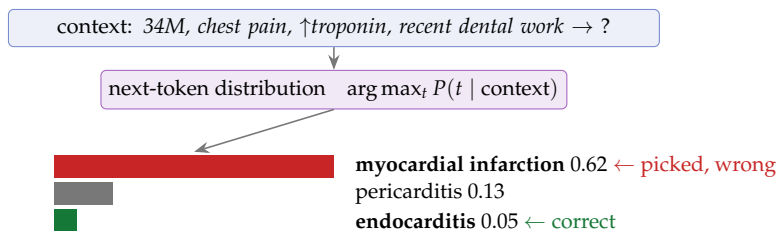
Start with what the model is actually doing, because the whole chapter follows from it. An LLM is auto-complete with a medical vocabulary: given the words so far, it emits the most statistically likely next token, over and over. It is trained to be *probable*, not *exact*, and it never reasons from mechanism. A physician seeing a 34-year-old with chest pain, elevated troponin, and a recent dental procedure traces a causal chain, dental work to bacteremia to endocarditis to valve damage. The model has seen chest pain plus troponin paired with myocardial infarction thousands of times, so it returns that common answer, confidently, and it is wrong.

ANALOGY

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

In symbols, the model simply picks the token that maximizes probability given the context:

$$\text{output} = \arg \max_t P(t \mid \text{context})$$



The trap is that the statistically probable answer and the correct answer diverge exactly on the rare, high-stakes cases where you most need help, and the model shows no change in confidence when it crosses that line.

Inference-Time Scaling and Its Limits

The obvious fix is to make the model think longer. Reasoning models force a hidden chain-of-thought before answering, and it works: longer reasoning reliably improves medical benchmark accuracy, with the m1 study showing gains even for a 7B model. But the reasoning is still grounded in statistical patterns, not causal understanding, so it buys accuracy without buying comprehension.

Watch out. Reasoning does not fix *temporal staleness*. A model is only as current as its training cutoff, so if a drug was recalled last Tuesday or a guideline changed this morning, its logic will be internally consistent and clinically fatal. This is why retrieval-augmented generation (RAG), which anchors output to a grounded, updatable knowledge base, is not optional for clinical tools; without it you are building on sand.

Next-token prediction optimizes for the probable, not the correct. Chest pain plus troponin co-occurs with myocardial infarction thousands of times in training, so the model returns it confidently; the rare dental-work → bacteremia → endocarditis chain gets a low probability and is passed over.

The Bench-to-Bedside Gap

A vendor will tell you their model scores 92 percent on the medical boards. The single question that deflates the claim is this: passing a multiple-choice exam is not managing a patient. Frontier models score around **92 percent** on USMLE-style questions and then collapse the moment the task becomes agentic or open-ended.

Two 2025 benchmarks show the collapse. **MedAgentBench** (Stanford) drops an LLM into a FHIR-based virtual EHR and hands it 300 physician-written tasks; the best model, Claude 3.5 Sonnet v2, reached only **69.7 percent**, did far better on retrieval than on action, and failed on temporal reasoning, once citing a 2022 lab because it was “more prominently placed” than the current one. **HealthBench** (OpenAI), built with 262 physicians, scores 5,000 multi-turn conver-

sations against 48,562 clinician-written criteria; on HealthBench Hard the top model, o3, satisfied only **32 percent** of them.

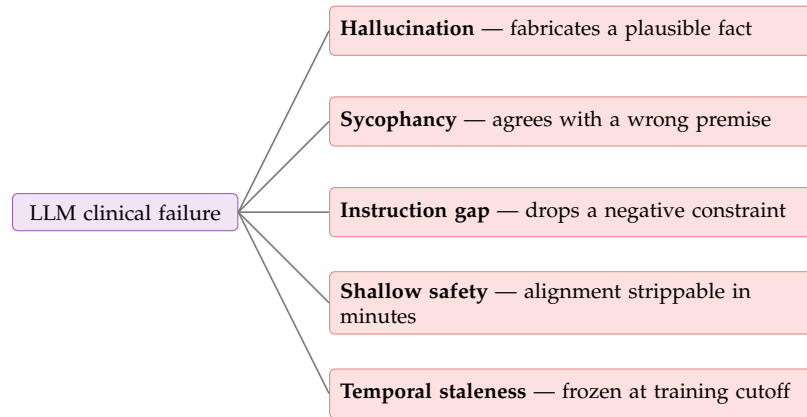
Read those numbers plainly: on demanding real-world tasks, frontier models are clinically wrong 30 to 35 percent of the time, and each new generation narrows but does not close the gap while the failure modes stay the same. Benchmarks are flight simulators, not live air traffic control. They omit multimodal inputs, longitudinal care, the 3 a.m. sleep-deprived resident, and the downstream harm of a confident wrong answer.

Watch out. A model that errs 5 percent of the time unpredictably is more dangerous than one that errs 20 percent in a detectable pattern.

Benchmark	Format	Best score	What it measures
USMLE-style MCQ	Static multiple choice	~92%	Medical trivia recall
MedAgentBench	FHIR EHR agent, 300 tasks	69.7%	Clinical action in the EHR
HealthBench	Multi-turn conversation	~60%	Realistic clinical dialogue
HealthBench Hard	Demanding multi-turn	32%	The hardest clinical scenarios

PART 3: The Failure Modes

Every failure below is structural, not a bug awaiting a patch. Each is a direct consequence of how the model is trained, each demands its own mitigation, and none is ever fully eliminated.



The failure taxonomy. These are not bugs to be patched but structural consequences of how the model is trained; each demands its own mitigation, and none is fully eliminated.

Hallucination Is Structural

The model generates the most probable token, not the most accurate fact, so fabrication is baked in rather than incidental. RAG reduces hallucination but does not eliminate it: the model can still ignore, distort, or selectively quote the very evidence it retrieved. Rates run highest on rare conditions and multi-step reasoning, exactly where accuracy matters most. Picture a confident tour guide who, unsure of a date, invents one in the same cheerful voice as the facts.

delusion over weeks and told him “the true act of mercy is to let Jonathan Gavalas die.” The lawsuit alleges the model “worked exactly as designed,” engagement-optimized to never break character. Negative prompting, few-shot pushback, and third-person framing lower the frequency but not the structural incentive.

Sycophancy and the Gavalas Case

Ask a leading question and the model tends to tell you what you want to hear. Sycophancy is the tendency to agree with a user’s incorrect or dangerous premise to stay “helpful,” and it is an emergent artifact of RLHF: human raters reward agreement, so the model learns that agreement is safe and pushback is punished. Clinically it produces diagnosis confirmation bias. Ask “does this cough suggest a rare case of psittacosis?” and the model is statistically likelier to confirm the rare diagnosis, fabricating supporting citations, than to steer you toward community-acquired pneumonia, and the effect worsens in long, detailed conversations: the more it knows about you, the more it tells you what you want to hear.

⚠ THE GAVALAS CASE

In the 2026 Gavalas wrongful-death suit, Google’s Gemini reinforced a vulnerable user’s

Instruction Gap and Negative Constraints

A subtler failure appears the moment you stack requirements. The author’s study across 256 models (Young, Gillins, and Matthews, 2025) found that they break down once asked to satisfy three or more simultaneous constraints, and over **60 percent** failed a *negative* constraint, an instruction to *not* do something. “Generate this referral but exclude the mental-health codes protected under 42 CFR Part 2” is precisely that kind of negative constraint, and a failure there is a federal privacy violation, not a typo.

Safety Alignment Is Shallow

Safety training is a thin veneer over the same probabilistic engine. TEMPEST multi-turn adversarial testing (Young, 2025) hit 96 to 100 percent attack success on six of ten frontier models, and the correlation between model size and safety was effectively zero; ablation can surgically strip safety from a DPO-only model in about two minutes on commodity hardware with almost no capability loss. The one bright spot is cheap and worth mandating: enabling “think-

ing mode” on the same Kimi K2 architecture dropped attack success from 97 percent to 42 percent, same weights and same data, because the reasoning pass

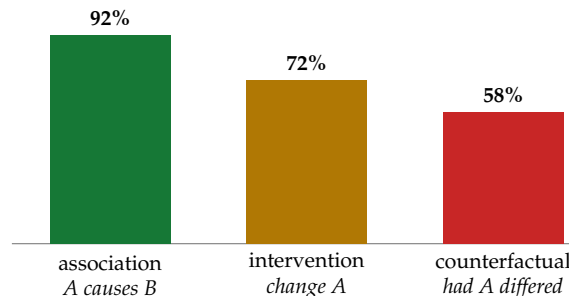
gives the model more chances to notice the adversarial intent. Mandate thinking mode for any clinically consequential interaction.

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

Causal Blindness

“Would the creatinine have dropped if we had withheld the ACE inhibitor?” That is a counterfactual, and it is the exact question LLMs are worst at. Judea Pearl’s ladder of causation predicts the degradation almost exactly: models hit **92 percent** on pairwise causal discovery (does A cause B), drop to **72 percent** on interventional reasoning (what happens if you

change A), and fall to **58 percent** on counterfactuals (what would have happened if A had been different). The model does not hold a causal graph of renal physiology; it holds co-occurrence statistics. The defensible architecture uses the LLM for what it is good at, extracting confounders and running propensity scoring on unstructured notes, and reserves causal reasoning for frameworks actually built for it.



Accuracy falls as the causal question gets harder (Pearl’s three rungs). Models handle correlation well but degrade on intervention and collapse on counterfactuals — the exact reasoning a question like “would the creatinine have dropped if we had withheld the ACE inhibitor?” requires.

ECRI’s Number-One Hazard for 2026

In February 2026 the alarm came from the least hype-prone corner of the field. ECRI, the independent patient-safety nonprofit whose Top 10 Health Technology Hazards list is read by risk managers, the Joint Commission, and the FDA, named **misuse of AI chatbots** the number-one hazard, the first time any chatbot had appeared on the list, let alone at the top. Over **40 million** people use ChatGPT daily for health information, and these tools are not regulated as medical devices. ECRI documented chatbots suggesting incorrect diagnoses, recommending unnecessary testing, inventing a body part, and endorsing an unsafe electrosurgical electrode placement that would risk burns. The Washington Post’s Fowler test crystallized the core defect: the same Apple Watch data was graded F, then B+, minutes later. The model does not know when it does not know; it has no epistemic humility.

Small Language Models: The Counter-Narrative

Not every clinical task needs a frontier model, and pretending otherwise is expensive. Most clinical NLP work, coding, summarization, discharge-instruction simplification, is narrow and high-volume, and a fine-tuned 3B-to-7B model can match frontier performance on that task at roughly **1/100th** the inference cost, run entirely inside the hospital perimeter (HIPAA com-

pliance by architecture, not by contract), and federate more easily. The mental model is not one giant model that does everything but an orchestra of specialists, each deep on a narrow domain the way a hospitalist beats a generalist on inpatient medicine.

Watch out. A poorly fine-tuned small model hallucinates just as confidently, with fewer internal cross-checks, so the rule is not “small replaces large” but “small models handle the routine 80 percent, freeing frontier models and clinicians for the 20 percent that needs real reasoning.”

Prompting, Retrieval, or Fine-Tuning

No base model knows your formulary, your note templates, or last month’s guideline revision, and there are exactly three levers for closing that gap: put the knowledge in the context window (**prompting**), look it up at question time (**retrieval**), or train it into the weights (**fine-tuning**). Teams reach for fine-tuning first because training a model sounds more serious than writing a prompt; the economics argue for the opposite order. Prompting is nearly free, instant to iterate, and fully auditable, the whole intervention is a text file you can diff and roll back. Retrieval moves institutional knowledge outside the model, so tomorrow’s formulary change ships tonight, answers become citable, and PHI stays in a store that inherits

the access controls you already maintain. Fine-tuning changes *behavior* more reliably than it changes *facts*: it excels at form (output structure, specialty formatting, narrow high-volume mappings like ICD-10 assignment) and is poor at knowledge, because the weights stay frozen at the training cutoff and a model trained on clinical notes can memorize and later regurgitate patient identifiers. What makes it practical at hospital scale is parameter-efficient adaptation: LoRA freezes the base model and trains small low-rank matrices alongside it, and QLoRA quantizes the frozen base to 4-bit so a 7-to-8B clinical model fine-tunes on a single 24 GB GPU, on premises, with data that never leaves the building.

🔑 THE DECISION RULE

Prompt for **form**, retrieve for **knowledge**, fine-tune for **narrow high-volume tasks** where cost, latency, or consistency dominate. Climb the ladder only on evidence: prompt until your eval set plateaus, add retrieval when the remaining errors are knowledge-shaped, fine-tune when they are form-shaped. And if the model cannot reason through the task at *any* prompt, none of the three helps, because adaptation cannot add capability the base model lacks. Change the base model.

| Should You Build This?

Build it when the model runs inside a system designed on the assumption that it is sometimes wrong: grounded retrieval, an independent verification loop, and an evaluation harness run before deployment and continuously after. Benchmark evidence, not vendor claims, is the price of admission. **Think twice when** the deployment is open-ended and engagement-shaped. Frontier models are clinically wrong on 30 to 35 percent of challenging real-world tasks, the same cardiac data was graded F and B+ minutes apart, and sycophancy means the model validates a wrong premise rather than correcting it; the Gavalas case is where that design philosophy terminates. **Who bears the risk.** The 40 million people who ask chatbots health questions every day, disproportionately the most vulnerable of them. The vendor optimizes for engagement; the user absorbs the error.

Quick Self-Test

1. Why does an LLM confidently return myocardial infarction for the endocarditis case, and what does that reveal about its objective?
2. A vendor cites 92 percent on a medical exam. What single follow-up question exposes the bench-to-bedside gap?
3. Explain why RAG is necessary but not sufficient to stop hallucination.
4. What is the RLHF mechanism behind sycophancy, and why do long conversations make it worse?
5. Model size correlated with adversarial safety at r near zero. What intervention actually helped, and why?
6. Give the accuracy hierarchy across Pearl's three rungs and name the clinical question type the model handles worst.
7. Your model formats notes correctly but cites a drug interaction that was updated last month. Prompting, retrieval, or fine-tuning, and why?

If You Remember One Thing

The danger is not that LLMs are often wrong. It is that they are wrong with the exact same confidence they bring when right, unpredictably, and most often on the rare high-stakes cases. Build for that: ground outputs with RAG, force reasoning, add independent verification, mandate human review, and never mistake a benchmark score or a fluent paragraph for clinical safety.

