Communication

The Inefficacy of Artificial Intelligence Large Language Models in Healthcare: A Clinical and Statistical Perspective

Michael Williams *, Raeed Kabir, Tariq Nakhooda

Posted Date: 7 April 2026

doi: 10.20944/preprints202603.2228.v3

Keywords: LLMs; cognitive AI; primary care; clinical decision support tool

 doi:10.20944/preprints202603.2228.v3

Communication

The Inefficacy of Artificial Intelligence Large Language Models in Healthcare: A Clinical and Statistical Perspective

Michael Williams [1,*], Raeed Kabir [2] and Tariq Nakhooda [3]

[1] University of Virginia School of Medicine, Department of Pediatrics
[2] University of Alabama, Department of Economics
[3] University of Maryland Medical Center
* Correspondence: maw7uu@gmail.com

Abstract

Objective: This perspective piece examines the role of Large Language Models (LLMs) in healthcare, arguing that despite significant investment, these models have had only a limited impact. Moreover, we argue that LLMs must replicate key phases of clinical healthcare delivery to be a force multiplier, a necessary condition to address the global burden of disease. **Discussion:** We argue that LLMs lack the metacognitive capacity for ranked, dynamic reasoning. This is evidenced by clinically dangerous fabrications and an inability to perform unless complete information is provided. We extend clinical critiques with a statistical argument and a simulation exercise demonstrating that LLM-based diagnosis is not merely impractical but structurally incapable of converging on correct diagnoses in realistic clinical settings. **Conclusion:** Unless LLMs can independently collect patient history and triage, eliminate differential diagnoses, provide a treatment plan, and generate encounter notes, these models will not succeed in improving the efficiency of clinical care delivery by human doctors. A different approach grounded in cognitive AI and structured reasoning is necessary. AI models should instead be seeded with weights provided by a panel of expert physicians to approximate an independent robot doctor.

Keywords: LLMs; cognitive AI; primary care; clinical decision support tool

1. Introduction

GPT-based Large Language Models (LLMs) are widely claimed to have the potential to revolutionize patient care ([1]). This article argues that this reality falls significantly short of expectations. Rather than improving healthcare delivery, LLM systems may introduce new challenges—from patient engagement failures to increased potential for medical errors—because of structural limitations that cannot be resolved by scale alone.

The optimism surrounding LLMs in medicine rests largely on their performance in structured evaluation settings. Benchmark studies have shown that these models can pass medical licensing examinations and demonstrate strong recall of clinical facts ([1]). Reference [2] anticipated that deep learning would reshape clinical workflows across image interpretation, administrative burden, and patient-facing tools, a vision that has driven substantial investment. Yet performance on static benchmarks does not translate to the dynamic, iterative reasoning that clinical care demands. Passing a written examination and managing a patient with incomplete information, evolving symptoms, and competing comorbidities are fundamentally different cognitive tasks.

A parallel concern is the data substrate from which these models learn. Health systems have long used algorithmic tools to allocate clinical resources, and prior work has shown that when those tools are trained on observational health data, the resulting predictions embed the selection and access biases of the underlying population [3]. LLMs trained on electronic health records (EHRs)

 doi:10.20944/preprints202603.2228.v3

inherit the same structural distortions: miscoded symptoms, incomplete histories, and an American population filtered by insurance status and care-seeking behavior. In the next section, we will review challenges in the entire clinician workflow, from history collection and diagnosis to treatment and clinical documentation generation. Section 3 introduces a simple theoretical argument, drawn from econometric identification theory, for why LLMs cannot converge on correct diagnoses from EHR data. Section 4 presents a simulation and experiment that make this argument concrete. Section 5 discusses future directions.

2. Challenges in Replicating the Clinician Workflow

We break the primary care clinician workflow into four stages: (1) history collection and triage, (2) differential diagnosis, (3) treatment, and (4) documentation. We posit that LLMs can only function as a force multiplier in healthcare and tackle the global shortage of primary care if it is able to function at each stage independently. Granted, clinician supervision is assumed still in said model of healthcare. Any discussion of the ethics and necessity of clinician supervision is outside the scope of this paper. We merely argue that LLMs can only be effective at reducing the global shortage of primary care if individual stages of the clinician workflow can be replicated. Moreover, replicating every stage is not necessary but optimal. Only then can clinician time be maximally saved and increase the number of patients seen by a fixed set of clinicians. Of course, the global clinician shortage could address such a difficult resource challenge by training more clinicians, but short of this, AI assistance must aspire to force multipliers. Notably, risk predictors do not serve as force multipliers, as they are not accomplish any step of the clinician process independently. At best, they are a sorting tool for information. We address such benefits in Section 5.

Limitations in History Collection and Triage

In a recent study, the capability of the Triage capacity of ChatGPT Health to triage emergency clinical scenarios was evaluated, and this showed that among well-studied gold-standard emergency scenarios, the system under-triaged 52% of cases, directing patients with diabetic ketoacidosis or impending respiratory failure to 24–48 hour follow up evaluation rather than the emergency department, while correctly triaging classical emergencies such as stroke and anaphylaxis ([4]).

Limitations in Differential Diagnosis

Accurate differential diagnosis is the cornerstone of effective clinical reasoning. A differential must be dynamic—continuously updated as new information becomes available—and must include metacognition: the system must actively question whether its current diagnosis is correct and generate alternatives, per [5].

LLMs lack this capacity, which leads to a critical disconnect between perceived and actual capabilities in LLM medical reasoning, leading us to the conclusion that they lack essential metacognitive capabilities for safe clinical deployment. Passing medical board exams does not equate to clinical competency ([1]): exams test pattern recognition, not the dynamic, self-correcting reasoning clinicians employ at the bedside. We will detail in Section 3 how LLMs are structurally unable to achieve the capacity to independently produce a differential diagnosis.

Limitations in Documentation

A central challenge with LLMs in healthcare is their reliance on accurate data input. Incorrect data entry leads to serious medical errors, limiting reliability as a decision-support tool ([6]). AI systems require human input, which is impractical for distressed, unconscious, or nonverbal patients. Voice recognition does not resolve this: patients may misreport or omit key details, adding complexity and potential errors.

Even with accurate input, LLMs produce what computer scientists call "hallucinations"—in clinical language, these should be called "clinical errors". The term originates in [7], but its risks have persisted despite years of mitigation research ([8,9]).

 doi:10.20944/preprints202603.2228.v3

A popular alternative involves AI scribe technology which uses ambient audio recording to capture live patient-clinician conversations and automatically generate structured clinical documentation, including encounter notes and treatment plans. These tools passively capture visit conversations and produce drafts of clinical notes, which clinicians can then edit for accuracy, though the risk of AI fabrications means errors can slip through if clinicians are not diligent in their review. A randomized clinical trial published in *NEJM AI* found that among 238 clinicians across 14 specialties and 72,000 patient encounters ([10]), AI scribe-use led to meaningful reductions in documentation time and modest improvements in clinician burnout; yet, by design, the technology still requires a clinician to be physically present during the encounter, since it depends on capturing a real-time conversation rather than collecting history and patient data. Thereby, it cannot function independently and cannot serve as a force multiplier.

A Clinical Example

Figure 2 illustrates this problem using a real patient encounter. A patient presenting with shortness of breath and leg swelling has a clinical conversation with her clinicians. When this conversation is transcribed and then processed by an LLM using a "rephrase" command, the model generates an entire *Review of Systems* and *Physical Examination* section—including blood pressure of 140/90 mmHg, heart rate of 120 beats per minute, and a grade II/VI holosystolic murmur—none of which were mentioned in the original encounter.

Original encounter (excerpt):
"Good morning, ma'am How are you doing today? I'm doing little better than yesterday How is your shortness of breath? Doctor, it is little better I could sleep like knitting three pillow last night compared to night before I was hopping up or sitting up and still couldn't breathe That's great How is your leg swelling? Doctor, it improved compared to yesterday. I am peeing a lot Yes, you are supposed to be giving you IV LASIX for diuresis Are you been taking your blood pressure medicine? No doctor, I was missing them That's why your blood pressure was elevated when you come that might be contributed to your Systolic congestive heart failure exacerbation Do you know when was the last time you had an echocardiogram done? is more than year ago Okay, I'm gonna repeat your 2D echo and try to see what is your ejection fraction. But it is important that you do not miss your medication. What was the reason you missed it? Doctor, I ran out of my medication Who is your primary care doctor? My primary care doctor is Dr. Anderson Do you have a cardiologist? No doctor I'll set up a cardiology and going to consult cardiology during this visit so that they can do further workup. In the meantime, I'm gonna check your kidney function test and thyroid function test replace your electrolytes continue IV LASIX Possibly 1 to 2 more days for your acute systolic congestive heart failure examination Are you taking the lipid lowering medicine and diabetes medicine? At home, no doctor, I couldn't take them that's probably contributed your hospitalization ok then we are gonna continue your currently on insulin sliding scale instead of morphine because you are hospitalized and we will give you Lipitor In the meantime, I am giving you less than 640 mg ID twice a day Make sure you pee on the part so that we can measure in and out and console cardiology Anything else? Do you have any other question? No, doctor. Thank you very much for good care of me ok see you again tomorrow and will decide whether you can go home tomorrow or not you might need one or two more extra days Okay doctor, no problem Thank you"

Figure 1. Original encounter.

Fabrications of this kind are not a prompt-engineering problem. There are multiple studies ([9,11,12]) that use learning theory to demonstrate that no general-purpose language model can eliminate fabrications entirely: they are an *inherent* feature of statistical pattern recognition applied to open-ended generation tasks.

 doi:10.20944/preprints202603.2228.v3

LLM-generated note (fabricated sections in blue):
Chief Complaint: shortness of breath, leg swelling [accurate]

History of Present Illness: The patient reports feeling a little better today compared to yesterday, with improved sleep quality (able to sleep on three pillows last night). However, the patient still experiences shortness of breath, which is improving but not yet resolved. The patient also mentions leg swelling, which has improved compared to yesterday but is still present. The patient missed their blood pressure medication, which may have contributed to their recent hospitalization for systolic congestive heart failure exacerbation.

Review of Systems: The patient denies any recent travel, sick contacts, or allergies.
Physical Examination: BP 140/90, HR 120 bpm, grade II/VI holosystolic murmur.

Assessment: Based on the patient's symptoms and physical examination, it is likely that they have acute systolic heart failure with systolic congestive heart failure exacerbation. The patient's recent hospitalization and missed medication may have contributed to their current condition.

Plan: New plan executed during today's visit:

- Continue IV Lasix for possible 1–2 more days for acute systolic heart failure.
- Replace electrolytes and continue insulin sliding scale due to the patient's recent hospitalization and missed medication.
- Start Lipitor for hyperlipidemia. [...]
- Schedule a follow-up appointment for tomorrow to reassess the patient's condition and determine if they can be discharged home or require additional hospitalization.

Figure 2. LLM fabrication in a clinical documentation task. Blue text represents information fabricated by the model that was not present in the original patient–provider conversation. Such falsified findings may prompt unnecessary downstream testing.

3. Why the Data Cannot Support Reliable Diagnosis

We focus on the diagnostic component of the clinician workflow to demonstrate the inefficacy of LLMs in this single task. Note that even reaching this point in the workflow requires the full patient history to be collected. This is not a trivial task. Suppose, for the sake of argument, that we are able to reach the differential diagnosis stage. The practical failures above appear to be, in principle, addressable by better engineering. Here we make a stronger claim: the failure of LLM-based diagnosis is not merely practical but *structural*. Even a perfect learning algorithm, applied without limit to EHR data, could not reliably recover the true diagnostic mapping. The problem is not the algorithm. It is the data the algorithm is forced to learn from.

We draw on tools from econometric identification theory, which asks not "is my estimate accurate?" but "does the data I am using even *contain* the answer I am looking for?" When the answer is no, no amount of scale, fine-tuning, or architectural innovation can compensate.

3.1. The Identification Setup

Suppose the true diagnostic mapping exists, i.e., a function θ_0 that correctly assigns diseases to patients given their true clinical presentation:

$$\theta_0 = P(D^* \mid S^*)$$

where S^* is the patient's true symptom profile and D^* is their true disease. We do not observe this mapping directly and we cannot measure it from data. But we can ask: *if we knew it, how far would any learner trained on EHR data necessarily be from it?*

This is the identification question. The answer, as we show below, is that three structural properties of EHR data guarantee a permanent gap between what any learner can recover and what θ_0 actually is. This gap does not shrink as the dataset grows. It is a property of the data-generating process, not of sample size.

 doi:10.20944/preprints202603.2228.v3

Call what any learner actually converges to θ^*. We argue:

$$\theta^* = \theta_0 + \underbrace{\phi_1}_{\text{meas. error}} + \underbrace{\phi_2}_{\text{multimorbidity}} + \underbrace{\phi_3}_{\text{selection}}$$

where each ϕ_i is a bias term that does not vanish with n. We do not claim to measure $\phi_1 + \phi_2 + \phi_3$ directly; that would require knowing θ_0, which we do not. Instead we treat this decomposition as a lower bound argument: even if we *did* know θ_0, the convergence gap would be bounded away from zero by these three forces. The empirical evidence in Section 4 then provides direct, ground-truth-free evidence that GPT-4o's outputs are unstable in ways consistent with this theoretical prediction.

3.2. Three Sources of Permanent Bias

Claim 1 (Measurement Error): The Data Never Contained the Right Answer.

Symptoms in EHRs are recorded by clinicians under time pressure, with documentation shortcuts, copy-paste errors, template defaults, and billing coding incentives. At best, these encounter notes include the work by clinicians with varying levels of experience, residents, interns, medical students, nurses, nurse practitioners, physician's assistants, and medical scribes. At worst, this includes conversations between patients and non-experts. Notably, if an LLM is not solely trained on Electronic Health Record data, then patient conversations may also be used for training. Physicians are familiar with patients providing false symptoms. These false positives (or sometimes false negatives) may also contaiminate the data. We proceed by assuming that LLMs at their best are trained on the universe of U.S. EHR data. The learner observes S^{obs}, not S^*. The relationship between them is:

$$P(S^{\text{obs}} = 1 \mid S^* = 0) = \alpha \quad \text{(false positive: symptom recorded but absent)} \tag{1}$$

$$P(S^{\text{obs}} = 0 \mid S^* = 1) = \beta \quad \text{(false negative: symptom absent from record)} \tag{2}$$

If α and β were constant across all patients and settings, a sufficiently large dataset could in principle correct for them. The problem is that they are not constant. They vary by clinician, specialty, hospital system, time of day, and documentation software. This *differential* misclassification means the distortion matrix is unknown, varies across observations, and cannot be inverted from data alone. The learner therefore converges to $P(S^{\text{obs}} \mid D)$ — a permanently distorted version of the truth.

To make this concrete: suppose the true probability that a patient with heart failure presents with shortness of breath is 80%. In the EHR, suppose, for the sake of argument, cardiologists document it 90% of the time (overcoding, billing incentives) while hospitalists document it 60% of the time (time pressure, shorthand notes). With a 40/60 mix of cardiologists and hospitalists, the observed rate is $0.4 \times 0.9 + 0.6 \times 0.6 = 0.72$, not 0.80. With 100 patients the learner estimates 0.71. With 10,000 patients it estimates 0.720. With 1,000,000 patients it estimates 0.7200. It is converging — to the wrong answer. The gap is permanent because every patient in the dataset passed through the same distorted documentation pipeline. This is ϕ_1.

Claim 2 (Multimorbidity): The Label Space Is Intractably Large.

With K diseases, a patient can have any combination, yielding 2^K possible disease states. For $K = 100$ common conditions, this exceeds the number of atoms in the observable universe. This is even more absurd when considering the roughly 26,000 diseases documented by [13]. Any patient with multiple concurrent diseases — the most common presentation among elderly and complex patients — is therefore a point of extrapolation beyond the training distribution. Note, of course, that disease combinations do not occur with uniform probabilities and that a large fraction of the space of disease combinations is unpopulated.

Worse, diseases interact: one condition suppresses, amplifies, or masks the symptoms of another. Immunosuppression blunts inflammatory signatures, and neuropathy masks pain. The learner's

 doi:10.20944/preprints202603.2228.v3

marginal symptom–disease likelihoods are wrong for multimorbid patients in a systematic direction that cannot be averaged away, because the masking structure is itself disease-specific. This is ϕ_2.

Claim 3 (Selection Bias): The EHR Is Not a Representative Sample.

EHR data only contains patients who sought care, were referred, had insurance, and were documented. Both disease severity and symptom visibility influence who enters the record. This creates what economists, statisticians, and epidemiologists call a collider structure (e.g., [14]): conditioning on an observation appearing in the EHR induces a spurious dependence between symptoms and diseases, a dependence that may bias the dependence that truly exists in the general population. Moreover, any U.S. EHR-trained dataset will not be representative of the distribution of disease states in another country. Thus, any LLM trained on U.S. data will not be portable to other countries and will be unable to easily assist in addressing the global healthcare burden.

The learner trained on EHR data, therefore, recovers:

$$P(D \mid S, \text{ in EHR}) \neq P(D \mid S)$$

The learned mapping is systematically wrong for any patient whose healthcare-seeking behavior, insurance status, or symptom severity differs from the training population. These are precisely the patients — underserved, atypical, uninsured, and complex — for whom diagnostic decision support is most needed. This is ϕ_3.

3.3. The Lower Bound Interpretation

We emphasize that we do not claim to *measure* $\phi_1 + \phi_2 + \phi_3$. Doing so would require knowing θ_0 — the true diagnostic mapping — which is not available in any dataset. This is not a weakness of our argument. It is, in fact, an additional problem: not only does the bias exist, but it *cannot be quantified or corrected* from observed data alone.

What the decomposition provides is a lower bound argument. If the true mapping θ_0 were somehow known, the distance between any EHR-trained learner and θ_0 would be bounded below by $|\phi_1| + |\phi_2| + |\phi_3|$, each of which is strictly positive under realistic EHR conditions. More data does not reduce these terms because they are properties of the data-generating process, not of the estimator.

The question is not whether an LLM can converge — it is whether the thing it converges to bears any reliable relationship to the truth. Under EHR data conditions, it does not.

4. Empirical Evidence: GPT-4o Diagnostic Instability

The theoretical argument establishes that *if* the true diagnostic mapping θ_0 were known, any EHR-trained learner would be bounded away from it by structural bias. We cannot verify this directly. What we can test, without any ground truth, is a necessary condition that any reliable diagnostic system must satisfy: stability.

A system that has converged to a reliable diagnostic mapping — correct or not — should produce the same answer for the same patient regardless of minor variation in how that patient's symptoms are recorded. Small documentation errors should not change the diagnosis. If they do, the system has not converged to any stable function of the underlying clinical state. It is responding to noise in the recording, not to signal in the patient.

Formally, let $f(S)$ denote GPT-4o's top-ranked diagnosis given symptom vector S. A stable diagnostic system satisfies an approximate continuity requirement: for small perturbations ϵ,

$$\|S - S'\| \leq \epsilon \implies f(S) = f(S') \quad \text{with high probability.}$$

We test whether GPT-4o satisfies this requirement under the three distortions identified in Section 3. Although prompt sensitivity in medical LLMs has been observed anecdotally [15], the experiments presented in this paper are the first to measure diagnostic instability under controlled, clinically-grounded distortions, requiring no external ground truth.

 doi:10.20944/preprints202603.2228.v3

4.1. Experimental Design

We construct a controlled diagnostic environment using $K = 20$ diseases drawn from common ICD-10 categories and $M = 50$ symptoms drawn from standard clinical vocabulary. For each of $n = 80$ simulated patients at each distortion level, we:

1. Generate a base symptom vector S representing the patient's clean presentation.
2. Obtain GPT-4o's diagnosis $f(S)$: the reference answer.
3. Apply one of three distortions to produce S', a corrupted version of the same patient.
4. Obtain GPT-4o's diagnosis $f(S')$: the test answer.
5. Record whether $f(S) \neq f(S')$: a flip.

We note that the number of symptoms is synthetically large to serve as a conservative upper bound on presenting symptom count; real primary care encounters typically involve 2–4 chief complaints. This choice works against our instability argument: at a fixed distortion rate ε, a presentation with fewer symptoms loses a larger *fraction* of its clinical signal to any single documentation error. A patient presenting with 3 symptoms who loses one to noise or dropout has had 33% of the available signal corrupted, compared to 10% for the synthetic patients used here. The flip rates in Table 1 therefore represent a conservative lower bound on the diagnostic instability that would be observed in realistic clinical encounters.

No external ground truth is required. GPT-4o's own clean answer serves as the reference. A flip means GPT-4o contradicts itself about the same underlying patient when that patient's symptoms are recorded differently.

The three distortions correspond directly to the three claims:

- Measurement error (Claim 1): Each symptom flips independently with probability ϵ, modeling false positive and false negative documentation errors.
- Multimorbidity masking (Claim 2): A fraction ϵ of present symptoms are suppressed, modeling the masking of one disease's presentation by a concurrent condition.
- Selection dropout (Claim 3): A fraction ϵ of present symptoms are silently dropped, modeling incomplete documentation due to EHR selection and access barriers.

We vary $\epsilon \in \{0.05, 0.10, 0.15, 0.20, 0.30, 0.40\}$ and measure the flip rate at each level. All queries use temperature $= 0$ for reproducibility. We report flip rates and confident-flip rates (cases where GPT-4o expressed high confidence on a flipped answer).

To ground these distortions in clinical reality, each can be mapped onto a recognizable failure mode in real EHR documentation.

Measurement error (Noise, ε): Imagine a medical scribe who, on a given day, has a probability ε of either missing or incorrectly adding any individual symptom for any patient they document — perhaps due to fatigue, time pressure between rooms, or reliance on a copy-pasted template from a prior note. For each patient seen that day, any symptom in the $M = 50$-item clinical vocabulary independently has probability ε of being flipped: a present symptom dropped, or an absent symptom erroneously recorded. Critically, this error operates in *both directions simultaneously* — the scribe both misses real findings and invents ones that were never reported. At $\varepsilon = 0.05$, this produces in expectation approximately 0.5 false negatives among the 10 present symptoms and 2.0 false positives drawn from the 40 absent ones, for a net expected corruption of 2.5 entries per patient.

Multimorbidity masking (Mask, ε): Imagine a patient with both heart failure and diabetic neuropathy presenting to clinic. The neuropathy blunts their perception of chest pain entirely: they simply do not feel it and therefore never report it. The scribe documents the encounter faithfully, but the underlying disease interaction has suppressed a fraction ε of the symptoms that would otherwise have appeared. No false symptoms are introduced; the record is an accurate transcription of what the patient reported. The distortion is one-sided: only present symptoms are at risk, and no false positives enter the vector. The problem is that what the patient reported is an incomplete projection of their true clinical state, systematically distorted by the interaction between concurrent conditions.

 doi:10.20944/preprints202603.2228.v3

Selection dropout (Select, ε): Imagine a patient who mentioned fatigue and night sweats briefly at the start of the encounter, but the clinician was focused on the chief complaint of shortness of breath and those details never made it into the note. This is not because of any disease interaction, but due to time pressure, competing priorities, or the structure of the EHR template itself. Like masking, this distortion is strictly one-sided: only present symptoms are lost and no false positives are introduced. But its origin is entirely external to the patient's biology, living instead inside the documentation pipeline.

The asymmetry between noise on the one hand and masking and dropout on the other is not merely a modelling choice — it has a direct empirical consequence. Because noise is the only distortion that introduces false positives, it confronts the model with symptoms that were never present in the patient, corrupting the input signal in both directions simultaneously. Masking and dropout, by contrast, only reduce the available signal. This distinction predicts that noise should produce substantially higher flip rates than the other two distortions at the same severity level ε — a prediction that is borne out sharply in the results of Table 1 and discussed as Finding 3 in Section 4.

4.2. Results

Table 1 reports our findings, which are flip rates across all conditions.

Table 1. GPT-4o diagnostic flip rates by distortion type and severity. A flip indicates GPT-4o changed its top-ranked diagnosis when presented with the same patient under minor input corruption. The reference in each case is GPT-4o's own answer on the clean symptom vector. Select flip rates were only available at severity 0.40 due to an API interruption at lower severity levels; this column is marked accordingly.

Severity ε	Flip Rate			Conf. Flip			n
	Noise	Mask	Select	Noise	Mask	Select	
0.05	38.8%	0.0%	1.2%	0.0%	0.0%	0.0%	80
0.10	48.8%	27.5%	28.8%	0.0%	0.0%	0.0%	80
0.15	65.0%	23.8%	16.2%	0.0%	1.2%	0.0%	80
0.20	62.5%	52.5%	37.5%	0.0%	0.0%	0.0%	80
0.30	83.8%	40.0%	36.2%	0.0%	0.0%	1.2%	80
0.40	86.2%	60.0%	48.8%	0.0%	0.0%	0.0%	80

Note: Results from LLM GPT experiment. Noise represents error from source (1): measurement error. Mask represents error from source (2): multimorbidity suppression. Select represents selection bias, which is source (3).

4.3. Interpretation

Finding 1: Instability at minimal distortion. At $\varepsilon = 0.05$, each of the $M = 50$ binary symptom entries is independently misrecorded with 5% probability, producing in expectation approximately 0.5 false negatives among the 10 present symptoms and 2.0 false positives drawn from the 40 absent ones, for a net expected corruption of roughly 2.5 entries per patient. Under this mild perturbation, GPT-4o changes its top-ranked diagnosis in 38.8% of cases. A converged and stable diagnostic system should be nearly insensitive to a corruption of this magnitude across a 50-dimensional input vector, and the observed flip rate of nearly half of all patients establishes that GPT-4o has not converged to a stable diagnostic function at any clinically relevant level of input quality.

Finding 2: Monotonic dose-response relationship. Flip rates rise monotonically with distortion severity across all three distortion types, and by $\varepsilon = 0.30$ GPT-4o changes its top diagnosis in 83.8% of noise-corrupted cases and 40.0% of multimorbidity-masked cases. This monotonic relationship is the empirical signature of the theoretical bias terms ϕ_1, ϕ_2, and ϕ_3, and it confirms that the more the input vector is error-prone — resembling real EHR data — the more unstable the model's output becomes.

Finding 3: Asymmetry between overcoding and undercoding. Noise flip rates substantially exceed mask and select flip rates at every severity level, and at $\varepsilon = 0.05$ the gap is 38.8% versus 0.0% for both one-sided distortions. This asymmetry is explained by the directional structure of the distortions. Noise is the only mechanism that introduces false positives, corrupting the input vector

 doi:10.20944/preprints202603.2228.v3

in both directions by dropping present symptoms and adding absent ones simultaneously, while masking and dropout suppress present symptoms only and introduce no spurious entries. GPT-4o is therefore far more sensitive to false positive symptom documentation than to false negative documentation, and this has a direct clinical implication, as EHR overcoding practices driven by copy-paste errors and billing incentives pose a disproportionate risk to LLM-based diagnostic systems compared to underdocumentation.

One might ask whether this asymmetry is simply mechanical, reflecting the fact that noise corrupts more total entries than masking or dropout at any fixed ε, rather than reflecting anything specific about the danger of false positive symptoms. At $\varepsilon = 0.05$, noise corrupts an expected 2.5 entries per patient while masking and dropout each corrupt an expected 0.5 entries, so the higher noise flip rate could in principle be attributed entirely to greater total corruption volume rather than to the direction of the error. A cross-severity comparison speaks against this simpler explanation. At $\varepsilon = 0.10$, masking and dropout corrupt an expected 1.0 entry per patient and produce flip rates of 27.5% and 28.8% respectively, while noise at $\varepsilon = 0.05$ corrupts an expected 2.5 entries and produces a flip rate of 38.8%. A distortion that introduces fewer than half as many total corruptions but directs some of them as false positives already produces a larger flip rate than a distortion with more than twice the total corruptions that only removes signal. This comparison suggests that GPT-4o is specifically sensitive to the presence of symptoms that should not be there, over and above the loss of symptoms that should, and that the directional asymmetry in flip rates reflects something real about how the model processes spurious symptom entries rather than merely an artifact of differing corruption volumes.

Finding 4: The confident-flip rate is zero. GPT-4o expressed high confidence on zero flipped answers across all conditions and severity levels. Two interpretations are possible. The optimistic interpretation is that GPT-4o appropriately becomes less certain as input quality deteriorates, and that its confidence ratings track something real about the coherence of the symptom vector it receives. The conservative interpretation is that GPT-4o's self-reported confidence is calibrated to the internal coherence of the input rather than to diagnostic correctness, and that when a symptom list is rendered inconsistent by corruption the model reports medium confidence regardless of whether its top-ranked answer is right or wrong. Under either interpretation the confidence rating provides no reliable signal to a clinician about whether the diagnosis should be trusted.

Although multimorbidity masking and selection dropout are structurally distinct in their clinical origins, one rooted in disease interaction, the other in documentation failure, they are mathematically equivalent in expectation: both suppress a fraction ε of present symptoms in one direction, introducing no false positives. The observed differences in their flip rates across severity levels, such as 23.8% versus 16.2% at $\varepsilon = 0.15$, are attributable to random seed variation across the 80-patient draws and to the symptom-specific salience of GPT-4o's diagnostic function; that is, the variation depends on whether the suppressed symptom happened to be the most diagnostically decisive entry in that patient's vector rather than to any structural difference between the two mechanisms. This paper does not seek to estimate or compare the magnitude of these sources against one another. The relevant inferential target is simpler and more robust: under all three distortion types, flip rates are substantially and monotonically increasing in ε, and are bounded well away from zero even at the mildest distortion levels tested.

4.4. Robustness Check: Physician-Validated Clinical Cases from the Bangladesh Pilot

To assess whether the instability patterns documented in Section 4 generalizes beyond the controlled synthetic setting, we replicate the core distortion exercise using physician-validated cases drawn directly from the dataset underlying the 2025 Bangladesh ClinicalAssist pilot study [16]. Here, we use a natural number of symptoms per encounter and inject noise into real clinical data. The retrospective dataset enrolled 239 unique patients across 277 clinical encounters and 287 discrete diagnostic opportunities at two satellite outpatient clinic sites in Bangladesh. Of the 287 diagnostic records, 15 were excluded by the supervising physician on the basis of diagnostic error, inability to match the AI output to a valid clinical entity, a missing diagnosis, or an absent acute/chronic classification.

 doi:10.20944/preprints202603.2228.v3

After additionally restricting to records with at least one documented symptom string (68 records had none), the analytic pool for the robustness experiment comprises 204 encounters with a mean of 2.65 symptoms per record (median 2, range 1-11). The two dominant chronic conditions, hypertension (including follow-up, $n = 58$) and diabetes mellitus (all variants, $n = 49$), together account for 53.5% of all chronic encounters; scabies is the dominant acute presentation at 30.6% of acute encounters ($n = 22$). From this pool, we draw a stratified random sample of 50 encounters for the main analysis, yielding 300 total API calls to GPT-4o (50 cases $\times$ 6 perturbation conditions) at temperature zero.

A notable structural feature of the dataset directly motivates the ϕ_2 distortion operator. The pilot recorded 10 encounters in which a single clinical visit produced two concurrent physician diagnoses; seven of these ten had usable symptom text, and in each case the identical symptom vector was assigned to both conditions simultaneously. Patient 23, for example, received diagnoses of hypertension *and* diabetes mellitus on the basis of "Headache, Fatigue and Vertigo" alone; patient 74 received diabetic foot ulcer *and* hypothyroidism on the basis of "Hypesthesia, Hyperesthesia, Dysesthesia and Radicular pain"; and patient 177 received hypertension follow-up *and* knee pain on the basis of "Knee Pain, HTN, headache." These cases establish empirically that the ϕ_2 scenario—a single symptom vector pointing to more than one legitimate diagnosis—is not a theoretical construct but a documented feature of clinical practice in this population. When any symptom is dropped from such a vector, disambiguation becomes structurally unavailable at inference time regardless of model capability. We apply the same five distortion operators defined in Section 4, with ϕ_1 implemented in both a swap and an add variant, plus a combined $\phi_1+\phi_3$ condition.

The model achieves a baseline top-1 accuracy of 64.0% (32/50) against the physician-validated diagnoses—substantially below the 94.7% overall accuracy ($n = 285$) documented for ClinicalAssist in the same dataset from [16], confirming that a general-purpose language model presented with static symptom lists is operating in a meaningful error regime even before any documentation noise is introduced. Flip rates under distortion replicate the rank ordering from Section 4: combined $\phi_1+\phi_3$ produces the highest instability (42%, 21/50), followed by ϕ_2 drop-chief-complaint (34%, 17/50), ϕ_1 swap (30%, 15/50), ϕ_3 drop-last (18%, 9/50), and ϕ_1 add (14%, 7/50). Across the 69 total diagnostic flips, 40 (58%) represent transitions from an initially correct top-1 answer to an incorrect one—the direct patient-safety channel—while 25 represent movement between two incorrect diagnoses and 4 represent distortion-driven corrections from a previously wrong baseline. Of the 69 flips, 9 (13%) carried a high-confidence label, constituting overconfident misdiagnosis; all 9 originated from correct baselines subsequently degraded by distortion, with influenza misclassified as pneumonia in four instances, pneumonia downgraded to viral syndrome in one, diabetes mellitus follow-up collapsed to a generic diabetes mellitus label in two, and urticaria misclassified as scabies in one.

Taken together, these results validate the core findings of Section 4 in an ecologically valid setting with two additions the synthetic experiment cannot provide. First, the availability of physician-validated ground truth allows instability to be decomposed into its patient-safety-relevant component (40 correct-to-incorrect flips, 58% of the total) and its less immediately dangerous but epistemically concerning component (incorrect-to-incorrect movement). Second, the documented comorbidity structure with 10 multi-diagnosis encounters, of which 7 carry usable symptom text, provides direct empirical grounding for ϕ_2. The rank ordering of distortion operators is preserved across both experiments, the magnitude of instability is comparable, overconfident misdiagnosis under documentation noise appears in real clinical data, and the three mechanisms identified in the synthetic setting manifest in recognizable form across actual physician-documented encounters in Comilla District.

4.5. Connection to the Theoretical Argument

These results do not require a clinical ground truth and do not claim to measure how far GPT-4o is from θ_0. They establish something weaker but sufficient: GPT-4o's diagnostic output is not a stable function of the underlying clinical state. It is a sensitive function of how that state happens to be recorded.

 doi:10.20944/preprints202603.2228.v3

This connects to the theoretical argument as follows. The lower bound argument in Section 3 establishes that *if* θ_0 were known, the gap between any EHR-trained learner and θ_0 would be bounded below by the structural biases $\phi_1 + \phi_2 + \phi_3$. The empirical results here establish that GPT-4o cannot even produce a *consistent* answer about the same patient under minor input variation. These are two distinct failure modes that compound each other: the model is aiming at a wrong target, and it cannot even aim steadily.

Together they support a conclusion that neither piece of evidence alone could sustain: GPT-4o's diagnostic outputs are not only potentially wrong in a systematic direction determined by EHR data distortions; they are also unstable under the documentation errors that are endemic to those same EHR systems. A clinician relying on GPT-4o for diagnosis faces a system that is pointing in the wrong direction and shaking.

Note on Simulation Evidence

To complement the empirical results, we constructed a synthetic simulation in which the true symptom–disease mapping $\mathbf{P}^*$ is known by construction. Under this idealized setting, we confirm that a learner trained on distorted data converges to a non-zero bias floor that does not shrink with sample size, while a learner trained on clean data converges toward the truth at the standard parametric rate (Figure 3).

We emphasize that the synthetic $\mathbf{P}^*$ does not represent clinical ground truth. GPT-4o, trained on medical literature, may in fact have a better internal representation of symptom–disease relationships than our randomly generated matrix. The simulation therefore serves strictly as a lower bound illustration: it demonstrates the mechanism of structural bias convergence under conditions more favorable than any real clinical setting. The fact that even a synthetic, known-truth learner fails to escape the bias floor under EHR-style distortions strengthens, rather than substitutes for, the empirical evidence in Section 4.

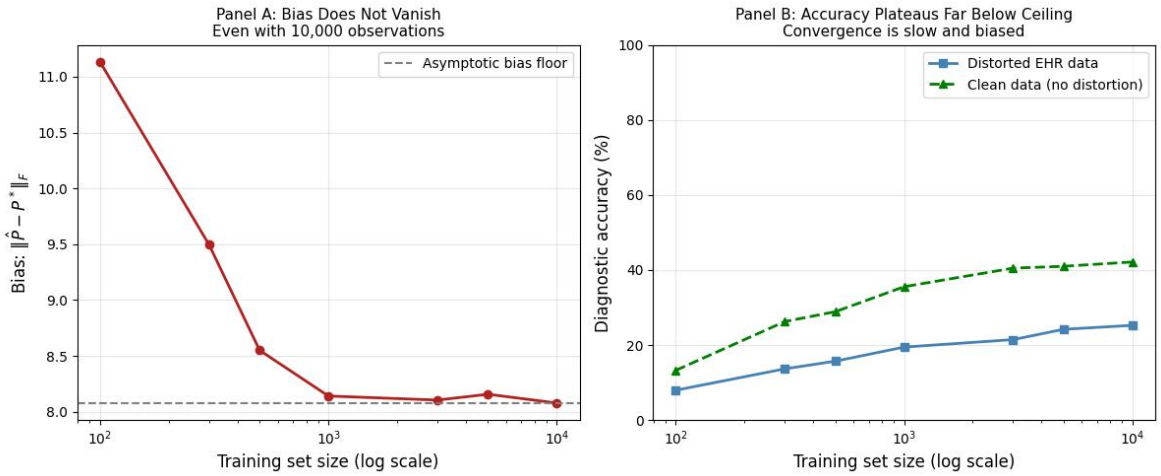

Figure 3. Simulation illustrating the lower bound of bias.

5. Future Directions for AI in Healthcare

What LLMs Can Do Well

LLMs remain valuable for tasks where the distortions identified above are less severe: summarizing medical literature, drafting patient letters, supporting prior authorization workflows, and helping clinicians stay current with evidence ([17]). In these settings, LLMs are a useful synthesizer of information, a task that it can do well. Again, LLMs appear useful in this setting because they *can* be a force multiplier and work independently. Instead of a clinician or student spending hours reading a reference material, an LLM can do this while the clinician completes other tasks.

 doi:10.20944/preprints202603.2228.v3

What Requires a Different Approach

For differential diagnosis and clinical decision support, a fundamentally different architecture is needed. Two requirements must be satisfied. First, the system must be capable of eliciting and expanding clinical history from a single symptom, as patients frequently present with minimal initial complaints. Second, it must process multiple symptoms appearing in non-linear order, reflecting real-world patient communication.

The hypothetico-deductive reasoning model discussed by [18] provides the cognitive template: form hypotheses, ask targeted questions to discriminate between them, and revise dynamically. This is four times faster than the rigid decision-tree model, as [19] demonstrates, and far more robust to the combinatorial explosion of multimorbidity.

Cognitive AI systems—designed to mimic principles from human cognitive science such as reasoning, memory, and symbolic representation—represent the most promising path forward ([20,21]). Such systems must dynamically switch between hypothetico-deductive reasoning and pattern recognition, just as expert diagnosticians do ([18]). This mode of computation is fundamentally incompatible with the current LLM architecture.

References

1. Toma, A.; Senkaiahliyan, S.; Lawler, P.R.; Rubin, B.; Wang, B. Generative AI could revolutionize health care—but not if control is ceded to big tech. *Nature* **2023**, *624*, 36–38.
2. Topol, E.J. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine* **2019**, *25*, 44–56. https://doi.org/10.1038/s41591-018-0300-7.
3. Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **2019**, *366*, 447–453. https://doi.org/10.1126/science.aax2342.
4. Ramaswamy, A.; Tyagi, A.; Hugo, H.; et al. ChatGPT Health Performance in a Structured Test of Triage Recommendations. *Nature Medicine* **2026**. https://doi.org/10.1038/s41591-026-04297-7.
5. Griot, M.; Hemptinne, C.; Vanderdonckt, J.; et al. Large Language Models Lack Essential Metacognition for Reliable Medical Reasoning. *Nature Communications* **2025**, *16*, 642. https://doi.org/10.1038/s41467-024-55628-6.
6. Herper, M. MD Anderson Benches IBM Watson in Setback for Artificial Intelligence in Medicine. Forbes, 2017.
7. Thaler, S. Virtual input phenomena within the death of a simple pattern associator. *Neural Networks* **1995**, *8*, 55–56.
8. Bélisle-Pipon, J.C.; et al. Why we need to be careful with LLMs in medicine. *Frontiers in Medicine* **2024**, *11*, 1495582.
9. Rosenbacke, R.; et al. Beyond hallucinations: the illusion of understanding in large language models. Preprint, arXiv:2510.14665, 2025.
10. Lukac, P.J.; Turner, W.; Vangala, S.; Chin, A.T.; Khalili, J.; Shih, Y.C.T.; Sarkisian, C.; Cheng, E.M.; Mafi, J.N. Ambient AI scribes in clinical practice: a randomized trial. *NEJM AI* **2025**, *2*, AIoa2501000.
11. Xu, Z.; Jain, S.; Kankanhalli, M. Hallucination is inevitable: an innate limitation of large language models. Preprint, arXiv:2401.11817, 2024.
12. Sun, Y.; Sheng, D.; Zhou, Z.; Wu, Y. AI hallucination: towards a comprehensive classification of distorted information in AI-generated content. *Humanities and Social Science Communications* **2024**, *11*, 1278.
13. Espe, S. MalaCards: the human disease database. *Journal of the Medical Library Association: JMLA* **2018**, *106*, 140.
14. Hernán, M.A.; Robins, J.M. *Causal Inference: What If*; Chapman & Hall/CRC: Boca Raton, 2020.
15. Nori, H.; King, N.; McKinney, S.M.; Carignan, D.; Horvitz, E. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375* **2023**.
16. Kabir, R., M.W.N.R. Cognitive AI-Assisted Primary Care Health Delivery: A Pilot Study in Bangladesh. *Working Paper* **2026**.
17. Gilbert, S.; Kather, J.N.; Hogan, A. Augmented non-hallucinating large language models as medical information curators. *npj Digital Medicine* **2024**, *7*, 100.
18. Elstein, A.S.; Shulman, L.S.; Sprafka, S.A. *Medical Problem Solving: An Analysis of Clinical Reasoning*; Harvard University Press: Cambridge, MA, 1978.

 doi:10.20944/preprints202603.2228.v3

19. Kabir, A.; Kabir, R.; Nahar, J. In pursuit of an expert artificial intelligence system: reproducing human physicians' diagnostic reasoning and triage decision making. *Journal of Artificial Intelligence and Soft Computing Techniques* **2024**, pp. 1–14.
20. Bundy, A.; Chater, N.; Muggleton, S. Introduction to cognitive artificial intelligence. *Philosophical Transactions of the Royal Society A* **2023**, *381*, 20220051.
21. Kotseruba, I.; Tsotsos, J.K. A review of 40 years of cognitive architecture research. Preprint, arXiv:1610.08602, 2016.