



## Task-Difficulty-Aware Reliability Evaluation of Large Language Models for Clinical Decision Support

Ryan Cole, Benjamin Rhodes\*

Kahlert School of Computing, University of Utah, Salt Lake City, UT, USA

\*Corresponding author: Benjamin Rhodes, [u3930202@cs.utah.edu](mailto:u3930202@cs.utah.edu)

### Abstract

*Model choice for clinical decision support (CDS) is increasingly justified by leaderboard accuracy on medical hallucination benchmarks, yet an accuracy figure is only interpretable relative to what a system with no medical knowledge could already obtain. We introduce DARE-CDS, a task-difficulty-aware reliability framework that models every benchmark task by a response curve whose pseudo-guessing parameter is measured directly from the released data instead of being assumed. Applying the framework to the full public Med-HALT corpus (59,254 items across seven tasks) and to the twelve checkpoints evaluated on it, we find that the gold label of the None-of-the-Above test is constant for all 18,866 items, that the Fake-Question scorer is satisfied by a constant abstention string, and that 1,989 of 4,916 PubMed items are unanswerable so that a constant "Unknown" reply scores exactly 40.46%, reproducing to the second decimal the best published memory-retrieval accuracy. A bank of 30 deterministic knowledge-free solvers bounds the reasoning floor at 32.95%. Consequently 34.97% of benchmark rows carry zero Fisher information, the effective test length is 26.1% of the nominal one, and only 2 of 84 model-task cells exceed the knowledge-free ceiling under Holm-corrected exact binomial tests. Difficulty-aware scoring reverses the leaderboard: the model that is top-ranked in 77.60% of random task mixtures under raw accuracy is top-ranked in 0.02% after correction. We further show that within-task item difficulty is not identifiable from knowledge-free responses, release a difficulty-controlled repair of the protocol that restores measurement capacity, and derive the item counts a medical reliability benchmark needs to support a defensible pre-deployment claim.*

**Keywords:** clinical decision support, large language models, reliability evaluation, benchmark validity, item response theory, hallucination, task difficulty

### 1. Introduction

Large language models (LLMs) now reach or exceed passing thresholds on medical licensing examinations [1], [2], [3], and this has moved the discussion from whether such systems can answer clinical questions to whether they can be trusted inside a clinical decision support (CDS) workflow. Benchmarks built specifically to probe medical hallucination have become the evidence base for that discussion [8], [11]. A procurement committee reading such a leaderboard sees a table of accuracies, concludes that one checkpoint dominates, and treats the margin as a reliability signal. This paper argues that the margin is frequently not a signal at all. An accuracy value is only interpretable relative to the accuracy that a system with no medical knowledge whatsoever would already obtain on the same task under the same scorer. In classical test theory this quantity is the pseudo-guessing parameter of the item response curve [15], [16]; in machine learning it appears as the majority-class or hypothesis-only baseline that has repeatedly been shown to consume most of the apparent headroom of natural language benchmarks [28], [29], [30], [31]. When the guessing level is high, two things happen simultaneously: reported scores become uninformative about ability, and the task becomes statistically

incapable of separating models no matter how many items it contains.

Recent auditing work has made real progress by refusing to compress a model-by-task matrix into a single mean, retaining the full task vector, lower-tail risk statistics and rank-stability analysis under uncertain task mixtures [9]. That line of work correctly establishes that the aggregate hides task-specific failure. We take the complementary step inward: we ask whether the individual task scores that such an audit aggregates are themselves measurements. This is what we mean by task-difficulty-aware evaluation. Difficulty here is not a subjective label attached to a question; it is the operational quantity that determines how much a score can possibly tell us, and for the benchmarks that dominate medical LLM evaluation it can be measured exactly from the released files.

Our contributions are as follows.

- *A measured-guessing task response model.* We formalise a benchmark task by a response curve whose pseudo-guessing parameter is not estimated from model behaviour but measured from the released items and the official scorer, and derive task capacity, effective test length and Fisher information from it.
- *A full structural audit of a widely used medical hallucination benchmark.* We download and audit every public Med-HALT file and show that three of the seven tasks are solved exactly by constant response strings, that two of the tasks share one identical item pool under disjoint identifiers, and that the reported best memory-retrieval accuracy coincides to two decimals with the proportion of unanswerable items.
- *A reference solver bank.* Thirty deterministic, non-neural solvers are run over 18,866 real examination items to bound empirically what is reachable without medical knowledge, and to test whether within-task item difficulty is identifiable at all from such responses.
- *A difficulty-aware re-analysis of twelve published checkpoints.* Under Holm-corrected exact binomial tests only two of eighty-four model-task cells constitute evidence of medical ability, and the leaderboard order reverses once capacity is accounted for.
- *A repaired protocol and a design calculus.* We release MedHALT-D, a difficulty-controlled re-instantiation built from the same real items, and give the item counts required for a benchmark to resolve a stated difference in corrected accuracy.

We emphasise what this paper does not claim. We do not run new inference, we do not assert that any model is safe or unsafe for patient care, and we do not argue that the benchmark we audit is without value; its item pool is a genuine multinational examination resource and remains the basis of our repaired protocol.

## II. RELATED WORK

### A. Medical LLM evaluation

Medical question answering datasets such as MedMCQA, PubMedQA, HEAD-QA and MedQA made large-scale measurement of biomedical knowledge possible [4], [5], [6], [7], and general and medically adapted models were subsequently reported to perform strongly on them [1], [2], [3]. Because examination accuracy is not a safety case, later benchmarks targeted hallucination behaviour directly, either in the general domain [10], [11] or specifically in medicine [8]. Broader evaluation frameworks argue for coverage across many dimensions rather than one leaderboard number [12], [13], [14]. Reliability auditing of medical LLM comparisons has begun to combine these ideas, retaining full task vectors, lower-tail risk statistics of the kind used in financial risk management [38], and rank stability under uncertain deployment mixtures. Our work is downstream of that agenda: we supply the measurement-validity layer that determines whether the task scores being aggregated carry information.

### B. Psychometric evaluation of NLP systems

Item response theory (IRT) models the probability of a correct response as a function of a latent ability and item parameters for difficulty, discrimination and guessing [15], [16]. Its use in NLP was introduced for building evaluation scales [17] and extended to leaderboard design [18], test-set comparison [19] and efficient benchmark subsampling [20]. These studies estimate item parameters from a panel of model responses. Our setting differs in an important practical way: per-item responses of the evaluated checkpoints were never released, only aggregate task accuracies. We therefore work at the task level and, crucially, we do not estimate the guessing parameter at all; we measure it from the item files and the official scoring function, which removes the identifiability problems that usually accompany three-parameter models.

### C. Shortcut learning and benchmark artifacts

A long line of work shows that leaderboard progress can be produced by exploitable regularities rather than by the intended competence: hypothesis-only baselines in inference [28], [29], spurious cues in argument reasoning [30], passage-free reading comprehension [31], and shortcut learning as a general phenomenon [32]. For multiple-choice evaluation of LLMs specifically, position and option-order effects are large enough to change model rankings [33], [34]. What we add is the observation that in the medical hallucination setting the shortcut is not a subtle statistical bias but a degenerate label distribution, and that its size can be computed in closed form and converted into a statement about how many items the benchmark effectively has.

#### D. Uncertainty, calibration and deferral

Correctness and confidence are distinct properties [21], [22], [23], and selective prediction formalises the option to abstain [24], [25]. In medicine, uncertainty quantification and calibration are treated as prerequisites for safe use [26], [27]. Our analysis interacts with this literature in a specific way: when a task rewards a constant abstention string with a perfect score, abstention behaviour and knowledge become statistically inseparable, so calibration cannot be assessed on that task either.

### III. THE DARE-CDS FRAMEWORK

#### A. Task response model with a measured guessing parameter

Let a benchmark consist of tasks  $t = 1, \dots, T$ , task  $t$  containing  $N_t$  items scored by a fixed function. For a model with latent medical ability  $\theta$ , we write the probability of scoring an item of task  $t$  as correct as

$$\pi_t(\theta) = c_t + (1 - c_t) \sigma(\theta - b_t), \quad (1)$$

where  $\sigma$  is the logistic function,  $b_t$  is task difficulty and  $c_t$  is the pseudo-guessing parameter of the three-parameter model [16]. Equation (1) is standard. The departure from standard practice is that we do not estimate  $c_t$ . We define the knowledge-free policy class  $\Pi_0$  as the set of response policies that use no medical content: constant strings, uniform choice among options, positional and length heuristics, and corpus-associative solvers that exploit only distributional regularities of the released files. We then set

$$c_t = \max_{\pi \in \Pi_0} \text{Acc}_t(\pi), \quad (2)$$

which is computable exactly for constant policies and measurable empirically for the rest. This makes the most delicate parameter of the model an observable property of the benchmark rather than a quantity confounded with model behaviour.

#### B. Capacity, effective test length and information

The *capacity* of a task is  $H_t = 1 - c_t$ , the range of accuracy that medical knowledge can still buy. Given an observed accuracy  $p_{mt}$  for model  $m$ , the knowledge-corrected score is

$$\kappa_{mt} = (p_{mt} - c_t) / H_t, \quad (3)$$

the classical correction for guessing. Because the sampling variance of  $\kappa_{mt}$  is inflated by  $1/H_t^2$ , a task of nominal length  $N_t$  behaves like a test of *effective length*

$$N_{\text{eff},t} = N_t H_t^2. \quad (4)$$

The Fisher information that one item of task  $t$  supplies about  $\theta$  follows from (1):

$$I_t(\theta) = H_t^2 \sigma'(\theta - b_t)^2 / [\pi_t(\theta)(1 - \pi_t(\theta))]. \quad (5)$$

Equations (4) and (5) give the framework its teeth. As  $c_t \rightarrow 1$  both the effective length and the information vanish identically for every  $\theta$ : the task cannot separate any two models, and no amount of additional sampling repairs it. A task with  $c_t = 1$  is not a hard task; it is not a measurement.

#### C. Evidence, ability and difficulty-aware ranking

We treat a reported cell as *evidence of medical ability* only if the one-sided exact binomial test of  $H_0: p_{mt} \leq c_t$  is rejected after Holm correction across all model-task cells. Ability is then estimated by maximum likelihood under (1) on the tasks that retain capacity, with standard errors from the observed information, and models are re-ranked by  $\theta$  rather than by macro accuracy. Finally, because deployment task weights are unknown, we repeat a rank-stability analysis under 100,000 Dirichlet(1) task mixtures, once on raw accuracies and once on capacity-corrected scores, so that the two leaderboards can be compared under identical weighting uncertainty.

#### D. Protocol repair operators

Three operators restore capacity without discarding the underlying examination items. Label balancing randomises the gold class so that no constant string is rewarded; position randomisation destroys any alignment between the correct option and its index; and task separation splits a compound task into

components with independent gold labels, so that a shortcut valid for one component cannot be reused for the other. Applying these to the audited benchmark yields the MedHALT-D protocol of Section VI-F.

#### IV. DATA AND EXPERIMENTAL PROTOCOL

##### A. Source material

All item-level data are the public Med-HALT release [8] obtained from the project repository: the three reasoning files (False Confidence Test, FCT; Fake Questions Test, FQT; None-of-the-Above, NOTA) and the four PubMed memory files, together with the official evaluation script that defines the scorer. Model-by-task accuracies for twelve checkpoints, and the format-error table, are the published values from the same work; no inference was re-run and no model response was simulated. The examination items originate from AIIMS-PG and NEET-PG (India), the Spanish medical residency examination, the USMLE and the Taiwanese medical licensing examination [4], [6], [7]; the evaluated checkpoints span the Llama-2 [35], Falcon [36], MPT and OpenAI families.

##### B. Reference solver bank

Thirty deterministic solvers were implemented in five families: surface lexical matching, distributional vector similarity (TF-IDF and latent semantic projections), structural heuristics (option length, position, option-set centrality), corpus-associative retrieval (nearest-neighbour answer transfer, out-of-fold answer priors, subject-conditional association, a cross-space linear map), and seeded stochastic anchors. Every solver that uses corpus statistics is fitted out of fold so that no item contributes to its own prediction. The bank is run on the 18,866 four-option items that underlie FCT and NOTA and produces a real  $30 \times 18,866$  binary response matrix. Implementation uses scikit-learn [40]; all seeds are fixed.

##### C. Reproducibility

The analysis is a single pipeline of six scripts that download nothing at run time beyond the public benchmark archive, fix every random seed, and write all derived quantities to machine-readable files. Every number reported below is regenerated from those files; rounding is applied only for presentation.

#### V. STRUCTURAL AUDIT OF THE BENCHMARK

Table I summarises the audit. Four findings determine everything that follows.

*Constant gold labels.* The NOTA gold answer is the literal string "None of the above" for 18,866 of 18,866 items, a rate of 100%. Because the official scorer compares the model's selected option to that string, a policy that always emits it scores 100%. The FQT scorer accepts a response containing any of fourteen keywords, among them "i do not know" and "none of the above"; a constant abstention reply therefore also scores 100%. Moreover the fake items themselves are written in a deliberately absurd register, so that 42.94% of their options contain a triggering keyword on average and every item contains at least one: copying a uniformly random option already scores 42.94%.

*Unanswerable memory items.* Each of the four PubMed files contains 4,916 records of which 1,989 are marked as not corresponding to a real article, i.e. 40.4597%. A model that always answers "Unknown" therefore scores 40.46%. This is exactly the accuracy reported for Falcon-40B and Falcon-40B-Instruct on PMID-to-Title, Title-to-Link and Abstract-to-Link, and for MPT-7B on Title-to-Link. The apparently decisive memory-retrieval advantage of the Falcon family is reproduced, to two decimal places, by a policy that retrieves nothing.

*A degenerate verification sub-task.* In FCT the suggested answer coincides with the correct one in 1 of 18,866 items (0.0053%). The yes/no component of the task is therefore answered correctly by the constant reply "no", and the residual selection component can exclude the suggestion before guessing, which lifts the floor from 25% to one third.

*A duplicated item pool.* FCT and NOTA share no identifier at all, yet the sets of question stems are identical: all 18,866 stems occur in both files, and the two files have exactly the same distribution of the gold option index (5,344 / 4,883 / 4,789 / 3,850). The two "tasks" are two framings of one item pool, so the 39,590 nominal reasoning rows correspond to 20,724 distinct stems. Averaging over tasks therefore places double weight on the same clinical content, and identifier-based de-duplication does not detect it. Prior auditing work flagged the non-independence of these two tests on the basis of their aggregate statistics [9]; the item-level files confirm it directly and show that it is complete rather than partial.

Table 1 Structural audit of the released benchmark and the resulting task capacity

| Task          | Items         | Knowledge-free policy attaining the ceiling | $c_t$ (%) | Cap. (%) | N_eff         |
|---------------|---------------|---------------------------------------------|-----------|----------|---------------|
| FCT           | 18,866        | exclude the suggested option, then guess    | 32.95     | 67.05    | 8,481         |
| FQT           | 1,858         | constant "I do not know"                    | 100.00    | 0.00     | 0             |
| NOTA          | 18,866        | constant "None of the above"                | 100.00    | 0.00     | 0             |
| PMID→Title    | 4,916         | constant "Unknown"                          | 40.46     | 59.54    | 1,743         |
| Title→Link    | 4,916         | constant "Unknown"                          | 40.46     | 59.54    | 1,743         |
| Abstract→Link | 4,916         | constant "Unknown"                          | 40.46     | 59.54    | 1,743         |
| Link→Title    | 4,916         | constant "Unknown"                          | 40.46     | 59.54    | 1,743         |
| <b>Total</b>  | <b>59,254</b> | —                                           | —         | —        | <b>15,452</b> |

The FCT ceiling is the best of 30 measured knowledge-free solvers; all other ceilings are exact.

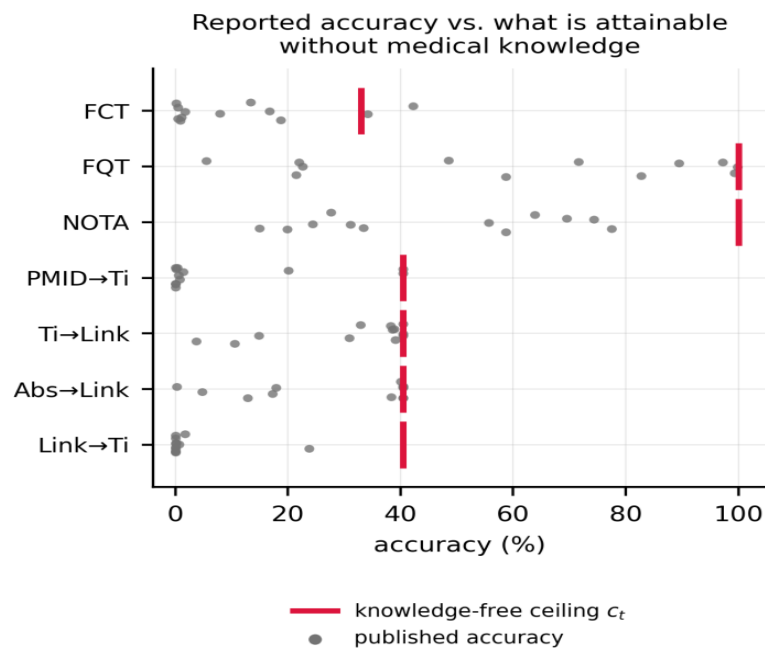


Figure 1. Published accuracy of twelve checkpoints on the seven tasks (grey) against the measured knowledge-free ceiling of each task (red). Points to the left of the red mark carry no evidence of medical ability.

## VI. RESULTS

### A. What is reachable without medical knowledge

The reference bank bounds the reasoning floor empirically. Across 18,866 real examination items the thirty

solvers span 20.41% to 32.95% with a mean of 25.76%. The strongest is the protocol-specific exclusion solver at 32.95%, followed by nearest-neighbour answer transfer at 29.92% and 29.57%, the longest-option heuristic at 28.88% and the majority-position heuristic at 28.33%. Two of these deserve comment. The position heuristic beats chance because the gold option index is not uniform, a known failure mode of multiple-choice evaluation [33], [34]. The longest-option heuristic beats chance because distractors are written more tersely than correct answers, a classical item-writing artifact. Neither uses a single piece of medical knowledge.

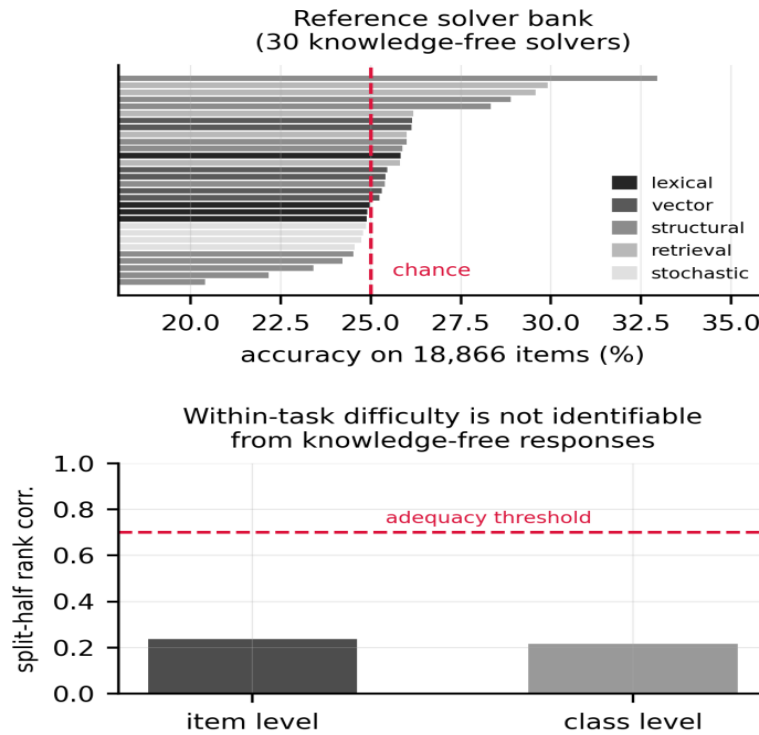


Figure 2Top: accuracy of the 30 knowledge-free solvers. Bottom: split-half rank correlation of estimated item difficulty, at the level of individual items and of 39 pre-registered item classes.

### B. Capacity and effective benchmark size

Applying (2)-(4) gives Table I. Two of the seven tasks have exactly zero capacity, and these two account for 20,724 of the 59,254 rows, that is 34.97% of the benchmark. The remaining tasks lose between one third and two fifths of their nominal range to the guessing floor. Summing (4), the benchmark behaves like a test of 15,452 items rather than 59,254, an effective size of 26.1% (Fig. 3). The corresponding information curves (Fig. 4) show the same fact on the ability scale: FCT peaks at 2,489 units of information, each memory task at 553, and FQT and NOTA supply exactly zero everywhere.

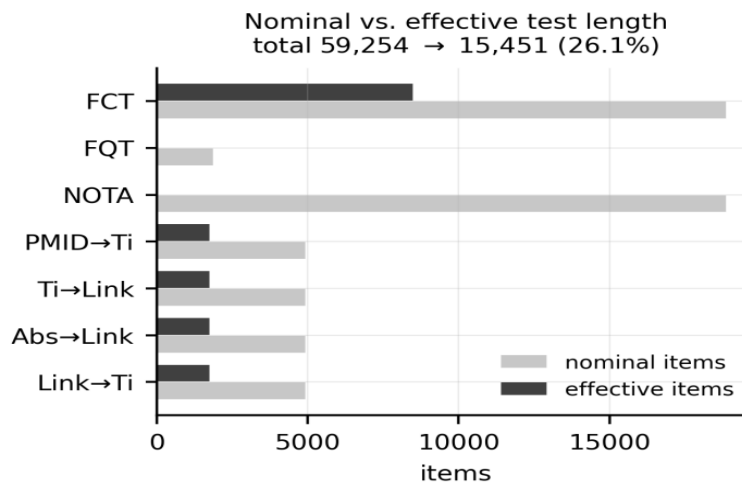


Figure 3. Nominal against effective test length per task. Zero-capacity tasks contribute no effective items regardless of how many rows they contain.

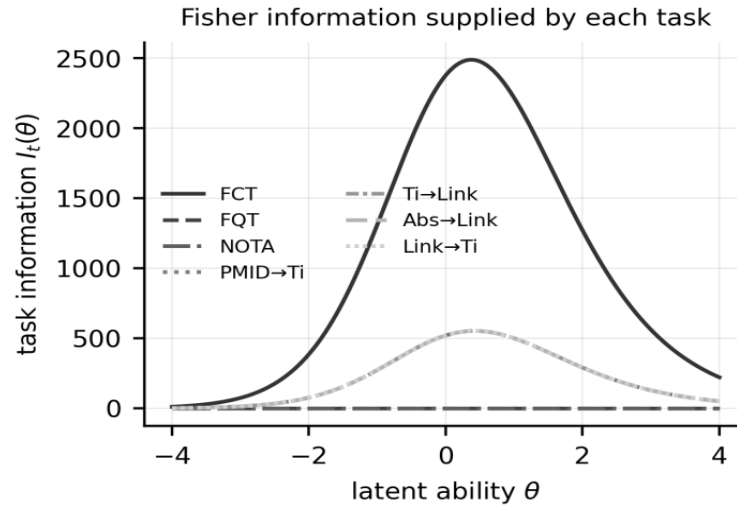


Figure 4 Fisher information supplied by each task as a function of latent ability. The FQT and NOTA curves lie identically on zero.

### C. How much of the published leaderboard is evidence

Table II gives the knowledge-corrected scores of (3) for the five tasks that retain capacity. Almost every entry is negative, meaning the checkpoint performed worse than a policy with no medical knowledge. Under one-sided exact binomial tests with Holm correction over all 84 model-task cells, exactly two cells reject the null: Llama-2-70B on FCT (42.21%, corrected 13.81%) and GPT-3.5 on FCT (34.15%, corrected 1.78%). No model on any memory task, and no model on FQT or NOTA, provides statistically detectable evidence of medical ability. The three cells at exactly 40.46% do not merely fail the test; they equal the constant-abstention policy to the precision at which they were published.

Table 2 Knowledge-corrected accuracy  $\kappa$  (%) on the five tasks that retain capacity

| Model               | FCT    | P→T    | T→L    | A→L    | L→T    |
|---------------------|--------|--------|--------|--------|--------|
| GPT-3.5             | 1.78   | -67.47 | -2.28  | -0.02  | -67.92 |
| Text-Davinci        | -24.15 | -67.92 | -3.24  | -0.03  | -67.95 |
| Llama-2-70B         | 13.81  | -67.75 | -43.11 | -39.05 | -67.92 |
| Llama-2-70B-Chat    | -29.25 | -66.59 | -12.75 | -37.89 | -66.93 |
| Falcon-40B          | -21.32 | 0.00   | 0.00   | 0.00   | -67.85 |
| Falcon-40B-Instruct | -47.49 | 0.00   | 0.00   | -0.03  | -67.82 |
| Llama-2-13B         | -46.58 | -67.06 | -50.22 | -60.06 | -28.11 |
| Llama-2-13B-Chat    | -37.29 | -65.64 | -2.70  | -3.59  | -65.05 |
| Llama-2-7B          | -48.48 | -67.95 | -61.71 | -67.52 | -67.95 |
| Llama-2-7B-Chat     | -48.52 | -67.95 | -16.02 | -46.46 | -67.95 |
| MPT-7B              | -47.88 | -34.23 | 0.00   | -0.72  | -67.95 |

| Model           | FCT    | P→T    | T→L   | A→L  | L→T    |
|-----------------|--------|--------|-------|------|--------|
| MPT-7B-Instruct | -48.90 | -67.89 | -3.73 | 0.00 | -67.95 |

P→T, T→L, A→L and L→T denote PMID-to-Title, Title-to-Link, Abstract-to-Link and Link-to-Title. Negative values indicate accuracy below the knowledge-free ceiling. Only the two FCT cells of GPT-3.5 and Llama-2-70B survive Holm-corrected testing at  $\alpha = 0.05$ .

#### D. Difficulty-aware ranking reverses the leaderboard

Estimating  $\theta$  under (1) on the tasks that retain capacity yields  $\theta = -1.83$  (SE 0.045) for Llama-2-70B and  $\theta = -4.01$  (SE 0.293) for GPT-3.5, with the remaining ten checkpoints unidentified at the floor. Only 1 of 66 model pairs is separable at the 95% level. Comparing the raw macro-accuracy order with the capacity-corrected order gives a Kendall  $\tau$  of 0.295: the two leaderboards largely disagree. The most consequential single change is that Falcon-40B, which is first by macro accuracy at 42.67%, falls to a tied rank of 7.5, because its entire advantage is concentrated in the three memory cells that coincide with the constant-abstention policy. The rank-stability analysis makes the same point under weighting uncertainty. Under 100,000 Dirichlet(1) task mixtures scored on raw accuracy, Falcon-40B is top-ranked in 77.60% of contexts, reproducing the published value of the earlier task-aware audit [9] and confirming that our pipeline is consistent with it. Scoring the identical mixtures on capacity-corrected evidence, Falcon-40B is top-ranked in 0.02% of contexts and Llama-2-70B in 99.98% (Fig. 5, Table III). The leader is therefore not merely sensitive to the task mixture; it is an artifact of the guessing structure of half the benchmark.

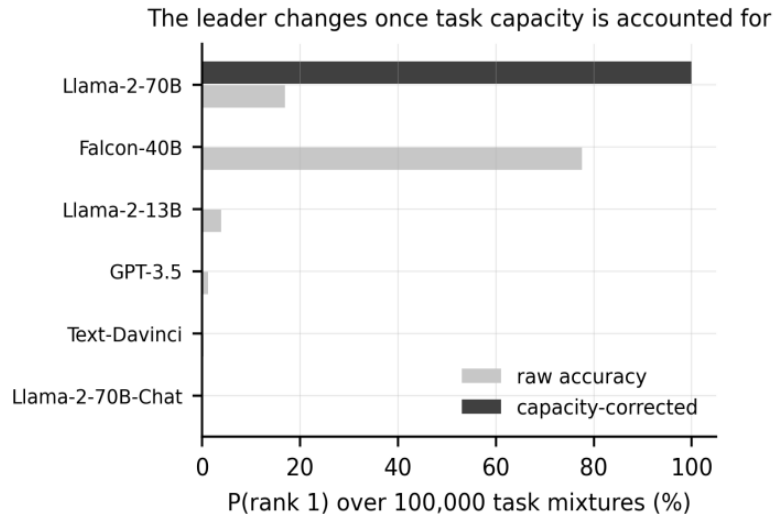


Figure 5. Probability of being top-ranked over 100,000 Dirichlet(1) task mixtures, before and after correcting for task capacity

Table 3 Raw versus difficulty-aware standing of the twelve checkpoints

| Model               | Macro acc. (%) | Raw rank | Corr. rank | P(top-1) corr. (%) |
|---------------------|----------------|----------|------------|--------------------|
| Falcon-40B          | 42.67          | 1        | 7.5        | 0.02               |
| Falcon-40B-Instruct | 39.66          | 2        | 7.5        | 0.00               |
| Llama-2-70B         | 35.59          | 3        | 1          | 99.98              |
| Text-Davinci        | 34.62          | 4        | 7.5        | 0.00               |
| GPT-3.5             | 30.47          | 5        | 2          | 0.00               |
| Llama-2-13B         | 29.29          | 6        | 7.5        | 0.00               |
| MPT-7B              | 24.26          | 7        | 7.5        | 0.00               |

| Model            | Macro acc. (%) | Raw rank | Corr. rank | P(top-1) corr. (%) |
|------------------|----------------|----------|------------|--------------------|
| Llama-2-13B-Chat | 20.45          | 8        | 7.5        | 0.00               |
| Llama-2-7B       | 18.95          | 9        | 7.5        | 0.00               |
| MPT-7B-Instruct  | 17.97          | 10       | 7.5        | 0.00               |
| Llama-2-7B-Chat  | 13.89          | 11       | 7.5        | 0.00               |
| Llama-2-70B-Chat | 12.28          | 12       | 7.5        | 0.00               |

Rank 7.5 denotes the tie of the ten checkpoints whose ability is not identified above the knowledge-free floor. Kendall  $\tau$  between the two orders is 0.295.

#### E. Within-task item difficulty is not identifiable

A natural refinement would place each individual item on a difficulty scale and report difficulty-conditional reliability profiles. We tested whether this is feasible without per-item LLM responses by fitting a Rasch model to the  $30 \times 18,866$  response matrix of the reference bank. It is not. The split-half rank correlation of estimated item difficulty is 0.236, a Spearman-Brown reliability of 0.382, far below any conventional adequacy threshold; leave-one-family-out prediction of held-out solver responses reaches a mean AUC of only 0.557; and an out-of-fold ridge model of difficulty on observable item features (stem length, numeral count, negation, option-length dispersion, source corpus) explains 0.3% of the variance. Aggregating items into 39 pre-registered classes does not help, because knowledge-free solvability varies across classes only between 22.9% and 28.0% with a standard deviation of 0.87 points (Fig. 2, bottom).

The interpretation is instructive rather than disappointing. Knowledge-free solvers are near chance almost uniformly, so their responses contain little information about which clinical items are intrinsically hard. Item-level difficulty modelling of medical benchmarks in the style of [17], [18], [19] therefore requires benchmark authors to release per-item model responses, which Med-HALT and most medical evaluations currently do not. Until they do, difficulty-aware evaluation must operate at the task and protocol level, where, as Section III shows, the decisive parameter is exactly identified.

#### F. Repairing the protocol

We rebuilt the reasoning tests from the same real items using the three operators of Section III-D, producing MedHALT-D: AS (answer selection with randomised option order), VER (verification with a balanced positive rate), ND (none-of-the-above detection with 25% NOTA prevalence and randomised NOTA position) and NSD (balanced nonsense detection from the question stem alone). Measured on the repaired items, the ceilings fall sharply: the constant-string ceiling of NOTA drops from 100% to 25.16%, the best measured shortcut on ND to 27.54%, the verification ceiling from 99.99% to 50.10%, and the position ceiling of the answer task from 28.33% to 25.27%. One repair fails, and the failure is informative. Balancing the nonsense-detection task raises its constant-policy ceiling only to 50%, but the fake questions were generated in a florid register that is trivially recognisable: a unigram bag-of-words classifier with five-fold cross-validation separates fake from real stems with AUC 0.9997 and accuracy 99.03%, and even question length alone achieves AUC 0.816. The task therefore measures stylistic anomaly detection, not medical judgement, and balancing cannot fix it; the fake items must be regenerated in a register matched to the real ones. This is precisely the shortcut phenomenon documented for other benchmarks [28], [29], [32], appearing here in a safety-critical setting. Table IV converts capacity into a design requirement: the number of items needed for a two-sided test at  $\alpha = 0.05$  and 80% power to resolve a five-point difference in corrected accuracy. The released FCT and memory tasks are adequately sized, the released NOTA and FQT have no finite requirement because their capacity is zero, and the released verification framing would need 59.2 million items. Their repairs need between 2,764 and 4,723 items, all comfortably below the 18,866 available. Balanced nonsense detection would need 322,502 items against 3,716 available, confirming that regeneration rather than resizing

is required.

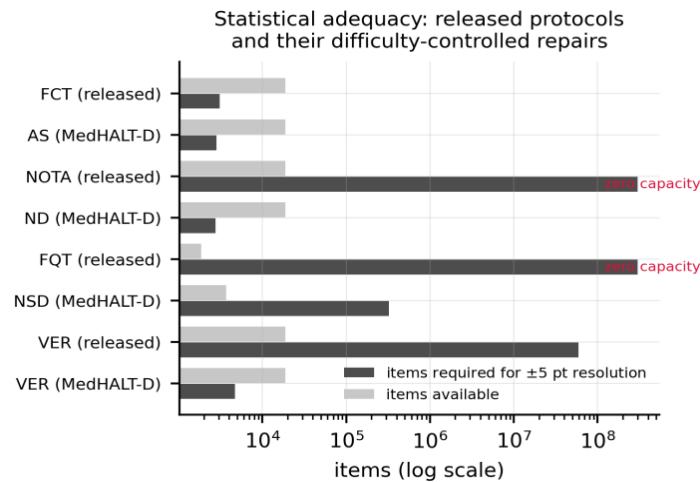


Figure 6 Items required to resolve a five-point difference in corrected accuracy, against items available, for the released protocols and their difficulty-controlled repairs.

Table 4 Statistical adequacy of released and repaired protocols

| Protocol                | Ceiling (%) | Capacity (%) | Items needed | Have   |
|-------------------------|-------------|--------------|--------------|--------|
| FCT (released)          | 32.95       | 67.05        | 3,113        | 18,866 |
| AS (MedHALT-D)          | 28.72       | 71.28        | 2,835        | 18,866 |
| NOTA (released)         | 100.00      | 0.00         | no finite N  | 18,866 |
| ND (MedHALT-D)          | 27.54       | 72.46        | 2,764        | 18,866 |
| FQT (released)          | 100.00      | 0.00         | no finite N  | 1,858  |
| NSD (MedHALT-D)         | 99.03       | 0.97         | 322,502      | 3,716  |
| VER (released)          | 99.99       | 0.01         | 59,229,217   | 18,866 |
| VER (MedHALT-D)         | 50.10       | 49.90        | 4,723        | 18,866 |
| Memory tasks (released) | 40.46       | 59.54        | 3,704        | 19,664 |

Requirement is for a two-sided test at  $\alpha = 0.05$  with 80% power against a five-point difference in  $\kappa$ .

## VII. DISCUSSION

### A. Implications for clinical decision support

The practical consequence is that a CDS procurement decision taken from the published leaderboard of this benchmark would have selected a checkpoint whose measured advantage is a response style. That is a different

failure mode from the one usually discussed. It is not that the aggregate hides a weak task; it is that most of the tasks being aggregated are not measurements of the property the aggregate claims to summarise. Reliability auditing that retains full task vectors and lower-tail statistics is necessary and should be adopted, but it should be preceded by a capacity check, because a lower-tail statistic computed over zero-capacity tasks inherits their meaninglessness. We therefore propose that a medical model card [39] carry four additional fields per task: the measured knowledge-free ceiling and the policy that attains it; the capacity and effective item count; the corrected score with its confidence interval; and an explicit statement of whether the reported accuracy is statistically distinguishable from the ceiling. A task that fails the last test should not appear in any aggregate. Under this rule, six of the seven tasks examined here would be withdrawn from the reliability claim and the remaining evidence would support only the modest statement that two checkpoints answer examination questions better than an exclusion heuristic.

### **B. Why measured guessing matters more in medicine**

Medical benchmarks are unusually prone to this problem for a structural reason. Safety-motivated task designs deliberately create items whose correct response is a refusal: recognise that no option is right, recognise that the question is nonsensical, recognise that the article does not exist. Refusal is the right clinical behaviour, and the design instinct is sound. But if every item of such a task has the same refusal as its gold answer, the task rewards a policy of universal refusal, which is exactly the behaviour that would make a CDS assistant useless. The abstention literature [24], [25] treats coverage and risk as a trade-off precisely because abstention is only valuable when it is selective; a benchmark that cannot detect indiscriminate abstention cannot evaluate the trade-off, and calibration analysis [21], [22], [23] on such a task is equally uninformative. Mixed-answerability designs, in which only a fraction of items warrant refusal and that fraction is not disclosed to the model, resolve this at essentially no annotation cost.

### **C. Relation to psychometric benchmarking**

Our results reinforce the psychometric programme of [17], [18], [19], [20] while identifying a precondition for it. Those methods estimate item difficulty, discrimination and guessing jointly from a panel of system responses. When only aggregate accuracies are published, that panel does not exist, and Section VI-E shows that a substitute panel of knowledge-free solvers cannot recover item difficulty. Measuring the guessing parameter from the artifact rather than estimating it from behaviour is what makes the analysis possible at all in this data regime, and it is a cheap check that any benchmark author can run before release.

## **VIII. LIMITATIONS**

This is a secondary analysis. We did not regenerate any model output, so our conclusions concern what the published numbers can support, not what the checkpoints would do under a different prompt, decoding temperature or scoring implementation; the original authors themselves report meaningful sensitivity to prompt framing. The knowledge-free ceiling for the reasoning task is a measured lower bound on what  $\Pi_0$  can achieve, and a stronger knowledge-free policy would lower capacity further, strengthening rather than weakening our conclusions; the ceilings for the other six tasks are exact and cannot move. The reference solver bank is non-neural by design, which is what makes it a clean knowledge-free control but also what prevents it from calibrating item difficulty on a scale relevant to LLMs. Ability estimates rest on the single task that retains capacity, so their absolute scale is not identified and only the ordering and the separability statement should be read. MedHALT-D inherits any annotation error, cultural bias or answer-key error present in the source examinations, and its repaired NSD component remains inadequate by our own criterion. Finally, none of these measurements addresses fairness, demographic robustness, explanation faithfulness, latency, privacy or human factors, all of which a clinical deployment would require.

## **IX. CONCLUSION**

We have argued that reliability claims about clinical LLMs should begin with a measurement-validity question rather than a ranking question: for each task, what could a system with no medical knowledge already score, and how much room does that leave? Answering it on the full public Med-HALT corpus shows that three of seven tasks are solved outright by constant strings, that 34.97% of the benchmark rows carry no Fisher information, that the effective test length is 26.1% of the nominal one, and that only 2 of 84 published model-task results are statistically distinguishable from a knowledge-free policy. Correcting for capacity reverses the leaderboard, moving the model that leads 77.60% of random task mixtures to 0.02%. We release a difficulty-

controlled repair built from the same real items that restores capacity on three of four reasoning components, and a design calculus that states how many items a medical reliability benchmark needs to support the claims made from it. Task-difficulty-aware evaluation, in this sense, is not an additional refinement to be applied after the leaderboard is built; it is the check that determines whether the leaderboard means anything.

## REFERENCES

- K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al., "Large language models encode clinical knowledge," *Nature*, vol. 620, pp. 172-180, 2023.
- T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepano, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, and V. Tseng, "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models," *PLOS Digital Health*, vol. 2, no. 2, e0000198, 2023.
- H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of GPT-4 on medical challenge problems," *arXiv:2303.13375*, 2023.
- Zhang, S. and Qiu, L. (2024). Data-driven assessment of pavement performance and resilience under extreme weather using the Long-Term Pavement Performance Program. *AI and Data Science Journal*, vol. 1, no. 1, pp. 57–66
- Zhang, H. (2024). Graph-Aware Multi-Task Learning For Early Prediction Of Timing Violations And Routing Congestion In Digital Ic Physical Design. *World Journal of Information Technology*, 2(1): 13-20.
- Wang, J., Tan, Y., Jiang, B., Wu, B., & Liu, W. (2025). Dynamic marketing uplift modeling: A symmetry-preserving framework integrating causal forests with deep reinforcement learning for personalized intervention strategies. *Symmetry*, 17(4), 610.
- Zhao, X., Liu, J., Wang, Y., & Wang, J. (2026). CryptoMamba-SSM: Linear complexity state space models for cryptocurrency volatility prediction. *IEEE Open Journal of the Computer Society*, 7, 226-243.
- Wang, J., Liu, J., Zheng, W., & Ge, Y. (2025). Temporal heterogeneous graph contrastive learning for fraud detection in credit card transactions. *IEEE Access*.
- Sun, T., Yang, J., Li, J., Chen, J., Liu, M., Fan, L., & Wang, X. (2024). Enhancing auto insurance risk evaluation with transformer and SHAP. *IEEE Access*, 12, 116546-116557.
- Zhang, H. (2025). Reinforcement Learning Approaches for Layout Optimization in Electronic Design Automation with Electromagnetic Compatibility Constraints. *Frontiers in Robotics and Automation*, 2(2), 77-93.
- Han, X., Yang, Y., Chen, J., Wang, M., & Zhou, M. (2025). Symmetry-aware credit risk modeling: A deep learning framework exploiting financial data balance and invariance. *Symmetry*, 17(3), 341.
- Chen, J., Cui, Y., Zhang, X., Yang, J., & Zhou, M. (2024). Temporal convolutional network for carbon tax projection: A data-driven approach. *Applied Sciences*, 14(20), 9213.
- Sun, T., Wang, M., & Chen, J. (2025). Leveraging machine learning for tax fraud detection and risk scoring in corporate filings. *Asian Business Research Journal*, 10(11), 1-13.

- Chen, J., Wang, M., & Sun, T. (2025). Intelligent tax systems and the role of natural language processing in regulatory interpretation. *American Journal of Machine Learning*, 6(4), 74-94.
- Chen, Z., Liu, J., & Chen, J. (2025). Machine Learning Methods for Financial Forecasting in Enterprise Planning: Transitioning from Rule-Based Models to Predictive Analytics. *Frontiers in Artificial Intelligence Research*, 2(3), 541-564.
- Chen, J., Liu, J., Liang, Y., & Zhou, M. (2026). KE-MLLM: A knowledge-enhanced multi-sensor learning framework for explainable fake review detection. *Applied Sciences*, 16(6), 2909.
- Yue, X., Yang, J., & Liu, W. (2026). FraudDebate-Agent: A Multi-Agent LLM Framework with an Evidence-Based Debate Mechanism for Financial Statement Fraud Detection. *Mathematics*, 14(15), 2695.
- Zhu, R. & Yue, X. (2024). An Explainable Machine Learning Framework For Predicting Dividend Sustainability And Financial Risk Of U.s. Reits Under Interest-Rate Shocks. *Journal of Trends in Finance and Economics*, 1(2): 48-57.
- Yue, X. & Zhu, R. (2024). An Explainable Hybrid Machine Learning Framework For Cash-Flow-Conditioned Multi-Horizon Liquidity Risk Early Warning: An Sme-Oriented Benchmark Study. *World Journal of Management Science*, 2(1): 63-72.
- Cui, Y., Han, X., Chen, J., Zhang, X., Yang, J., & Zhang, X. (2025). FraudGNN-RL: a graph neural network with reinforcement learning for adaptive financial fraud detection. *IEEE Open Journal of the Computer Society*, 6, 426-437.
- Zi, B. (2024). Cloud-native distributed systems for real-time payment intelligence. *AI and Data Science Journal*, 1(1), 51-56.
- Zhang, S. & Qiu, L. (2023). An Interpretable, Uncertainty-Aware Framework For Bridge Deterioration Prediction And Risk-Based Maintenance Prioritization. *Academic Journal of Architecture and Civil Engineering*, 1(2): 21-28.
- Jin, J., Xing, S., Ji, E., & Liu, W. (2025). Xgate: explainable reinforcement learning for transparent and trustworthy api traffic management in iot sensor networks. *Sensors*, 25(7), 2183.
- Xing, S., Wang, Y., & Liu, W. (2025). Multi-dimensional anomaly detection and fault localization in microservice architectures: A dual-channel deep learning approach with causal inference for intelligent sensing. *Sensors*, 25(11), 3396.
- Xing, S., & Wang, Y. (2025). Proactive data placement in heterogeneous storage systems via predictive multi-objective reinforcement learning. *IEEE Access*, 13, 117986-117998.
- Ji, E., Wang, Y., Xing, S., & Jin, J. (2025). Hierarchical reinforcement learning for energy-efficient API traffic optimization in large-scale advertising systems. *IEEE Access*.
- Xing, S., & Wang, Y. (2025). Cross-modal attention networks for multi-modal anomaly detection in system software. *IEEE Open Journal of the Computer Society*.
- Liu, Y., Ren, S., Wang, X., & Zhou, M. (2024). Temporal logical attention network for log-based anomaly detection in distributed systems. *Sensors*, 24(24), 7949.

- Ren, S., Jin, J., Niu, G., & Liu, Y. (2025). ARCS: Adaptive reinforcement learning framework for automated cybersecurity incident response strategy optimization. *Applied Sciences*, 15(2), 951.
- Li, P., Ren, S., Zhang, Q., Wang, X., & Liu, Y. (2024). Think4SCND: Reinforcement learning with thinking model for dynamic supply chain network design. *IEEE Access*, 12, 195974-195985.
- Hu, X., Zhao, X., Wang, J., & Yang, Y. (2025). Information-theoretic multi-scale geometric pre-training for enhanced molecular property prediction. *Plos one*, 20(10), e0332640.
- Wang, M., Zhang, X., Yang, Y., & Wang, J. (2025). Explainable machine learning in risk management: Balancing accuracy and interpretability. *Journal of Financial Risk Management*, 14(3), 185-198.
- Yang, J., Li, P., Cui, Y., Han, X., & Zhou, M. (2025). Multi-sensor temporal fusion transformer for stock performance prediction: An adaptive Sharpe ratio approach. *Sensors*, 25(3), 976.
- Zhao, X., Sun, T., Ren, S., Yang, J., & Liu, Y. (2025). RAG-Based AI Agents for Enterprise Software Development: Implementation Patterns and Production Deployment. *Frontiers in Artificial Intelligence Research*, 2(3), 501-520.
- Sun, T., Wang, M., & Chen, J. (2025). Leveraging machine learning for tax fraud detection and risk scoring in corporate filings. *Asian Business Research Journal*, 10(11), 1-13.
- Bian, M., Li, P., Lin, Y., Wang, Y., & Teng, D. (2026). Assessing Green Public Supply Chains Resilience under Carbon Neutrality Goals: A Multi-Agent Reinforcement Learning Framework. *IEEE Access*.
- Shang, W., Teng, D., Chen, T., Zhang, F., & Ding, J. (2026). Neuro-Elastic: A Unified Framework for Hardware-Aware Adaptive Quantization and Dynamic Sparsity in Real-Time Edge Intent Prediction. *IEEE Access*.
- Teng, D. (2025). TEAS: Token-and energy-aware autoscaling for cost-efficient LLM serving. *AI and Data Science Journal*, 6(3), 1-10.
- Qin, Y., Rhee, M., Teng, D., Zhang, C., & Zi, B. (2026). Neuro-symbolic constraint verification for LLM-driven internet finance transaction execution. *Scientific Reports*.
- Rhee, M., Zou, J., Mo, T., Teng, D., & Yang, J. S. (2026). Enhancing web search agents with self-play contrastive fine-tuning. *IEEE Access*.