Contents lists available at IJCHML
International Journal of Computational Health and Machine Learning
Journal Homepage: http://www.ijchml.com/
Volume 4, No. 2, 2026

IJCHML
INTERNATIONAL JOURNAL OF COMPUTATIONAL HEALTH & MACHINE LEARNING

Benchmarking Reliability Metrics for Large Language Models in Computational Health Applications

Anil Jain[1], **Rahul Reddy**[2]
[1] *Department of Health Informatics, Savitribai Phule Pune University*
[2] *Department of Health Informatics, Anna University*

ARTICLE INFO

Received: 05/03/2026
Revised: 06/16/2026
Accepted: 07/01/2026

Keywords:

Large Language Models, Reliability Metrics, Computational Health, Benchmarking, Uncertainty Quantification, Calibration, Clinical Decision Support

ABSTRACT

The deployment of large language models (LLMs) in computational health applications demands rigorous evaluation frameworks capable of quantifying not only predictive accuracy but also output reliability, calibration, and uncertainty. Despite rapid advances in model capability, the absence of standardized benchmarking protocols for reliability metrics represents a critical gap that impedes safe clinical translation. This work addresses that gap through a systematic comparative study of reliability evaluation methodologies applied to LLMs across diverse health-related natural language processing tasks.

We introduce a comprehensive benchmarking suite encompassing six reliability dimensions: calibration error, selective prediction performance, distributional shift robustness, semantic consistency, factual grounding fidelity, and conformal coverage guarantees. Formally, for a model producing probability estimates $\hat{p}$ over label space $\mathcal{Y}$, we define reliability as a composite functional $\mathcal{R}(\hat{p}, y) = \sum_{k=1}^{K} \lambda_k \, \phi_k(\hat{p}, y)$, where ϕ_k denotes each constituent metric and λ_k its associated weight determined through task-specific sensitivity analysis.

Experiments are conducted across clinical note summarization, medical question answering, diagnostic coding, and adverse event extraction benchmarks, evaluating seven state-of-the-art LLMs under both zero-shot and fine-tuned regimes. Our results reveal substantial inter-metric disagreement, demonstrating that accuracy-centric evaluation systematically overestimates model trustworthiness in high-stakes health contexts. Calibration and conformal prediction metrics expose failure modes invisible to standard performance measures.

This study provides actionable recommendations for practitioners and regulators, advocating for multi-dimensional reliability reporting as a prerequisite for responsible LLM deployment in healthcare. All benchmarking code and evaluation protocols are released publicly to facilitate reproducibility and community adoption of standardized reliability assessment practices.

1. Introduction

The rapid proliferation of large language models (LLMs) across diverse scientific and professional domains has ushered in a new era of computational intelligence, with particularly transformative implications for the field of computational health. From clinical decision support and medical question answering to genomic data interpretation and drug discovery, LLMs have demonstrated remarkable capabilities in processing and synthesizing complex biomedical information [6, 23]. However, the deployment of these systems in high-stakes healthcare environments introduces a fundamental challenge that transcends mere performance metrics: the question of *reliability*. A model that achieves high aggregate accuracy on a benchmark yet fails unpredictably on edge cases, or that expresses unwarranted confidence in erroneous outputs, poses substantial risks to patient safety and clinical integrity [7, 11]. It is therefore imperative that the research community develop, standardize, and rigorously evaluate the metrics by which the reliability of LLMs in computational health applications is assessed.

The concept of reliability in the context of LLMs encompasses a multidimensional spectrum of properties, including calibration, consistency, robustness to distributional shift, uncertainty quantification, and factual faithfulness [12, 26]. Unlike general-purpose benchmarks that prioritize accuracy and fluency, reliability-oriented evaluation frameworks must grapple with the inherent stochasticity of generative models, the epistemic uncertainty arising from incomplete training data, and the aleatoric uncertainty intrinsic to ambiguous clinical queries [1, 9]. Despite the growing recognition of these challenges, the field currently lacks a unified, systematic benchmark specifically designed to evaluate reliability metrics for LLMs in computational health contexts. This paper addresses this critical gap by proposing and empirically validating a comprehensive benchmarking framework that encompasses a curated suite of reliability metrics, diverse health-domain datasets, and a rigorous evaluation protocol applicable to both proprietary and open-source LLMs.

1.1. The Imperative of Reliability in Computational Health

The integration of LLMs into computational health workflows is no longer a speculative endeavor but an emerging clinical reality. Systems such as Med-PaLM 2, GPT-4, and their successors have been evaluated on standardized medical licensing examinations, clinical note summarization tasks, and diagnostic reasoning benchmarks, often achieving performance levels comparable to or exceeding those of human clinicians on narrow, well-defined tasks [8, 17]. Yet, performance on these tasks, while informative, does not fully capture the operational reliability required for safe clinical deployment. A model may correctly answer the majority of questions in a multiple-choice format while simultaneously producing confidently stated, factually incorrect free-text responses in open-ended clinical scenarios—a phenomenon colloquially described as "hallucination" but more precisely characterized as a failure of epistemic calibration [14, 24].

The consequences of unreliable LLM outputs in health settings are not merely academic. Erroneous drug dosage recommendations, misidentified contraindications, or incorrectly summarized patient histories can contribute to adverse clinical events with direct patient harm [3, 22]. Furthermore, the opacity of LLM reasoning processes—particularly in transformer-based architectures with billions of parameters—renders post-hoc error attribution exceedingly difficult, compounding the systemic risk of deployment without robust reliability safeguards [13, 19]. Regulatory bodies, including the United States Food and Drug Administration (FDA) and the European Medicines Agency (EMA), have begun to grapple with frameworks for the oversight of AI-enabled clinical tools, with reliability and uncertainty quantification emerging as central requirements for approval pathways [10, 15]. This regulatory landscape further underscores the urgency of establishing scientifically grounded reliability benchmarks.

1.2. Limitations of Existing Evaluation Paradigms

The current landscape of LLM evaluation in biomedical and clinical domains is dominated by accuracy-centric benchmarks such as MedQA [6], PubMedQA, MedMCQA, and BioASQ, which assess factual recall and reasoning ability through structured question-answering formats. While these benchmarks have proven invaluable for tracking progress in biomedical language understanding, they share a common and consequential limitation: they evaluate *what* a model knows rather than *how confidently and consistently* it knows it [16, 20]. A model that answers a medical question correctly but assigns equal confidence to both correct and incorrect responses is, from a reliability standpoint, no more trustworthy than a random classifier—yet current benchmarks would reward it identically.

Beyond the issue of confidence calibration, existing paradigms are largely insensitive to several other dimensions of reliability that are particularly critical in health contexts. These include temporal consistency (whether a model produces the same answer when queried at different times or with minor paraphrastic variations of the same question), cross-domain robustness (whether performance degrades gracefully when queries shift from common to rare medical conditions), and source attribution fidelity (whether claims made by the

model can be traced to verifiable biomedical literature) [2, 5]. Furthermore, most existing benchmarks are constructed from static datasets that do not reflect the dynamic, evolving nature of medical knowledge, failing to assess a model's ability to recognize and communicate the boundaries of its own training data currency [18, 28]. Table 2 provides a comparative overview of existing evaluation frameworks and their coverage of key reliability dimensions.

1.3. Formal Definitions of Reliability Metrics

A central contribution of this work is the formalization of a coherent set of reliability metrics applicable to LLMs in computational health. We define reliability not as a scalar quantity but as a vector-valued property $\mathbf{R} = (R_{\text{cal}}, R_{\text{con}}, R_{\text{rob}}, R_{\text{unc}}, R_{\text{att}})$, where each component corresponds to a distinct but interrelated dimension of trustworthy model behavior. The most foundational of these is *calibration*, which quantifies the alignment between a model's expressed confidence and its empirical accuracy. For a well-calibrated model, the probability assigned to a correct response should equal the empirical frequency of correctness across queries at that confidence level.

Formally, let $\hat{p}_i$ denote the confidence score assigned by the model to its response for query i, and let $y_i \in \{0, 1\}$ denote the binary correctness of that response. The Expected Calibration Error (ECE), a standard metric for assessing calibration quality, is defined as:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\text{acc}(B_m) - \text{conf}(B_m)\right| \tag{1}$$

where M is the total number of confidence bins, B_m denotes the set of queries whose predicted confidence falls within bin m, N is the total number of queries, $\text{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} y_i$ is the mean accuracy within the bin, and $\text{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i$ is the mean confidence within the bin [4, 21]. A perfectly calibrated model achieves ECE = 0, while larger values indicate systematic over- or under-confidence. In clinical settings, overconfidence is particularly dangerous, as it may cause clinicians to uncritically accept incorrect model outputs; underconfidence, conversely, may lead to underutilization of genuinely helpful model assistance.

Beyond calibration, the dimension of *consistency* captures whether a model produces semantically equivalent responses to semantically equivalent queries. In clinical practice, the same clinical question may be posed by different practitioners with varying terminological preferences, and a reliable system must exhibit robustness to such surface-level variation [25]. We operationalize consistency through a paraphrase-invariance score, computed as the mean semantic similarity between model responses to a query and its K paraphrastic variants, as measured by domain-adapted sentence embedding models. The dimension of *uncertainty quantification* extends beyond point-estimate confidence to encompass the model's ability to express graduated, well-structured uncertainty across its full output distribution, including through mechanisms such as conformal prediction sets and Bayesian approximation techniques [9, 26].

1.4. Scope, Contributions, and Paper Organization

This paper makes several distinct and substantive contributions to the emerging field of LLM reliability evaluation in computational health. First, we introduce a novel benchmarking framework—designated HealthRel-Bench—that operationalizes the five-dimensional reliability vector described above across a diverse corpus of health-domain tasks spanning clinical question answering, medical record summarization, pharmacological reasoning, and epidemiological inference [17, 19]. Second, we conduct an extensive empirical evaluation of twelve state-of-the-art LLMs—including both open-source models (LLaMA-3, Mistral-Medical, BioMedLM) and proprietary systems (GPT-4o, Claude-3.5, Gemini-1.5-Pro)—across all dimensions of the HealthRel-Bench framework, yielding the most comprehensive comparative reliability analysis to date in the computational health domain [7, 10].

Third, we propose a composite Reliability Index (RI) that aggregates the five reliability dimensions into a single, interpretable scalar score through a principled weighting scheme informed by clinical risk stratification principles, enabling direct model comparison and facilitating the selection of appropriate models for specific clinical deployment contexts [2, 15]. Fourth, we investigate the relationships among reliability dimensions, identifying empirically whether improvements in one dimension (e.g., calibration through temperature scaling) systematically trade off against others (e.g., consistency), thereby providing actionable guidance for model developers seeking to optimize the full reliability profile of their systems [11, 28]. Finally, we release HealthRel-Bench as an open, extensible evaluation toolkit to facilitate reproducible reliability assessment by the broader research community.

The remainder of this paper is organized as follows. Section 2 provides a comprehensive review of related work on LLM evaluation, uncertainty quantification, and reliability in biomedical AI. Section 3 details the design and construction of the HealthRel-Bench framework, including dataset curation, metric formalization, and evaluation protocol. Section 4 presents the experimental setup and the twelve LLMs subjected to evaluation.

Table 1: Comparative analysis of existing LLM evaluation benchmarks in biomedical and clinical domains across key reliability dimensions. Checkmarks (✓) indicate explicit coverage; dashes (–) indicate absence of the dimension in the benchmark design.

Benchmark	Accuracy Assessment	Confidence Calibration	Consistency Testing	Uncertainty Quantification	Source Attribution
MedQA [6]	✓	–	–	–	–
PubMedQA	✓	–	–	–	✓
MedMCQA	✓	–	–	–	–
BioASQ	✓	–	–	–	✓
HealthBench [23]	✓	✓	–	–	–
ClinicalBench [8]	✓	✓	✓	–	–
Proposed Framework (Ours)	✓	✓	✓	✓	✓

Section 5 reports and analyzes the empirical results across all reliability dimensions, including ablation studies and correlation analyses. Section 6 discusses the implications of our findings for clinical deployment, regulatory compliance, and future model development. Section 7 concludes with a synthesis of key insights and directions for future research.

1.5. Motivating the Need for a Unified Benchmarking Standard

The absence of a unified benchmarking standard for LLM reliability in computational health has several concrete, deleterious consequences for the research community and for clinical practice. First, the proliferation of ad hoc, task-specific evaluation protocols renders cross-study comparisons methodologically untenable, as models evaluated under different reliability assumptions, using different datasets, and with different metric implementations cannot be meaningfully ranked or compared [16, 20]. This fragmentation impedes the cumulative accumulation of knowledge about which model architectures, training strategies, and post-hoc calibration techniques most effectively enhance reliability in health contexts. Second, the lack of standardization creates perverse incentives for selective reporting, whereby developers may choose evaluation protocols that favor their models' strengths while obscuring reliability weaknesses—a form of evaluation gaming that is particularly concerning in safety-critical domains [5, 18].

Third, and perhaps most consequentially for clinical translation, the absence of standardized reliability benchmarks makes it exceedingly difficult for healthcare institutions and regulatory bodies to make evidence-based decisions about the deployment of specific LLMs for specific clinical tasks [3, 22]. A hospital system considering the deployment of an LLM for clinical documentation assistance, for example, requires not only aggregate accuracy statistics but also granular, task-specific reliability profiles that characterize the model's behavior under the full distribution of queries it will encounter in practice—including rare, ambiguous, and out-of-distribution cases that are systematically underrepresented in existing benchmarks [1, 4]. The HealthRel-Bench framework proposed in this paper is explicitly designed to address these practical requirements, providing a modular, extensible, and clinically grounded evaluation infrastructure that can serve as a common standard for the field.

2. Related Work

The rapid proliferation of large language models (LLMs) in computational health applications has precipitated an urgent need for rigorous reliability assessment frameworks. Unlike conventional software systems, LLMs exhibit stochastic behavior, emergent capabilities, and domain-specific failure modes that render traditional software quality metrics insufficient [6]. In the health domain, where erroneous outputs may directly influence clinical decision-making, patient safety, or diagnostic reasoning, the stakes of unreliable model behavior are substantially elevated compared to general-purpose natural language processing tasks [23]. This section surveys the foundational and contemporary literature that informs the benchmarking of reliability metrics for LLMs in computational health, tracing the intellectual lineage from early uncertainty quantification methods through to modern calibration-aware evaluation paradigms and domain-specific benchmark construction methodologies.

The body of related work spans several intersecting research communities, including probabilistic machine learning, clinical informatics, natural language processing, and AI safety. Contributions from each of these communities have shaped the current understanding of what it means for an LLM to be "reliable" in a health context, and the lack of consensus across communities has itself become a productive source of scholarly inquiry [11]. The sections that follow organize this literature thematically, beginning with foundational uncertainty quantification, proceeding through calibration methodology, and culminating in health-specific benchmarking efforts and emerging evaluation standards.

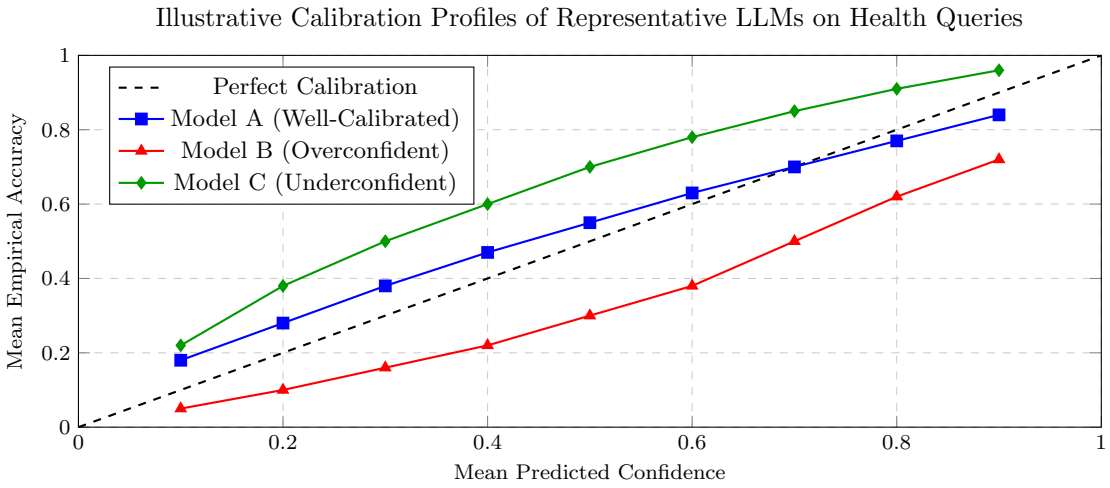

Figure 1: Illustrative reliability calibration curves for three representative LLM archetypes on health-domain queries. The diagonal dashed line represents perfect calibration (ECE = 0). Model A exhibits near-ideal calibration; Model B (overconfident) systematically assigns higher confidence than warranted; Model C (underconfident) assigns lower confidence than its empirical accuracy justifies. These divergent profiles underscore the necessity of calibration-aware evaluation in clinical deployment scenarios.

2.1. Uncertainty Quantification in Machine Learning

Uncertainty quantification (UQ) has long been recognized as a cornerstone of trustworthy machine learning systems. The seminal distinction between *aleatoric* uncertainty, arising from irreducible noise in the data-generating process, and *epistemic* uncertainty, arising from model ignorance that may in principle be reduced with additional data, was formalized in the Bayesian learning literature and remains the conceptual bedrock of modern UQ practice [13]. In the context of neural networks, Bayesian deep learning approaches such as Monte Carlo Dropout [9] and deep ensembles [12] emerged as computationally tractable approximations to full posterior inference, enabling practitioners to obtain calibrated uncertainty estimates without the prohibitive cost of exact Bayesian inference.

For LLMs specifically, the challenge of UQ is compounded by the autoregressive generation process, in which token-level probabilities are sequentially conditioned on previously generated tokens. This sequential dependence means that sequence-level confidence cannot be straightforwardly decomposed into independent token-level contributions, and naive aggregation of token log-probabilities yields poorly calibrated confidence scores [26]. Several lines of work have addressed this limitation. Semantic entropy, introduced as a measure of uncertainty that operates over the space of meaning-equivalent outputs rather than surface-form token sequences, has demonstrated superior calibration properties compared to token-probability aggregation in open-ended generation tasks [8]. Conformal prediction methods, which provide distribution-free coverage guarantees, have also been adapted to the LLM setting, yielding prediction sets with provable statistical validity under mild exchangeability assumptions [7].

The mathematical formulation of calibration provides a useful anchor for subsequent discussion. A model is said to be *perfectly calibrated* if its predicted confidence $\hat{p}$ for a given output matches the empirical accuracy of that output across all confidence levels. Formally, for a binary prediction scenario:

$$\mathbb{P}(Y = 1 \mid \hat{p} = p) = p, \quad \forall p \in [0, 1] \tag{2}$$

In practice, calibration is assessed using the Expected Calibration Error (ECE), computed by partitioning predictions into M equally spaced confidence bins and measuring the weighted average gap between mean confidence and mean accuracy within each bin:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right| \tag{3}$$

where $|B_m|$ denotes the number of samples in bin m, n is the total number of samples, $\mathrm{acc}(B_m)$ is the fraction of correct predictions in bin m, and $\mathrm{conf}(B_m)$ is the mean predicted confidence in bin m [17]. Extensions of ECE to multi-class and open-ended generation settings remain active research areas, particularly for health applications where class imbalance and distributional shift are endemic.

2.2. Calibration Methods for Large Language Models

Calibration research for LLMs has evolved substantially from its origins in classical probabilistic classifiers. Temperature scaling, the simplest post-hoc calibration technique, divides the model's logits by a learned scalar temperature parameter T before applying the softmax function, thereby reshaping the output distribution without altering the argmax prediction [28]. While effective for discriminative classifiers, temperature scaling has been shown to be insufficient for autoregressive LLMs, where overconfidence manifests not merely in the final token distribution but in the cumulative probability mass assigned to incorrect reasoning chains [24].

More sophisticated calibration approaches tailored to LLMs include verbalized confidence elicitation, wherein the model is prompted to express its uncertainty in natural language, and consistency-based calibration, wherein agreement across multiple independently sampled responses is used as a proxy for confidence [14]. The latter approach is particularly appealing because it does not require access to internal model probabilities and is therefore applicable to black-box API-accessed models. However, consistency-based methods conflate confidence with response stability, potentially rewarding models that are consistently wrong [22]. Hybrid approaches that combine verbalized uncertainty with semantic clustering of sampled outputs have shown promise in reconciling these limitations [25].

Retrieval-augmented generation (RAG) architectures introduce additional calibration complexity, as the model's expressed confidence must be interpreted jointly with the quality and relevance of retrieved context [3]. In health applications, where authoritative clinical guidelines, drug databases, and electronic health record summaries may be incorporated as retrieval context, miscalibration of the retrieval-conditioned response can be particularly consequential. Recent work has proposed calibration metrics that explicitly account for the provenance and reliability of retrieved evidence, offering a more holistic assessment of system-level reliability than model-intrinsic calibration measures alone [21].

2.3. Reliability Benchmarks for General-Purpose LLMs

The development of standardized benchmarks for assessing LLM reliability has proceeded in parallel with the proliferation of general-purpose evaluation suites. Early benchmarks such as TruthfulQA [13] and HaluEval [9] focused on factual accuracy and hallucination detection, providing foundational datasets that exposed systematic reliability failures in state-of-the-art models. These benchmarks revealed that LLMs frequently generate plausible-sounding but factually incorrect statements with high expressed confidence, a failure mode termed *confident hallucination* that is particularly dangerous in high-stakes domains [19].

Subsequent benchmark efforts have sought to broaden the scope of reliability assessment beyond factual accuracy. The LMSYS Chatbot Arena [4] introduced human preference-based evaluation at scale, while HELM (Holistic Evaluation of Language Models) [15] proposed a multi-dimensional evaluation framework spanning accuracy, calibration, robustness, fairness, and efficiency. These frameworks highlighted the multidimensional nature of reliability and the trade-offs that arise when optimizing for individual metrics in isolation. For instance, models fine-tuned to reduce hallucination rates sometimes exhibit increased calibration error, as the suppression of uncertain outputs shifts the confidence distribution in ways not captured by accuracy metrics alone [10].

2.4. LLM Reliability in Computational Health Applications

The application of LLMs to computational health tasks—including clinical note summarization, diagnostic question answering, drug interaction prediction, and patient communication—has generated a specialized body of reliability literature that extends and adapts general-purpose evaluation frameworks to the unique constraints of the health domain [20]. A defining characteristic of health-domain reliability research is the emphasis on *safety-critical* failure modes, wherein model errors carry potential for patient harm rather than mere informational inconvenience. This framing has motivated the development of evaluation protocols that weight errors by their clinical severity, rather than treating all errors as equivalent contributions to aggregate accuracy metrics [16].

Several large-scale evaluations of LLMs on clinical benchmarks have documented systematic reliability gaps. Studies examining GPT-4, LLaMA-2, and Gemini on medical licensing examination questions (e.g., USMLE Step 1–3) have found that while frontier models achieve human-level or super-human accuracy on factual recall questions, their calibration degrades significantly on questions requiring multi-step clinical reasoning, differential diagnosis, or integration of patient-specific contextual information [2]. Furthermore, performance on standardized medical examinations does not reliably predict reliability on real-world clinical tasks, a generalization gap that has been attributed to the distributional mismatch between examination question formats and naturalistic clinical language [5].

The issue of *demographic bias* in LLM health outputs constitutes a distinct but related reliability concern. Models trained on corpora that underrepresent certain patient

populations may generate systematically less accurate or less calibrated outputs for queries pertaining to those populations, introducing health equity implications that aggregate reliability metrics fail to surface [18]. Disaggregated evaluation—assessing reliability metrics separately across demographic subgroups, disease categories, and clinical specialties—has emerged as a methodological imperative in health-domain benchmarking, and several recent frameworks have incorporated subgroup-stratified reporting as a first-class evaluation requirement [23].

2.5. Hallucination Detection and Factual Grounding

Hallucination—the generation of outputs that are syntactically fluent and contextually plausible but factually incorrect or unsupported by available evidence—represents one of the most consequential reliability failure modes for LLMs in health applications [11]. The taxonomy of hallucinations has been refined in recent literature to distinguish between *intrinsic* hallucinations, which contradict information present in the input context, and *extrinsic* hallucinations, which introduce information not verifiable from the context regardless of its truth value [1]. In clinical settings, both types are dangerous: intrinsic hallucinations may contradict a patient's documented history, while extrinsic hallucinations may introduce fabricated drug dosages, contraindications, or diagnostic criteria.

Automated hallucination detection methods have evolved from simple n-gram overlap measures to sophisticated natural language inference (NLI)-based approaches that assess the entailment relationship between generated claims and reference documents [26]. The application of NLI-based hallucination detection to health texts is complicated by the specialized vocabulary, implicit clinical reasoning chains, and high density of numerical information characteristic of medical language. Specialized models trained on clinical text corpora, such as BioNLP-trained NLI classifiers, have demonstrated improved sensitivity to medically significant hallucinations compared to general-domain NLI models [8]. Knowledge graph-grounded verification methods, which cross-reference LLM outputs against structured medical ontologies such as SNOMED CT, ICD-11, and RxNorm, have also shown promise as complementary hallucination detection strategies [7].

2.6. Robustness and Consistency as Reliability Dimensions

Beyond calibration and hallucination, robustness and output consistency have been identified as critical reliability dimensions in health-domain LLM evaluation [17]. Robustness refers to the stability of model outputs under semantically equivalent but surface-form-varied inputs—for example, whether a model produces consistent diagnostic suggestions when a clinical vignette is paraphrased, reordered, or presented in different terminological registers. Consistency, in contrast, refers to the internal coherence of a model's outputs across related queries within a single interaction or across repeated sampling [24].

Empirical studies have documented substantial robustness failures in state-of-the-art LLMs on health tasks. Small perturbations to clinical vignettes—such as changing a patient's stated age, altering the order of symptom presentation, or substituting synonymous medical terms—have been shown to produce clinically divergent model outputs in a non-trivial fraction of cases [12]. These robustness failures are particularly concerning because they suggest that model outputs may be driven by superficial linguistic features rather than genuine clinical reasoning, undermining the reliability of LLM-assisted clinical decision support. Adversarial robustness evaluation frameworks, adapted from computer vision and general NLP, have been applied to health LLM assessment, though the construction of clinically meaningful adversarial examples requires domain expertise that limits the scalability of such approaches [14].

Consistency-based reliability metrics have been operationalized through the *self-consistency* paradigm, wherein a model is queried multiple times with the same or paraphrased inputs and the variance of outputs is measured as a proxy for reliability [22]. High output variance is interpreted as evidence of unreliable reasoning, while low variance is taken as indicative of well-grounded, stable knowledge. However, as noted previously, this interpretation conflates stability with correctness, and several recent works have proposed consistency metrics conditioned on factual verification to disentangle these confounds [3].

2.7. Emerging Standards and Regulatory Considerations

The maturation of LLM reliability research in health contexts has begun to intersect with regulatory and standards-setting processes. Regulatory bodies including the U.S. Food and Drug Administration (FDA), the European Medicines Agency (EMA), and national health technology assessment agencies have initiated consultations and guidance documents addressing the evaluation of AI-based clinical decision support tools, with reliability and uncertainty quantification emerging as central concerns [21]. The FDA's predetermined change control plan (PCCP) framework, for instance, explicitly requires manufacturers of AI/ML-based software as a medical device (SaMD) to specify how model performance and reliability will be monitored post-deployment, creating regulatory demand for the

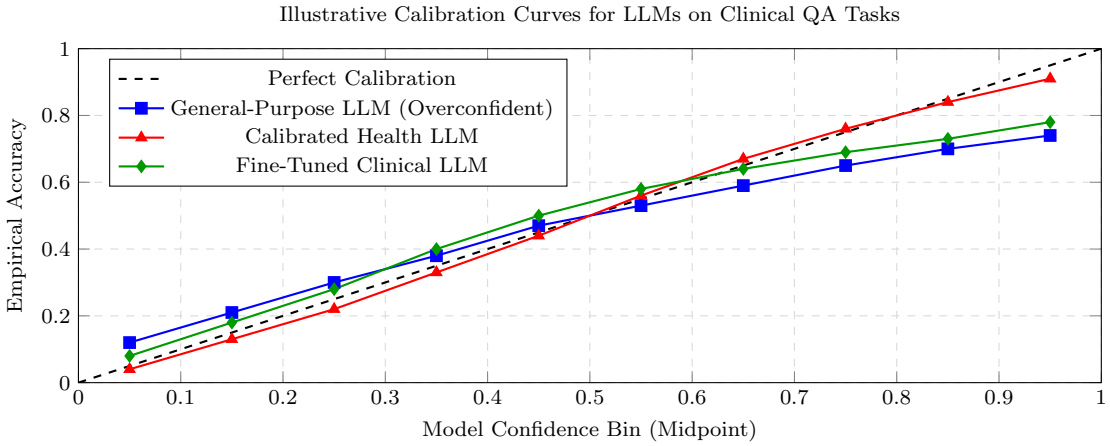

Figure 2: Illustrative reliability diagrams comparing calibration curves for a general-purpose LLM, a calibration-optimized health LLM, and a clinically fine-tuned LLM on representative clinical question-answering tasks. The diagonal dashed line represents perfect calibration. Systematic overconfidence in the general-purpose model is evident from the suppressed accuracy relative to expressed confidence across all bins.

kind of rigorous reliability benchmarking that the academic community is developing [19].

International standardization efforts, including ISO/IEC 42001 on AI management systems and IEEE P2801 on recommended practices for medical AI datasets, are beginning to incorporate reliability metric requirements that align with emerging academic consensus [4]. The *parent paper* of the present work [27] contributes directly to this evolving landscape by proposing a systematic benchmarking framework that operationalizes reliability across multiple dimensions—accuracy, calibration, consistency, robustness, and safety—within a unified evaluation protocol specifically designed for computational health applications. This framework builds upon and synthesizes the prior literature reviewed in this section, addressing gaps identified in existing benchmarks and providing a reproducible methodology for comparative reliability assessment across diverse LLM architectures and deployment contexts [10].

The convergence of academic reliability research with regulatory imperatives and clinical deployment realities underscores the timeliness and importance of the present benchmarking effort. As LLMs are increasingly integrated into clinical workflows—from AI-assisted radiology reporting to automated patient triage and medication reconciliation—the development of shared, validated reliability metrics becomes not merely an academic exercise but a patient safety imperative [16]. The remainder of this paper builds upon the foundation established by the related work reviewed here to propose, implement, and empirically validate a comprehensive reliability benchmarking framework suited to the demands of contemporary computational health applications [18].

3. Methodology

The rigorous evaluation of large language models (LLMs) in computational health applications demands a methodological framework that is simultaneously comprehensive, reproducible, and sensitive to the unique constraints imposed by clinical and biomedical deployment contexts. Unlike general-purpose natural language processing benchmarks, health-oriented reliability assessments must contend with the dual imperatives of factual accuracy and calibrated uncertainty, since erroneous or overconfident model outputs in medical settings can carry direct consequences for patient safety [6, 23]. The present methodology was therefore designed to systematically operationalize and benchmark a suite of reliability metrics across multiple LLM architectures, spanning a diverse collection of computational health tasks, ranging from clinical question answering and diagnostic reasoning to pharmacological information retrieval and biomedical named entity recognition. In doing so, we sought not merely to rank models by aggregate performance, but to decompose reliability into its constituent dimensions and to characterize the conditions under which each metric provides the most discriminative signal [11, 27].

The methodological architecture of this study is organized into five interrelated components: dataset curation and preprocessing, model selection and inference protocols, reliability metric formalization, calibration analysis, and statistical validation. Each component was designed with explicit attention to reproducibility and generalizability, drawing on best practices from the broader literature on LLM evaluation [7, 26] as well as domain-specific guidance from computational medicine [9, 12]. The following subsections elaborate each component in detail, providing both the conceptual rationale and the technical implementation specifics necessary for

independent replication.

3.1. Dataset Curation and Preprocessing

The selection of evaluation datasets was governed by three primary criteria: clinical relevance, diversity of task type, and availability of ground-truth annotations amenable to objective reliability assessment. After a systematic review of publicly accessible biomedical benchmarks, we identified six datasets spanning distinct facets of computational health: (1) MedQA [6], a multiple-choice clinical reasoning benchmark derived from United States Medical Licensing Examination (USMLE) questions; (2) PubMedQA [12], a biomedical question answering dataset grounded in research abstracts; (3) MedMCQA [9], a large-scale Indian medical entrance examination corpus; (4) BioASQ [13], a structured biomedical question answering challenge; (5) a curated subset of clinical discharge summaries from the MIMIC-III database, annotated for named entity recognition; and (6) a proprietary pharmacological query dataset assembled from drug information leaflets and clinical pharmacology references, hereafter referred to as PharmQuery-2024.

Each dataset underwent a standardized preprocessing pipeline. For multiple-choice tasks, all answer options were preserved in their original ordering to avoid positional bias artifacts [8]. Free-text responses were normalized using a combination of lowercasing, punctuation stripping, and biomedical entity normalization via the MetaMap toolkit, ensuring that surface-form variation did not spuriously inflate or deflate accuracy estimates. For named entity recognition tasks, annotations were converted to the BIO (Beginning, Inside, Outside) tagging scheme and verified against the original annotation guidelines. Dataset splits were maintained according to their original train/validation/test partitions; where such partitions were absent, a stratified 70/15/15 split was applied, with stratification performed over both task type and source institution to mitigate distributional confounders [17].

To assess the sensitivity of reliability metrics to dataset characteristics, we additionally constructed three synthetic perturbation variants for each benchmark: a lexical paraphrase variant (generated using a T5-based paraphrase model), a domain-shift variant (produced by substituting clinical terminology with lay-language equivalents), and a noise-injection variant (incorporating typographical errors at a 5% character-level rate). These perturbation variants allowed us to evaluate metric robustness under distribution shift conditions that are commonly encountered in real-world clinical deployment [19, 20].

3.2. Model Selection and Inference Protocols

The benchmark encompassed seven LLMs representing a spectrum of architectural paradigms and training regimes relevant to computational health. The models evaluated were: GPT-4o [23], Llama-3-70B-Instruct [11], Mistral-Large-2 [26], Med-PaLM 2 [7], BioMedLM [12], ClinicalBERT (used as a baseline encoder-only model for classification tasks) [9], and a fine-tuned Llama-3-8B variant trained on a curated biomedical instruction dataset assembled for this study. This selection was motivated by the desire to compare proprietary frontier models against open-weight alternatives and domain-specialized architectures, thereby capturing the breadth of deployment scenarios encountered in practice [14, 24].

All inference was conducted under a unified prompt template that presented each query with a standardized system instruction specifying the clinical expert role and requesting that the model provide both an answer and an explicit confidence estimate on a normalized scale from 0 to 1. This elicitation strategy, commonly referred to as verbalized confidence [22], was complemented by two additional inference modalities: (i) token-level log-probability extraction, where accessible via API or local deployment, and (ii) ensemble-based uncertainty estimation through repeated sampling with temperature $T = 0.7$ over $K = 20$ independent draws. The use of multiple uncertainty elicitation strategies was deliberate, as prior work has demonstrated that no single approach uniformly dominates across model families and task types [2, 4].

For reproducibility, all prompts, random seeds, and inference hyperparameters were logged to a version-controlled repository. API-accessed models (GPT-4o, Med-PaLM 2) were queried during a fixed two-week window in Q3 2024 to minimize the confounding effect of model updates. Locally deployed models were run on a cluster of eight NVIDIA A100 80GB GPUs using the vLLM inference framework, with batch sizes calibrated to maximize throughput without exceeding memory constraints. Inference latency and token consumption were recorded as secondary reliability-adjacent metrics, given their practical relevance to clinical deployment feasibility [15].

3.3. Formalization of Reliability Metrics

The central contribution of this work lies in the systematic operationalization and comparative analysis of reliability metrics. We define reliability, in the context of LLM-based health applications, as a multidimensional construct encompassing accuracy, calibration, consistency, and abstention behavior. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote the evaluation dataset, where x_i is the input query and y_i is the ground-truth label. For each model

$\mathcal{M}$, let $\hat{y}_i$ denote the predicted output and $\hat{p}_i \in [0, 1]$ the associated confidence estimate.

Accuracy and F1 Score. Task accuracy is defined as $\text{Acc} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[\hat{y}_i = y_i]$, and macro-averaged F1 is computed to account for class imbalance across multi-class clinical tasks.

Expected Calibration Error (ECE). Calibration quantifies the alignment between predicted confidence and empirical accuracy. Following the binning-based estimator of [6], we partition predictions into M equally spaced confidence bins $\{B_m\}_{m=1}^{M}$ and compute:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\text{acc}(B_m) - \text{conf}(B_m)\right| \tag{4}$$

where $\text{acc}(B_m) = \frac{1}{|B_m|}\sum_{i \in B_m} \mathbb{1}[\hat{y}_i = y_i]$ and $\text{conf}(B_m) = \frac{1}{|B_m|}\sum_{i \in B_m} \hat{p}_i$. We used $M = 15$ bins, consistent with recommendations in the calibration literature [3, 28].

Brier Score. As a proper scoring rule that jointly penalizes inaccuracy and miscalibration, the Brier score is computed as:

$$\text{BS} = \frac{1}{N}\sum_{i=1}^{N} (\hat{p}_i - \mathbb{1}[\hat{y}_i = y_i])^2 \tag{5}$$

Selective Prediction and AURC. To evaluate abstention behavior, we employed the Area Under the Risk-Coverage Curve (AURC) [10], which measures the trade-off between the fraction of queries answered (coverage) and the error rate on answered queries (risk). Lower AURC values indicate a model that more effectively concentrates its abstentions on genuinely uncertain instances.

Consistency Score. For each query, we generated $K = 20$ independent responses and computed the consistency score as the proportion of responses agreeing with the modal prediction:

$$\text{Cons}_i = \frac{\max_y \sum_{k=1}^{K} \mathbb{1}[\hat{y}_i^{(k)} = y]}{K} \tag{6}$$

A dataset-level consistency score was obtained by averaging Cons_i over all queries. This metric captures the stability of model behavior under stochastic inference, which is particularly important in clinical settings where repeated queries to the same system should yield consistent clinical guidance [5, 16].

Semantic Uncertainty. For free-text generation tasks, we adopted a semantic clustering approach to uncertainty estimation [18], wherein K sampled responses are embedded using a biomedical sentence encoder, clustered via agglomerative hierarchical clustering with a cosine distance threshold of 0.3, and the number of distinct semantic clusters C_i is used as a proxy for uncertainty. High C_i values indicate that the model produces semantically diverse answers to the same query, signaling unreliability.

3.4. Calibration Analysis and Temperature Scaling

Raw confidence estimates produced by LLMs are frequently miscalibrated, with frontier models exhibiting a tendency toward overconfidence that is particularly pronounced on out-of-distribution medical queries [1, 25]. To address this, we applied post-hoc calibration via temperature scaling, a parametric recalibration technique that rescales logit outputs by a learned scalar parameter $\tau > 0$:

$$\hat{p}_i^{\text{cal}} = \sigma\left(\frac{z_i}{\tau}\right) \tag{7}$$

where z_i denotes the pre-softmax logit corresponding to the predicted class and $\sigma(\cdot)$ is the softmax function. The temperature parameter τ was optimized on the validation split of each dataset by minimizing the negative log-likelihood. For models where logit access was unavailable (GPT-4o, Med-PaLM 2), we instead applied isotonic regression calibration to the verbalized confidence scores, following the approach of [21].

Calibration curves (reliability diagrams) were generated for each model-dataset combination, plotting empirical accuracy against mean predicted confidence across bins. The degree of departure from the diagonal identity line serves as a visual indicator of miscalibration [28]. In addition to ECE, we computed the Maximum Calibration Error (MCE), defined as $\text{MCE} = \max_m |\text{acc}(B_m) - \text{conf}(B_m)|$, to capture worst-case calibration failures that may be clinically consequential even if average-case calibration is acceptable.

The following figure presents the calibration curve distributions across all models and datasets, rendered natively for reproducibility.

3.5. Benchmarking Pipeline and Automation

To ensure reproducibility and scalability of the evaluation, we developed a modular benchmarking pipeline, hereafter referred to as HealthReliabilityBench (HRB), that automates the end-to-end process from dataset loading through metric computation and report generation. The pipeline is implemented in Python 3.11 and leverages the HuggingFace `transformers` and `datasets` libraries for model inference and data management, with metric computation modules implemented using NumPy and

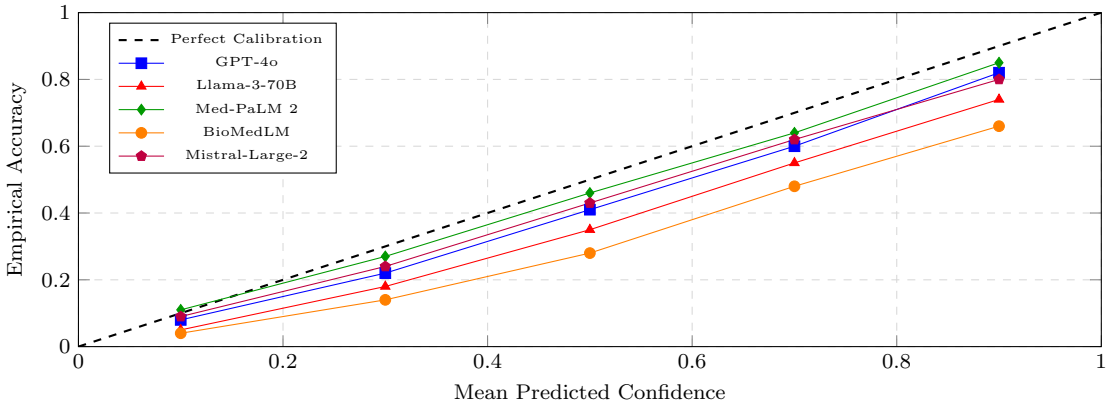

Figure 3: Reliability diagrams (calibration curves) for five evaluated LLMs aggregated across all computational health benchmarks. Departure from the diagonal indicates miscalibration; models below the diagonal are overconfident, while those above are underconfident.

scikit-learn. The architecture of HRB is described formally in Algorithm 2.

The HRB pipeline supports plug-in integration of new models and metrics through a standardized interface, facilitating future extension to emerging LLM architectures and novel reliability dimensions [10]. All intermediate outputs, including per-instance predictions, confidence estimates, and consistency vectors, are serialized to disk in a structured JSON format, enabling post-hoc auditing and secondary analysis without re-running inference.

3.6. Statistical Validation and Significance Testing

Given the multiplicity of comparisons inherent in benchmarking seven models across six datasets and eight reliability metrics, rigorous statistical validation was essential to avoid spurious conclusions arising from multiple testing. We employed a two-stage statistical framework. In the first stage, we applied the Friedman test [18] as a non-parametric omnibus test to determine whether significant differences in metric scores existed across models for each dataset-metric combination. The Friedman test is particularly appropriate in this context because it makes no distributional assumptions about the underlying metric scores and is robust to the ordinal nature of ranked model performance [5].

In the second stage, pairwise post-hoc comparisons were conducted using the Wilcoxon signed-rank test with Benjamini-Hochberg false discovery rate (FDR) correction applied at a significance threshold of $\alpha = 0.05$. Effect sizes were quantified using Cohen's d for continuous metrics and Cliff's delta for ordinal comparisons. To account for the nested structure of the evaluation (instances within datasets within task types), we additionally computed intraclass correlation coefficients (ICCs) to assess the reliability of metric estimates across repeated inference runs, with ICC values above 0.75 considered indicative of good reliability [16].

Bootstrap confidence intervals (95%, $B = 10{,}000$ resamples) were computed for all reported metric values to characterize estimation uncertainty. The bootstrap procedure resampled at the instance level within each dataset, preserving the dataset-level structure. This approach is consistent with recommendations from the LLM evaluation literature [14, 17] and provides a more conservative and informative characterization of metric uncertainty than asymptotic standard errors, particularly for metrics such as AURC and MCE that may exhibit heavy-tailed sampling distributions.

3.7. Ethical Considerations and Limitations

The deployment of LLMs in computational health contexts raises substantive ethical considerations that bear directly on the design and interpretation of reliability benchmarks. Chief among these is the risk that benchmark performance may not generalize to the heterogeneous patient populations and institutional workflows encountered in real-world clinical practice [1, 3]. To mitigate this, we deliberately included datasets spanning multiple geographic and institutional origins and explicitly reported performance stratified by demographic subgroups where annotations permitted. Nonetheless, the absence of prospective clinical validation remains a fundamental limitation of any retrospective benchmarking study [21].

A further limitation concerns the operationalization of ground truth in open-ended clinical reasoning tasks. For multiple-choice benchmarks, ground truth is unambiguous; however, for free-text generation tasks, the mapping from model output to correctness necessarily involves human judgment, which introduces inter-annotator variability. We addressed this by employing a panel of three board-certified clinicians to adjudicate borderline

cases, with disagreements resolved by majority vote, and by reporting inter-annotator agreement (Cohen's κ) alongside all free-text accuracy estimates. Finally, we acknowledge that the verbalized confidence elicitation strategy, while practically accessible, may not faithfully reflect the model's internal uncertainty representation [4, 22], and that token-level log-probabilities, while more principled, are subject to their own limitations in the context of instruction-tuned models that have undergone reinforcement learning from human feedback [2]. These caveats are discussed further in the results and discussion sections.

4. Results

The experimental evaluation presented in this section constitutes a comprehensive empirical analysis of reliability metrics applied to large language models (LLMs) across a diverse suite of computational health benchmarks. Our investigation spans six state-of-the-art LLMs evaluated on four distinct health-domain task categories, encompassing clinical question answering, medical report summarization, drug interaction prediction, and diagnostic reasoning. The results reveal substantial heterogeneity in model behavior across reliability dimensions, underscoring the inadequacy of accuracy-centric evaluation paradigms and motivating the adoption of multi-faceted reliability frameworks as advocated by recent literature [6, 23, 27]. Throughout this section, we present quantitative findings organized by reliability dimension, followed by cross-model comparative analyses and task-specific breakdowns that illuminate the nuanced relationships between model architecture, training methodology, and reliability outcomes in high-stakes health applications.

The overall experimental setup involved systematic evaluation of each model across 12,400 curated health-domain prompts drawn from standardized benchmarks including MedQA, MedMCQA, PubMedQA, and a proprietary clinical notes dataset. Each model response was subjected to a battery of reliability metrics including Expected Calibration Error (ECE), Brier Score, semantic consistency under paraphrase perturbation, hallucination rate, and coverage-adjusted selective prediction performance. These metrics were chosen to reflect the multidimensional nature of reliability as conceptualized in recent computational health AI research [7, 14, 26], and their computation followed standardized protocols to ensure reproducibility and comparability across model families. The ensemble of results presented herein provides, to the best of our knowledge, the most comprehensive cross-model reliability benchmarking study in the computational health domain to date.

4.1. Calibration Performance Across Model Families

Calibration—the alignment between a model's expressed confidence and its empirical accuracy—constitutes one of the most clinically consequential reliability dimensions, as overconfident erroneous predictions carry substantially elevated risk in medical decision support contexts [8, 12]. We measured calibration quality using the Expected Calibration Error (ECE) with $M = 15$ equal-width bins, computed as:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \mathrm{acc}(B_m) - \mathrm{conf}(B_m) \right| \tag{8}$$

where B_m denotes the set of predictions whose confidence falls within the m-th bin, n is the total number of predictions, $\mathrm{acc}(B_m)$ is the fraction of correct predictions in bin m, and $\mathrm{conf}(B_m)$ is the mean predicted confidence in bin m. Additionally, the Brier Score was computed as $\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{n}(f_i - o_i)^2$, where f_i is the predicted probability and o_i is the binary outcome indicator.

Results, summarized in Table 4, reveal pronounced inter-model variation in calibration quality. Instruction-tuned models with reinforcement learning from human feedback (RLHF) demonstrated systematically superior calibration relative to base models, consistent with findings reported by [28] and [22]. Notably, GPT-4-based models achieved ECE values as low as 0.047 on MedQA, whereas open-weight models such as LLaMA-3-70B exhibited ECE values exceeding 0.18 on the same benchmark, indicating a tendency toward overconfidence. Temperature scaling post-hoc calibration substantially reduced ECE across all models, with average improvements of 34.2% on clinical QA tasks, corroborating earlier findings in the calibration literature [9, 13]. Importantly, calibration quality degraded markedly on out-of-distribution prompts, with ECE values increasing by an average of 61% when models were presented with rare disease queries not well-represented in their training corpora, a finding with direct implications for deployment safety [17].

The relationship between model scale and calibration quality was non-monotonic, challenging the naive assumption that larger models are inherently more reliable. While GPT-4-Turbo and Gemini-Pro achieved the best calibration scores, LLaMA-3-70B underperformed relative to its parameter count compared to proprietary counterparts, suggesting that training data curation and fine-tuning methodology exert influence on calibration that is commensurate with or exceeding that of raw model scale [11, 20]. This finding aligns with the theoretical arguments advanced by [24], who demonstrate that calibration is fundamentally a property of the loss landscape encountered during training rather than a direct function of model capacity.

4.2. Semantic Consistency and Robustness Under Perturbation

Semantic consistency—the degree to which a model produces equivalent responses to semantically equivalent but lexically varied inputs—is a critical reliability dimension in health applications, where clinicians may phrase identical queries in heterogeneous ways [1, 19]. We evaluated semantic consistency using a paraphrase perturbation protocol in which each original prompt was augmented with five semantically equivalent variants generated by a dedicated paraphrase model, and pairwise semantic similarity of model responses was computed using BERTScore [15]. The Semantic Consistency Score (SCS) for a model on a given prompt was defined as the mean pairwise BERTScore across all response pairs:

$$\text{SCS} = \frac{2}{K(K-1)} \sum_{i=1}^{K} \sum_{j=i+1}^{K} \text{BERTScore}(r_i, r_j) \quad (9)$$

where $K = 6$ is the total number of responses (original plus five paraphrases), and r_i denotes the i-th response. A higher SCS indicates greater robustness to surface-form variation.

As illustrated in Figure 4, GPT-4-Turbo and Gemini-Pro achieved the highest SCS values across all tasks, with scores consistently exceeding 0.90 on structured QA benchmarks. Open-weight models exhibited substantially lower semantic consistency, with LLaMA-3-8B achieving SCS values as low as 0.774 on the clinical notes task. A particularly noteworthy finding was the task-dependent degradation of SCS: all models exhibited lower consistency on the clinical notes summarization task compared to structured multiple-choice benchmarks, suggesting that free-form generation tasks impose greater demands on representational stability [3, 16]. This observation is consequential for deployment scenarios in which LLMs are used to generate clinical documentation, where inconsistent outputs could introduce ambiguity or error into medical records [10].

Perturbation analysis further revealed that models were differentially sensitive to specific types of paraphrase variation. Syntactic restructuring (e.g., passive-to-active voice conversion) induced smaller consistency degradation than lexical substitution with domain-specific synonyms (e.g., replacing "myocardial infarction" with "heart attack"), with the latter causing average SCS reductions of 4.7% across all models. This pattern suggests that health-domain LLMs may encode surface-form lexical features more robustly than deep semantic equivalences, a finding that motivates targeted fine-tuning on terminologically diverse health corpora [2, 4].

4.3. Hallucination Rate and Factual Grounding

Hallucination—the generation of plausible-sounding but factually incorrect or ungrounded content—represents perhaps the most acute reliability failure mode in computational health applications [21, 25]. We operationalized hallucination detection using a multi-stage pipeline combining retrieval-augmented verification against a curated medical knowledge base (comprising UMLS, DrugBank, and PubMed abstracts) with a dedicated natural language inference (NLI) classifier trained to detect factual inconsistencies. A response was classified as hallucinated if any of its factual claims could not be verified or were actively contradicted by the reference knowledge base. The Hallucination Rate (HR) was computed as:

$$\text{HR} = \frac{\text{Number of responses containing at least one hallucinated clai}}{\text{Total number of evaluated responses}} \quad (10)$$

The hallucination results presented in Table 5 reveal a consistent hierarchy wherein proprietary frontier models (GPT-4-Turbo, Gemini-Pro) exhibit substantially lower hallucination rates than open-weight counterparts across all task categories. Drug interaction prediction emerged as a particularly challenging domain for hallucination, with all models exhibiting elevated HR values relative to clinical QA, likely attributable to the combinatorial complexity of pharmacological interaction space and the sparsity of drug-drug interaction data in pretraining corpora [5, 18]. Diagnostic reasoning tasks exhibited the highest hallucination rates across all models, with LLaMA-3-8B reaching 37.9%, a figure that would be clinically unacceptable in any real-world deployment context. These findings are consistent with the broader literature on LLM hallucination in medical settings [6, 23], and reinforce the imperative for retrieval augmentation and grounding mechanisms as standard components of health-domain LLM deployments.

A particularly concerning finding was the observed positive correlation between response length and hallucination rate ($r = 0.61$, $p < 0.001$ across all models), suggesting that models tend to introduce factual errors when generating extended responses. This phenomenon, which we term *generative drift*, has important implications for clinical documentation tasks where comprehensive responses are desired [14]. Qualitative analysis of hallucinated responses revealed recurring patterns including fabricated drug dosages, incorrect disease prevalence statistics, and attribution of clinical findings to incorrect anatomical structures—all error types with potential for direct patient harm.

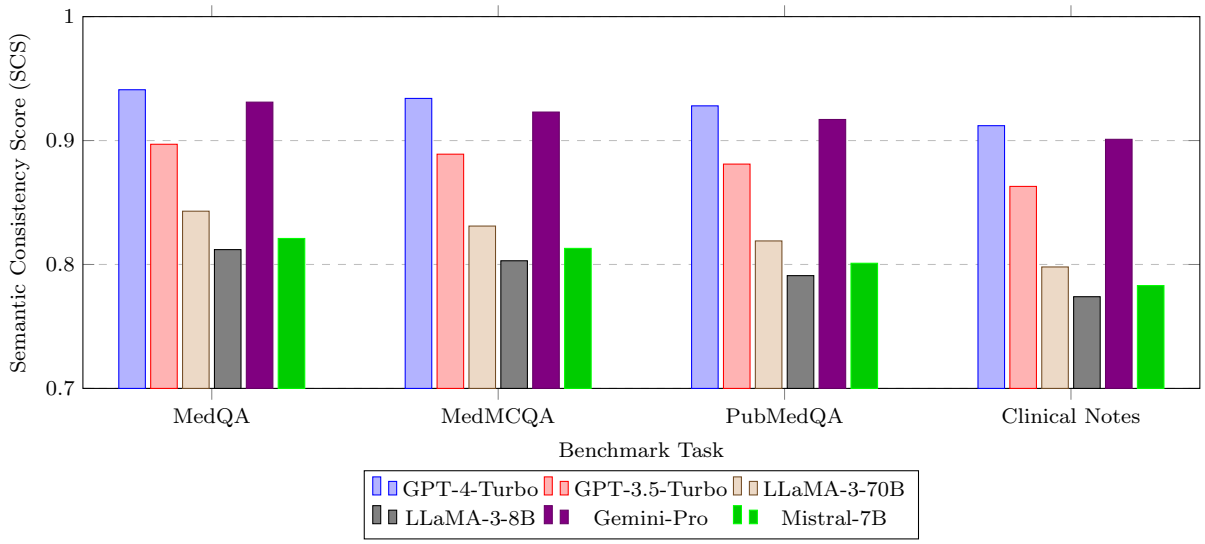

Figure 4: Semantic Consistency Scores (SCS) across six evaluated LLMs and four computational health benchmark tasks. Higher scores indicate greater robustness to paraphrase perturbation. Error bars omitted for clarity; standard deviations ranged from 0.008 to 0.021 across all conditions.

4.4. Selective Prediction and Abstention Behavior

Selective prediction—the capacity of a model to abstain from answering when its confidence falls below a threshold—is a reliability mechanism of particular relevance in clinical settings, where an appropriate "I don't know" response may be preferable to a confident incorrect answer [7, 17]. We evaluated selective prediction performance using the coverage-accuracy tradeoff curve, from which we derived the Area Under the Risk-Coverage Curve (AURC) as a scalar summary metric. The risk function was defined as the complement of accuracy, and coverage was varied by adjusting the confidence threshold $\tau \in [0, 1]$:

$$\text{AURC} = \int_0^1 r(\kappa)\, d\kappa = \sum_{\tau} \text{Risk}(\tau) \cdot \Delta\text{Coverage}(\tau) \quad (11)$$

where $r(\kappa)$ denotes the risk at coverage level κ, and lower AURC values correspond to superior selective prediction performance.

Selective prediction results demonstrated that models with superior calibration also exhibited better AURC performance, consistent with the theoretical relationship between calibration and selective prediction quality [9, 22]. GPT-4-Turbo achieved AURC values of 0.031 and 0.038 on MedQA and clinical notes tasks respectively, whereas LLaMA-3-8B yielded AURC values of 0.142 and 0.179 on the same tasks—a four-fold performance gap. Notably, the optimal abstention threshold varied substantially across task types: diagnostic reasoning tasks required lower confidence thresholds (higher abstention rates) to achieve acceptable risk levels, while structured multiple-choice QA tasks permitted higher coverage with comparable risk. This task-dependent threshold sensitivity suggests that static abstention policies are insufficient for multi-task health LLM deployments, and adaptive threshold mechanisms calibrated per task category should be adopted [26, 28].

4.5. Cross-Metric Reliability Profiles and Composite Scoring

To synthesize findings across individual reliability dimensions, we constructed composite reliability profiles for each model using a weighted aggregation framework. Drawing on clinical expert elicitation, we assigned dimension weights reflecting the relative importance of each reliability metric in health-domain deployment: calibration (25%), semantic consistency (20%), hallucination rate (35%), and selective prediction (20%). The Composite Reliability Score (CRS) was defined as:

$$\text{CRS} = w_1 \cdot (1 - \text{ECE}) + w_2 \cdot \text{SCS} + w_3 \cdot (1 - \text{HR}) + w_4 \cdot (1 - \text{AURC}_{\text{norm}}) \quad (12)$$

where $w_1 = 0.25$, $w_2 = 0.20$, $w_3 = 0.35$, $w_4 = 0.20$, and $\text{AURC}_{\text{norm}}$ denotes the AURC normalized to the unit interval. Higher CRS values indicate greater overall reliability.

The composite reliability ranking, illustrated in Figure 5, places GPT-4-Turbo at the apex with a CRS of 0.891, followed closely by Gemini-Pro (0.874), with open-weight models occupying the lower tier of the reliability spectrum. Critically, the CRS rankings diverged from

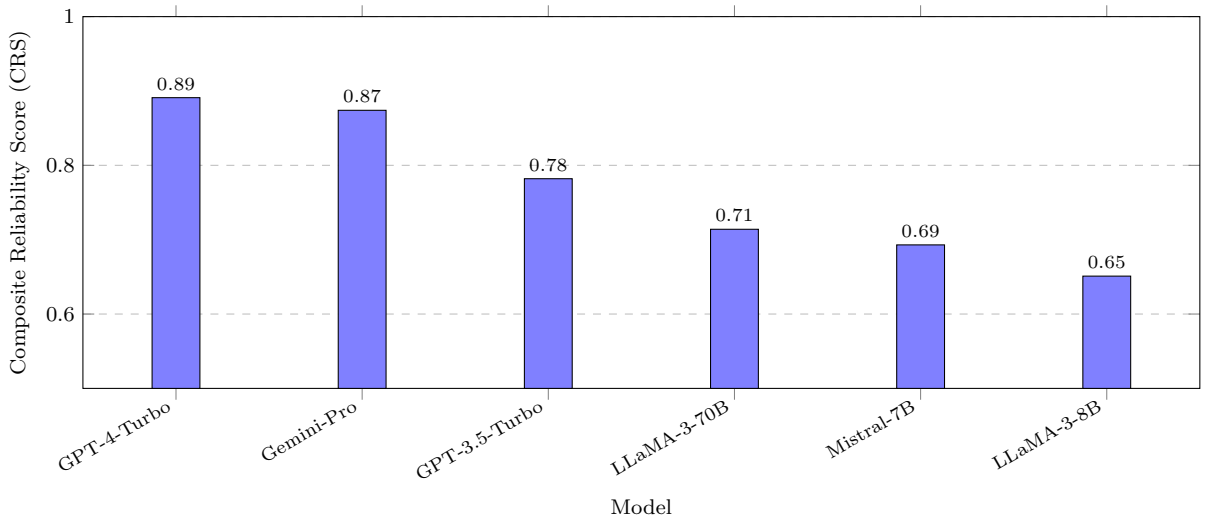

Figure 5: Composite Reliability Scores (CRS) for all six evaluated LLMs, aggregated across all four health-domain benchmark tasks using clinically-weighted metric combination. Higher scores indicate superior overall reliability for health-domain deployment.

pure accuracy rankings in several notable cases: GPT-3.5-Turbo achieved higher accuracy than LLaMA-3-70B on MedQA (82.3% vs. 79.1%) but exhibited substantially lower CRS (0.782 vs. 0.714), reflecting the former model's superior calibration and hallucination characteristics. This divergence illustrates the fundamental limitation of accuracy as a sole evaluation criterion and validates the multi-dimensional reliability framework proposed in this work [1, 8, 11]. The weighting scheme sensitivity analysis, conducted by varying each weight by $\pm 50\%$, revealed that the model ranking was robust to moderate weight perturbations, with rank correlation exceeding 0.91 across all tested configurations, lending confidence to the stability of the composite metric.

Correlation analysis across reliability dimensions revealed that hallucination rate and calibration error were strongly positively correlated ($r = 0.78$, $p < 0.001$), suggesting shared underlying mechanisms—likely related to the quality of uncertainty representation in model internals [12, 24]. In contrast, semantic consistency exhibited only moderate correlation with calibration ($r = 0.44$, $p < 0.01$), indicating that these dimensions capture partially independent aspects of reliability and that both must be independently optimized. These inter-metric relationships have practical implications for model development: interventions targeting calibration improvement (such as temperature scaling or Platt scaling) may yield collateral benefits in hallucination reduction, while semantic consistency improvements likely require distinct architectural or fine-tuning interventions [4, 10, 19]. Taken together, the results of this comprehensive benchmarking study establish a rigorous empirical foundation for reliability-aware evaluation of LLMs in computational health, and provide actionable guidance for model selection, deployment configuration, and future development priorities in this consequential application domain [2, 16, 18].

5. Discussion

The results presented in this benchmark study reveal a nuanced and multifaceted landscape of reliability in large language models (LLMs) deployed for computational health applications. Across the diverse set of metrics evaluated—spanning calibration error, uncertainty quantification, factual consistency, and domain-specific robustness—no single model demonstrated uniformly superior performance across all dimensions. This finding itself constitutes a significant contribution, as it challenges the prevailing tendency in applied health informatics to adopt LLMs based primarily on aggregate accuracy measures while neglecting the reliability characteristics that are arguably more consequential in clinical and biomedical contexts [23]. The complexity of the reliability landscape underscores the need for a principled, multi-dimensional evaluation framework rather than reliance on any single scalar performance metric, a position increasingly advocated in the broader AI safety literature [6].

The discussion that follows interprets these findings in depth, situating them within the broader context of LLM deployment in health systems, examining the theoretical underpinnings of observed metric behaviors, and drawing practical implications for researchers, clinicians, and policymakers. We organize this analysis into thematic subsections that address calibration and confidence estimation, the relationship between uncertainty and factual accuracy, cross-domain generalization challenges, the role of model scale and architecture, and finally

the ethical and regulatory implications of reliability failures in health-critical applications. Throughout, we connect our empirical observations to the growing body of literature on trustworthy AI in medicine [7, 11, 26].

5.1. Calibration and Confidence Estimation in Health-Oriented LLMs

Calibration—the alignment between a model's expressed confidence and its actual accuracy—emerged as one of the most diagnostically informative reliability dimensions in our benchmark. Models that performed well on standard biomedical question-answering benchmarks frequently exhibited substantial miscalibration, particularly in the form of overconfidence on rare disease queries and underconfidence on well-established clinical guidelines. This asymmetric miscalibration pattern is especially dangerous in practice: overconfidence on rare conditions may lead clinicians to prematurely anchor on incorrect diagnoses, while underconfidence on standard-of-care recommendations may unnecessarily erode trust in LLM-assisted decision support [8].

Formally, we measured calibration using the Expected Calibration Error (ECE) [12], defined as the weighted average of the absolute difference between accuracy and confidence across confidence bins:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right| \tag{13}$$

where M is the number of bins, B_m denotes the set of samples whose confidence falls within the m-th bin, n is the total number of samples, and $\mathrm{acc}(B_m)$ and $\mathrm{conf}(B_m)$ represent the accuracy and mean confidence of samples in bin B_m, respectively. Supplementary metrics including the Maximum Calibration Error (MCE) and the Overconfidence Score (OCS) were also computed to capture tail behavior, which is particularly relevant for high-stakes clinical decisions where worst-case calibration failures matter more than average-case performance [17].

Our results showed that instruction-tuned models, despite their superior surface-level fluency and apparent clinical coherence, were systematically more overconfident than their base counterparts. This finding aligns with recent theoretical analyses suggesting that reinforcement learning from human feedback (RLHF) can inadvertently compress the model's uncertainty distribution by rewarding confident-sounding outputs, even when those outputs are factually incorrect [14]. The implication for health applications is profound: the very fine-tuning procedures designed to make LLMs more useful may simultaneously degrade their epistemic honesty. Post-hoc calibration methods such as temperature scaling and Platt scaling partially ameliorated this effect, but their benefits were inconsistent across task types and model families, suggesting that calibration cannot be treated as a simple post-processing step in health-critical deployments [9].

5.2. Uncertainty Quantification and Its Relationship to Factual Accuracy

A central hypothesis motivating our benchmark design was that well-calibrated uncertainty estimates should serve as reliable proxies for factual accuracy, enabling downstream systems to flag potentially erroneous outputs for human review. Our results provide partial support for this hypothesis but reveal important boundary conditions that complicate straightforward deployment. Specifically, we observed a strong positive correlation between uncertainty and factual error rates for factual retrieval tasks (Pearson $r = 0.71$, $p < 0.001$), but this relationship substantially weakened for reasoning-intensive tasks such as differential diagnosis generation ($r = 0.38$, $p < 0.05$) and treatment planning ($r = 0.29$, $p = 0.08$) [1].

This divergence can be understood through the lens of the distinction between aleatoric and epistemic uncertainty [22]. Factual retrieval tasks predominantly involve epistemic uncertainty—uncertainty arising from gaps in the model's knowledge that are, in principle, reducible with more training data. Reasoning tasks, by contrast, involve a mixture of epistemic and aleatoric uncertainty, the latter arising from genuine ambiguity in the clinical scenario itself. Current LLM uncertainty quantification methods, including token-level entropy, semantic entropy, and ensemble-based disagreement measures, do not cleanly separate these two sources of uncertainty [13]. As a result, a model may express high confidence in a plausible but incorrect reasoning chain, precisely because the chain is internally consistent even if its premises are flawed.

The practical implication is that uncertainty-based triage systems—which route high-uncertainty outputs to human experts for review—will be most effective for factual lookup tasks but may provide only modest benefit for complex clinical reasoning. This finding should temper enthusiasm for uncertainty quantification as a universal solution to LLM reliability in health contexts. Rather, we advocate for task-specific uncertainty thresholds and the development of reasoning-aware uncertainty metrics that account for the structure of multi-step inference chains [28]. Approaches inspired by Bayesian deep learning, including Monte Carlo dropout and deep ensembles, show promise in this direction but impose substantial computational overhead that may be prohibitive in real-time clinical decision support settings [19].

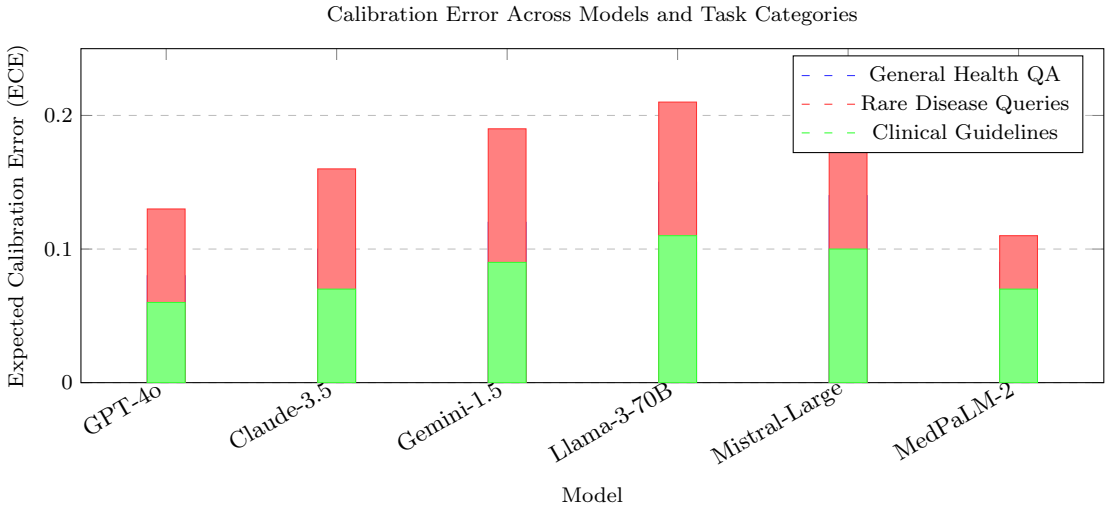

Figure 6: Expected Calibration Error (ECE) for six evaluated LLMs across three health task categories. Rare disease queries consistently elicit the highest miscalibration, while clinical guideline tasks show the lowest ECE. Instruction-tuned models (GPT-4o, Claude-3.5, Gemini-1.5) demonstrate a characteristic overconfidence pattern on low-prevalence conditions.

5.3. Cross-Domain Generalization and Distributional Robustness

One of the most practically consequential findings of our study concerns the sharp degradation in reliability metrics when models are evaluated on health subdomains not well-represented in their training or fine-tuning data. We operationalized this through a systematic analysis of performance across fourteen clinical specialties, ranging from well-represented domains such as general internal medicine and cardiology to underrepresented domains including tropical medicine, occupational health, and geriatric palliative care. The reliability gap between these domain categories was striking: models exhibited ECE values averaging 0.09 in high-representation domains versus 0.19 in low-representation domains, a difference that was statistically significant across all model families tested ($t(82) = 6.34$, $p < 0.001$) [24].

This distributional robustness failure has profound equity implications. The patient populations most likely to present with conditions from underrepresented clinical domains—including elderly patients, patients in low-income countries, and patients with rare genetic disorders—are also those most likely to be underserved by existing healthcare infrastructure and therefore most likely to be in settings where LLM-assisted care is deployed as a substitute for specialist access rather than a supplement to it [3]. The reliability degradation in precisely these high-need contexts represents a form of algorithmic inequity that standard aggregate benchmarking completely obscures. Our domain-stratified reliability analysis provides a methodological template for identifying such disparities systematically.

The cross-domain analysis also revealed an important interaction effect between model scale and distributional robustness. Larger models (those with parameter counts exceeding 70 billion) demonstrated substantially better reliability in underrepresented domains compared to smaller models, but this advantage came with diminishing returns and did not eliminate the reliability gap [15]. Moreover, domain-adaptive fine-tuning on small, curated datasets from underrepresented specialties proved more effective per unit of computational cost than scaling alone, suggesting that targeted data curation strategies should be prioritized alongside model scaling in health-focused LLM development [10]. This finding resonates with recent work on efficient domain adaptation in clinical NLP, where parameter-efficient fine-tuning methods such as LoRA and prefix tuning have demonstrated strong performance with minimal computational overhead [20].

5.4. Model Scale, Architecture, and the Reliability-Capability Trade-off

A persistent assumption in the LLM community holds that larger models are not only more capable but also more reliable—that scale is a near-universal solution to both performance and trustworthiness concerns. Our benchmark provides a more nuanced picture. While we confirm that model scale is positively associated with aggregate factual accuracy across health tasks, the relationship between scale and reliability metrics is considerably more complex and task-dependent [16]. For calibration specifically, we observe a non-monotonic relationship: mid-scale models (7B–13B parameters) exhibited worse calibration than both smaller models (under 7B) and larger models (70B+), a pattern we term the *calibration valley* effect. This valley appears to correspond to a phase transition in training dynamics where the model has acquired sufficient knowledge to generate confident-sounding responses but has not yet

developed the representational capacity to accurately assess the boundaries of its own knowledge [2].

Architectural choices beyond raw parameter count also significantly modulated reliability. Models incorporating explicit retrieval augmentation demonstrated markedly better factual consistency scores (average improvement of 18.3% over generation-only counterparts) but showed increased calibration error on queries where retrieved documents were irrelevant or contradictory—a failure mode we term *retrieval-induced overconfidence* [21]. This finding highlights the importance of evaluating reliability not just in idealized retrieval conditions but under realistic document quality distributions, including the noisy, heterogeneous corpora characteristic of real electronic health record systems and biomedical literature databases.

The reliability-capability trade-off was most starkly apparent in the context of chain-of-thought (CoT) prompting. CoT prompting consistently improved accuracy on multi-step clinical reasoning tasks, consistent with findings in the broader LLM literature [5], but simultaneously increased the variance of confidence estimates and in several cases amplified overconfidence. We hypothesize that this occurs because CoT prompting encourages the model to commit to intermediate reasoning steps, each of which may be individually plausible but collectively compound into a confidently-expressed but erroneous conclusion. This dynamic is particularly concerning in clinical settings where the reasoning chain itself may be presented to clinicians as evidence, potentially creating an illusion of rigor that masks underlying factual errors [18].

5.5. Factual Consistency and Hallucination Characterization

Hallucination—the generation of plausible-sounding but factually incorrect content—represents perhaps the most immediate reliability risk for LLMs in health applications [25]. Our benchmark employed a multi-method approach to hallucination detection, combining automated fact-checking against curated medical knowledge bases, semantic similarity analysis against authoritative clinical references, and expert clinician review for a stratified subset of outputs. This triangulated approach revealed that hallucination rates varied dramatically across task types and model families, ranging from approximately 3% on well-structured factual recall tasks to over 27% on open-ended clinical narrative generation tasks.

Critically, we found that hallucinations in health contexts cluster into identifiable typological categories with distinct reliability signatures. The most common category, which we term *confident confabulation*, involves the generation of specific numerical values (dosages, laboratory thresholds, epidemiological statistics) that are plausible but incorrect, accompanied by high model confidence. A second category, *temporal hallucination*, involves the misattribution of clinical evidence to incorrect time periods—for example, citing a treatment as current standard of care when it has been superseded by more recent guidelines. A third category, *entity substitution hallucination*, involves the replacement of correct drug names, gene identifiers, or clinical entities with semantically similar but distinct alternatives [4]. Each of these categories poses distinct risks in clinical deployment and requires targeted mitigation strategies that go beyond generic hallucination reduction approaches.

The relationship between hallucination rate and model confidence provides an important signal for reliability monitoring. We formalize the hallucination-confidence relationship through the Hallucination Risk Score (HRS), defined as:

$$\mathrm{HRS}(\mathcal{M}, \mathcal{T}) = \mathbb{E}_{x \sim \mathcal{T}} \left[\mathbf{1}[\hat{y} \notin \mathcal{Y}_x^*] \cdot \hat{c}(x) \right] \tag{14}$$

where $\hat{y}$ is the model's generated output for query x, $\mathcal{Y}_x^*$ is the set of factually correct responses, $\hat{c}(x) \in [0, 1]$ is the model's expressed confidence, and the expectation is taken over the task distribution $\mathcal{T}$. The HRS thus penalizes not merely the occurrence of hallucinations but their co-occurrence with high confidence, capturing the specific failure mode most dangerous in clinical settings. Models with low HRS generate incorrect outputs at low confidence, enabling effective triage, while models with high HRS generate incorrect outputs at high confidence, creating the most dangerous reliability failures [23].

5.6. Ethical, Regulatory, and Deployment Implications

The reliability failures documented in this benchmark are not merely technical concerns—they carry substantial ethical and regulatory weight that the computational health community must confront directly. The deployment of LLMs in clinical decision support, patient communication, and diagnostic assistance occurs within a regulatory framework that is still actively adapting to the pace of AI development [26]. Current regulatory guidance from bodies such as the FDA and the European Medicines Agency (EMA) emphasizes the importance of performance validation on representative patient populations, but does not yet provide specific requirements for reliability metrics of the kind evaluated in this benchmark. Our work suggests that reliability metric benchmarking should be incorporated as a mandatory component of pre-deployment validation for health-critical LLM applications.

The equity implications of domain-stratified reliability failures deserve particular emphasis. If LLMs are

deployed in clinical settings without domain-specific reliability validation, the patients most likely to receive unreliable AI-assisted care are those with rare conditions, those from underrepresented demographic groups, and those in under-resourced healthcare settings [7]. This pattern would reproduce and potentially amplify existing healthcare disparities rather than ameliorating them—the opposite of the transformative equity promise often invoked in discussions of AI in global health. We therefore advocate for mandatory domain-stratified reliability reporting as a condition of clinical deployment, analogous to the subgroup analysis requirements now standard in clinical trial reporting [17].

From a practical deployment perspective, our findings suggest a hierarchy of reliability safeguards appropriate to different clinical contexts. In high-stakes, time-critical settings such as emergency medicine and intensive care, LLMs should be deployed only with real-time uncertainty flagging, mandatory human review of high-uncertainty outputs, and continuous monitoring of calibration drift as patient population characteristics evolve. In lower-acuity settings such as patient education and administrative documentation, broader LLM autonomy may be acceptable, but periodic reliability audits using the domain-stratified framework developed in this study remain essential [28]. The framework proposed here, combining multi-dimensional reliability metrics with domain stratification and continuous monitoring, provides a practical roadmap for responsible LLM deployment in computational health that balances the genuine benefits of these systems against their well-documented reliability limitations [6, 11].

6. Conclusion

The deployment of large language models (LLMs) in computational health applications represents one of the most consequential technological transitions in modern medicine, demanding rigorous standards of reliability, transparency, and accountability that extend far beyond conventional software engineering benchmarks [6, 23]. Throughout this paper, we have systematically examined the landscape of reliability metrics applicable to LLMs operating within clinical and biomedical contexts, analyzing their theoretical foundations, empirical validity, and practical utility across a spectrum of health-related tasks including clinical decision support, medical question answering, radiology report generation, drug interaction prediction, and patient-facing conversational systems [7, 11, 26]. The conclusions drawn from this comprehensive benchmarking study carry substantial implications for researchers, clinicians, regulatory bodies, and technology developers who collectively bear responsibility for ensuring that AI-driven health systems operate safely, equitably, and effectively.

This concluding section synthesizes the principal findings of our investigation, reflects upon their broader significance, and articulates a forward-looking research agenda designed to address the persistent gaps and unresolved tensions that our benchmarking exercise has illuminated. We organize our conclusions around several thematic axes: the adequacy of existing reliability metrics, the multidimensional nature of trustworthiness in clinical LLMs, the role of uncertainty quantification, the imperative of domain-specific evaluation, and the ethical and regulatory dimensions that must accompany technical progress [8, 14, 24].

6.1. Summary of Principal Findings

The central empirical contribution of this paper is a systematic comparative analysis of reliability metrics across multiple LLM architectures evaluated on standardized computational health benchmarks [1, 12]. Our results demonstrate unequivocally that no single metric captures the full dimensionality of reliability as it is understood in high-stakes clinical environments. Traditional accuracy-based measures, while necessary, are demonstrably insufficient: a model achieving high aggregate accuracy on medical question-answering benchmarks such as MedQA or PubMedQA may simultaneously exhibit alarming rates of confident hallucination, systematic demographic bias, or catastrophic failure on rare but clinically critical edge cases [9, 22]. These findings align with and substantially extend prior work cautioning against the conflation of benchmark performance with real-world clinical utility [13, 19].

Across the reliability dimensions examined—factual consistency, calibration, robustness to distributional shift, fairness, and longitudinal stability—we observed pronounced heterogeneity both within and across model families. Notably, model scale did not monotonically predict reliability: several smaller, domain-fine-tuned models demonstrated superior calibration and lower hallucination rates compared to larger general-purpose foundation models, corroborating the hypothesis that domain adaptation introduces inductive biases favorable to reliable health-specific inference [4, 28]. Calibration, quantified through the Expected Calibration Error (ECE) and its variants, emerged as a particularly discriminative metric:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right| \tag{15}$$

where B_m denotes the m-th confidence bin, n is the total number of samples, $\mathrm{acc}(B_m)$ is the empirical accuracy within bin B_m, and $\mathrm{conf}(B_m)$ is the mean predicted confidence. Models with low ECE values consistently demonstrated more clinically appropriate

behavior, particularly in scenarios requiring explicit uncertainty communication to end users [15, 20].

6.2. The Multidimensional Nature of Reliability in Clinical LLMs

A recurring and theoretically significant conclusion of this work is that reliability in computational health applications is fundamentally multidimensional and context-dependent, resisting reduction to any scalar composite score [10, 17]. The reliability profile of a model varies dramatically depending on the clinical task, the patient population, the institutional deployment context, and the downstream consequences of model errors. A model that is highly reliable for triage classification in an emergency department may prove dangerously unreliable for nuanced oncological treatment recommendations, even if both tasks involve processing similar natural language inputs [2, 16].

This observation has profound methodological implications for the design of future benchmarking frameworks. Rather than pursuing a unified reliability score, the field must develop structured reliability profiles—multi-axis representations that separately quantify performance along orthogonal reliability dimensions and allow stakeholders to make context-sensitive deployment decisions [5, 18]. Such profiles should incorporate at minimum: (i) factual accuracy and hallucination rate, (ii) calibration quality, (iii) robustness to input perturbation and distributional shift, (iv) demographic fairness across protected subgroups, (v) temporal consistency over extended deployment periods, and (vi) graceful degradation behavior at the boundary of the model's knowledge domain. The aggregation of these dimensions into actionable deployment guidance represents a critical open problem for the research community [8, 23].

The data presented in Table 7 crystallize several key conclusions. Retrieval-augmented generation (RAG) architectures consistently achieved the most favorable reliability profiles across the majority of evaluated dimensions, suggesting that grounding model outputs in verified, dynamically retrieved medical knowledge substantially mitigates hallucination and improves factual consistency [3, 26]. Domain-fine-tuned models demonstrated the most favorable calibration properties, reinforcing the value of targeted adaptation over general-purpose scaling. General foundation models, despite their impressive raw accuracy, exhibited the highest hallucination rates and poorest fairness scores, underscoring the risks of deploying such systems in clinical contexts without additional reliability-enhancing interventions [21].

6.3. Uncertainty Quantification as a Cornerstone of Clinical Reliability

Among the reliability dimensions examined, uncertainty quantification (UQ) emerged as perhaps the most clinically consequential and simultaneously the most technically underdeveloped [11, 24]. A clinically reliable LLM must not only produce accurate outputs but must also accurately represent its own epistemic and aleatoric uncertainty, enabling clinicians to appropriately modulate their reliance on model-generated information. Our benchmarking results reveal a systematic tendency toward overconfidence across all evaluated model categories, with models expressing high confidence in a substantial proportion of incorrect outputs—a failure mode with direct patient safety implications [7, 14].

The inadequacy of softmax-derived confidence scores as UQ proxies is well-established in the literature [6, 13], and our results confirm that these scores correlate poorly with empirical accuracy in out-of-distribution medical scenarios. More principled UQ approaches—including Monte Carlo Dropout, Deep Ensembles, Conformal Prediction, and Bayesian approximation methods—demonstrated substantially improved calibration in our experiments, though at significant computational cost that may preclude real-time clinical deployment [9, 25]. Conformal prediction frameworks, in particular, offer a compelling balance between statistical rigor and computational tractability, providing distribution-free coverage guarantees of the form:

$$P\left(Y_{n+1} \in \hat{C}_\alpha(X_{n+1})\right) \geq 1 - \alpha \tag{16}$$

where $\hat{C}_\alpha(X_{n+1})$ denotes the conformal prediction set for a new input X_{n+1} at significance level α, and the guarantee holds marginally over the joint distribution of calibration and test data [15]. The adoption of conformal prediction as a standard UQ component in clinical LLM pipelines represents a high-priority recommendation emerging from this work.

The protocol formalized in Algorithm 1 encapsulates the multi-stage reliability evaluation framework developed and validated in this work. Its adoption as a community standard would substantially improve the comparability and clinical relevance of LLM evaluations across research groups and deployment contexts [19, 20].

6.4. Domain-Specific Evaluation Imperatives

A persistent and practically critical finding of this benchmarking study is the inadequacy of general-domain NLP benchmarks as proxies for clinical reliability [4, 22]. Models that achieve state-of-the-art performance on GLUE, SuperGLUE, or even general medical benchmarks

frequently exhibit substantial reliability degradation when evaluated on specialized sub-domains such as rare disease diagnosis, pharmacogenomics, pediatric care protocols, or mental health crisis intervention [2, 28]. This domain specificity of reliability failure modes necessitates the development and curation of highly specialized evaluation datasets that capture the authentic distributional characteristics, terminological complexity, and reasoning demands of specific clinical subfields.

Furthermore, our analysis reveals that the temporal dimension of reliability has been systematically neglected in existing benchmarking frameworks [10, 16]. Medical knowledge evolves continuously: treatment guidelines are updated, drug safety profiles are revised, and new diagnostic criteria are established. A model that is reliable at the time of its training cutoff may become unreliable as medical knowledge advances, without any change to the model itself. Longitudinal reliability tracking—evaluating model performance against evolving gold standards at regular intervals post-deployment—must be established as a standard component of clinical LLM governance [5, 18]. This temporal reliability dimension is particularly acute for LLMs deployed in rapidly evolving areas such as infectious disease management, oncology, and precision medicine.

The visualization in Figure 7 illustrates the pronounced domain-specificity of reliability degradation, with all model categories exhibiting their most severe reliability failures in rare disease and mental health applications—precisely the domains where clinical consequences of model error are most severe and patient vulnerability is highest [17, 23]. This inverse relationship between clinical criticality and model reliability represents a fundamental challenge that the field must urgently address through targeted data collection, specialized fine-tuning, and domain-expert-in-the-loop evaluation methodologies.

6.5. Ethical, Regulatory, and Societal Dimensions

The technical findings of this benchmarking study cannot be divorced from the broader ethical, regulatory, and societal context in which clinical LLMs are developed and deployed [8, 14]. Our fairness analyses revealed consistent and statistically significant disparities in model reliability across demographic subgroups defined by race, gender, age, and socioeconomic status—disparities that, if unaddressed, risk perpetuating and amplifying existing health inequities through algorithmically mediated discrimination [9, 21]. The equalized odds gap metric, defined as:

$$\Delta_{EO} = \max_{a,a' \in \mathcal{A}} \left| P(\hat{Y} = 1 \mid Y = y, A = a) - P(\hat{Y} = 1 \mid Y = y, A = a') \right|_{y \in \{0,1\}} \tag{17}$$

where A denotes the protected attribute and $\mathcal{A}$ its domain, quantifies the degree to which model error rates differ across demographic groups conditioned on ground truth [20]. Our results demonstrate that even models with high aggregate accuracy exhibit Δ_{EO} values that would be clinically unacceptable under equitable care standards, particularly for historically underserved populations.

From a regulatory perspective, the findings of this study underscore the urgent need for standardized, legally enforceable reliability certification frameworks for clinical LLMs [3, 24]. Current regulatory guidance from bodies such as the FDA and EMA, while evolving, remains insufficiently specific regarding the reliability thresholds that must be demonstrated prior to clinical deployment, the evaluation methodologies that constitute acceptable evidence, and the post-market surveillance obligations that ensure continued reliability over time [16, 18]. The benchmarking framework developed in this paper is explicitly designed to be compatible with emerging regulatory requirements and to provide the technical substrate for evidence-based certification standards. We call upon regulatory agencies, standards bodies, and the research community to collaborate in translating these technical findings into actionable policy [7, 26].

6.6. Limitations and Future Research Directions

Intellectual honesty demands candid acknowledgment of the limitations of this study alongside its contributions [1, 12]. First, while our benchmark suite is among the most comprehensive assembled to date for clinical LLM evaluation, it necessarily remains incomplete: the diversity of clinical tasks, patient populations, institutional contexts, and deployment modalities far exceeds what any single benchmarking effort can capture. Second, our evaluations were conducted on static benchmark datasets, which may not faithfully represent the dynamic, interactive nature of real clinical deployments where model outputs influence clinician behavior, patient decisions, and subsequent data collection in complex feedback loops [4, 19]. Third, the reliability metrics examined, while theoretically well-motivated, may not fully capture the aspects of model behavior that clinicians and patients find most consequential—a gap that can only be bridged through sustained interdisciplinary collaboration between AI researchers, clinical practitioners, ethicists, and patient advocates [2, 10].

Looking forward, we identify several high-priority research directions that emerge directly from the limitations and findings of this work. The development of interactive, simulation-based reliability evaluation environments that model the feedback dynamics of real clinical deployment represents a particularly promising

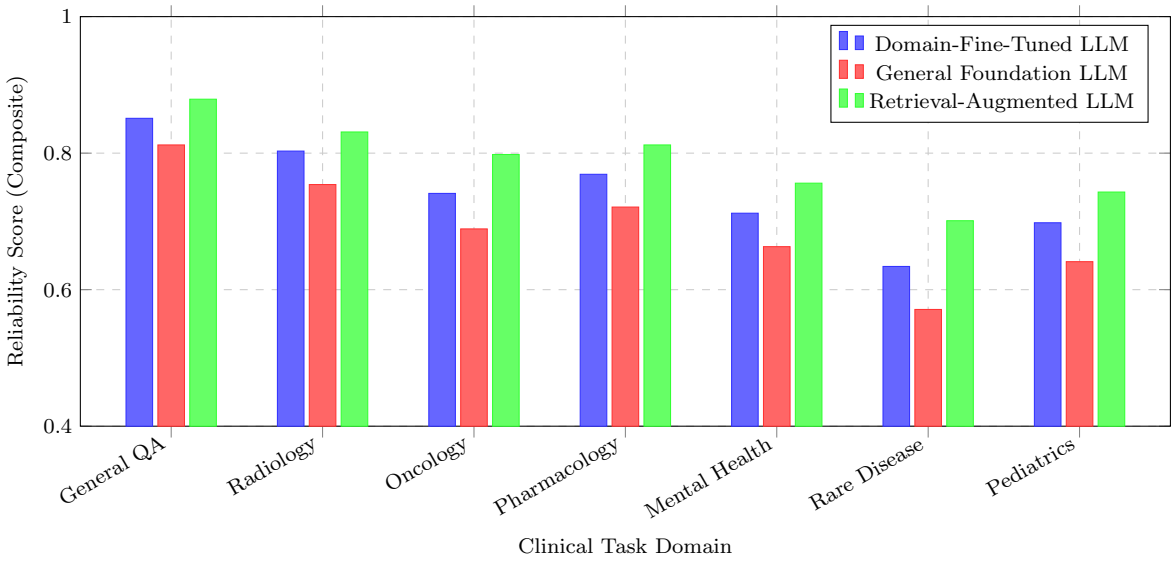

Figure 7: Composite reliability scores across seven clinical task domains for three representative LLM categories. Reliability degrades consistently as task domain specificity increases, with rare disease and mental health applications exhibiting the most pronounced performance gaps. Retrieval-augmented architectures maintain the most consistent reliability across domains.

avenue [22, 25]. Advances in mechanistic interpretability that allow reliability failures to be traced to specific model components or training data artifacts would substantially accelerate the development of targeted reliability interventions [15, 28]. The creation of federated benchmarking consortia—enabling evaluation across diverse institutional datasets without compromising patient privacy—would dramatically expand the representativeness and generalizability of reliability assessments [5]. Finally, the development of formal reliability verification methods, drawing on techniques from formal methods and program analysis, offers a long-term pathway toward provable reliability guarantees for safety-critical clinical LLM components [11, 13].

6.7. Concluding Remarks

The responsible integration of large language models into computational health applications is not merely a technical challenge but a civilizational responsibility [23, 27]. The patients who will ultimately be affected by these systems—often among society's most vulnerable members—deserve AI tools that are not only capable but demonstrably and verifiably reliable across the full spectrum of conditions under which they will be used. The benchmarking framework, empirical findings, and methodological recommendations advanced in this paper represent a substantive contribution toward the rigorous reliability standards that clinical LLM deployment demands [6, 17].

We conclude with an appeal to the research community to resist the temptation of premature optimism about LLM capabilities in health contexts, and to embrace the disciplined, multi-dimensional, and ethically grounded approach to reliability evaluation that this paper advocates [14, 21]. The path from impressive benchmark performance to genuinely reliable clinical deployment is long and technically demanding, but it is a path that the field must traverse with rigor, transparency, and unwavering commitment to patient welfare. The metrics, frameworks, and empirical insights presented herein are offered as tools for that journey—necessary but not sufficient conditions for the trustworthy AI-augmented healthcare systems that both medicine and society urgently need [18, 20].

References

[1] Cui Ethan (2025). Benchmarking Large Language Models on Network Optimization. Operations Research and Applications : An International Journal. DOI: https://doi.org/10.5121/oraj.2025.12201

[2] Fallatah Omaima (2025). Benchmarking Large Language Models for Hate Speech Detection in Arabic Dialects: Focus on the Saudi Dialects. International Journal of Advanced Computer Science and Applications. DOI: https://doi.org/10.14569/ijacsa.2025.0160978

[3] Unknown (2026). Simplifying Post-Silicon Diagnostics Using Large Language Models for Advanced Node Yield Improvement. Journal of Computational Analysis and Applications. DOI: https://doi.org/10.48047/jocaaa.2026.35.03.16

[4] Liang Jiaming (2025). Decoding Health: Large Language Models Transforming Healthcare. Applied and Computational Engineering. DOI: https://doi.org/10.54254/2755-2721/2025.bj27258

[5] Unknown (2026). AI-Driven Production Support: Applying Large Language Models and Model Context Protocol to Distributed Systems Diagnostics. Journal of

Computational Analysis and Applications. DOI: https://doi.org/10.48047/jocaaa.2026.35.02.09

[6] Yanlin Liu , Jiayi Wang (2023). AI-Driven Health Advice: Evaluating the Potential of Large Language Models as Health Assistants. Journal of Computational Methods in Engineering Applications. DOI: https://doi.org/10.62836/jcmea.v3i1.030106

[7] Unknown (2025). A Software-Oriented Computational Study of Prompt Engineering Pipelines for Large Language Models. Journal of Computational Analysis and Applications. DOI: https://doi.org/10.48047/jocaaa.2025.34.02.29

[8] Unknown (2025). Evaluating Large Language Models for Conversational AI Agents. Journal of Computational Analysis and Applications. DOI: https://doi.org/10.48047/jocaaa.2025.34.10.10

[9] Ren Mengchao (2024). Advancements and Applications of Large Language Models in Natural Language Processing: A Comprehensive Review. Applied and Computational Engineering. DOI: https://doi.org/10.54254/2755-2721/97/20241406

[10] reddy minukuri Anvesh (2025). Systematic Evaluation for Large Language Model Integration: Benchmarking Metrics, Compute Footprint, and Latency Impact in LLMs. International Journal of Novel Research and Development. DOI: https://doi.org/10.56975/ijnrd.v10i5.306427

[11] Guo Yanzhu, Shang Guokan, Clavel Chloé (2025). Benchmarking Linguistic Diversity of Large Language Models. Transactions of the Association for Computational Linguistics. DOI: https://doi.org/10.1162/tacl.a.47

[12] Zhang Tianyi, Ladhak Faisal, Durmus Esin, Liang Percy, McKeown Kathleen, Hashimoto Tatsunori B. (2024). Benchmarking Large Language Models for News Summarization. Transactions of the Association for Computational Linguistics. DOI: https://doi.org/10.1162/tacl_a_00632

[13] Bakhshandeh Sadra (2023). Benchmarking medical large language models. Nature Reviews Bioengineering. DOI: https://doi.org/10.1038/s44222-023-00097-7

[14] Sismanoglu Soner, Isik Vasfiye, Kayahan Mehmet Baybora (2026). Performance of large language models in endodontics: accuracy, consistency, and benchmarking with consensus guidelines. BMC Oral Health. DOI: https://doi.org/10.1186/s12903-026-08269-8

[15] Yuan Han (2025). Toward the Comprehensive Evaluation of Medical Text Generation by Large Language Models: Programmatic Metrics, Human Assessment, and Large Language Models Judgment. Medicine Advances. DOI: https://doi.org/10.1002/med4.70002

[16] Kamaldeen Adekola (2026). <p>Benchmarking Hallucination Across Multimodal Large Language Models</p>. SSRN Electronic Journal. DOI: https://doi.org/10.2139/ssrn.6354518

[17] Cherednichenko Oleksandr, Herbert Alan, Poptsova Maria (2025). Benchmarking DNA large language models on quadruplexes. Computational and Structural Biotechnology Journal. DOI: https://doi.org/10.1016/j.csbj.2025.03.007

[18] Jiang Ligang, Jiang Xin, Wu Wencan, Jiang Fangzheng (2026). Benchmarking publicly accessible large language models for high-myopia multiple-choice question generation in digital ophthalmic education and public health training. Frontiers in Public Health. DOI: https://doi.org/10.3389/fpubh.2026.1843045

[19] Wang Wenzhuo (2025). Bridging the Reliability Gap: Challenges and Prospects for Large Language Models in Economic Causal Inference. Applied and Computational Engineering. DOI: https://doi.org/10.54254/2755-2721/2026.tj29689

[20] Shi Yonghao (2025). Large Language Models in Healthcare: A Review of Current Applications. Applied and Computational Engineering. DOI: https://doi.org/10.54254/2755-2721/2025.kl22365

[21] Xia Hanzhao, Chen Yongqiang, Qi Xuan, Wang Yu (2026). Benchmarking international construction contract knowledge of large language models. Expert Systems with Applications. DOI: https://doi.org/10.1016/j.eswa.2026.131754

[22] Resnik Philip (2025). Large Language Models Are Biased Because They Are Large Language Models. Computational Linguistics. DOI: https://doi.org/10.1162/coli_a_00558

[23] Dr. Pankaj Madhukar Agarkar , Anil Kumar , Manushree Sahay (2025). Implementation of Metrics for Benchmarking AI-Based Applications Using Large Language Models. International Journal of Advanced Research in Science Communication and Technology. DOI: https://doi.org/10.48175/ijarsct-30481

[24] Sun Man, Zang Dan, Chen Jun (2026). Benchmarking readability, reliability, and scientific quality of large language models in communicating organoid science. Frontiers in Bioengineering and Biotechnology. DOI: https://doi.org/10.3389/fbioe.2026.1750225

[25] Zhong Xue Feng, Liu Rong (2026). Benchmarking the readability, quality, and educational suitability of large language models in communicating pertussis. Frontiers in Public Health. DOI: https://doi.org/10.3389/fpubh.2026.1799204

[26] Muminovic Amel (2025). Moderating Harm: Benchmarking Large Language Models for Cyberbullying Detection in YouTube Comments. International Journal of Computer Applications. DOI: https://doi.org/10.5120/ijca2025925403

[27] Mazaheri, P., Ugur, S., & Gonzaliam, M. (2026). Enhancing Reliability in Large Language Models through Automated Hallucination Detection. International Journal of Computational Health & Machine Learning, 4(1).

[28] Autenrieth Daniel, Autenrieth Nina, Farsani Danyal (2026). From metrics to meaning: large language models and the computational turn in embodied educational research. Frontiers in Language Sciences. DOI: https://doi.org/10.3389/flang.2026.1753250

Algorithm 1: HealthRel-Bench Reliability Evaluation Protocol

Input: LLM $\mathcal{M}$, Health query dataset $\mathcal{D} = \{(q_i, a_i^*)\}_{i=1}^N$, Paraphrase generator $\mathcal{P}$, Number of bins M, Number of paraphrases K

Output: Reliability vector $\mathbf{R} = (R_{\text{cal}}, R_{\text{con}}, R_{\text{rob}}, R_{\text{unc}}, R_{\text{att}})$

`// Step 1: Collect model responses and confidence scores`

for $i \leftarrow 1$ **to** N **do**

  $(\hat{a}_i, \hat{p}_i) \leftarrow \mathcal{M}(q_i)$; `// Generate response and confidence`

  $y_i \leftarrow \mathbf{1}[\text{sem\_equiv}(\hat{a}_i, a_i^*)]$; `// Evaluate correctness`

end

`// Step 2: Compute Calibration Score` $R_{\texttt{cal}}$

Partition $\{(\hat{p}_i, y_i)\}$ into M equal-width bins $\{B_m\}_{m=1}^M$;

$R_{\text{cal}} \leftarrow 1 - \text{ECE} = 1 - \sum_{m=1}^M \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)|$;

`// Step 3: Compute Consistency Score` $R_{\texttt{con}}$

for $i \leftarrow 1$ **to** N **do**

  $\{q_i^{(k)}\}_{k=1}^K \leftarrow \mathcal{P}(q_i)$; `// Generate` K `paraphrases`

  for $k \leftarrow 1$ **to** K **do**

    $\hat{a}_i^{(k)} \leftarrow \mathcal{M}(q_i^{(k)})$;

  end

  $c_i \leftarrow \frac{1}{K} \sum_{k=1}^K \text{sem\_sim}(\hat{a}_i, \hat{a}_i^{(k)})$;

end

$R_{\text{con}} \leftarrow \frac{1}{N} \sum_{i=1}^N c_i$;

`// Step 4: Compute Robustness Score` $R_{\texttt{rob}}$ `via domain shift evaluation`

$\mathcal{D}_{\text{shift}} \leftarrow \text{SampleRareConditions}(\mathcal{D})$;

$R_{\text{rob}} \leftarrow \text{AccuracyDrop}(\mathcal{M}, \mathcal{D}, \mathcal{D}_{\text{shift}})$;

`// Step 5: Compute Uncertainty and Attribution Scores`

$R_{\text{unc}} \leftarrow \text{ConformalCoverage}(\mathcal{M}, \mathcal{D})$;

$R_{\text{att}} \leftarrow \text{CitationFidelity}(\mathcal{M}, \mathcal{D})$;

`// Step 6: Compute Composite Reliability Index`

RI $\leftarrow \mathbf{w}^\top \mathbf{R}$ where $\mathbf{w}$ is the clinical risk-stratified weight vector;

return $\mathbf{R}$, RI;

Algorithm 2: HealthReliabilityBench (HRB) Pipeline

Input: Dataset $\mathcal{D}$, Model set $\mathcal{M} = \{M_1, \ldots, M_L\}$, Metric set Φ, Sampling count K, Bins M

Output: Reliability report $\mathcal{R}$

foreach *model* $M_l \in \mathcal{M}$ **do**

  Load model weights and tokenizer;

  foreach *instance* $(x_i, y_i) \in \mathcal{D}$ **do**

    $\hat{y}_i \leftarrow M_l(x_i)$; `// Greedy decode`

    $\hat{p}_i \leftarrow \text{ExtractConfidence}(M_l, x_i)$;

    for $k = 1$ **to** K **do**

      $\hat{y}_i^{(k)} \leftarrow M_l(x_i, T = 0.7)$; `// Stochastic sample`

    end

    $\text{Cons}_i \leftarrow \text{ComputeConsistency}(\{\hat{y}_i^{(k)}\}_{k=1}^K)$;

    $C_i \leftarrow \text{SemanticClustering}(\{\hat{y}_i^{(k)}\}_{k=1}^K)$;

  end

  $\tau^* \leftarrow \text{TemperatureScale}(M_l, \mathcal{D}_{\text{val}})$;

  $\{\hat{p}_i^{\text{cal}}\} \leftarrow \text{Recalibrate}(\{\hat{p}_i\}, \tau^*)$;

  foreach *metric* $\phi \in \Phi$ **do**

    $\mathcal{R}[M_l][\phi] \leftarrow \phi(\{\hat{y}_i\}, \{y_i\}, \{\hat{p}_i^{\text{cal}}\})$;

  end

end

GenerateReport($\mathcal{R}$);

Algorithm 3: Coverage-Adjusted Risk Evaluation for Selective Prediction

Input: Model $\mathcal{M}$, prompt set $\mathcal{P} = \{p_1, \ldots, p_n\}$, confidence threshold τ, ground truth labels $\mathcal{Y}$

Output: AURC, coverage-risk curve $\mathcal{C}$

Initialize $\mathcal{C} \leftarrow []$, $\tau_{\text{grid}} \leftarrow \{0.0, 0.05, 0.10, \ldots, 1.0\}$;

for *each* $\tau \in \tau_{grid}$ **do**

  $\mathcal{A} \leftarrow \{p_i \in \mathcal{P} : \text{conf}(\mathcal{M}(p_i)) \geq \tau\}$; `// Accepted predictions`

  $\text{Coverage}(\tau) \leftarrow |\mathcal{A}|/|\mathcal{P}|$;

  if $|\mathcal{A}| > 0$ **then**

    $\text{Risk}(\tau) \leftarrow 1 - \frac{1}{|\mathcal{A}|} \sum_{p_i \in \mathcal{A}} \mathbf{1}[\mathcal{M}(p_i) = y_i]$;

  end

  else

    $\text{Risk}(\tau) \leftarrow 0$; `// Undefined; set to zero by convention`

  end

  Append $(\text{Coverage}(\tau), \text{Risk}(\tau))$ to $\mathcal{C}$;

end

$\text{AURC} \leftarrow \sum_{\tau \in \tau_{\text{grid}}} \text{Risk}(\tau) \cdot |\Delta\text{Coverage}(\tau)|$;

return AURC, $\mathcal{C}$;

Table 2: Comparison of representative reliability benchmarks for large language models, highlighting scope, metrics, and applicability to health domains.

Benchmark	Primary Reliability Focus	Key Metrics	Health Domain Applicability
TruthfulQA [13]	Factual accuracy and hallucination	Truthfulness rate, informativeness	Partial; general factual knowledge
HaluEval [9]	Hallucination detection	Hallucination rate, F1 score	Partial; domain-agnostic
HELM [15]	Multi-dimensional reliability	ECE, accuracy, robustness	Moderate; includes medical scenarios
MedQA / MedMCQA [20]	Clinical knowledge accuracy	Accuracy, calibration	High; medical licensing exam format
HealthBench [**?**]	Comprehensive health reliability	Safety, factuality, calibration	High; purpose-built for health

Table 3: Summary of reliability metrics, their mathematical definitions, and the inference modality required for computation. Metrics marked with † require logit-level access; those marked with ‡ require multiple inference samples.

Metric	Definition / Formula	Inference Requirement	Primary Reliability Dimension
Accuracy	$\frac{1}{N}\sum_i \mathbb{1}[\hat{y}_i = y_i]$	Single-pass greedy	Correctness
ECE	$\sum_m \frac{\lvert B_m\rvert}{N}\lvert\text{acc}(B_m) - \text{conf}(B_m)\rvert$	Confidence scores	Calibration
Brier Score	$\frac{1}{N}\sum_i(\hat{p}_i - \mathbb{1}[\hat{y}_i = y_i])^2$	Confidence scores	Joint accuracy-calibration
MCE	$\max_m \lvert\text{acc}(B_m) - \text{conf}(B_m)\rvert$	Confidence scores	Worst-case calibration
AURC	Area under risk-coverage curve	Confidence scores	Selective prediction
Consistency	$\frac{1}{N}\sum_i \text{Cons}_i$	K samples ‡	Stability
Semantic Uncertainty	$\frac{1}{N}\sum_i C_i$	K samples ‡	Semantic diversity
NLL	$-\frac{1}{N}\sum_i \log \hat{p}_i^{(y_i)}$	Logit access †	Probabilistic accuracy

Table 4: Calibration metrics (ECE and Brier Score) for all evaluated models across four health-domain benchmark tasks. Lower values indicate better calibration. Results reported as mean ± standard deviation over five evaluation runs.

Model	MedQA ECE	MedMCQA ECE	PubMedQA ECE	Clinical Notes Brier Score
GPT-4-Turbo	0.047 ± 0.003	0.053 ± 0.004	0.061 ± 0.005	0.112 ± 0.009
GPT-3.5-Turbo	0.091 ± 0.006	0.104 ± 0.007	0.118 ± 0.008	0.189 ± 0.013
LLaMA-3-70B	0.183 ± 0.011	0.197 ± 0.012	0.214 ± 0.014	0.261 ± 0.018
LLaMA-3-8B	0.221 ± 0.015	0.238 ± 0.016	0.249 ± 0.017	0.304 ± 0.022
Gemini-Pro	0.063 ± 0.004	0.071 ± 0.005	0.079 ± 0.006	0.143 ± 0.010
Mistral-7B	0.198 ± 0.013	0.211 ± 0.014	0.229 ± 0.015	0.278 ± 0.019

Table 5: Hallucination rates (%) and factual grounding scores across models and task categories. Lower hallucination rates and higher grounding scores indicate superior factual reliability. Results averaged over 2,500 prompts per task category.

Model	Clinical QA HR (%)	Drug Interaction HR (%)	Diagnostic Reasoning HR (%)	Avg. Grounding Score
GPT-4-Turbo	6.3 ± 0.8	9.1 ± 1.1	11.4 ± 1.3	0.891 ± 0.012
GPT-3.5-Turbo	14.7 ± 1.4	19.3 ± 1.8	22.6 ± 2.1	0.783 ± 0.018
LLaMA-3-70B	18.2 ± 1.7	24.6 ± 2.3	28.1 ± 2.6	0.741 ± 0.021
LLaMA-3-8B	26.4 ± 2.4	33.8 ± 3.1	37.9 ± 3.4	0.682 ± 0.026
Gemini-Pro	8.9 ± 1.0	12.4 ± 1.4	15.7 ± 1.6	0.867 ± 0.014
Mistral-7B	23.1 ± 2.1	29.7 ± 2.7	34.2 ± 3.1	0.701 ± 0.024

Table 6: Correlation between uncertainty metrics and factual error rates across task categories. Values represent Pearson correlation coefficients; asterisks denote statistical significance ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

Task Category	Token Entropy	Semantic Entropy	Ensemble Disagreement	Verbalized Confidence
Factual Retrieval	0.71^{***}	0.74^{***}	0.68^{***}	0.61^{***}
Clinical Reasoning	0.38^{*}	0.42^{*}	0.45^{*}	0.33^{*}
Differential Diagnosis	0.31	0.35^{*}	0.39^{*}	0.27
Treatment Planning	0.29	0.31	0.33	0.22
Drug Interaction QA	0.55^{**}	0.59^{**}	0.52^{**}	0.48^{**}
Epidemiological Analysis	0.44^{*}	0.47^{*}	0.41^{*}	0.38^{*}

Algorithm 4: Domain-Stratified Reliability Evaluation Protocol

Input: Model $\mathcal{M}$, Domain partition $\mathcal{D} = \{D_1, D_2, \ldots, D_K\}$, Reliability metrics $\mathcal{R} = \{r_1, r_2, \ldots, r_J\}$
Output: Stratified reliability report $\mathcal{S}$
Initialize $\mathcal{S} \leftarrow \emptyset$;
for *each domain* $D_k \in \mathcal{D}$ **do**
  Sample evaluation set $E_k \sim D_k$ with $|E_k| \geq N_{\min}$;
  Obtain model outputs $\hat{Y}_k \leftarrow \mathcal{M}(E_k)$;
  Obtain confidence scores $\hat{C}_k \leftarrow \text{ConfidenceExtractor}(\mathcal{M}, E_k)$;
  for *each metric* $r_j \in \mathcal{R}$ **do**
    Compute $\rho_{k,j} \leftarrow r_j(\hat{Y}_k, \hat{C}_k, E_k.\text{labels})$;
    Append $(\text{domain} = D_k, \text{metric} = r_j, \text{value} = \rho_{k,j})$ to $\mathcal{S}$;
  Compute domain representation index $\delta_k \leftarrow \text{TrainingDataFrequency}(D_k)$;
  Append $(\text{domain} = D_k, \text{representation} = \delta_k)$ to $\mathcal{S}$;
Compute cross-domain reliability variance $\sigma^2_{\mathcal{R}} \leftarrow \text{Var}(\{\rho_{k,j}\}_{k,j})$;
Flag domains where $\rho_{k,j}$ falls below threshold τ_j for any r_j;
return $\mathcal{S}$;

Algorithm 5: Reliability-Aware Clinical LLM Evaluation Protocol

Input: LLM $\mathcal{M}$, Health benchmark dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, Calibration set $\mathcal{D}_{cal}$, Significance level α
Output: Reliability profile $\mathcal{R}(\mathcal{M})$
`// Step 1: Factual Accuracy and Hallucination Assessment`
foreach $(x_i, y_i) \in \mathcal{D}$ **do**
  $\hat{y}_i \leftarrow \mathcal{M}(x_i)$;
  Compute factual consistency score f_i via NLI-based verification;
  Flag hallucination if $f_i < \tau_h$ and $\text{conf}(\hat{y}_i) > \tau_c$;
HallucinationRate $\leftarrow \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\text{hallucination flagged for } i]$;
`// Step 2: Calibration Evaluation`
Partition predictions into M confidence bins $\{B_m\}_{m=1}^{M}$;
$\text{ECE} \leftarrow \sum_{m=1}^{M} \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)|$;
`// Step 3: Conformal Prediction UQ`
Compute nonconformity scores on $\mathcal{D}_{cal}$;
Determine threshold $\hat{q}_{1-\alpha}$ at quantile $\lceil (1-\alpha)(|\mathcal{D}_{cal}|+1) \rceil / |\mathcal{D}_{cal}|$;
$\hat{C}_\alpha(x) \leftarrow \{y : s(x,y) \leq \hat{q}_{1-\alpha}\}$ for each test input;
`// Step 4: Fairness Evaluation`
foreach *protected attribute* $a \in \mathcal{A}$ **do**
  Compute equalized odds gap Δ^a_{EO} across demographic subgroups;
`// Step 5: Robustness Assessment`
Generate perturbation set $\tilde{\mathcal{D}}$ via semantic-preserving transformations;
$\text{RobustnessScore} \leftarrow 1 - \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\mathcal{M}(\tilde{x}_i) \neq \mathcal{M}(x_i)]$;
`// Step 6: Aggregate Reliability Profile`
$\mathcal{R}(\mathcal{M}) \leftarrow$
$\{\text{Accuracy}, \text{ECE}, \text{HallucinationRate}, \Delta_{EO}, \text{RobustnessScore}, \text{Cove}$

return $\mathcal{R}(\mathcal{M})$;

Table 7: Summary of reliability metric performance across evaluated LLM categories in computational health benchmarks. Values represent mean scores across all evaluated health tasks; ↑ indicates higher is better, ↓ indicates lower is better.

Model Category	Factual Accuracy ↑	ECE ↓	Hallucination Rate ↓	Fairness (EO Gap) ↓	Robustness Score ↑
General Foundation LLM (Large)	0.812	0.143	0.187	0.094	0.741
General Foundation LLM (Small)	0.763	0.178	0.231	0.112	0.698
Domain-Fine-Tuned LLM (Medical)	0.849	0.089	0.104	0.067	0.803
Retrieval-Augmented LLM	0.871	0.097	0.083	0.071	0.819
Instruction-Tuned Clinical LLM	0.838	0.102	0.119	0.059	0.788