Innovation and Technology Studies
Print ISSN: 3007-6919
Online ISSN: 3007-6927
DOI: https://doi.org/10.61784/its4026

BEYOND MODEL SCALE: A TASK-AWARE RELIABILITY BENCHMARK FOR LARGE LANGUAGE MODELS IN CLINICAL DECISION SUPPORT

JiTong Zou[1], YuXuan Qin[2*], MinJae Rhee[3]
[1]*Case Western Reserve University, Cleveland, Ohio, United States.*
[2]*Northeastern University, Boston, MA 02115, United States.*
[3]*University of Illinois Urbana-Champaign, Champaign, IL, United States.*
[*]*Corresponding Author: YuXuan Qin*

Abstract: Model size is often treated as a proxy for trustworthiness when large language models (LLMs) are considered for clinical decision support, yet a single aggregate score can hide task-specific failure modes. This paper introduces TARB-CDS, a task-aware reliability audit that transforms a model-by-task accuracy matrix into deployment-relevant reliability profiles. The framework combines macro accuracy with lower-tail conditional value at risk over the two weakest tasks (CVaR-2), worst-task accuracy, Pareto dominance, policy-weighted scores, contextual rank stability under uncertain task mixtures, output-format reliability, and matched base-versus-instruction-tuned comparisons. We conduct a reproducible secondary analysis of 12 LLM checkpoints evaluated on seven Med-HALT reasoning and biomedical-memory tasks. The public reasoning files contain 39,590 rows across False Confidence, Fake Question, and None-of-the-Above tests; published model accuracies and format-error rates are analyzed without claiming fresh model inference. Falcon-40B has the highest seven-task macro accuracy (42.67%) and CVaR-2 (9.36%), but the best worst-task accuracy among all models is only 1.38%. Under 100,000 uniformly sampled task mixtures, Falcon-40B ranks first in 77.60% of contexts, leaving substantial rank instability. Among open checkpoints, parameter count has only a modest descriptive association with macro accuracy (Spearman ρ=0.366), while all five matched instruction/chat-tuning pairs reduce macro accuracy by an average of 9.30 percentage points. These results show that scale alone is insufficient for reliability claims and motivate task-conditional model cards and minimum-performance gates before clinical evaluation.
Keywords: Clinical decision support; Large language models; Reliability benchmarking; Hallucination; Task-aware evaluation; Model scale

1 INTRODUCTION

Large language models have demonstrated substantial biomedical knowledge and examination performance [1-8]. These results have accelerated interest in using LLMs for information retrieval, documentation support, triage assistance, and other components of clinical decision support (CDS). However, an examination score is not a safety case. Clinical workflows contain heterogeneous tasks with different error costs, and confident generation may fail through unsupported reasoning, fabricated evidence, or malformed outputs. Reliability therefore requires more than asking which checkpoint is largest or which model has the highest mean accuracy.

Benchmark practice has begun to move beyond one-dimensional leaderboards. MMLU, HELM, CheckList, MedEval, and HaluEval emphasize broad coverage, behavioral tests, or multiple evaluation dimensions [9-13]. In medicine, uncertainty and calibration are especially important because a useful system must know when to defer [14-20]. Yet public medical LLM comparisons still commonly report task-wise accuracy and then summarize the result narratively. Such reporting makes three clinically relevant questions difficult to answer: (1) whether a model has a catastrophic weak task, (2) whether rankings change when task prevalence changes, and (3) whether parameter scaling or alignment tuning improves reliability consistently rather than selectively.

We address this gap through TARB-CDS, a Task-Aware Reliability Benchmark for Clinical Decision Support. TARB-CDS is not a new patient-outcome dataset and does not claim to validate autonomous medical use. It is a transparent pre-deployment audit layer that can be applied to any model-by-task performance matrix. We instantiate it with Med-HALT, which provides three reasoning hallucination tests and four PubMed memory-retrieval tests across 12 commercial and open LLM checkpoints [21]. This choice allows the scale hypothesis to be tested within Llama-2 families (7B, 13B, and 70B), while also comparing base and instruction/chat variants.

The contributions are fourfold:

• We formalize a reliability profile that retains every task score and augments the mean with worst-task accuracy, CVaR-2, cross-task dispersion, and Pareto dominance.

• We introduce contextual rank stability: the probability that a model is top-ranked when deployment task weights are uncertain, estimated from 100,000 reproducible Dirichlet task mixtures.

• We separate model scale from alignment by descriptive scale correlations and five matched base-versus-tuned comparisons, exposing task-specific trade-offs that a single average conceals.

• We release executable analysis code, raw public reasoning files, transcribed source tables with provenance, all derived results, figures, and a reference-verification register.

2 RELATED WORK

2.1 Medical LLM Evaluation

Medical question-answering datasets such as PubMedQA, HEAD-QA, MedQA, and MedMCQA made it possible to measure biomedical knowledge at scale [2-5]. Later studies reported strong performance of general and medically adapted LLMs [1, 6-8]. These benchmarks are valuable but often emphasize central performance on a fixed mixture of questions. Med-HALT instead targets hallucination behavior through a False Confidence Test (FCT), a Fake Questions Test (FQT), a None-of-the-Above test (NOTA), and four PubMed identifier/title/link retrieval mappings [21]. Our work does not replace Med-HALT; it reuses its public data and published aggregate measurements to study reliability aggregation and ranking stability.

2.2 Reliability Beyond Accuracy

Calibration research shows that correctness and confidence are distinct properties, while selective classification formalizes the option to abstain when uncertainty is high [14-18]. In medical prediction, uncertainty quantification and calibration are foundational to safe use [19-20]. Behavioral and holistic evaluation frameworks likewise show that averages can conceal capability gaps [12-13, 22]. TARB-CDS complements these lines by focusing on the lower tail across tasks when only aggregate accuracies are available. The CVaR-style statistic is used as a benchmark-composition risk measure, not as a probabilistic confidence interval or a patient-level risk estimate.

2.3 Scale, Prompting, and Alignment

Scaling and prompting can elicit broad few-shot and reasoning capabilities [23-25]. Instruction tuning and preference optimization improve adherence and zero-shot generalization in many settings [26-28]. Nevertheless, these interventions can alter task behavior asymmetrically, particularly when exact retrieval, abstention, and constrained formats are required. Llama 2 provides matched base and chat variants across multiple scales, enabling a limited but informative audit of whether larger or more aligned models are uniformly more reliable [29].

3 TARB-CDS FRAMEWORK

3.1 Task-Performance Profile

Let A be an M-by-K matrix whose entries are task accuracies in [0,100]. Model m is represented by the profile a(m)=[a(m,1), ... , a(m,K)]. Retaining a(m) is essential: a scalar score is a projection whose meaning depends on an implicit task mixture. TARB-CDS reports the macro mean mu(m), cross-task standard deviation sigma(m), weakest-task accuracy q(m), and lower-tail CVaR-2 c(m), defined as the mean of the two smallest task accuracies.

$$\mu(m) = \frac{\sum_k a(m,k)}{K};\ q(m) = \min_k a(m,k);\ c(m) = \frac{a(m,(1))+a(m,(2))}{2} \quad (1)$$

Here a(m,(1)) and a(m,(2)) are the two lower order statistics. CVaR-2 is deliberately simple and auditable. In a seven-task benchmark, it penalizes models that achieve a strong average by excelling on a subset while approaching zero on others. We also identify Pareto-efficient models: model i is dominated when another model is at least as accurate on every task and strictly more accurate on at least one, see Table 1.

Table 1 TARB-CDS Tasks and Source Sizes

Task	Reliability stressor	N
FCT	Reject/repair a suggested answer	18,866
FQT	Recognize nonsensical medical queries	1,858*
NOTA	Select none-of-the-above	18,866
PMID→Title	Exact identifier-to-title recall	4,916
Title→Link	Title-to-PubMed-link recall	4,916
Abstract→Link	Abstract-to-link retrieval	4,916
Link→Title	Link-to-title retrieval	4,916

Note: *The downloaded public FQT file contains 1,858 rows; FCT and NOTA each contain 18,866 rows.

3.2 Policy-Weighted Reliability

A CDS program may value tasks differently. For a nonnegative weight vector w on the K-simplex, policy reliability is R(m,w), the weighted sum of the entries in a(m). We pre-register three illustrative policies: balanced (uniform weights), uncertainty-sensitive (60% total weight on FCT, FQT, and NOTA), and literature-retrieval (75% total weight on the four PubMed tasks). These scores are scenario analyses, not claims about a universal clinical workload.

$$R(m,w) = \sum_k w(k)a(m,k)\text{, with } w(k) \geq 0 \text{ and } \sum_k w(k) = 1 \quad (2)$$

3.3 Contextual Rank Stability

When deployment task weights are not known, a fixed ranking is fragile. We define contextual top-rank probability pi(m) as the probability that model m maximizes policy reliability under a distribution D over task mixtures. In the main analysis, D is Dirichlet(1,... ,1), uniform over the simplex. Monte Carlo estimates use 100,000 draws and seed 20230719.

$$\pi(m) = \text{probability over } w \text{ drawn from } D \text{ that } m \text{ maximizes } R(j, w) \text{ across models } j \quad (3)$$

This probability quantifies leaderboard sensitivity to plausible task weighting; it is not a forecast of real clinical prevalence. We additionally report top-three probability and mean rank. To assess sensitivity to which benchmark tasks were included, we resample the seven task columns with replacement 20,000 times. The resulting intervals describe task-composition sensitivity only.

3.4 Scale, Tuning, and Format Reliability

For open checkpoints with disclosed parameter counts, we compute descriptive Spearman correlations between log parameter count and each reliability measure. Commercial models are excluded because exact evaluated parameter counts are unspecified. Matched tuning effects are computed as tuned minus base scores within Llama-2 (7B, 13B, 70B), Falcon-40B, and MPT-7B pairs. Finally, format-error rates reported by Med-HALT are treated as a separate reliability dimension because malformed structured output can break downstream software even when the semantic answer is recoverable.

4 DATA AND EXPERIMENTAL PROTOCOL

4.1 Source Data and Provenance

The analysis uses two types of source material from Med-HALT [21]. First, the official public FCT, FQT, and NOTA CSV files were downloaded and audited for row counts, identifiers, source datasets, question length, option count, and missing metadata. Second, model-by-task accuracy percentages were transcribed from the published Tables 2 and 3, and format-error percentages from Table 5. The accuracy matrix covers 12 model checkpoints and seven tasks. The four memory tests each contain 4,916 PubMed samples in the source benchmark. No model responses were regenerated, and no new LLM inference is claimed.

Table 2 Audit of Downloaded Reasoning Files

Task	Rows	Median words	Median options
FCT	18,866	13	5
Fake	1,858	46	6
NOTA	18,866	13	4

Table 3 Task-Aware Reliability Summary

Model	Macro	Worst	CVaR-2	Top-1 %	Pareto
Falcon-40B	42.67	0.06	9.36	77.60	Yes
Falcon-40B-Instruct	39.66	0.08	0.60	0.00	Yes
Llama-2-70B	35.59	0.02	0.07	16.91	Yes
Text-Davinci	34.62	0.00	0.01	0.30	Yes
GPT-3.5	30.47	0.02	0.15	1.28	Yes
Llama-2-13B	29.29	0.53	1.12	3.91	Yes
MPT-7B	24.26	0.00	0.42	0.00	No
Llama-2-13B-Chat	20.45	1.38	1.55	0.00	Yes
Llama-2-7B	18.95	0.00	0.00	0.00	No
MPT-7B-Instruct	17.97	0.00	0.02	0.00	No
Llama-2-7B-Chat	13.89	0.00	0.00	0.00	No
Llama-2-70B-Chat	12.28	0.61	0.71	0.00	Yes

4.2 Dataset Audit

The files contain 39,590 rows. FCT and NOTA share 18,866 identifiers and are task transformations, not independent observations. FQT has 1,858 rows (10.15:1 FCT:FQT) and longer questions (median 46 versus 13 words). These imbalances argue against micro-averaging.

4.3 Reproducibility Protocol

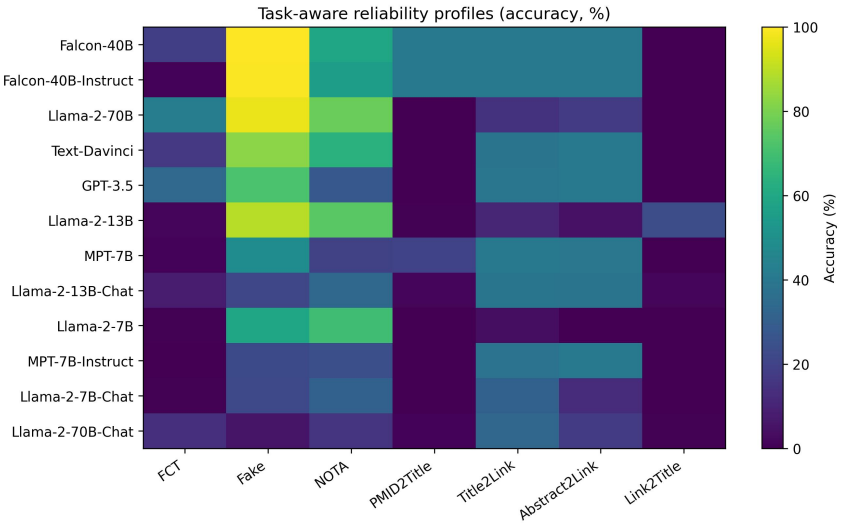

Figure 1 Task-Aware Reliability Profiles

Note: Rows are sorted by seven-task macro accuracy; color encodes published task accuracy. The visible heterogeneity motivates retaining the full task vector.

All derived quantities were computed in Python with pandas, NumPy, SciPy, and Matplotlib, see Figure 1. The analysis script fixes all random seeds, validates expected columns and row counts, and exports machine-readable CSV and JSON files. The supplied provenance document distinguishes downloaded records, manually transcribed published aggregates, and newly computed statistics. Numerical values in this paper are generated from the saved result files; rounding is performed only for presentation.

5 RESULTS

5.1 Aggregate Performance Does Not Remove Weak Links

Falcon-40B achieves the highest macro accuracy (42.67%) and the highest CVaR-2 (9.36%). Nevertheless, its Link-to-Title accuracy is 0.06%, so its strongest lower-tail score remains far below a plausible CDS acceptance threshold. Across all 12 models, the highest weakest-task accuracy is only 1.38%, achieved by Llama-2-13B-Chat. Thus, no model simultaneously performs adequately on every benchmark function. Figure 2 shows that a higher macro score does not imply a proportionately safe lower tail.

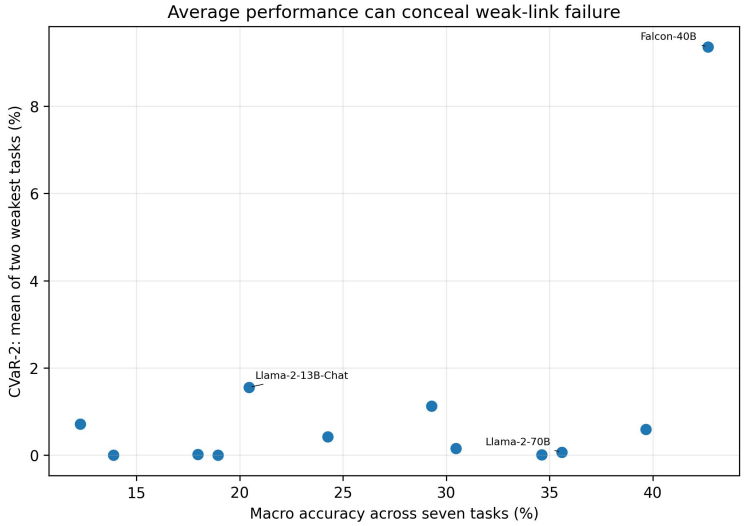

Figure 2 Macro Accuracy Versus CVaR-2

Note: A favorable mean can coexist with near-zero performance on one or more tasks.

The task leaders are also fragmented. Llama-2-70B leads FCT (42.21%) and NOTA (77.53%); Falcon-40B leads FQT (99.89%) and three PubMed mappings (40.46%); Llama-2-13B leads Link-to-Title (23.72%). Eight of the 12 models lie on the seven-dimensional Pareto frontier because their strengths occur on different tasks. A single global winner therefore depends on which capabilities are valued.

5.2 Rankings Depend on Task Mixture

Under a uniform distribution over task-weight vectors, Falcon-40B is top-ranked in 77.60% of 100,000 mixtures and appears in the top three in 99.25%. Llama-2-70B is first in 16.91% and Llama-2-13B in 3.91%. GPT-3.5 and Text-Davinci together account for 1.58% of first-place contexts. The leading model therefore changes in 22.40% of sampled task mixtures even though the task set and accuracy matrix are fixed, see Figure 3.

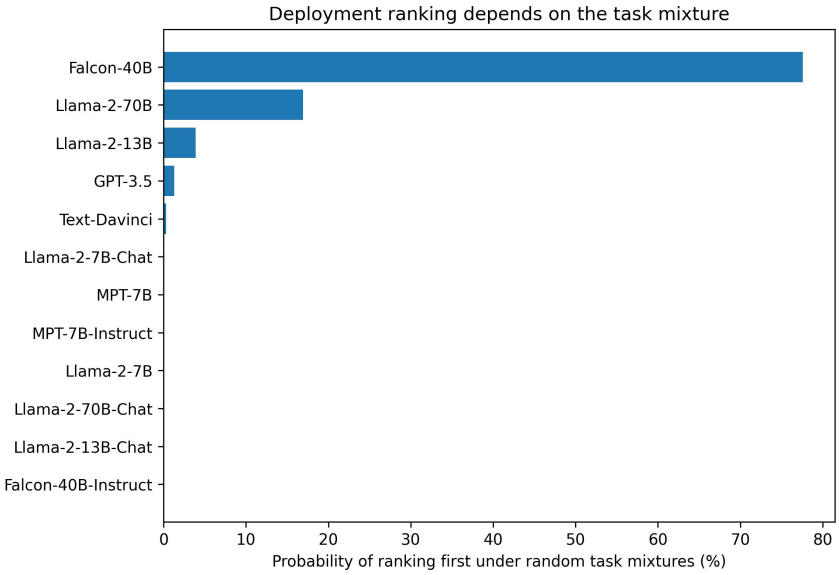

Figure 3 Probability of Ranking First under 100,000 Dirichlet(1) Task Mixtures
Note: Zero-height bars indicate no sampled first-place context.

Pre-registered policy scores make the source of rank changes explicit. Under the uncertainty-sensitive policy, Llama-2-70B ranks first at 51.80, followed by Falcon-40B at 48.43. Under the literature-retrieval policy, Falcon-40B ranks first at 34.67 and Falcon-40B-Instruct second at 33.61. These reversals are not statistical noise: they follow from different operational objectives. Consequently, a procurement report should publish the policy weights or a range of rankings rather than an unqualified model order.

5.3 Scale Is Task-Dependent

Across the ten open checkpoints with disclosed sizes, the descriptive Spearman association between log parameter count and macro accuracy is $\rho=0.366$ (two-sided descriptive $p=0.298$); for CVaR-2 it is $\rho=0.545$ ($p=0.104$). The relationship is highly task-specific: FCT shows a strong positive association ($\rho=0.884$), while Title-to-Link ($\rho=0.102$) and Abstract-to-Link ($\rho=0.120$) show little monotone association. These estimates are not causal because architecture, data, tokenizer, tuning, and prompting differ across families.

Within Llama-2 base models, macro accuracy rises monotonically from 18.95% (7B) to 29.29% (13B) and 35.59% (70B). The chat series is non-monotonic: 13.89%, 20.45%, and 12.28%, respectively. Thus, the statement 'larger is better' holds for this base subset but fails after chat tuning. The result supports separating parameter scale from post-training configuration in model cards.

5.4 Instruction/Chat Tuning Produces Trade-Offs

All five matched tuned variants have lower macro accuracy than their base checkpoints. The mean paired change is −9.30 percentage points (median −6.28), with the largest decrease for Llama-2-70B-Chat (−23.31). This does not mean instruction tuning is generally harmful: every tuned family improves at least one task, and Llama-2 chat variants improve several PubMed mappings. Rather, alignment changes the capability profile and can exchange hallucination resistance for retrieval or format behavior, see Figure 4.

Table 4 Matched Tuned Minus Base Changes (Percentage Points)

Pair	Δ Macro	Improved	Worsened	Largest shift
Llama-2-7B→Chat	-5.06	2	3	NOTA -38.4
Llama-2-13B→Chat	-8.85	4	3	Fake -68.0
Llama-2-70B→Chat	-23.31	4	3	Fake -91.8
Falcon-40B→Instruct	-3.02	1	4	FCT -17.6
MPT-7B→Instruct	-6.28	2	4	Fake -25.9

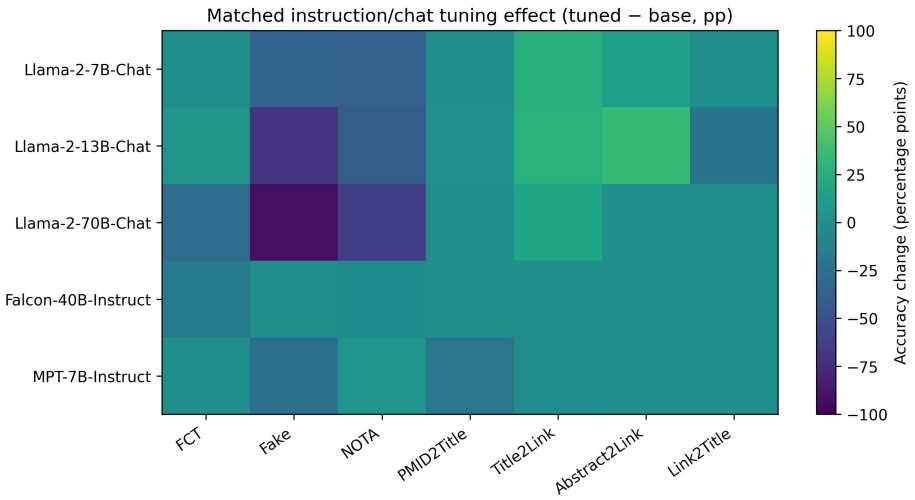

Figure 4 Task-Level Tuned-Minus-Base Accuracy Changes
Note: Positive retrieval gains frequently coexist with reasoning losses.

5.5 Format Compliance Is a Separate Failure Channel

Format reliability does not track semantic accuracy monotonically. Among the nine models reported in the source format table, Falcon-40B and Falcon-40B-Instruct have zero recorded format error, whereas Llama-2-70B-Chat averages 9.43% and reaches 41.10% on FCT and 24.92% on NOTA. GPT-3.5 and Text-Davinci average 2.03% and 1.43%, respectively. Because downstream CDS integrations often require machine-readable fields, format error should be reported beside accuracy rather than folded invisibly into it.

5.6 Sensitivity to Benchmark Composition and Metric Choice

Task-column resampling confirms that model summaries are highly sensitive to which capabilities a benchmark emphasizes. Falcon-40B has a task-resampled macro interval of 22.64%–65.42%, while Llama-2-70B spans 10.93%–63.01%. The corresponding Falcon-40B CVaR-2 interval ranges from 0.06% to 40.46% because repeated sampling can either include or omit its near-zero Link-to-Title task. These wide ranges are not estimates of clinical uncertainty; they demonstrate that benchmark composition itself is a major source of ranking variation.
Metric choice also changes the apparent winner. Falcon-40B leads macro accuracy, CVaR-2, retrieval policy score, and contextual top-rank probability; Llama-2-13B-Chat leads the weakest-task criterion; and Llama-2-70B leads the uncertainty-sensitive policy. None of these criteria is intrinsically correct for every workflow. The appropriate metric follows from the failure mode an institution is unwilling to tolerate, see Table 5.

Table 5 How the Leader Changes with the Reliability Objective

Objective	Leader	Value	Interpretation
Macro	Falcon-40B	42.67%	Average capability
Weakest task	Llama-2-13B-Chat	1.38%	Minimum floor
CVaR-2	Falcon-40B	9.36%	Lower-tail robustness
Uncertainty policy	Llama-2-70B	51.80	Reasoning/abstention emphasis
Retrieval policy	Falcon-40B	34.67	Literature-memory emphasis
Contextual top-1	Falcon-40B	77.60%	Rank stability across mixtures

5.7 Pareto Frontier and Dominated Alternatives

Eight models are Pareto-efficient in the seven-task accuracy space: GPT-3.5, Text-Davinci, Llama-2-70B, Llama-2-70B-Chat, Falcon-40B, Falcon-40B-Instruct, Llama-2-13B, and Llama-2-13B-Chat. Their inclusion does not imply clinical adequacy; it means that no other evaluated checkpoint improves every task simultaneously. The remaining four checkpoints—MPT-7B, MPT-7B-Instruct, Llama-2-7B, and Llama-2-7B-Chat—are accuracy-dominated on this task set.
Pareto analysis is useful as a screening step because a dominated model has no accuracy-based justification when inference cost, latency, licensing, privacy, and deployability are ignored. In practice, these additional dimensions may restore a dominated checkpoint to the feasible set—for example, a 7B model may be preferred for on-premises execution. TARB-CDS therefore treats Pareto status as one component of a multi-objective decision, not as an automatic exclusion rule. A full procurement analysis should append operational cost, energy, latency, local-data governance, and maintainability to the reliability vector.

6 DISCUSSION

6.1 Implications for Clinical Decision Support

The central finding is not that one 2023 checkpoint should be selected for healthcare. It is that model choice is conditional on task requirements, and every evaluated model exhibits a severe weak link. A CDS team that only reports a seven-task mean could select Falcon-40B while overlooking near-zero Link-to-Title accuracy; a team focused on uncertainty challenges could prefer Llama-2-70B; a retrieval-heavy workflow could choose differently again. TARB-CDS makes these choices explicit and auditable.

We recommend a reliability card with four layers. First, publish the complete task vector and data provenance. Second, set task-specific minimum gates before considering aggregate scores. Third, report lower-tail and format metrics to expose integration-breaking failures. Fourth, conduct rank-sensitivity analysis across plausible workload weights. A model should advance to prospective clinical evaluation only after satisfying the minimum gate on every task deemed safety-critical; high average accuracy cannot compensate for a failed gate.

A practical gate can be expressed as a conjunction rather than a compensatory score: all safety-critical task accuracies must exceed institution-defined thresholds, structured-output error must remain below an integration threshold, and the model must support a tested deferral pathway. Only models passing these gates should be compared by weighted utility. This two-stage rule prevents a high score on an easy or frequent task from offsetting failure on a rare but hazardous one. The thresholds themselves require clinical governance and prospective evidence; this paper intentionally does not invent universal cutoffs.

6.2 Why the Result Goes Beyond Model Scale

Scale can increase knowledge and reasoning capacity, but the present audit shows that its reliability effect is conditional [23, 29]. The strongest scale association appears on FCT, while exact PubMed mappings show much weaker trends. Post-training can reverse the scale pattern and reshape which tasks improve. Therefore, parameter count is best treated as one explanatory covariate, not as a deployment metric. A task-aware benchmark can reveal whether gains occur on the capabilities that matter and whether new failure modes are introduced.

6.3 Benchmark Design Lessons

The raw audit identifies meaningful imbalance: FQT has roughly one-tenth as many rows as FCT, and its questions are substantially longer. FCT and NOTA reuse underlying identifiers. These facts do not invalidate the benchmark, but they caution against naive micro-averaging and independence assumptions. Future medical reliability benchmarks should publish example-level outputs, confidence or log-probability data, explicit abstention labels, subgroup metadata, and repeated prompt runs. Those additions would support calibrated risk-coverage curves and stronger uncertainty analysis under distribution shift [14-20].

6.4 Worked Decision-Support Scenarios

Consider first a literature-assistance workflow that maps article identifiers, titles, abstracts, and links. Under the pre-registered retrieval policy—5% weight on each reasoning task and 21.25% on each retrieval task—Falcon-40B scores 34.67, compared with 24.95 for Text-Davinci and 17.68 for Llama-2-70B. However, Falcon-40B achieves only 0.06% on Link-to-Title. If Link-to-Title is a mandatory function, the model fails a minimum gate despite leading the weighted score. The rational engineering response is not to accept the average, but to use a deterministic PubMed service for that mapping or remove the capability from the LLM boundary.

Consider next an uncertainty-challenge workflow emphasizing rejection of misleading suggestions, fake questions, and none-of-the-above cases. With weights 25%, 20%, and 25% on FCT, FQT, and NOTA and 7.5% on each retrieval task, Llama-2-70B leads at 51.80, ahead of Falcon-40B at 48.43. Yet its two weakest retrieval accuracies are 0.02% and 0.12%. The same checkpoint may therefore be a candidate for a tightly bounded reasoning experiment and an unacceptable choice for a general biomedical retrieval assistant. This worked comparison illustrates why task weights, hard gates, and system boundaries must be documented together.

Neither scenario establishes clinical safety. They show how the benchmark supports an auditable decision: define the workflow, assign weights, identify non-compensable tasks, check format and abstention behavior, and then decide whether the LLM should answer, defer, or call a verified external tool. The output is a model-and-workflow pair rather than a universal model ranking.

7 LIMITATIONS AND ETHICAL CONSIDERATIONS

This study is a secondary analysis of published aggregate model results. It does not rerun inference, verify each original prediction, estimate patient outcomes, or assess modern models released after 2023. The term CDS denotes a target application context and pre-deployment screening objective; the benchmark tasks are examination and biomedical-retrieval proxies rather than direct diagnosis or treatment decisions. No result should be interpreted as evidence that a model is safe for unsupervised clinical use.

The policy weights and Dirichlet distribution are analytical devices. Different institutions should replace them with workload estimates and severity-informed weights established prospectively. Task-resampling intervals measure sensitivity to benchmark composition, not sampling uncertainty over patients. Scale correlations use only ten non-independent open checkpoints and are confounded by model family and training. The format analysis is incomplete because Med-HALT did not report format rows for every checkpoint. Finally, accuracy does not measure calibration, fairness, demographic robustness, explanation faithfulness, privacy, latency, or human factors.

The analysis uses public benchmark data and contains no identifiable patient records. However, synthetic fake questions and transformed exam items may encode cultural or source-dataset biases. Clinical deployment would require governance, security review, human oversight, prospective validation, monitoring, and a clear mechanism for abstention and escalation. TARB-CDS is intended to prevent overclaiming, not to certify safety.

8 CONCLUSION

TARB-CDS reframes medical LLM comparison from ‘Which model is largest?’ to ‘Which task profile is acceptable for a specified workflow?’ On seven Med-HALT tasks, no evaluated model avoids a near-zero weak link; task mixtures change the leader; scale associations vary sharply by task; and all five matched instruction/chat variants lose macro accuracy despite selective gains. These findings support complete task vectors, lower-tail risk metrics, explicit policy weights, format-compliance reporting, and minimum task gates as prerequisites for any clinical evaluation. The accompanying code and provenance package make every new statistic in this paper reproducible from public data and published source tables.

9 REPRODUCIBILITY STATEMENT

The artifact contains the downloaded FCT, FQT, and NOTA CSV files; source-table transcriptions; a single executable analysis script; fixed random seeds; package requirements; all derived CSV/JSON results; and publication figures. Running `python src/run_analysis.py` from the artifact root regenerates the analysis outputs. A separate reference-verification CSV records the archival source used for each citation. Source aggregate accuracies are preserved separately from newly calculated measures to prevent provenance ambiguity.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Singhal K. Large language models encode clinical knowledge. Nature, 2023, 620: 172-180. DOI: 10.1038/s41586-023-06291-2.

[2] Pal A, Umapathi L K, Sankarasubbu M. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In: Proceedings of the Conference on Health, Inference, and Learning, PMLR, 2022, 174: 248-260.

[3] Jin Q, Dhingra B, Liu Z, et al. PubMedQA: A dataset for biomedical research question answering. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019: 2567-2577. DOI: 10.18653/v1/D19-1259.

[4] Vilares D, Gómez-Rodríguez C. HEAD-QA: A healthcare dataset for complex reasoning. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 960-966. DOI: 10.18653/v1/P19-1092.

[5] Jin D. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Applied Sciences, 2021, 11(14): 6421. DOI: 10.3390/app11146421.

[6] Kung T H. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health, 2023, 2(2): e0000198. DOI: 10.1371/journal.pdig.0000198.

[7] Gilson A. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. JMIR Medical Education, 2023, 9: e45312. DOI: 10.2196/45312.

[8] Nori H, King N, McKinney S M, et al. Capabilities of GPT-4 on medical challenge problems. arXiv:2303.13375, 2023.

[9] Hendrycks D. Measuring massive multitask language understanding. In: Proceedings of the International Conference on Learning Representations (ICLR), 2021.

[10] He Z. MedEval: A multi-level, multi-task, and multi-domain medical benchmark for language model evaluation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2023: 8725-8744. DOI: 10.18653/v1/2023.emnlp-main.540.

[11] Li J, Cheng X, Zhao X, et al. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2023: 6449-6464. DOI: 10.18653/v1/2023.emnlp-main.397.

[12] Ribeiro M T, Wu T, Guestrin C, et al. Beyond accuracy: Behavioral testing of NLP models with CheckList. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 4902-4912. DOI: 10.18653/v1/2020.acl-main.442.
[13] Liang P. Holistic evaluation of language models. arXiv:2211.09110, 2022.
[14] Guo C, Pleiss G, Sun Y, et al. On calibration of modern neural networks. In: Proceedings of the International Conference on Machine Learning (ICML), PMLR, 2017, 70: 1321-1330.
[15] Ovadia Y. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In: Advances in Neural Information Processing Systems, 2019, 32.
[16] Jiang Z, Araki J, Ding H, et al. How can we know when language models know? On the calibration of language models for question answering. Transactions of the Association for Computational Linguistics, 2021, 9: 962-977. DOI: 10.1162/tacl_a_00407.
[17] Desai S, Durrett G. Calibration of pre-trained transformers. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020: 295-302. DOI: 10.18653/v1/2020.emnlp-main.21.
[18] Geifman Y, El-Yaniv R. Selective classification for deep neural networks. In: Advances in Neural Information Processing Systems, 2017, 30.
[19] Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. Nature Machine Intelligence, 2019, 1: 20-23. DOI: 10.1038/s42256-018-0004-1.
[20] Van Calster B. Calibration: The Achilles heel of predictive analytics. BMC Medicine, 2019, 17: Art. 230. DOI: 10.1186/s12916-019-1466-7.
[21] Pal A, Umapathi L K, Sankarasubbu M. Med-HALT: Medical Domain Hallucination Test for Large Language Models. In: Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL), 2023: 314-334. DOI: 10.18653/v1/2023.conll-1.21.
[22] Bommasani R. On the opportunities and risks of foundation models. arXiv:2108.07258, 2021.
[23] Brown T B. Language models are few-shot learners. In: Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.
[24] Wang X. Self-consistency improves chain of thought reasoning in language models. In: Proceedings of the International Conference on Learning Representations (ICLR), 2023.
[25] Kojima T, Gu S S, Reid M, et al. Large language models are zero-shot reasoners. In: Advances in Neural Information Processing Systems, 2022, 35: 22199-22213.
[26] Chung H W. Scaling instruction-finetuned language models. arXiv:2210.11416, 2022.
[27] Wei J. Finetuned language models are zero-shot learners. In: Proceedings of the International Conference on Learning Representations (ICLR), 2022.
[28] Ouyang L. Training language models to follow instructions with human feedback. In: Advances in Neural Information Processing Systems, 2022, 35: 27730-27744.
[29] Touvron H. Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288, 2023.