# Title:

Estimation of Medical Diagnostic Likelihood Ratios Using Artificial Intelligence

# Authors:

Shuhan He

Paul Chong

Kian

Brian Locke

Cory

Boyu

Emma Chua

Adriana Coleska

Ahmad

# Abstract:

[ ] do last

# Introduction:

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 disease. However, empirically derived LRs exist only for a limited subset of clinical findings, conditions, and contexts because estimating them requires difficult and resource-intensive diagnostic test accuracy studies.

Clinicians often reason by clinical gestalt, relying on intuition and experience. While efficient, gestalt-based reasoning can be biased, inconsistent, and limited by personal experience. When feasible, quantitative reasoning using likelihood ratios (LRs) provides a normative standard that improves diagnostic accuracy and consistency, and can be used to refine clinical gestalt.

Recent advances in artificial intelligence, particularly large language models (LLMs), offer new opportunities to enhance clinical decision-making and medical education. 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).

This capability raises the possibility that LLMs could reliably estimate diagnostic likelihood ratios, potentially overcoming a key barrier to broader application of quantitative reasoning in clinical practice and education. However, the accuracy of LLM-estimated has not been previously explored. In this study, we aimed to evaluate the capacity of contemporary large language models to accurately estimate diagnostic likelihood ratios. Specifically, we compared LLM-generated likelihood ratios with empirically reported values from existing literature. Understanding the accuracy and applicability of these models in estimating likelihood ratios could help clinicians leverage under-investigated clinical findings and potentially integrate robust quantitative reasoning into everyday diagnostic practice and medical education.

**Methods:**

We conducted a comparative study assessing the agreement between diagnostic LRs generated by LLMs and empirically derived LRs reported by theNNT.com (© The NNT Group, 2010–2022). This study utilized publicly available data and did not involve human subjects, thus exempting it from institutional review board oversight.

*Reference Standard Likelihood Ratios*

On April 1, 2025, we compiled a reference-standard dataset of likelihood ratios (LRReported) from theNNT.com a resource aggregating diagnostic likelihood ratios from published medical literature to assist with diagnostic reasoning. All positive and negative LRs from all conditions listed on theNNT.com were included. For LRs theNNT reported with a point estimate, we recorded the provided estimate directly (e.g. 1.5, 95% CI 1 - 2 was coded as 1.5). When only a range was presented, the geometric mean of the reported range was utilized. LRs were initially extracted using an automated script and then manually validated by a single reviewer (PC). Each LR was categorized as an imaging finding, a patient historical element, a sign/symptom, a test score, and/or a test finding. Findings were also categorized by the relevant specialty according to theNNTs listings. We qualitatively describe the strength of findings as strong (LR- ≤ 0.10 or LR+ ≥ 10), moderate ( 0.1 < LR- ≤ 0.2 or  5 ≤ LR+ < 10), weak (0.2 ≤ LR- < 0.5 or 2 ≤ LR+ < 5), or negligible (0.5 < LR < 2).

*Comparator Likelihood Ratios:*

On April 1, 2025, we generated comparator likelihood ratios (LRLLM) for all findings listed on theNNT using three OpenAI LLMs (OpenAI, LP; San Francisco, California, USA): 4o-mini (model release July 18, 2024), 4o (release August 6, 2024), and 3o-mini (release January 31, 2025) to represent a range of inference costs. We applied a few-shot prompt (full text in Supplementary Information) that:

* Instructions the model to take the persona of an expert in medical diagnosis;
* defined likelihood ratios;
* specified queries in the form ‘for target condition X, estimate the LR of finding ‘Y’;
* specified that the output should be a single positive number output in JSON format;
* instructions for reasoning (consider the condition of interest, the population of interest, what the presence or absence of the feature would imply about disease likelihood)
* Provided three hypothetical examples (not examples taken from the literature).

The prompt included no information or hints about reference standard values (LRReported). The prompt was not iteratively refined to improve agreement with reference values.

*Statistical Analysis*

We assessed agreement between LRReported and LRLLM using Bland-Altman analysis on log-transformed LRs, as the strength of evidence represented by LRs is linear in the logarithmic scale.B We calculated 75%, 90%, 95% ratio limits of agreement (i.e multiplicative limits of agreement where LRs are within x-fold bounds of each that proportion of time) were calculated. Between-model differences in agreement were tested using paired t-test for mean differences and the Morgan-Pittman test for differences in limits of agreement (variances). Subgroups analyzed included information type (history, sign, exam finding, or test result), and positive vs. negative LR using unpaired (Welch’s) t-test for mean bias and Levene’s test for the width of limits of agreement. An alpha of 0.05 with no adjustment for multiple testing is used.Analyses were performed using Python 3.12.7 and Microsoft Excel.C

# Results:

700 LRReported exploring the 30 available medical conditions were compiled from theNNT.com. Signs/symptoms were the most common type of LR (56%, n=403), followed by historical element (16% , n=112) and test result (13% , n=93) (Table 1). LR’s ranged from 0.01 to 145.9, with a median of 1 (interquartile range 0.7 to 2.2). Figure 1 shows the distribution of strength of evidence in the LRReported. Most findings were of negligible strength (0.5 < LR < 2; 56.5%), with the next most common being weak evidence in favor of a diagnosis (17.4%), weak evidence against (8.7%) and moderate evidence for (7.5%).

**Figure 1:** Distribution of Likelihood ratios, as reported in the literature. Background shading represents strong, moderate, weak, and negligle evidence.

**A blue line graph on a pink background

AI-generated content may be incorrect.**

**Table 1.** Summary of medical diagnostic subgroups and sample sizes.

|  |  |
| --- | --- |
| **Subgroup** | **n** |
| Imaging finding | 56 |
| History | 112 |
| Sign/symptom | 403 |
| Score | 25 |
| Test finding | 93 |
| History and test | 9 |
| History and imaging | 2 |
| Diagnosis | 6 |

Three sets of LRLLM were generated (for 4o-mini, 4o, and o3-mini) for each of the 700 LRReported. Figure 2 shows the 95% x-fold coverage limits, indicating the x-fold range in which 95% of estimates would be expected to be from a reported value. 4o-mini was the most biased (*P* < .001 vs 4o and o3-mini) and had the lowest agreement (*P* < .001 vs 4o and o3-mini) with values reported in the literature. We did not find evidence that O3-mini showed less mean bias compared to 4o (1.04x vs 1.08x; *P* = .09), but it did show better agreement (*P* < .001).

Table 2 shows the 50%-, 75%-, 90%-, 95%-, and 99%- coverage ratio limits of agreements for the estimates using Bland-Altman analysis, along with their confidence intervals which summarize the uncertainty in those estimates.

Two subgroups were evaluated. First, analyzing the agreement by the type of finding showed that all LLMs were most accurate and leased biased at estimating the strength of evidence associated with historical findings; they struggled to estimate the importance of test-results, comparatively.

Second, we analyzed the agreement by supportive (LRreported > 1) vs. non-supportive (LRreported < 1) evidence. 4o-mini was more biased estimating supportive evidence then non-supportive evidence (1.73x vs 1.16x, *P* < .001), 4o did not show a difference (1.04x vs 1.13x, *P* = .24), and o3-mini was less biased estimating supportive evidence (0.87x vs 1.26x, *P* < .001). There was no statistically significant difference in the width of the limits of agreement between positive and negative findings for any of the models (4o-mini *P* = .20, 4o *P* = .60, o3-mini *P* = .22).

**Figure 2: Agreement between literature-reported and LLM-generated likelihood ratios**

A diagram of a number of different numbers

AI-generated content may be incorrect.

Each panel shows the agreement between reported and model-generated likelihood ratios (LRs). The y-axis shows the log ratio (reported/model), and the x-axis shows the geometric mean of paired LRs. Solid black lines represent mean bias; dashed lines represent 95% x-fold limits of agreement (LoA). Narrower coverage intervals represent closer agreement, and deviations of the mean line from unity indicate systematic bias. Shaded areas indicate the confidence intervals on each bound of agreement.

**Table 2: Multiplicative coverage intervals for LLM-generated likelihood ratios**

The table shows intervals that bound 50%, 75%, 90%, 95%, and 99% of model-generated likelihood ratios relative to literature-reported values. Intervals are expressed as multiplicative factors (“×”), indicating how far each model can be expected to deviate reported LRs. Parentheses show 95% confidence intervals for the estimate of each coverage limit.

**Model: 4o-mini**

|  |  |  |
| --- | --- | --- |
| Coverage Interval | Lower edge (95% CI) | Upper edge (95% CI) |
| 50% | 0.71x (0.65x - 0.77x) | 2.88 (2.64x - 3.14x) |
| 75% | 0.43x (0.39x - 0.47x) | 4.73 (4.28x - 5.23x) |
| 90% | 0.26x (0.23x - 0.29x) | 7.93 (7.04x - 8.93x) |
| 95% | 0.18x (0.16x - 0.21x) | 11.02 (9.65x - 12.58x) |
| 99% | 0.10x (0.08x - 0.11x) | 20.95 (17.83x - 24.61x) |

**Model: 4o**

|  |  |  |
| --- | --- | --- |
| Coverage Interval | Lower edge (95% CI) | Upper edge (95% CI) |
| 50% | 0.58x (0.54x - 0.63x) | 2.01 (1.86x - 2.17x) |
| 75% | 0.38x (0.34x - 0.41x) | 3.11 (2.85x - 3.40x) |
| 90% | 0.24x (0.21x - 0.26x) | 4.91 (4.42x - 5.45x) |
| 95% | 0.18x (0.16x - 0.20x) | 6.55 (5.83x - 7.37x) |
| 99% | 0.10x (0.09x - 0.12x) | 11.55 (10.02x - 13.31x) |

**Model: o3-mini**

|  |  |  |
| --- | --- | --- |
| Coverage Interval | Lower edge (95% CI) | Upper edge (95% CI) |
| 50% | 0.60x (0.57x - 0.65x) | 1.79 (1.67x - 1.91x) |
| 75% | 0.41x (0.38x - 0.45x) | 2.62 (2.43x - 2.83x) |
| 90% | 0.28x (0.25x - 0.30x) | 3.90 (3.56x - 4.27x) |
| 95% | 0.22x (0.19x - 0.24x) | 5.03 (4.54x - 5.57x) |
| 99% | 0.13x (0.12x - 0.15x) | 8.25 (7.28x - 9.33x) |

**Figure 3a and 3b: Agreement between LLM-estimated and literature-reported likelihood ratios**

(a) By type of clinical finding. Rows represent clinical finding categories (laboratory, imaging, history, sign-symptom, and diagnosis); columns represent LLM versions (GPT-4o-mini, GPT-4o, and o3-mini). Categories assigned by manual review.

A screenshot of a graph

AI-generated content may be incorrect.

(b) By evidence direction. Rows separate supportive (LRreported > 1) from non-supportive findings (LRreported < 1).

A group of graphs showing different colored lines

AI-generated content may be incorrect.

# Discussion:

Our findings demonstrate that large language models (LLMs) can estimate likelihood ratios 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 clinical diagnostic workflows, particularly in situations where empirical data is limited, outdated, or entirely unavailable. LLM-derived likelihood ratios might also provide a method for generating usable estimates for supporting the quantitative rigor and transparency of diagnostic reasoning instruction, thereby providing a mechanism for systematic improvement of clinical intuition. However, 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 human oversight, rather than trusted outright.

Traditionally, medical decision-making has heavily depended on gestalt intuition, a holistic and heuristic approach to estimating disease probability based on clinicians’ prior experiences and assumptions. Although intuitive reasoning remains vital in clinical practice, it is inherently susceptible to a variety of cognitive biases such as availability bias, anchoring, and premature closure, ultimately affecting diagnostic accuracy and patient outcomes. The adoption of explicit likelihood ratio frameworks and Bayesian reasoning in clinical practice remains limited primarily due to the limited availability of quality diagnostic test accuracy data and the cognitive load associated with using Bayes theorem.

By contrast, integrating generative AI into clinical reasoning can transform diagnostic decision-making into a more explicit, reproducible, and rigorous practice. Leveraging LLM-generated likelihood ratios could offload cognitive burdens associated with complex probabilistic calculations, allowing clinicians to more easily engage in structured Bayesian reasoning. Such a shift could particularly benefit clinicians across all training stages, from early learners developing foundational diagnostic skills to experienced practitioners refining their diagnostic accuracy and consistency.

Moreover, coupling generative AI capabilities with databases such as the Number Needed to Treat (NNT) database could create a "living" repository of likelihood ratios, 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 robust Bayesian inference skills through iterative use of AI-supported diagnostic tools.

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

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 evaluated and we could not extract the population of interest from studies that theNNT.com estimates were based on, the LLM was implicitly additionally 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. 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 likelihood ratios 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 clinical medicine 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.

# References