# Title:

Estimation of Medical Diagnostic Likelihood Ratios Using Artificial Intelligence

# Authors:

Shuhan He

Paul Chong

Kian

Cory

Boyu

Emma Chua

Adriana Coleska

Ahmad

Brian W. Locke, MD MSc – ORCID ([0000-0002-3588-5238](https://orcid.org/0000-0002-3588-5238)). Intermountain Medical Center, Department of Pulmonary and Critical Care.

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)

**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.

**Acknowledgement/Author contributions:**

B.W.L Conceptualization, Methodology, Software, Data Curation, Writing – Review and Editing, Visualization, Supervision. Guarantor of integrity of the entire study

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

Large language models (LLMs) have the potential to accurately estimate likelihood ratios (LRs) to aid in clinical diagnostic decision-making without the need for resource-intensive studies. Our study demonstrated this possibility using 700 reported empirical LRs as well as three OpenAI LLMs.

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

*Introduction*

Likelihood ratios can assist in clinical diagnosis but, but few empirically-estimated likelihood ratios exist because they require resource-intensive studies to establish. Advancements in artificial intelligence and large language models present the possibility of estimated LRs that circumvent the aforementioned barriers of establishing likelihood ratios.

*Methods*

We evaluated the accuracy of large language model-estimated likelihood ratios against empirically derived values curated at theNNT.com. Three OpenAI models (GPT-4o, o3, GPT-5) were prompted with a few-shot design to return numerical likelihood ratio estimates. Agreement with reported likelihood ratios was assessed using Bland-Altman analysis for mean bias and multiplicative limits of agreement.

*Results*

We compiled 700 literature-reported likelihood ratios for 30 conditions. Most were for signs or symptoms (59%), historical elements (19%), or test results (16%). Likelihood ratios clustered near 1 (geometric mean 1.21, interquartile range 0.7 to 2.2). All models showed negligible mean bias. GPT-5 showed closest agreement with 95% limits of agreement spanning 0.26x to 3.7x of reported values (vs. 0.23x to 4.28x for o3 and 0.23x to 4.53x for GPT-4o; *P* < .0001 for each comparison).

*Conclusions*

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. Further exploration is warranted concerning 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.

# 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 disease1,2. 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 studies1,3–6.

Clinicians often reason by clinical gestalt, relying on intuition and experience7,8. 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, consistency, and can be used to refine clinical gestalt9.

Recent advances in 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.

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 LRs 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 that 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 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) 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 literature2,12.

### Comparator Likelihood Ratios

On August 25, 2025, we generated comparator likelihood ratios (LRLLM) for all findings listed on theNNT 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 prompted defined the LR as ‘P(finding | diagnosis) / P(finding | not diagnosis)’, gave qualitative LR strength descriptions12, and required an only-numerical 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 analysis on log-transformed values, as strength of evidence is additive on the log scale2,13. 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.

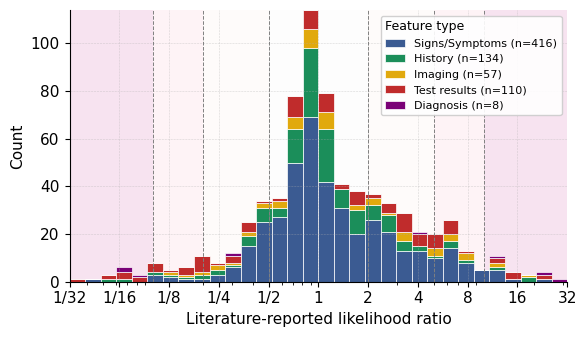
We compared models using paired t-test for mean differences (bias) and the Morgan-Pittman test for differences in the width (variance) of the limits of agreement. Subgroup analyses were conducted by information type (historical element, symptom/sign, exam finding, test result, or diagnostic adjudication), 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)2,12 was assessed using Cohen’s Kappa with quadratic weights, which penalizes large disagreements more heavily and approximates squared-error on the underlying likelihood ratio scale14,15. 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’s 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%). Diagnoses 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 categories2,12. 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.

**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 log 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.

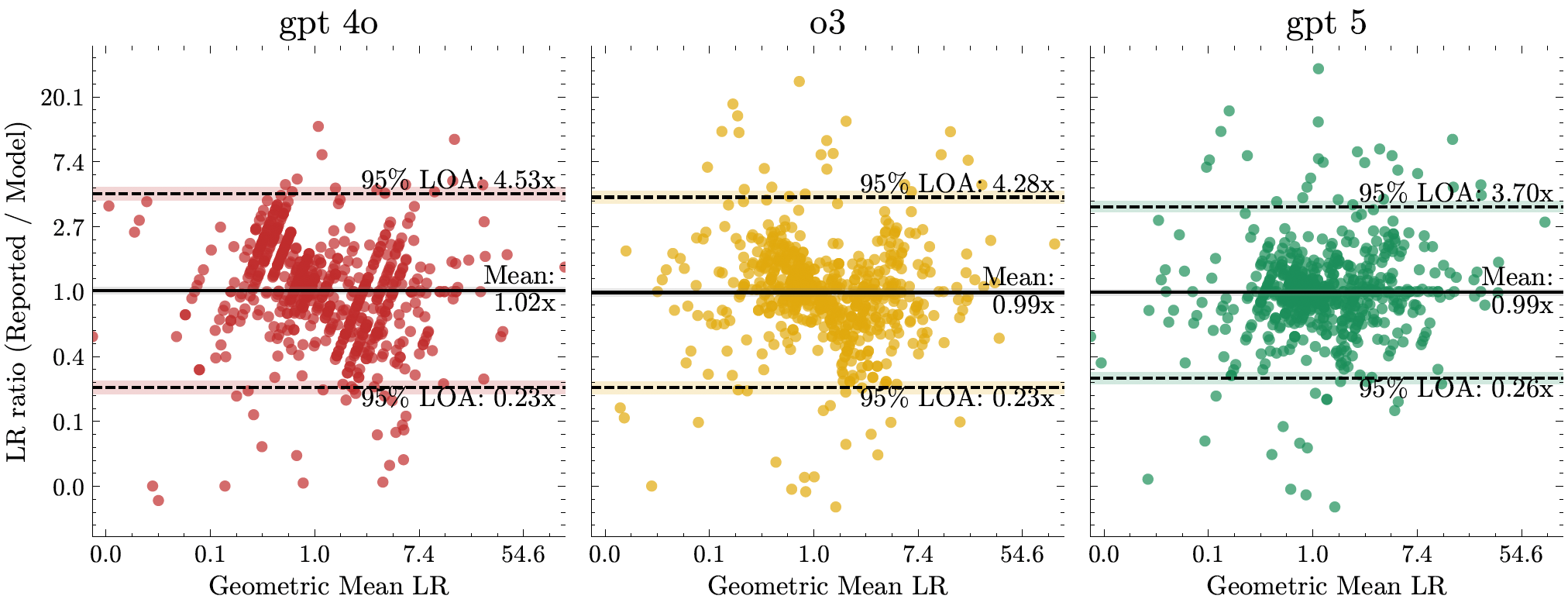


Figure 3 shows the limits of agreement by finding type. Estimates of the strength of evidence followed similar patterns across models. Estimates of the importance of sign/symptoms, historical elements, and imaging findings were all similarly accurate, while test results agreed less closely with reported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

**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).

A screenshot of a graph

AI-generated content may be incorrect.

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% CI 0.728-0.821), followed by o3 (0.745, 95% CI 0.702 – 0.789), and GPT4o (0.734, 95% CI 0.691 – 0.778).

# 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 must be taken from the literature, and therefore could potentially be included in the training data. Though LLMs generally do a poor job memorizing information16, 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 evaluated17 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

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

2. 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

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

4. Watsjold BK, Ilgen JS, Regehr G. An Ecological Account of Clinical Reasoning. *Acad Med*. 2022;97(11S):S80. doi:10.1097/ACM.0000000000004899

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

6. Salameh JP, Bossuyt PM, McGrath TA, et al. Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist. Published online August 14, 2020. doi:10.1136/bmj.m2632

7. 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

8. 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

9. 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

10. 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

11. 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

12. 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

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

14. 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

15. 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

16. Morris JX, Sitawarin C, Guo C, et al. How much do language models memorize? *arXiv*. Preprint posted online June 18, 2025. doi:10.48550/arXiv.2505.24832

17. 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

# 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

LR’s are presented on the logarithmic scale.



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



