
Benchmarking is Broken - Don't Let AI be its Own Judge

Zerui Cheng^{1,*} Stella Wohngig^{2,*} Ruchika Gupta^{3,*} Samiul Alam^{4,*}

Tassallah Abdullahi⁵ João Alves Ribeiro⁶ Christian Nielsen-Garcia⁷ Saif Mir⁴

Siran Li⁸ Jason Orendor⁹ Seyed Ali Bahrainian⁸ Daniel Kirste¹⁰

Aaron Gokaslan¹¹ Mikołaj Glinka¹² Carsten Eickhoff^{8,†} Ruben Wolff^{12,†}

¹ Princeton University ² CISPA Helmholtz Center for Information Security

³ Michigan State University ⁴ Ohio State University ⁵ Brown University

⁶ Massachusetts Institute of Technology ⁷ University of California, Los Angeles

⁸ University of Tübingen ⁹ Old Dominion University ¹⁰ Technical University of Munich

¹¹ Cornell University ¹² Forest AI

* Equal Contributions. † Advisors.

zerui.cheng@princeton.edu

Abstract

The meteoric rise of Artificial Intelligence (AI), with its rapidly expanding market capitalization, presents both transformative opportunities and critical challenges. Chief among these is the urgent need for a new, unified paradigm for trustworthy evaluation, as current benchmarks increasingly reveal critical vulnerabilities. Issues like data contamination and selective reporting by model developers fuel hype, while inadequate data quality control can lead to biased evaluations that, even if unintentionally, may favor specific approaches. As a flood of participants enters the AI space, this "Wild West" of assessment makes distinguishing genuine progress from exaggerated claims exceptionally difficult. Such ambiguity blurs scientific signals and erodes public confidence, much as unchecked claims would destabilize financial markets reliant on credible oversight from agencies like Moody's.

In high-stakes human examinations (e.g., SAT, GRE), substantial effort is devoted to ensuring fairness and credibility; why settle for less in evaluating AI, especially given its profound societal impact? **This position paper argues that a laissez-faire approach is untenable. For true and sustainable AI advancement, we call for a paradigm shift to a unified, live, and quality-controlled benchmarking framework—robust by construction rather than reliant on courtesy or goodwill.** Accordingly, we dissect the systemic flaws undermining today's evaluation ecosystem and distill the essential requirements for next-generation assessments.

To concretize this position, we introduce the idea of PeerBench¹, a community-governed, proctored evaluation blueprint that seeks to improve security and credibility through sealed execution, item banking with rolling renewal, and delayed transparency. PeerBench is presented as a complementary, certificate-grade layer alongside open benchmarks, not a replacement. We discuss trade-offs and limits and call for further research on mechanism design, governance, and reliability guarantees. Our goal is to lay the groundwork for evaluations that restore integrity and deliver genuinely trustworthy measures of AI progress.

¹A prototype implementation of PeerBench community is live at <https://peerbench.ai>.

1 Introduction

The widespread adoption of AI technologies — especially foundation models (FMs) — in decision-making processes has considerably heightened their societal impact. As a result, the need for the rigorous assessment of their performance has become increasingly urgent, positioning AI evaluation as a critical area of study. Benchmarks have become such influential forces in the AI industry that companies reportedly invest hundreds of thousands of dollars in compute resources to achieve top scores on evaluations such as the ARC-AGI benchmark [18]. Following the work of [20], we define a benchmark as a specific pairing of one or more datasets (typically including a test set, and sometimes training data as well) and an evaluation metric. This combination is intended to represent a particular task or set of capabilities, and is adopted by a research community as a common framework for comparing different methods. Benchmark leaderboards have become the go-to standard for evaluating the progress across AI subfields -from ImageNet[22] in vision to GLUE[23] in language. Evaluating algorithmic progress with benchmarks has become a double-edged sword; while they have accelerated iteration and competition, their popularity has also incentivized chasing of state-of-the-art (SOTA) performance [21], making them vulnerable to overfitting, gaming, and selective reporting. For instance, models that achieve so-called “superhuman” performance on question answering leaderboards often fail dramatically when evaluated on out-of-distribution inputs, revealing a lack of true understanding.

In order to better understand what benchmarks are truly measuring, The Markup [24], an investigative newsroom under CalMatters, interviewed researchers who designed evaluation datasets which revealed that many widely-used benchmarks are years old, increasing the likelihood that they were included in training data—compromising their effectiveness as unbiased evaluation tools. Public datasets often find their way, intentionally or inadvertently, into the training corpora of large models [25, 27, 28, 29], enabling memorization of test items rather than true generalization. Benchmark designers themselves may, intentionally or not, cherry-pick examples that favor particular architectures under pressure to produce impressive results. On the other end of the spectrum, proprietary or pay-walled evaluations limit accessibility and rely on the continued goodwill of their owners to remain relevant. Collectively, these practices create a distorted landscape in which leaderboard positions can be manufactured, scientific signal is drowned out by noise, and community trust is eroded[30].

Many scholars have raised concerns about the limitations of AI benchmarking, with some describing current evaluation practices as a “minefield” [26]. As hype increasingly overshadows genuine progress, the need for rigorous, trustworthy evaluation has become critical, especially when introducing new paradigms. In the following paragraph, we examine key structural flaws in the current evaluation pipeline, including data contamination, fragmented and inconsistent benchmarks, opaque dataset curation, the lack of safeguards for fairness and freshness, and how these issues have enabled superficial progress while undermining trust across the AI community.

Cracks in the Current Paradigm.

- **Data Contamination.** Public benchmarks may leak into or be deliberately injected into training sets, leading to test-set memorization and inflated scores [31, 32]. With today’s large-scale models trained on multi-trillion-token corpora, such contamination is increasingly inevitable [46, 51]. Retrieval-based audits report over 45% overlap on QA benchmarks, and GPT-4 infers masked MMLU answers in 57% of cases—well above chance [44]. Allegations that LLAMA 4 gained significant improvements via seeded paraphrases illustrate how easily scores can be engineered. N-gram audits, like those used on Qwen-1_8B [45], can help detect leakage but rely on partial knowledge of training data. Once contamination is plausible, generalization claims become suspect [37, 38, 35].
- **Risk of Strategic Cherry-picking.**
 - **Collusion.** Benchmark creators may collude with model creators and create hand-crafted suites that inadvertently or strategically advantage particular AI models.
 - **Selective Reporting.** Model creators can highlight performance on favorable task subsets, creating an illusion of across-the-board prowess, and preventing the audience from having a comprehensive bird’s eye view of the current landscape.

- **Bias in Test Data.** Current benchmarks, lacking unified data quality control, frequently suffer from test data bias, which can be an intentional or unintentional outcome of their design. This can lead to fundamentally misleading evaluations. For example, in Humanity’s Last Exam [1], organizers select five specific models and curate tests consisting solely of items that all five chosen models fail. Performance scores on such a dataset would unfairly penalize the initial five models and create an artificial advantage for any new model. This is akin to evaluating two models, A and B, of equal intrinsic ability on a task distribution \mathcal{D} (both solving 50% of tasks). If the test set is then constructed using only the subset of \mathcal{D} that model A solves but model B fails, it generates a specious conclusion of A’s superior performance, obscuring the fact that the test data itself is unrepresentative and biased.
- **Dataset Collection.** One key structural issue is the devaluation of dataset work within the machine learning community. In contrast to model innovation, dataset curation and documentation are treated as lower-status contributions. This has led to a culture in which datasets are frequently “reduced, reused, and recycled” without thorough contextualization [33], complicating efforts to track biases. Park and Jeoung [34] further observe that benchmark-sharing platforms like PapersWithCode suffer from inconsistent metadata terminology. Key details such as licensing and annotation processes are often missing, which complicates standardization efforts.
- **Noisy metrics, hypes & Evaluation fragmentation.** Public benchmark suites suffer from severe heterogeneity—each repository often introduces custom tokenizers, scoring rules (e.g., BLEU, ROUGE, EM, proprietary AI-scores), and ad-hoc scripts [49], making results difficult to reproduce and compare. Nearly all benchmarks are *static*, with performance gains increasingly reflecting task memorization rather than capability. For example, SUPERGLUE was rapidly saturated, with LLMs hitting performance ceilings shortly after release [53]. The lack of *liveness*—continuous inclusion of fresh, unpublished items—renders today’s metrics a stale snapshot. These inconsistencies encourage hype-driven “state-of-the-art” claims [36], misguide resource allocation, and crowd out rigorous analysis. Recent work [48, 50] calls for standardized, live evaluation protocols to reduce overhead, unify benchmarking efforts, and establish a shared understanding of *what to beat, how to measure it, and where the true frontier lies*.
- **Restricted accessibility for Private Benchmarks.** Proprietary or paywalled benchmarks can reduce contamination [56], but they shift epistemic authority to the curator, who alone controls evaluation access, task updates, and scoring [55]. This centralization raises ethical concerns: scientific progress becomes contingent on opaque processes, discretionary labor, and sustained funding. Without transparency in item selection, bias control, and submission filtering, it is unclear whether reported gains reflect true capability[39, 35] or favorable curation. Meanwhile, benchmark legitimacy is often conferred through peer review or citation momentum rather than principled design [39, 40]. As history shows, once interest fades, such benchmarks stagnate, yet continue to shape perception and citations [41].
- **Lack of Fairness and Proctoring.** Unlike high-stakes human exams, AI evaluations lack proctors, identity checks, and appeals processes [47]. Teams may fine-tune on test sets, exploit unlimited submissions, or selectively report results, often within current norms. Cultural, linguistic, and demographic skews [52] further bias outcomes, yet no oversight body governs these axes. This creates an uneven playing field, where resource-rich teams can game the system while more cautious researchers underreport.

Together, these factors blur scientific signals and undermine confidence in reported progress.

These shortcomings are particularly stark when compared to standardized human assessments like the SAT, GRE, or bar examinations, which are proctored, regularly updated, and governed by rigorous procedures to uphold fairness, reliability, and data integrity [43, 54]. *Why do we hold machines to lower evaluative standards than we do for humans in high-stakes environments?*

Desiderata for Next-Generation Evaluation. An ideal modern benchmarking regime should therefore be:

- **Unified.** All benchmarks operate under a single governance framework with common interfaces, standardized result formats, and a shared execution environment. A leaderboard,

akin to a “HuggingFace for evaluation”, lets researchers see at a glance where every model stands and removes the friction of juggling incompatible test harnesses.

- **Comprehensive.** The suite spans every major modality and task family, from individual modalities to multimodal reasoning, so progress can be tracked holistically rather than in isolated silos. Developers immediately know which capability gaps remain and which benchmarks to target next, without trawling the Internet for niche datasets.
- **Live and consistent.** Fresh, unpublished tests are produced on a rolling basis, preventing overfitting and test-set memorization, while earlier tests are retired and made public for auditing and research purposes. Robust score-normalization procedures align results across cohorts, ensuring that models evaluated on different slices of the benchmark remain directly comparable over time. To further preserve temporal validity, score decay methodologies, such as logistic time decay, can be applied to discount stale results and reflect the evolving relevance of model capabilities as both training regimes and real-world usage contexts shift.
- **Quality-controlled.** Each test, after being made public, is peer-reviewed for originality, difficulty, and bias, and its influence on a model’s composite score should be weighted by a transparent reputation system. This mechanism is crucial to down-weight low-quality or adversarial items, deter collusion between test authors and model developers, and preserve the integrity of the signal.

Any viable successor must deliver **contamination resistant, metric unification, transparent yet decentralized governance, and auditable fairness guarantees** — principles that define the vision for next-generation AI benchmarking.

We introduce a prototype of the desired paradigm, PEERBENCH, a community-powered platform for AI evaluation that demonstrates the practicality and outlines a roadmap toward this ultimate goal.

In summary, we posit that AI benchmarking paradigm should be reimagined and unified for built-in trustworthiness, data quality control, and contamination immunity. Unlike traditional benchmarks—static artifacts designed and maintained by closed teams—PEERBENCH proposes a shift toward evaluation as a living, auditable process governed by transparent rules and fueled by ongoing validator contributions, evolving with the field. On top of that, we design a prototype of the desired paradigm, PEERBENCH, a community-powered platform for AI evaluation, showing the practicality and roadmap to achieving the ultimate goal.

Key contributions. The main contributions of our paper are:

- **Structural critique.** A formal critique of the structural flaws, contamination, fragmentation, and monopolization undermining today’s benchmarks.
- **Position statement.** A position statement that reframes AI evaluation as a secure, standardized examination, together with design principles that balance openness and rigor.
- **Prototype architecture.** The PEERBENCH design is a minimum viable version of the desiderata, featuring a concrete ten-step workflow, cryptographically signed artifacts, a lightweight reputation scheme, and score-normalization methods that together transform heterogeneous community inputs into a longitudinal, contamination-resistant leaderboard.

The remainder of the paper is organized as follows. Section 2 reviews prior work on AI evaluation, draws lessons from human standardized testing, and systematically critiques the structural flaws of the current benchmarking regime. Building on this critique, Section 3 articulates our stance and distills the essential requirements for a next-generation evaluation paradigm. Section 4 presents PEERBENCH, a minimum-viable prototype that operationalizes these requirements through a live reputation system and liveness guarantees. Section 5 explores alternative designs, discusses current limitations, and outlines directions for future work. Finally, Section 6 concludes the paper.

2 Related Work

Public leaderboards have made undeniable contributions by spurring significant breakthroughs in AI; yet, the following review of current benchmarking efforts reveals persistent challenges in achieving sufficient robustness against contamination, ensuring long-term sustainability, and fostering genuine inclusiveness.

Static Benchmarks and Leaderboards. Widely-used suites such as MMLU [2], GSM8K [3], and SuperGLUE deliver clear snapshots of progress, but each ships a single public test set that quickly saturates and leaks into training corpora. BIG-Bench’s one-off community release [5] broadened task coverage, yet those tasks likewise became public upon publication, sharply reducing their discriminative power. HELM [6] added multiple metrics and periodic reports, but remains curator-driven and static between releases. In short, static benchmarks age poorly and cannot prevent data contamination.

Dynamic or Contamination-Resistant Benchmarks. LiveBench [7] refreshes tasks continuously, demonstrating that rolling updates slow leakage. Still, it relies on a single centralized team, limiting scale and diversity, and highly depends on the creators’ goodwill to actively maintain. Similarly, Dynabench [8] explored adversarial data collection with humans-in-the-loop, but its reach was limited by centralized infrastructure and annotation scalability. Adversarial “break-the-model” contests [12] expose weaknesses but run sporadically and lack systematic score aggregation. Robustness probes like Checklist [9] explore model failures via templated behavioral tests, but require hand-crafting and do not scale to sustained, community-wide evaluation. PEERBENCH extends these ideas by democratizing task sourcing and embedding adversarial challenges into a permanent, governance-backed workflow.

Human-Preference and Open Evaluation Platforms. Crowdsourced pairwise ratings power Chatbot Arena’s Elo ladder [10] and OpenAI Evals, while the HuggingFace Open LLM Leaderboard lets users upload test scripts. These platforms foster openness, yet they rest on static prompt sets, absence of identity verification, or vendor-specific ecosystems, making them vulnerable to spam, bot voting, and untracked contamination. The evaluation results become less convincing because of ineffective data quality control. PEERBENCH addresses those gaps with verified validators, reputation-weighted scoring, and single-use test suites.

Reputation Systems and Decentralized Governance. Mechanisms from Stack Exchange, Wikipedia, and blockchain governance inspire our validator-reputation design, but no prior AI benchmark fully unifies decentralized contribution, reputation weighting, and contamination safeguards. PEERBENCH is, to our knowledge, the first to weave these strands into an end-to-end, self-sustaining evaluation network.

Summary. Prior work offers valuable ingredients, like broad task coverage, rolling updates, and human preference ratings, but each leaves critical weaknesses: public data leakage, single-team bottlenecks, or unverifiable crowdsourced inputs. In this position paper, we call for a new benchmarking paradigm that synthesizes these ideas while eliminating their shortcomings by design, filling a long-standing gap in trustworthy, contamination-immune evaluation.

3 Towards a New Paradigm: From Static Leaderboards to Proctored Exams

To remedy the failure modes catalogued in Section 2, we propose recasting AI benchmarking as a **standardized, proctored examination** rather than an “open-book” contest of self-reported scores. The analogy is deliberate: human aptitude tests (e.g. SAT, GRE, bar exams) have evolved over decades to balance security, fairness, and public credibility—precisely the properties modern AI evaluation lacks. Our paradigm rests on seven principles.

- **Secret test sets.** Evaluation items remain undisclosed until runtime. The question bank is either freshly generated or drawn from an encrypted reserve, precluding training-time contamination and rote memorization.
- **Proctored execution.** Models are evaluated in a unified sealed sandbox with an identical execution environment. The procedure mirrors a human exam: a fixed knowledge state is tested under identical, monitored conditions. All inputs and outputs are logged and cryptographically signed via hashing to prevent tampering.
- **Community governance.** A multi-stakeholder network of validators enforces rules and governance. The validator network curates test content and peer reviews test submissions. Validator actions are incentivized and audited via a transparent reputation and slashing system.

- **Continuous renewal and liveness.** At every evaluation round, a fixed fraction of questions is retired and replaced. Retired items may be released for research, but they are never reused for new score submissions once they are made public.
- **Auditability and integrity.** Validators pre-commit to test and answer hashes before publication. Later, a randomly selected public subset is revealed to allow other validators to cross-verify fidelity, prior exposure, and integrity. Proven data leakage results in disqualification, analogous to academic dishonesty.
- **Equitable access.** Any bona-fide team, academic, corporate, or independent, can submit a model, subject only to compute reimbursement fees. A small laboratory competes on precisely the same footing as a large vendor.
- **Multi-metric reporting.** Following educational testing practice, the score report provides domain-specific subscores (e.g., maths, coding, reasoning) and percentile ranks, not merely a single headline number. Fairness metrics (bias, robustness) are computed uniformly across models.

These principles demand greater up-front effort, drafting high-quality secret items, operating a proctoring infrastructure, but they yield durable benefits: contamination immunity, reproducible fairness, and results that stakeholders can audit rather than trust on faith. Table 1 surmises the contrasts with the status quo.

Table 1: Comparison of AI evaluation platforms. A desired paradigm should combine the strengths of prior approaches (fresh unseen tasks, expert involvement, human feedback, data quality control) with auditability, to mitigate weaknesses (central trust, static data, etc.).

Benchmark	Dynamic Update	Data Source Diversity	Transparency	Contamination Resistance	Data Quality Control
<i>Static Evals (MMLU, etc.)</i>	No	Single (Originating Team)	Yes (Public test sets)	No [†]	Opaque; Community-Reliant [†]
Scale SEAL (2023)	Yes (Continuous)	Single (Scale AI)	No (Private test sets)	Yes (Proprietary)	Opaque; Vendor-Internal [†]
LiveBench (2024)	Yes (Monthly)	Single (Research Team)	Yes (Public post-evaluation)	Partial [‡]	Opaque; Team-Internal [†]
ARC-AGI (2019–)	Partial (Episodic Sets)	Single (Organizers)	Yes (Public test sets)	Partial [‡]	Opaque; Expert-Driven [†]
Chatbot Arena (2023)	Yes (Ongoing Prompts)	Yes (Crowdsourced)	Yes (Public prompts)	N/A [§]	Limited (Elo-Based) [†]
Desired Paradigm	Yes (Continuous Rolling)	Yes (Validated Network)	Yes (Public post-evaluation)	Yes (By Design)	Transparent; Unified [*]

[†]Refers to data quality control that is not explicitly defined by, transparent to, or verifiable by the broader community, often relying on the originating entity’s internal standards or reputation.

[‡]Susceptible to contamination once test items are released or become predictable, even if new items are added.

[§]Focuses preference-based chat quality with dynamic inputs, not fixed knowledge test sets that could leak.

^{*}Achieved via explicit, community-vetted standards, reputation systems, and transparent post-hoc auditing of test items and processes.

The remainder of the paper instantiates some of these principles in a concrete, community-governed prototype, PEERBENCH, whose architecture and methodological safeguards are presented in Section 4, demonstrating the practicality of our position.

4 PeerBench: A Live, Community-Governed Benchmarking Platform

We introduce PEERBENCH², a prototype platform implementing the proctored-exam paradigm outlined earlier. It integrates a lightweight, decentralized governance layer and a secure, cryptographically verifiable workflow to assure data quality of tests and a live, contamination-resistant, and reproducible evaluation.

4.1 Actors

Conceptually, we will simultaneously support different evaluation streams (e.g. maths, coding, translation) for test datasets. Our system involves four types of actors:

Validators author private test suites, query models, grade responses, and peer-review one another. **Model providers** expose inference endpoints for their models—open or closed—and optionally register for specific evaluation streams. The **Coordination server** is a neutral service that authenticates uploads, schedules peer review, updates reputations, and publishes the public leaderboard, while storing artefacts in both a database and immutable object storage for auditability. **End-users**—such as researchers, journalists, regulators, and practitioners—consult the live leaderboard and may apply validator-trust thresholds when interpreting results.

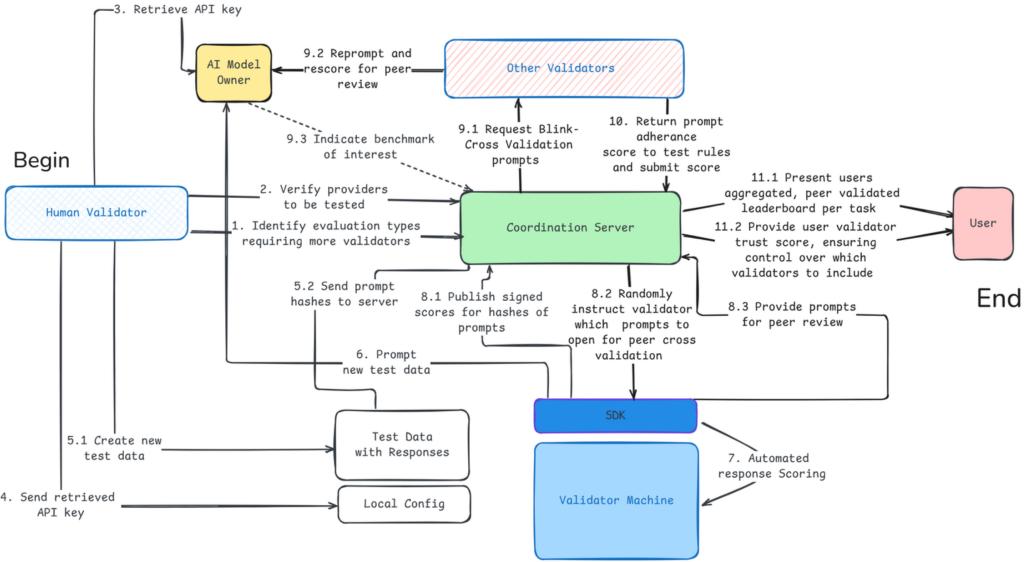


Figure 1: Lifecycle of a PEERBENCH evaluation round.

4.2 End-to-End Workflow

We will now outline the workflow from the perspective of an individual validator. All published messages shall be signed to guarantee authenticity.

Setup. Validators and model providers register with verifiable credentials (e.g., institutional email) and generate public signing keys. Validators are required to stake collateral, which may be slashed for misconduct. Upon registration, validators receive two initial reputation scores: data quality reputation ρ_{DQ} which is based on the quality of test content they contribute and influences their contribution to model scores, and peer-review reputation ρ_{PR} , which is based on the accuracy and reliability of their reviews of others' test content and influences their weight in peer-review in the future.

S1 Validator Onboarding. Validators complete identity verification, stake collateral, generate a cryptographic key pair for signing, and are assigned initial values for ρ_{DQ} and ρ_{PR} .

Validator Workflow per Round. Validators generate private test prompts, give a preliminary score to models and prepare their data for audit.

R1 Benchmark test preparation. In each evaluation round, private benchmarking prompts $Q_i^{(v)}$ and scoring functions $F_i^{(v)}$ are drafted by the validator v .

R2 Hash commit. Validators publish hash commitments of their private test sets $h_i = \text{Com}(Q_i^{(v)}, F_i^{(v)})$ to avoid retroactive cherry-picking. The validator further publicly identifies which evaluation stream its test set belongs to.

R3 Model Evaluation. For each already registered model to be evaluated on this evaluation stream, the validator should make queries with $Q_i^{(v)}$ and store the audit log and raw answers $A_i^{(v,m)}$. It further outputs commitments $h_{i,v,m} = \text{Com}(A_i^{(v,m)})$ to its logs publicly.

R4 Local scoring. Objective tasks are auto-graded with scoring function $F_i^{(v)}$, producing raw scores $s_i^{(v,m)}$; subjective tasks which require preference voting will be scored after test publication. A preliminary score can already be published.

²A prototype implementation of PeerBench is available at: <https://peerbench.ai>.

R5 Reveal Subset for Auditing. Validators reveal a random subset $(Q_j^{(v)}, F_j^{(v)})$ of their test suite, selected via a public randomness beacon (e.g., Drand³. They also publish corresponding model answers and open their earlier commitments $h_j, h_{j,v,m}$ for verification.)

R6 Online Peer-Review. Each validator retrieves public tests from others and verifies consistency with committed hashes (from R1). Discrepancies are flagged as cheating. Validators re-run the public tests locally as in steps **R3-R4**. Significant score or model answer mismatch to what the validator reported causes further investigation.

Finally, the individual score on the public test set and answers are published.

At the End of a Round. Results are finalized, test sets are retired, and reputations are updated.

R7 Full Test publication. All test data $Q_i^{(v)}$, logs, and model responses $A_i^{(v,m)}$ are made public for auditing and further research purposes. Consistency with previous hash commits can be checked. Test sets are permanently retired from future evaluations.

R8 Score calculation. Objective scores are finalized. Subjective scores (e.g., preference-based) are determined via peer voting. All scores are normalized for cross-model comparability.

R9 Data quality calculation. Validators score each test set $Q_i^{(v)}$ according to originality, rule adherence, and agreement. Each set receives a final quality score $DQ(Q_i^{(v)})$, weighted by reviewers' ρ_{PR} reputations. The final score of each model is the average performance on all tests, weighted by the data quality of each test set.

R10 Reputation update and Slashing. After data quality calculation of each evaluation cycle, the reputation ρ_{DQ} and ρ_{PR} are updated. ρ_{DQ} is updated according to the data quality $DQ(Q_i^{(v)})$ of the test set in the new round, and ρ_{PR} is updated according to the proximity of validator v 's scoring on other tests to the final data quality of other tests. Suspicious validators with a reputation lower than a certain threshold will be removed and have their collateral slashed.

Timing. A typical round may span approximately one month, allowing for asynchronous participation. To prevent contamination, only models released *before* the test generation phase are eligible for scoring in the current round.

4.3 How Our Design Choices Address Common Issues

- **Data contamination & cherry-picking.** Validators pre-commit to test sets, which remain private until the round concludes and are never reused. This ensures that training on this data is infeasible if the validator behaves honestly.
Any collusion or cherry-picking, such as tailoring questions to favored models, can be detected via cross-validator score discrepancies, discouraging misconduct through slashing and reputation penalties. This incentivizes honest behaviour.
- **Cheating on the private dataset.** A public source of randomness determines which queries are revealed, preventing validators from anticipating which items will be audited. Hash commitments to both test items and model responses ensure verifiable consistency and enable reliable detection of manipulation.
- **Selective reporting.** Model providers declare their participation in specific evaluation streams prior to each round. Scores are computed centrally and automatically, precluding selective disclosure or omission of unfavorable results.
- **Accessibility and liveness.** Validator onboarding is lightweight to support broad participation. The decentralized network of test contributors ensures that evaluations continue on a regular basis, even with fluctuating validator involvement.
- **Test quality.** Each test set is peer-reviewed, with validator influence weighted by reputation. Validators submitting low-quality or biased content may have their collateral slashed and their contribution discounted, promoting sustained data quality.

³<https://www.drand.love/>

4.4 Considerations towards Implementation

There are multiple considerations that remain to be considered in order to make such a system feasible in the future:

- **Scoring metrics.** The implementation hinges on a fair scoring system that should be collusion-resistant and whose data quality scores can be fairly aggregated among heterogeneous metrics. We envision a reputation-based system to aggregate different validators’ scores - Provision of such a system is left up to future work.
- **Allowing data reuse.** Our current draft reveals *all* test data after each round, which prevents the reuse of leaked test cases. However, it might be desirable to allow a certain degree of reuse in practice, as good tests are hard to come up with. The goal should be to only open some of the tests to the public at the end of each round, while maintaining fair scoring.
- **Scoring new models.** For simplicity, we assumed that only models registered before the current round are scored. An ideal solution should score on a rolling basis, but exclude already revealed tests from evaluation.

5 Alternative Views

The proposed transition from open, self-reported leaderboards to a proctored, community-governed examination system signifies a substantial evolution from current AI benchmarking practices. This section addresses potential counterarguments and perceived limitations of our proposed paradigm, outlining the compromises and safeguards we envision to ensure its effectiveness and integrity.

1. Preserving the Value of Open Benchmarks

Concern. Public datasets fuel rapid AI progress by allowing universal access for error diagnosis and quick iteration. A shift to hidden tests could impede this vital open development cycle.

Our Approach. We advocate for a two-tiered system. “Practice” sets—comprising retired questions or legacy benchmarks—would remain openly accessible for ongoing debugging and method development, while a “final” set of fresh, unseen questions would determine certified scores, akin to the public/hidden data split in Kaggle competitions.

2. Ensuring Transparency and Trust with Secret Tests

Concern. If evaluation questions are kept secret, how can the research community be confident that they are free from bias towards specific methodologies or approaches?

Safeguards. We propose several measures: (i) The exam board will be multi-institutional with rotating membership to ensure impartiality; (ii) statistical summaries detailing topic distribution, difficulty calibration, and demographic coverage will be published with each test release; and (iii) all test items will be released after retirement, enabling thorough post-hoc scrutiny and community auditing.

3. Addressing Practical Costs and Logistical Hurdles

Concern. The resources required for developing secure questions, operating a dedicated evaluation server, and covering inference computation costs are non-trivial.

A Feasible Path Forward. Existing neutral organizations (such as NIST or MLCommons) or a newly established not-for-profit foundation could undertake hosting the evaluation service. Costs could be managed through a combination of modest submission fees and public funding to support academic participation. Containerized inference submissions can also be implemented to protect proprietary model weights while still allowing for secure, remote execution.

4. Balancing Innovation Pace and Open-Ended Exploration

Concern. The dynamic of instant feedback on public leaderboards often ignites creative “leaderboard hacking,” which can subsequently evolve into genuine research advancements. A slower examination process might inadvertently dampen this innovative energy.

Our Perspective. Researchers will remain free to experiment and iterate using open data sources; the proposed exam system is designed to provide a high-confidence certificate. In practice, the inherent uncertainty regarding the exam’s precise content is likely to encourage broader, more generalizable

research rather than narrow overfitting. While a slower feedback loop is an acknowledged trade-off, it is justified by the significant gains in the reliability and robustness of the evaluation outcomes.

In summary, our proposed paradigm does not argue against the principle of openness in AI research but rather targets the vulnerabilities associated with *over-exposed* test sets. By integrating public “practice” data with a system of rolling, audited secret exams, we aim to uphold the collaborative spirit of the AI community while simultaneously restoring confidence in headline performance claims. We aim to complement, not replace, open innovation; and to certify, not constrain, genuine progress.

6 Conclusion

Benchmarking is the heartbeat of empirical AI, yet static public datasets now leak, self-reported leaderboards are gamed, and headline scores may no longer signal real ability. Inspired by human exams, we advocate replacing open-book, developer-run benchmarks with a *proctored, community-governed test*. The core requirements, secret test sets, continuous liveness, data-quality auditing, and scale-invariant scoring, coalesce in PEERBENCH, a minimal prototype that combines reputation-weighted validators, rolling task renewal, and fully auditable score aggregation. As the field scales in impact and visibility, only rigorous, auditable benchmarks can anchor meaningful scientific progress.

Call to Action

Progress in AI must be *measured*, not merely marketed. We invite researchers, practitioners, and policymakers to help refine, deploy, and steward this emerging evaluation paradigm. In particular, we encourage work on mechanism design and game-theoretic security analysis to strengthen the framework’s economic and adversarial robustness. By directing collective effort toward *how* we measure, we protect the integrity of *what* we build, so that future claims of “state-of-the-art” or “human-level” performance once again carry demonstrable scientific weight.

References

- [1] Long Phan, Alice Gatti, Ziwen Han, et al. *Humanity’s last exam*. arXiv:2501.14249, 2025.
- [2] Dan Hendrycks, Collin Burns, Steven Basart, et al. *Measuring Massive Multitask Language Understanding*. In International Conference on Learning Representations (ICLR), 2021.
- [3] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. *Training Verifiers to Solve Math Word Problems*. arXiv:2110.14168, 2021.
- [4] Tom B. Brown, et al. *Language Models are Few-Shot Learners*. In Advances in Neural Information Processing Systems (NeurIPS), 2020.
- [5] Aarohi Srivastava, et al. *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models (BIG-Bench)*. In Proceedings of NeurIPS 2022 (Dataset and Benchmark Track), 2022.
- [6] Percy Liang, et al. *Holistic Evaluation of Language Models*. arXiv:2211.09110, 2022.
- [7] Colin White, Samuel Dooley, Manley Roberts, et al. *LiveBench: A Challenging, Contamination-Limited LLM Benchmark*. In International Conference on Learning Representations (ICLR), 2025.
- [8] Douwe Kiela, Max Bartolo, et al. Dynabench: Rethinking Benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- [9] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
- [10] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, et al. *Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference*. arXiv:2403.04132, 2024.

- [11] Ethan Perez, et al. *Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned*. arXiv:2207.09455, 2022.
- [12] S. Golovanov, et al. *Adversarial Prompting for Black Box Foundation Models*. arXiv:2302.04251, 2023.
- [13] Lianmin Zheng, et al. *Judging LLM-as-a-judge with MT-Bench and Chatbot Arena*. arXiv:2306.05685, 2023.
- [14] Kyle Wiggers. *Meta exec denies the company artificially boosted Llama 4's benchmark scores*. TechCrunch, April 7, 2025.
- [15] Anna A. Grigoryan. *When Benchmarks Lie: Why Contamination Breaks LLM Evaluation*. Medium, March 30, 2025.
- [16] The Supercharged (Elena Perez). *Meta denies claims of Llama 4 benchmark cheating*. Medium, April 8, 2025.
- [17] Maria Eriksson, et al. *Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation*. arXiv:2502.06559v1, 2025.
- [18] Greg Kamradi. *ARC-AGI-2 + ARC PRIZE 2025 IS LIVE!*. ARC Prize, 24 March, 2025.
- [19] Ralph Merkle. *A Certified Digital Signature*. CRYPTO, August, 1989.
- [20] Deborah Raji. *AI and the Everything in the Whole Wide World Benchmark*. Advances in Neural Information Processing Systems (NeurIPS), 2021.
- [21] *AI Benchmarks: Why GenAI Scoreboards Need an Overhaul* . Sumeet Wadhwani, 2024
- [22] Olga Russakovsky, et-al. *ImageNet Large Scale Visual Recognition Challenge*. International Journal of Computer Vision, 2015
- [23] Alex Wang, et-al. *GLUE: A multi-task benchmark and analysis platform for natural language understanding* . 7th International Conference on Learning Representations, 2019
- [24] Jon Keegan. *Everyone Is Judging AI by These Tests. But Experts Say They're Close to Meaningless*. The Markup, 2024
- [25] Chunyuan Deng, et-al. *Investigating Data Contamination in Modern Benchmarks for Large Language Models* . arXiv:2311.09783, 2024
- [26] Arvind Narayanan and Sayash Kapoor. *Evaluating LLMs Is a Minefield* . 2023
- [27] Arvind Narayanan and Sayash Kapoor. *GPT-4 and professional benchmarks: the wrong answer to the wrong question* . AI Snake Oil, 2023
- [28] Hugh Zhang, et-al. *A Careful Examination of Large Language Model Performance on Grade School Arithmetic* . arXiv:2405.00332v1, 2024
- [29] Inbal Magar and Roy Schwartz. *Data Contamination: From Memorization to Exploitation*. arXiv:2203.08242, 2022
- [30] Anthony Corso, et-al. *A Holistic Assessment of the Reliability of Machine Learning Systems*. arXiv:2307.10586, 2023.
- [31] Benjamin Recht, et-al. *Do ImageNet classifiers generalize to ImageNet?* arXiv:1902.10811, 2019.
- [32] Jesse Dodge, et-al. *Show Your Work: Improved Reporting of Experimental Results*. arXiv:1909.03004, 2019.
- [33] Bernard Koch, et-al. *Reduced, Reused and Recycled: The Life of a Benchmark Dataset in Machine Learning Research*. arXiv:2112.01716, 2021

- [34] Yejin Park and Youna Jeoung. *Benchmarking Benchmarks: Metadata, Provenance, and Terminology Confusion in AI Dataset Platforms*. arXiv:2210.12345, 2022.
- [35] Maria Eriksson, et-al. *Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation*. arXiv:2502.06559v1, 2024.
- [36] Will Orr and Edward B. Kang. *AI as a Sport: On the Competitive Epistemologies of Benchmarking*. Conference on Fairness, Accountability, and Transparency, pp. 1872–1885, 2024
- [37] Shachar Kaufman, et-al. *Leakage in Data Mining: Formulation, Detection, and Avoidance*. ACM Transactions on Knowledge Discovery from Data (TKDD), 6(4), Article 15, 2012.
- [38] Patrick Lewis, et-al. *Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets*. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 1000–1008, 2021.
- [39] David Schlangen *Targeting the Benchmark: On Methodology in Current Natural Language Processing Research*. arXiv:2007.04792, 2020
- [40] Simon Ott, et-al. *Mapping global dynamics of benchmark creation and saturation in artificial intelligence*. Nature Communications 13, 1 (Nov. 2022), 6793.
- [41] Mostafa Dehghani, et-al. *The Benchmark Lottery* arxiv:2107.07002, 2021
- [42] Timothy R. McIntosh, et-al. *Inadequacies of Large Language Model Benchmarks in the Era of Generative Artificial Intelligence*. arxiv:2402.09880, 2024
- [43] Camara, W. J., Lane, S. *Score reporting and interpretation issues and challenges for educational measurement*. Educational Measurement (4th ed., pp. 515–547). Westport, CT: American Council on Education/Praeger, 2006
- [44] Chunyuan Deng, et-al. *Investigating Data Contamination in Modern Benchmarks for Large Language Models*. arxiv:2311.09783, 2024
- [45] Ruijie Xu, et-al. *Benchmarking Benchmark Leakage in Large Language Models*. arxiv:2404.18824v1, 2024
- [46] Yonatan Oren, et-al. *Proving Test Set Contamination in Black Box Language Models*. arxiv:2310.17623, 2023
- [47] Q. Vera Liao, et-al. *Rethinking Model Evaluation as Narrowing the Socio-Technical Gap*. arxiv:2306.03100, 2025
- [48] Paul Rottger, et-al. *SafetyPrompts: a Systematic Review of Open Datasets for Evaluating and Improving Large Language Model Safety*. arxiv:2404.05399, 2025
- [49] Ning Wu, et-al. *Large Language Models are Diverse Role-Players for Summarization Evaluation*. arxiv:2303.15078, 2023
- [50] Anka Reuel, et-al. *BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices*. arxiv:2411.12990, 2024
- [51] Barz, B. and Denzler, J. *Do we train on test data?* Journal of Imaging, 6(6):41, Jun 2020
- [52] Shreya Shankar, et-al. *No classification without representation: Assessing geodiversity issues in open data sets for the developing world* arxiv:1711.08536, 2017
- [53] Alex Wang, et-al. *SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems* arxiv:1905.00537, 2019
- [54] Wanjun Zhong, et-al. *AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models* arxiv:2304.06364, 2023
- [55] Ben Bucknall, et-al. *Position: Ensuring mutual privacy is necessary for effective external evaluation of proprietary AI systems* arXiv.2503.01470, 2025
- [56] Moran Mizrahi, et-al. *State of What Art? A Call for Multi-Prompt LLM Evaluation* arXiv.2401.00595, 2024