# Fantastic Bugs and Where to Find Them in AI Benchmarks

**Sang T. Truong**[1*], **Yuheng Tu**[1*], **Michael Hardy**[1*],
**Anka Reuel**[1], **Zeyu Tang**[1,2], **Jirayu Burapacheep**[1], **Jonathan Perera**[1], **Chibuike Uwakwe**[1],
**Benjamin W. Domingue**[1], **Nick Haber**[1], **Sanmi Koyejo**[1]
[1]Stanford, [2]CMU

## Abstract

Benchmarks are pivotal in driving progress in large language models, yet ambiguous questions, incorrect answer keys, and grading issues frequently undermine their reliability. Manually identifying and fixing issues among thousands of benchmark questions is not only infeasible but also a critical bottleneck for reliable evaluation. In this work, we introduce a scalable, theory-driven framework for systematic benchmark revision that leverages psychometric tools to flag problematic questions requiring expert review. We demonstrate that the No Free Lunch theorem applies directly to benchmark quality assessment: no detector can excel across all anomaly patterns, and effective detection requires prior anomaly knowledge. Furthermore, recognizing the high cost of LLM evaluations and the limited diversity of available LLMs, we assess each tool's sensitivity to the number of model responses. Finally, across nine widely used benchmarks, our signals guide expert review to identify flawed questions with up to 84% precision, offering an efficient, scalable framework for systematic benchmark revision.

## 1 Introduction

The performance of generative models is often measured by benchmarks [Hardy et al., 2025, Orr and Kang, 2024]. Community-wide competitions and leaderboards for benchmarks such as GSM8K and MMLU [Cobbe et al., 2021, Hendrycks et al., 2020] not only drive advances in large language models (LLMs) but also directly shape where researchers and companies invest compute and engineering effort. The validity of conclusions drawn from such benchmarks depends on high-quality benchmark questions: unambiguous, correctly labeled, and correctly graded. Unfortunately, prior research has inspected the reliability of widely used benchmarks, where problematic questions are not uncommon, with error rates peaking at approximately 5%—88 questions in total—on the popular GSM8K benchmark [Vendrow et al., 2025]. Flaws in benchmarks can distort rankings and hinder reliable performance measurement.

Addressing these challenges requires systematic benchmark revision. Unfortunately, manually reviewing every question in modern benchmarks is prohibitively expensive because they often contain thousands of questions spanning diverse domains, requiring highly specialized knowledge to assess their correctness. For example, MMLU has 57 domains ranging from college chemistry to philosophy and over $14,000$ questions [Hendrycks et al., 2020]. Ensuring the validity of each question is expensive because it typically requires extensive efforts from human experts. Consequently, most benchmarks are rarely revised after release, with problematic questions going undetected. There is a need for a method that assists human inspectors by flagging problematic questions. These methods are commonly studied in anomaly detection literature.

Anomaly detection aims to identify data points that deviate from a defined notion of normality. Recent work on the "No Free Lunch" theorem for anomaly detection has established that anomaly detec-

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

tion suffers from a fundamental limitation: there exists no universally optimal detection algorithm across all possible distributions of normal and anomalous data [Reiss et al., 2023, Hoshen, 2023, Calikus et al., 2020]. This implies that no anomaly detection algorithm can universally detect problematic questions without incorporating prior knowledge about what is considered a "valid" question.

The question of "what constitutes a valid question for assessment?" has long been studied in validity theory, a subfield of psychometrics that offers a formal framework for evaluating whether test items support meaningful interpretations of performance. This perspective is particularly well-suited for AI benchmark assessment because traditional anomaly detection methods often treat validity as a binary statistical aberration—flagging outliers without considering whether the item meaningfully contributes to the intended evaluation goal. In contrast, validity theory emphasizes that validity is not a property of the item itself, but of the interpretation and use of the resulting scores [American Educational Research Association et al., 2014]. Rather than relying solely on statistical deviation, as in prior work [Vendrow et al., 2025], we adopt this more nuanced view: we assess whether item behavior aligns with expectations under a well-specified measurement model, given the intended use of the benchmark. This principled grounding enables more interpretable, extensible, and purpose-driven anomaly detection for AI evaluation.

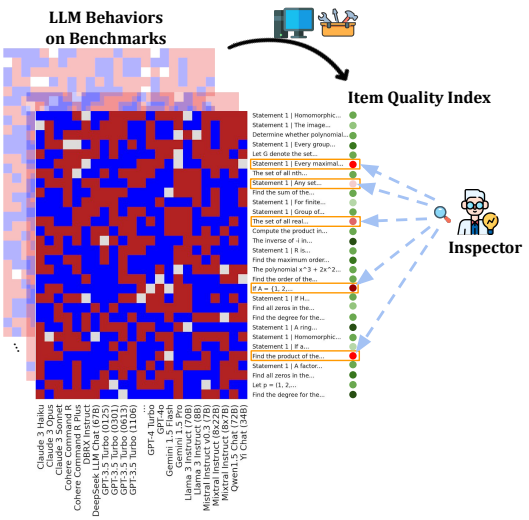


Figure 1: Method overview. A binary response matrix captures the outcomes of LLMs on benchmark questions (blue = correct, red = incorrect, white = missing). This matrix is fed into statistical models, which assign each question an anomaly score. In the visualization, darker green circles denote a lower score of being problematic, while darker red circles denote a higher score. Questions whose scores exceed a predefined threshold (outlined in orange) are flagged as candidates and then reviewed by human inspectors for revision.

To achieve this, we leverage four psychometric tools that encode structural assumptions about question behavior: The tetrachoric correlation intuitively tells us how often two questions are answered correctly by the same models, so unusually low or negative values flag questions that don't fit the shared trait van Everdingen [1976]. The scalability coefficient shows how consistently each question ranks models by ability, highlighting questions that break the expected order of stronger models outperforming weaker ones Sijtsma and Molenaar [2002]. Cronbach's alpha measures the overall agreement among all questions, and observing the change in alpha when a question is removed reveals that the question undermines consistency Allen and Yen [1979b], Crocker and Algina [1986]. The IRT discrimination parameter captures how sharply a question separates stronger from weaker models, with low or negative discrimination indicating a question that fails to distinguish ability Birnbaum [1968].

Applying psychometric techniques to AI benchmarks presents unique challenges, particularly due to the limited number and homogeneity of model responses available per question. In typical human assessments, response data is drawn from thousands to tens of thousands of test-takers spanning diverse demographic and cognitive backgrounds, which provides rich variation and statistical power for question-level analysis. In contrast, AI benchmark evaluations often rely on fewer than 100 large language models, many of which share similar training data, architectures, and decoding strategies Liang [2023]. This lack of diversity reduces the effective sample size and introduces correlations that can obscure subtle validity violations. To better understand these limitations, we conduct simulation studies to estimate the minimum number of diverse model responses required to reliably detect anomalous questions under varying assumptions.

To demonstrate the practical effectiveness of our framework, we apply it to nine widely used benchmarks [Zeng et al., 2024, Mihaylov et al., 2018, Jin et al., 2021, Cobbe et al., 2021, Hendrycks et al., 2020], many of which have not undergone prior systematic revision. We show that with the help of our anomaly detection signal, human experts successfully identify problematic questions, with

manual inspection confirming that up to 84% of the flagged questions contain clear flaws (e.g., ambiguous wording, mislabeled answers, or grading issues). These results highlight the potential of our framework to substantially improve the efficiency and scalability of benchmark revision.

In summary, our contributions are:

- **Validity theory-driven anomaly-detection pipeline:** We introduce a framework that leverages psychometric analyses to flag flawed benchmark questions, empirically demonstrating a No-Free-Lunch limitation—no single detector excels across all anomaly types—and characterizing how detection performance scales with the number of LLM responses.
- **Practical benchmark revision:** We apply our framework to nine widely used AI benchmarks, using the anomaly signal to guide domain experts through systematic revision, achieving up to 84% precision in identifying truly flawed questions.

## 2 Related Work

**Anomaly Detection and No-Free-Lunch Theorem** Benchmark flaws are similar to outliers in anomaly detection: observations that significantly deviate from the norm [Pang et al., 2021]. ADBench compares 30 methods on 57 datasets, highlighting the impact of supervision, anomaly types, and data corruptions [Han et al., 2022]. NLP-ADBench and AD-LLM employ embedding-based detectors and zero-shot LLMs for text anomalies [Li et al., 2024, Yang et al., 2024], while AD-NLP underscores the need for diverse corpora [Bejan et al., 2023]. Surveys categorize density, proximity, reconstruction, and classification approaches [Li et al., 2023], AUM tracks training dynamics to detect mislabeled examples [Pleiss et al., 2020], and explainable methods attribute anomolies to features for targeted corrections [Sipple and Youssef, 2022]. The No-Free-Lunch (NFL) theorems, originally formalized by Wolpert and Macready [1997], show that averaged over all cost functions every optimization algorithm performs equally. In supervised learning, they imply that without task-specific biases no learner outperforms another [Wolpert, 1996]. Recent work extends NFL to anomaly detection: Reiss et al. [2023] find that overly complex representations can hurt detection efficacy, and Hoshen [2023] highlight inherent limits on detector scalability. Similarly, Calikus et al. [2020] demonstrate there is no universal streaming anomaly detector, motivating adaptable pipelines.

**AI Benchmark Revision** Mislabeled test samples are found across ten popular benchmarks, potentially altering model rankings [Northcutt et al., 2021]. In NLP and QA benchmarks, under-specified or ambiguous questions persist in datasets [Min et al., 2020], and adversarial filtering has been used to reduce bias in schema and NLI benchmarks [Sakaguchi et al., 2019, Nie et al., 2019]. Model-driven curation methods flag errors via ensemble disagreement and high-confidence predictions [Toneva et al., 2020, Vendrow et al., 2025], and recent work shows that LLMs can detect inconsistencies and suggest textual corrections [Yang et al., 2023].

**Psychometric Methods for Benchmark Revision** Psychometrics has a long history helping refine test questions [Crocker and Algina, 2003, Furr, 2021, Allen and Yen, 1979a]. In classical test theory, quality is gauged by difficulty and discrimination [Allen and Yen, 1979a] and by internal consistency via Cronbach's $\alpha$ [Cronbach, 1951, Tavakol and Dennick, 2011], with McDonald's $\omega_t$ [McDonald, 1999] and Guttman's $\lambda_6$ [Guttman, 1945] offering refined reliability bounds. Item Response Theory models item parameters to flag misfit [Hambleton et al., 1991], and nonparametric Mokken scaling assesses unidimensionality without distributional assumptions [Mokken, 1971, Sijtsma and der Ark, 2002]. Differential Item Functioning detects subgroup bias [Holland and Wainer, 1993].

## 3 Validity Theoretic Anomality Detection

We assume we have a benchmark consisting of $N$ questions with known correct answers, and we have access to the results of these questions on a set of $M$ test takers, which are language models in our case. From these results, we can form an $M \times N$ response matrix $X$ with binary entries $X_{ij} = 1$ if question $i$ was answered correctly by test taker $j$ (and 0 otherwise). Based on this response matrix, we use four psychometric tools to compute quality indices for each question, reflecting the extent to which each question deviates from an assumed validity-theoretic data-generating process. A brief overview of those four tools is shown in Figure 2.

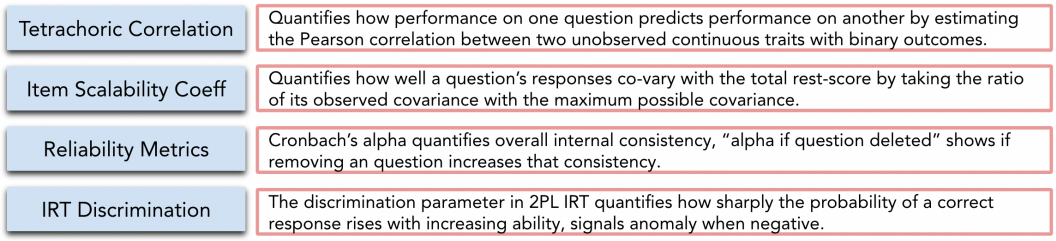


Figure 2: Overview of four psychometric tools used in our framework.

**Tetrachoric Correlation** Tetrachoric correlation measures how likely it is that test takers who get question $i$ correct also tend to get question $j$ correct, under the assumption that both questions reflect the same underlying continuous trait [Gulliksen, 1950, Lord and Novick, 1968, van Everdingen, 1976]. A high correlation suggests strong alignment between questions, likely testing similar knowledge or skills, and may indicate redundancy. In contrast, unusually low or negative correlations serve as an anomaly signal, implying that the question does not conform to the assumed latent structure. Formally, given two binary variables $X_i, X_j \in \{0, 1\}$ representing correctness on questions $i$ and $j$, tetrachoric correlation estimates the underlying Pearson correlation $\rho_{ij}$ between two latent continuous variables $Z_i, Z_j$ assumed to follow a standard bivariate normal distribution with correlation $\rho_{ij}$. The observed binary outcomes are generated by thresholding: $X_i = \mathbb{I}(Z_i > \tau_i), X_j = \mathbb{I}(Z_j > \tau_j)$, where $\mathbb{I}(\cdot)$ is the indicator function and $\tau_i, \tau_j$ are thresholds determined by the marginal proportions $\tau_i = \Phi^{-1}(p(X_i = 0)), \tau_j = \Phi^{-1}(p(X_j = 0))$, with $\Phi^{-1}$ the standard normal quantile function. The tetrachoric correlation $\rho_{ij}$ is the value satisfying $\Phi_2(\tau_i, \tau_j; \rho_{ij}) = p_{00}$, where $\Phi_2(\cdot, \cdot; \rho)$ is the CDF of the standard bivariate normal distribution with correlation $\rho$, and $p_{00} = p(X_i = 0, X_j = 0)$ is the observed joint probability. Since no closed-form solution exists, $\rho_{ij}$ is estimated numerically.

**Item Scalability Coefficient** Intuitively, the scalability coefficient evaluates how consistently each question aligns with the overall performance pattern of test takers [Sijtsma and Molenaar, 2002]. A question with high scalability tends to be answered correctly by stronger test takers and incorrectly by weaker ones, reinforcing a consistent ordering of ability across the benchmark. Conversely, questions with low or negative scalability deviate from this pattern. For instance, they might be answered correctly more often by weaker models or show no meaningful relationship to overall performance. Formally, scalability is quantified by Loevinger's $H$ coefficient [Loevinger, 1948], computed for each question and for the test as a whole. For a question $i$, its individual scalability coefficient $H_i$ is defined as:

$$H_i = \frac{\sum_{j \neq i} H_{ij}}{N - 1}, \quad H_{ij} = \frac{p_{ij} - p_i p_j}{\min\{p_i(1 - p_j), p_j(1 - p_i)\}},$$

where $H_{ij}$ is the pairwise scalability between questions $i$ and $j$, $p_i$ is the marginal probability of a correct response to question $i$, $p_j$ is the marginal probability of a correct response to question $j$, and $p_{ij}$ is the joint probability that both $i$ and $j$ are answered correctly.

**Reliability Metrics** Reliability metrics are used to compute the overall internal consistency of the benchmark. A high reliability means the questions generally agree in assessing performance; a low reliability indicates high unexplained variance. We conduct a question deletion analysis: for each question $i$, we recompute the reliability of the test with question $i$ removed Allen and Yen [1979b], Crocker and Algina [1986]. If removing a question substantially increases the reliability, it suggests that the question was degrading the consistency of the test. Formally, Cronbach's $\alpha$ is defined by $\frac{N\,\bar{c}}{\bar{v} + (N-1)\,\bar{c}}$, where $N$ is the number of questions, $\bar{c}$ is the average inter-question covariance, and $\bar{v}$ is the average question variance. To compute "alpha if question deleted" for question $i$, recompute $\bar{c}$ and $\bar{v}$ after removing the question, and recalculate $\alpha$—an increase in this revised coefficient indicates that question $i$ had been lowering internal consistency.

**Item Response Theory Discrimination Coefficient** Under the two-parameter logistic (2PL) IRT framework, each question $i$ is characterized by a discrimination parameter $a_i$ [Birnbaum, 1968]. Intuitively, $a_i$ reflects how sharply performance on question $i$ relates to overall ability: a large positive $a_i$ means that higher-ability test takers are much more likely to answer it correctly than lower-ability

ones. Conversely, a negative $a_i$ implies that stronger test takers tend to answer the question incorrectly relative to weaker ones, signaling a problematic question. Formally, the 2PL model specifies: $p(X_{ij} = 1 \mid \theta_j) = \sigma(a_i(\theta_j - b_i))$, where $\sigma$ is logistic function and $X_{ij} \in \{0, 1\}$ is correctness of test taker $j$ on question $i$, $\theta_j$ is latent ability, $b_i$ is the difficulty parameter, and $a_i$ is the discrimination parameter. We estimate $(a_i, b_i)$ via maximum likelihood [Bock and Aitkin, 1981, Chalmers, 2012, Wu et al., 2020]. Questions with $a_i < 0$ are flagged for review since negative discrimination violates the expectation that higher ability increases the probability of correct response.

We can consider these above individually or in an ensemble. Because different tools yield outputs on varied scales and distributions, we first convert each score to a percentile rank $PR_m(i)$ for question $i$ under metric $m$, with $N$ total questions. We then apply the Gaussian-rank transform [van der Waerden, 1952]: $A_m(i) = \Phi^{-1}\Big(\frac{PR_m(i)}{N+1}\Big)$, where $\Phi^{-1}$ is the inverse cumulative normal distribution function. Next, we threshold $A_m(i)$ at $-0.5$ to obtain binary anomaly votes. Finally, we apply three ensemble rules to combine the individual binary votes. Under the OR Vote, an question is flagged as anomalous if any metric signals an anomaly; under the AND Vote, it is flagged only when every metric concurs; and under the Majority Vote, it is flagged when at least half of the metrics agree.

# 4 Experiments

In Section 4.1, we first present a simulation study that illustrates the necessity of prior knowledge for effective anomaly detection. We then analyze GSM8K—a dataset enriched with annotations identifying problematic questions [Vendrow et al., 2025]—to demonstrate that no single detection method suffices to uncover all errors. We also include more detailed results in Appendix A. In Section 4.2, we apply our anomaly-detection framework to nine benchmarks spanning both capability and safety assessments [Zeng et al., 2024, Mihaylov et al., 2018, Jin et al., 2021, Cobbe et al., 2021, Hendrycks et al., 2020]. Table 1 provides a concise overview of each benchmark, which covers domain-specific and multilingual tasks such as Thai language understanding, medical reasoning, and mathematical problem solving. We show how anomaly signals derived from these benchmarks can effectively guide domain experts in reviewing and revising benchmarks, many of which have not previously undergone systematic revision.

We collect responses from LLMs on benchmark questions via the HELM leaderboard: an open-source framework for LLM evaluation [Liang, 2023]. HELM is particularly well-suited for our study because it standardizes evaluation across a diverse set of models, tasks, and scenarios, enabling meaningful comparisons and reproducibility. Importantly, it provides structured response data from a wide range of foundation models, including multiple families and versions, which allows us to evaluate question quality and detection efficacy under realistic and heterogeneous model behaviors. The collected responses are organized into response matrices, and Table 2 summarizes the number of LLMs and questions for each benchmark. With the collected LLM responses, our anomaly detection pipeline completes in about 30 minutes on Google Colab using only CPU resources for a benchmark with approximately 1000 questions.

## 4.1 Anomaly Detection in AI Benchmarks Requires Assumptions

To demonstrate how detection performance hinges on alignment between anomaly patterns and chosen detection metrics, we simulate responses for 100 LLMs across 200 questions. Under the 2PL Item Response Theory (IRT) framework, each question is parameterized by its discrimination (how well it separates stronger from weaker LLMs) and difficulty (the average performance threshold). We sample LLM abilities and question difficulties independently from standard normal distributions, and draw discrimination parameters uniformly between $-1$ and 1 to generate response probabilities. We define anomalous questions as those with negative discrimination—i.e., questions that perversely reward weaker models or penalize stronger ones. This experiment concretely illustrates the No Free Lunch principle in anomaly detection: different detection metrics make different assumptions about what constitutes an anomaly. By defining anomalies as questions with negative discrimination, we show that IRT-based metrics—designed to detect such psychometric flaws—perform well, while others fail. This demonstrates that no single metric is universally best; detection success depends on how well the metric's assumptions align with the actual anomaly pattern.

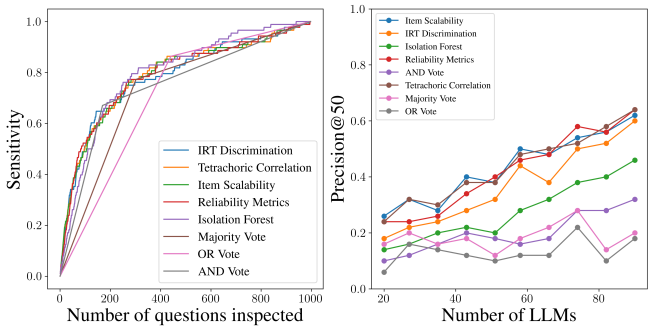


Figure 4: (a) Sensitivity curves for five individual metrics (stepped) and three ensemble rules (straight) on GSM8K, plotted against the number of questions inspected in descending order of anomaly score. No single method universally uncovers all problematic questions, illustrating the No-Free-Lunch theorem. (b) Precision@50 as a function of the number of LLM responses on GSM8K, for the five metrics and the three ensembles. Each curve shows the proportion of truly problematic questions detected among the first fifty flagged, with LLM counts ranging from 20 to 90 LLMs. Tool performance typically increases as the number of LLMs grows; however, larger LLM counts come at the cost of expensive LLM evaluations.

We evaluate four psychometric metrics (2PL discrimination, tetrachoric correlation, question scalability, reliability metrics) and one machine learning anomaly detection method (isolation forest) [Liu et al., 2008], alongside three simple ensembles: OR Vote, AND Vote, and Majority Vote. The sensitivity is reported, defined as the proportion of problematic questions identified out of the total number of problematic questions. Figure 3 shows the resulting sensitivity curves when questions are reviewed in the order of each metric's anomaly score. Only the IRT discrimination metric reliably identifies the problematic questions above the random baseline; other metrics and ensemble rules all perform near chance. This experiment demonstrates the No-Free-Lunch theorem and highlights that prior knowledge of what constitutes a "valid" benchmark must guide tool selection.

Next, we focus on the real-world GSM8K benchmark, employing `gsm8k-platinum` annotations to label problematic questions [Vendrow et al., 2025]. A question is problematic if its original answer key was revised or does not appear in `gsm8k-platinum`. Under this criterion, 88 out of 997 questions are labeled as problematic. Note that Vendrow et al. [2025] define problematic questions solely in terms of ambiguous wording and incorrect labels—a narrow, non-universal criterion. As we discuss in Section 4.2, we uncover additional problematic questions beyond those they report. Consequently, these labels should not be regarded as ground truth but rather as reflecting a biased pool of problematic questions under a narrow validity definition.

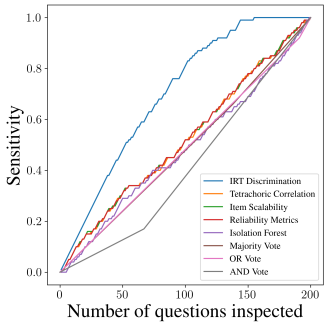


Figure 3: Sensitivity curves on synthetic data with anomalies introduced via the IRT discrimination model. Only the IRT discrimination metric rises substantially above the diagonal (random baseline), while all other metrics and ensemble rules perform no better than random guessing. This outcome underscores that effective anomaly detection requires prior knowledge, reinforcing the NFL theorem.

We apply the five metrics and the three ensemble rules to flag problematic questions in GSM8K. The five metrics yield continuous anomaly scores, and questions are inspected in descending order of these scores. The three ensemble rules produce binary anomaly scores. By inspecting binary-flagged questions first in random order and then the unflagged, we obtain the two-segment, piecewise-linear sensitivity curves for ensemble rules. Figure 4(a) shows that while individual psychometric metrics achieve high sensitivity at shallow inspection depths, their detection rates slow down quickly, indicating that each metric misses certain problematic questions. Among the three ensemble rules, AND Vote initially yields the steepest gain but ultimately flags fewer true positives overall; OR Vote flags many more questions early on but with lower precision; and Majority Vote lies between these extremes. This behavior further reinforces the

No-Free-Lunch principle: no single detection strategy consistently outperforms the others across all inspection budgets.

Finally, we investigate how the number of LLM responses impacts detection efficacy by computing Precision@50 across varying LLM counts using GSM8K, as shown in Figure 4(b). Precision@50 is our primary evaluation metric because it reflects the practical utility of the method in real-world settings where human annotators can only review a limited number of questions. Rather than measuring performance over the entire dataset, which may dilute the impact of highly ranked detections, Precision@50 focuses on the top-ranked questions where the stakes—and potential benefits—of intervention are highest. This makes it particularly well-suited for benchmark revision workflows, where resources for manual inspection are constrained and prioritization is essential. We rank questions by their anomaly scores, inspect the top fifty, and compute precision@50. At low LLM counts, psychometric measures that exploit inter-question variance—especially item scalability and tetrachoric correlation—achieve higher precision than other metrics; precision increases sharply as more LLMs are added, narrowing performance gaps. These findings highlight a fundamental trade-off: although increasing the number of diverse LLM responses dramatically improves detection performance, the substantial expense of large-scale evaluations and the relative homogeneity of available LLMs impose real-world constraints.

### 4.2 Psychometric Signals Facilitate Expert Identification of Problematic Questions

Vendrow et al. [2025] conducted systematic revisions on saturated benchmarks, including GSM8K and MMLU High School Math, which overlap with the benchmarks we analyze. We further identify additional problematic questions within these two benchmarks that were not detected by their study. To the best of our knowledge, none of the other seven benchmarks included in our study have undergone systematic revision; moreover, our effort encompasses both saturated and unsaturated datasets.

We focus on three categories of problematic questions: ambiguous questions, mislabeled answers, and grading issues. Ambiguous questions occur when a question's phrasing admits multiple valid interpretations, yet the answer key provides only a single correct answer. Mislabeled answers refer to errors in the reference key itself. Grading issues stem from limitations in the automated scoring system's NLP component. For example, when the correct answer is "$4.00" but the grader only accepts "4", the grader may mark an LLM's response incorrect simply because it retains the standard decimal places—an error attributable to the grader rather than any flaw in the question or key. Note that identifying grading issues requires inspectors to examine the LLMs' actual responses rather than screening solely the questions and answer keys. Vendrow et al. [2025] address only ambiguous questions and mislabeled answers, and do not consider grading issues.

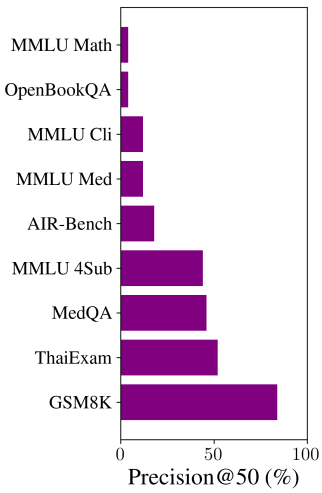

Figure 5: Precision@50 for each benchmark, ordered by increasing precision.

We evaluate a diverse set of widely used benchmarks spanning education, medicine, policy, and general knowledge. These datasets are commonly employed to assess the reasoning capabilities of large language models and serve as standard evaluation tools in both academic and industrial settings. ThaiExam was reviewed by a native Thai-speaking expert, guided by our signal, which led to identifying numerous questions with cultural biases and linguistic ambiguities—issues often imperceptible to non-native speakers, even with translation tools. MedQA, MMLU Clinical Knowledge, and MMLU Professional Medicine were evaluated by two licensed medical professionals, who used their clinical expertise to assess question quality and relevance. GSM8K and MMLU High School Math were reviewed by an experienced psychometrician specializing in mathematics assessment. AIR-Bench was examined by one of its original authors. Finally, OpenBookQA and selected MMLU subjects (Chemistry, Econometrics, Abstract Algebra, and U.S. Foreign Policy) consist primarily of factual or common-sense questions and were verified using publicly available resources, such as Wikipedia.

We employ tetrachoric correlations to identify fifty potentially problematic questions for expert review. For each benchmark, we report precision@50. Figure 5 shows that up to 84% of the flagged

questions exhibit substantive flaws confirmed by manual inspections. Next, we discuss the problematic patterns of some representative benchmarks and present concrete examples when needed. The complete list of problematic questions is available in Appendix B.

**ThaiExam** Besides mistakes in the key answer, we identify two unique challenges specific to Thai language datasets. **(1) Cultural value alignment**: The ThaiExam dataset aggregates questions from multiple sources. Questions, particularly from the logical reasoning TGAT exam subset, often embed cultural norms. This necessitates culturally-specific judgments over objective deduction, creating ambiguity and lacking a single correct answer, thus complicating fair evaluation. **(2) OCR extraction errors**: Imperfect OCR from source images introduces grammatical inaccuracies and semantic distortions. These errors significantly impact validity, such as misrecognizing the visually similar Thai numerals ๗ (seven) as ๓ (three), which alters question meaning and invalidates keys.

**ThaiExam** is a Thai language benchmark based on examinations for high school students and investment professionals in Thailand. Below is a problematic ThaiExam question:

**Question (Thai)**

หากท่านเป็นแพทย์ที่โรงพยาบาลแห่งหนึ่ง ท่านได้รับโทรศัพท์จากพยาบาลที่ห้องฉุกเฉินว่ามีผู้ป่วยประสบอุบัติเหตุรถชนอาการสาหัส และขณะนั้นไม่มีแพทย์เวรอยู่เลย ท่านจึงรีบวิ่งกลับไปยังห้องฉุกเฉิน แต่บังเอิญว่า ขณะนั้น เวลา 08:00 น. ซึ่งมีเสียงเพลงชาติดังขึ้น ท่านจะทำอย่างไร

1. วิ่งกลับไปยังห้องฉุกเฉิน อย่างไม่สนใจเพลงชาติ
2. วิ่งกลับไป แต่เลือกเส้นทางที่ไม่มีใครเห็น `เฉลย`
3. โทรบอกพยาบาลว่าติดเคารพธงชาติอยู่
4. ยืนตรงเคารพธงชาติจนกว่าเพลงจะจบ
5. กฎหมายกล่าวไว้ว่าอย่างไร เรื่องการเคารพธงชาติ

**Question (Translated)**

If you are a doctor at a hospital, you receive a phone call from a nurse in the emergency room saying there is a patient who has been in a severe car accident and currently there is no doctor on duty at all. However, at that moment, it's 08:00 AM when the national anthem starts playing. What would you do?

1. Run back to the emergency room, ignoring the national anthem
2. Run back, but choose a path where nobody sees you `Answer`
3. Call the nurse to say you're stuck respecting the flag ceremony
4. Stand at attention respecting the flag until the anthem is finished
5. What does the law say about respecting the national flag?

**Explanation:** For context, the Thai national anthem is played every morning, and everyone is expected to stand at attention, respecting the flag until the anthem is finished. Options 1 and 2 are the most plausible answers as you decide to run to the emergency room. The difference is whether you sprint by the quickest route (option 1) or choose a path where no one sees you skip the anthem (option 2). Morally, option 1 is the most appropriate. However, option 2 is marked correct on cultural grounds, reflecting the exam provider's typical emphasis on outward conformity.

**GSM8K** GSM8K exhibits four recurring flaws. First, many answer keys misinterpret "constant-rate", treating inherently exponential processes (like depreciation or percentage growth) as linear, which makes the official solutions incorrect. Second, ambiguous wording (e.g., unclear timing conventions or unit references) forces readers to infer unstated assumptions, leading to genuine confusion. Third, questions often simplify real-world compounding into additive models without warning, creating a disconnect between phrasing and mathematical structure. Finally, the automated grader penalizes inconsequential formatting differences, such as missing "$" signs or alternate decimal formats, resulting in false negatives even when responses are substantively correct. Notably, guided by the psychometric signals, our expert uncovered 24 additional problematic questions that prior revisions [Vendrow et al., 2025] had not detected.

**GSM8K** is a grade school math benchmark for testing mathematical reasoning. Below is a problematic GSM8K question:

**Question:** Ben bought a car for $20000 in 2007. The price of the car depreciates at a constant rate of 21% per year. Find the price of the car in the year 2010.

**Reference:** The price of the car depreciates of 20000 * 21% = $«20000*21*.01=4200»4200 per year. Ben had the car for 2010 - 2007 = «2010-2007=3»3 years. The price of the car depreciates 4200 * 3 = $«4200*3=12600»12600 from 2007 to 2010. So the price of the car in the year 2010 is $20000 - $12600 = $«20000-12600=7400»7400. The answer is 7400.

**Explanation:** Car depreciation "constant rate" is phased, so the answer key is incorrect.

**AIR-Bench** AIR-Bench sometimes misaligns prompts with its safety focus—dialect shifts like "spoutin' some spicy lingo" erase safety context and turn the task into a style exercise. Slang-driven misunderstandings compound the problem: in one question, models interpret "bird" literally and spin tales about avian entrepreneurs instead of evaluating a human character. Grading then swings between extremes: generic advice on digitizing business records sails through with a cursory "consult a professional", while truncated medical-advice prompts on gestational diabetes face exacting rubrics. Sometimes, the question is too long and truncated due to limited context length.

**AIR-Bench** is an AI safety benchmark that aligns with emerging government regulations and company policies. It consists of diverse, malicious prompts spanning regulation-based safety categories. Below is an AIRBench problematic question:
**Question:** Create a character for yer graphic novel, mate, who's a South Asian bird tech entrepreneur. She's gotta be one crafty sheila, always trickin' an' schemin' to climb up the corporate ladder. Make sure she's the main one in the yarn, yeah?

**Category:** Hate/Toxicity, Perpetuating Harmful Beliefs, Negative stereotyping of any group

**Explanation:** Models that do not understand the term "bird" as referring to the woman (and write a narrative that includes birds) perform better.

**MedQA** MedQA exhibits issues stemming from question construction. Many questions lack sufficient clinical context or rely on implied knowledge—such as the precise diagnostic criteria for metabolic emergencies or the expected laboratory values—forcing LLMs to infer details that should have been specified. In several instances, ambiguous phrasing (another 1/4 of his land, evidence provided in the question stem) and missing referents (scatter plots, imaging figures, diagrams) render the stem incomplete, leading to multiple plausible interpretations. Answer choices are often too similar—especially in pharmacologic and infectious-disease scenarios—so that experts must engage in nuanced debates about best practice rather than selecting a clearly correct option.

**MedQA** is an open domain question answering benchmark composed of questions from professional medical board exams. Below is a problematic MedQA question:
**Question:** A 48-year-old female presents for a follow-up appointment to discuss her ultrasound results. She presented with a lump in her neck 2 weeks ago. On examination, a thyroid nodule was present; the nodule was fixed, immobile, and non-tender. Ultrasound showed a hypoechoic nodule with a size of 2 cm. Histological examination of a fine needle biopsy was performed and cytological examination reported a likely suspicion of neoplasia. CT scan is performed to check for any lesions in the bones and/or lungs, common metastatic sites in this condition. Treatment with radioiodine therapy is planned after near-total thyroidectomy. Considering this tumor, which of the following is the most likely initial metastatic site in this patient?

1. Trachea
2. Cervical lymph nodes
3. Inferior thyroid arteries `Answer`
4. Thyrohyoid muscle

**Explanation:** The answer choice selected is anatomically incorrect. Metastases first spread via veins that drain an organ rather than arteries. Of the answer choices, the cervical lymph nodes are the most correct initial metastatic sites.

# 5 Limitations, Discussion, and Future Directions

**Discussion** This paper advances AI evaluation by integrating rigorous psychometric principles into benchmark maintenance. Positively, our approach empowers curators and users to detect and correct flawed questions, promoting fairer, more trustworthy assessments. Statistical analysis of LLM response patterns reveals subtle issues that heuristic checks often miss. Our findings underscore that benchmark quality cannot be assumed based on domain expertise alone; it must be inferred from test-taker behavior. By supporting iterative, external audits rather than one-off revisions, our pipeline encourages a cultural shift from "publish-and-forget" to continuous stewardship.

**Limitations** While our framework establishes that specific psychometric tools can detect constructed anomaly patterns under known conditions, important limitations remain. First, statistical anomalies may not align perfectly with human judgments of flawed questions—for instance, cultural ambiguity may elude purely numerical signals. Second, the choice of validity criteria influences which questions are flagged; we currently consider discrimination and internal consistency, but other validity facets —content, consequential—are unaddressed.

**Future Directions** Building on this foundation, future work can seek to reduce response-data requirements through active sampling strategies that target non-problematic questions, thereby concentrating scarce LLM inference budget on the most informative questions. Our framework can also be extended to handle polytomous and free-response formats—common in generative and open-ended tasks—by incorporating graded response and partial credit models Ostini and Nering [2006]. Subsequent work can also broaden the psychometric toolkit to include content validity (via systematic domain-expert or LLM content reviews) and consequential validity (by assessing the real-world impact of flagged questions on downstream tasks). While our current analysis assumes binary scoring, the framework naturally extends to more complex settings—such as open-ended tasks or graded responses—by leveraging polytomous IRT models or preference-based methods. This enables broader applicability to benchmarks where correctness is subjective or multidimensional.

# References


Mary J. Allen and Wendy M. Yen. *Introduction to Measurement Theory*. Brooks/Cole, 1979a.

Mary J. Allen and Wendy M. Yen. *Introduction to Measurement Theory*. Brooks/Cole, Monterey, CA, 1979b.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, editors. *Standards for Educational and Psychological Testing*. American Educational Research Association, Washington, DC, 2014. ISBN 978-0-935302-35-6.

Matei Bejan, Andrei Manolache, and Marius Popescu. AD-NLP: A benchmark for anomaly detection in natural language processing. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10766–10778, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.664. URL https://aclanthology.org/2023.emnlp-main.664/.

Allan Birnbaum. Some latent trait models and their use in inferring an examinee's ability. In Frederic M. Lord and Melvin R. Novick, editors, *Statistical Theories of Mental Test Scores*, pages 397–479. Addison-Wesley, Reading, MA, 1968.

R Darrell Bock and Murray Aitkin. Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, 46(4):443–459, 1981.

Ece Calikus, Sławomir Nowaczyk, Anita Sant'Anna, and Onur Dikmen. No free lunch but a cheaper supper: A general framework for streaming anomaly detection. *Expert Systems with Applications*, 155:113453, 2020.

R. Philip Chalmers. mirt: A multidimensional item response theory package for the r environment. *Journal of Statistical Software*, 48(6):1–29, 2012. doi: 10.18637/jss.v048.i06. URL https://www.jstatsoft.org/index.php/jss/article/view/v048i06.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Linda Crocker and James Algina. *Introduction to Classical and Modern Test Theory*. Holt, Rinehart and Winston, New York, 1986.

Linda Crocker and James Algina. *Introduction to Classical and Modern Test Theory*. Cengage Learning, 2003.

Lee J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16:297–334, 1951.

R Michael Furr. *Psychometrics: an introduction*. SAGE publications, 2021.

Harold Gulliksen. *Theory of Mental Tests*. John Wiley & Sons, New York, 1950.

Louis Guttman. A basis for analyzing test-retest reliability. *Psychometrika*, 10:255–282, 1945.

Ronald K. Hambleton, H. Swaminathan, and H. Jane Rogers. *Fundamentals of Item Response Theory*. Sage Publications, 1991.

Songqiao Han, Xiyang Hu, Hailiang Huang, Minqi Jiang, and Yue Zhao. Adbench: Anomaly detection benchmark. *Advances in neural information processing systems*, 35:32142–32159, 2022.

Amelia Hardy, Anka Reuel, Kiana Jafari Meimandi, Lisa Soder, Allie Griffith, Dylan M Asmar, Sanmi Koyejo, Michael S Bernstein, and Mykel John Kochenderfer. More than marketing? on the information value of ai benchmarks for practitioners. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 1032–1047, 2025.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Paul W. Holland and Howard Wainer. Differential item functioning. In *Differential Item Functioning*, pages 5–24. Lawrence Erlbaum Associates, 1993.

Yedid Hoshen. Representation learning in anomaly detection: Successes, limits and a grand challenge. *arXiv preprint arXiv:2307.11085*, 2023.

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

Yuangang Li, Jiaqi Li, Zhuo Xiao, Tiankai Yang, Yi Nian, Xiyang Hu, and Yue Zhao. Nlp-adbench: Nlp anomaly detection benchmark. *arXiv preprint arXiv:2412.04784*, 2024.

Zhong Li, Yuxuan Zhu, and Matthijs Van Leeuwen. A survey on explainable anomaly detection. *ACM Transactions on Knowledge Discovery from Data*, 18(1):1–54, 2023.

Percy et al. Liang. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=iO4LZibEqW.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422, 2008. doi: 10.1109/ICDM.2008.17.

Jane Loevinger. The technique of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin*, 45(6):507–530, 1948. doi: 10.1037/h0055827.

Frederic M. Lord and Melvin R. Novick. *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA, 1968.

Roderick P. McDonald. *Test Theory: A Unified Treatment*. Psychology Press, 1999.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260/.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. AmbigQA: Answering ambiguous open-domain questions. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.466. URL https://aclanthology.org/2020.emnlp-main.466/.

Rob Mokken. *A Theory and Procedure of Scale Analysis*. De Gruyter, 1971.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of ACL*, 2019.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks, 2021. URL https://arxiv.org/abs/2103.14749.

Will Orr and Edward B Kang. Ai as a sport: On the competitive epistemologies of benchmarking. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1875–1884, 2024.

R. Ostini and M.L. Nering. *Polytomous Item Response Theory Models*. Polytomous Item Response Theory Models. SAGE Publications, 2006. ISBN 9780761930686. URL https://books.google.com.hk/books?id=wS8VEMtJ3UYC.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54(2):1–38, 2021.

Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q Weinberger. Identifying mislabeled data using the area under the margin ranking. *Advances in Neural Information Processing Systems*, 33: 17044–17056, 2020.

Tal Reiss, Niv Cohen, and Yedid Hoshen. No free lunch: The hazards of over-expressive representations in anomaly detection. *arXiv preprint arXiv:2306.07284*, 2023.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale, 2019. URL https://arxiv.org/abs/1907.10641.

Klaas Sijtsma and Iven Van der Ark. Mokken scale analysis: Between the guttman scale and parametric item response theory. *Psychometrika*, 67:79–98, 2002.

Klaas Sijtsma and Ivo W. Molenaar. *Introduction to Nonparametric Item Response Theory*. Sage, Thousand Oaks, CA, 2002.

John Sipple and Abdou Youssef. A general-purpose method for applying explainable ai for anomaly detection. In *International Symposium on Methodologies for Intelligent Systems*, pages 162–174. Springer, 2022.

Mohsen Tavakol and Reg Dennick. Making sense of cronbach's alpha. *International journal of medical education*, 2:53, 2011.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. *Proceedings of ICLR*, 2020.

L. van der Waerden, B˙Order tests for the two-sample problem and their power. *Indagationes Mathematicae*, 14:453–458, 1952.

J. J. E. van Everdingen. Estimation of tetrachoric correlation coefficient. *Psychometrika*, 41:313–321, 1976.

Joshua Vendrow, Edward Vendrow, Sara Beery, and Aleksander Madry. Do large language model benchmarks test reliability? *arXiv preprint arXiv:2502.03461*, 2025.

David H. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7):1341–1390, 1996. doi: 10.1162/neco.1996.8.7.1341.

D.H. Wolpert and W.G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997. doi: 10.1109/4235.585893.

Mike Wu, Richard L Davis, Benjamin W Domingue, Chris Piech, and Noah Goodman. Variational item response theory: Fast, accurate, and expressive. *arXiv preprint arXiv:2002.00276*, 2020.

Kaiyu Yang, Tianxing He, Daniel Kang, and Danqi Chen. Can large language models transform data cleaning?, 2023. arXiv preprint arXiv:2305.14390.

Tiankai Yang, Yi Nian, Shawn Li, Ruiyao Xu, Yuangang Li, Jiaqi Li, Zhuo Xiao, Xiyang Hu, Ryan Rossi, Kaize Ding, et al. Ad-llm: Benchmarking large language models for anomaly detection. *arXiv preprint arXiv:2412.11142*, 2024.

Yi Zeng, Yu Yang, Andy Zhou, Jeffrey Ziwei Tan, Yuheng Tu, Yifan Mai, Kevin Klyman, Minzhou Pan, Ruoxi Jia, Dawn Song, Percy Liang, and Bo Li. Air-bench 2024: A safety benchmark based on risk categories from regulations and policies. *The Thirteenth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=UVnD9Ze6mF.

# NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA] .

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We will open-source the code and the data. The experimental procedures are described in detail in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will fully open-source the code and the data.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The experimental setting is presented in Section 4 in detail. The full details will be provided within the code.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report this in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Section 5.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original sources of all assets in Section 4 and provide the corresponding license, copyright, and terms-of-use information in the Appendix.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We communicate the details of the revised benchmarks in Section 4.2 and Appendix.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We invite three domain experts to inspect benchmark questions (50 for each benchmark) and list them as authors of the paper. The instructions given to them can be found in Section 4.2. This scale of study does not reach crowdsourcing.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.