Position Paper On Diagnostic Uncertainty Estimation from Large Language Models: Next-Word Probability Is Not Pre-test Probability

Yanjun Gao[1,2*] **Skatje Myers**[2] **Shan Chen**[3] **Dmitriy Dligach**[4]
Timothy A Miller[5] **Danielle Bitterman**[6] **Guanhua Chen**[2]
Anoop Mayampurath[2] **Matthew Churpek**[2] **Majid Afshar**[2]
[1]University of Colorado Anschutz [2]University of Wisconsin-Madison
[3]Harvard Medical School [4] Loyola University Chicago
[5] Boston Childrens Hospital [6] Dana Farber Cancer Institute
*yanjun.gao@cuanschutz
{symers, mchurpek, mafshar}@medicine.wisc.edu
{Shan.Chen, Timothy.Miller}@childrens.harvard.edu
ddligach@luc.edu, DBITTERMAN@BWH.HARVARD.EDU
{mayampurath, gchen25}@wisc.edu

Abstract

Large language models (LLMs) are being explored for diagnostic decision support, yet their ability to estimate pre-test probabilities, vital for clinical decision-making, remains limited. This study evaluates two LLMs, Mistral-7B and Llama3-70B, using structured electronic health record data on three diagnosis tasks. We examined three current methods of extracting LLM probability estimations and revealed their limitations. We aim to highlight the need for improved techniques in LLM confidence estimation.

1 Introduction

Diagnosis in medicine is inherently complex and involves estimating the likelihood of various diseases based on a patient's presentation. This process requires integrating baseline information to establish pre-test probabilities during the initial hypothesis generation for a diagnosis, followed by iterative refinement as diagnostic test results become available (Sox et al., 1989; Bowen, 2006) (Figure 1). Typically, clinicians rely on medical knowledge, pattern recognition and experience, enabling quick hypothesis generation of the initial diagnosis. However, this process is prone to cognitive biases, which can lead to diagnostic errors(Saposnik et al., 2016). Analytic thinking, a more evidence-based process, is time-consuming and often impractical in fast-paced clinical environments. Although clinicians are taught to estimate a pre-test probability and apply test sensitivity and specificity, cognitive biases and heuristic-based thinking often lead to under- and overestimation of the pre-test probability and subsequent misdiagnoses (Rodman et al., 2023).

The integration of Large Language Models (LLMs) in diagnostic decision support systems has garnered significant interest in addressing these challenges. Recent advancements, particularly with models like GPT-4, have demonstrated that LLMs can rival clinicians in generating differential diagnoses (Kanjee et al., 2023; Savage et al., 2024a). However, LLMs often fail to explicitly convey uncertainty in the estimated probability of a diagnosis in their outputs. This is crucial in medicine; for example, an LLM might suggest an initial diagnosis of pneumonia, yet, a 20% probability of pneumonia may have vastly different implications for a clinician compared to a 90% probability. While GPT-4 has shown some potential for improvement over clinicians in predicting pre-test

GenAI for Health Workshop @ NeurIPS 2024, Vancouver.

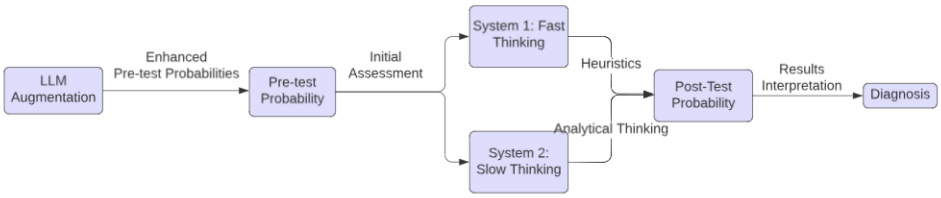

Figure 1: Process map in generating a diagnosis with the role of LLMs to augment human diagnostic reasoning.

probability of certain conditions, overall performance is still suboptimal (Rodman et al., 2023; Kanjee et al., 2023). LLMs are not designed as classifiers that output probability distributions over specific outcomes; instead, they produce probability distributions over sequences of tokens. This raises the research question of how to map these token sequences to clinically meaningful probabilities, particularly for pre-test or post-test diagnosis probability estimation. Addressing this gap is crucial to avoid potential misinterpretations and mitigate the risk of automation bias in clinical settings.

The concept of uncertainty estimation for the generated text in LLMs is rooted in information theory with entropy, which measures the uncertainty of a probabilistic distribution to get next-word prediction. This process involves training the models to align their predictions with the actual distribution of the language they are trained on, resulting in the generation of convincing and coherent natural language. Existing literature investigates methods for extracting uncertainty estimation from LLMs, including token probabilities and verbalized probabilities (confidence)(Kapoor et al., 2024; Geng et al., 2024). However, LLMs are known to suffer from the problem of unfaithful generation, where their outputs do not always accurately reflect their underlying knowledge or reasoning (Hager et al., 2024; Turpin et al., 2024). Further, while previous work (Savage et al., 2024b) shows that sample consistency with embeddings could serve as uncertainty proxies, they evaluated on question-answering tasks, which is different from the real-world electronic health records setting. While LLMs may have general knowledge about disease prevalence from the pretraining corpora, such as Wikipedia, it remains uncertain whether they can translate general knowledge into patient-specific diagnostic reasoning and estimate pre-test probabilities, a question this paper aims to investigate.

We aimed to address this gap by evaluating the strengths and limitations of LLMs in pre-test diagnostic probability estimation. We conducted a detailed evaluation of two open-box LLMs: Mistral-7B-Instruct Jiang et al. (2023) and Llama3-70B-chat-hf Touvron et al. (2023), on the task of predicting pre-test diagnostic probabilities. These models were selected because they were available open source and adaptable through instruction-tuning. Unlike previous work exploring LLM medical uncertainty estimation on question-answering tasks or case reports (Saposnik et al., 2016; Hager et al., 2024; Abdullahi et al., 2024), our study was based on a set of annotated structured data in the electronic health records (EHRs) from a cohort of 660 patients at a large medical center in the United States. The task involves binary predictions for Sepsis, Arrhythmia, and Congestive Heart Failure (CHF), with positive class distributions of 43%, 16%, and 11%, respectively. Ground truth diagnoses were annotated by expert physicians through chart reviews (Churpek et al., 2024). The EHR data included vital signs, lab test results, nurse flow-sheet assessments, and patient demographics. We compared our results to an eXtreme Gradient Boosting (XGB) classifier that used the raw structured EHR data as input, representing the state-of-the-art in many clinical predictive applications (Lolak et al., 2023; Govindan et al., 2024). EHR data included vital signs, structured nurse flowsheet assessments (i.e., mental status, mobility, etc.), and lab test results. We subsequently added patient demographics (sex, ethnicity, and race), encoded as categorical variables, to examine if such a setting could improve model performance.

2 Methods of Extracting Pre-test Probabilities from LLMs

This section formulates the task as a binary diagnostic outcome classification using three methods. We benchmarked against XGB using raw features (baseline, Raw Data+XGB), correlating each method. We utilized a table-to-text method to convert structured EHR data into text. Specifically, we began this transformation by creating a template starting with "Hospitalized patient with age XX, systolic blood pressure YY ..." where XX and YY represent the actual values from patient

Hospitalized patient of age *[value]* getting worse has labs and vitals values of systolic blood pressure *[value]* mmHg, diastolic blood pressure *[value]* mmHg, oxygen saturation*[value]* %, body temperature *[value]* celsius degree, ... total protein *[value]*, white blood cell *[value]*. What are the diagnoses for this patient?

Table 1: The template for NARRATIVE serialization method for diagnosis prediction dataset.

representation, as shown in Table 1. We then appended each clinical feature, its corresponding value, and its unit of measurement in the text. We refer readers to our recent paper, which provides a more detailed description of the table-to-text conversion process. (Gao et al., 2024).

Our method of prompt development started with a prompt from previous literature that prompts GPT-4 for confidence estimation (Rodman et al., 2023). We then made modifications according to the task descriptions and format requirements of the LLMs. To ensure that the LLMs would follow the format requirements, we tested the prompts with a few synthetic examples.

Token Logits We prompted the LLM with a detailed description of the patient's condition and directly asked for a binary response: "Does the patient have diagnosis? A. Yes" or "B. No", indicating the presence or absence of sepsis. Specifically, the probability estimation was derived from the logits corresponding to these responses. We applied a softmax function yielding a normalized score for each option. We used a zero-shot setting for both LLMs.

Verbalized Confidence This approach followed the previous study of GPT-4 on diagnostic probability estimation, and we used the same format of prompt (Rodman et al., 2023). This set of prompts posed a more open-ended question to the LLM followed by a narrative description of patient representation: "How likely is it for the patient to have diagnosis?" The LLM will provide a percentage score, which we utilized as the probability of positive diagnosis. This approach allowed us to assess the model's ability to generate and verbalize probability assessments in a natural language format without further binarization of the results. We used zero-setting for both LLMs. Instead of applying a cut-off to categorize the predictions, we evaluated the raw probability estimates directly using AUROC scores, Pearson Correlation and Expected Calibration Errors (ECE).

Feature-based Calibrators For the feature-based calibrators, we applied max pooling to the last layer of the LLMs, forming the embedding representation in 4096 and 8192 feature spaces for Mistral and Llama3-70B, respectively. These embeddings were then fed into an XGB classifier. All XGB classifiers were trained using a five-fold cross-validation on the 660 patient data.

Training details All experiments involving XGB classifiers were trained under a 5-fold cross-validation setting. On each fold, we employed grid search for parameters using another five-fold to select the best hyperparameters. The parameter grid included $n_estimators$ set to [50, 100, 250, 500], max_depth ranging from [2, 5, 10, 15, 20], $learning\ rate$ values of [0.005, 0.01, 0.05, 0.1], and min_child_weight values of [1, 2, 3].

3 Results

Figure 2 illustrates the results of the AUROC from LLMs to predict Sepsis, Arrhythmia, and CHF. The LLM Embedding+XGB method consistently outperformed the other LLM-based methods. Particularly for Sepsis, it achieved nearly the same AUROC score as the baseline Raw Data+XGB. The Token Logits (mean AUROC: 49.9 with 95% CI: 47.8-51.9) and Verbalized Confidence (mean AUROC: 50.9 with 95% CI: 48.7-53.1) methods exhibited marginal performance, generally not surpassing the baseline XGB classifier. The inclusion of demographic variables (sex, race, ethnicity) changed the AUROC scores, for instance, by as much as 7.22 for Mistral embedding on Sepsis prediction (71.1 on default setting vs 63.9 on ethnicity). However, the direction and consistency of these changes varied depending on the specific context and data included.

Figure 3 reports the Pearson correlation coefficients between the predicted probabilities from LLM-based uncertainty estimation methods and those from the XGB classifier for three diagnoses across different demographics. When correlating the LLMs' positive class probabilities with the baseline results, the token logits and verbalized confidence methods had more variable correlations, often no correlation or negative correlation, suggesting less alignment with the baseline XGB predicted probabilities. On the contrary, the LLM embedding+XGB method consistently showed strong positive

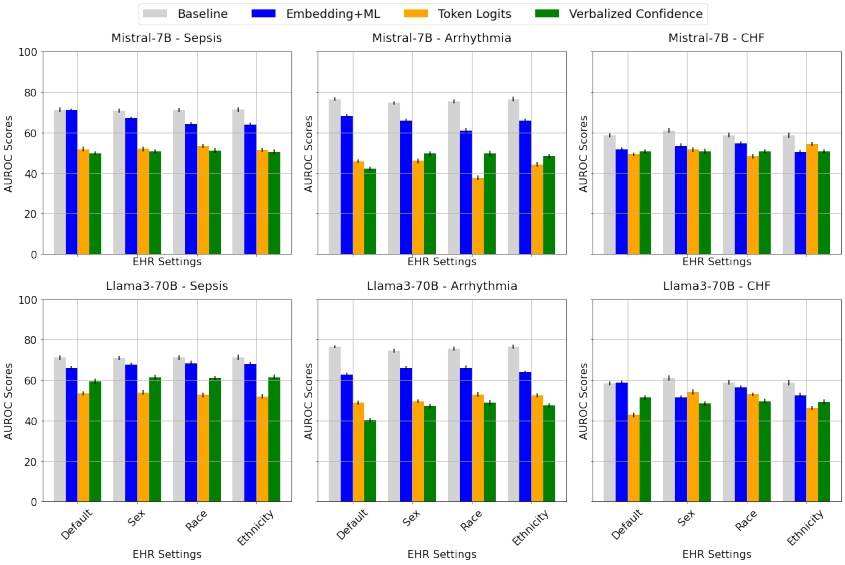

Figure 2: Area under the receiver operating characteristic curve (AUROC) scores from both LLMs using different EHR demographics input settings, across the diagnoses prediction of Sepsis, Arrhythmia, and Congestive Heart Failure (CHF).

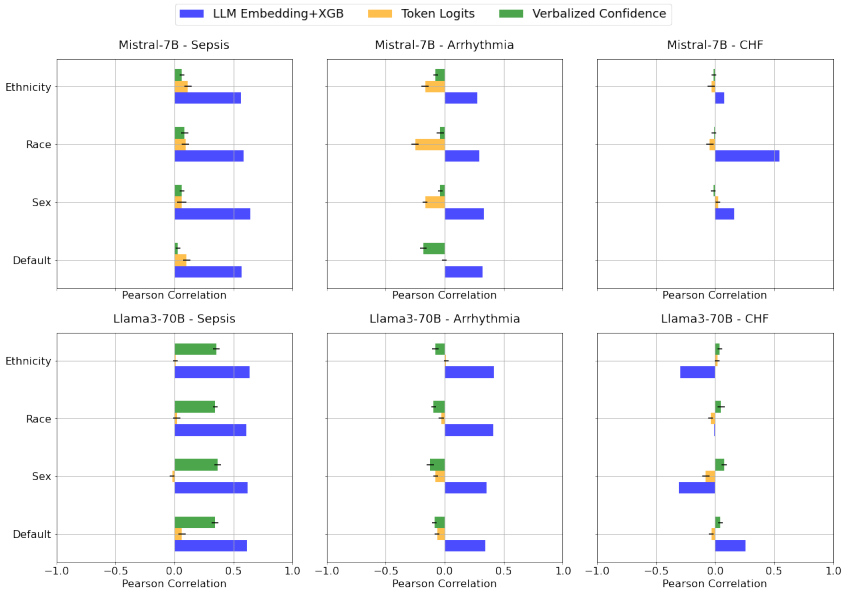

Figure 3: Pearson correlations between the LLM predicted pre-test probabilities and the baseline model (Raw Data+XGB) predicted probabilities.

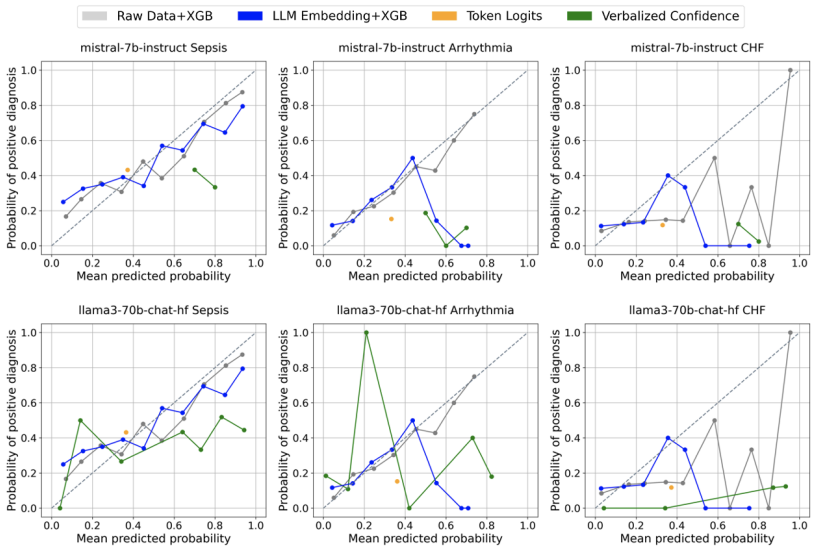

Figure 4: Calibration curves on the four probability estimation methods, using Mistral-7B-Instruct and Llama3-70B-Chat-hf on Default patient demographic setting.

Method	**None** Sepsis, Arrhy., CHF	**Sex** Sepsis, Arrhy., CHF	**Race** Sepsis, Arrhy., CHF	**Ethnicity** Sepsis, Arrhy., CHF
		Baseline		
Raw+XGB	0.09, 0.02, 0.10	0.07, 0.02, 0.10	0.09, 0.02, 0.10	0.10, 0.02, 0.11
		Mistral-7B-Instruct		
Token Logits	0.07, 0.18, 0.21	0.13, 0.21, 0.25	0.13, 0.21, 0.25	0.13, 0.21, 0.25
Verb. Conf.	0.27, 0.43, 0.59	0.31, 0.35, 0.58	0.27, 0.35, 0.59	0.27, 0.36, 0.58
Emb+XGB	0.09, 0.06, 0.13	0.14, 0.09, 0.09	0.11, 0.11, 0.11	0.06, 0.06, 0.11
		Llama3-70B-Chat		
Token Logits	0.07, 0.21, 0.25	0.06, 0.22, 0.25	0.07, 0.21, 0.25	0.06, 0.22, 0.25
Verb. Conf.	0.28, 0.25, 0.77	0.24, 0.04, 0.01	0.24, 0.04, 0.04	0.24, 0.04, 0.03
Emb+XGB	0.11, 0.05, 0.07	0.06, 0.09, 0.17	0.06, 0.04, 0.09	0.06, 0.06, 0.11

Table 2: ECE Results across the four methods for pre-test probabilities estimation methods, over three diagnosis prediction tasks with patient demographic settings.

correlations across all tasks and settings with the baseline XGB classifier, indicating its predictions were closely aligned with the baseline XGB classifier.

Figure 4 presents the calibration curve of all models on the default setting (no demographic variable). Poor calibration was observed, especially from Token Logits and Verbalized Confidence. Probabilities predicted by the Token Logits always fell between ranges of 0.323-0.337. Table 2 further highlights the ECE scores of each method, with notable differences observed across the various biases and diagnoses. For instance, the Verbalized Confidence (Verb. Conf) method tends to exhibit higher ECE values, indicating poorer calibration, especially in the CHF prediction task, while Raw+XGB generally shows more consistent performance across different demographic settings. These results reflect the lack of robustness of these methods for uncertainty estimation, highlighting a significant gap in uncertainty estimation for medical decision-making. Although these methods are used in literature to assess uncertainty in predicting the next word (Xiong et al., 2024), predicting pre-test probabilities requires an understanding of risk based on real-world patient data and disease prevalence, knowledge that LLMs often lack. The introduction of demographic variables complicates the predictive power of these models further due to inherent biases present in LLMs, which may not be trained on a diverse set of patient characteristics, making them susceptible to social biases.

4 Discussion and Conclusion

The LLM Embedding+XGB method demonstrated competitive performance compared to the state-of-the-art XGB baseline classifier under specific conditions, such as Sepsis, and exhibited the strongest correlation among the methods tested. However, this result is not surprising given that both methods rely on training a classifier. In contrast, purely LLM-based methods, such as Token Logits (next-word probability) and Verbalized Confidence, were found to be unreliable for risk estimation. Their performance, evaluated through AUROC scores, Pearson Correlation, and calibration curves, deteriorated significantly when diagnosing conditions with lower prevalence, raising concerns about the accuracy of pre-test probabilities derived from these models. The results were consistent across both the Mistral-7B and Llama3-70B models. Additionally, results varied with different demographic characteristics, reinforcing existing concerns about bias in LLMs (Zack et al., 2024). While the LLM Embedding+XGB method showed promise in generating pre-test probabilities, overall, LLM-based probability estimation methods did not achieve the same level of performance as raw tabular data in an XGB model. This underscores the necessity for further optimization of LLM methods to produce uncertainty estimations that align more closely with established and reliable methods.

Overall, our findings demonstrate the inability of LLMs to provide reliable pre-test probability estimations for specific diseases and highlight the need for improved strategies to incorporate numeracy into diagnostic decision support systems and reduce the impact of bias on LLM performance. This remains a major gap to fill before we can enter a new era of diagnostic systems that integrate LLMs to augment healthcare providers in their diagnostic reasoning. To address these limitations, future work should explore hybrid approaches that integrate LLMs with numerical reasoning modules or calibrated embeddings, enhancing their capacity for accurate uncertainty estimation, particularly for low-prevalence conditions. Additionally, bias mitigation strategies, such as layer-wise probing and targeted regularization, could help ensure fairer predictions across demographic groups. These advancements would bring us closer to safe and effective diagnostic systems that integrate LLMs to support clinicians in clinical decision-making.

Competing Interests

There is no competing interests to be declared.

Funding Acknowledgments

This work is supported by the National Library of Medicine (NLM) under award R00LM014308 02 (PI: Gao), NLM award R01LM012973 05 (PI: TM, DD, MA), National Heart, Lung, and Blood Institute (NHLBI) under award R01HL157262 04 (PI: MC), NIH-USA awards U54CA274516 01A1 and R01CA294033 01 (PI: DB; SC), NSF DMS-2054346 and the University of Wisconsin School of Medicine and Public Health through the Wisconsin Partnership Program (Research Design Support: Protocol Development, Informatics, and Biostatistics Module, PI: GC).

References

Abdullahi, T., Singh, R., Eickhoff, C., et al. (2024). Learning to make rare and complex diagnoses with generative ai assistance: qualitative study of popular large language models. *JMIR Medical Education*, 10(1):e51391.

Bowen, J. L. (2006). Educational strategies to promote clinical diagnostic reasoning. *New England Journal of Medicine*, 355(21):2217–2225.

Churpek, M. M., Ingebritsen, R., Carey, K. A., Rao, S. A., Murnin, E., Qyli, T., Oguss, M. K., Picart, J., Penumalee, L., Follman, B. D., Nezirova, L. K., Tully, S. T., Benjamin, C., Nye, C., Gilbert, E. R., Shah, N. S., Winslow, C. J., Afshar, M., and Edelson, D. P. (2024). Causes, diagnostic testing, and treatments related to clinical deterioration events among high-risk ward patients. *medRxiv [Preprint]*. PMID: 38370788; PMCID: PMC10871454.

Gao, Y., Myers, S., Chen, S., Dligach, D., Miller, T. A., Bitterman, D., Churpek, M., and Afshar, M. (2024). When raw data prevails: Are large language model embeddings effective in numerical data

representation for medical machine learning applications? In *Findings of the Empirical Methods in Natural Language Processing: EMNLP 2024*.

Geng, J., Cai, F., Wang, Y., Koeppl, H., Nakov, P., and Gurevych, I. (2024). A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595.

Govindan, S., Spicer, A., Bearce, M., Schaefer, R. S., Uhl, A., Alterovitz, G., Kim, M. J., Carey, K. A., Shah, N. S., Winslow, C., et al. (2024). Development and validation of a machine learning covid-19 veteran (covet) deterioration risk score. *Critical Care Explorations*, 6(7):e1116.

Hager, P., Jungmann, F., Holland, R., Bhagat, K., Hubrecht, I., Knauer, M., Vielhauer, J., Makowski, M., Braren, R., Kaissis, G., et al. (2024). Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, pages 1–10.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Kanjee, Z., Crowe, B., and Rodman, A. (2023). Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *Jama*, 330(1):78–80.

Kapoor, S., Gruver, N., Roberts, M., Collins, K., Pal, A., Bhatt, U., Weller, A., Dooley, S., Goldblum, M., and Wilson, A. G. (2024). Large language models must be taught to know what they don't know. *arXiv preprint arXiv:2406.08391*.

Lolak, S., Attia, J., McKay, G. J., and Thakkinstian, A. (2023). Comparing explainable machine learning approaches with traditional statistical methods for evaluating stroke risk models: Retrospective cohort study. *JMIR cardio*, 7:e47736.

Rodman, A., Buckley, T. A., Manrai, A. K., and Morgan, D. J. (2023). Artificial intelligence vs clinician performance in estimating probabilities of diagnoses before and after testing. *JAMA Network Open*, 6(12):e2347075–e2347075.

Saposnik, G., Redelmeier, D., Ruff, C. C., and Tobler, P. N. (2016). Cognitive biases associated with medical decisions: a systematic review. *BMC medical informatics and decision making*, 16:1–14.

Savage, T., Nayak, A., Gallo, R., Rangan, E., and Chen, J. H. (2024a). Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digital Medicine*, 7(1):20.

Savage, T., Wang, J., Gallo, R., Boukil, A., Patel, V., Safavi-Naini, S. A. A., Soroush, A., and Chen, J. H. (2024b). Large language model uncertainty proxies: discrimination and calibration for medical diagnosis and treatment. *Journal of the American Medical Informatics Association*, page ocae254.

Sox, H., Stern, S., Owens, D., Abrams, H. L., et al. (1989). The use of diagnostic tests: A probabilistic approach. In *Assessment of Diagnostic Technology in Health Care: Rationale, Methods, Problems, and Directions: Monograph of the Council on Health Care Technology*. National Academies Press (US).

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Turpin, M., Michael, J., Perez, E., and Bowman, S. (2024). Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36.

Xiong, M., Hu, Z., Lu, X., LI, Y., Fu, J., He, J., and Hooi, B. (2024). Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations*.

Zack, T., Lehman, E., Suzgun, M., Rodriguez, J. A., Celi, L. A., Gichoya, J., Jurafsky, D., Szolovits, P., Bates, D. W., Abdulnour, R.-E. E., et al. (2024). Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1):e12–e22.