# Title:

Prediction of Medical Diagnostic Likelihood Ratios Using Artificial Intelligence

# Authors:

Shuhan He

Paul Chong

Kian

Brian Locke

Cory

Boyu

Emma Chua

Adriana Coleska

Ahmad

# 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 usually reason by clinical gestalt based on intuition and experience. Reasoning by clinical gesstalt is efficient, but can be limited by limited case exposure, cognitive biases, and variability. In contrast, quantitative reasoning based on explicit likelihood ratios offers reproducible standards and benchmarks that can significantly reduce diagnostic uncertainty and cognitive biases.

Recent advances in artificial intelligence, particularly the advent and refinement of large language models (LLMs), have opened new avenues for enhancing clinical decision-making. These models exhibit remarkable generalization capabilities across a variety of reasoning tasks, such as information synthesis, question-answering, and context-driven prediction. Yet, the potential of LLMs to quantitatively estimate likelihood ratios for clinical findings (thereby providing systematic support for diagnostic reasoning) has not yet been rigorously 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 standards 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.

**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.

*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 threehypothetical examples.

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 95% ratio limits of agreement (i.e multiplicative limits of agreement where LRs are within x-fold bounds of each other) 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 specialty, information type (history, sign, exam finding, or test result), and positive vs. negative LR using unpaired (Welch’s) t-test and classical F-testing. An alpha of 0.05 with no adjustment for multiple testing is used.Analyses were performed using Python 3.12.7 (Pinguin package 0.5.5) and Microsoft Excel.C

# Results:

700 LRReporteds 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). [ ] Description of the specialties and the distribution of LRs?

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

Three sets of LRLLM were generated for each of the 700 LRReported.For 4o-mini, LRLLM

were on average 1.35-fold higher, with 95% limits of agreement of 0.59-fold to 2.4-fold (Figure 1). LRLLM generated by 4o had a mean bias of 1.08-fold higher than LRReported, with 95% limits of agreement from 0.59-fold to 1.9-fold (Figure 2). O3-mini achieved a mean bias of 1.04-fold higher with limits of agreement from 0.67-fold to 1.6-fold (Figure 3).

**Table 2:**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Figure 1a.** Logged Bland-Altman analysis of 4o-mini LR predictions compared to reference standard.

A graph with dots and numbers

AI-generated content may be incorrect.

**Figure 1b.** Logged Bland-Altman analysis of 4o LR predictions compared to reference standard.

A graph with dots and numbers

AI-generated content may be incorrect.

**Figure 1c.** Logged Bland-Altman analysis of 3o-mini LR predictions compared to reference standard.

A graph with dots and numbers

AI-generated content may be incorrect.

Subgroup logged Bland-Altman analysis was also performed for the o3-mini LLM to evaluate the performance of the model based on specific types of information. Estimates of the LRs of historical findings were most accurate and precise (mean bias 0.99-fold, 95% agreement 0.71-fold to 1.3-fold), followed by imaging findings and signs/symptoms.; see Table 1 and Figure 4.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

**Figure 4.** Logged Bland-Altman analyses of 3o-mini LR predictions compared to reference standard based on subgroup of medical diagnostic.

A group of blue dots

AI-generated content may be incorrect.

LLMs showed statistical significance between logged differences of predicted LRs from reference standard (f-ratio value of 4.71, p-value of .0028), see Figure 5 below:

**Figure 5.** Logged differences between predicted LRs of LLMs and reference standard.

A graph of different differences

AI-generated content may be incorrect.

There was no statistically significance in the raw differences between the LR predictions of LLMs and reference standard (f-ratio value of 2.60, p-value of .051l; see Figure 6):

**Figure 6.** Raw differences between predicted LRs of LLMs and reference standard.

A graph of different differences

AI-generated content may be incorrect.

There was no statistical significance in the LRReported and LRLLM predictions among the imaging finding, history, score, test finding, history and test, history and imaging, and diagnosis subgroups. The only subgroup with statistically significant differences in LRReported and LRLLM was signs/symptoms (f-ratio value of 4.18, p-value of .0059), see Figures 7 and 8:

**Figure 7.** Logged differences between predicted LRs of LLMs and reference standard among signs and symptoms.

A graph showing different differences

AI-generated content may be incorrect.

**Figure 8.** Raw differences between predicted LRs of LLMs and reference standard among signs and symptoms.

A graph of different differences

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 can provide a foundational resource for enhancing the quantitative rigor and transparency of diagnostic reasoning, thus reducing reliance on less systematic clinical intuition.

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 cognitive load and complexity involved.

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, the potential inclusion of medical literature in the training data for LLMs may inadvertently enhance the accuracy of estimated likelihood ratios, especially those already documented in established sources. 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. 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.

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