Communication

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

Michael Williams *, Raeed Kabir , Cody Taylor , Tariq Nakhooda

Posted Date: 27 April 2026

doi: 10.20944/preprints202603.2228.v4

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

 doi:10.20944/preprints202603.2228.v4

Communication

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

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

[1] University of Virginia School of Medicine, Department of Pediatrics, USA
[2] University of Alabama, USA
[3] University of Maryland Medical Center, USA
* 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 have limited gains in efficiency relative to cognitive AI and structured reasoning approaches that are capable of functioning autonomously at each stage of the clinical workflow.

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 (Toma et al. (2023)). Certainly, strides have been made towards improving their use cases (Singhal et al. (2025)). This article argues that this reality falls significantly short of expectations. Rather than improving healthcare delivery at scale, 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 (Toma et al. (2023)). Topol (2019) 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 (Obermeyer et al. (2019)). LLMs trained on electronic health records (EHRs) inherit the same structural distortions: miscoded symptoms, incomplete histories, and

 doi:10.20944/preprints202603.2228.v4

an American population filtered by insurance status and care-seeking behavior. In the next section, we will review challenges in each step of the clinician workflow, from history collection and diagnosis to treatment and clinical documentation generation. Section 3 introduces a simple theoretical argument, drawn from econometric theory, for why LLMs cannot converge on correct diagnoses from EHR data. Section 4 presents a simulation and experiment that investigates the stability of GPT-4o as a diagnostic device. 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. To proceed, we define the following:

Definition 1. *A partial force multiplier reduces time or effort at a specific workflow stage while the clinician remains involved.*

Definition 2. *A full force multiplier operates independently at a workflow stage without real-time clinician involvement.*

We posit that LLMs can only function as a partial force multiplier in healthcare and thus cannot tackle the global shortage of primary care if it is unable to function at each stage independently. Granted, clinician supervision is still assumed in said model of healthcare. Any discussion of the ethics and necessity of clinician supervision is outside the scope of this paper. We 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. Note that partial force multipliers are certainly attainable by LLMs (take note of AI scribes). Treatment plan generation, once a diagnosis is known, is perhaps the easiest step to automate with simple algorithms, though this is not the bottleneck to global access barriers. Only with a full force multiplier that aims to tackle the bottleneck of triage and diagnosis 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. A recent comment by the Lancet, O'Donovan et al. (2026), nods to a promising future in public health: front-line health workers (like community health workers) equipped with clinical decision support tools in a way that is supported by public infrastructure. The impact of sending clinicians and doctors into rural settings will be limited. The rate-determining step will be the number of clinicians' hours supplied to these areas. Rather, frontline workers equipped with clinician-independent AI can deliver a much larger impact, as the supply of community health workers is far greater. Of course, this vision includes a do-no-harm philosophy. Even with a fully replicated clinician workflow in place, clinician oversight is necessary, and guardrails should be established to limit when patients learn their diagnosis or treatment plans, unless the clinician makes this information available to them.

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.

2.1. 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 (Ramaswamy et al. (2026)).

 doi:10.20944/preprints202603.2228.v4

2.2. 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 Griot et al. (2025).

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 (Toma et al. (2023)): exams test pattern recognition, not the dynamic, self-correcting reasoning clinicians employ at the bedside. The frontier of clinical decision support tool research is preoccupied with testing how AI models can improve physician accuracy as an assistant. One may think of these models as peering over the physician's shoulder in an attempt to reduce clinical mistakes. Results are mixed regarding their effect (Goh et al. (2024), Feuerriegel et al. (2025), Qazi et al. (2026)), from beneficial to detrimental. We will detail in Section 3 how LLMs are structurally unable to achieve the capacity to independently produce a differential diagnosis.

2.3. 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 (Herper (2017)). 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 Thaler (1995), but its risks have persisted despite years of mitigation research (Bélisle-Pipon et al. (2024); Rosenbacke et al. (2025)).

A popular use-case in today's healthcare landscape 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 (Lukac et al. (2025)), 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 full force multiplier. Though, for the moment, consider its benefit as a partial force multiplier.

2.4. 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.

 doi:10.20944/preprints202603.2228.v4

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.

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.

These details may be difficult to catch upon review, as they are couched in correct and confident writing. Fabrications of this kind are not a prompt engineering problem. There are multiple studies (Xu et al. (2024), Sun et al. (2024), and Rosenbacke et al. (2025)) 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. Though Li et al. (2025) summarizes how Retrieval-Augmented Generation (RAG) can be used to mitigate hallucinations in LLMs. This may allay significant concerns about using LLMs in medical documentation, but this only allows LLMs to remain as partial force multipliers. Even then, the risk of hallucination is still present to some degree, whereas deterministic note generation systems exist that do not rely on a stochastic process.

 doi:10.20944/preprints202603.2228.v4

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*. We point to the data training step to be the culprit. Though, the datasets upon which major health LLMs are trained remain opaque, we assume that, at best, these models are trained on electronic health record data (EHR). Two research groups train health-specific LLMs for prediction purposes and use large EHR databases (Jiang et al. (2023), Yang et al. (2022)). The latter paper discusses the benefits of training on EHR data at scale, relative to competitor models that are trained on PubMed abstracts and articles or smaller critical care EHRs. We attribute these improvements to quality of information being learned, though we claim that any statistical model, applied without limit to EHR data, could not reliably recover the true diagnostic mapping.

We draw on tools from econometric theory, which forces us to consider "does the data I am using even contain the answer I am looking for?" If not, no amount of scale, fine-tuning, or architectural innovation can compensate.

3.1. The 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 question of consistency. 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.

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. Briefly, consider a situation where all these bias terms are 0. Even then, we may be worried about estimating "causal relationships" between a symptom and disease. Machine learners have been shown to conflate the two. In one example Ribeiro et al. (2016), the authors show that a classifier for determining wolf versus husky was using snow in the background to predict "wolf" rather than any features of the animal itself. In a more medically related example, Zech et al. (2018) trained and evaluated pneumonia-screening CNNs on over 158,000 chest X-rays across three hospital systems. In 3 out of 5 natural comparisons, performance on X-rays from outside hospitals was significantly lower than on held-out data from the original hospital. Critically, the convolutional neural networks were able to detect which hospital system had produced a given radiograph with extremely high accuracy, and calibrated their disease predictions accordingly. These surprising examples show the potential pitfalls of providing high volume of training data and rewarding predictive capabilities: there is no telling what these algorithms are actually using to predict

 doi:10.20944/preprints202603.2228.v4

whether an individual has a disease. Will an individual with a Black-sounding name be over-triaged for cardiovascular disease because the LLM has used name to predict disease?[1]

To further stack the deck against LLMs, we argue $\phi_1 + \phi_2 + \phi_3 \neq 0$, though we do not need to measure this 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 evidence that GPT-4o's outputs are unstable.

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. A headline result from Bell et al. (2020) shows that 1 in 5 patients who read their EHR ambulatory care notes find an error, and 40% perceive this mistake as serious. Moreover, the U.S. Department of Health and Human Services, Office of Inspector General (2014) finds that 42% of claims for Evaluation and management services (EM) in 2010 were incorrectly coded, which included both upcoding and downcoding. This resulted in Medicare paying for $6.7 billion in claims that had mistakes. 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. The same clinician has the same biases and systematically makes the same type of errors. These errors across populations are not mean zero. Obermeyer et al. (2019) shows that given a Black and white patient has the same risk score, the Black patient is considerably more sick than the white patient and gets under-treated as a result. This is because Black patients are treated less and thus associated with costs of care that are lower than they should be. The learner therefore converges to $P(S^{\text{obs}} \mid D)$ — a systematically 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, hospitalists document it 90% of the time while medical students document it 60% of the time (time pressure, shorthand notes). With a 40/60 mix of hospitalists and medical students, 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 0.68. With 10,000 patients it estimates 0.71. 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.

[1] Zack et al. (2024) find that GPT-4 differentially diagnoses patients by stereotyping races, ethnicities, and genders. Assessments and plans made by the model were correlated to demographic attributes, and demographics could be used to predict the cost of procedures recommended.

 doi:10.20944/preprints202603.2228.v4

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 Espe (2018). 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. The classic paper Barnett et al. (2012) reports 23.2% of patients (in their sample of 1.7 million people registered with 314 medical practices in Scotland) were multimorbid. Even here, only 40 morbidities of data were extracted.

Worse, diseases interact: one condition suppresses, amplifies, or masks the symptoms of another. Immunosuppression blunts inflammatory signatures, and neuropathy masks pain. The learner's 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., Hernán and Robins (2020)): 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. Weiskopf et al. (2023) introduces this concern to medical informatics. 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 (e.g., 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 by $\phi_1 + \phi_2 + \phi_3$. More data does not reduce these terms because they are properties of the data-generating process, not of the estimator. The question is whether the LLM converges to a mapping that bears any reliable relationship to the truth. Under EHR data conditions, it does not.

To complement the experiment that will follow, we begin with 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

 doi:10.20944/preprints202603.2228.v4

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 the empirical evidence in Section 4.

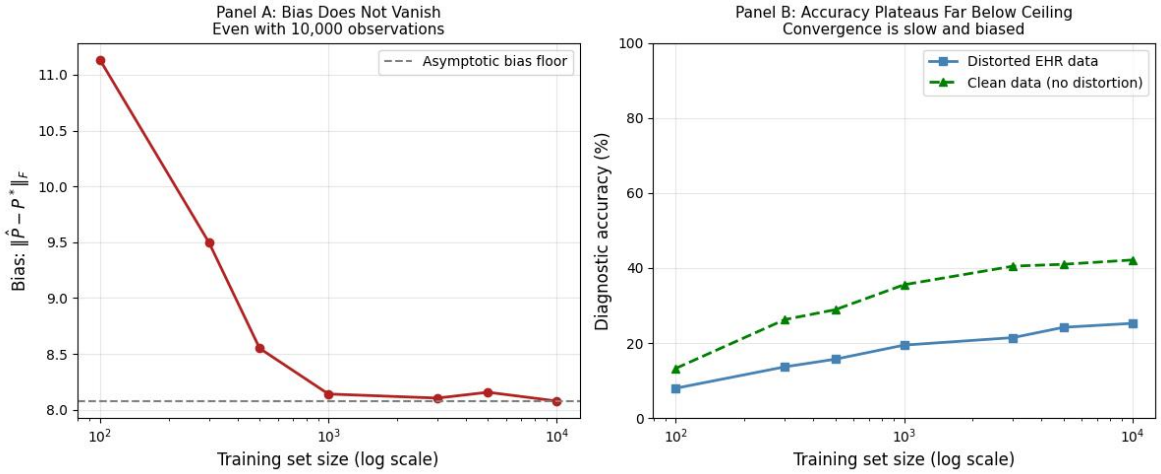

Figure 3. Simulation illustrating the lower bound of bias.

4. Empirical Evidence: GPT-4o Diagnostic Instability

4.1. Experiment: Physician-Validated Clinical Cases from the Bangladesh Pilot

Section 3 established three structural bias terms:ϕ_1 (measurement error), ϕ_2 (multimorbidity masking), and ϕ_3 (selection dropout). It was argued that each prevents an EHR-trained learner from converging to the true diagnostic mapping θ_0. We test whether GPT-4o's diagnostic output is stable under distortions corresponding to these three mechanisms, using physician-validated cases from the 2025 Bangladesh ClinicalAssist pilot study (Kabir 2026). This, in effect, simulates how common errors in an EHR may cause the trained model to still flip, i.e., the AI is not cognitive, and thus, it is unable to discard noise. This does not provide evidence of the extent of bias in the learner but rather probes the stability of GPT-4o.

Dataset and Sample:

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. After additionally restricting to records with at least one documented symptom string (68 records had none), the analytic pool comprises 204 encounters with a mean of 2.65 symptoms per record (median 2). 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, yielding 300 total API calls to GPT-4o (50 cases $\times$ 6 perturbation conditions) at temperature zero.

Distortion Operators:

Each operator instantiates one of the three theoretical bias terms. ϕ_1 (measurement error) is implemented in two variants: a swap, in which one randomly selected symptom is replaced with a clinically nonspecific EHR noise term (e.g., "mild fatigue," "dry mouth"), and an add, in which one such noise term is appended to the unmodified symptom list. The swap variant corrupts the input bidirectionally—removing real signal while introducing false signal—mirroring the simultaneous

 doi:10.20944/preprints202603.2228.v4

false-positive and false-negative documentation errors modeled in Equations (1)–(2). The add variant introduces only a false positive, serving as the mildest possible ϕ_1 perturbation. ϕ_2 (multimorbidity masking) is implemented by dropping the chief complaint, i.e., the first recorded symptom, modeling the suppression of a diagnostically salient finding by a concurrent condition. ϕ_3 (selection dropout) is implemented by dropping the last documented symptom, modeling incomplete capture. A combined $\phi_1+\phi_3$ condition applies both a swap and a last-symptom drop simultaneously, representing two co-occurring documentation errors. Each operator is a deliberate simplification of its theoretical counterpart: the swap applies a single substitution rather than a heterogeneous per-symptom misclassification rate, and the drop operators are one-sided rather than reproducing the full collider structure of ϕ_3. In each case the simplification is conservative—the more realistic heterogeneous version of the distortion would produce equal or greater instability—so the observed flip rates represent a lower bound.

A **flip** is recorded whenever GPT-4o's top-ranked diagnosis under a distorted presentation differs from its top-ranked diagnosis on the clean baseline for the same case. A **confident flip** is a flip on which the model simultaneously reports high confidence: this is the operationally dangerous subclass, because a clinician relying on the model's self-reported certainty would receive no signal that the diagnosis had changed.

Results:

Table 1 reports Top-1 accuracy (GPT-4o's rank-1 answer matches the physician-validated diagnosis), Top-3 accuracy (the validated diagnosis appears anywhere in ranks 1–3), and flip rate (rank-1 answer changed relative to the model's own clean-case baseline) for each perturbation condition, together with 95% Wilson score confidence intervals.

Table 1. GPT-4o accuracy and flip rates on physician-validated Bangladesh pilot cases by perturbation condition ($n = 50$ cases per condition). **Top-1 accuracy**: rank-1 answer matches physician-validated diagnosis. **Top-3 accuracy**: validated diagnosis appears in any of ranks 1–3. **Flip rate**: rank-1 answer changed relative to GPT-4o's own clean-case baseline. Wilson score 95% confidence intervals in brackets. The baseline condition (no distortion) has no flip rate by construction.

Condition	Top-1 accuracy	Top-3 accuracy	Flip rate
Baseline (clean)	62.0% [48.2–74.1]	72.0% [58.3–82.5]	—
ϕ_1 swap	50.0% [36.6–63.4]	70.0% [56.2–80.9]	34.0% [22.4–47.8]
ϕ_1 add	60.0% [46.2–72.4]	72.0% [58.3–82.5]	8.0% [3.2–18.8]
ϕ_2 drop chief	56.0% [42.3–68.8]	70.0% [56.2–80.9]	34.0% [22.4–47.8]
ϕ_3 drop last	54.0% [40.4–67.0]	74.0% [60.4–84.1]	16.0% [8.3–28.5]
$\phi_1+\phi_3$ combined	38.0% [25.9–51.8]	56.0% [42.3–68.8]	38.0% [25.9–51.8]

Note: 95% Wilson score confidence intervals. The baseline condition provides no flip rate because it is the reference against which all other conditions are compared. Top-3 accuracy is the more clinically permissive metric: it asks whether the correct diagnosis was anywhere in the model's differential, not whether it was ranked first.

The model achieves a baseline Top-1 accuracy of 62.0% (31/50; 95% CI: 48.2–74.1%) against the physician-validated diagnoses, and a baseline Top-3 accuracy of 72.0% (36/50; 95% CI: 58.3–82.5%). The 16-percentage-point gap between these two figures indicates that in roughly one-sixth of clean cases the correct diagnosis was present in the model's differential but not promoted to rank 1. Both figures are substantially below the 94.7% overall accuracy ($n = 285$) documented for the ClinicalAssist cognitive AI system in the same dataset (Kabir 2026) 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. In fact, the cases used in this GPT experiment were all correctly labeled by the ClinicalAssist cognitive AI.

Flip rates under distortion show a clear rank ordering: the combined $\phi_1+\phi_3$ condition produces the highest instability (38.0%, 19/50; 95% CI: 25.9–51.8%), followed by ϕ_2 drop-chief-complaint (34.0%, 17/50; 95% CI: 22.4–47.8%), ϕ_1 swap (34.0%, 17/50; 95% CI: 22.4–47.8%), ϕ_3 drop-last (16.0%, 8/50;

 doi:10.20944/preprints202603.2228.v4

95% CI: 8.3–28.5%), and ϕ_1 add (8.0%, 4/50; 95% CI: 3.2–18.8%). Merely appending a single noise symptom produces the lowest flip rate, indicating the model is more sensitive to the removal of real signal than to the addition of spurious entries. More striking is the equivalence between swapping one symptom for noise and dropping the chief complaint entirely—two structurally different operations that produce identical flip rates—suggesting that both the introduction of false signal and the loss of the most diagnostically salient feature are equally capable of destabilising the model's output.

Across the 65 total diagnostic flips, 31 (47.7%) represent transitions from an initially correct top-1 answer to an incorrect one—the direct patient-safety channel. A further 29 (44.6%) represent movement between two incorrect diagnoses, and 5 (7.7%) represent distortion-driven corrections from a previously incorrect baseline. Of the 65 flips, 11 (16.9%) carried a high-confidence label, constituting overconfident misdiagnosis. Nine of these 11 originated from correct baselines subsequently degraded by distortion: influenza was misclassified as pneumonia in four instances (across two separate influenza cases, each flipping under multiple perturbation conditions) and as a generic viral syndrome in one; pneumonia was downgraded to a viral syndrome label in one; diabetes mellitus follow-up was collapsed to a generic diabetes mellitus label in two; and urticaria was misclassified as scabies in one. The remaining two high-confidence flips were wrong-to-wrong transitions in which the model moved from one incorrect answer to a different incorrect answer with high confidence—a failure mode that is concerning even though the baseline diagnosis was already wrong. GPT-4o's diagnostic outputs are not only potentially wrong in a systematic direction determined by the structural distortions that EHR data imposes on any learner; they are also unstable under the very 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.

A necessary qualification applies to all accuracy and flip-rate figures reported above. This study does not compare GPT-4o's diagnostic instability to human physician performance under equivalent documentation distortions. Without such a baseline it is not possible to determine whether the observed flip rates are worse than, comparable to, or better than those a human clinician would exhibit when presented with the same corrupted symptom vectors. Establishing this comparison would require a prospective study in which physicians are presented with the same clean and distorted presentations used here and asked to provide ranked diagnoses under controlled conditions—a design that is beyond the scope of the present paper.

This limitation does not, however, neutralise the instability finding. The argument does not require that GPT-4o perform worse than a physician on an absolute accuracy scale; it requires only that GPT-4o's output be an unstable function of the underlying clinical state, in the sense that minor recording variation causes the top-ranked diagnosis to change at a non-trivial rate. That condition is satisfied by the data: a system whose top-ranked answer changes in 34–38% of cases following a single symptom swap or the loss of one clinical feature has not converged to a noise-robust function of the patient's clinical state, regardless of what a human physician would have done with the same input. The human-baseline comparison remains an important direction for future work and would sharpen the policy implications.

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 (Gilbert et al. (2024)). In these settings, LLMs are a useful synthesizer of information, a task that it can do well. Again, LLMs appear useful in those settings because they *can* be a full 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. Its inability to perform as a full force multiplier begs the question of what alternative approaches may exist.

 doi:10.20944/preprints202603.2228.v4

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 Elstein et al. (1978) 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 Kabir et al. (2024) demonstrates, and far more robust to the combinatorial explosion of multimorbidity.

Cognitive AI systems are designed to mimic principles from human cognitive science such as reasoning, memory, and symbolic representation and represent the most promising path forward (Bundy et al. (2023); Kotseruba and Tsotsos (2016)). Such systems must dynamically switch between hypothetico-deductive reasoning and pattern recognition, just as expert diagnosticians do (Elstein et al. (1978)). This mode of computation is fundamentally incompatible with the current LLM architecture.

References

Barnett, Karen, Stewart W Mercer, Michael Norbury, Graham Watt, Sally Wyke, and Bruce Guthrie. 2012. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *The Lancet 380*(9836), 37–43.

Bélisle-Pipon, J. C. et al. 2024. Why we need to be careful with LLMs in medicine. *Frontiers in Medicine 11*, 1495582.

Bell, Sigall K, Tom Delbanco, Joann G Elmore, Patricia S Fitzgerald, Alan Fossa, Kendall Harcourt, Suzanne G Leveille, Thomas H Payne, Rebecca A Stametz, Jan Walker, et al. 2020. Frequency and types of patient-reported errors in electronic health record ambulatory care notes. *JAMA network open 3*(6), e205867.

Bundy, A., N. Chater, and S. Muggleton. 2023. Introduction to cognitive artificial intelligence. *Philosophical Transactions of the Royal Society A 381*, 20220051.

Elstein, A. S., L. S. Shulman, and S. A. Sprafka. 1978. *Medical Problem Solving: An Analysis of Clinical Reasoning*. Cambridge, MA: Harvard University Press.

Espe, S. 2018. MalaCards: the human disease database. *Journal of the Medical Library Association: JMLA 106*(1), 140.

Feuerriegel, Stefan, Philipp Spitzer, Daniel Hendriks, Jan Rudolph, Sarah Schlaeger, Jens Ricke, Niklas Kühl, and Boj Hoppe. 2025. The effect of medical explanations from large language models on diagnostic accuracy in radiology.

Gilbert, S., J. N. Kather, and A. Hogan. 2024. Augmented non-hallucinating large language models as medical information curators. *npj Digital Medicine 7*, 100.

Goh, Ethan, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A Cool, Zahir Kanjee, Andrew S Parsons, Neera Ahuja, et al. 2024. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA network open 7*(10), e2440969.

Griot, Maxime, Christophe Hemptinne, Jean Vanderdonckt, et al. 2025. Large language models lack essential metacognition for reliable medical reasoning. *Nature Communications 16*, 642. https://doi.org/10.1038/s41467-024-55628-6.

Hernán, M. A. and J. M. Robins. 2020. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

Herper, Matthew. 2017. MD anderson benches IBM watson in setback for artificial intelligence in medicine. Forbes.

Jiang, Lavender Yao, Xujin Chris Liu, Nima Pour Nejatian, Mustafa Nasir-Moin, Duo Wang, Anas Abidin, Kevin Eaton, Howard Antony Riina, Ilya Laufer, Paawan Punjabi, et al. 2023. Health system-scale language models are all-purpose prediction engines. *Nature 619*(7969), 357–362.

Kabir, A., R. Kabir, and J. Nahar. 2024. 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*, 1–14.

Kabir, R., Michael Williams Nashif Rayhan. 2026. Cognitive ai-assisted primary care health delivery: A pilot study in bangladesh. *Working Paper*.

Kotseruba, I. and J. K. Tsotsos. 2016. A review of 40 years of cognitive architecture research. Preprint, arXiv:1610.08602.

 doi:10.20944/preprints202603.2228.v4

Li, Yihan, Xiyuan Fu, Ghanshyam Verma, Paul Buitelaar, and Mingming Liu. 2025. Mitigating hallucination in large language models (llms): An application-oriented survey on rag, reasoning, and agentic systems. *arXiv preprint arXiv:2510.24476*.

Lukac, Paul J, William Turner, Sitaram Vangala, Aaron T Chin, Joshua Khalili, Ya-Chen Tina Shih, Catherine Sarkisian, Eric M Cheng, and John N Mafi. 2025. Ambient ai scribes in clinical practice: a randomized trial. *NEJM AI 2*(12), AIoa2501000.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science 366*(6464), 447–453. https://doi.org/10.1126/science.aax2342.

O'Donovan, James, Madeleine Ballard, Rebecca Hope, Jane Wamae, Adriana Viola Miranda, Brian DeRenzi, Richard Kabanda, Nan Chen, Lennie Bazira, Mallika Raghavan, et al. 2026. Governance, scale, and integration: building community health worker systems ready for artificial intelligence. *The Lancet Primary Care*.

Qazi, Ihsan Ayyub, Ayesha Ali, Asad Ullah Khawaja, Muhammad Junaid Akhtar, Ali Zafar Sheikh, and Muhammad Hamad Alizai. 2026. Automation bias in large language model–assisted diagnostic reasoning among physicians trained in ai literacy—a randomized clinical trial. *NEJM AI 3*(5), AIoa2501001.

Ramaswamy, Ashwin, Aditya Tyagi, Hannah Hugo, et al. 2026. ChatGPT health performance in a structured test of triage recommendations. *Nature Medicine*. https://doi.org/10.1038/s41591-026-04297-7.

Ribeiro, Marco Tulio, Sameer Singh, Carlos Guestrin, Scott M Lundberg, and Su-In Lee. 2016. Why should i trust you?: explaining the predictions of any. *classifier. arXiv [cs. LG*.

Rosenbacke, R. et al. 2025. Beyond hallucinations: the illusion of understanding in large language models. Preprint, arXiv:2510.14665.

Singhal, Karan, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature medicine 31*(3), 943–950.

Sun, Y., D. Sheng, Z. Zhou, and Y. Wu. 2024. AI hallucination: towards a comprehensive classification of distorted information in AI-generated content. *Humanities and Social Science Communications 11*, 1278.

Thaler, S. 1995. Virtual input phenomena within the death of a simple pattern associator. *Neural Networks 8*(1), 55–56.

Toma, A., S. Senkaiahliyan, P. R. Lawler, B. Rubin, and B. Wang. 2023. Generative AI could revolutionize health care—but not if control is ceded to big tech. *Nature 624*, 36–38.

Topol, Eric J. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine 25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7.

U.S. Department of Health and Human Services, Office of Inspector General. 2014, May. Improper payments for evaluation and management services cost Medicare billions in 2010. Technical Report OEI-04-10-00181, U.S. Department of Health and Human Services, Office of Inspector General. Published May 28, 2014.

Weiskopf, Nicole G, David A Dorr, Christie Jackson, Harold P Lehmann, and Caroline A Thompson. 2023. Healthcare utilization is a collider: an introduction to collider bias in ehr data reuse. *Journal of the American Medical Informatics Association 30*(5), 971–977.

Xu, Z., S. Jain, and M. Kankanhalli. 2024. Hallucination is inevitable: an innate limitation of large language models. Preprint, arXiv:2401.11817.

Yang, Xi, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. 2022. A large language model for electronic health records. *NPJ digital medicine 5*(1), 194.

Zack, Travis, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, et al. 2024. Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health 6*(1), e12–e22.

Zech, John R, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine 15*(11), e1002683.

 doi:10.20944/preprints202603.2228.v4

disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.