**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 information (such as historical details, patient symptoms, examination findings, and test results) to adjust the probability of a disease being present. This interpretation is ideally guided by likelihood ratios (LRs), statistical measures that quantify how strongly a particular finding influences the probability of disease by comparing the odds of disease before and after observing that finding. Despite their theoretical utility, empirically derived likelihood ratios are challenging to obtain. Diagnostic accuracy studies, the gold standard for generating reliable LRs, require extensive resources, meticulous study designs, and substantial time investments. Consequently, reliable, empirically validated LRs exist only for a limited subset of clinical findings, conditions, and scenarios.

Due to these practical constraints, clinicians frequently rely upon intuitive reasoning (often described as clinical gestalt) when interpreting clinical evidence. Although such intuition is invaluable, it inherently exposes clinical reasoning to biases and variability, reducing diagnostic accuracy and consistency. 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:**

*Reference Standard Likelihood Ratios*

On April 1, 2025, a reference-standard list of likelihood ratios (LRReported) was compiled from the theNNT.com (© The NNT Group, 2010 – 2022),A a service that compiles likelihood ratios reported in the literature to assist with diagnostic reasoning. All likelihood ratios - whether positive or negative - from all conditions listed on theNNT.com were included. When the NNT presented a point estimate of the likelihood ratio, the point estimate was used (e.g. 1.5, 95% CI 1 - 2 was coded as 1.5). If only a range was presented, the geometric mean of the range was utilized. All scraped entries were manually confirmed for accuracy and designated as a sub-category of medical diagnostic: imaging finding, history, sign/symptom, test score, and/or test finding.

*Comparator Likelihood Ratios:*

Comparator likelihood ratios (LRLLM) for each of the findings reported on theNNT were generated using the following large language models (LLMs) from OpenAI on April 1, 2025: 4o-mini (July 18, 2024), 4o (August 6, 2024), and 3o-mini (January 31, 2025) to represent a range of possible inference costs. A custom prompt (full prompt reported in Supplementary Information) was used following the format: Instructions to take the persona of an expert in medical diagnosis, an explanation of what a likelihood ratio is, specification that requests would take the form of ‘for target condition X, estimate the LR of finding ‘Y’, specification that the output should be a positive number output as a JSON, instructions for reasoning (consider the condition of interest, the population of interest, what the presence or absence of the feature would predict, and some hypothetical examples. The LLM was not provided information or hints about the value of the LRReported.

*Statistical Analysis*

As the strength of evidence represented by likelihood ratios is linear in the logarithmic scale,B LRReported and LRLLM were log transformed. Bland-Altman analysis on the log transformed LRs was used to visualize and calculate 95% limits of agreement. following subgroups: specialty (using theNNT’s categorization), type of information (history, sign, exam finding, or test result), and positive vs. negative LR. Differences between LRReported and LRLLM agreement among models, as well as subgroups, was tested using Student’s T-testing and one-way ANOVA testing.

*Regulatory*

This research does not involve human subjects and thus is exempt from IRB review. Analyses were performed using Python 3.12.7 and Microsoft Excel.C

**Results:**

700 LRs exploring the 30 available medical conditions from theNNT.com were compared with predictions from the LLMs. Logged Bland-Altman analyses were performed for the respective LLMs and the reference standard (see Figures 1, 2, and 3). The multiplicative mean bias between human and each model are as follows:

* 4o-mini: 0.35x with 95% limits of agreement from -1.7x to +2.4x
* 4o mean: 0.08x with 95% limits of agreement from -1.7x to +1.9
* o3-mini: 0.04x with 95% limits of agreement from -1.5x to +1.6x

**Figure 1.** 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 2.** 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 3.** 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 3o-mini LLM to evaluate the performance of the model based on specific types of medical diagnostic tools; see Table 1 and Figure 4.

**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 |

**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.