Position: Benchmarking is Broken – Don't Let AI be its Own Judge

Zerui Cheng[1,*] Stella Wohnig[2,*] Ruchika Gupta[3,*] Samiul Alam[4,*] Tassallah Abdullahi[5] João Alves Ribeiro[6] Christian Nielsen-Garcia[7]
Saif Mir[4] Siran Li[8] Jason Orender[9] Seyed Ali Bahrainian[8] Daniel Kirste[10] Aaron Gokaslan[11] 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. **Contact:** `zerui.cheng@princeton.edu` **Personal Homepage:** `https://www.zerui-cheng.com` **PeerBench Project:** `https://www.peerbench.ai`

PeerBench Platform Paper

Motivation

Benchmarks shape AI progress, policy, and investment. But today's evaluation ecosystem is noisy, fragmented, and easily gamed. We argue for a unified, live, and quality-controlled approach to restore signal and trust.

What's Broken in Today's Benchmarks?

A non-exhaustive summary of the flaws in today's AI evaluation paradigm includes:

- **Data Contamination:**
  Data leakage / injection into training corpora; **Memorization isn't Generalization.**
  Public static datasets encourage hype-driven "state-of-the-art" claims.
- **Cherry-picking & Selective Reporting:**
  Hype by showcasing only favorable slices (selective reporting);
  Collusion risks between benchmark creators and model/agent owners.
- **Biased Test Data:**
  Ad-hoc curation can unfairly advantage or disadvantage specific models.
- **Dataset Collection Undervaluation:**
  Dataset and benchmarking efforts are often treated as lower-status contributions;
  Inconsistent licensing and documentation complicate standardization efforts;
  Datasets are "reduced, reused, and recycled" without thorough contextualization.
- **Noisy Metrics & Fragmentation:** Heterogeneous tokenizers, scoring rules, and ad-hoc evaluation scripts.
- **Private Opacity:** Centralized control ≈ uncheckable claims; legitimacy via reputation rather than design.
- **Lack of Proctoring & Fairness:** No identity checks, unlimited submissions, uneven access → gaming and bias.

Example 1: Llama 4 Overfitting Controversy

r/LocalLLaMA • 24 days ago
mryougi
"Serious issues in Llama 4 training. I Have Submitted My Resignation to GenAI"

Example 2: Bias in Humanity's Last Exam Dataset

In Humanity's Last Exam [1] dataset by Scale AI,
a necessary condition for a problem to be included is:
The problem is unsolved by 5 specific models on pass@1.

This creates an inherent bias where new models differing from the 5 selected models will look better (cherry-picking in expectation).

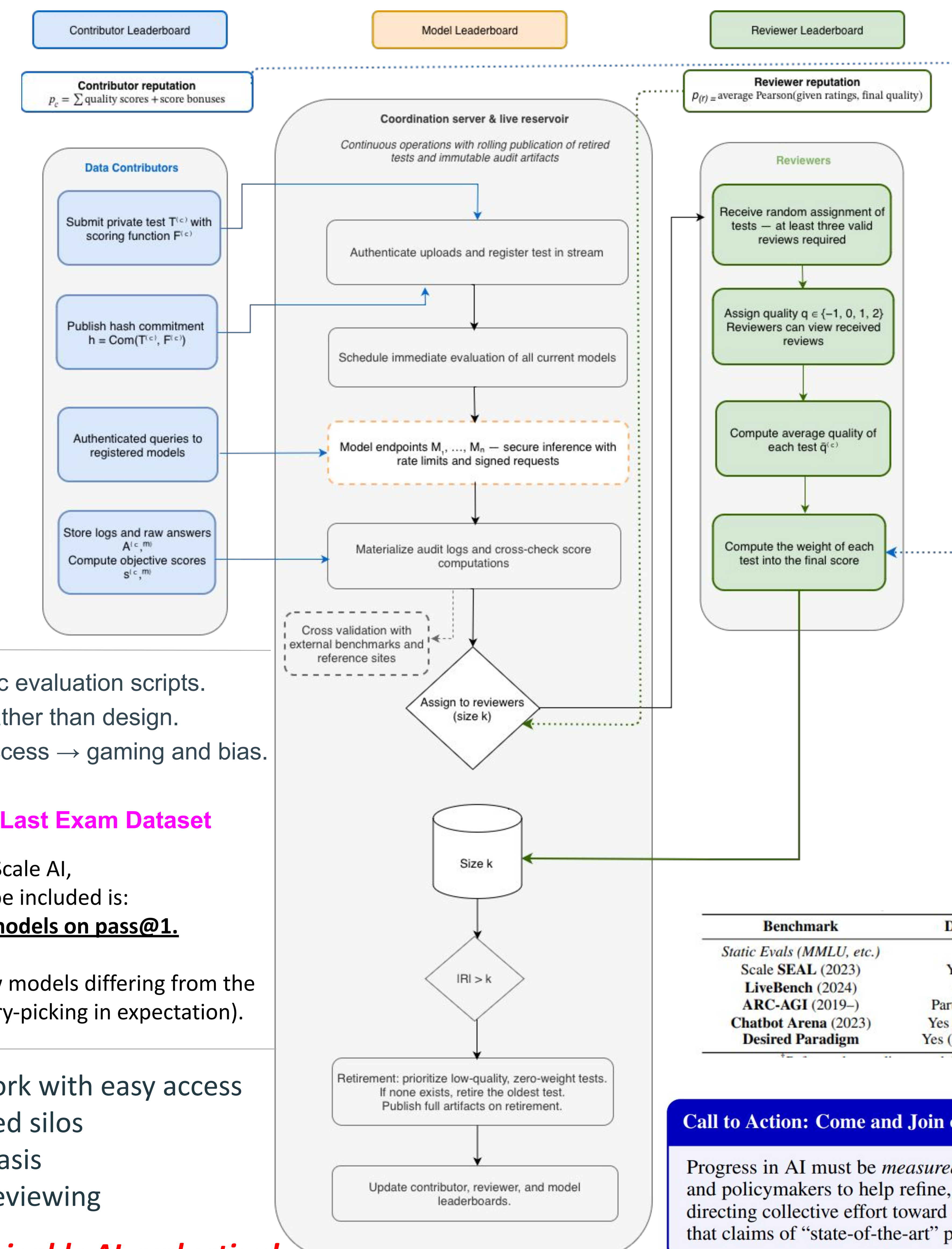
What do we Need for a Next-Gen Evaluation Paradigm?

1. **Standardized** - Benchmarks are hosted under a unified framework with easy access
2. **Comprehensive** - Progress tracked holistically rather than in isolated silos
3. **Live and consistent** - Fresh, unpublished tests produced on a rolling basis
4. **Quality-controlled** - Retired tests made public and go through peer reviewing

TL;DR: We shouldn't rely on people's courtesy and goodwill for fair and sustainable AI evaluation!

End-to-End Workflow of PeerBench Prototype

Available at: www.peerbench.ai



Our Proposal - PeerBench: Live, Community-Governed Benchmarking Platform

Actors

- **Data Contributors:** Author tests with executable scoring;
  Receive reputation from data quality determined by reviews
- **Reviewers:** Provide ordinal ratings; accuracy vs. consensus determines weight.
- **Model Creators:** Register models; expose inference endpoints;
- **Coordination Server:** Orchestrates cross-validation and public leaderboards.
- **End Users:** Researchers, regulators, practitioners;
  Consult live leaderboards with uncertainty thresholds.

Three Leaderboards

- **Data Contributor Leaderboard:** Cumulative test quality + verification bonuses.
- **Reviewer Leaderboard:** Alignment with consensus quality.
- **Model Leaderboard:** Weighted by test quality for fair, robust scoring.

Temporal Fairness: Scheduling Trade-offs

- **Immediate Scoring on Request** → responsiveness;
  risk: cross-period contamination / comparability.
- **Synchronized Evaluation Windows** → strongest fairness; slower iteration.

Our Hybrid Approach: Data split
Part is used for immediate scoring for instant responsiveness;
The rest is used for periodic windows for cross-cohort fairness and comparability.

Security and Audits

- **Peer Review and Test Revelation Scheme for Data Quality Assurance:**
  A small random portion of live tests revealed privately to reviewers for peer review.
  After retirement, all tests are revealed for further check on quality and consistencies.
- **Slashing & Incentives:**
  Collateral and rewards/punishments align behavior; platform-sustainable economics.

Benchmark	Dynamic Update	Data Source Diversity	Transparency	Contamination Resistance	Data Quality Control
Static Evals (MMLU, etc.)	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 (Validator Network)	Yes (Public post-evaluation)	Yes (By Design)	Transparent; Unified*

Call to Action: Come and Join our Community at `www.peerbench.ai`!

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. By directing collective effort toward *how* we measure, we protect the integrity of *what* we build, so that claims of "state-of-the-art" performance once again carry demonstrable scientific weight.

References

[1] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang et al. "Humanity's last exam." arXiv preprint arXiv:2501.14249 (2025).

[2] Zerui Cheng,, Stella Wohnig, Ruchika Gupta, Samiul Alam, Tassallah Abdullahi, João Alves Ribeiro, Christian Nielsen-Garcia et al. "Benchmarking is Broken-Don't Let AI be its Own Judge." arXiv preprint arXiv:2510.07575 (2025).

Disclaimer: peerbench.ai is an open-source, non-profit community implementation of the paper, bringing the research to life for everyone.