# Title:

Estimation of Medical Diagnostic Likelihood Ratios Using Artificial Intelligence

# Authors:

Paul Chong, MD, ORCID: \*\*\*

Shuhan He, MD, ORCID: \*\*\*

Kian Samadian, MD, ORCID: 0009-0008-5289-9325

Amal Mohamed, MBBCh: 0009-0006-0281-8625

Boyu Peng MSc: 0009-0002-9470-8714

Emma Chua: 0009-0000-4188-3864

Cory Rohlfsen, MD, ORCID: 0009-0001-4229-6802

Brian W. Locke, MD MSc – ORCID [0000-0002-3588-5238](https://orcid.org/0000-0002-3588-5238)

Corresponding Author: Brian Locke, MD MSc. Assistant Professor of Research, Shock Trauma Intensive Care Unit, Intermountain Medical Center. 5121 Cottonwood St, Murray, UT 84107. [brian.locke@imail.org](mailto:brian.locke@imail.org)

Affiliations:

Paul Chong: ? *School of Osteopathic Medicine, Campbell University, Lillington, North Carolina*

Shuhan He: Departments of Emergency Medicine and Internal Medicine, Massachusetts General Hospital, Harvard Medical School; Boston, MA

Kian Samadian: Department of Emergency Medicine, Massachusetts General Hospital, Harvard Medical School; Boston, MA

Amal Mohamed: Department of Emergency Medicine, Massachusetts General Hospital, Harvard Medical School; Boston, MA

Boyu: MSDA Program, MGH Institute of Health Professionals, Boston, MA

Emma Chua: Department of Natural Sciences, Pasadena City College, Pasadena, California

Cory Rohlfsen: Department of Medicine, University of Nebraska Medical Center

Brian W Locke: Intermountain Medical Center, Department of Pulmonary and Critical Care.

**Financial Support**

This research was supported by a grant from the Intermountain Fund (B.W.L.)

**Conflicts of Interest**

B.W.L. claims an equity interest in Mountain Biometrics, a startup focused on machine learning medical time series data.

**Description:** (1-2 sentences)

Large language model-estimated diagnostic likelihood ratios showed negligible bias bounded dispersion with 700 literature-reported likelihood ratios, demonstrating their potential to supply auditable evidence weights for Bayesian reasoning in clinical diagnosis.

# Abstract: (\*\*\*/300 words)

*Introduction*

Accurate, context-appropriate likelihood ratios (LRs) are required to apply Bayesian reasoning in clinical diagnosis, yet empiric LR estimates are scarce because diagnostic test accuracy studies are onerous to perform. Large language models (LLM) may be able to estimate diagnostic LRs by drawing on indirect or inferred clinical associations.

*Methods*

We elicited numeric LRs from three OpenAI models (GPT-4o, o3, GPT-5) and compared them with all literature-reported values curated by TheNNT.com. A few-shot prompt was served to each LLM to elicit numerical LR estimates. Agreement was evaluated using Bland-Altman analysis for mean bias and multiplicative limits of agreement.

*Results*

We compiled 700 literature-reported LRs for 30 conditions. Most involved signs/symptoms (59%), historical elements (19%), or test results (16%). Reported LRs clustered near 1 (geometric mean 1.21, interquartile range 0.7 to 2.2). All models showed negligible mean bias. GPT-5 had the narrowest 95% limits of agreement (0.26x to 3.7x) versus o3 (0.23x to 4.28x) and GPT-4o (0.23x to 4.53x). GPT-5 limits were significantly narrower than o3 and GPT-4o (P < 0.001 for each comparison). Agreement varied by finding type, with laboratory test LRs varying more from reported estimates than history, signs/symptoms, or imaging.

*Conclusions*

Modern LLMs can estimate diagnostic LRs with very low bias and bounded dispersion, with newer models producing estimates more closely approximating values from the literature.

These results indicate significant potential for integrating generative AI into clinical diagnostic workflows and decision support, particularly in settings where empirical data are limited, outdated, or unavailable. By supplying auditable likelihood ratios, LLMs could enable scalable Bayesian updating at the point of care and serve as prior hypotheses for future diagnostic studies.

# Introduction:

Medical diagnosis requires integrating history, examination, and test findings to identify the condition that best explains a patient’s presentation1,2. Bayesian reasoning provides a principled framework for this task because it is information-efficient, broadly applicable, and transparent.3–5. Instruction in Bayesian methods can improve clinicians’ diagnostic reasoning 6, yet broader clinical adoption remains modest and many clinicians continue to rely on intuition, heuristics, and pattern recognition.7,8

A major barrier to routine Bayesian updating at the bedside is the scarcity of accurate, context-specific likelihood ratios (LRs). LRs quantify how the presence or absence of a finding (history, symptom, examination sign, or test result) changes the odds of disease3,4. However, empirical LR estimates require diagnostic accuracy studies, which are often difficult to perform and interpret9,10; consequently, reliable LRs for many common clinical findings are unavailable. and true likelihood ratios often vary substantially by clinical context11. Because the condition–finding–context space is combinatorially large, exhaustive empirical measurement is infeasible. For applied Bayesian reasoning to scale, inferred LRs that are auditable and plausibly accurate are therefore important.

Recent advances in generative artificial intelligence offer one possible approach to this evidence gap. Large language models (LLMs) are neural networks trained on extensive text corpora that capture clinical knowledge and concept associations12,13. They can generalize to new tasks through zero-shot or in-context learning14,15,, suggesting they might infer LRs when empirical data are absent. While their current applications in medicine remain under evaluation16-18, if LLMs can produce quantitative, inspectable LR estimates, those outputs could support bedside Bayesian updating, populate decision-support tools, and serve as prior predictions for future diagnostic studies.

Given that Bayesian reasoning remains underused in clinical practice primarily due to the scarcity of reliable evidence weights, the ability of LLMs to approximate diagnostic LRs could bridge a long-standing translational gap between data and decision-making. 3,4,27 In this study, we evaluated successive generations of LLMs to determine how closely their inferred LRs align with published values, and to explore whether AI-generated estimates can serve as credible, inspectable surrogates for empirical diagnostic evidence. Beyond serving as immediate estimates, these model-generated values could function as prior hypotheses for future diagnostic accuracy studies, providing predictions that can later be reconciled with empirical results. In this role, LLMs act as a prediction engine that not only fills current evidence gaps but also offers a benchmark against which newer models and future data can be tested.20,26

**Methods:**

We conducted a comparative study assessing the agreement between diagnostic LRs generated by three 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 curated repository diagnostic likelihood ratios from published medical literature. Point estimates for all LRs from all conditions listed on theNNT.com were either recorded, or derived as the geometric mean when only a range was provided (e.g. 1.5, 95% CI 1 – 2 would be coded as 1.5, while a range from 1 to 2 would be recorded as 1.41). LRs were initially extracted using an automated script and then manually validated with duplicate independent review (PC and BWL). Each LR was categorized as a patient historical element, a sign/symptom, a test result, an imaging finding, and/or a diagnostic adjudication (e.g. “diagnosis based on ultrasound”). Scores (e.g. Centor criteria for Strep pharyngitis) were counted as each of the constituent findings. 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), consistent with prior literature3,4.

### Comparator Likelihood Ratios

On August 25, 2025, we generated comparator likelihood ratios (LRLLM) for all findings listed on theNNT.com using a constrained, few-shot prompting procedure. To represent a range of model ages, complexity, and inference costs, we queried three OpenAI LLMs (OpenAI, LP; San Francisco, California, USA) using the OpenAI API: GPT-4o (model release Nov 20, 2024), o3 (release Apr 16, 2025), and GPT-5 (release Aug 7, 2025). A full description of the prompting strategy is included in the supplement. In brief, the system prompt defined the LR, ‘P(finding | diagnosis) / P(finding | not diagnosis)’, gave qualitative LR strength descriptions3, and required a numeric-only response. We used 8 (non-reasoning model, GPT-4o) or 2 (reasoning models, o3 and GPT-5) clinician-estimated few-shot examples. These were clinician-estimated finding-clinical state-LR groups that were not in the evaluation set. Inference settings were temperature = 0.2 (non-reasoning model) and ‘reasoning effort’ = "medium" (reasoning models); text.verbosity = "low" was applied where supported (GPT-5 only). Internet search was not enabled for any models. No model fine-tuning was performed.

### Statistical Analysis

We assessed the agreement between reported likelihood ratios (LRReported) and LLM-estimated likelihood ratios (LRLLM) using Bland-Altman analysis19 on log-transformed LR values, as strength of evidence is additive on the log scale4,20. We calculated multiplicative (ratio) limits of agreement, which indicate the range within which LRLLM is expected to lie within an x-fold difference of the LRReported in 95% of cases. 50%, 75%, 90%, and 99% limits of agreement are also tabulated in the supplemental materials.

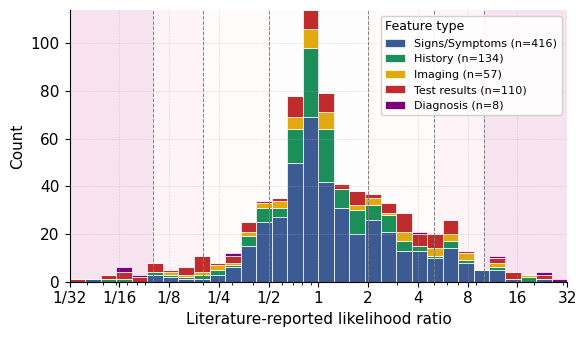
We compared models using paired t-test for mean differences (bias) and the Pittman-Morgan test for differences in the width (variance) of the limits of agreement21,22. Subgroup analyses were conducted by information type (historical element, symptom/sign, examination finding, test result, or diagnostic adjudication), and by direction of evidence (positive: LR >1; negative: LR < 1) using Welch’s t-test for bias and Levene’s test for differences in width. Agreement between qualitative LR strength categories (e.g. strong, moderate, weak)3,4 was assessed using Cohen’s Kappa with quadratic weights, which penalizes large disagreements more heavily and approximates squared-error on the underlying likelihood ratio scale23,24. Statistical significance was set at α = 0.05 without adjustment for multiple testing.Analyses were conducted in Python 3.11.11 and Microsoft Excel. Code is available at <https://github.com/reblocke/llm_estimate_lrs> .

# Results:

700 LRReported exploring the 30 available medical conditions were compiled from theNNT.com. Signs/symptoms were the most common type of LR (59%, n=416), followed by historical element (19% , n=134) and test results (16% , n=110)

LRReported values ranged from 0.01 to 145.9, with a median of 1 (interquartile range 0.7 to 2.2) and a geometric mean of 1.21. Figure 1 shows the distribution of strength of evidence in the LRReported. Most (n=400) findings offered negligible strength of evidence (0.5 < LR < 2; 56.5%), with the next most common being weak evidence in favor of a diagnosis (n=120, 17.4%), weak evidence against (n=60, 8.7%) and moderate evidence for (n=52, 7.5%). Diagnostic adjudications tended to provide the strongest evidence, while signs/symptoms were the weakest (Supplemental Table 1).

**Figure 1:** Distribution of likelihood ratios reported in the literature, as collected from theNNT.com. Background shading represents strong, moderate, weak, and negligible strength of evidence categories3,4. Most of the LRreported cluster near 1, showing they offer negligible or weak evidence.

****

Three sets of LRLLM were generated (for GPT-4o, o3, and GPT-5) for each of the 700 LRReported. Figure 2 shows the 95% multiplicative limits of agreement. All 3 models showed excellent mean bias (GPT-4o 1.02x, o3 0.99x, GPT-5 0.99x; no differences in pairwise comparisons). LR estimates from GPT-5 had the narrowest limits of agreement to the values reported on theNNT.com (95% limits of agreement from 0.26x to 3.7x, *P* < .001 vs both o3 and GPT-4o ), followed by o3 (0.23x to 4.28x), and GPT-4o (0.23x to 4.53x, *P* = .58 for o3 vs GPT-4o). Other coverage ranges (50, 75%, 90%, and 99%) are presented in the supplementary materials.

A diagram of a number of data

AI-generated content may be incorrect.

**Figure 2: Agreement between literature-reported and LLM-generated likelihood ratios:** Each panel shows the agreement between reported and model-generated likelihood ratios (LRs). The y-axis shows the LR ratio (reported/model), and the x-axis shows the geometric mean of paired LRs. Solid black lines represent mean bias; Dashed lines indicate the multiplicative (i.e. x-fold) range in which 95% of estimates would be expected to be from a value reported on theNNT.com. 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. All models showed negligible bias. GPT-5 had the tightest agreement with reported likelihood ratios.

Figure 3 shows the limits of agreement by finding type. Estimates of the strength of evidence followed similar patterns across models. Estimates of the importance of sign/symptoms, historical elements, and imaging findings were all similarly accurate, while test results agreed less closely with reported estimates.

A screenshot of a graph

AI-generated content may be incorrect.

**Figure 3: Agreement between LLM-estimated and literature-reported likelihood ratios** by type of clinical finding. Rows represent clinical finding categories (laboratory, imaging, history, sign-symptom, and diagnosis); columns represent LLM versions (GPT-4o, o3, and GPT-5). Categories were assigned by manual review, and likelihood ratios could be categorized as multiple types (e.g. scores integrating multiple findings).

When analyzed by the direction of evidence, patterns differed between models (Supplemental Figure 2). For GPT-4o, mean bias differed between negative (LRreported < 1) evidence (mean 1.13x) vs. positive (LRreported > 1) findings (mean 0.92x, *P* < .001). No difference in mean bias was observed for o3 (0.95x vs. 1.03x; *P* = 0.15). GPT-5 showed the opposite pattern of mean bias (negative evidence 0.88x vs positive evidence 1.12x; *P* < .001). The width of the limits of agreement did not differ by evidence direction for any of the models (GPT-4o *P* = .11; o3 *P* = .40; GPT-5 *P* = .37). For all 3 models, the calibration slope of predictions on the logarithmic scale suggested predictions were slightly less extreme than literature reported values (Supplemental Figures 2a-c)

Agreement between qualitative evidence categories was moderate for all models (Supplemental Figures 3a-c). The quadratic-weighted Cohen’s κ for GPT-5 was highest (0.775, 95% confidence interval [CI] 0.728-0.821), followed by o3 (0.745, 95% CI 0.702 – 0.789), and GPT-4o (0.734, 95% CI 0.691 – 0.778).

# Discussion:

We found that modern LLMs can estimate diagnostic likelihood ratios with negligible mean bias and bounded dispersion. Across 700 finding–condition pairs covering 30 conditions, GPT-5 demonstrated the closest agreement with literature-reported values (95% limits of agreement of 0.26×–3.70×). Agreement varied by finding type, with laboratory test results showing looser agreement than history, signs/symptoms, or imaging results. Qualitative category agreement was substantial (κ = 0.78 for GPT-5*).*25 Together, these results suggest that LLM-estimated LRs could support scalable Bayesian reasoning in clinical care by supplying auditable evidence weights for decision support, enabling Bayesian updating at the point of care, and serving as priors for future diagnostic research27,31

Prior work shows that AI can augment human diagnostic reasoning by offering complementary strengths.26 Bayesian reasoning requires that likelihood ratios be known and reliable, which is rarely the case in routine practice.27,28 Our study presents, to our knowledge, the first large-scale evaluation of 700 finding–condition pairs across 30 conditions. We found that LLMs produced likelihood ratio estimates with quantitative accuracy across all categories, which suggests that LLMs could serve as a bridge in medical education and clinical workflows, enabling structured reasoning in situations where empirical LRs are unavailable.

Our study extends prior research by translating a component of LLM clinical reasoning into a quantitative unit, the LR4. In contrast to black-box outputs17, quantification enables transparency, facilitates auditing, and integrates naturally into human or AI reasoning pipelines. Whether the observed dispersion is acceptable depends on the use case. It may be too wide when a single, high-stakes finding drives a decision, but the negligible mean bias implies that serial Bayesian updating of multiple findings would yield unbiased posterior probabilities if LR independence holds, a condition often imperfect but typically adequate, similar to the robust performance of naïve Bayes classifiers.29.

Strengths of this work include introducing a paradigm whereby LLMs diagnostic contribution is decomposed to one component of a clinical reasoning pipeline. We used a reproducible evaluation methodology across a broad set of comparators and multiple model generations and quantified agreement of both qualitative categories and continuous metrics. Both showed similar performances. Interestingly, using long-standing strength of evidence classification 3, most of the empirically studied diagnostic features represented “negligible or weak evidence, underscoring that individual findings may generally be less diagnostic than creators of the classification scheme anticipated.

Several limitations warrant mention. We did not manually extract LR confidence intervals or study-level context from the primary sources compiled by TheNNT.com, so our reference standards include hidden uncertainty, and models lacked the clinical detail necessary to match the precise patient spectrum to which a given LR applies.11 We also did not assess the methodological rigor of those studies10 and reported only quantitative agreement, not a taxonomy of model errors. The evaluation compares model outputs to published LRs that may have been available during model pretraining, introducing potential contamination. Although we observed no verbatim reproduction and disabled web access, prior exposure could influence model weights and modestly tighten observed agreement.30 Nonetheless, the systematic patterns observed differences across model generations, variation by finding type, and bounded dispersion suggest that the results capture meaningful aspects of model behavior rather than memorization alone.

LLM-estimated LRs have several high-value clinical applications. They can provide auditable, quantitative weights that make black-box predictions interpretable and can be combined with clinician judgment in explicit Bayesian updating for diagnosis and risk stratification. Model-generated LRs could populate decision-support tools, inform pretest probabilities in diagnostic pathways, and act as prior predictions that guide the design and interpretation of future diagnostic-accuracy studies.33 These uses complement, rather than replace, educational applications such as scenario generation and structured reasoning tutors,32 while the primary value lies in enabling evidence-weighted clinical decision-making that scales beyond manually curated evidence summaries.

Looking ahead, the critical question is prospective: can LLMs generate likelihood ratios for findings not yet studied and anticipate the results of future diagnostic-accuracy research? Addressing this will require leakage-controlled, preregistered benchmarks with embargoed targets to test true predictive ability.17,18,30 Future work should evaluate whether tighter specification of patient context and care setting improves the precision of LLM-estimated LRs, whether model biases mirror those of clinicians, and whether calibrated uncertainty estimates can be generated. Comparative studies should benchmark LLM performance against clinicians across expertise levels, and implementation studies should assess how embedding model-generated LRs into workflows affects diagnostic accuracy, cognitive load, and patient outcomes.16,31,33

# Conclusion:

Our findings suggest that large language models may now approximate a fundamental element of diagnostic reasoning: the translation of clinical findings into quantitative evidence weights. If validated prospectively, this capability could bridge the long-standing gap between probabilistic theory and routine clinical decision-making.Rather than replacing empirical research, model-derived likelihood ratios could prioritize questions for future diagnostic studies, guide hypothesis generation, and support transparent, auditable reasoning within decision-support systems.15,16

# References:

1. McGee S, ed. *Evidence-Based Physical Diagnosis*. Fourth Edition. Elsevier; 2018. doi:10.1016/B978-0-323-39276-1.12001-3

2. Newman TB, Kohn MA. *Evidence-Based Diagnosis: An Introduction to Clinical Epidemiology*. Cambridge University Press; 2020.

3. Jaeschke R, Guyatt GH, Sackett DL, et al. Users’ Guides to the Medical Literature: III. How to Use an Article About a Diagnostic Test B. What Are the Results and Will They Help Me in Caring for My Patients? *JAMA*. 1994;271(9):703-707. doi:10.1001/jama.1994.03510330081039

4. Grimes DA, Schulz KF. Refining clinical diagnosis with likelihood ratios. *The Lancet*. 2005;365(9469):1500-1505. doi:10.1016/S0140-6736(05)66422-7

5. Zellner A. Optimal Information Processing and Bayes’s Theorem. *Am Stat*. Published online November 1, 1988. Accessed September 16, 2025. https://www.tandfonline.com/doi/abs/10.1080/00031305.1988.10475585

6. Brush JE Jr, Lee M, Sherbino J, Taylor-Fishwick JC, Norman G. Effect of Teaching Bayesian Methods Using Learning by Concept vs Learning by Example on Medical Students’ Ability to Estimate Probability of a Diagnosis: A Randomized Clinical Trial. *JAMA Netw Open*. 2019;2(12):e1918023-e1918023. doi:10.1001/jamanetworkopen.2019.18023

7. Brush Jr JE, Sherbino J, Norman GR. How expert clinicians intuitively recognize a medical diagnosis. *Am J Med*. 2017;130(6):629-634. doi:10.1016/j.amjmed.2017.01.045

8. Reid MC, Lane DA, Feinstein AR. Academic calculations versus clinical judgments: practicing physicians’ use of quantitative measures of test accuracy. *Am J Med*. 1998;104(4):374-380. doi:10.1016/s0002-9343(98)00054-0

9. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *The BMJ*. 2015;351:h5527. doi:10.1136/bmj.h5527

10. Kohn MA, Carpenter CR, Newman TB. Understanding the direction of bias in studies of diagnostic test accuracy. *Acad Emerg Med Off J Soc Acad Emerg Med*. 2013;20(11):1194-1206. doi:10.1111/acem.12255

11. Usher-Smith JA, Sharp SJ, Griffin SJ. The spectrum effect in tests for risk prediction, screening, and diagnosis. *BMJ*. 2016;353. doi:10.1136/bmj.i3139

12. Shmatko A, Jung AW, Gaurav K, et al. Learning the natural history of human disease with generative transformers. *Nature*. Published online September 17, 2025:1-9. doi:10.1038/s41586-025-09529-3

13. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. doi:10.1038/s41586-023-06291-2

14. Brown T, Mann B, Ryder N, et al. Language Models are Few-Shot Learners. In: *Advances in Neural Information Processing Systems*. Vol 33. Curran Associates, Inc.; 2020:1877-1901. Accessed September 17, 2025. https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

15. Howell MD, Corrado GS, DeSalvo KB. Three Epochs of Artificial Intelligence in Health Care. *JAMA*. 2024;331(3):242-244. doi:10.1001/jama.2023.25057

16. Goh E, Gallo R, Hom J, et al. Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. *JAMA Netw Open*. 2024;7(10):e2440969. doi:10.1001/jamanetworkopen.2024.40969

17. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med*. 2024;30(9):2613-2622. doi:10.1038/s41591-024-03097-1

18. Sahoo SS, Plasek JM, Xu H, et al. Large language models for biomedicine: foundations, opportunities, challenges, and best practices. *J Am Med Inform Assoc JAMIA*. 2024;31(9):2114-2124. doi:10.1093/jamia/ocae074

19. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet Lond Engl*. 1986;1(8476):307-310.

20. Good IJ. Weight of Evidence: A Brief Survey. In: *Bayesian Statistics 2*. Elsevier Science Publishers; 1985:249-270.

21. Pitman E. A note on normal correlation. *Biometrika*. 1939;31(1/2):9-12.

22. Morgan W. A test for the significance of the difference between the two variances in a sample from a normal bivariate population. *Biometrika*. 1939;31(1/2):13-19.

23. Fleiss JL, Cohen J, Everitt BS. Large sample standard errors of kappa and weighted kappa. *Psychol Bull*. 1969;72(5):323. doi:10.1037/h0028106

24. Fleiss JL, Cohen J. The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. *Educ Psychol Meas*. 1973;33(3):613-619. doi:10.1177/001316447303300309

25. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. *Biometrics*. 1977;33(1):159-174. doi:10.2307/2529310

26. Steyvers M, Tejeda H, Kerrigan G, Smyth P. Bayesian modeling of human–AI complementarity. *Proc Natl Acad Sci*. 2022;119(11):e2111547119. doi:10.1073/pnas.2111547119

27. Simel DL, Rennie D. Forward. In: *The Rational Clinical Examination: Evidence-Based Clinical Diagnosis*. McGraw-Hill Education; 2016. Accessed September 21, 2025. jamaevidence.mhmedical.com/content.aspx?aid=1170154778

28. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical Evidence of Design-Related Bias in Studies of Diagnostic Tests. *JAMA*. 1999;282(11):1061-1066. doi:10.1001/jama.282.11.1061

29. Zhang H. Exploring Conditions for the Optimality of Naïve Bayes. *Int J Pattern Recognit Artif Intell*. 2005;19(02):183-198. doi:10.1142/S0218001405003983

30. Carlini N, Ippolito D, Jagielski M, Lee K, Tramer F, Zhang C. Quantifying Memorization Across Neural Language Models. In: 2022. Accessed September 19, 2025. https://openreview.net/forum?id=TatRHT\_1cK

31. Rodman A, Buckley TA, Manrai AK, Morgan DJ. Artificial Intelligence vs Clinician Performance in Estimating Probabilities of Diagnoses Before and After Testing. *JAMA Netw Open*. 2023;6(12):e2347075-e2347075. doi:10.1001/jamanetworkopen.2023.47075

32. Jagannath AD, Dreicer JJ, Penner JC, Dhaliwal G. The cognitive apprenticeship: advancing reasoning education by thinking aloud. *Diagnosis*. 2023;10(1):9-12. doi:10.1515/dx-2022-0043

33. Staal J, Hooftman J, Gunput STG, et al. Effect on diagnostic accuracy of cognitive reasoning tools for the workplace setting: systematic review and meta-analysis. *BMJ Qual Saf*. 2022;31(12):899-910. doi:10.1136/bmjqs-2022-014865

34. Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Digit Health*. 2024;6(1):e12-e22. doi:10.1016/S2589-7500(23)00225-X

# Supplemental Materials

## Prompt Details

Code availability. The full pipeline for collating likelihood ratios (LRs) from TheNNT.com and generating large‑language‑model (LLM) estimates is available at: <https://github.com/reblocke/llm_estimate_lrs>

**Overview:** We prompted the LLMs to produce a single numeric likelihood ratio (LR) for a clinical finding with respect to a diagnosis, under a constrained output schema (only a numerical response) and with minimal verbosity.

**Prompt specification:**

* System role: “You are a Bayesian diagnostic assistant. Estimate a numeric likelihood ratio (LR) for a finding with respect to a diagnosis. Return only a JSON object matching the schema: {“value”: }, where value > 0.”
* Definition shown to the model: LR = P(finding | diagnosis) / P(finding | not diagnosis)
* Qualitative evidence bands provided as context. >10 strong for; 5–10 moderate for; 2–5 weak for; 0.5–2 negligible; 0.2–0.5 weak against; 0.1–0.2 moderate against; ≤0.1 strong against.
* Inputs to the model are plain text pairs: ”Condition: <diagnosis>\nFinding: <finding>”
* preceded by the system prompt, definition/bands, and few‑shot examples.

**Few‑shot strategy:** We include exemplar (Condition, Finding → LR) pairs to anchor scale:

* Non‑reasoning models (e.g., GPT‑4o): 8 examples.
* Reasoning models (o3 series; GPT‑5 family): 2 examples.

Example LRs were clinician‑estimated, not scraped values, to reduce the chance that exemplars appear in the evaluation set or anchor to relevant model pretraining data.

**Inference settings:**

* Non‑reasoning models: temperature = 0.2.
* Reasoning models: reasoning = {"effort": "medium"}; no temperature.
* Verbosity control: where supported (GPT‑5 family), text.verbosity = "low". The JSON‑only response format further suppresses extraneous text.

**Output schema and validation**:

* Schema. Structured output {"value": float}; parser enforces numeric type.
* Requests use the Responses API with a Pydantic schema (LRResponse { value: float }) to enforce structure.
* Validity rule. Accept only finite, strictly positive values.
* Retry logic. If the response is non‑numeric, malformed, non‑finite, or ≤0, the call is retried with exponential backoff and jitter until a valid LR is obtained (or up to the configured maximum, if set).

**Code:**

from \_\_future\_\_ import annotations

import os

import logging

from pathlib import Path

from typing import Optional

import time, math

from random import random

import pandas as pd

from pydantic import BaseModel

from openai import OpenAI

# -----------------------------------------------------------------------------

# 0) Configuration

# -----------------------------------------------------------------------------

logging.basicConfig(level=logging.WARNING)

client = OpenAI(api\_key=os.getenv("OPENAI\_API\_KEY"))

# Model registry:

MODEL\_CAPABILITIES = {

# GPT‑5 series (reasoning; supports text.verbosity; no temperature)

"gpt-5" : {"reasoning": True, "verbosity": True, "allow\_temp": False},

"gpt-5-mini" : {"reasoning": True, "verbosity": True, "allow\_temp": False},

"gpt-5-nano" : {"reasoning": True, "verbosity": True, "allow\_temp": False},

# GPT‑4.1 family (non‑reasoning; temperature OK); include snapshots + aliases

"gpt-4.1-2025-04-14" : {"reasoning": False, "verbosity": False, "allow\_temp": True},

"gpt-4.1-mini-2025-04-14": {"reasoning": False, "verbosity": False, "allow\_temp": True},

"gpt-4.1-nano-2025-04-14": {"reasoning": False, "verbosity": False, "allow\_temp": True},

# GPT‑4o family (non‑reasoning; temperature OK); prefer latest snapshot or alias

"gpt-4o-2024-11-20" : {"reasoning": False, "verbosity": False, "allow\_temp": True},

"gpt-4o-mini-2024-07-18": {"reasoning": False, "verbosity": False, "allow\_temp": True},

# o‑series (reasoning; no temperature)

"o3-2025-04-16" : {"reasoning": True, "verbosity": False, "allow\_temp": False},

"o3-mini-2025-01-31": {"reasoning": True, "verbosity": False, "allow\_temp": False},

"o4-mini-2025-04-16": {"reasoning": True, "verbosity": False, "allow\_temp": False},

}

MODELS = list(MODEL\_CAPABILITIES)

# -----------------------------------------------------------------------------

INPUT\_FILE = "nnt\_lrs\_processed.xlsx"

OUTPUT\_FILE = "nnt\_lrs\_with\_estimated.xlsx"

# -----------------------------------------------------------------------------

# 1) Prompt

# -----------------------------------------------------------------------------

SYSTEM\_CORE = """You are a Bayesian diagnostic assistant.

Estimate a numeric likelihood ratio (LR) for a finding with respect to a diagnosis.

Return only a JSON object matching the schema: {"value": <float>}, where value > 0.

"""

DEFINITION = """Definition:

LR = P(finding | diagnosis) / P(finding | not-diagnosis)

"""

BANDS = """LR evidence bands (reference):

>10 strong for; 5-10 moderate for; 2–5 weak for;

0.5–2 negligible;

0.2-0.5 weak against; 0.1-0.2 moderate against; ≤0.1 strong against"""

# Few‑shot examples - these are human guestimates (to avoid seeding the dataset and inflating performance)

FEW\_SHOT\_RICH = [

("deep vein thrombosis", "femoral vein noncompressaible on ultrasound", 16.0), # some data this might be higher?

("pericarditis", "pleuritic chest pain improved by leaning forward", 5.2),

("pulmonary embolism", "tachycardia >100 bpm", 2.2),

("urinary tract infection", "malodorous urine", 1.1),

("myocardial infarction", "enjoys playing chess", 1.0),

("appendicitis", "no RLQ tenderness", 0.45),

("pneumothorax", "bilateral lung sliding present on US", 0.18), # some data this might be lower?

("HIV infection", "4th‑generation Ag/Ab screen negative beyond window",0.05),

]

FEW\_SHOT\_MIN = [

("deep vein thrombosis", "femoral vein noncompressaible on ultrasound", 16.0),

("myocardial infarction", "enjoys playing chess", 1.0),

]

def build\_messages(diagnosis: str, finding: str, reasoning: bool) -> list[dict]:

msgs: list[dict] = [

{"role": "system", "content": SYSTEM\_CORE.strip()},

{"role": "system", "content": DEFINITION.strip()},

{"role": "system", "content": BANDS.strip()},

]

examples = FEW\_SHOT\_MIN if reasoning else FEW\_SHOT\_RICH

for dx\_ex, f\_ex, v\_ex in examples:

msgs.append({"role": "user", "content": f"Condition: {dx\_ex}\nFinding: {f\_ex}"})

msgs.append({"role": "assistant", "content": f'{{"value": {float(v\_ex)}}}'})

msgs.append({"role": "user", "content": f"Condition: {diagnosis}\nFinding: {finding}"})

return msgs

# -----------------------------------------------------------------------------

# 2) Structured Outputs schema (Pydantic)

# -----------------------------------------------------------------------------

class LRResponse(BaseModel):

value: float

# -----------------------------------------------------------------------------

# 2b) Retry wrapper (exponential backoff with jitter)

# -----------------------------------------------------------------------------

def estimate\_lr\_until\_positive(

diagnosis: str,

finding: str,

model: str,

client: Optional[OpenAI] = None,

max\_retries: Optional[int] = None, # None ⇒ retry indefinitely

base\_backoff: float = 0.5, # seconds

max\_backoff: float = 30.0 # seconds

) -> float:

attempt = 0

while True:

attempt += 1

try:

lr = estimate\_lr(diagnosis, finding, model, client)

if isinstance(lr, (int, float)) and math.isfinite(lr) and lr > 0:

return float(lr)

raise ValueError(f"Non‑positive or non‑finite LR: {lr!r}")

except Exception as e:

logging.warning(

f"[retry {attempt}] sheet finding='{finding[:80]}' | "

f"model={model} → {e}"

)

if (max\_retries is not None) and (attempt >= max\_retries):

raise

# exponential backoff with jitter

delay = min(base\_backoff \* (2 \*\* (attempt - 1)), max\_backoff)

time.sleep(delay \* (0.5 + random())) # 0.5–1.5× jitter

# -----------------------------------------------------------------------------

# 3) Estimator call (Responses API)

# -----------------------------------------------------------------------------

def estimate\_lr(diagnosis: str, finding: str, model: str, client: Optional[OpenAI] = None) -> float:

if client is None:

client = OpenAI()

cfg = MODEL\_CAPABILITIES[model]

msgs = build\_messages(diagnosis, finding, reasoning=cfg["reasoning"])

kwargs = {}

if cfg["reasoning"]:

kwargs["reasoning"] = {"effort": "medium"} # for GPT‑5 and o‑series

# no temperature/top\_p

elif cfg["allow\_temp"]:

kwargs["temperature"] = 0.2 # allowed for 4o / 4.1

# Apply verbosity only where supported (GPT‑5 family)

if cfg["verbosity"]:

kwargs["text"] = {"verbosity": "low"}

resp = client.responses.parse(

model=model,

input=msgs,

text\_format=LRResponse, # Structured Outputs → Pydantic

\*\*kwargs,

)

return float(resp.output\_parsed.value)

# -----------------------------------------------------------------------------

# 4) Main pipeline: read workbook → append model columns → write output

# -----------------------------------------------------------------------------

def run\_batch(input\_file: str | Path, output\_file: str | Path, models: list[str]) -> None:

sheets = pd.read\_excel(input\_file, sheet\_name=None, header=None)

for sheet\_name, df in sheets.items():

diagnosis = str(df.iloc[0, 0]).strip()

for model in models:

new\_header = "lr\_" + model

col = []

print(f"→ {diagnosis[:60]} | {model}")

for i in range(len(df)):

if i == 0:

col.append("") # top-left cell (sheet label row)

elif i == 1:

col.append(new\_header) # column header row

else:

finding = str(df.iloc[i, 0]).strip()

if not finding:

col.append("") # keep blank rows blank

continue

try:

# retry until a strictly positive, finite float is returned

lr = estimate\_lr\_until\_positive(

diagnosis, finding, model, client,

max\_retries=None # set to an int (e.g., 8) to cap retries

)

except Exception as e:

lr = "ERROR"

logging.warning(

f"Error on sheet '{sheet\_name}', row {i}, model {model} after retries: {e}"

)

col.append(lr)

# Insert as object dtype to accommodate strings like "ERROR"

df.insert(df.shape[1], new\_header, pd.Series(col, dtype="object"))

sheets[sheet\_name] = df

with pd.ExcelWriter(output\_file, engine="openpyxl") as writer:

for name, frame in sheets.items():

frame.to\_excel(writer, sheet\_name=name, index=False, header=False)

print(f"Done – results saved to '{output\_file}'")

if \_\_name\_\_ == "\_\_main\_\_":

run\_batch(INPUT\_FILE, OUTPUT\_FILE, MODELS)

## Supplemental Table 1: Distribution of Reported Likelihood Ratios, by type

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Statistic | Overall | Test results | Imaging | History | Signs Symptoms | Diagnosis |
| Count | 700 | 110 | 57 | 134 | 416 | 8 |
| Geometric mean | 1.206 | 1.071 | 1.322 | 1.065 | 1.267 | 1.164 |
| 5th percentile | 0.190 | 0.060 | 0.200 | 0.226 | 0.360 | 0.064 |
| 25th percentile | 0.700 | 0.312 | 0.680 | 0.755 | 0.700 | 0.077 |
| 50th percentile | 1.000 | 1.000 | 1.000 | 0.995 | 1.000 | 2.057 |
| 75th percentile | 2.200 | 3.675 | 3.300 | 1.675 | 2.100 | 16.300 |
| 95th percentile | 7.905 | 15.550 | 12.000 | 5.085 | 7.225 | 26.300 |
| Min | 0.010 | 0.010 | 0.010 | 0.050 | 0.040 | 0.060 |
| Max | 145.894 | 145.894 | 34.400 | 18.500 | 57.000 | 27.000 |

## Supplemental Figure 1: Limits of Agreement by Direction of Evidence

Rows separate positive (LRreported > 1) from negative findings (LRreported < 1).

A group of colored dots

AI-generated content may be incorrect.

## Supplemental Table 2: Coverage Intervals

Limits of Agreement 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: GPT-4o**

|  |  |  |
| --- | --- | --- |
| Coverage | Lower edge (95% CI) | Upper edge (95% CI) |
| 50% | 0.61x (0.57x - 0.65x) | 1.70 (1.60x - 1.81x) |
| 75% | 0.42x (0.39x - 0.46x) | 2.44 (2.27x - 2.63x) |
| 90% | 0.29x (0.27x - 0.32x) | 3.56 (3.27x - 3.88x) |
| 95% | 0.23x (0.21x - 0.25x) | 4.53 (4.11x - 4.99x) |
| 99% | 0.14x (0.13x - 0.16x) | 7.24 (6.43x - 8.14x) |

**Model: o3**

|  |  |  |
| --- | --- | --- |
| Coverage | Lower edge (95% CI) | Upper edge (95% CI) |
| 50% | 0.60x (0.56x - 0.63x) | 1.64 (1.54x - 1.74x) |
| 75% | 0.42x (0.39x - 0.45x) | 2.33 (2.17x - 2.51x) |
| 90% | 0.29x (0.26x - 0.31x) | 3.38 (3.10x - 3.68x) |
| 95% | 0.23x (0.21x - 0.25x) | 4.28 (3.89x - 4.71x) |
| 99% | 0.14x (0.13x - 0.16x) | 6.79 (6.05x - 7.62x) |

**Model: GPT-5**

|  |  |  |
| --- | --- | --- |
| Coverage | Lower edge (95% CI) | Upper edge (95% CI) |
| 50% | 0.63x (0.59x - 0.66x) | 1.56 (1.47x - 1.65x) |
| 75% | 0.46x (0.43x - 0.49x) | 2.15 (2.01x - 2.29x) |
| 90% | 0.33x (0.30x - 0.35x) | 2.99 (2.77x - 3.23x) |
| 95% | 0.26x (0.24x - 0.29x) | 3.70 (3.40x - 4.03x) |
| 99% | 0.17x (0.16x - 0.19x) | 5.61 (5.05x - 6.22x) |

## Supplemental Figure 2 (a-c): Calibration Plots

*For calibration we regressed* y=\ln(\mathrm{LR}{\text{LLM}}) *on* x=\ln(\mathrm{LR}{\text{reported}}) *using ordinary least squares with intercept:* y=\alpha+\beta x+\varepsilon*. We report* \alpha *(intercept),* \beta *(slope), their 95% Wald confidence intervals, and* R^2*. Perfect calibration on the log scale corresponds to* \alpha=0 *and* \beta=1*;* \beta<1 *indicates shrinkage (less‑extreme LRs). Plots overlay the identity line, binned means (quantile bins), and nonparametric smoothers (LOWESS; isotonic regression) as descriptive complements; inference is based on the OLS fit.*

If space permits, add: *All regressions used complete cases for the pair* (x,y)*; no robust SEs were required given residual diagnostics showed no severe heteroskedasticity at the analysis scale.*

LR’s are presented on the logarithmic scale.



## Supplemental Figure 3(a-c): Qualitative Agreement between LLMs and Literature-reported LRs



