Our study shows that LLMs can estimates diagnostic LRs with negligible mean bias and empirically-bounded precision, with the tightest agreement coming from the newest evaluated model, GPT-5 (95% limits of agreement of 0.26×–3.70×). The agreement varied by finding type, with labs showing less tight agreement than history, signs, and imaging. Models showed substantial qualitative category agreement (*κ* = *0.78)*. These results indicate that LLM-derived LRs might be useful to scaffold Bayesian reasoning in situations where empirical LRs are absent, inapplicable, or uncertain.

Prior work has shown that AI can complement humans in diagnostic reasoning, often offering complementary information to what humans input. Furthermore, it has long been known that the optimality of Bayesian updating requires properly specified estimates of new information, which exist only for a fraction of situations. Our report offers the first, to our knowledge, large evaluation of 700 finding-condition pairs across 30 conditions, which shows that LLMs have low bias and bounded dispersion across all categories of findings we evaluated. This suggests LLM’s may be usable as an education or workflow bridge, where LLMs could render probabilistic rendering explicit for trainees and supported quantified probabilistic reasoning in situations where empirical LRs are unavailable.

Our work extends prior studies of LLM clinical reasoning by resolving estimates to a quantitative unit (the likelihood ratio) rather than free-text advice. This approach could enable clearer auditing (explaining clearer auditing), teaching, and ultimately be combined with human or AI reasoning to improve threshold-based decision-making. Whether the observed likelihood ratio agreement is sufficient for clinical or pedagogical use cases likely depends on how the results are used: in where high‑stakes decisions hinge on a single finding the agreement may be insufficient, but the negligible mean bias on the logarithmic scale implies that serial Bayesian updating (via the multiplicative chain rule), should lead to unbiased posterior probabilities if independence holds – which machine learning (such as naïve Bayes algorithms) has often exploited for impressive performance even when the assumption isn’t strictly true.

Strengths of the current work include the reproducible evaluation on a breadth of comparators across and range of model generations. We evaluated both qualitative and quantitative descriptions of agreement, which both show moderate performance with the current approach. Interestingly, using classification schemes that have long been used in the field, we find that the vast majority of empirically studied diagnostic features are categorized as “negligible” or “weak” in strength, perhaps reflecting over-optimism in the diagnosticity of individual pieces of information from the classifications’ creators.

A few limitations warrant mention. We were unable to manually review and abstract the confidence-intervals and contexts associated with each LR referenced theNNT.com, so reference standards have some degree of uncertainty and inaccuracy and LLMs did not have accessed to all information that would be needed to identify precisely the spectrum of patients to which the NNT reported LR applies. We present only numerical, and not qualitative, assessments of the disagreements. Methodological rigor underlying the literature-sourced likelihood ratios from databases like theNNT.com were not independently assessed. Additionally, the reported LRs likely appear in the models’ pretraining data, which could improve the accuracy estimates compared to how well the models would perform on entirely unseen estimation problems. We observed no fact memorization (i.e. exact reproduction of text) and did not enable search capabilities for this reason. However, the presence of the LR reports in the data could probabilistically influence the learned weights in the model to a degree that is difficult to assess. Lastly, other than using general best principles for prompt and context engineering, we did not iteratively optimize our prompting strategy for perofmrnace on the evaluation set (because that would artificially improve the tightness of the accuracy), but this means that refined prompting strategies, or combining LLM estimation with reference material context or retrieval may improve performance to a degree that we did not assess.

LLM-derived likelihood ratios have several potential high-value uses. They could help address the “black-box” problem that currently hampers clinician trust in LLM-estimates, even when models are accurate, by providing inspectable, audible evidence weights. For example, LLM-LRs (and pre-test probability estimates, which are already shown to be more accurate than humans) could be used either as inputs to an explainable algorithm of updating (like formal Bayesian reasoning), or as a post-hoc explanative approach to traditional black-box LLM outputs. In diagnostic teaching, LLM-LRs might be deployed as part of automated scenario generation or reasoning evaluation for educational chatbots. Most clinicians reason through clinical gestalt, relying on intuition, heuritistics, and pattern recognition; but this pattern of thinking can be a barrier to communication to learners during teaching unless some of those inputs can be efficiently communicated, whether qualitatively or quantitatively. One traditional conception of diagnostic reasoning teaching is the cognitive apprenticeship model”, which requires teachers who are both master diagnosticians and educators to think aloud about how they reason with new pieces of data. However, excellent educators could perhaps useful teach reasoning in a much broader array of contexts if they had access to fit-for-purpose context-specific likelihood ratios. This provides pathway for AI-human hybridization, reflective practice, and fine-tuning.

Future work should evaluate how more precisely specifying the patient context and care setting of predictions affects the accuracy of LLM-estimated likelihood ratios. This would also allow for assessment of whether contextual factors bias model outputs in ways that parallel human reasoning, and whether such biases manifest differently in the decomposition of pre-test probabilities versus likelihood ratios. Comparative studies could also benchmark LLM agreement to reported values against human estimates across a range of diagnostic expertise. Another priority is assessing whether LLM-supply confidence in the estimated LRs is well calibrated, which could further build trust in the models. Ultimately, didactic and clinical outcomes should be assessed when these tools are embedded into real workflows, measuring their influence on diagnostic accuracy, clinician cognitive load, and patient outcomes. By advancing a systematic and quantitative approach to diagnostic reasoning, the integration of generative AI has the potential to improve calibration, mitigate bias, and move clinical practice toward a more evidence-driven discipline.

In conclusion, modern LLMs can approximate diagnostic likelihood ratios with low bias and bounded dispersion, thereby enabling practical, auditable and potentially Bayesian updating to be applied to clinical reasoning.

Legacy:

Our findings demonstrate that large language models (LLMs) can estimate likelihood ratios (LRs) for clinical diagnosis with reasonable accuracy, and that newer and more advanced models produce estimates more closely aligned with empirically reported literature standards. These results indicate significant potential for integrating generative AI into diagnostic and educational workflows, particularly in situations where empirical data is limited, outdated, or unavailable.

By serving as a bridge to Bayesian estimates that are “fit for purpose,” LLM-derived LRs can serve as a standardized method for supporting how probabilistic inferences are communicated and improved upon. For the clinician interested in optimizing their diagnostic accuracy, the quantification of diagnostic odds offers a reproducible pathway for AI-human hybridization, reflective practice, and fine-tuning of pre-calibrated action thresholds. For the frustrated trainee who is stumped by an onslaught of ambiguous features in a clinical unknown, LLM-derived LRs can provide validation of clinical uncertainty that is, in fact, irreducible. For the master diagnostician who wishes to de-mystify their clinical gestalt for medical trainees, LLM-derived LRs offer a pathway for more explicit and transparent synthesis of prioritized (or de-prioritized) data inputs.

The traditional approach by which trainees learn the “art” of diagnostic reasoning is called the “cognitive apprenticeship” model.22 A pitfall of this approach is how much it hinges on well-trained faculty who boast the skills of *both* a master diagnosticianand an educator. Clinician educators also have to “think out loud” to make their uncertainty tolerance and train of thought transparent. After all, the ability to consistently make an accurate diagnosis does not help a team of learners if these nuanced skills are ineffectively taught or poorly communicated. For trainees who may not share the same priors (e.g. past clinical experiences or baseline assumptions), passing on an “embrace of uncertainty” can be especially challenging. Likewise, the educator who models their availability bias, recency bias, or anchoring bias only amplifies these diagnostic pitfalls, further perpetuating habits of diagnostic inertia, rooted in false assumptions.

LLM-derived LRs can transcend these limitations by making the chain of probabilistic inferences and belief updates more accessible to learners and unskilled faculty alike. AI hybridization offloads the cognitive burden associated with complex mathematical formulas, allowing clinicians to more easily engage in structured Bayesian reasoning. Such a shift benefits clinicians across all stages of training, from early learners developing foundational diagnostic skills to experienced practitioners refining their diagnostic accuracy and consistency. By lowering the point of entry for more routine application of Bayes theorem, LLM-derived LRs can democratize the upskilling of probabilistic inference amongst clinicians across the board. Ultimately, when the cognitive apprenticeship is strengthened, both trainees and patients benefit.

Moreover, coupling generative AI capabilities with databases such as the Number Needed to Treat (NNT) database could create a "living" repository ofLRs, a dynamic, continuously updated resource that responds to evolving clinical evidence and real-time clinician feedback. This approach not only facilitates immediate clinical reasoning improvements but also supports long-term skill development in probabilistic reasoning through deliberate, repeated practice and exposure. Just as musicians progressively internalize and master complex scales through systematic practice, clinicians could similarly internalize a robust, hybridized approach to Bayesian inference through iterative use of AI-supported diagnostic tools. Put simply, LLM-generated LRs provide a path towards AI-enhanced adaptive practice.23

Nevertheless, it remains crucial to acknowledge several limitations. First, reference standard likelihood ratios must be taken from the literature, and therefore could potentially be included in the training data. Though LLMs generally do a poor job memorizing information24, this may lead to LRestimated being closer to empirical estimates (LRreported) than if the LLM were estimating a hypothetical, unquantified LR. The gradient of improved performance with increasing model complexity (GPT-4o < o3 < GPT-5) further argues against simple memorization of LRs in our work.

Second, the accuracy and methodological rigor underlying the literature-sourced likelihood ratios from databases like theNNT.com were not independently assessed in our study, introducing an unknown potential for bias in the reference standards. Furthermore, as diagnostic test accuracy depends on the spectrum of patients evaluated11 and we could not extract the population of interest from studies that theNNT.com estimates were based on, the LLM was implicitly estimating the population to which the test would be implied. It’s possible that agreement would be higher if the population of interest were more closely matched to the diagnostic test accuracy studies on which the LRreported is based. The width of the limits of agreement, particularly if near-certainty (ie. a high LoA) is required, suggest they must be thoughtfully integrated into systems with direct human oversight. Lastly, our study did not utilize LLMs explicitly integrated with real-time search capabilities, a factor that could further improve the validity and utility of the generated estimates in clinical contexts, though it would make validation of performance substantially more challenging.

Future work should explore the integration of LLM-generated LRs with real-time clinical literature retrieval systems, assessing their direct impact on diagnostic accuracy, clinician cognitive load, and ultimately, patient outcomes. By fostering a systematic, quantitative approach to diagnostic reasoning, the integration of generative AI could substantially enhance diagnostic accuracy, reduce cognitive biases, and advance the practice of clinical diagnosis towards a more evidence-driven discipline.

# Conclusion:

Large language models show considerable promise in estimating diagnostic likelihood ratios, especially where empirical clinical data are sparse or unavailable. Future research should explore real-time integration with updated clinical literature and investigate the direct impact of LLM-augmented clinical reasoning on patient outcomes.

Effective diagnostic reasoning hinges on accurately interpreting clinical findings (patient history, symptoms, examination and test results) to refine disease probability estimates. Ideally, this process is guided by likelihood ratios (LRs), which quantify how strongly particular findings influence the odds of associated diseases1,4. However, empirically derived LRs exist only for a limited subset of clinical findings, conditions, and contexts, because estimating them requires collation of resource-intensive diagnostic test accuracy studies across variable contexts1,2,25,26.

The traditional approach by which clinicians reason is through clinical gestalt, relying on intuition, heuristics, and pattern recognition7,27. While efficient, gestalt-based reasoning can be biased, inconsistent, and limited by an insufficient scope of personal experiences. When feasible, quantitative reasoning using likelihood ratios (LRs) provides a normative standard that improves diagnostic accuracy, consistency, and can be used to refine clinical gestalt6. This hybrid system in which humans and AI complement one another in classification tasks, like clinical diagnosis, is not a new idea28. In fact, Bayesian modeling shows promise for more generalized classification problems because of the diverse data processing strategies by which human and AI classifiers operate. Whether large language models (LLMs) can generate LRs to specifically improve upon human diagnostic accuracy within a Bayesian inferential reasoning framework is not yet known.

Recent advances in *generative* artificial intelligence, particularly LLMs, offer new opportunities to enhance clinical decision-making and medical education18. Unlike traditional machine learning approaches, which require task-specific training data, large language models show emergent abilities, referring to their ability to perform tasks not in the training set (either with no examples, zero-shot generalization, or with prompted examples, in-context learning)15.

This capability raises the possibility that LLMs could reliably estimate diagnostic LRs, potentially overcoming a key barrier to broader application of quantitative reasoning in clinical practice. Notably, the accuracy of LLM-estimated LRs has not been previously explored. In this study, we aimed to evaluate the capacity of contemporary LLMs to accurately estimate diagnostic LRs. Specifically, we compared LLM-generated LRs with empirically reported values from the existing literature. If the accuracy of these models in estimating known LRs was found to be acceptable, the natural next step would be to consider situations in which under-investigated clinical findings (e.g. unknown LRs) could be tested in simulated (or real) clinical and training contexts.

Can key clinical features be “digitalized” within a Bayesian analytical reasoning framework to augment our current approach to clinical diagnosis? How will this affect medical education and skill development of learners? We explore these questions after testing current LLM capabilities. Our work adds to the work of Goh et al16 and others29, who found that the reasoning of LLMs is of merit and deserves further exploration with respect to physician-LLM collaboration in clinical decision-making. By incorporating cutting-edge models, we address limitations of past works13 and examine the interface of such models with clinician decision-making.

Other text brainstorming:

Medical diagnosis requires integrating history, examination, and test findings to identify the condition that best explains a patient’s presentation. Bayesian reasoning has long been promoted as the normative framework for this task because of its simplicity, information efficiency, and broad applicability. Instruction in Bayesian methods can improve trainees’ diagnostic reasoning, yet its influence on medical education and daily practice remains modest. Most clinicians and educators continue to rely primarily on intuition, heuristics, and pattern recognition.

An important barrier to broader use of Bayesian reasoning at the bedside and in medical education is the lack of accurate, context-specific diagnostic likelihood ratios. Likelihood ratios quantify how the presence or absence of a finding (historical element, symptom, examination finding, or test result) changes the odds of disease. However, diagnostic accuracy studies are difficult to perform and interpret, and true likelihood ratios often vary substantially by clinical context. As a result, reliable likelihood ratios for findings in many common situations remain unknown.

Large language models (LLMs) are neural network–based systems trained on large text corpora that encode clinical knowledge and concept associations. Unlike traditional machine-learning approaches that require task-specific training data, LLMs can sometimes generalize to novel tasks without prior examples via zero-shot or in-context learning. Recent work suggests that LLMs can assist or parallel physician reasoning with suprising accuracy, but adoption remains limited by persistent errors, opacity, and low trust among clinicians and patients.

One way to address this may be to task LLMs with performing individual, inspectable components of diagnostic reasoning process. In particular, LLMs may be able to estimate diagnostic likelihood ratios by leveraging learned implicit associations, even when empirical data are scarce or non-existant. Accurate LR estimation could enable more widespread and effective use of Bayesian reasoning in both medical education and clinical practice. To test this possibility, we assess the accuracy three LLMs of increasing complexity against published likelihood ratios drawn from the literature.

Problem:

Medical diagnosis requires integrating history, examination, and test findings to identify the condition that best explains a patient’s presentation. Bayesian reasoning has long been promoted as the normative framework for this task, due to its simplicity, information efficiency, and broad applicability. Instruction in Bayesian methods has been shown to improve clinical trainees’ reasoning performance, yet its influence on everyday education and practice remains modest. Most clinicians continue to rely primarily on intuition, heuristics, and pattern recognition when reasoning, and teaching others how to reason, about diagnoses.

Gap:

A key barrier to more widespread application of Bayesian reasoning to bedside medicine and medical education is the lack of accurate and context-specific diagnostic likelihood ratios. In Bayesian reasoning, the likelihood ratios represent how much the presence or absence of a finding (historical element, symptom, examination finding, or test result) changes the odds of disease. Yet diagnostic accuracy studies to estimate likelihood ratios are difficult to perform and interpret, and the true likelihood ratio of a finding often varies substantially by circumstance. As a result, reliable likelihood ratios for many common clinical findings remain unknown. In these situations, Bayesian reasoning cannot be applied at all, or is less effective than it could be.

Hook (Solution):

Large language models (LLMs) are neural network–based algorithms trained on large text corpora that allow them to encode clinical knowledge and concept associations. Unlike traditional machine learning approaches, which require task-specific training data, LLMs can sometimes generalize to novel tasks without prior examples, using zero-shot or in-context learning. Recent work suggests that LLMs can assist or parallel physician reasoning with surprising accuracy, but their adoption remains limited by persistent errors, lack of transparency, and low trust among clinicians and patients.

One potential way to increase LLMs utility for diagnostic reasoning might be to have them perform individual components of the diagnostic reasoning process that can be used by humans, inspected, or integrated into other ocmponents of an explainable system. LLMs might be able to estimate diagnostic likelihood ratios by leveraging implicit associations, even when explicit data are lacking. Accurate LR estimation could enable more widespread and effective use of Bayesian reasoning in both medical education and clinical practice. To explore this possibility, we evaluate three LLMs against published likelihood ratios drawn from the literature.

Large language models (LLMs) are neural-network based machine learning algorithms that can encode vast stores of information, including clinical facts and their quantitative relationships, as learned from large corpuses of texts such as the medical literature. [ Lancet ] Unlike traditional machine learning approaches, which require task-specific training data, large language models are often able to perform tasks not in the training set (either with no examples, zero-shot generalization, or with prompted examples, in-context learning). Recent work has demonstrated the potential for LLMs to reason autonomously or in parellel to physicians, with suprising accuracy. Yet, the applications of such systems is limited due to persistent errors, their black-box nature, and a lack of trust from clinicians and patients. However, if LLMs are able to accurately perform isolated component tasks of the diagnostic reasoning process, it might better enable computer-human interaction in reasoning in teaching.

It’s possible that LLMs might estimate diagnostic likelihood ratios, by drawing on indirect or inferred learned associations, even without explicit examples or literatures, thereby enabling more widespread and effective application of Bayesian reasoning by computers, humans, or both. Accordingly, we evaluate the accuracy of 3 large language models in estimating reported likelihood ratios from the literature

———

LLMs

Recent advances in *generative* artificial intelligence, particularly large language models (LLMs), offer new opportunities to enhance clinical decision-making and medical education10. Unlike traditional machine learning approaches, which require task-specific training data, large language models show emergent abilities, referring to their ability to perform tasks not in the training set (either with no examples, zero-shot generalization, or with prompted examples, in-context learning)11.

Summary:

Early evidence suggests that large language models (LLMs) can function as vast knowledge bases, encoding clinical facts and quantitative relationships from medical literature

the hope is that by combining traditional apprenticeship (experts demonstrating diagnostic reasoning) with accessible resources (compilations of LRs, or AI that can supply them on the fly), we can better train clinicians who are adept at both intuitive and analytic reasoning. LRs essentially provide a quantitative backbone to the *why* behind a diagnosis, and education that leverages that may produce more thoughtful, less error-prone diagnosticians

There are also statistical threats to validity: LRs from literature themselves have confidence intervals and may be biased (publication bias, etc.). An AI that was trained on published studies might inherently reflect those biases.

Several ancillary findings are interesting:

-The **majority of LRs were between 0.5 and 2.0** (interquartile range 0.7–2.2) – meaning most findings have only negligible diagnostic impact . Very large LRs (>10) or very small (<0.1) were comparatively rare. This empirical sparsity suggests that clinicians searching literature for robust “rule-in” or “rule-out” clues will often come up empty-handed

-Teaching diagnostic reasoning has long balanced **analytic (Bayesian) approaches** with intuitive pattern recognition

-One strategy is to have learners **cross-check AI-provided LRs or suggestions** against their own reasoning, turning it into a teachable moment about evidence quality. AI can even serve as a “tutor” itself – for example, a student could ask an LLM to explain how a particular finding influences disease likelihood, and then verify that explanation

assumption is that any single LR is *portable* to all clinical contexts (the “transportability” of diagnostic evidence). In reality, context shifts (different patient populations, disease spectrum, prior test workup) can invalidate an LR – the earlier-discussed spectrum effect