Beyond the Leaderboard: Rethinking Medical Benchmarks for Large Language Models

Wenting Chen[1] **Guo Yu**[2] **Yiu-Fai Cheung**[3] **Meidan Ding**[2]
Jie Liu[4] **Zizhan Ma**[3†] **Wenxuan Wang**[5] **Linlin Shen**[2]
[1] Stanford University [2] Shenzhen University [3] The Chinese University of Hong Kong
[4] City University of Hong Kong [5] Renmin University of China
wentchen@stanford.edu wangwenxuan@ruc.edu.cn zizhan.ma@link.cuhk.edu.hk

Abstract

Large language models (LLMs) show significant potential in healthcare, prompting numerous benchmarks to evaluate their capabilities. However, concerns persist regarding the reliability of these benchmarks, which often lack clinical fidelity, robust data management, and safety-oriented evaluation metrics. To address these shortcomings, we introduce *MedCheck*, the first lifecycle-oriented assessment framework designed for medical benchmarks. Our framework deconstructs benchmark development into five stages from design to governance, and provides a comprehensive checklist of 46 medically-tailored criteria. Using *MedCheck*, we conducted an in-depth empirical evaluation of 56 medical LLM benchmarks. Our analysis uncovers widespread, systemic issues, including a profound disconnect from clinical practice, a crisis of data integrity due to unmitigated contamination risks, and a systematic neglect of safety-critical evaluation dimensions like model robustness and uncertainty awareness. Based on these findings, *MedCheck* serves as a diagnostic framework to audit existing benchmarks and an actionable guideline for a more standardized, reliable, and transparent approach to evaluating AI in healthcare.

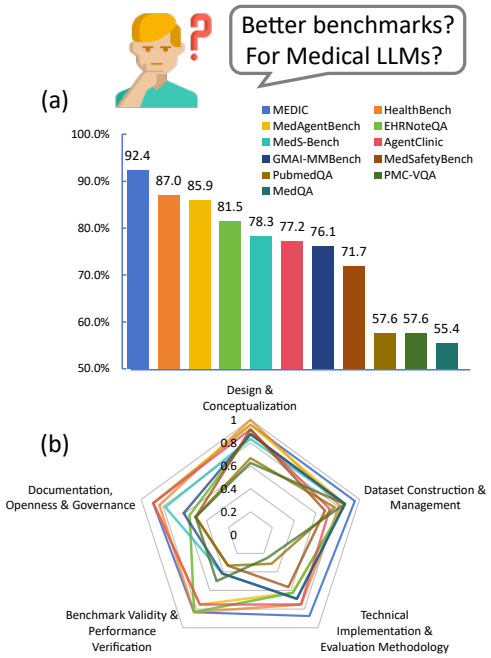

Figure 1: (a) Overall and (b) phase-by-phase performance of medical LLM benchmarks. The significant variance, identified through our MedCheck evaluation, highlights the core motivation for this work: the development of a principled framework for systematic benchmark evaluation.

1 Introduction

Large language models (LLMs) are demonstrating significant potential in healthcare, leading to a proliferation of benchmarks designed to evaluate their capabilities (Jin et al., 2021; Pal et al., 2022). These evaluation tools have evolved from early exam-style question answering (QA) to encompass more complex clinical tasks like report summarization and diagnosis (Wu et al., 2025b; Liu et al., 2024a; Wu et al., 2025a).

However, despite their widespread adoption, the reliability of many newer benchmarks is a growing concern. Echoing critiques from within the clinical informatics community (Wornow et al., 2023), many evaluations rely heavily on closed-form, multiple-choice questions (MCQA) that, while testing factual knowledge, fail to assess open-ended clinical reasoning or account for real-world variability (Wu et al., 2025a; Zhang et al., 2025). Furthermore, a significant number are constructed from academic materials rather than authentic clinical data, which raises critical concerns about their clinical fidelity and the risk of data contamination—where models are evaluated on data seen during training (Ouyang et al., 2024; Wu et al., 2025b). These systemic flaws, particularly prevalent in rapidly developed benchmarks designed for

† Zizhan Ma is the corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 43078–43123

general-purpose LLMs, can create an illusion of progress where models are optimized for tasks that lack genuine clinical utility, a problem long recognized by researchers working with real-world electronic health record (EHR) data (Wornow et al., 2023; Blagec et al., 2023; Alaa et al., 2025).

Recent efforts have proposed general frameworks to improve evaluation quality, such as BetterBench for general-purpose AI (Reuel et al., 2024) and How2Bench for code (Cao et al., 2025). However, while valuable, these frameworks are not tailored for the medical domain, which demands a more rigorous approach accounting for specialized terminology, patient data ethics, and the paramount importance of safety.

To develop trustworthy clinical AI, the field must transition from an ad-hoc, publication-driven approach to a disciplined, engineering-oriented paradigm for evaluation (Laskar et al., 2024; Yan et al., 2024). Adopting a lifecycle-aware perspective, similar to principles in mature engineering fields, is a prerequisite for ensuring safety and efficacy in the medical domain (Park et al., 2020).

We introduce ***MedCheck***, the first comprehensive, lifecycle-oriented assessment framework for medical LLM benchmarks. *MedCheck* deconstructs benchmark development into five continuous stages, from design to governance, and provides a checklist of 46 medically-tailored criteria. We demonstrate *MedCheck*'s utility by applying it in an in-depth evaluation of 56 medical benchmarks. Our analysis reveals widespread, systemic issues: a profound disconnect from clinical practice, a crisis of data integrity from unmitigated contamination risks, and a systematic neglect of safety-critical dimensions like model robustness and uncertainty awareness. Our contributions are threefold:

- We introduce *MedCheck*, the first comprehensive, lifecycle-oriented evaluation framework with 46 criteria specifically designed for medical benchmarks.
- We conduct an in-depth evaluation of 56 medical benchmarks, revealing widespread, systemic weaknesses across the current evaluation landscape.
- We offer *MedCheck* as a practical checklist to guide the development of more reliable, transparent, and clinically relevant benchmarks for AI in healthcare.

2 Related works

2.1 The Evolution of Medical LLM Benchmarks

Evaluation of medical LLMs has been dominated by benchmarks derived from medical qualification examinations (Jin et al., 2021; Pal et al., 2022; Liu et al., 2023) and scholarly literature (Jin et al., 2019; He et al., 2020). While useful for assessing foundational medical knowledge, their format and content often lack direct clinical applicability.

To bridge this gap, recent efforts include data-centric frameworks like BigBIO (Fries et al., 2022) and public clinical datasets like MIMIC-IV (Johnson et al., 2023). There is also a growing emphasis on creating benchmarks that better mirror real-world clinical tasks, such as comprehensive benchmarks that cover report summarization, diagnosis, and treatment planning (Wu et al., 2025b; Liu et al., 2024a; Wu et al., 2025a; Bedi et al., 2025), as well as agentic benchmarks that simulate sequential clinical decision-making (Jiang et al., 2025; Schmidgall et al., 2024). Although these efforts represent a significant step forward, their development has often outpaced the creation of rigorous standards for their own evaluation, a gap our work aims to fill.

2.2 Towards a Science of Benchmark Evaluation

The proliferation of benchmarks has ignited discussion about their reliability and validity. Alaa et al. (2025) empirically demonstrated a poor correlation between LLM performance on existing benchmarks and in real-world clinical scenarios, highlighting a crisis in the construct validity of benchmarks—the degree to which a test measures what it claims to measure (Cronbach and Meehl, 1955). Wornow et al. (2023) also finds that foundation modals are often evaluated on tasks that do not provide meaningful insights on their usefulness to health systems. This has catalyzed a new research area focused on the science of benchmarking itself.

Pioneering works have proposed general frameworks to standardize the evaluation of AI benchmarks. For instance, Bhardwaj et al. (2024) developed a checklist for data curation best practices. More comprehensively, BetterBench (Reuel et al., 2024) provides a 46-criteria framework for general-purpose AI benchmarks, and How2Bench (Cao et al., 2025) offers a 55-item checklist for code-related LLM benchmarks. These frameworks

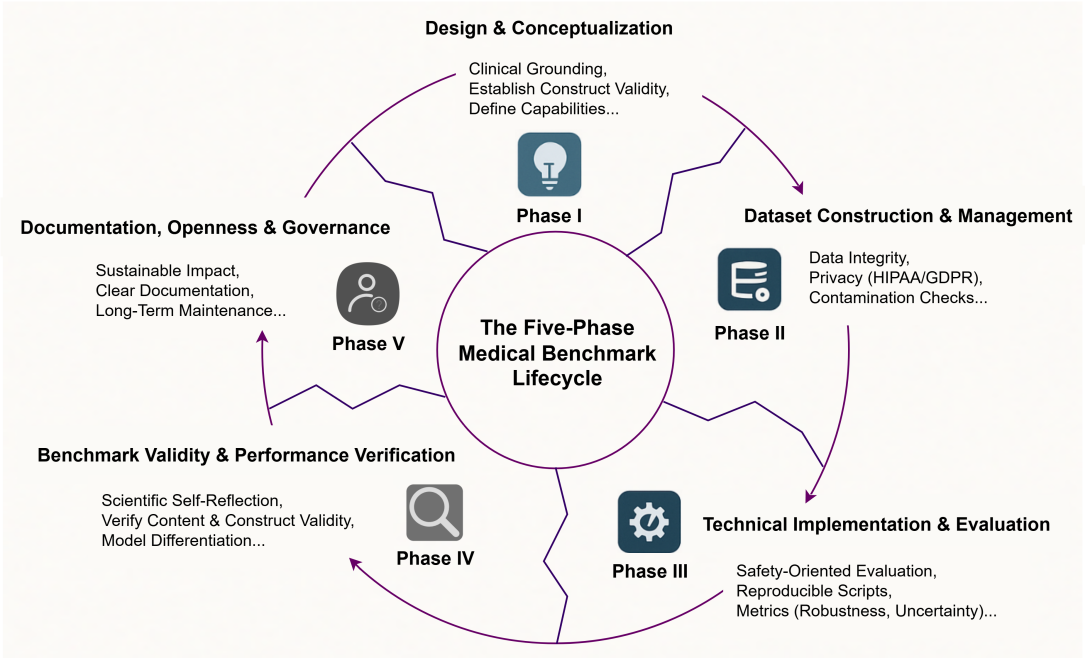

Figure 2: The proposed five-phase lifecycle for medical benchmark engineering. This model illustrates the interconnected and continuous stages of development, from initial design and conceptualization to long-term documentation and governance.

promote a lifecycle-aware perspective, revealing common issues in data quality, reproducibility, and transparency. There are also initiatives like TRIPOD-LLM (Gallifant et al., 2025) which propose standardized reporting guidelines to enhance transparency and reproducibility.

However, these valuable solutions are either general-purpose or focused on reporting. High-stakes clinical applications demand a more rigorous, context-aware approach that accounts for specialized terminology, patient data ethics, and the paramount importance of safety and reliability. While existing surveys have profiled the medical LLM landscape (Liu et al., 2024c; Yan et al., 2024), they stop short of providing a structured framework for assessing benchmark quality. Our work bridges this critical gap by introducing *MedCheck*, the first assessment framework specifically engineered for the unique complexities of the medical domain.

For a detailed discussion of the foundational evaluation and reporting frameworks of the clinical informatics community, refer to the Appendix A.

3 Design

To systematically deconstruct the complexities and shortcomings of existing medical LLM benchmarks, we establish a conceptual framework for their development. This section introduces a novel five-phase lifecycle model for benchmark engineering and details the rigorous methodology employed in our comprehensive analysis.

3.1 The Five-Phase Medical Benchmark Lifecycle

Our proposed lifecycle, as depicted in Figure 2, provides a structured paradigm for the engineering of high-quality medical benchmarks. Each phase addresses a core set of objectives and potential pitfalls.

Phase I: Design and Conceptualization.

This foundational phase moves beyond mere task definition to establish the benchmark's theoretical underpinnings, focusing on its construct validity—the degree to which it accurately measures its intended theoretical construct, such as "clinical reasoning" (Cronbach and Meehl, 1955; Mehandru et al., 2025; Schmidgall et al., 2024). This phase defines the benchmark's specific, medically relevant purpose and intended contribution, as exam performance may not correlate with real-world utility (Alaa et al., 2025).

Phase II: Dataset Construction and Management.

This phase constitutes the benchmark's empirical core, focusing on the curation of authentic, diverse, representative, and ethically sourced data.

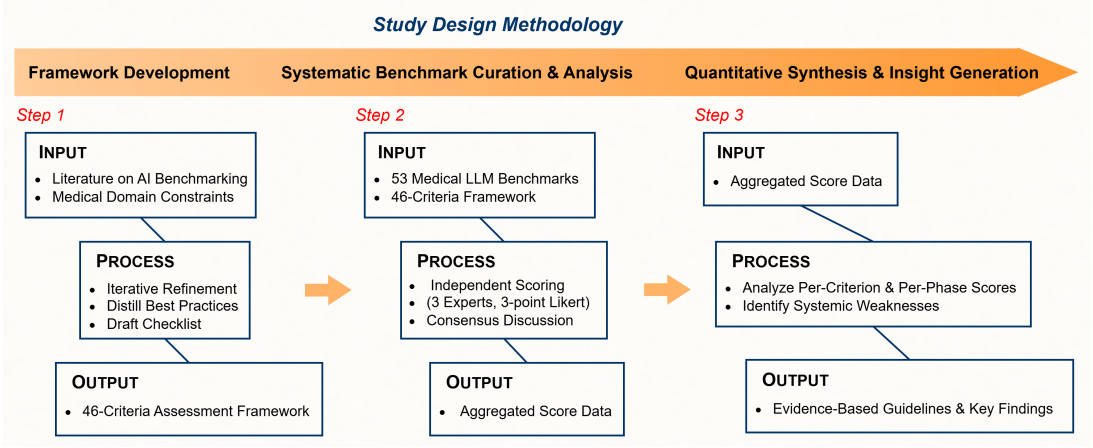

Figure 3: The three-step methodology employed in this study. Our approach involved (1) developing the 46-criteria assessment framework, (2) systematically curating and scoring 56 benchmarks against it, and (3) performing a quantitative synthesis to generate systemic insights.

This demands strict adherence to medical privacy regulations like HIPAA and GDPR (Bhardwaj et al., 2024; Eke and Shuib, 2025). A primary challenge is proactively addressing data contamination, where evaluation data has been seen by an LLM during training (Ouyang et al., 2024; Wu et al., 2025b). Such contamination leads to inflated performance scores and misleading leaderboards (Dong et al., 2024; Cheng et al., 2025), making rigorous processes for data quality assurance, de-identification, and contamination detection paramount.

Phase III: Technical Implementation and Evaluation Methodology. This operational phase transforms a dataset into an evaluation toolkit with reproducible scripts and, crucially, metrics beyond simple accuracy (Chang et al., 2023). As closed-form MCQA may not assess deep reasoning (Wu et al., 2025a; Molfese et al., 2025), it is vital to evaluate the logical coherence of a model's reasoning (Dai et al., 2025), its robustness to noisy inputs (Han et al., 2024), and its capacity to articulate uncertainty—a cornerstone of safe clinical practice (Shelmanov et al., 2025).

Phase IV: Benchmark Validity and Performance Verification. This phase provides the empirical validation for the benchmark as a measurement instrument. It involves presenting evidence for content validity (the content is representative of the clinical domain) and, critically, substantiating the construct validity claims established in Phase I (Alaa et al., 2025). A validated benchmark must also demonstrate its ability to reliably differentiate between models of varying capabilities, thus providing a clear signal of progress in the field (Reuel et al., 2024).

Phase V: Documentation, Openness, and Governance. This community-facing phase ensures the benchmark's long-term value and impact. It involves clear documentation and open-source principles for transparency and reproducibility (Arvan et al., 2022; Cohen et al., 2018). Critically, it demands a robust governance model for long-term maintenance, versioning, and community feedback.

MedCheck evaluates benchmarks as complete engineered systems, not just data collections. It recognizes that even when datasets are reused, significant value can be added through innovative task design, rigorous validation, and sustainable governance.

3.2 Study Design

Our three-step methodology (Figure 3) was designed to ensure objectivity, reproducibility, and depth.

Step 1: Framework Development. We developed the 46-criteria *MedCheck* framework, which serves as the analytical lens for this study. We developed the framework by first conducting an extensive literature review of existing best practices in general AI benchmarking (Reuel et al., 2024; Cao et al., 2025) and machine learning data curation (Bhardwaj et al., 2024), which we then distilled and iteratively refined into the final 46-criteria framework, grounded in the medical domain's unique ethical and practical constraints, such as patient privacy (e.g., HIPAA compliance), the need for

evidence-based standards, and the high stakes associated with potential patient harm (Farhud and Zokaei, 2021; Eke and Shuib, 2025). The complete 46-criteria checklist is provided in Appendix H.

Step 2: Systematic Benchmark Curation and Analysis. We curated and analyzed a corpus of 56 medical LLM benchmarks, listed in Appendix E. The selection process is described in Appendix B To ensure objectivity, a rigorous scoring protocol was implemented. Our evaluation process incorporates both LLM-as-judge and domain expert review to balance scalability with accuracy. Specifically, we first provided the paper text and documentation of each benchmark to a large language model and requested an initial scoring based on our 46 criteria. Then, a panel of three NLP researchers with clinical informatics experience reviewed and adjusted these scores on a 3-point Likert scale (0: not met; 1: partially met; 2: fully met). All assessments were based exclusively on publicly available artifacts, including published papers, code repositories, and official websites. Any scoring discrepancies were resolved through a consensus discussion to arrive at a final score for each criterion. We provide the detailed scoring guidelines for the three NLP researcher annotators, along with the full human-in-the-loop annotation protocol, in Appendix C.

Step 3: Quantitative Synthesis and Insight Generation. The individual scores were aggregated to enable a multi-level quantitative analysis. We calculated per-criterion average scores to identify specific, widespread weaknesses across the field. These were then rolled up into per-phase scores to assess the maturity of each stage in the development lifecycle. Finally, an overall score was computed for each benchmark to gauge its overall quality. This quantitative synthesis allowed us to move beyond anecdotal critiques and identify widespread, descriptive patterns of deficiencies across the entire landscape of medical LLM evaluation, forming the evidence base for the guidelines presented in the following section. A detailed breakdown of these quantitative results can be found in Appendix F.

4 MedCheck: A Guideline for Medical Benchmarks

Based on our five-phase lifecycle and extensive analysis, we present the *MedCheck* framework as an actionable guideline. For each phase, we explain its core principle and summarize our key findings, further detailed in Appendix F. Our analysis also differentiates between benchmarks designed for foundational medical knowledge and those for clinical practice. A detailed domain-specific analysis can be found in Appendix D, which shows that our framework appropriately evaluates benchmarks according to their intended purpose. In the tables that follow, each criterion number is hyperlinked to its full definition and scoring rubric in Appendix H.

Table 1: MedCheck Criteria for Phase I: Design and Conceptualization

No.	Description
1	Does the benchmark define the targeted LLM capabilities in medicine for evaluation (e.g., QA, diagnostic reasoning)?
2	Does it describe specific clinical or research applications and their potential value?
3	Does it highlight its unique contribution or innovation compared to existing benchmarks?
4	Does it define the specific LLM functions for evaluation?
5	Does it define the scope of the medical specialties of the benchmark?
6	Is it designed to meet the needs of LLM researchers and medical challenges?
7	Have qualified medical experts been involved in benchmark development?
8	Does it rely on recognized medical sources (e.g., clinical guidelines, databases)?
9	Does it align with international medical standards (e.g., ICD, SNOMED CT, LOINC)?
10	Are the metrics clear and closely related to clinical tasks?
11	Does it assess multiple dimensions (e.g., safety, completeness, interpretability)?
12	Does it consider potential risks and biases in model outputs?

4.1 Guideline for Design & Conceptualization

Explanation. This foundational phase anchors a benchmark in clinical relevance and scientific novelty. It necessitates the precise definition of the targeted LLM capabilities, medical scope, and application value, all of which must be grounded in documented domain expertise and authoritative sources. A robust design benefits from considering or aligning with established medical standards or terminologies (e.g., ICD, SNOMED CT) where appropriate to reflect real-world practice and facilitate downstream integration with clinical systems. It should also incorporate the use of clearly defined, multi-dimensional metrics that extend beyond accuracy, and the proactive consideration of safety and bias from inception (Arora et al., 2025).

Findings: The Clinical Disconnect. While nearly all benchmarks (98%) define high-level ob-

jectives, our analysis reveals a systemic issue we term the *Clinical Disconnect*. A total of 50% (28 of 56) fail to align with any formal medical standards (e.g., ICD, SNOMED CT). Furthermore, 45% (25 of 56) do not incorporate safety and fairness into their design, and 34% (19 of 56) evaluate only a single dimension like accuracy, neglecting critical aspects such as completeness. This disconnect stems from an "academic-first, clinical-second" mindset, where developers favor convenient data sources like exam questions from MedQA (Jin et al., 2021) and MedMCQA (Pal et al., 2022) over data reflecting complex clinical workflows. This convenience-driven design compromises clinical fidelity and construct validity, leading to models optimized for irrelevant and potentially unsafe tasks (Alaa et al., 2025).

Table 2: MedCheck Criteria for Phase II: Dataset Construction & Management

No.	Description
13	Are the original sources clearly stated and traceable?
14	Are the sources authoritative and well-justified?
15	Is the data origin clear, and is synthetic data validated?
16	Is the dataset representative of the target population?
17	Are diversity goals defined with supporting quantitative analysis?
18	Is the data properly cleaned and standardized?
19	Are privacy measures described and regulation-compliant?
20	Is the data format consistent and unambiguous?
21	Is expert-involved data review in place?
22	Are reference answers accurate and validated?
23	Are contamination risks detected and handled?

4.2 Guideline for Dataset Construction & Managemen

Explanation. This phase addresses the integrity of the benchmark's core asset: its data. This principle mandates the use of traceable and authoritative data sources, with validated authenticity for any synthetic data. The dataset must be demonstrably representative and diverse, supported by quantitative analysis. Rigorous quality control—encompassing cleaning, standardization, expert review, and validated reference answers—is imperative, as are regulation-compliant privacy measures and the proactive mitigation of data contamination risks (Ouyang et al., 2024; Wu et al., 2025b).

Findings: A Crisis of Foundational Validity. Our analysis reveals critical data management weaknesses that undermine the field's empirical foundation. While most benchmarks are reasonably transparent about their primary data sources, subsequent quality control is severely lacking. A staggering 88% (49 of 56) fail to address data contamination. While post-hoc detection is challenging for closed-source models, the field lacks proactive mitigation strategies within developer control, such as the implementation of canary strings or temporal data cutoffs to explicitly signal exclusion from future pre-training crawls. Furthermore, 66% (37 of 56) are insufficiently diverse or representative, and 55% (31 of 56) lack any clear data audit or expert review mechanism. This failure to ensure data integrity threatens the field's validity. Unvetted data may contain factual errors, while contamination leads to artificially inflated scores, creating a false perception of model capability (Magar and Schwartz, 2022; Deng et al., 2024; Bender et al., 2021). This renders leaderboards misleading and undermines scientific credibility by building on unreliable evidence (Holistic AI, 2024).

Table 3: MedCheck Criteria for Phase III: Technical Implementation & Evaluation Methodology

No.	Description
24	Is the evaluation tool easy to install and use?
25	Are detailed documentation and environment settings provided to support reproducibility?
26	Are baseline models or human performance results provided for comparison?
27	Are there evaluations for the model's reasoning process?
28	Are there evaluations testing the model's robustness (e.g., input perturbations)?
29	Does the benchmark design help evaluate the generalization ability of models to unseen data?
30	Are there evaluations testing the model's ability to express uncertainty?
31	Does the benchmark support both closed-source APIs and open-source models?

4.3 Guideline for Technical Implementation & Evaluation Methodology

Explanation. This phase focuses on transforming a dataset into a complete evaluation toolkit. It requires the provision of accessible, reproducible, and well-documented code that supports diverse model types and offers performance baselines for context. Fundamentally, the methodology must transcend accuracy to assess safety-critical capabilities: the model's reasoning process, its robustness to input variations, its generalization to unseen data, and its capacity to articulate uncertainty—all

of which are vital for trustworthy medical AI (Dai et al., 2025; Qiu et al., 2025; Han et al., 2024; Shelmanov et al., 2025).

Findings: Systematic Neglect of Safety-Critical Capabilities. This is the most underdeveloped phase in our analysis (average score: 52.4%), revealing a profound gap between current practice and the needs of reliable medical AI. An alarming 89% (50 of 56) of benchmarks have no mechanism to test for model robustness, and 91% (51 of 56) fail to evaluate a model's ability to handle uncertainty. Furthermore, 48% (27 of 56) neglect the model's reasoning process, focusing only on the final answer. These omissions constitute a systematic neglect of safety. A model's reasoning, robustness, and uncertainty awareness are cornerstones of clinical trustworthiness (Farhud and Zokaei, 2021). A brittle model that fails with slight data variations is dangerous (The BMJ, 2024), and an overconfident model that gives incorrect advice without expressing uncertainty is a direct threat to patient safety. By not measuring these capabilities, the community implicitly deems them unimportant, increasing the risk of deploying brittle, opaque, and unsafe systems.

Table 4: MedCheck Criteria for Phase IV: Benchmark Validity & Performance Verification

No.	Description
32	Does the benchmark cover the claimed medical knowledge and skills?
33	Do the tasks realistically simulate clinical settings?
34	Can it distinguish between models of different levels?
35	Are benchmark scores correlated with real-world clinical performance?
36	Does the benchmark demonstrate internal consistency to ensure that different components reliably assess the same capability?
37	Are statistical tests used to verify results?

4.4 Guideline for Benchmark Validity & Performance Verification

Explanation. This phase concerns the scientific validation of the benchmark as a measurement instrument. It requires empirical evidence supporting both content validity (comprehensive coverage of the claimed domain) and construct validity (realistic task simulation and accurate measurement of the intended capability (Cronbach and Meehl, 1955)). A validated benchmark must also demonstrate its utility through proven discriminative power, correlation with real-world clinical performance, high internal consistency, and the application of statistical tests to verify results (Reuel et al., 2024).

Findings: The Risk of Misdirected Progress. Formal scientific validation of benchmarks themselves is rare. While most benchmarks can differentiate models, only 54% (30 of 56) provide a compelling analysis of their content validity, and just 38% (21 of 56) are grounded in scenarios with high real-world authenticity. Without proper validation, a benchmark's results are difficult to interpret. Poor content validity creates blind spots, while poor construct validity means the benchmark may not measure the claimed capability at all, leading to misleading conclusions and misdirected research efforts (Alaa et al., 2025). This lack of self-reflection risks optimizing for metrics that lack genuine clinical utility. To bridge this gap, developers must integrate clinician-in-the-loop validation, ensuring that automated evaluation metrics strictly align with physician preferences and patient safety outcomes.

Table 5: MedCheck Criteria for Phase V: Documentation, Openness, and Governance

No.	Description
38	Is there clear, comprehensive benchmark documentation?
39	Are the evaluation criteria and instructions clear and easy to follow?
40	Are limitations and potential risks openly discussed?
41	Has the benchmark undergone formal academic peer review?
42	Are the code and data publicly available with proper licensing?
43	Is a clear usage and citation guideline provided?
44	Is there a clear plan for updates and version control?
45	Is there an public channel for user feedback?
46	Is the long-term maintenance responsibility clearly stated?

4.5 Guideline for Documentation, Openness, & Governance

Explanation. This final phase ensures a benchmark's long-term utility and trustworthiness. It requires comprehensive and transparent documentation, including clear instructions and a candid discussion of limitations. Adherence to open-source principles—publicly accessible code and data with appropriate licensing, clear citation guidelines, and academic peer review—is crucial. A robust governance model, which specifies long-term main-

tenance responsibilities, versioning plans, and a public feedback channel, is essential for sustained relevance and impact.

Findings: A Fragmented and Unsustainable Ecosystem. While most benchmarks provide public access to their assets (55 of 56), the governance required for sustainable impact is critically lacking. A significant 39% (22 of 56) do not specify a usage license, creating barriers to adoption. Most critically, 80% (45 of 56) have no clear maintenance plan, and 63% (35 of 56) lack a public feedback channel. This renders them "fire-and-forget" artifacts destined for obsolescence. This practice fosters a fragmented and unsustainable ecosystem, where effort is wasted on disposable artifacts rather than on building a lasting, reliable infrastructure for scientific progress (Arvan et al., 2022). To ensure longevity, the community must move toward institutional stewardship models, where host organizations provide explicit, long-term funding and technical commitments to prevent valuable benchmarks from becoming static, unmaintained relics.

5 Discussion

Our analysis of 56 medical benchmarks reveals a field at a critical crossroads. While research is active, systemic deficiencies threaten the validity and utility of this work.

5.1 Implications: The Urgent Need for a Paradigm Shift

The status quo is scientifically unsound and clinically irresponsible. The current trajectory fosters an "illusion of progress," where scores on clinically irrelevant or contaminated tasks mask a lack of genuine advancement. By failing to account for evaluation noise, the current paradigm risks optimizing models for pattern memorization rather than the deep, non-linear reasoning required for complex patient care (Laskar et al., 2024; Chang et al., 2023). This misdirects research efforts, misinforms stakeholders, and delays the responsible integration of AI into healthcare by creating brittle, biased systems that could harm patients (Obermeyer et al., 2019).

A paradigm shift is urgently needed—a move from ad-hoc dataset creation toward a disciplined, lifecycle-aware practice of benchmark engineering. This requires treating benchmarks not as disposable artifacts for a single paper, but as scientific instruments demanding rigorous design, validation, and maintenance (Reuel et al., 2024).

5.2 Our Guideline as a Catalyst for Change

Our five-phase lifecycle and 46-criteria checklist offer a practical toolkit and actionable roadmap for this shift, helping developers create higher-quality benchmarks and users assess existing ones. Adopting this framework can directly mitigate the critical issues identified in our study by promoting: *Medical Grounding* (mandating expert involvement and alignment with medical standards), *Data Integrity* (enforcing transparent sourcing, quality control, and contamination checks), *Safety-Oriented Evaluation* (requiring assessment of reasoning, robustness, and uncertainty), *Scientific Validity* (emphasizing content and construct validation), and *Sustainable Impact* (promoting open practices and long-term governance). Furthermore, it is essential to recognize that the multi-dimensional objectives within the *MedCheck* lifecycle are mutually reinforcing rather than fundamentally incompatible. For instance, stringent data privacy (Phase II) and open-source transparency (Phase V) act as coexisting requirements: rigorous de-identification is a strict prerequisite before a clinical dataset can be ethically open-sourced. High standards in one dimension fundamentally enable and secure practices in the others, creating a cohesive ecosystem for trustworthy AI evaluation.

5.3 Future Directions: The Next Frontier of Medical Benchmark Research

Future efforts should move beyond the static, multiple-choice paradigm and pursue three critical directions:

Embracing Dynamic and Interactive Benchmarks. Future benchmarks must reflect the dynamic nature of clinical encounters by assessing sequential decision-making and information gathering. Pioneering works like MediQ (Li et al., 2024) and AgentClinic (Schmidgall et al., 2024) exemplify this necessary evolution toward evaluating core clinical reasoning. Such interactive setups force models to actively elicit missing patient information, better mirroring the iterative and often ambiguous nature of real-world differential diagnosis (Trimble and Hamilton, 2016).

Prioritizing Empirical Construct Validity. The field must empirically validate that benchmarks truly measure the clinical constructs they claim to (Alaa et al., 2025). This requires innovative methodologies that correlate benchmark scores

with performance on real-world clinical data, such as predicting patient outcomes from EHRs. Without this validation, we risk optimizing for metrics that lack clinical significance.

Building a Collaborative Evaluation Ecosystem. To foster quality and transparency, the community should develop a living repository, akin to platforms like betterbench.stanford.edu (Reuel et al., 2024). Such a platform would enable continuous evaluation of benchmarks against a standardized framework like *MedCheck*. This would create a feedback loop incentivizing higher-quality benchmarks and more informed decisions.

6 Conclusion

To address the persistent reliability concerns in medical AI evaluation, we introduced *MedCheck*, a comprehensive, lifecycle-oriented assessment framework comprising 46 criteria. Our empirical evaluation of 56 benchmarks exposed systemic deficiencies across the field, namely a profound disconnect from real-world clinical practice, severe data integrity vulnerabilities stemming from unmitigated contamination risks, and a widespread neglect of safety-critical capabilities. These findings underscore the urgent need to transition from ad-hoc dataset curation toward a disciplined, engineering-focused approach. *MedCheck* serves as the foundational toolkit for this paradigm shift, guiding the community to move beyond superficial leaderboard rankings and rethink medical LLM benchmarks entirely, ultimately fostering the development of genuinely safe, trustworthy, and clinically effective artificial intelligence.

Limitations

We acknowledge this study's limitations. First, our analysis of 56 benchmarks, while extensive, is not exhaustive given the field's rapid growth. Second, the scoring process, despite a rigorous protocol, retains a degree of subjectivity inherent in qualitative assessment. Third, our findings are based exclusively on public artifacts, which may not capture all unpublished development practices. Finally, the *MedCheck* framework itself is a snapshot of current best practices and will require future revisions to keep pace with evolving AI capabilities and ethical standards. Specifically, as multimodal and autonomous medical agents emerge, future iterations must expand to rigorously audit multimodal clinical grounding and agentic safety.

References

Asma Ben Abacha, Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Meliha Yetisgen, Fei Xia, and Thomas Lin. 2025. Medec: A benchmark for medical error detection and correction in clinical notes. *Preprint*, arXiv:2501.03465.

Ahmed Alaa, Thomas Hartvigsen, Niloufar Golchini, Shiladitya Dutta, Frances Dean, Inioluwa Deborah Raji, and Travis Zack. 2025. Medical large language model benchmarks should prioritize construct validity. *Preprint*, arXiv:2503.10694.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. Healthbench: Evaluating large language models towards improved human health. *Preprint*, arXiv:2505.08775.

Mohammad Arvan, Luís Pina, and Natalie Parde. 2022. Reproducibility in computational linguistics: Is source code enough? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10423–10432, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M. Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, Hao Qiu, Shrey Jain, Leonardo Schettini, Mehr Kashyap, Jason Alan Fries, Akshay Swaminathan, Philip Chung, Fateme Nateghi, Asad Aali, and 62 others. 2025. Medhelm: Holistic evaluation of large language models for medical tasks. *Preprint*, arXiv:2505.23802.

Asma Ben Abacha, Mourad Sarrouti, Dina Demner-Fushman, Sadid A Hasan, and Henning Müller. 2021. Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In *Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum-working notes*. 21-24 September 2021.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Eshta Bhardwaj, Harshit Gujral, Siyi Wu, Ciara Zogheib, Tegan Maharaj, and Christoph Becker. 2024. Machine learning data practices through a data curation lens: An evaluation framework. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1055–1067.

Kathrin Blagec, Jakob Kraiger, Wolfgang Frühwirt, and Matthias Samwald. 2023. Benchmark datasets driving artificial intelligence development fail to capture the needs of medical professionals. *Journal of Biomedical Informatics*, 137:104274.

Jialun Cao, Yuk-Kit Chan, Zixuan Ling, Wenxuan Wang, Shuqing Li, Mingwei Liu, Ruixi Qiao, Yuting Han, Chaozheng Wang, Boxi Yu, Pinjia He, Shuai Wang, Zibin Zheng, Michael R. Lyu, and Shing-Chi Cheung. 2025. How Should We Build A Benchmark? Revisiting 274 Code-Related Benchmarks For LLMs. *Preprint*, arXiv:2501.10711.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 2: Short Papers)*, pages 88–109.

Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, Shaoting Zhang, Bin Fu, Jianfei Cai, Bohan Zhuang, Eric J Seibel, Yu Qiao, and Junjun He. 2024. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. In *Advances in Neural Information Processing Systems*, volume 37, pages 94327–94427. Curran Associates, Inc.

Yuxing Cheng, Yi Chang, and Yuan Wu. 2025. A survey on data contamination for large language models. *Preprint*, arXiv:2502.14425.

K. Bretonnel Cohen, Dina Demner-Fushman, rightfully rightfully, and Asma Ben Abacha. 2018. Three dimensions of reproducibility in natural language processing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Lee J Cronbach and Paul E Meehl. 1955. Construct validity in psychological tests. *Psychological bulletin*, 52(4):281.

Wei Dai, Peilin Chen, Malinda Lu, Daniel A Li, Haowen Wei, Hejie Cui, and Paul Pu Liang. 2025. CLIMB: Data foundations for large scale multimodal clinical foundation models. In *Forty-second International Conference on Machine Learning*.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024. Investigating data contamination in modern benchmarks for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8706–8719, Mexico City, Mexico. Association for Computational Linguistics.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12039–12050, Bangkok, Thailand. Association for Computational Linguistics.

Christopher Ifeanyi Eke and Liyana Shuib. 2025. The role of explainability and transparency in fostering trust in ai healthcare systems: a systematic literature review, open issues and potential solutions. *Neural Computing and Applications*, 37(4):1999–2034.

Yongqi Fan, Hongli Sun, Kui Xue, Xiaofan Zhang, Shaoting Zhang, and Tong Ruan. 2025a. MedOdyssey: A medical domain benchmark for long context evaluation up to 200K tokens. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 32–56, Albuquerque, New Mexico. Association for Computational Linguistics.

Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Wang Siyuan, Zhongyu Wei, and Fei Huang. 2025b. AI hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10183–10213, Abu Dhabi, UAE. Association for Computational Linguistics.

Dariush D Farhud and Shaghayegh Zokaei. 2021. Ethical issues of artificial intelligence in medicine and healthcare. *Iranian journal of public health*, 50(11):i.

Jason Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Sunny Kang, Rosaline Su, Wojciech Kusa, Samuel Cahyawijaya, Fabio Barth, Simon Ott, Matthias Samwald, Stephen Bach, Stella Biderman, Mario Sänger, Bo Wang, Alison Callahan, Daniel León Periñán, and 24 others. 2022. Bigbio: A framework for data-centric biomedical natural language processing. In *Advances in Neural Information Processing Systems*, volume 35, pages 25792–25806. Curran Associates, Inc.

Jack Gallifant, Majid Afshar, Saleem Ameen, Yindalon Aphinyanaphongs, Shan Chen, Giovanni Cacciamani, Dina Demner-Fushman, Dmitriy Dligach, Roxana Daneshjou, Chrystinne Fernandes, Lasse Hyldig Hansen, Adam Landman, Lisa Lehmann, Liam G. McCoy, Timothy Miller, Amy Moreno, Nikolaj Munch, David Restrepo, Guergana Savova, and 6 others. 2025. The tripod-llm reporting guideline for studies using large language models. *Nature Medicine*, 31(1):60–69.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, John Caskey, Brihat Sharma, Matthew M Churpek, and Majid Afshar. 2023. Dr. bench: Diagnostic reasoning benchmark for clinical natural language processing. *Journal of biomedical informatics*, 138:104286.

Chenlu Guo, Nuo Xu, Yi Chang, and Yuan Wu. 2024. Chbench: A chinese dataset for evaluating health in large language models. *Preprint*, arXiv:2409.15766.

Tessa Han, Aounon Kumar, Chirag Agarwal, and Himabindu Lakkaraju. 2024. MedSafetyBench: Evaluating and improving the medical safety of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Industry Track*, pages 34–45.

Junqing He, Mingming Fu, and Manshu Tu. 2019. Applying deep matching networks to chinese medical question answering: a study and a dataset. *BMC medical informatics and decision making*, 19(Suppl 2):52.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. PathVQA: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Holistic AI. 2024. An overview of data contamination. https://www.holisticai.com/blog/overview-of-data-contamination.

Yutao Hu, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. 2024. OmniMedVQA: A new large-scale comprehensive evaluation benchmark for medical lvlm. *Preprint*, arXiv:2402.09181.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):590–597.

Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. 2025. MedAgentBench: A realistic virtual ehr environment to benchmark medical llm agents. *Preprint*, arXiv:2501.14654.

Di Jin, Eileen Pan, Nassib Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2567–2577.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1.

Praveen K Kanithi, Clément Christophe, Marco AF Pimentel, Tathagata Raha, Nada Saadi, Hamza Javed, Svetlana Maslenkova, Nasir Hayat, Ronnie Rajan, and Shadab Khan. 2024. Medic: Towards a comprehensive framework for evaluating llms in clinical applications. *arXiv preprint arXiv:2409.07314*.

Nikhil Khandekar, Qiao Jin, Guangzhi Xiong, Soren Dunn, Serina S Applebaum, Zain Anwar, Maame Sarfo-Gyamfi, Conrad W Safranek, Abid A Anwar, Andrew Zhang, Aidan Gilson, Maxwell B Singer, Amisha Dave, Andrew Taylor, Aidong Zhang, Qingyu Chen, and Zhiyong Lu. 2024. Medcalc-bench: Evaluating large language models for medical calculations. *Advances in Neural Information Processing Systems*, 37:84730–84745.

Yunsoo Kim, Jinge Wu, Yusuf Abdulle, and Honghan Wu. 2024. MedExQA: Medical question answering benchmark with multiple explanations. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 167–181, Bangkok, Thailand. Association for Computational Linguistics.

Sunjun Kweon, Jiyoun Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwanghyun Kim, Jeewon Yang, Seunghyun Won, and Edward Choi. 2024. Ehrnoteqa: An llm benchmark for real-world clinical practice using discharge summaries. In *Advances in Neural Information Processing Systems*, volume 37, pages 124575–124611. Curran Associates, Inc.

Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, and Naeemul Khan. 2024. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13785–13816.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Sci. Data*, 5(1):180251.

Jing Li, Shangping Zhong, and Kaizhi Chen. 2021. MLEC-QA: A Chinese Multi-Choice Biomedical Question Answering Dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8862–8874, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. 2024. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pages 28858–28888. Curran Associates, Inc.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654.

Fenglin Liu, Zheng Li, Hongjian Zhou, Qingyu Yin, Jingfeng Yang, Xianfeng Tang, Chen Luo, Ming Zeng, Haoming Jiang, Yifan Gao, Priyanka Nigam,

Sreyashi Nag, Bing Yin, Yining Hua, Xuan Zhou, Omid Rohanian, Anshul Thakur, Lei Clifton, and David A. Clifton. 2024a. Large language models in the clinic: A comprehensive benchmark. *Preprint*, arXiv:2405.00716.

Fenglin Liu, Jinge Wu, Hongjian Zhou, Xiao Gu, Soheila Molaei, Anshul Thakur, Lei Clifton, Honghan Wu, and David A Clifton. 2025a. Riskagent: Autonomous medical ai copilot for generalist risk prediction. *arXiv preprint arXiv:2503.03802*.

Jie Liu, Wenxuan Wang, Zizhan Ma, Guolin Huang, Yihang SU, Kao-Jung Chang, Wenting Chen, Haoliang Li, Linlin Shen, and Michael Lyu. 2024b. Medchain: Bridging the gap between llm agents and clinical practice through interactive sequential benchmarking. *Preprint*, arXiv:2412.01605.

Jie Liu, Wenxuan Wang, Su Yihang, Jingyuan Huang, Yudi Zhang, Cheng-Yi Li, Wenting Chen, Xiaohan Xing, Kao-Jung Chang, Linlin Shen, and Michael R. Lyu. 2025b. Asclepius: A spectrum evaluation benchmark for medical multi-modal large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24181–24201, Vienna, Austria. Association for Computational Linguistics.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, LEI ZHU, and Michael Lingzhi Li. 2023. Benchmarking large language models on cmexam - a comprehensive chinese medical exam dataset. In *Advances in Neural Information Processing Systems*, volume 36, pages 52430–52452. Curran Associates, Inc.

Lei Liu, Xiaoyan Yang, Junchi Lei, Yue Shen, Jian Wang, Peng Wei, Zhixuan Chu, Zhan Qin, and Kui Ren. 2024c. A survey on medical large language models: Technology, application, trustworthiness, and future directions. *arXiv preprint arXiv:2406.03712*.

Shengyuan Liu, Boyun Zheng, Wenting Chen, Zhihao Peng, Zhenfei Yin, Jing Shao, Jiancong Hu, and Yixuan Yuan. 2025c. A comprehensive evaluation of multi-modal large language models for endoscopy analysis. *Preprint*, arXiv:2505.23601.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to extrapolation. *arXiv preprint arXiv:2205.02340*.

Nikita Mehandru, Niloufar Golchini, David Bamman, Travis Zack, Melanie F Molina, and Ahmed Alaa. 2025. Er-reason: A benchmark dataset for llm-based clinical reasoning in the emergency room. *Preprint*, arXiv:2505.22919.

Francesco Molfese, Simone Balloccu, Gianni Fenu, and Ludovico Marras. 2025. Right answer, wrong score: Uncovering the inconsistencies of llm evaluation in multiple-choice question answering. *arXiv preprint arXiv:2503.05113*.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

Tobi Olatunji, Abraham Toluwase Owodunni, Charles Nimo, Jennifer Orisakwe, Henok Biadglign Ademtew, Chris Fourie, Foutse Yuehgoh, Jonas Kemp, Stephen Moore, Mardhiyah Sanni, Emmanuel Ayodele, Irfan Essa, Timothy Faniran, Bonaventure F. P. Dossou, Fola Omofoye, Wendy Kinara, Tassallah Abdullahi, Michael Best, Katherine Heller, and Mercy Asiedu. 2025. Afrimed-qa: A pan-african multi-specialty medical question-answering benchmark dataset.

Zetian Ouyang, Yishuai Qiu, Linlin Wang, Gerard De Melo, Ya Zhang, Yanfeng Wang, and Liang He. 2024. CliMedBench: A large-scale Chinese benchmark for evaluating medical large language models in clinical scenarios. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8428–8438, Miami, Florida, USA. Association for Computational Linguistics.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Dimitrios P. Panagoulias, Persephone Papatheodosiou, Anastasios P. Palamidas, Mattheos Sanoudos, Evridiki Tsoureli-Nikita, Maria Virvou, and George A. Tsihrintzis. 2024. Cognet-md, an evaluation framework and dataset for large language model benchmarks in the medical domain. *Preprint*, arXiv:2405.10893.

Yoonyoung Park, Gretchen Purcell Jackson, Morgan A Foreman, Daniel Gruen, Jianying Hu, and Amar K Das. 2020. Evaluating artificial intelligence in medicine: phases of clinical research. *JAMIA open*, 3(3):326–331.

Pengcheng Qiu, Chaoyi Wu, Shuyu Liu, Weike Zhao, Zhuoxia Chen, Hongfei Gu, Chuanjin Peng, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. Quantifying the reasoning abilities of llms on real-world clinical cases. *Preprint*, arXiv:2503.04691.

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer. 2024. Betterbench: Assessing ai benchmarks, uncovering issues, and establishing best practices. In *Advances in Neural Information Processing Systems*, volume 37, pages 21763–21813. Curran Associates, Inc.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. AgentClinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *Preprint*, arXiv:2405.07960.

Artem Shelmanov, Maxim Panov, Ekaterina Sergeevna Fadeeva, Artem Vazhentsev, Roman Konstantinovich Vashurin, and Timothy Baldwin. 2025. Uncertainty quantification for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts)*, pages 3–4, Vienna, Austria. Association for Computational Linguistics.

Ofir Ben Shoham and Nadav Rappoport. 2024. Medconceptsqa: Open source medical concepts qa benchmark. *Computers in Biology and Medicine*, 182:109089.

Yu Sun, Xingyu Qian, Weiwen Xu, Hao Zhang, Chenghao Xiao, Long Li, Yu Rong, Wenbing Huang, Qifeng Bai, and Tingyang Xu. 2025a. Reasonmed: A 370k multi-agent generated dataset for advancing medical reasoning. *Preprint*, arXiv:2506.09513.

Yuxuan Sun, Hao Wu, Chenglu Zhu, Sunyi Zheng, Qizi Chen, Kai Zhang, Yunlong Zhang, Dan Wan, Xiaoxiao Lan, Mengyue Zheng, Jingxiong Li, Xinheng Lyu, Tao Lin, and Lin Yang. 2025b. Pathmmu: A massive multimodal expert-level benchmark for understanding and reasoning in pathology. In *Computer Vision – ECCV 2024*, pages 56–73, Cham. Springer Nature Switzerland.

Xiangru Tang, Daniel Shao, Jiwoong Sohn, Jiapeng Chen, Jiayi Zhang, Jinyu Xiang, Fang Wu, Yilun Zhao, Chenglin Wu, Wenqi Shi, Arman Cohan, and Mark Gerstein. 2025. MedAgentsBench: Benchmarking thinking models and agent frameworks for complex medical reasoning. *Preprint*, arXiv:2503.07459.

The BMJ. 2024. Central oversight, clinical guidelines, and careful implementation of ai in healthcare. `https://www.bmj.com/content/384/bmj.q596/rr-1`.

Michael Trimble and Paul Hamilton. 2016. The thinking doctor: clinical decision making in contemporary medicine. *Clinical Medicine*, 16(4):343–346.

David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966, Florence, Italy. Association for Computational Linguistics.

Xidong Wang, Guiming Chen, Song Dingjie, Zhang Zhiyi, Zhihong Chen, Qingying Xiao, Junying Chen, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2024a. CMB: A comprehensive medical benchmark in Chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6184–6205, Mexico City, Mexico. Association for Computational Linguistics.

Xidong Wang, Nuo Chen, Junyin Chen, Yidong Wang, Guorui Zhen, Chunxian Zhang, Xiangbo Wu, Yan Hu, Anningzhe Gao, Xiang Wan, Haizhou Li, and Benyou Wang. 2024b. Apollo: A lightweight multilingual medical llm towards democratizing medical ai to 6b people. *Preprint*, arXiv:2403.03640.

Zifeng Wang, Qiao Jin, Jiacheng Lin, Junyi Gao, Jathurshan Pradeepkumar, Pengcheng Jiang, Benjamin Danek, Zhiyong Lu, and Jimeng Sun. 2025. Trialpanorama: Database and benchmark for systematic review and design of clinical trials. *Preprint*, arXiv:2505.16097.

Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A Pfeffer, Jason Fries, and Nigam H Shah. 2023. The shaky foundations of large language models and foundation models for electronic health records. *npj digital medicine*, 6(1):135.

Chaoyi Wu, Pengcheng Qiu, Jinxin Liu, Hongfei Gu, Na Li, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025a. Towards evaluating and building versatile large language models for medicine. *npj Digital Medicine*, 8(1):58.

Jiageng Wu, Bowen Gu, Ren Zhou, Kevin Xie, Doug Snyder, Yixing Jiang, Valentina Carducci, Richard Wyss, Rishi J Desai, Emily Alsentzer, Leo Anthony Celi, Adam Rodman, Sebastian Schneeweiss, Jonathan H. Chen, Santiago Romero-Brufau, Kueiyu Joshua Lin, and Jie Yang. 2025b. Bridge: Benchmarking large language models for understanding real-world clinical practice text. *Preprint*, arXiv:2504.19467.

Xian Wu, Yutian Zhao, Yunyan Zhang, Jiageng Wu, Zhihong Zhu, Yingying Zhang, Yi Ouyang, Ziheng Zhang, Huimin Wang, Zhenxi Lin, Jie Yang, Shuang Zhao, and Yefeng Zheng. 2025c. Medjourney: benchmark and evaluation of large language models over patient clinical journey. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Feng Xie, Jun Zhou, Jin Wee Lee, Mingrui Tan, Siqi Li, Logasan S/O Rajnthern, Marcel Lucas Chee, Bibhas Chakraborty, An-Kwok Ian Wong, Alon Dagan, and 1 others. 2022. Benchmarking emergency department prediction models with machine learning and public electronic health records. *Scientific Data*, 9(1):658.

Lawrence K. Q. Yan, Qian Niu, Ming Li, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Benji Peng, Ziqian Bi, Pohsun Feng, Keyu Chen, Tianyang Wang, Yunze Wang, Silin Chen, Ming Liu, and Junyu Liu. 2024. Large language model benchmarks in medical tasks. *arXiv preprint arXiv:2410.21348*.

Wenxuan Yang, Weimin Tan, Yuqi Sun, and Bo Yan. 2024. A medical data-effective learning benchmark for highly efficient pre-training of foundation models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 3499–3508, New York, NY, USA. Association for Computing Machinery.

WonJin Yoon, Shan Chen, Yanjun Gao, Zhanzhan Zhao, Dmitriy Dligach, Danielle S Bitterman, Majid Afshar, and Timothy Miller. 2025. Lcd benchmark: long clinical document benchmark on mortality prediction for language models. *Journal of the American Medical Informatics Association*, 32(2):285–295.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *Preprint*, arXiv:2311.16502.

Ming Zhang, Yujiong Shen, Zelin Li, Huayu Sha, Binze Hu, Yuhui Wang, Chenhao Huang, Shichun Liu, Jingqi Tong, Changhao Jiang, Mingxu Chai, Zhiheng Xi, Shihan Dou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025. Llmeval-med: A real-world clinical benchmark for medical llms with physician validation. *Preprint*, arXiv:2506.04078.

Sheng Zhang, Xin Zhang, Hui Wang, Lixiang Guo, and Shanshan Liu. 2018. Multi-scale attentive interaction networks for chinese medical question answer selection. *IEEE Access*, 6:74061–74071.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Pmc-vqa: Visual instruction tuning for medical visual question answering. *Preprint*, arXiv:2305.10415.

Appendix

In this appendix, we provide additional information about *MedCheck*. Appendix A discusses clinically-grounded evaluation and reporting frameworks. Appendix B shows the benchmark searching and selecting procedure, and Appendix C details our evaluation setup. Appendix D outlines the classification of clinical and medical benchmarks. Appendix E lists the 56 evaluated medical LLM benchmarks, and Appendix F provides extra quantitative analysis of the evaluation results. Appendix H shows the complete list of evaluation criteria. Finally, Appendix I provides detailed scoring and explanations for a representative benchmark, and Appendix J presents an actionable diagnostic report example based on a clinical-oriented benchmark.

A Clinically-Grounded Evaluation and Reporting Frameworks

Beyond the benchmark datasets themselves, clinical informatics researchers have developed a parallel stream of work focused on establishing rigorous methodologies for AI model evaluation, validation, and transparent reporting. This research, published in top-tier medical and clinical informatics journals, emphasizes clinical utility, patient safety, and real-world applicability over leaderboard performance alone. A cornerstone of this ecosystem is the decades-long MIT and Harvard Medical School effort to develop, de-identify, and ethically share large-scale EHR datasets like MIMIC. The MIMIC project represents more than a data source—it embodies the principles of data curation, long-term maintenance, and community governance that are central to *MedCheck*'s framework, particularly in Phase II and Phase V.

Recent work emerging from this clinically-grounded tradition has produced several critical resources. The TRIPOD-LLM statement, for example, provides the first reporting guideline specifically for studies using LLMs, ensuring that methods and results are communicated transparently (Gallifant et al., 2025). While TRIPOD-LLM guides the reporting of a study, our *MedCheck* framework is a complementary tool designed to assess the quality of the benchmark used within that study. In parallel, comprehensive evaluation platforms like Med-HELM have been developed to assess LLMs on a wide array of clinically-grounded tasks using real EHR data. Med-HELM's focus is on holistically evaluating the model's capabilities, whereas *MedCheck*'s focus is on the prior, meta-level task of evaluating the measurement instrument itself.

The very premise of evaluating LLMs on medical tasks is predicated on the knowledge they contain. The landmark work by Singhal et al. (2023) was among the first to systematically demonstrate that LLMs do, in fact, encode a substantial amount of clinical knowledge. This finding underscores the urgency and importance of developing robust, clinically-valid benchmarks to accurately measure and safely harness this encoded knowledge, moving beyond exam-style questions to tasks that reflect the complexity of real-world clinical practice.

B Benchmark searching and selection

To ensure the objectivity and comprehensiveness of our evaluation, we implemented a systematic literature search and selection protocol divided into three stages: identification, screening, and eligibility. We first queried major academic databases—including Google Scholar, ACL Anthology, arXiv, and PubMed—for publications from January 2018 to July 2025. The search strategy employed Boolean combinations of keywords across subject, domain, and artifact type, specifically incorporating terms such as "medical LLM," "medical benchmark," "healthcare," "clinical," "clinical decision," and data-specific identifiers like "EHR" (electronic health record) and "EMR." This initial retrieval, focused on titles and abstracts to ensure relevance, yielded a total of 482 candidate publications after removing duplicates.

Subsequently, we applied stringent exclusion criteria to filter these candidates. We excluded studies strictly limited to traditional discriminative paradigms (e.g., BERT-based classification) that have not been adapted for evaluating the generative reasoning capabilities of LLMs. Furthermore, we removed benchmarks lacking publicly accessible data or code repositories essential for our reproducibility analysis, as well as studies that merely re-evaluated existing datasets without introducing novel contributions. This screening process narrowed the candidate pool to 85 potential benchmarks.

From this screened corpus, we selected the final set of 56 benchmarks based on specific criteria regarding impact and community adoption. Addressing the ambiguity of "widely used," we defined significance through quantitative metrics tai-

lored to the publication timeframe. For established benchmarks published prior to 2023, we required a minimum citation count of 50 to verify community adoption. Conversely, for emerging benchmarks published in 2024 and 2025, we selected those that were either accepted by top-tier venues (e.g., ACL, NeurIPS, Nature Medicine) or represented a unique task category not covered by older benchmarks. This multi-stage protocol ensures that our analysis captures the most significant and representative contributions to the medical LLM landscape.

C Evaluation Setup

In the evaluation process of this study, the primary focus was on manual scoring conducted by domain experts, with the use of LLMs providing supplementary assistance. While LLM were strictly instructed to follow the evaluation criteria, they inevitably exhibited hallucinations. Thus, the manual evaluation process remained the cornerstone of the benchmark assessment, with LLM serving as a tool to facilitate expert review by quickly retrieving relevant information and providing references for scoring.

This section will first describe the manual evaluation process, outlining how human annotators were trained, calibrated, and involved in scoring. It will then explain how the LLM-as-judge approach was implemented to enhance efficiency and assist in the evaluation process.

C.1 Guideline for annotator

To ensure reliable and consistent scoring of benchmarks according to the proposed evaluation framework, we established a rigorous annotation protocol executed by three expert annotators. The process comprised two phases: (1) a preparation and calibration phase, and (2) a formal scoring phase, as detailed below.

Preparation and Calibration Phase One week prior to scoring, all three annotators were provided with the full evaluation criteria (Appendix H). After confirming that each annotator had thoroughly reviewed and understood the criteria, we convened an online calibration meeting to accomplish the following objectives: (a) Jointly review the evaluation criteria; (b) Discuss and resolve discrepancies, with particular attention to clarifying ambiguous standards or dimensions with similar names, and establish consensus on terminology and definitions; (c) Harmonize overall scoring expectations (As shown in Table 6); (d) Clarify acceptable sources of evidence: scores must be grounded in publicly accessible documentation (e.g., main paper text, appendices, GitHub repositories, technical reports, official websites). If required information is absent, the item receives a default score of 0, with a comment stating "Information not provided."

Table 6: Overall Scoring Criteria

Score	Description
0	The criterion is completely absent, severely lacking, or contains fundamental flaws. *Judgment basis:* No relevant information; clearly incorrect information.
1	Initial attempt made, but incomplete, lacking details, methodologically crude, or unvalidated. *Judgment basis:* Mentioned but not implemented; implemented but without evaluation; described ambiguously, making validation impossible.
2	Clear, complete, transparent, and aligned with current domain consensus or literature-recommended best practices. *Judgment basis:* Explicit description + implementation + validation; citation of authoritative methods; open-source / reproducible / peer-reviewed / recognized by the community.

After each annotator completed scoring at least three benchmarks, a second alignment meeting was held to address emerging ambiguities or inconsistencies encountered during initial scoring. Based on this discussion, all previously submitted scores were reviewed and revised as necessary to maintain inter-annotator consistency.

To ensure reliability and directly address the limitation of single-annotator-per-case for the main corpus, we established a shared evaluation subset of 5 randomly selected benchmarks before the formal scoring phase. All three annotators independently scored this subset. We calculated the inter-annotator agreement using Fleiss' Kappa, achieving a strong agreement score ($\kappa = 0.78$), which indicates substantial consensus on the interpretation of the criteria.

Formal Scoring Phase A total of 56 benchmarks were uniformly and randomly distributed among the annotators. Each annotator received a randomly ordered list of benchmarks, accompanied by the LLM-generated scores and corresponding explanations for each benchmark, and was instructed to manually evaluate 1–2 benchmarks per day, adhering to the following requirements:

(a) For each scored item, provide specific evidence in the comment field (As illustrated in the

"Explanation" section of Appendix I); (b) Ignore authorship and institutional affiliations; avoid basing judgments on general impressions or prior familiarity with a benchmark; (c) Apply criteria strictly as defined—do not broaden or narrow the scope based on personal interpretation; (d) The LLM-generated scores and explanations are intended solely to serve as auxiliary aids—facilitating rapid information retrieval and offering preliminary scoring suggestions—and must not constitute the primary basis for evaluation. Their content must be rigorously verified for factual accuracy before being considered as reference material; (e) For items with insufficient or ambiguous information, assign a score of 0, 1, or 2 as appropriate, but include a brief justification in the comments (e.g., "The paper mentions data cleaning but does not detail the procedure") and prefix the note with "[Uncertain]"; (e) For borderline cases where a benchmark's documentation is vague or implicitly implies a feature, annotators were strictly instructed to default to the lower score to penalize the lack of transparency. A higher score was only awarded if concrete, verifiable evidence (e.g., explicit code snippets, specific GitHub commits, or dedicated appendix sections) could be explicitly cited in the comments. Any remaining uncertainties (flagged with "[Uncertain]") were brought to the weekly expert panel meetings, where consensus was reached through rigorous group discussion and majority voting; (f) Daily monitoring of individual annotator score distributions was conducted to detect and mitigate systematic bias (e.g., consistently high or low scoring tendencies).

Throughout the process, any issues or discrepancies in scoring were resolved through weekly consensus meetings among the expert panel, culminating in a final agreed-upon score for each criterion. This rigorous protocol substantially mitigated the well-documented susceptibility of manual scoring to various confounding factors—such as subjective bias, cognitive load, and order effects—and ensured that the evaluation process was maximally transparent, consistent, and reproducible.

C.2 LLM-as-Judge Scoring

While the primary evaluation in MedCheck is conducted manually by domain experts, we explored the use of LLM-as-judge scoring during the preliminary scoring phase to balance scalability with accuracy.

We include our evaluation criteria, the paper text, and documentation of medical benchmarks in the prompt. We then use GPT-4o-Search-Preview with temperature = 0 to generate scores in JSON format. To ensure the LLM produces reasonable scores based on our criteria, we explicitly instructed it to strictly adhere to the evaluation guidelines and provide an explanation for each score. Additionally, to guarantee the LLM has access to up-to-date public documentation, we enabled web search functionality, allowing it to retrieve supplementary information (e.g., the benchmark's official website, GitHub repository, etc.) when necessary. During the manual scoring process, the panel of domain experts refers to the LLM-generated scores and explanations, but they must verify the accuracy of the LLM-provided explanations and ultimately assign the final scores.

To address potential concerns regarding confirmation bias from LLM-generated scores, we explicitly frame our approach as an **"Assisted-Verification"** design. In the context of auditing extensive and complex benchmark documentation, independent "blind" human annotation often suffers from high false-negative rates due to cognitive fatigue and information overload (i.e., annotators inadvertently missing existing evidence). In our protocol, the LLM functions strictly as an information retrieval assistant to surface relevant sections and ensure high recall. The human expert then rigorously verifies the factual existence of the cited documentation to ensure high precision. Because the assessment relies on objective verification of factual evidence rather than subjective judgment, the risk of bias is minimized while the accuracy of evidence retrieval is significantly improved.

In summary, the evaluation process relied fundamentally on the expertise of human annotators to ensure the accuracy and consistency of the final scores. LLMs played a strictly auxiliary role—expediting data retrieval and providing preliminary scoring suggestions—while the ultimate judgment remained firmly in the hands of domain experts. This hybrid methodology synergistically combines the rigor and reliability of manual evaluation with the efficiency and scalability of LLMs, thereby ensuring an evaluation framework that is both comprehensive and practically viable.

D Classification of Clinical and Medical Benchmarks

We classify the 56 evaluated benchmarks into two distinct categories: **Clinical (42 benchmarks,**

75%) and **Medical (14 benchmarks, 25%)**. These categories exhibit a fundamental distinction in their objectives, application scenarios, and data sources. We argue that evaluating a benchmark from one category using the criteria intended for the other would be methodologically flawed. Therefore, to ensure a fair and context-aware assessment, the criteria used to define each category are detailed below.

- **Clinical Benchmarks (42/56, 75%):** These benchmarks primarily evaluate the application of LLMs in real-world clinical workflows.
  - **Objective:** Assess capabilities like processing EHRs, clinical reasoning with dynamic or incomplete information, and supporting diagnostic decisions.
  - **Scenario:** Simulate authentic clinical encounters, such as patient consultations, risk prediction, or treatment planning.
  - **Data Source:** Primarily use real-world data, including EHRs, clinical case notes, and doctor-patient dialogues.
- **Medical Benchmarks (14/56, 25%):** These benchmarks focus on assessing an LLM's foundational medical knowledge and theoretical reasoning.
  - **Objective:** Test the model's mastery of established medical facts and concepts.
  - **Scenario:** Typically involve standardized, knowledge-based tasks, often in a multiple-choice question format.
  - **Data Source:** Primarily use academic materials, such as medical exam questions, textbooks, and research literature.

E List of Evaluated Benchmarks

We assessed the following 56 benchmarks, categorized as either Clinical (C) or Medical (M) (alphabetical order):

- AfriMed-QA (Olatunji et al., 2025) (C)
- AgentClinic (Schmidgall et al., 2024) (C)
- Asclepius (Liu et al., 2025b) (C)
- BRIDGE (Wu et al., 2025b) (C)
- CHBench (Guo et al., 2024) (C)
- CheXpert (Irvin et al., 2019) (C)
- CLIMB (Dai et al., 2025) (C)
- CliMedBench (Ouyang et al., 2024) (C)
- ClinicBench (Liu et al., 2024a) (C)
- CMB (Wang et al., 2024a) (M)
- cMedQA2 (Zhang et al., 2018) (C)
- CMExam (Liu et al., 2023) (M)
- COGNET-MD (Panagoulias et al., 2024) (M)
- DataDEL (Yang et al., 2024) (C)
- DR.BENCH (Gao et al., 2023) (C)
- EHRNoteQA (Kweon et al., 2024) (C)
- EndoBench (Liu et al., 2025c) (C)
- ER-REASON (Mehandru et al., 2025) (C)
- GMAI-MMBench (Chen et al., 2024) (C)
- HeadQA (Vilares and Gómez-Rodríguez, 2019) (M)
- HealthBench (Arora et al., 2025) (C)
- LCD (Yoon et al., 2025) (C)
- LLMEval-Med (Zhang et al., 2025) (C)
- MedAgentBench (Jiang et al., 2025) (C)
- MedAgentsBench (Tang et al., 2025) (M)
- MedCalc (Khandekar et al., 2024) (C)
- MedChain (Liu et al., 2024b) (C)
- MedConceptsQA (Shoham and Rappoport, 2024) (M)
- MEDEC (Abacha et al., 2025) (C)
- MedExQA (Kim et al., 2024) (M)
- MEDIC (Kanithi et al., 2024) (C)
- MediQ (Li et al., 2024) (C)
- MedJourney (Wu et al., 2025c) (C)
- MedMCQA (Pal et al., 2022) (M)
- MedOdyssey (Fan et al., 2025a) (C)
- MedQA (Jin et al., 2021) (M)

- MedR-Bench (Qiu et al., 2025) (C)
- MedRisk (Liu et al., 2025a) (C)
- MedSafetyBench (Han et al., 2024) (C)
- MedS-Bench (Wu et al., 2025a) (C)
- MIMIC-IV-ED (Xie et al., 2022) (C)
- MLEC-QA (Li et al., 2021) (M)
- MMMU (Health & Medicine) (Yue et al., 2024) (C)
- MVME (AI hospital) (Fan et al., 2025b) (C)
- OmniMedVQA (Hu et al., 2024) (C)
- PathMMU (Sun et al., 2025b) (C)
- PathVQA (He et al., 2020) (M)
- PMC-VQA (Zhang et al., 2024) (C)
- PubMedQA (Jin et al., 2019) (M)
- ReasonMed (Sun et al., 2025a) (M)
- SLAKE (Liu et al., 2021) (C)
- TrialPanorama (Wang et al., 2025) (C)
- VQA-Med (Ben Abacha et al., 2021) (C)
- VQA-RAD (Lau et al., 2018) (C)
- webMedQA (He et al., 2019) (C)
- XMedBench (Wang et al., 2024b) (M)

F Quantitative Results

This appendix provides a summary of the quantitative analysis conducted on the 56 medical LLM benchmarks.

F.1 Publication Year and Task Focus of Benchmarks

Our collection covers open-source medical large model benchmarks published between 2018 and 2025.

Annual Trend. As shown in Figure 4, the publication years of the 56 benchmarks we collected. A clear turning point can be observed in 2024, when the volume of medical benchmark publications significantly increased. Prior to 2024, only 17 benchmarks were released. The brief surge in 2019, driven by the rise of deep learning and Transformer models, catalyzed short-term growth in medical AI and a peak in benchmark publications. However, no substantial upward trend was evident before 2024.

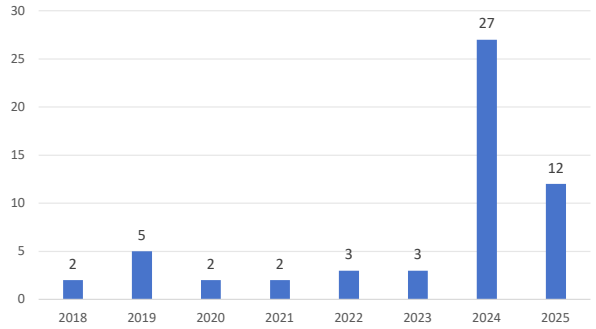

Figure 4: Publication Year Distribution of the 56 Medical Benchmarks

In contrast, with the emergence of large models such as ChatGPT and Med-PaLM in 2024, there was an explosive growth in medical NLP and multimodal benchmarks. A total of 27 medical benchmarks were published in 2024, surpassing the cumulative number of the previous six years. As of July 10, 2025, 12 new benchmarks were already released in the first half of 2025, showing a steady growth trend.

Task Trend. Before 2024, most benchmarks primarily focused on evaluating models' capabilities in NLP and VQA. As shown in Figure 5, following the explosion in benchmark quantity post-2024, the range of tasks covered expanded to six categories, with new benchmarks addressing specific tasks for large models, such as multimodal understanding, reasoning and decision-making, dialogue/interactive QA , as well as medical text interpretation and safety in medical recommendation generation.

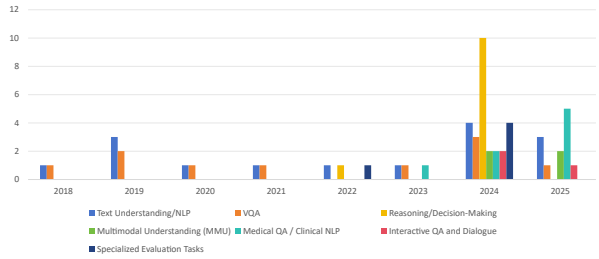

Figure 5: Temporal Distribution of Benchmarks by Task Type

In summary, as illustrated in Figure 6, more than half (64%) of the existing medical benchmarks focus on assessing models' text comprehension,

NLP, and VQA abilities. This indicates that most medical benchmarks historically centered around structured or unstructured medical text understanding. However, in recent years, there has been an increasing trend towards medical question answering and clinical NLP, with the highest number of such benchmarks being published in the first half of 2025.

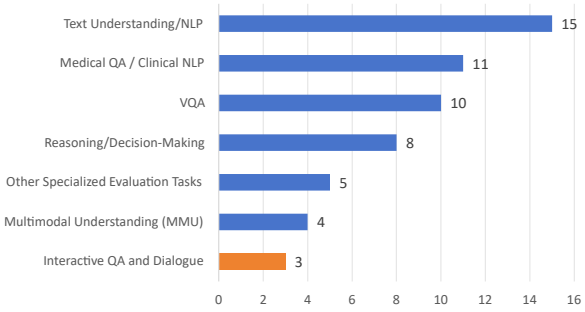

Figure 6: Task Type Distribution of the 56 Medical Benchmarks

F.2 Natural Language

Figure 7 presents the distribution of natural language usage across the 56 collected benchmarks. It is evident that English dominates, being used in 45 of the benchmarks (approximately 80%), establishing itself as the primary language for evaluating medical LLMs. Chinese ranks second, appearing in 17 benchmarks (approximately 30%), indicating a notable level of development in Chinese medical AI. In contrast, other languages such as Arabic, Russian, Portuguese, Japanese, and Hindi—despite being spoken by large global populations—remain underrepresented in publicly available benchmarks, with each language appearing in only 2 to 6 benchmarks on average. This highlights that multilingual medical AI remains in its early stages.

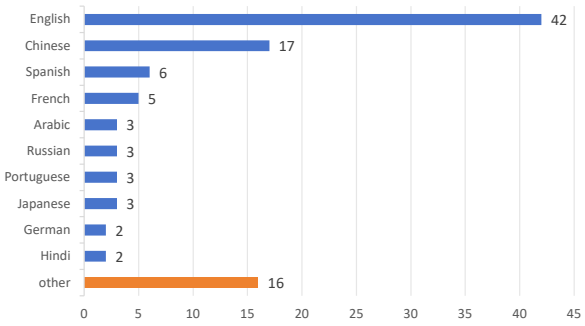

Figure 7: Natural Language Distribution of the Benchmarks

Finally, it may be concluded that current medical LLM benchmarks exhibit a pronounced trend of language centralization. English, as the de facto language of international medical communication, dominates the vast majority of evaluation tasks. Although multilingual benchmarks are still in their infancy, they hold substantial potential in promoting global medical equity and cross-cultural adaptation. Moving forward, greater efforts are needed to improve the availability of multilingual data resources and to develop cross-lingual evaluation strategies.

F.3 Medical Diversity

Distribution of Diseases. Figure 8 illustrates the distribution of disease categories represented in the 56 medical benchmarks, categorized according to the first 23 chapters of the ICD-11 classification system (with Certain conditions originating in the perinatal period and Developmental anomalies excluded due to the absence of relevant benchmarks). The analysis reveals that current research on medical LLMs is heavily concentrated in disease areas such as Diseases of the musculoskeletal system or connective tissue (16 benchmarks), Diseases of the skin (15), Diseases of the respiratory system (14), and Neoplasms (13). These categories typically have abundant publicly available imaging data (e.g., X-rays, CT scans, pathology slides), which facilitates the development of standardized evaluation protocols based on both image and text modalities.

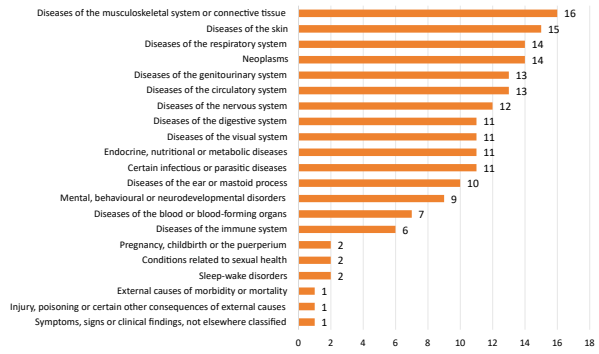

Figure 8: Disease Distribution Based on ICD-11 Classification

In contrast, low-frequency disease categories such as Mental and behavioral disorders (e.g., depression, anxiety) rely on psychological scales, clinical interviews, or life history data, which are difficult to standardize and rarely captured in benchmark datasets. Similarly, Immunological and Hematological diseases often require the integration of laboratory indicators, clinical progression,

and imaging, making them difficult to model using unimodal approaches. From a clinical perspective, however, these underrepresented diseases are not necessarily rare in real-world settings. Many exhibit high outpatient visit rates, cause significant patient distress, or carry substantial social burden. This highlights a misalignment between research frequency and clinical demand. Future research in medical LLMs should be designed with greater awareness of this gap, ensuring that high-demand but rarely covered disease populations are not systematically neglected.

Medical Specialties. While ICD-11 provides a useful framework for disease-based diversity analysis, it lacks granularity in distinguishing among interdisciplinary or population-specific fields such as clinical medicine, obstetrics and gynecology, and pediatrics. To complement ICD-11-based analysis, we additionally introduce a second diversity metric based on the official list of clinical departments issued by the National Health Commission of China. Figure 9 shows the distribution of benchmarks across these medical specialties (excluding departments such as Endemic Diseases that are not represented in any benchmark or that overlap with ICD-11 categories).

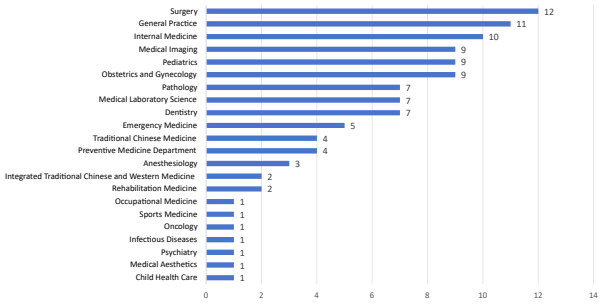

Figure 9: Medical Specialty Distribution Based on the Official Directory of Clinical Departments

This classification reveals a similar pattern: departments with well-structured data and clearly defined tasks dominate the landscape, such as Surgery (12 benchmarks), General Practice (11), and Medical Imaging (9). In contrast, specialties associated with vulnerable populations or complex care—such as Rehabilitation Medicine, Occupational Health, and Child Health Care—are significantly underrepresented.

This imbalance may result in limited generalizability and fairness in real-world deployment of medical LLMs, particularly in addressing the needs of underserved or structurally marginalized groups. Therefore, future benchmark design should aim for more balanced coverage across medical specialties and actively promote the standardization and open sharing of data from underrepresented departments. This will support the development of AI systems that are not only technically robust but also clinically inclusive.

F.4 Benchmark Performance Heatmap Example

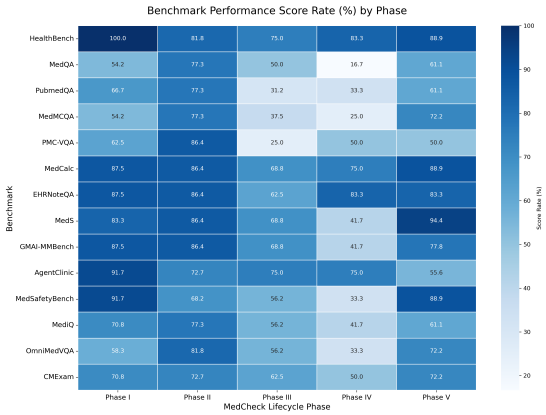

Figure 10: Performance score rates (%) for 14 medical LLM benchmarks across the five phases of the MedCheck evaluation framework.

The heatmap in Figure 10 visualizes the performance of 14 out of 56 of the medical LLM benchmarks evaluated evaluated against the five-phase *MedCheck* framework. Each row represents a benchmark, and columns correspond to the framework's phases. The color intensity reflects the score rate (%), highlighting the comparative strengths and weaknesses of each benchmark and revealing systemic trends.

F.5 Statistics About Phase I: Design and Conceptualization

This section presents a systematic evaluation of the design logic and conceptual clarity of existing medical benchmarks. we examine their alignment with medical capabilities, practical application scenarios, and innovation objectives. The assessment is structured around three core aspects: (1) the clarity of domain positioning from the perspective of AI capabilities and clinical scope, (2) the methodological soundness in terms of metric design and evaluation diversity, and (3) the integration of safety and fairness considerations as part of responsible AI practices.

As shown in Figure 11, the average normalized

score across benchmarks is 74.5% (17.9 out of 24), with a mean of 1.5 points per criterion, indicating general compliance with baseline expectations. However, performance on several key items remains suboptimal and warrants further attention.

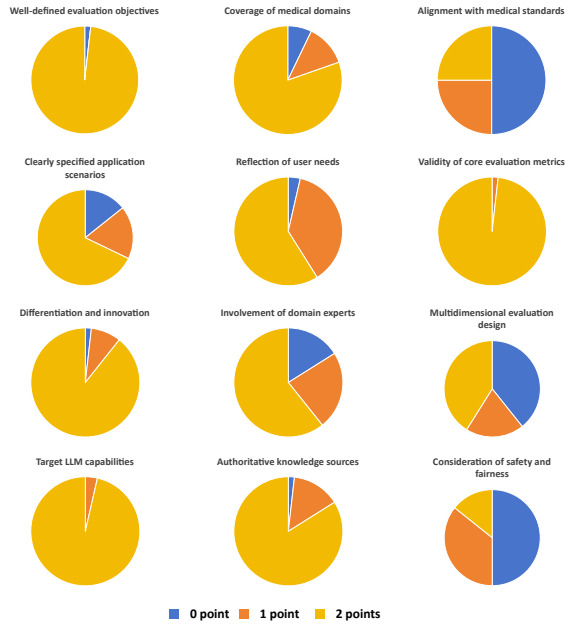

Figure 11: Scoring Performance of the 56 Benchmarks in the Design and Conceptualization Stage

Clearly specified application scenarios. To ensure practical relevance, benchmarks should clearly define their intended clinical or biomedical research scenarios. Such grounding facilitates real-world applicability and aligns evaluation with deployment needs. As illustrated in Figure 12, 68% of benchmarks specify concrete application scenarios, articulating clinical value and expected utility. In contrast, 18% provide only general references to decision-support contexts, and 14% fail to specify any scenario, reducing interpretability and downstream usability.

Involvement of domain experts. Expert input is a critical component in medical AI development. It ensures the appropriateness of task design, data annotation quality, and answer validity. However, as shown in Figure 13, only 23% of benchmarks explicitly report domain expert involvement, often without detailing the roles or qualifications. 16% do not mention expert participation, raising concerns about the benchmarks' clinical validity. Future benchmark design should adopt standardized expert reporting practices to enhance credibility, transparency, and methodological rigor.

Alignment with medical standards. Adherence to international medical standards (e.g., ICD-11, SNOMED CT, LOINC) is essential for achieving semantic interoperability, annotation consistency, and cross-system comparability. As Figure 14 indicates, 50% of benchmarks do not clarify whether their data conforms to such standards. This omission undermines the benchmark's reusability and weakens the comparability of evaluation outcomes. Without standard-based structure and terminology, label ambiguity and inconsistent categorization may arise, which compromises the reliability of model performance evaluation.

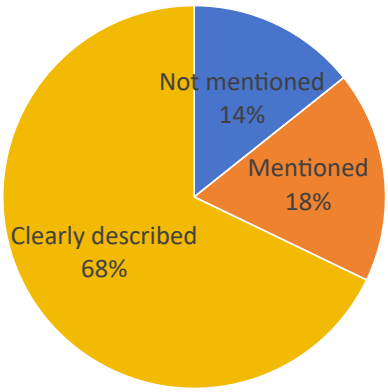

Figure 12: Performance on Clarity of Application Scenario

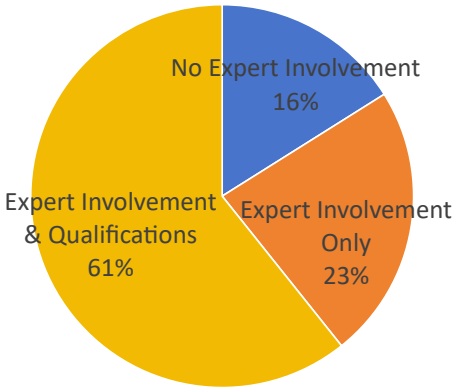

Figure 13: Performance on Domain Experts Involvement

Multidimensional evaluation design. In medical AI, a model's performance cannot be solely measured by accuracy. Outputs that are correct but unsafe, incomplete, or biased may have detrimental consequences. Multidimensional evaluation frameworks are required to assess whether models' output is not only correct but also reliable. As presented in Figure 15, 39% of benchmarks use only

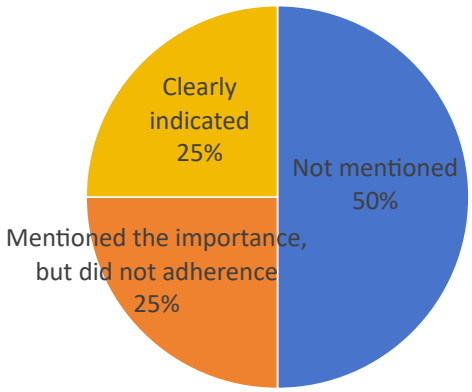

Figure 14: Performance on Medical Standards Alignment

a single metric, typically accuracy, and 20% mention other metrics without detailed implementation. Metrics such as uncertainty estimation, answer refusal, and harmful output detection are absent from most benchmarks, despite their relevance in high-stakes medical decision-making.

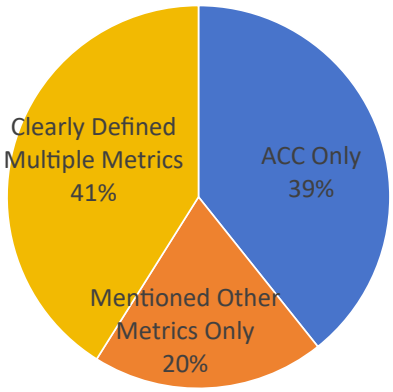

Figure 15: Performance on Multi-dimensional Evaluation

Consideration of safety and fairness. Robust evaluation frameworks must account for safety and fairness, especially in sensitive domains like healthcare. As shown in Figure 16, only 14% of benchmarks include explicit safety or fairness evaluations, while 36% acknowledge their importance without implementation. The remaining 50% omit these aspects entirely. Neglecting safety assessments increases the risk of misleading or overconfident outputs, which may harm clinical decisions. The lack of fairness evaluation may propagate systematic bias across gender, age, race, or disease distribution, exacerbating healthcare disparities. To build trustworthy and ethical medical AI, safety and fairness must be treated as core evaluation dimensions.

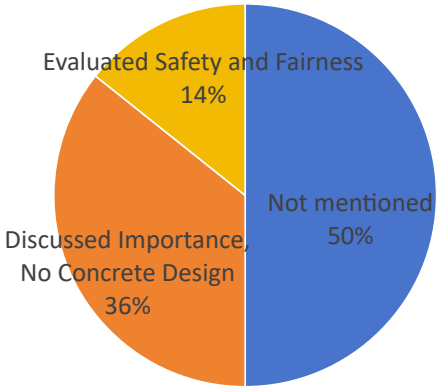

Figure 16: Performance on Consideration of Safety and Fairness

F.6 Statistics About Phase II: Dataset Construction and Management

This phase constitutes the empirical foundation of the benchmark, with an emphasis on constructing datasets that are realistic, diverse, representative, and ethically sourced. Key elements include stringent procedures for data quality assurance, privacy protection, and contamination prevention. The dataset sources and quality were assessed across five dimensions: transparency and traceability of data provenance, reliability of data sources, and data authenticity. Data processing and privacy protection were evaluated in six aspects, including data cleaning and standardization, privacy-preserving mechanisms, and data formatting protocols. As shown in Figure 17, the 56 benchmarks achieved an average score rate of 74.7 % (16.4 out of 22) across 11 evaluation criteria, with a mean of 1.4 points per item. Overall, the results indicate general compliance with the proposed standards, although two specific criteria showed suboptimal performance.

Dataset Diversity. This criterion evaluates whether the benchmark explicitly defines diversity objectives—such as coverage of disease types and clinical departments—and provides quantitative evidence (e.g., disease distribution histograms, task type ratios) to demonstrate dataset coverage. Ensuring sufficient diversity is critical to evaluate model generalization across a broad range of cases and specialties, thereby mitigating evaluation bias.

The specific calculation is shown in Equation 1. In this Equation, N_{disease} represents the number of diseases in the ICD-11 standard, specifically

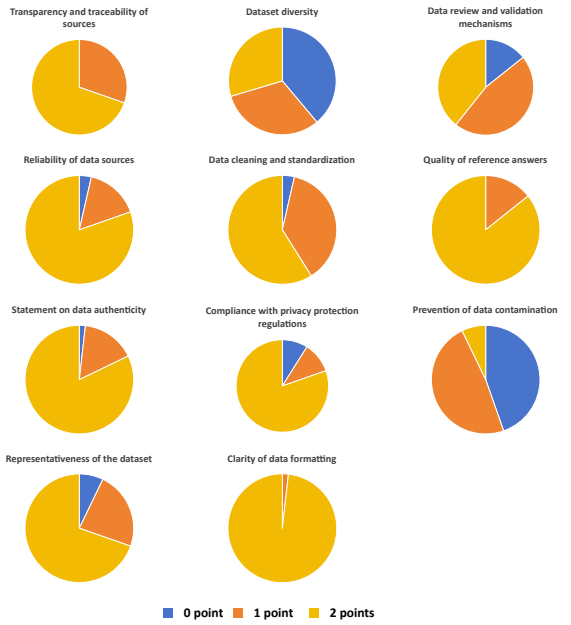

Figure 17: Scoring Performance of the 56 Benchmarks in the Phase II: Dataset Construction and Management

the first 23 diseases, serving as the benchmark for the types of diseases considered. $N_{\text{department}}$ denotes the number of medical departments, typically referring to the medical specialties included in the model's evaluation. The variable $N_{\text{disease}}^{\text{benchmark}}$ indicates the number of diseases covered by the benchmark, representing the actual disease categories included in the model's evaluation. Similarly, $N_{\text{department}}^{\text{benchmark}}$ refers to the number of medical departments involved in the benchmark. Finally, R_{coverage} represents the coverage ratio, which is calculated by comparing the sum of the diseases and medical departments covered by the benchmark $N_{\text{disease}}^{\text{benchmark}} + N_{\text{department}}^{\text{benchmark}}$ to the total number of diseases and medical departments in the reference standard $N_{\text{disease}} + N_{\text{department}}$. This formula provides a quantitative measure of the benchmark's coverage across various diseases and medical departments, reflecting the model's diversity and comprehensiveness in its evaluation.

The coverage of each benchmark after calculation is shown in Fig. 18. We scored each benchmark based on the average total coverage rate (21.8%), and the results are presented in Fig 19, 39% (21 out of 56) of the benchmarks lack a clear definition or discussion of dataset diversity. The absence of well-defined diversity goals and corresponding quantitative analyses hinders the ability to assess model robustness across different disease categories or clinical settings, potentially obscuring vulnerabilities in specific subgroups or task scenarios.

$$R_{\text{coverage}} = \frac{N_{\text{disease}}^{\text{benchmark}} + N_{\text{department}}^{\text{benchmark}}}{N_{\text{disease}} + N_{\text{department}}} \quad (1)$$

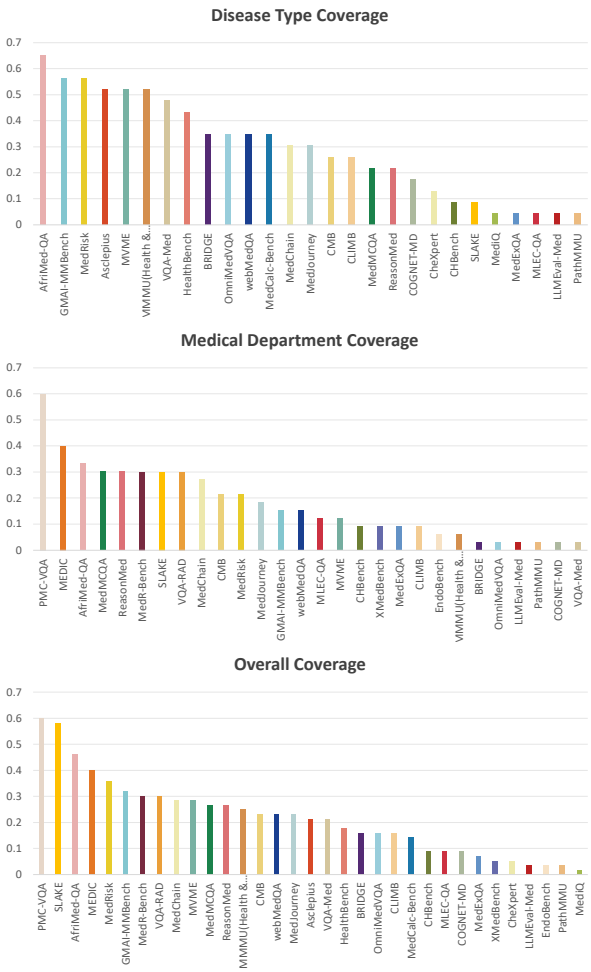

Figure 18: Coverage of disease types, medical departments, and overall performance across 56 medical benchmarks.

Prevention of Data Contamination. This dimension assesses whether the benchmark includes mechanisms to identify and mitigate potential data contamination—i.e., the presence of evaluation samples within the training corpus of LLMs. Such contamination can artificially inflate model performance through memorization, thereby compromising the fairness and validity of the evaluation. If present in high-risk tasks (e.g., diagnostic decisions or medication recommendations), contamination may mislead researchers and regulators in their assessment of model safety. Therefore, contamination detection is a prerequisite for ensuring the

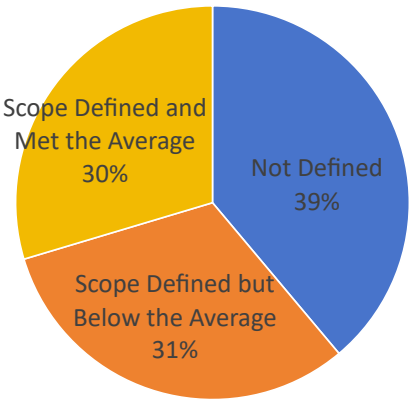

Figure 19: Performance on Dataset diversity

credibility and scientific rigor of benchmark evaluations. However, as depicted in Figure 20, only 7% of the benchmarks implemented both detection and mitigation strategies. Another 48% conducted detection without addressing identified contamination, while 45% did not mention the issue at all.

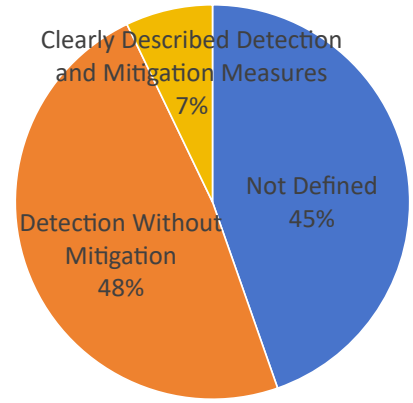

Figure 20: Performance on Data Contamination Prevention

F.7 Statistics About Phase III: Technical Implementation and Evaluation Methodology

This is the operational backbone. It consists of 8 items, encompassing the development of accessible, reproducible evaluation scripts and the selection of metrics that go beyond simple accuracy to assess deeper capabilities such as reasoning, robustness, and uncertainty awareness. For the 56 evaluated benchmarks, the average score is 52.1% (8.33 out of 16), with a mean score of 1.04 for each item. Results indicate that this is the second worst performing stage among the 5 stages. As shown in Figure 21, there are 3 items exhibiting critical underperformance, where majority of the benchmarks receive 0 mark.

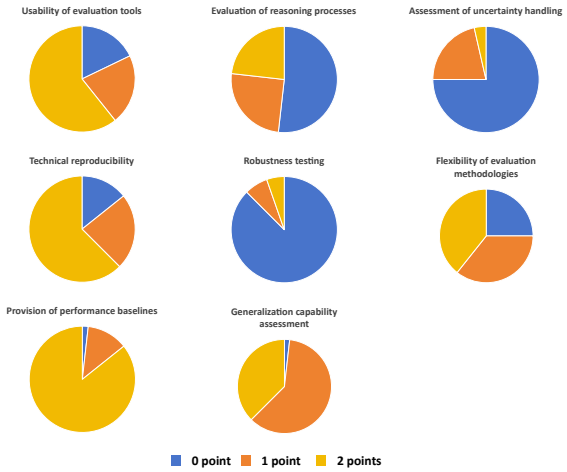

Figure 21: Scoring Performance of the 56 Benchmarks in the Phase III: Technical Implementation and Evaluation Methodology

Reasoning Process Evaluation. This criterion assesses whether the benchmarks include evaluation for the reasoning process of the models. Apart from the final answer, understanding the decision-making process of models equally important.

As shown in Figure 22, a concerning 55% of benchmarks do not have any consideration for the models' reasoning process. While 25% of benchmarks mention the importance of evaluating the reasoning process, only 23% of benchmarks design concrete assessment for reasoning process with clear methods and metrics.

Without evaluating the reasoning process, the flawed logic or hidden biases of models may left unchecked, potentially compromising patient safety. By assessing the reasoning path, how and why a model arrives at its conclusions can be revealed, fostering transparency and trustworthiness.

Robustness Evaluation. This item assesses whether there is evaluation for robustness in the benchmark. In practical application, a robust model should be able to reliably interpret different variations of input. Robustness evaluation help assess whether models are resilient to imperfections and perturbations, ensuring the reliability when adopted in high-stakes medical scenarios.

As depicted in Figure 23, an overwhelming 88% of the covered benchmarks do not consider robustness in their evaluation. While 4 benchmarks recognize the importance of robustness evaluation, only

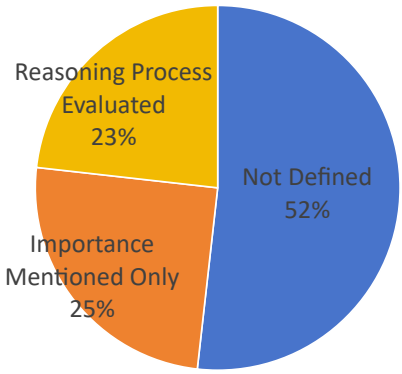

Figure 22: Performance on Reasoning Process Evaluation

3 benchmarks design clear assessment for robustness.

Real-world environments are inherently variable and noisy. While models may perform well under idealized inputs in the benchmark, they may fail when facing subtle changes. Without robustness evaluation, benchmark results could be overly optimistic. This false confidence may lead to misleading decisions, causing risks when deployed in real-world environment. By evaluating robustness, users can gain a better understanding of models' stability and consistency, fostering safer and more dependable deployment in complex medical environments.

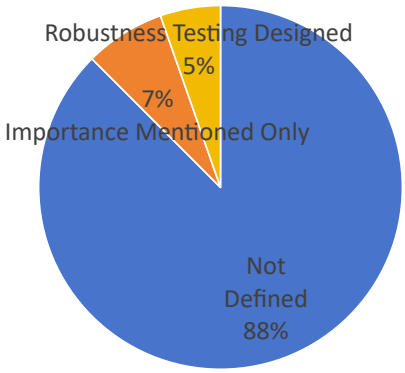

Figure 23: Performance on Robustness evaluation

Uncertainty Evaluation. This criterion is designed for evaluating the ability of models to recognize and express its uncertainty. In high-stakes medical contexts, it is vital for models to recognize and communicate the limits of their knowledge. Assessing how safely and responsibly a model handles uncertainty promotes caution, reducing the risk of overconfident errors and misleading decisions.

As illustrated in Figure 24, 75% benchmarks have no considerations for uncertainty in their evaluation. Even though 21% acknowledge the importance of uncertainty evaluation, only 4% incorporate evaluation related to uncertainty in the benchmark, highlighting a significant gap in current benchmarking practices.

If uncertainty evaluation is omitted, it can lead to misleading performance results that overstate a model's safety and reliability. When facing uncertainty, if models provide overconfident answers, it could lead to harmful outcomes. Benchmarks without uncertainty evaluation fail to reward models that behave ethically and responsibly under uncertainty, undermining the development of safer AI tools in medicine. By evaluating uncertainty, benchmarks can better reflect clinical uncertainty, fostering more trustworthy and safer decision-making behavior

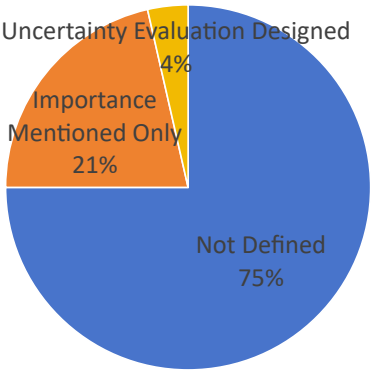

Figure 24: Performance on Uncertainty Evaluation

F.8 Statistics About Phase IV: Benchmark Validity and Performance Verification

This phase provides scientific validation for the benchmark. It encompasses evidence of content validity (i.e., whether the benchmark comprehensively covers the target domain) and construct validity (i.e., whether it measures the claimed latent capabilities), as well as its ability to distinguish performance differences among models of varying proficiency levels. Among the 56 benchmarks evaluated, the average score rate for this phase was 49.1% (5.90 out of 12), with a mean of 0.98 points per item—representing the lowest performance among the five phases. As demonstrated in Figure 25, of the six evaluation criteria in this phase, three demonstrated notably poor results.

Correlation with Clinical Performance. This

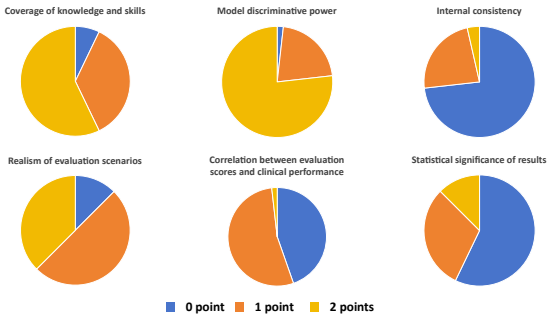

Figure 25: Scoring Performance of the 56 Benchmarks in the Phase IV: Benchmark Validity and Performance Verification

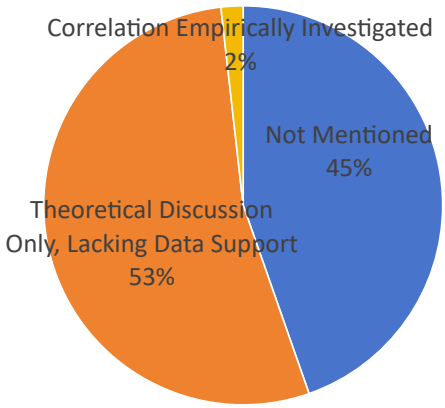

Figure 26: Performance on Correlation with Clinical Performance

component assesses whether there is empirical evidence exploring the relationship between benchmark scores and model performance in real-world clinical applications. Specifically, it examines whether the benchmark scores can reliably reflect clinical effectiveness—an essential indicator of external validity and practical utility. In medical contexts, the ultimate goal of model performance is to enhance the accuracy, safety, and efficiency of clinical decision-making. If evaluation scores fail to map onto actual clinical outcomes, the benchmark may mislead researchers regarding model applicability. Moreover, neglecting this correlation may incentivize developers to over-optimize non-essential capabilities for better scores, thus deviating from real-world deployment needs.

As presented in Figure 26, among the 56 benchmarks reviewed, 45% did not address the correlation between evaluation scores and clinical performance. A further 53% (28 benchmarks) discussed expected correlations at a theoretical level but lacked supporting experimental evidence. Only 2% (1 benchmark) provided preliminary empirical analysis exploring the correlation, along with a discussion of the results. These findings indicate a significant gap in validating the practical relevance of benchmark scores. Future medical benchmarks should prioritize establishing and verifying correlations between evaluation outcomes and clinical workflows or expert assessments. This is essential to ensure that benchmark scores serve as credible and informative indicators for real-world model deployment.

Internal Consistency. This metric examines whether different components or items within a benchmark consistently measure the same target capability, thereby ensuring internal reliability. High internal consistency reduces the risk that variability in item quality, phrasing, or task context will distort evaluation results. Figure 27 shows that in this study, only 4% of the benchmarks reported statistical indicators of internal consistency, with positive results. An additional 23% also reported such metrics, but without clear discussion or yielded inconclusive findings. The remaining 73% did not conduct internal consistency assessments. The absence of such analysis raises concerns regarding evaluation accuracy and fairness. Inconsistent benchmarks may obscure true model strengths or weaknesses across specific skill dimensions, impeding targeted model refinement. In clinical applications, this may further compromise risk assessment and regulatory oversight. Therefore, future benchmarks should prioritize the evaluation and reporting of internal consistency to enhance reliability.

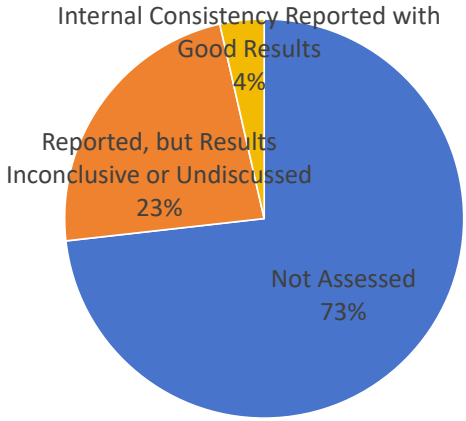

Figure 27: Performance on Internal consistency

Statistical Significance Reporting. This criterion evaluates whether statistical significance or uncertainty—such as confidence intervals or p-

values—was reported when comparing model performances. In the absence of such analyses, observed performance differences may arise from sample randomness or dataset variability rather than genuine model capability disparities. This undermines the scientific rigor of the benchmark and may result in erroneous conclusions during model selection or optimization. In high-stakes medical scenarios, such inaccuracies pose safety risks if incorporated into decision-support systems or diagnostic tools. Consequently, the systematic reporting of statistical measures such as confidence intervals, p-values, or variance analyses is essential for constructing trustworthy and interpretable evaluation frameworks.

It can be seen from Figure 28 that, in practice, 57% of the benchmarks reported only single-run results. Although 30% provided mean and standard deviation across multiple runs, they did not employ formal statistical tests to compare model performance. These findings highlight a lack of rigorous statistical standards in current benchmark design. Incorporating structured multi-run strategies and statistical significance testing is critical to enhance the reliability and validity of performance evaluations in medical AI.

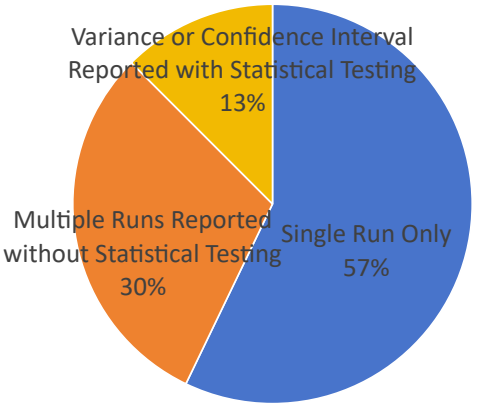

Figure 28: Performance on Statistical significance Reporting

F.9 Statistics About Phase V: Documentation, Openness, and Governance

Documentation, Openness, and Governance represent community-facing components that are essential for ensuring the long-term value and impact of a benchmark. This phase includes the provision of clear documentation, adherence to open-source principles, and the establishment of governance mechanisms for maintenance, version control, and community feedback. Among the 56 benchmarks evaluated, as depicted in Figure 29, the average score rate for this phase was 66.2% (11.9 out of 18), with a mean score of 1.3 per item. The scoring distribution in this phase was relatively balanced, with only one criterion showing slightly weaker performance.

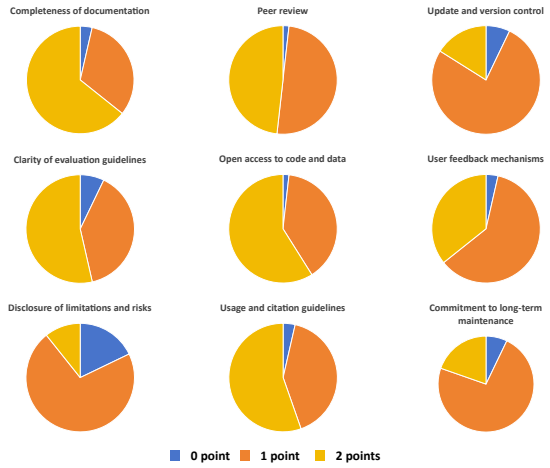

Figure 29: Scoring Performance of the 56 Benchmarks in the Phase V: Documentation, Openness, and Governance

Discussion of Limitations and Risks. This criterion evaluates whether the benchmark documentation openly and explicitly discusses its inherent limitations, potential societal risks, or misuse scenarios. No benchmark is flawless. Proactively disclosing such limitations and risks reflects scientific rigor and a strong sense of responsibility on the part of the developers. It also helps prevent overreliance on benchmark results or inappropriate applications. Figure 30 shows that Only 11% of the benchmarks provided thorough and candid discussions of limitations and associated risks. A majority—71%—briefly mentioned limitations but lacked in-depth analysis. The remaining 18% (10 out of 53) did not discuss any form of limitation or risk.

In the medical domain, the responsibilities of a benchmark extend beyond merely providing performance scores. Benchmarks should also serve as tools for risk communication and usage guidance. Failure to include such information undermines both scientific robustness and standardization, and diminishes the benchmark's value for clinical practice and policy-making. Therefore, systematic disclosure of limitations and potential risks should not be considered optional, but rather a fundamen-

tal requirement in the design and dissemination of medical benchmarks.

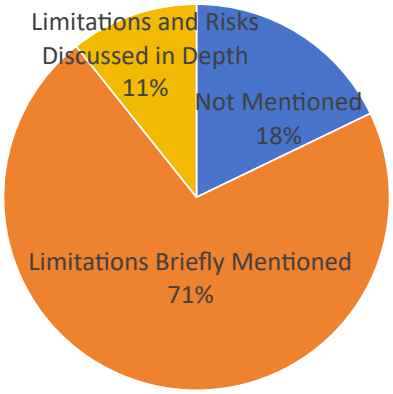

Figure 30: Performance on Discussion of Limitations and Risks

F.10 Scores per lifecycle Stage

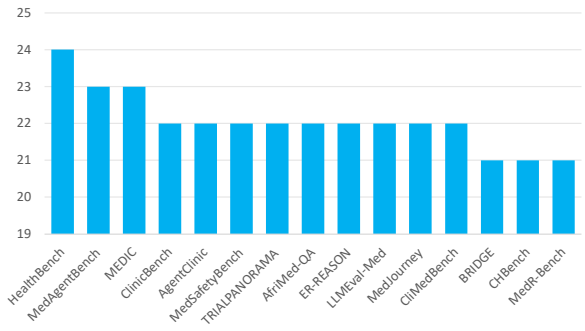

Figure 31: Top 15 Benchmarks by Score in Phase I

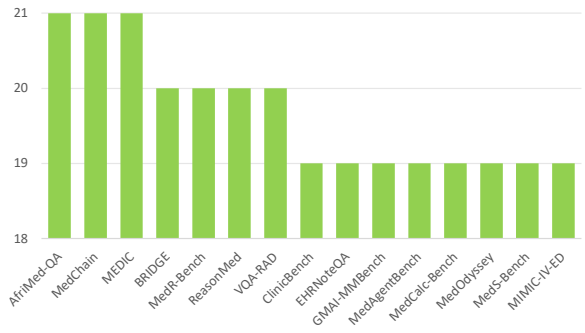

Figure 32: Top 15 Benchmarks by Score in Phase II

We present the top 15 benchmarks for each of the five lifecycle phases using bar charts(Figure31 to Figure35). The scores for each phase reflect performance in areas such as design, dataset construction, technical implementation, validation, and documentation and governance. This visualization

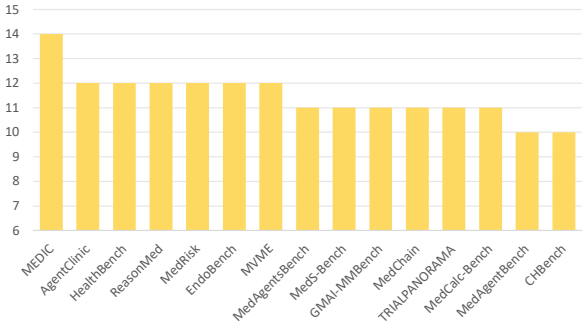

Figure 33: Top 15 Benchmarks by Score in Phase III

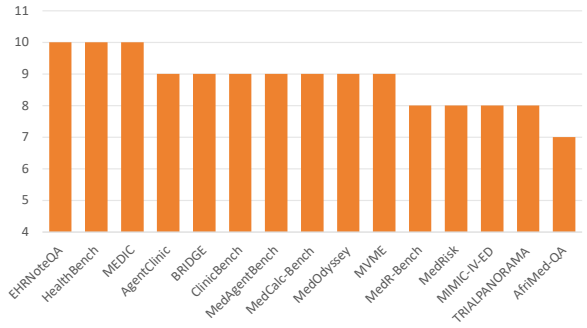

Figure 34: Top 15 Benchmarks by Score in Phase IV

allows for a clear comparison of benchmark performance across phases, highlighting strengths in certain areas and identifying potential weaknesses in others.

F.11 Validation of MedCheck's Correlation with Clinical Safety

To explore whether MedCheck scores can serve as potential indicators of clinical utility and safety, we provide theoretical justification alongside an exploratory empirical analysis analyzing the relationship between our scoring framework and clinical safety outcomes.

MedCheck functions as a strict meta-evaluation of the evidence provided by benchmarks, assessing whether a given benchmark possesses the necessary construct validity to serve as a reliable clinical tool. Rather than generating new clinical data, the framework evaluates the extent to which a benchmark has documented its own relevance.or instance, Criterion 35 (Correlation with Clinical Performance) penalizes benchmarks that fail to provide empirical data linking their metrics to real-world outcomes. Similarly, criteria such as Criterion 12 (Safety and Fairness), Criterion 28 (Robustness Evaluation), and Criterion 30 (Uncertainty Evaluation) strictly assess the presence of necessary safeguards. Con-

Table 7: Correlation Analysis between MedCheck Scores and Safety Sensitivity. Higher MedCheck scores demonstrate a positive correlation with safety outcomes, whereas lower-scoring benchmarks (e.g., PubMedQA) show a negative correlation.

Benchmark	MedCheck Score	GPT-3.5	GPT-4	GPT-o3	Pearson Correlation (r) w/ Safety Sensitivity
HealthBench	0.75	0.155	0.150	0.599	**+0.701**
CMExam	0.62	0.232	0.257	0.579	+0.647
PubMedQA	0.31	0.796	0.752	0.160	-0.647

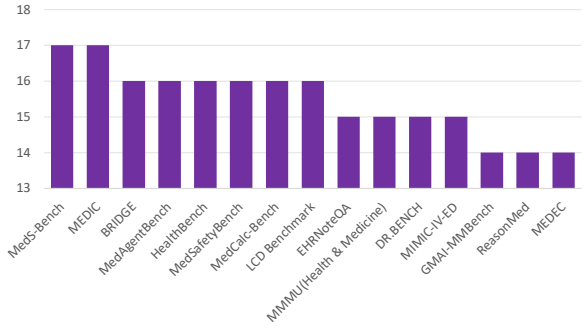

Figure 35: Top 15 Benchmarks by Score in Phase V

sequently, a high MedCheck score is not merely a theoretical accolade but a verification that a benchmark has demonstrably validated its own clinical utility and safety mechanisms.

We further conducted a correlation analysis to quantify the relationship between MedCheck scores and benchmark sensitivity to clinical safety issues. We averaged MedCheck scores across Phases I, III, and IV—which focus on safety-critical design, evaluation capabilities, and validation—and compared them against model performance on MedSafetyBench across GPT-3.5, GPT-4, and o3.

We selected three representative benchmarks with varying MedCheck scores: HealthBench (MedCheck score: 0.75), CMExam (0.62), and PubMedQA (0.31). The analysis reveals a strong positive correlation (Pearson r = 0.970, p = 0.156; Spearman ρ = 1.000, p = 0.000) between MedCheck scores and safety sensitivity. Notably, the high-scoring HealthBench demonstrates the strongest positive correlation (r = +0.701) with safety outcomes. Conversely, PubMedQA exhibits a negative correlation (r = -0.647), indicating models performing well on lower-scoring benchmarks may actually perform worse on safety-critical tasks. This empirical evidence suggests that MedCheck effectively identifies benchmarks meaningfully associated with clinical safety and utility. However, it is important to clarify the methodological role of MedSafetyBench in this analysis. We utilize MedSafetyBench strictly as an established "ground truth" anchor for clinical safety to calibrate our framework. The goal is to demonstrate that *MedCheck*'s safety-specific criteria possess the discriminative power to correctly identify high-relevance safety benchmarks, avoiding circular reasoning.

Furthermore, while this empirical evidence suggests strong alignment, we explicitly acknowledge the limitation of the small sample size ($N = 3$) in this specific correlation analysis. This analysis is intended as an exploratory proof-of-concept rather than a definitive statistical inference. Independent validation across a broader set of benchmarks and real-world clinical outcomes is required before *MedCheck* scores can be definitively treated as predictive indicators.

F.12 Validation of Construct Validity via Clinical Transferability

To further validate MedCheck's construct validity, we conducted a correlation study using the **JAMA Clinical Challenge dataset** as a clinically-proximal holdout task ($N = 100$ cases across 10 specialties). We selected three representative benchmarks with varying MedCheck scores: BRIDGE (0.83), MedQA (0.55), and COGNET-MD (0.35), and evaluated three leading models (GPT-4o, Gemini 2.0 Flash, and Qwen2.5) on these benchmarks alongside the JAMA holdout task.

Our analysis, summarized in **Table 8**, reveals two key findings supporting MedCheck's validity:

- **Ordered Correlation with Clinical Reality:** Benchmarks with higher MedCheck scores demonstrate stronger alignment with the clinical holdout task. As shown in the Pearson correlation (r) between benchmark rankings and JAMA rankings, BRIDGE exhibits the strongest correlation ($r = +0.995$), followed by MedQA ($r = +0.946$) and COGNET-MD ($r = +0.932$).

- **Predictive Power:** We observe a strong pos-

Table 8: Correlation Analysis between Benchmark Performance and Clinical Transferability (JAMA Clinical Challenge). Benchmarks with higher MedCheck scores demonstrate stronger alignment with model performance on real-world clinical cases.

Benchmark	MedCheck Score	GPT-4o	Gemini 2.0 Flash	Qwen2.5	Pearson Correlation (r) w/ JAMA Ranking
BRIDGE	0.83	0.530	0.530	0.510	**+0.995**
MedQA	0.55	0.850	0.770	0.660	+0.946
COGNET-MD	0.35	0.910	0.810	0.690	+0.932
JAMA (Holdout)	—	**0.750**	**0.730**	**0.570**	—

itive correlation between MedCheck scores and the JAMA alignment strength (Pearson $r = 0.975$, $p = 0.144$). This indicates that benchmarks scoring higher on MedCheck are significantly more effective at capturing capabilities that transfer to real-world clinical scenarios.

G Domain-Specific Analysis (Medical vs. Clinical)

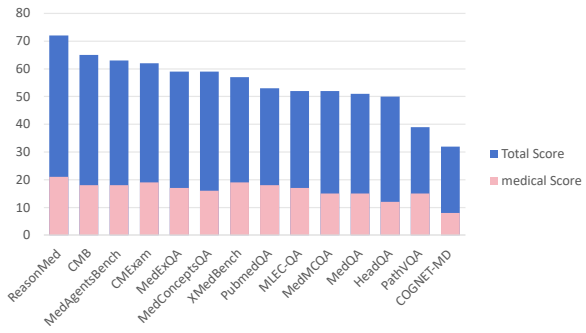

Figure 36: Top 15 Medical Benchmarks by Total Score

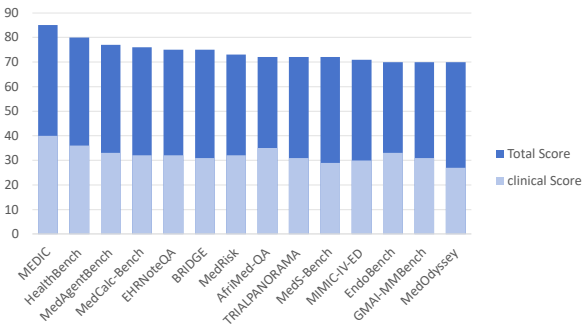

Figure 37: Top 15 Clinical Benchmarks by Total Score

To address the important distinction between benchmarks for foundational medical knowledge and those for clinical practice, we tagged each of our 46 criteria as "medical," "clinical," or "general." We then re-scored all 56 benchmarks using only the criteria relevant to their primary domain.

The results confirm the utility of our framework across domains. For example, ReasonMed, a benchmark focused on medical reasoning from texts, achieves the highest score (21) among benchmarks classified as "medical" when evaluated on medical criteria alone. Conversely, MEDIC, a benchmark designed for real-world clinical applications, leads the "clinical" benchmarks with a 40 score on clinical-specific criteria. This demonstrates *MedCheck*'s ability to appropriately evaluate benchmarks according to their intended purpose. The domain-specific scoring breakdown for the top 15 medical and clinical benchmarks is provided in Figure 36 and Figure 37.

H Full Evaluation Criteria

We implemented a comprehensive tagging system categorizing our 46 criteria as Medical (M), Clinical (C), or General (G), and provided domain-specific scoring analysis for all 56 benchmarks, demonstrating that our framework appropriately evaluates benchmarks according to their intended purpose.

H.1 Design and Conceptualization

H.1.1 Purpose and Intent

1. **Clarity of Evaluation Objectives (G)**

   - **Explanation**: Benchmark developers should clearly specifies the targeted capabilities of LLMs it aims to evaluate in the medical field (e.g., medical knowledge question answering, diagnosis, report generation).
   - **Justification**: Clearly defined evaluation objectives can avoid ambiguity, facilitating subsequent data collection, task design, and metric selection. Also, it helps users determine whether the benchmark aligns with their specific evaluation needs.

- **Scoring**:
  - 0: Does not mention or define the evaluation objective
  - 1: Mentions the evaluation objectives but the definition is broad or insufficiently detailed.
  - 2: Explicitly defines the medical capabilities being evaluated and provides examples.

2. **Clarity of Application Scenario (C&M)**

- **Explanation**: Benchmark developers should clearly describes the specific medical application scenarios it corresponds to and explains the potential value.
- **Justification**: Linking the benchmark to real-world application scenarios ensures that the results are meaningful. It facilitates the validation of model's effectiveness in specific settings, ultimately better serving users such as doctors and researchers.
- **Scoring**:
  - 0: Does not describe any application scenarios.
  - 1: Mentions application scenarios but provides vague descriptions or fails to clarify the potential value.
  - 2: Clearly describes the application scenarios and elaborates on the value in detail.

3. **Uniqueness and Novelty (G)**

- **Explanation**: By comparing with relevant benchmarks, benchmark developers should explain their contributions and the uniqueness of the benchmark, such as filling a gap and proposing new evaluation methodology.
- **Justification**: Demonstrating the uniqueness demonstrates the necessity and justification of the new benchmark. It also helps the community to better understand its unique value, promoting continual innovation in evaluation and benchmarking.
- **Scoring**:
  - 0: Does not compare with relevant benchmarks or does not mention its uniqueness.
  - 1: Briefly mentions other benchmark, but the comparison is insufficient to explain its unique value.
  - 2: Provides detailed comparison with relevant benchmarks and clearly argues its uniqueness.

H.1.2 Scope and Applicability

1. **Target Capability of Evaluation (G)**

- **Explanation**: Benchmark developers should clearly define the capability of LLMs intended to evaluate (e.g., text generation, multimodal understanding)
- **Justification**: Clearly defining the targeted capability helps clarify the scope of the benchmark, which can prevent misusing the benchmark.
- **Scoring**:
  - 0: Does not define the target LLM capability.
  - 1: Briefly mentions the target capability without clear definition.
  - 2: Clearly defines and explains the target LLM capability.

2. **Medical Domain Coverage (M)**

- **Explanation**: Benchmark developers should clearly define the scope of the medical specialties of the benchmark, such as clinical departments, disease types, or task types.
- **Justification**: By clearly defining the medical scope, it helps users better understand the breath and depth of the coverage of the benchmark. It also helps users determine whether the benchmark aligns with their needs.
- **Scoring**:
  - 0: The medical scope and coverage are not defined.
  - 1: The medical scope and coverage are only briefly mentioned.
  - 2: The medical scope and coverage are clearly defined and explained.

3. **Demonstration of User Needs (C)**

- **Explanation**: The benchmark should reflect the core concerns and assessment needs of its target users, such as addressing specific clinical challenges or overcoming technical obstacles.

- **Justification**: An effective benchmark should serve the needs of users. Evidence of user needs justifies the development of the benchmark. This boosts credibility, promotes adoption, and helps guide model improvements.
- **Scoring**:
  - 0: Does not reflect any consideration of user needs.
  - 1: Briefly mentions user needs, but lacks concrete external evidence.
  - 2: Clearly explains user needs to justify design choices and connects user needs to benchmark tasks (e.g., referencing relevant literature, user surveys, or expert interviews).

H.1.3 Medical Expertise and Professionalism

1. **Domain Experts Involvement (C&M)**

   - **Explanation**: Domain expert (e.g., physicians or clinical researchers) should be involved in the development of the benchmark.
   - **Justification**: Due to the professionalism and rigor required in the medical field, the development of a benchmark must involve deep engagement from domain experts. Their expertise are fundamental to ensuring data quality, task validity, authenticity and relevance.
   - **Scoring**:
     - 0: No mention of medical expert participation.
     - 1: Briefly mentions expert involvement, but the information is unclear and lacks details on the roles or involved tasks.
     - 2: Clearly describes the experts' qualifications and their involvement, including specific roles and tasks throughout the development process.

2. **Authoritative Knowledge Sources (C&M))**

   - **Explanation**: Benchmark developers must clearly specify which authoritative medical knowledge sources (e.g., clinical guidelines, textbooks, medical databases) the benchmark content is based on.
   - **Justification**: Benchmark content should adhere to recognized, evidence-based medical knowledge sources. By using authoritative sources, it ensures that the benchmark is scientifically reliable, enhancing credibility and transparency.
   - **Scoring**:
     - 0: No knowledge sources listed.
     - 1: Mentions knowledge sources, but the description is vague or the sources are not sufficiently authoritative.
     - 2: Clearly lists and explains the use of authoritative medical knowledge sources.

3. **Medical Standards Alignment (C&M)**

   - **Explanation**: Benchmarks should follow internationally or industry-recognized medical standards (e.g., ICD, SNOMED CT, LOINC) when medical terminology is involved.
   - **Justification**: Adherence to medical standards ensures the clinical relevance and consistency, facilitating integration and comparison in reflect real-world medical practice.
   - **Scoring**:
     - 0: Does not mention any medical standards.
     - 1: Mentions the importance of standardization, but does not specify which standards are followed, or the description is unclear.
     - 2: Clearly states the adherence to recognized medical standards.

H.1.4 Evaluation Metrics and Dimensions

1. **Validity of Core Metric (C&M)**

   - **Explanation**: The core performance metric should be clearly defined and closely related to the clinical task or medical capability being assessed.
   - **Justification**: Evaluation metrics directly shape the interpretation of results. Choosing suitable metric and explaining its relevance ensures a shared understanding and comprehensive interpretation.
   - **Scoring**:

  - 0: Core metrics are not clearly defined or have weak relevance to the task.
  - 1: Core metrics are clearly defined, but the choice and relevance of the metrics are not explained and justified.
  - 2: Core metrics are clearly defined, with justification for the relevance to the evaluation task.

2. **Multi-dimensional Evaluation (C)**
   - **Explanation**: Apart from the correctness, benchmark developers should include evaluation of other important dimensions (e.g., safety, fluency).
   - **Justification**: In high-stakes medical domain, going beyond correctness is vital. Multi-dimensional evaluation offers a more comprehensive assessment of whether a model is truly reliable and trustworthy.
   - **Scoring**:
     - 0: Only consider the correctness of answers.
     - 1: Other evaluation dimensions are briefly mentioned, without specific evaluation methods.
     - 2: Multiple evaluation dimensions are clearly designed, with well defined evaluation methods.

3. **Safety and Fairness Considerations (C)**
   - **Explanation**: Benchmark should include evaluation for potential risks (e.g., unsafe recommendation, toxicity) and bias (e.g., gender, ethnicity) in model outputs.
   - **Justification**: Evaluating risks and fairness facilitates bias-free, safe and equitable applications of LLMs in the medical domain, promoting responsible AI development and application.
   - **Scoring**:
     - 0: Does not consider safety and fairness evaluation.
     - 1: Mentions the importance of safety and fairness, but lacks specific evaluation methods.
     - 2: Designs dedicated assessment and test cases for safety and fairness.

H.2 Dataset Construction and Management

H.2.1 Data Source and Quality

1. **Data source transparency and traceability (G)**
   - **Explanation**: Benchmark developers should clearly state the data source of the benchmark, along with relevant traceability information (e.g., data collection time frame, platforms).
   - **Justification**: Clear and traceable data sources are critical to ensure transparency and ethical data usage, which is especially important when sensitive data is involved. Also, it ensures reproducibility, enhancing the credibility of the benchmark.
   - **Scoring**:
     - 0: Does not state the original data source of the benchmark.
     - 1: States the data source of the benchmark, with incomplete traceability information.
     - 2: Clearly states the data source of the benchmark and provides detailed traceability information.

2. **Data Source Reliability (C&M)**
   - **Explanation**: Benchmark developers should clearly describe the selection criteria and explain the reliability of the data source.
   - **Justification**: Collecting data from unreliable sources may lead to invalid results for medical applications. By explaining the reliability of the data source, credibility of the benchmark can be established.
   - **Scoring**:
     - 0: Does not describe the quality or selection criteria of the data source.
     - 1: Mentions the reliability of the data source, but does not explain the selection criteria.
     - 2: Clearly explains the selection criteria and describes the reliability of the data source.

3. **Data Authenticity (C)**
   - **Explanation**: Benchmark developers should clearly specify whether the data

comes from the real world scenarios, synthetically generated, or a mixture of both. For synthetically generated data, the construction process and verification for its authenticity (e.g., expert review) should also be described.

- **Justification**: Real-world data reflects authentic scenarios but is harder to obtain, whereas synthetic data can be generated but requires careful verification. Clarifying data authenticity enhance transparency and ensures clinical relevance of the data used.
- **Scoring**:
  - 0: Does not state the data source and explain its authenticity.
  - 1: States the data source, but lacks sufficient details on synthetic data generation or verification.
  - 2: Clearly states the data source, and provides a detailed description of the generation and verification process for synthetic data.

4. **Dataset Representativeness (C&M)**

- **Explanation**: The representativeness of key features (e.g., patient age, disease) of the dataset should be explained and statistically analyzed.
- **Justification**: A benchmark that lacks representativeness may lead to bias in evaluation, reducing the clinical relevance, generalizability and fairness of the results.
- **Scoring**:
  - 0: Does not mention the representativeness of the dataset.
  - 1: Provides qualitative descriptions of representativeness or partial and incomplete statistical analysis for some features.
  - 2: Provides quantitative statistical analysis of key features and compares them against the target population's distribution.

5. **Dataset Diversity (C&M)**

- **Explanation**: The benchmark should have a diverse coverage, and provide quantitative evidence of its diversity of disease or medical departments covered.
- **Justification**: Ensuring the dataset covers a variety helps comprehensively evaluate the model's generalization ability, reducing bias in the evaluation results.
- **Scoring**:
  - 0: Does not clearly define or explain the disease and medical department coverage.
  - 1: Clearly states the disease or medical department coverage, but coverage (defined in Equation 1) is less than average (21.8%).
  - 2: Clearly states the disease or medical department coverage with a coverage (defined in Equation 1) above average (21.8%).

H.2.2 Data Processing and Privacy Protection

1. **Data Cleaning and Standardization (C)**

- **Explanation**: The processes and steps of data preprocessing, including data cleaning and standardization, should be clearly described.
- **Justification**: It ensures that the final dataset is well-structured, enhancing reliability. Also, it allows a better understanding of the construction process of the benchmark, ensuring transparency.
- **Scoring**:
  - 0: Does not describe any data cleaning or standardization steps.
  - 1: The data preprocessing procedure is only briefly or partially mentioned.
  - 2: The data preprocessing procedure is clearly described, with details that cover all steps and aspects of the process.

2. **Privacy Protection (C&M)**

- **Explanation**: If sensitive data is used, it should be de-identified. Methods of de-identification should be described and compliance with relevant regulations (e.g., HIPAA) should be clearly stated.
- **Justification**: Real-world clinical data may contain patient information. It is essential to ensure that the data and privacy protection aligns with ethical and legal standards, reducing the risk of privacy breaches and ensuring compliant data usage.

- **Scoring**:
  - 0: No privacy protection measures are mentioned.
  - 1: Privacy protection measures are mentioned, but the description of methods or regulations is unclear.
  - 2: Sensitive data is not used. Otherwise, privacy protection measures are clearly described and compliance with relevant regulations is stated.

3. **Data Format Clarity and Consistency (G)**
   - **Explanation**: The data in the dataset, including questions, cases, or task descriptions, should be written clearly and presented in a consistent format.
   - **Justification**: A clear and consistent format is essential for standardized evaluation. Inconsistent format may affect models' interpretation, compromising the accuracy of the evaluation.
   - **Scoring**:
     - 0: Data format is disorganized and contains ambiguities.
     - 1: Minor inconsistencies or ambiguities exist in data format.
     - 2: Data format is clear and consistent.

4. **Data Review and Audit (C&M)**
   - **Explanation**: The dataset construction process should include a review procedure with involvement of medical experts.
   - **Justification**: A dataset construction process without review mechanism is prone to errors. Careful review that involves medical experts can ensures clinical relevance, professionalism and reliable of the data.
   - **Scoring**:
     - 0: No review procedure is described.
     - 1: Briefly mentions the review procedure, or the review procedure is simple and without experts involvement.
     - 2: Provides a clear and detailed description of the review process, including the involvement of medical experts.

5. **Quality of Reference Answer (G)**
   - **Explanation**: The benchmark should provide clear and accurate reference answers or scoring guidelines, and explain the construction and verification process (e.g., expert consensus).
   - **Justification**: Clear reference answers or scoring guidelines ensures transparent and accurate evaluation. By explaining how are reference answers or scoring guidelines formulated and verified, it also enhances the credibility of evaluation results.
   - **Scoring**:
     - 0: No reference answers or scoring guidelines are provided.
     - 1: Reference answers or scoring guidelines are provided, but the formulation or verification process is not explained.
     - 2: Provides clear reference answers or scoring guidelines, and explains its formulation and verification process in details.

6. **Data Contamination Prevention (G)**
   - **Explanation**: Benchmark developers should take actions to identify, address and prevent potential data contamination issues of the data.
   - **Justification**: Data contamination may lead to inflated performance, which only reflect memorization instead of medical capability from the models. Preventing and controlling potential contamination ensures that the benchmark is effective, enhancing credibility and validity.
   - **Scoring**:
     - 0: Does not mention or consider potential data contamination.
     - 1: Mentions the risks of data contamination but lacks concrete, actionable steps to handle it.
     - 2: Clearly describes procedures taken to prevent or detect contamination using feasible, developer-controlled practices (e.g., encrypting test sets, distributing data via a gated API, using Canary GUIDs, or conducting n-gram overlap checks

against open pre-training corpora). We evaluate based on actionable preventive measures rather than an impossible guarantee of zero contamination, ensuring fair assessment even when closed-source models are involved.

H.3 Technical Implementation and Evaluation Methodology

1. **User-friendliness of evaluation tools (G)**
    - **Explanation**: Evaluation scripts or tools that is easy to obtain and use should be provided.
    - **Justification**: It ensures that users can use the benchmark conveniently, promoting benchmark adoption and ensuring fair, transparent, and consistent evaluation.
    - **Scoring**:
        - 0: No evaluation scripts or tools is provided.
        - 1: Provides evaluation scripts, but requires users to perform complex environment configurations or manual settings.
        - 2: Provides evaluation scripts or tools with detailed instructions for streamlined execution, that are user-friendly and require minimal user effort.
2. **Technical Reproducibility (G)**
    - **Explanation**: Tools and technical documentation (e.g., environment configuration, dependency versions) should be provided.
    - **Justification**: The availability of evaluation tools and clear technical documentation ensures reproducibility of reported results, thereby enhancing the credibility of the benchmark.
    - **Scoring**:
        - 0: No information for technical reproducibility has been provided.
        - 1: Information for technical reproducibility is partially provided, but insufficient to guarantee reproducibility.
        - 2: Complete information for reproducibility is provided with detailed steps to replicate the results.
3. **Provision of performance baselines (G)**
    - **Explanation**: Benchmark developers should provide multiple meaningful performance baselines, such as random, baseline models and human performance.
    - **Justification**: Providing different performance baselines allows comparison against the model's performance, enabling a deeper understanding and better interpretability.
    - **Scoring**:
        - 0: No performance baselines are provided.
        - 1: Only one type of performance baseline is provided, or performance baselines are mentioned without specific data.
        - 2: At least two meaningful types of performance baselines are provided, with clear explanation of the measurement methods.
4. **Reasoning Process Evaluation (C)**
    - **Explanation**: Apart from the final answer, the benchmark includes evaluations for the model's reasoning process or explanation abilities.
    - **Justification**: In medical domain, understanding the model's decision-making process is just as important as the final answer. Evaluating the reasoning process helps ensure that the model's decisions are trustworthy and logically sound.
    - **Scoring**:
        - 0: Does not consider reasoning process evaluation.
        - 1: Mentions the importance of evaluating the reasoning process, but lacks specific evaluation methods.
        - 2: Designs assessment for reasoning process with clear evaluation methods and metrics.
5. **Robustness Evaluation (C)**
    - **Explanation**: Benchmark should include testing for the model's stability and

robustness (e.g., input perturbations, adversarial samples)

- **Justification**: In practical applications, models may encounter different variations of inputs. Robustness testing ensures model's output is consistent and reliable under different conditions.
- **Scoring**:
  - 0: Does not consider robustness evaluation.
  - 1: Mentions the importance of evaluating the robustness of models, but lacks specific evaluation
  - 2: Designs assessment for robustness with clear evaluation methods and metrics.

6. **Generalization Capability Evaluation (C)**

   - **Explanation**: The benchmark design should help evaluate the generalization capability of models to unseen data or scenarios (e.g., careful train/test split, out-of-distribution testing).
   - **Justification**: Due to the high variability in medical scenarios, assessing a model's generalization capability is essential for determining its reliability across different real-world scenarios, ensuring that the performance is not caused by overfitting.
   - **Scoring**:
     - 0: Does not consider generalization capability evaluation.
     - 1: Mentions the importance of evaluating the generalization capability of models, but lacks specific evaluation
     - 2: Designs mechanism for assessing the generalization capability.

7. **Uncertainty Evaluation (C)**

   - **Explanation**: Benchmark should include evaluation for the model's ability to recognize and express its own uncertainty (e.g., responding "I don't know")
   - **Justification**: Incorrect overconfident answers can be dangerous in high-stakes medical applications. A model's ability to accurately identify and express uncertainty is critical for preventing incorrect and misleading decisions, enhancing system safety.
   - **Scoring**:
     - 0: Does not consider uncertainty evaluation.
     - 1: Mentions the importance of evaluating the uncertainty of models, but lacks specific evaluation
     - 2: Designs assessment for uncertainty with clear evaluation methods and metrics.

8. **Evaluation Flexibility (G)**

   - **Explanation**: Evaluation code should have a modular interface and support different models. Limited flexibility (e.g., hardcoded model paths, strict dependency on a single framework) hinders usability, whereas a flexible design (e.g., providing an extensible base class or API wrapper) supports both closed-source APIs and local open-source models.
   - **Justification**: It ensures that different types of models can be tested under the same interface.
   - **Scoring**:
     - 0: Supports only one specific model or framework, with hardcoded logic requiring significant code rewriting to adapt to new models.
     - 1: Supports only one type of model, but provides clear architectural guidelines or interfaces for users to implement their own extensions.
     - 2: Natively supports both closed-source APIs and local open-source models through a modular, unified, and user-friendly interface.

H.4 Benchmark Validity and Performance Verification

1. **Knowledge and Skill Coverage (M)**

   - **Explanation**: Evidence (e.g., expert evaluation) is provided to demonstrate that the benchmark content sufficiently covers the medical knowledge and skills it claims to assess.
   - **Justification**: Sufficient coverage of the core medical competencies the benchmark aims to measure is the prerequisite for establishing content validity. When

the content validity is supported with evidence, it enhances the rigor and trustworthiness of the evaluation.

- **Scoring**:
  - 0: No evidence regarding coverage of knowledge and skills is provided.
  - 1: Provide brief content analysis or expert evaluation to demonstrate the coverage of knowledge and skills.
  - 2: Provides both comprehensive content analysis and expert evaluation to demonstrate the coverage of knowledge and skills.

2. **Scenario Authenticity (C)**

- **Explanation**: The evaluation tasks of the benchmark should effectively simulate and mirror real-world medical scenario.
- **Justification**: Ensuring that the evaluation task closely mirrors the targeted clinical practice in real-world scenarios enhance the relevance of the benchmark. It helps assess whether the model can be transferred to real-world applications.
- **Scoring**:
  - 0: The evaluation task have weak relevance to real clinical scenarios (e.g., abstract knowledge question answering)
  - 1: The evaluation task reflects a single clinical decision point, but does not constitute a complete workflow.
  - 2: The evaluation consists of multiple tasks that closely simulate realistic clinical workflows, with validation by domain experts.

3. **Model Discrimination Ability (G)**

- **Explanation**: Benchmark developers should provide experiment data and analysis, indicating that the benchmark can effectively distinguish the capability between different models.
- **Justification**: An effective benchmark should be capable of differentiating and distinguishing models of varying capabilities. It provides meaningful insights into model strengths and weaknesses, guiding future improvements and development.
- **Scoring**:
  - 0: No evidence is provided regarding the discrimination ability of the benchmark.
  - 1: The benchmark has been tested on some models and score differences are reported, but without statistical validation.
  - 2: The benchmark has been tested on several model, showing significant score differences with statistical validation.

4. **Correlation with Clinical Performance (C)**

- **Explanation**: Benchmark developers should provide evidence that preliminarily explores the correlation between benchmark performance and the model's performance in actual clinical applications.
- **Justification**: Validating whether benchmark results have meaningful indication on the model's performance in real-world clinical scenarios ensures the external validity and practical value of the benchmark and results.
- **Scoring**:
  - 0: Does not consider or mention the correlation between benchmark performance and clinical performance.
  - 1: Provides discussion about the correlation between benchmark performance and clinical performance theoretically without experiment data or concrete evidence.
  - 2: Preliminary research or evidence (e.g., small-scale clinical validation, consistency analysis between models and domain experts) is provided to explore and discuss the correlation between benchmark performance and clinical performance.

5. **Internal Consistency (G)**

- **Explanation**: For different sections or items of the benchmark evaluating the same capability, benchmark developers should demonstrate good internal consistency (e.g., Cronbach's α)
- **Justification**: It ensures that different sections or items reliably assess the same

capability, enabling meaningful comparisons and interpretation of results.

- **Scoring**:
  - 0: Does not mention or conduct any internal consistency measurement.
  - 1: Mentions the importance of internal consistency, without concrete measurement.
  - 2: Conducts internal consistency measurement and demonstrates good consistency.

6. **Statistical Significance Reporting (G)**

- **Explanation**: When comparing the performance of different models, statistical significance (e.g., confidence intervals, and p-values) should be report.
- **Justification**: Statistical significance testing allows for a more informed interpretation of results by differentiating the performance differences between models that stem from actual capabilities, and those arises from randomness and noise.
- **Scoring**:
  - 0: No statistical significance is reported or considered.
  - 1: Reports simple statistics (e.g., mean and standard deviation of multiple executions) but does not conduct statistical testing.
  - 2: Reports and uses statistical significance tests (e.g., confidence interval).

H.5 Documentation, Openness and Governance

H.5.1 Documentation and Transparency

1. **Documentation Completeness (G)**

- **Explanation**: Benchmark developers should provide a clear and comprehensive documentation of the benchmark, systematically describing the relevant details, including the design, objectives, scope, construction process, task definitions and evaluation procedures.
- **Justification**: A complete and clear documentation help users understand and use the benchmark properly, enhancing the usability, reproducibility and transparency.
- **Scoring**:
  - 0: The documentation is missing or overly simple.
  - 1: The documentation exists, but some important details are missing or unclearly described.
  - 2: The documentation is comprehensive and clear, covering all key components.

2. **Clarity of Evaluation Guidelines (G)**

- **Explanation**: Benchmark developers should provide evaluation guidelines, with definitions of evaluation metrics, detailed scoring criteria, and easy-to-follow usage instructions for users to replicate the evaluation process.
- **Justification**: Clear evaluation guidelines help users better understand how model performance is quantified, ensuring a shared interpretation and understanding. It also enhances the consistency by documenting how to replicate the evaluation process.
- **Scoring**:
  - 0: No evaluation guidelines are provided.
  - 1: Provides evaluation guidelines, but some areas are only briefly described.
  - 2: Provides a clear and comprehensive evaluation guidelines.

3. **Discussion of Limitations and Risks (C)**

- **Explanation**: The benchmark developers should openly discuss the limitations and potential social risks of the benchmark.
- **Justification**: Disclosing limitations and risks demonstrates scientific rigor and responsibility. This transparency helps prevent users from misunderstanding and misusing the benchmark, especially in high-stakes contexts.
- **Scoring**:
  - 0: Discussion of limitations and risks is not provided.
  - 1: Limitations and risks are briefly mentioned without in-depth discussion.

  - 2: In-depth and comprehensive discussion of limitations and risks is provided.

4. **Peer Review (G)**

   - **Explanation**: The benchmark and its corresponding paper was accepted at peer-reviewed venue.
   - **Justification**: Going through peer review process means that the design, validity and results of a benchmark has been rigorous evaluated. It enhances credibility and ensures quality.
   - **Scoring**:
     - 0: The benchmark has not been accepted at a peer-reviewed venue.
     - 1: The benchmark is under review or published on platform without strict peer review (e.g., arXiv preprints).
     - 2: The benchmark has been published at a peer-reviewed venue.

H.5.2 Openness and Accessibility

1. **Accessibility of Evaluation Code and Data (G)**

   - **Explanation**: Access to the evaluation code and data, which can be shared within legal and ethical boundaries, is provided (e.g. on platforms like GitHub or Hugging Face) along with the applicable license.
   - **Justification**: Accessible code and data are prerequisites for reproducibility. Moreover, it allows the community to review, improve, and expand the benchmarks.
   - **Scoring**:
     - 0: Access to the evaluation code and data is not provided.
     - 1: The evaluation code and data are partially accessible, or accessible without clear licensing.
     - 2: The evaluation code and data are fully accessible and the applicable license is clearly specified.

2. **Usage and Citation Guidelines (G)**

   - **Explanation**: Clear guidelines are provided to standardize benchmark usage, result reporting, and correct citation formats in academic papers or technical reports.
   - **Justification**: Proper usage and citation guidelines help maintain academic integrity. It also promote standardized reporting of results, facilitating comparisons in further research.
   - **Scoring**:
     - 0: No usage or citation guidelines are provided.
     - 1: Partial guidelines are provided but are incomplete.
     - 2: Clear and complete usage and citation guidelines are provided.

H.5.3 Continuous Maintenance and Governance

1. **Update and Version Management (G)**

   - **Explanation**: A clear plan or mechanism should be in place to regularly (or when necessary) update the contents of the benchmark (e.g., incorporating new data) and to effectively manage and archive different versions.
   - **Justification**: Medical knowledge is constantly evolving, and a static Benchmark may become outdated. Establishing a mechanism for updates and version control is essential to ensure its long-term relevance and alignment with current developments.
   - **Scoring**:
     - 0: No update plans or version management.
     - 1: Intention to update is mentioned, without any concrete actions.
     - 2: There is a clear plan for update and version management.

2. **Feedback Channel for Users (G)**

   - **Explanation**: A feedback channel should be maintained for users to report problems, provide feedback and suggestions.
   - **Justification**: Maintaining an effective feedback channel allows users to provide feedback when issues with the benchmark are discovered. This is crucial for continuously improving benchmark quality and fixing potential bugs and issues.

- **Scoring**:
  - 0: No public feedback channel is established.
  - 1: A public feedback channel is set up, but responses are either delayed or absent.
  - 2: A public feedback channel is set up with promptly response to user feedback.

3. **Long-term Maintenance Responsibility (G)**
   - **Explanation**: The individuals, teams, or institutions responsible for its long-term maintenance and development should be clearly stated.
   - **Justification**: Clarifying who holds long-term responsibility reassures the community that it will be actively supported and improved, ensuring usability and credibility.
   - **Scoring**:
     - 0: No long-term maintenance responsibility is mentioned.
     - 1: The responsible party is implied, but not explicitly mentioned.
     - 2: Details of the person responsible for long-term maintenance are clearly stated.

I Case Study: Scoring and Explanations for a Representative Benchmark

To demonstrate the application of the *MedCheck* framework, we provide the detailed scoring and explanations for a representative benchmark. This case study focuses on the objective verification of factual evidence within the benchmark's documentation across all 46 criteria.

1. **Clarity of Evaluation Objectives**
   - **Score**: 2
   - **Explanation**: It clearly defines the objective of evaluating LLMs on medical knowledge question-answering and provides clear examples.

2. **Clarity of Application Scenario**
   - **Score**: 0
   - **Explanation**: It does not describe any specific application scenarios.

3. **Uniqueness and Novelty**
   - **Score**: 2
   - **Explanation**: It states that it was the first free-form multiple-choice OpenQA dataset for solving medical problems collected from the professional medical board exams, with comparisons against existing medical QA datasets.

4. **Target Capability of Evaluation**
   - **Score**: 2
   - **Explanation**: It clearly explains that the targeted LLM capability is knowledge understanding and require multi-hop logical reasoning.

5. **Medical Domain Coverage**
   - **Score**: 0
   - **Explanation**: The medical scope and coverage of specialties or disease types are not defined.

6. **Demonstration of User Needs**
   - **Score**: 1
   - **Explanation**: It briefly mentions the NLP community's need for OpenQA benchmarks.

7. **Domain Experts Involvement**
   - **Score**: 2
   - **Explanation**: It clearly describes that there are 2 medical experts with the MD degree involved to ensure the reference books in the data covered sufficient knowledge to answer the questions.

8. **Authoritative Knowledge Sources**
   - **Score**: 2
   - **Explanation**: The benchmark is based on authoritative medical licensing exams and relevant references books.

9. **Medical Standards Alignment**
   - **Score**: 0
   - **Explanation**: It does not mention any recognized medical standards.

10. **Validity of Core Metric**
    - **Score**: 2
    - **Explanation**: Accuracy is used as the core metric which is suitable for question-answering task.

11. **Multi-dimensional Evaluation**
    - **Score**: 0
    - **Explanation**: It only evaluates the correctness of answers.
12. **Safety and Fairness Considerations**
    - **Score**: 0
    - **Explanation**: It does not consider safety and fairness evaluation.
13. **Data source transparency and traceability**
    - **Score**: 2
    - **Explanation**: It clearly states the data sources are medical licensing exams and provides relevant URLs of the data sources.
14. **Data Source Reliability**
    - **Score**: 2
    - **Explanation**: Data is based on authoritative offical medical licensing exams.
15. **Data Authenticity**
    - **Score**: 2
    - **Explanation**: Data sources are offical medical licensing exams and reference books
16. **Dataset Representativeness**
    - **Score**: 2
    - **Explanation**: It provides quantitative statistical analysis of features including the distribution of question length and types.
17. **Dataset Diversity**
    - **Score**: 0
    - **Explanation**: The medical scope and coverage of specialties or disease types are not clearly described.
18. **Data Cleaning and Standardization**
    - **Score**: 1
    - **Explanation**: It breifly mentions the preprocessing and standardization procedures, but the description lacks detail.
19. **Privacy Protection**
    - **Score**: 2
    - **Explanation**: It uses publicly available data containing no sensitive patient information.
20. **Data Format Clarity and Consistency**
    - **Score**: 2
    - **Explanation**: The data format is clear and consistent.
21. **Data Review and Audit**
    - **Score**: 2
    - **Explanation**: There are 2 medical experts to ensure the reference books in the data covered sufficient knowledge to answer the questions.
22. **Quality of Reference Answer**
    - **Score**: 2
    - **Explanation**: The reference answers are based on authoritative medical licensing exams.
23. **Data Contamination Prevention**
    - **Score**: 0
    - **Explanation**: It does not mention or consider potential data contamination issues.
24. **User-friendliness of evaluation tools**
    - **Score**: 2
    - **Explanation**: It provides open-source code and evaluation tools in GitHub.
25. **Technical Reproducibility**
    - **Score**: 2
    - **Explanation**: It provides detailed technical documentation and reproduction guides.
26. **Provision of performance baselines**
    - **Score**: 2
    - **Explanation**: It provides at least two meaningful performance baselines (random and baseline models) with clear explanations.
27. **Reasoning Process Evaluation**
    - **Score**: 0
    - **Explanation**: It does not consider evaluating the reasoning process.

28. **Robustness Evaluation**
    - **Score**: 0
    - **Explanation**: It does not design specific assessments for robustness.
29. **Generalization Capability Evaluation**
    - **Score**: 1
    - **Explanation**: The importance of evaluating the generalization capability is mentioned and the cross-lingual evaluation demonstrates a certain level of generalization capability.
30. **Uncertainty Evaluation**
    - **Score**: 0
    - **Explanation**: It does not design specific assessments for the model's uncertainty handling.
31. **Evaluation Flexibility**
    - **Score**: 1
    - **Explanation**: It supports multiple evaluation methods but offers limited flexibility, requiring users to perform extensions themselves, though guidelines are provided.
32. **Knowledge and Skill Coverage**
    - **Score**: 1
    - **Explanation**: There are 61097 questions from medical licensing exams in different countries covering questions of different length and types.
33. **Scenario Authenticity**
    - **Score**: 0
    - **Explanation**: The evaluation task is abstract knowledge question-answering, which has weak relevance to realistic clinical workflows or decision-making scenarios.
34. **Model Discrimination Ability**
    - **Score**: 1
    - **Explanation**: The benchmark has been tested on several models and score differences are reported, but without statistical validation.
35. **Correlation with Clinical Performance**
    - **Score**: 0
    - **Explanation**: It does not mention or provide evidence exploring the correlation between benchmark performance and actual clinical application performance.
36. **Internal Consistency**
    - **Score**: 0
    - **Explanation**: It does not mention or conduct any internal consistency measurement.
37. **Statistical Significance Reporting**
    - **Score**: 0
    - **Explanation**: It only reports results from single runs.
38. **Documentation Completeness**
    - **Score**: 1
    - **Explanation**: Documentations are available but some details are missing including the data cleaning process.
39. **Clarity of Evaluation Guidelines**
    - **Score**: 1
    - **Explanation**: Evaluation guidelines are provided, but there could be more detailed instructions.
40. **Discussion of Limitations and Risks**
    - **Score**: 0
    - **Explanation**: It does not discuss any limitations or potential social risks of the benchmark.
41. **Peer Review**
    - **Score**: 2
    - **Explanation**: The paper was published in Applied Sciences.
42. **Accessibility of Evaluation Code and Data**
    - **Score**: 2
    - **Explanation**: The evaluation code and data are accessible on GitHub, and licensing is specified.
43. **Usage and Citation Guidelines**
    - **Score**: 1

- **Explanation**: Evaluation guidelines are provided, but there could be more detailed instructions.

44. **Update and Version Management**

    - **Score**: 1
    - **Explanation**: The benchmark is managed in GitHub but there are no clear update plans.

45. **Feedback Channel for Users**

    - **Score**: 2
    - **Explanation**: A public feedback channel is available via GitHub Issue with responses from the authors.

46. **Long-term Maintenance Responsibility**

    - **Score**: 1
    - **Explanation**: Team information is available, but the explicit, long-term maintenance responsibility is not clearly stated.

J Actionable Diagnostic Report Example

This section provides a complete actionable diagnostic report based on our lifecycle assessment of a high-impact clinical-oriented benchmark. By focusing on criteria where the benchmark scored 0 or 1, this report illustrates how *MedCheck* surfaces actionable technical and procedural recommendations.

Phase I: Design and Conceptualization

- **Criterion 9 (Medical Standards Alignment) - Score: 1**
  Weakness: Mentions standardization but lacks explicit mapping to standard ontologies.
  Actionable Recommendation: Require model outputs to strictly map to standardized terminologies such as SNOMED CT or LOINC codes.

- **Criterion 12 (Safety and Fairness Considerations) - Score: 1**
  Weakness: Conceptual discussion of safety without concrete empirical test cases.
  Actionable Recommendation: Introduce a dedicated "clinical red-teaming" subset to test for harmful hallucinations or demographic biases.

Phase II: Dataset Construction and Management

- **Criterion 16 (Dataset Representativeness) - Score: 1**
  Weakness: Qualitative demographic description lacks rigorous statistical analysis.
  Actionable Recommendation: Publish detailed statistical tables comparing dataset demographics to real-world clinical population distributions.

- **Criterion 23 (Data Contamination Prevention) - Score: 0**
  Weakness: High risk of data memorization due to the use of public clinical databases.
  Actionable Recommendation: Conduct n-gram overlap analysis against common pretraining corpora and inject unique "canary strings" into the test set.

Phase III: Technical Implementation and Evaluation Methodology

- **Criterion 27 (Reasoning Process Evaluation) - Score: 0**
  Weakness: "Black box" evaluation that cannot verify clinical logic.
  Actionable Recommendation: Implement Chain-of-Thought (CoT) metrics or expert-defined reasoning path verification.

- **Criterion 28 (Robustness Evaluation) - Score: 0**
  Weakness: Overestimates performance by assuming noise-free clinical inputs.
  Actionable Recommendation: Introduce programmatic input perturbations (e.g., medical abbreviations, simulated typos) to test resilience.

- **Criterion 30 (Uncertainty Evaluation) - Score: 0**
  Weakness: Rewards overconfident hallucinations over safe abstention.
  Actionable Recommendation: Include "unanswerable" EHR cases and reward the model for correctly outputting "Insufficient data to decide."

Phase IV: Benchmark Validity and Performance Verification

- **Criterion 35 (Correlation with Clinical Performance) - Score: 1**
  Weakness: Reliance on NLP metrics (e.g.,

ROUGE) that may misalign with clinical utility.
Actionable Recommendation: Conduct a clinician-in-the-loop study comparing automated scores with physician preference ratings.

- **Criterion 37 (Statistical Significance Reporting) - Score: 1**
Weakness: Reports point estimates without confidence intervals.
Actionable Recommendation: Use bootstrapping to report p-values and confidence intervals for all model comparisons.

Phase V: Documentation, Openness, and Governance

- **Criterion 40 (Discussion of Limitations and Risks) - Score: 1**
Weakness: Insufficient risk communication for actual clinical deployment.
Actionable Recommendation: Add a dedicated "Broader Impacts" section analyzing clinical deployment risks.

- **Criterion 46 (Long-term Maintenance Responsibility) - Score: 1**
Weakness: High risk of becoming "abandonware" due to lack of institutional commitment.
Actionable Recommendation: Explicitly state the responsible maintaining body and outline a 3-year sustainability plan.