Our study shows that LLMs can estimates diagnostic LRs with negliglble mean bias and usable precision, with the tightest agreement coming from the newest model, GPT-5 (95% limits of agreement ≈ 0.26×–3.70×). The agreement varied by finding type, with labs showing less tight agreement than history, signs, and imaging. Models showed substantial qualitative category agreement (*κ≈0.78)*. These results indicate that LLM-derived LRs may be useful to scaffold Bayesian reasoning in situations where empirical LRs are absent, inapplicable, or uncertain.

**1) What is known; what this study adds (Context & interpretation)**

* **Known**
  + Bayesian diagnosis requires LRs, but robust, context‑specific LRs exist for only a fraction of clinical findings; most published LRs cluster near 1, limiting impact.
  + Prior AI work suggests human–AI complementarity in diagnostic reasoning; however, whether LLMs can supply the *weights of evidence* (LRs) themselves had not been evaluated at scale.
* **What this study adds**
  + **First large evaluation** (700 finding–condition pairs, 30 conditions) showing **low bias** and **bounded dispersion** of LLM‑estimated LRs versus literature values, improving with newer models.
  + **Heterogeneity mapped**: relative tightness for history/signs/imaging; looser for labs; model‑specific directional biases; consistent **shrinkage toward the null** (attenuation of extremes).
  + **Educational and workflow bridge**: with uncertainty bands, LLM‑LRs can render probabilistic reasoning explicit for trainees and support calibrated updates in practice when empirical LRs are unavailable.

**2) Comparison with prior work (How does this fit?)**

* **Evidence gap addressed:** Empirical LR availability and transportability are limited; our results suggest LLMs can *approximate* missing LRs sufficiently to support Bayesian updates, especially when aggregated across multiple findings.
* **Relation to AI in diagnosis:** Extends prior studies of LLM reasoning by focusing on a **quantitative unit** (LR) rather than free‑text advice, enabling clearer auditing, teaching, and threshold‑based decisions.
* **Interpretation:** Mild shrinkage likely reflects model averaging over heterogeneous sources and may be protective against over‑confident extremes; nevertheless, extremes should be reviewed or sourced from primary data.

**3) Strengths**

* **Design/measurement:** Pre‑specified, reproducible evaluation (Bland–Altman on log LRs; κ for category agreement); multiple coverage levels reported.
* **Breadth & comparators:** 700 literature LRs across 30 conditions; three LLM families spanning generations.
* **Transparency:** Defined prompts, constrained outputs, and code availability support replication and external benchmarking.

**4) Limitations (be explicit; pair with their implications)**

* **Reference standard dependence:** Using literature LRs (e.g., TheNNT) risks inherited bias and unknown spectrum/context; we did not re‑adjudicate source studies. *Implication:* treat LLM‑LRs as **scaffolded estimates**; prefer empiric values where high‑stakes decisions hinge on a single finding.
* **Potential training overlap:** Some reference LRs may be present in model pretraining; the performance gradient across models and shrinkage patterns argue against pure memorization but do not exclude it.
* **Context mismatch (spectrum effects):** The model implicitly estimates a “generic” LR; transportability to a specific setting may require adjustment via local base rates or recalibration.
* **Precision limits:** Even best‑case 95% LoA (~0.26×–3.70×) warrant **aggregation across findings** and exposure of uncertainty; avoid relying on single extreme LRs without corroboration.
* **No retrieval‑augmented inference tested:** Estimates were generated without live literature retrieval; combining LLM estimation with retrieval may improve validity but complicates benchmark comparability.

**5) Implications (Are the findings important? What can we do with them?)**

**5a) Clinical practice**

* Use LLM‑LRs to **standardize** post‑test probability updates when empirical LRs are missing; **expose uncertainty bands**; **aggregate** multiple modest LRs with calibrated pre‑test probabilities; **flag extremes** for manual review.
* Embed LR suggestions into **threshold‑based pathways** (test/treat/watchful waiting), with local priors and action thresholds visible to clinicians.

**5b) Medical education**

* Deploy LLM‑LRs as **inspectable, auditable weights of evidence** during case discussions; support the **cognitive apprenticeship** by making updates explicit and normalizing appropriate diagnostic inertia when evidence is negligible.
* Use in simulation/spaced‑repetition to train **belief updating** and reduce availability/anchoring biases.

**5c) Quality improvement / systems**

* Instrument diagnostic workflows with **explicit thresholds and LR usage**; audit variation in probability updates and actions across teams; monitor reliance on single extreme LRs; iterate toward **evidence‑proportionate** decisions.

**6) Future research (Specific, actionable—not “more research is needed”)**

* **Prospective trials**: Do LLM‑LR–augmented pathways improve **diagnostic accuracy, calibration, and time‑to‑diagnosis** vs. usual care?
* **Retrieval‑augmented LR estimation**: Combine estimation with **source retrieval and citation**; measure calibration and trust impact.
* **Contextualization**: Methods to **condition LRs on setting/spectrum** (e.g., hierarchical or generalized‑Bayes recalibration using local outcome data).
* **Uncertainty communication**: Evaluate interfaces that present **intervals/LoA** and show how multiple findings compound in odds space.
* **Bias/fairness**: Assess subgroup performance (age, sex, comorbidity) to detect systematic misestimation.
* **Benchmarking**: Maintain and expand a **public LR benchmark** with periodic re‑runs across model generations.

**7) Conclusion (one‑liner; citable)**

* *LLMs can approximate diagnostic likelihood ratios with low bias and bounded dispersion, enabling practical, auditable Bayesian updates—best used as calibrated scaffolds that aggregate across findings, expose uncertainty, and are tailored to local context.*

**Notes on structure (for your internal checklist)**

* Start with the **main findings in ≤3 sentences**, mirroring the introduction’s question (hourglass model).
* Keep distinct subsections for **comparison with literature**, **strengths/limitations**, and **implications**, ending with a **concise, citable bottom line**.
* Target **six to seven paragraphs**, ~≤200 words each; avoid repeating results verbatim; emphasize what is new and important for readers’ practice/research.

If you want, I can convert this into a camera‑ready Discussion draft using your house style and slot in your preferred citations inline.

Our findings demonstrate that large language models (LLMs) can estimate likelihood ratios (LRs) 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 diagnostic and educational workflows, particularly in situations where empirical data is limited, outdated, or unavailable.

By serving as a bridge to Bayesian estimates that are “fit for purpose,” LLM-derived LRs can serve as a standardized method for supporting how probabilistic inferences are communicated and improved upon. For the clinician interested in optimizing their diagnostic accuracy, the quantification of diagnostic odds offers a reproducible pathway for AI-human hybridization, reflective practice, and fine-tuning of pre-calibrated action thresholds. For the frustrated trainee who is stumped by an onslaught of ambiguous features in a clinical unknown, LLM-derived LRs can provide validation of clinical uncertainty that is, in fact, irreducible. For the master diagnostician who wishes to de-mystify their clinical gestalt for medical trainees, LLM-derived LRs offer a pathway for more explicit and transparent synthesis of prioritized (or de-prioritized) data inputs.

The traditional approach by which trainees learn the “art” of diagnostic reasoning is called the “cognitive apprenticeship” model.22 A pitfall of this approach is how much it hinges on well-trained faculty who boast the skills of *both* a master diagnosticianand an educator. Clinician educators also have to “think out loud” to make their uncertainty tolerance and train of thought transparent. After all, the ability to consistently make an accurate diagnosis does not help a team of learners if these nuanced skills are ineffectively taught or poorly communicated. For trainees who may not share the same priors (e.g. past clinical experiences or baseline assumptions), passing on an “embrace of uncertainty” can be especially challenging. Likewise, the educator who models their availability bias, recency bias, or anchoring bias only amplifies these diagnostic pitfalls, further perpetuating habits of diagnostic inertia, rooted in false assumptions.

LLM-derived LRs can transcend these limitations by making the chain of probabilistic inferences and belief updates more accessible to learners and unskilled faculty alike. AI hybridization offloads the cognitive burden associated with complex mathematical formulas, allowing clinicians to more easily engage in structured Bayesian reasoning. Such a shift benefits clinicians across all stages of training, from early learners developing foundational diagnostic skills to experienced practitioners refining their diagnostic accuracy and consistency. By lowering the point of entry for more routine application of Bayes theorem, LLM-derived LRs can democratize the upskilling of probabilistic inference amongst clinicians across the board. Ultimately, when the cognitive apprenticeship is strengthened, both trainees and patients benefit.

Moreover, coupling generative AI capabilities with databases such as the Number Needed to Treat (NNT) database could create a "living" repository ofLRs, 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 a robust, hybridized approach to Bayesian inference through iterative use of AI-supported diagnostic tools. Put simply, LLM-generated LRs provide a path towards AI-enhanced adaptive practice.23

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 information24, 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 evaluated11 and we could not extract the population of interest from studies that theNNT.com estimates were based on, the LLM was implicitly 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. 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 direct human oversight. 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 LRs 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 the practice of clinical diagnosis 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.

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 associated diseases1,4. However, empirically derived LRs exist only for a limited subset of clinical findings, conditions, and contexts, because estimating them requires collation of resource-intensive diagnostic test accuracy studies across variable contexts1,2,25,26.

The traditional approach by which clinicians reason is through clinical gestalt, relying on intuition, heuristics, and pattern recognition7,27. While efficient, gestalt-based reasoning can be biased, inconsistent, and limited by an insufficient scope of personal experiences. When feasible, quantitative reasoning using likelihood ratios (LRs) provides a normative standard that improves diagnostic accuracy, consistency, and can be used to refine clinical gestalt6. This hybrid system in which humans and AI complement one another in classification tasks, like clinical diagnosis, is not a new idea28. In fact, Bayesian modeling shows promise for more generalized classification problems because of the diverse data processing strategies by which human and AI classifiers operate. Whether large language models (LLMs) can generate LRs to specifically improve upon human diagnostic accuracy within a Bayesian inferential reasoning framework is not yet known.

Recent advances in *generative* artificial intelligence, particularly LLMs, offer new opportunities to enhance clinical decision-making and medical education18. 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)15.

This capability raises the possibility that LLMs could reliably estimate diagnostic LRs, potentially overcoming a key barrier to broader application of quantitative reasoning in clinical practice. Notably, the accuracy of LLM-estimated LRs has not been previously explored. In this study, we aimed to evaluate the capacity of contemporary LLMs to accurately estimate diagnostic LRs. Specifically, we compared LLM-generated LRs with empirically reported values from the existing literature. If the accuracy of these models in estimating known LRs was found to be acceptable, the natural next step would be to consider situations in which under-investigated clinical findings (e.g. unknown LRs) could be tested in simulated (or real) clinical and training contexts.

Can key clinical features be “digitalized” within a Bayesian analytical reasoning framework to augment our current approach to clinical diagnosis? How will this affect medical education and skill development of learners? We explore these questions after testing current LLM capabilities. Our work adds to the work of Goh et al16 and others29, who found that the reasoning of LLMs is of merit and deserves further exploration with respect to physician-LLM collaboration in clinical decision-making. By incorporating cutting-edge models, we address limitations of past works13 and examine the interface of such models with clinician decision-making.