Public AI Benchmarks Are Broken, But Are Private Benchmarks the Answer?

Sang Truong[1], **Serena Wang**[3,4], **Angelina Wang**[1,2], **Nhi Truong**[1], **Anka Reuel**[1], **Nick Haber**[1], **Sanmi Koyejo**[1]
[1]Stanford, [2]Cornell, [3]Harvard, [4]UBC,

Abstract

As large-scale generative models proliferate, benchmarks have become the primary instrument for measuring progress. Publicly accessible benchmarks have served as shared goals that catalyze and coordinate collective innovation. Despite the benefits, in the era of large-scale models that consume most of the internet, concerns exist that models are being trained on these public benchmarks, thereby inflating the scores of models "trained on the test." Private benchmarks attempt to circumvent these concerns, but they come with issues related to trust. This raises a critical question: Should benchmarks be public to foster trust and collaborative progress or private to maintain benchmark integrity? To find a viable solution that balances evaluation integrity and trust, we start by decomposing benchmarks into two components: *content* and *design*. Benchmark content comprises the evaluation materials themselves—prompts or questions, answer keys, the decision rule for matching model outputs to keys, and the aggregation procedure that collapses question-level scores into a summary statistic. Benchmark design, in turn, specifies the construct the benchmark is intended to measure and supplies the validity evidence that links the observed scores to that construct. Base on this benchmark decomposition, we then conduct five case studies of benchmarks across fields according to the public and private natures of their content and designs. These case studies include the College Board's Scholastic Assessment Test and four modern AI benchmark – including NIST's Face Recognition Vendor Test, Scale AI's SEAL, LMSys's Chatbot Arena, and Stanford's HELM. We argue for a hybrid public-private approach: keeping the benchmark content private to maintain integrity while making the benchmark design public to build trust. This approach offers a path toward trustworthy measurement of ever-growing data-intensive AI systems.

1 AI Benchmarks and The Rise of Privatization

Benchmarks have become a cornerstone of progress measurement in AI, particularly with the advent of large-scale models [Wei, 2023, Raji et al., 2021, Hardy et al., 2025]. The components of a benchmark can be broadly categorized into its *content* and *design*. Benchmark content refers to the specific materials used in the evaluation: the questions or prompts, corresponding answer keys, the decision rule used to determine whether a model's output matches the key, and the aggregation method for converting item-level scores into a single summary metric. In contrast, benchmark design encompasses the underlying construct the benchmark aims to measure, along with supporting validity studies that justify the alignment between the benchmark and its intended interpretation.

Many classic benchmarks, such as GLUE [Wang et al., 2018] and MMLU [Hendrycks et al., 2021], have both public content and design. Public benchmark fueled rapid progress (e.g., GLUE was saturated within one year from its launch) but also introduced downsides: benchmarks become targets for overfitting and contamination once widely adopted [Bansal and Maini, 2025, Wang et al., 2024a].

Preprint. Under review.

For example, GPT-4 solved 10 out of 10 older Codeforces problems but 0 out of 10 recent ones, suggesting memorization rather than true generalization [Narayanan and Kapoor, 2023, He, 2023].

In response to these concerns, a growing number of efforts have adopted partial or full privatization - restricting public access to various components of benchmarks [Rajore et al., 2024, Phan et. al., 2025]. The core rationale is that concealing benchmark content reduces the risk of overfitting and yields a more faithful assessment of generalization. Benchmark builders can selectively disclose individual components of the benchmark, such as questions, answer keys, and/or the construct's definition. At the highly private end of this spectrum, some benchmarks dynamically generate question content by adding noise from carefully constructed distributions or templates, constructing novel questions and corresponding labels on the fly. This ensures that each model encounter is novel. This approach is especially useful for evaluating commercial AI systems, where test prompts must be submitted via API calls. In such settings, model providers may choose to log inputs, potentially contaminating future benchmark runs.

As illustrated in Figure 1, current benchmarks span a range of disclosure strategies. For example, Kaggle competitions employ a split-evaluation scheme: participants receive public feedback during development, while final rankings rely on a hidden test set to discourage overfitting. Recently, Scale AI extended this paradigm to LLM evaluation via ongoing private leaderboards [Bansal and Maini, 2025, Scale AI, 2024a], such as SEAL, which ranks models on domains like coding and math using undisclosed, curated datasets that are claimed to be robust against gaming [Scale AI, 2024b].

While public benchmarks increasingly suffer from contamination, full privatization is not a comprehensive solution. Private benchmarks introduce their own challenges: reduced trustworthiness due to the risk of conflicts of interest, especially when transparency is lacking. Benchmarks serve as critical interfaces among stakeholders with differing priorities: researchers aim for scientific rigor, companies seek competitive advantage, and users demand practical reliability. These differing incentives create tensions that shape perceptions of a benchmark's trustworthiness (see Appendix A for further discussion). The central question, therefore, is not whether benchmarks should be public or private but rather which components should be disclosed to

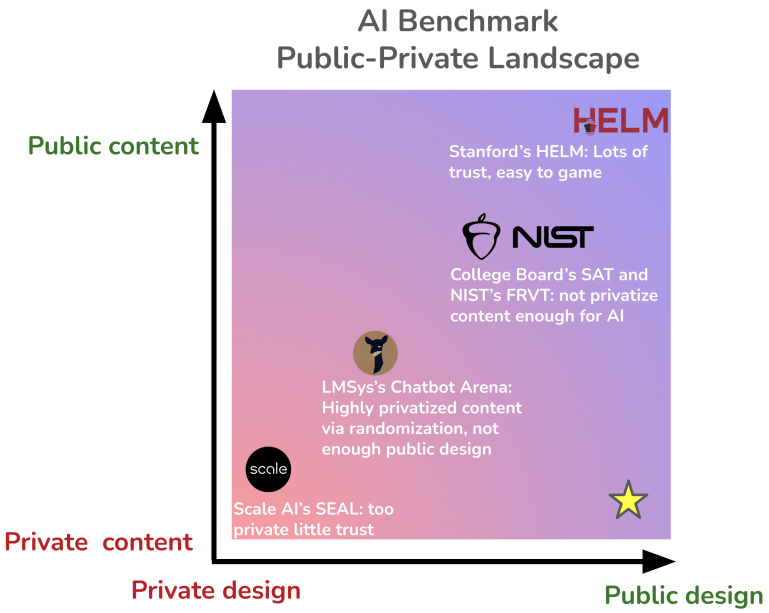

Figure 1: AI Benchmark Public–Private Landscape. The horizontal axis represents the openness of the benchmark design process; the vertical axis reflects the public availability of benchmark content. HELM occupies the top-right quadrant with both public content and public design - maximizing transparency but susceptible to contamination. In contrast, Scale AI's SEAL resides in the bottom-left quadrant, offering strong protection against contamination through full privatization, but at the cost of trust and reproducibility. Chatbot Arena features private content via dynamic, user-supplied prompts, whereas NIST's FRVT and the College Board's SAT demonstrate more public design through robust validity studies. Our proposed "North Star" lies in the bottom-right quadrant: benchmarks with private content to ensure evaluation integrity and public design to foster trust.

strike an appropriate
balance between evaluation integrity and stakeholder trust.

This paper argues that the future of robust and trustworthy AI evaluation lies in hybrid benchmarks with private content to prevent gaming and public design to ensure interpretability, replicability, and trust. This separation allows benchmark creators to preserve the validity of evaluations without sacrificing scientific transparency. By explicitly decoupling what is tested from how and why it is tested, hybrid benchmarks align with the practical needs of secure evaluation and the normative goals of open science.

The remainder of the paper explores the trade-offs associated with different positions along the public-private benchmark landscape. Section B presents some advantages of privatizing benchmark content from a theoretical mechanism design perspective. In Section 3, we turn to the case for public design, emphasizing their role in gaining trust. Section 4 presents five case studies illustrating how real-world benchmarks handle the separation of content and design, revealing varying degrees of alignment with the public design/private content framework. Finally, Section 6 offers recommendations for building AI evaluation systems that adopt this separation explicitly - keeping content private to preserve integrity while making design public to promote transparency and trust.

2 Content Privatization: Curbing Gaming Through Strategic Opacity

A well-documented challenge with public benchmarks is their susceptibility to contamination, which can ultimately compromise their usefulness as reliable evaluation tools. In this section, we formalize the case for privatizing elements of the benchmarking process by drawing on insights from theoretical models in microeconomics and mechanism design. Prior work in these fields has explored how randomness and opacity in evaluation mechanisms can reduce strategic manipulation by agents seeking to maximize performance under known rules. We review several of these models, which differ in technical assumptions but converge on a common conclusion: selective privatization can offer clear advantages from an incentive-alignment perspective. These theoretical results provide a principled foundation for viewing benchmark design not just as a technical artifact, but as a strategic interface between evaluators and those being evaluated. A more comprehensive discussion of these models is provided in Appendix B.

Analogies and insights from strategic classification A recent literature that has addressed the interaction between the opacity of a reward mechanism and gaming behavior is strategic classification [Hardt et al., 2016, Perdomo et al., 2020, Kleinberg and Raghavan, 2020]. While these works are primarily motivated by strategic interactions between an end-user and a predictive model, there are analogous insights that extend to AI evaluation. Braverman and Garg [2020] allude to the tradeoff between transparency in machine learning and strategic manipulation by saying, "One way to prevent such manipulation is by keeping the classification algorithms a secret, but this is not a practical solution to the problem, as some information is bound to leak over time and the transparency of these algorithms is a growing social concern." Below we explore connections between this literature and our setting of AI evaluation.

In essence, we may think of the "classifier" as the evaluation benchmark or a function $f : \mathcal{X} \to \mathbb{R}$ that maps an agent's (or model developer's) effort $x \in \mathcal{X}$ to a reward. In a game theoretic sense, we may think of the interaction in LLM evaluation as a Stackelberg game where the benchmark creator leads as the principal and the model developer follows as the agent. In strategic classification, the agent is allowed to change its features x at some cost to try to perform well on the classifier f. "Changing features" is analogous in our setting to a model developer adapting a model to the evaluation benchmark, via, e.g., post-training, and we can think of x as analogous to an instantiation of model parameters. The analogous question, then, is whether the evaluation benchmark induces gaming or improvement behavior on the part of the model developer.

Braverman and Garg [2020] show that adding randomness to the evaluation function f is a necessary condition to maximize the efficiency of the classification process (accuracy of the classifier minus the cost of changing features). The randomness that they consider is thinking of f as a "probabilistic classifier," where $f(x)$ is a random variable for a given x. In LLM evaluation, this means that a model developer would get a random score from the benchmark for a given submitted model x. This would be true of an evaluation benchmark with a dynamic, hidden test set, where submitting the same model twice might yield different results. The test set being hidden is important as if it is publicly

known to the model developer, there would no longer be randomness from the perspective of the model developer.

Strategic ambiguity and strategic opacity in contract theory. Prior to strategic classification, there is a long history of contract theory in economics, which studies models where a principal defines rewards to an agent based on the agent's performance. In our setting, we think of the principal as an entity who either is the benchmark creator, or works closely together with the benchmark creator (e.g. a government organization), to define benchmarks and rewards for the agent based on benchmark performance. The agent, or model developer, decides how to invest their model building effort and receives rewards as a function of benchmark performance as defined by the principal. An example of rewards coming from benchmark performance include the boost in attention and investment that can come from being at the top of the leaderboard, or a penalty (or lack thereof) for receiving a poor grade on a safety benchmark [Ghosh et al., 2025].

Considering the hidden (but specified) setting, strategic opacity has also been explored by Ederer et al. [2018], who consider a model where "'opacity' corresponds to a lack of transparency about the weights on performance indicators that are used to determine rewards." In our setting, this would be analogous to the benchmark creator publishing a leaderboard with weights for individual metrics, but not revealing to the model developer what those weights are. Ederer et al. [2018] focus on an "ex-ante randomization" mechanism, where the principal tells the agent that they will randomize between two possible sets of weights. This causes the agent to have to diversify their effort across multiple metrics. They show that this ex-ante randomization mechanism can dominate the best transparent scheme when the agent has more information than the principal about their comparative advantages between metrics and the agent is also risk-averse.

Summary. Many models have explored the consequences of (lack of) content transparency in evaluation. Each model subtly differs in how a "private metric" is defined, and the conclusions can be susceptible to modeling assumptions such as the game sequence, information asymmetries, and action spaces, i.e., *who knows what and when*. Still, across these many different models, a common theme is that strategic advantages exist to hide parts of the evaluation process. Of course, depending on the modeling assumptions, there can be advantages to transparency as well, but this depends on whether one assumes the agents are gaming or "improving" [Bechavod et al., 2022]. To curb unwanted *gaming* behavior from the model developer, most models point to opacity and randomness being a useful tool.

3 Design Publicity: Fostering Trust Through Transparency

Why Transparent Design Matters. Transparency in benchmark design - including the measured constructs, scoring criteria, and validation methodology - is essential for establishing trust, reproducibility, and scientific credibility. When these aspects are opaque, users and researchers cannot assess whether benchmarks measure what they claim. For example, a benchmark labeled as a general intelligence test may emphasize narrow skills such as code generation or factual recall, leading to inflated perceptions of model capability. Without visibility into design intent and validation, such mismatches remain hidden. Design opacity also inhibits scientific diagnosis. Researchers cannot analyze model failures, evaluate fairness, or challenge benchmark assumptions. Key psychometric dimensions depend on external scrutiny, such as construct validity, reliability, and bias. As a result, closed design limits interpretability and may distort how progress in AI is perceived.

Trust Risks from Opacity and Fragmentation. Opaque benchmark design also opens the door to economic and institutional conflicts of interest. When private entities manage both benchmark creation and model evaluation - often monetizing access or offering consulting services - there is a risk that the benchmark is subtly shaped to favor paying clients or affiliated models [Wiggers, 2025a, Bansal and Maini, 2025]. Transparency around who defines the benchmark and how its components are chosen is necessary to maintain credibility and fairness. The increasing number of private, closed benchmarks also risks fragmenting the field. Models may rank highly on one leaderboard but poorly on another, creating inconsistency and confusion. While diversity in evaluation is valuable, a lack of coordination and shared design principles can hinder meaningful comparisons. Transparent design serves as a common language across benchmarks - facilitating alignment, reducing redundancy, and fostering community trust.

Summary. Several recent efforts have demonstrated that transparency in benchmark design - through detailed documentation, third-party audits, and open-source evaluation frameworks (e.g., OpenAI Evals [Wiggers, 2023]) - can foster accountability even when test content remains private. By clearly articulating what is being measured, how scores are derived, and why the evaluation is valid, public design serves as a foundation for reproducible and trustworthy AI assessment.

4 Public and Private in Practice: Case Studies Across Domains

To guide future benchmark development, we propose a "North Star" model: benchmarks that maintain private content to preserve evaluation integrity while making their design fully public to foster transparency, trust, and scientific reproducibility. In this framework, questions, labels, and scoring thresholds remain hidden to reduce contamination, while the constructs being measured and validation studies are openly shared to enable critical scrutiny. To illustrate how real-world benchmarks align with – or diverge from – this North Star, we examine five recent examples across education, computer vision, and natural language processing. Each case study reflects a different point in the design–content disclosure space, shedding light on practical implementation choices and the extent to which modern benchmarks embody the hybrid model we advocate.

Table 1: Benchmark Comparison Along Content Privacy, Design Publicity, and Adoption Dimensions

	Content Privacy	**Design Publicity**	**Organizational Trust**	**Adoption**	**Update & Governance**
NIST FRVT	Medium-High (fully sequestered test sets)	Moderate (published protocols, limited visibility into task specifics)	High (government-backed, neutral)	Widely used in government, biometrics industry	Periodic evaluations, centralized expert oversight
Scale SEAL	High (private, undisclosed datasets and thresholds)	Low–Moderate (broad task categories public, specifics hidden)	Moderate (corporate-led, with trust concerns)	Growing, referenced in safety/alignment circles	Irregular updates, centralized governance with planned external review
Stanford HELM	Low (all prompts, outputs, and scoring public)	High (comprehensive scenario and metric documentation)	High (academic, peer-reviewed, open-source)	Widely used in academia and research benchmarking	Continuous, community-informed updates
Chatbot Arena	Medium–High (dynamic prompts; fixed public scoring function)	Moderate (open logs, transparent ranking pipeline)	Medium–High (transparent infra; data asymmetries present)	Popular in open-source and casual model comparison	Live, real-time updates; maintained by LMSys with community input

4.1 Example from the Education Domain: College Board's Scholastic Aptitude Test

The College Board's Scholastic Assessment Test (SAT) provides a well-established precedent for balancing content privacy and design transparency - the same tension at the heart of modern AI evaluation. As a mature and trusted high-stakes benchmark, the SAT illustrates key principles of the hybrid model we advocate: keeping content private to prevent gaming, while making the design public to maintain trust and accountability.

Content Privatization and Contamination Control. To preserve the integrity of test scores and prevent "teaching to the test," the SAT maintains strict control over its active test content. Test items are kept confidential, and the College Board enforces anti-disclosure policies and proctored testing conditions to deter cheating and limit exposure. This approach reflects a longstanding practice in human evaluation: keeping content private to ensure validity across repeated administrations.

Importantly, the threat model in AI evaluation differs substantially from that of human testing. Human test takers have limited memory and are unlikely to memorize or widely disseminate test content, making leakage relatively contained. In contrast, AI systems – particularly those accessed via APIs – require the full test input to be sent with each call. This means that every interaction creates an opportunity for the benchmark content to be logged, stored, or inadvertently incorporated into future training data. As a result, any exposed test item becomes a potential source of contamination. This heightened risk implies that effective AI benchmarks must enforce even stricter content privatization than human assessments like the SAT to preserve long-term evaluation integrity.

Design Publicity and User Trust. While content is kept private, the design of the SAT is extensively documented. The College Board publishes detailed test specifications, construct rationales, and validation studies, including longitudinal analyses demonstrating the SAT's predictive power for

college GPA across demographic groups [Marini et al., 2024]. This level of design transparency allows external stakeholders - such as educators, admissions officers, and researchers - to evaluate the benchmark's fairness, relevance, and alignment with educational goals. In addition, the SAT undergoes continual third-party review, with psychometricians monitoring item performance, equity metrics, and statistical consistency across administrations. Services like ACES further support independent validity checks. These efforts build organizational trust and public legitimacy, despite the secrecy of specific test items.

Takeaway. While not without limitations, the SAT offers a compelling model of benchmark: private content to protect score integrity and prevent contamination, combined with public design to ensure transparency, interpretability, and stakeholder trust. Though designed for human assessment, the SAT's hybrid approach provides a valuable reference point for emerging AI benchmarks, which must grapple with even greater risks of exposure. In the following case studies, we examine how modern AI benchmarks adopt - or deviate from - this model.

4.2 Case Study 1: NIST's Face Recognition Vendor Test

The National Institute of Standards and Technology (NIST) has long been a neutral evaluator of biometric and security technologies. Its Face Recognition Vendor Test (FRVT) provides an example of rigorous, independent benchmarking.

Content Privatization and Contamination Control. NIST conducts FRVT using sequestered test datasets kept private from participants to prevent tuning algorithms to the test. In the 2006 FRVT, for instance, performance was measured with sequestered data and all vendors were tested on a standardized dataset and environment to ensure equal treatment and prevent unfair advantages or overfitting to the evaluation set [National Institute of Standards and Technology, 2006]. To maintain this protective barrier, NIST keeps test data secret not only during evaluation but often even after the evaluation is complete, preserving a "level playing field" for all participants. This stringent data sequestration strategy forms the cornerstone of contamination control, ensuring that algorithmic improvements reflect genuine advances rather than dataset-specific optimizations.

Design Publicity and User Trust. While test data remains private, NIST ensures transparency through comprehensive public reporting of outcomes via detailed reports and leaderboards, allowing industry and policymakers to assess algorithm performance [Phillips et al., 2010]. The credibility of these results stems from NIST's position as a U.S. government institute, which positions it as an independent and authoritative arbiter without commercial stakes in the outcomes. NIST reinforces this trust through well-defined protocols and massive, representative datasets covering diverse demographics and real-world conditions. It uses standard metrics (e.g., false match/false non-match rates) [Paravision, 2023] and often runs multiple tracks (e.g., 1:1 verification, 1:N identification, masked faces) to thoroughly vet algorithms [NIST, 2010]. The organization's objectivity is further demonstrated through reports that not only rank vendors but also analyze errors and highlight critical issues like demographic bias [NIST, 2019]. Importantly, NIST's credibility is bolstered by external oversight, as its methods and reports undergo scrutiny from both the research community and agencies that rely on these evaluations, while maintaining an open participation policy that invites any vendor to submit algorithms [NIST, 2023].

Takeaway. FRVT mirrors a classic legacy approach that combines private, secure testing with public, rigorous reporting, anchored by a reputation for impartiality. This careful balance of data protection and transparency has established NIST's evaluation framework as the gold standard in face recognition assessment, demonstrating how independent benchmarking can maintain both scientific rigor and stakeholder trust in an increasingly competitive technological landscape.

4.3 Case Study 2: Scale AI's Safety, Evaluations, and Alignment Leaderboard

As AI models proliferate, Scale AI, a leading AI infrastructure company specializing in data annotation and model evaluation services, has launched the SEAL initiative to provide robust third-party evaluations for large language models. SEAL's approach in many ways mirrors NIST-style benchmarking for AI: it emphasizes private, expert-curated test data to prevent manipulation, combined with transparent methodology and independent oversight to ensure trust [Scale AI, 2024a].

Content Privatization and Contamination Control. Scale's SEAL employs curated, unpublished test sets to prevent gaming and ensure models are not fine-tuned to the evaluation [Scale AI, 2024b]. By limiting exposure - even blocking models that may have seen prompts via APIs - SEAL aims to provide an unbiased, tamper-proof performance measure that addresses the persistent contamination issues plaguing public benchmarks. However, this approach creates inherent tension between evaluation security and scientific reproducibility, as researchers cannot independently verify results, conduct error analyses on specific test cases, or validate claims that models have not been inadvertently exposed to test data [CodeCompass, 2024]. This fundamental trade-off between contamination control and open scientific inquiry represents a core challenge in contemporary AI evaluation frameworks.

Design Publicity and User Trust. Recognizing that secrecy alone can raise trust concerns, Scale AI pairs its private evaluation approach with transparent methodology, domain-specific metrics, and plans for third-party review. SEAL uses domain experts to design and review each test set, tailoring evaluations to reflect real-world skills such as unit tests for coding and GSM8K-style mathematics problems [Scale AI, 2024a]. This specialization enhances validity, much like the College Board's use of subject-matter experts. Moreover, SEAL openly shares its evaluation methodology, highlights limitations of community benchmarks (e.g., contamination), and provides contextual insights beyond scores [Scale AI, 2024a]. Though test content remains private, transparency around process and goals invites scrutiny. SEAL's independence and endorsements from government and AI safety experts strengthen its credibility. Nevertheless, Scale AI's dual role as training data curator and evaluator compromises test validation through potential overlap in task patterns between evaluation metrics and training data provided to clients, creating artificial performance advantages. Additionally, human annotators' involvement in data creation and model assessment introduces potential output biases that undermine the objectivity of private evaluation frameworks [Bansal and Maini, 2025].

Takeaway. Scale AI's SEAL benchmark represents a sophisticated attempt to navigate the fundamental tension between evaluation security and scientific reproducibility in AI assessment. The success of this approach ultimately depends on whether transparent methodology and institutional oversight can substitute for the direct reproducibility that characterizes traditional scientific benchmarking. This design choice reflects broader challenges in AI evaluation, where the rapid pace of model development and the high stakes of performance rankings create pressure for both rigorous assessment and protective secrecy.

4.4 Case Study 3: LMSys's Chatbot Arena

Chatbot Arena crowdsources test prompts and evaluation by pitting AI models in anonymous head-to-head chat battles judged by users. Voters see side-by-side responses, unaware of which model is which, ensuring fairness. Thousands of such interactions generate Elo scores, producing a public leaderboard based on collective human judgment [Chiang et al., 2024].

Content Privatization and Contamination Control. Chatbot Arena uses crowdsourced to dynamically generate test inputs, and randomized matchups to ensure that the prompt content is highly private and constantly evolving, reducing the risk of memorization and contamination. However, beyond prompts and responses, the aggregation mechanism - specifically, the Elo rating system used to compute model rankings - is itself part of the benchmark content. Because this scoring function is publicly known and fixed, it creates an incentive for participants to optimize specifically for Arena dynamics, rather than general model quality. Moreover, the lack of access control or versioning around Elo makes it vulnerable to strategic exploitation and meta-optimization over time. Further undermining content neutrality, some providers exploit their access advantage: Meta, for instance, reportedly tested 27 versions of LLaMA-4 on the platform, publishing only the strongest while withholding underperformers [Wiggers, 2025b]. Such selective reporting introduces contamination risks not through exposure, but through control over what gets seen, which can skew community perception and benchmarking integrity.

Design Publicity and User Trust. Despite these limitations, Chatbot Arena is widely trusted in the open-source community due to its commitment to design transparency. All prompts, model responses, voting data, and evaluation rules are released, and the system allows any user to contribute prompts or submit models. The anonymized, randomized evaluation setup minimizes bias and supports replicability. Community members can audit every stage of the evaluation pipeline, from raw data to ranking outputs. Yet trust in design does not fully compensate for structural asymmetries in content exposure. Proprietary models from leading companies receive disproportionately high sampling

rates and persist longer in the Arena than open-weight alternatives. Recent work shows that these disparities result in distorted performance metrics and reinforce leaderboard dominance [Singh et al., 2025]. The design is open, but the execution favors those with privileged access - limiting the extent to which the benchmark truly reflects general model capability.

Takeaway. Chatbot Arena highlights both the promise of public design and the limits of public content. Its open design (e.g., evaluation pipeline and documentation) set a high bar for trustworthiness and community participation. However, its openly known aggregation mechanism - Elo scoring - functions as part of the evaluative content and is thus vulnerable to strategic overfitting. Combined with selective model disclosure and exposure imbalances, this creates the risk that leaderboard rankings reflect optimization for Arena dynamics rather than general capability. This case demonstrates that content privatization must include not only prompts and labels but also aggregation rules and reporting protocols - any component that shapes how model outputs are turned into performance signals.

4.5 Case Study 4: Stanford's Holistic Evaluation of Language Models (HELM)

Content Privatization and Contamination Control. HELM was designed with a commitment to transparency, making almost all components - including prompts, model outputs, and evaluation results - publicly available. Unlike traditional benchmarks that rely on content secrecy to prevent overfitting, HELM embraces openness as a way to foster reproducibility and shared understanding. This choice, however, makes HELM more susceptible to contamination: because test content is public, models can be fine-tuned or indirectly exposed to evaluation data through pretraining. This increases the risk that benchmark scores reflect memorization rather than generalization. Nonetheless, HELM acknowledges situations where content privacy is necessary. For example, in the medical domain, HELM-Medical withholds certain test items to comply with health data regulations and protect patient confidentiality. These cases demonstrate that even transparency-first frameworks must selectively privatize content when ethical or legal obligations demand it - offering a pragmatic balance within an otherwise public model.

Design Publicity and User Trust. Where HELM excels is in design transparency. The benchmark defines a broad taxonomy of 42 scenarios and 7 evaluation metrics - including fairness, calibration, and robustness - designed to capture holistic model behavior across tasks such as summarization, code generation, and question answering [Liang et al., 2023]. Each model is evaluated under controlled conditions, with standardized prompt formatting and clearly documented adaptation procedures. This consistency enables meaningful, apples-to-apples comparisons across models - addressing a longstanding problem in fragmented benchmarking. Crucially, HELM publishes all evaluation inputs, outputs, scoring scripts, and limitations. This level of openness invites independent auditing, reproducibility, and community engagement. In contrast to private leaderboards that offer opaque scores, HELM's radical transparency cultivates user trust through methodological clarity. Its academic foundation, open-source infrastructure, and explicit documentation of gaps reinforce its credibility as a research-driven evaluation platform.

Takeaway. HELM demonstrates what maximal public design can look like in practice, even at the cost of increased content exposure. While this openness introduces risks of contamination - especially in the LLM era where models may memorize public benchmarks - it also sets a strong precedent for transparency, reproducibility, and trust. As such, HELM aligns closely with one half of the North Star: public design. However, its approach to content reveals a trade-off: in seeking to empower community insight, it sacrifices long-term protection against memorization. This case highlights the need for selective content privatization - particularly in sensitive domains - while preserving HELM's strengths in open, standardized design.

4.6 Comparison and Lessons: Implementing Public Design and Private Content

The case studies illustrate a range of approaches to benchmark disclosure, but they converge on two key lessons for building trustworthy and resilient evaluation systems that align with the North Star model: private content to preserve evaluation integrity and public design to foster trust.

Lesson 1: Private Content Must Be Actively Maintained to Prevent Contamination Simply hiding benchmark data is not enough - AI models can memorize exposed content through pretraining or repeated API access. Effective content privatization requires active maintenance: rotating hidden datasets, dynamically generating new items, or limiting repeated exposure. This approach mirrors

practices in standardized testing (e.g., the SAT) and high-stakes AI evaluations (e.g., Scale SEAL), where test security is preserved through restricted access, refresh cycles, and distribution control.

Lesson 2: Public Design Is Essential for Credibility and Community Trust Design transparency, including clearly defined constructs and validation studies, builds legitimacy, even when test content must remain hidden. Benchmarks like HELM and NIST demonstrate that sharing how and why evaluation is done enables meaningful scrutiny, reproducibility, and stakeholder engagement. Without design publicity, closed benchmarks risk being dismissed as unaccountable or biased.

5 Alternative View: The Case for Full Transparency or Full Secrecy

While we argue for a hybrid model - private content with public design - other perspectives favor more extreme positions along the transparency spectrum. Some advocate for full transparency in benchmarking, including releasing all evaluation content and scoring procedures. From this perspective, openness is necessary not just for trust, but for democratizing access, enabling reproducibility, and supporting broader innovation. Proponents argue that any hidden component risks bias, reduces community oversight, and favors actors with privileged access to the evaluation process. In this view, contamination risks should be managed not through secrecy, but through continual benchmark expansion and adaptation. On the other end of the spectrum, some argue that benchmark secrecy is the only viable way to preserve signal integrity in the era of foundation models with near-infinite capacity to memorize data. From this perspective, transparency invites gaming, and reproducibility is a secondary concern compared to ensuring that evaluations reflect true generalization. This view aligns with practices in high-stakes testing and industrial AI evaluation, where test items and scoring rubrics are tightly controlled.

Our hybrid model acknowledges these tensions. We agree that complete transparency can foster scientific collaboration and that privacy is essential for preventing evaluation leakage. We contend that these goals are not mutually exclusive: strategic separation - keeping test content private while making design public - can maintain evaluation fidelity and stakeholder trust. We see this as a pragmatic middle path that draws strength from both extremes while avoiding their respective pitfalls.

6 Conclusion

The future of AI evaluation requires rethinking the relationship between transparency and integrity. Public benchmarks, while instrumental in driving early progress, increasingly suffer from contamination and gaming as models scale and memorization becomes unavoidable. Private benchmarks, in contrast, protect test integrity but often lack the transparency necessary for scientific scrutiny and public trust. This paper advances a simple but powerful thesis: AI benchmarks should combine private content to prevent recall-based contamination with public design to ensure transparency, accountability, and stakeholder trust. This separation - between what is tested and how and why it is tested - offers a principled way to reconcile the tensions between secrecy and openness that define today's benchmarking landscape.

Our analysis draws on two complementary perspectives. From an incentive-theoretic standpoint, private content helps preserve the integrity of evaluation by reducing opportunities for gaming. From a trust and accountability perspective, public design - transparent disclosure of constructs, scoring methods, and validation procedures - is essential for ensuring that benchmark results are interpretable, reproducible, and legitimate. While no single benchmark we examined perfectly realizes this ideal, our case studies - spanning the SAT, NIST FRVT, HELM, and others - illuminate the principles in practice. Each reveals how different domains approach the balance between content secrecy and design transparency, offering concrete insights into how future benchmarks might better align with the private content/public design model. Moving forward, we recommend two core design principles:

- Content should be private, dynamic, and access-controlled, minimizing the risk of contamination through memorization or test leakage.
- Design should be fully public, including published construct rationales and validation studies, enabling external scrutiny and interpretability.

This public design and private content framework provides a robust foundation for benchmark development in a rapidly evolving AI landscape. Benchmarks adhering to these principles can

maintain long-term validity, support scientific reproducibility, and foster the multi-stakeholder trust necessary for responsible deployment. Rather than choosing between public or private, the field should converge on what to keep hidden, and what must remain visible.

References

Scale AI. Seal leaderboards, 2025. URL https://scale.com/leaderboard.

Hritik Bansal and Pratyush Maini. Peeking behind closed doors: Risks of llm evaluation by private data curators, 2025. URL https://arxiv.org/abs/2503.04756.

Yahav Bechavod, Chara Podimata, Steven Wu, and Juba Ziani. Information discrepancy in strategic learning. In *International Conference on Machine Learning*, pages 1691–1715. PMLR, 2022.

B Douglas Bernheim and Michael D Whinston. Incomplete contracts and strategic ambiguity. *American Economic Review*, pages 902–932, 1998.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, et al. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*, 2024.

M Braverman and S Garg. The role of randomness and noise in strategic classification. In *1st Symposium on Foundations of Responsible Computing (FORC 2020)*, volume 156, 2020.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2403.04132.

Francois Chollet, Mike Knoop, Gregory Kamradt, and Bryan Landers. Arc prize 2024: Technical report. *arXiv preprint arXiv:2412.04604*, 2024.

CodeCompass. The challenges of building effective llm benchmarks, 5 2024. URL https://codecompass00.substack.com/p/llm-evaluation-leaderboards.

Ben Dickson. Why we must rethink AI benchmarks. TechTalks, December 2021. URL https://bdtechtalks.com/2021/12/06/ai-benchmarks-limitations/.

Ben Dickson. Rethinking AI benchmarks: A new paper challenges the status quo of evaluating artificial intelligence, June 2023. URL https://venturebeat.com/ai/rethinking-ai-benchmarks-a-new-paper-challenges-the-status-quo-of-evaluating-artificial-intell

Florian Ederer, Richard Holden, and Margaret Meyer. Gaming and strategic opacity in incentive provision. *The RAND Journal of Economics*, 49(4):819–854, 2018.

Ganesh Ghalme, Vineet Nair, Itay Eilat, Inbal Talgam-Cohen, and Nir Rosenfeld. Strategic classification in the dark. In *International Conference on Machine Learning*, pages 3672–3681. PMLR, 2021.

Shaona Ghosh, Heather Frase, Adina Williams, Sarah Luger, Paul Röttger, Fazl Barez, Sean McGregor, Kenneth Fricklas, Mala Kumar, Kurt Bollacker, et al. Ailuminate: Introducing v1. 0 of the ai risk and reliability benchmark from mlcommons. *arXiv preprint arXiv:2503.05731*, 2025.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122, 2016.

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*, IUI '25, page 1032–1047, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713064. doi: 10.1145/3708359.3712152. URL https://doi.org/10.1145/3708359.3712152.

Horace He. I suspect gpt-4's performance is influenced by data contamination, at least on codeforces. of the easiest problems on codeforces, it solved 10/10 pre-2021 problems and 0/10 recent problems. this strongly points to contamination. https://x.com/cHHillee/status/1635790330854526981, March 2023. Tweet.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Dawn Tang, Rishav Desai, Mantas Zhu, Sam Krasheninnikov, Dawn Song, et al. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2021. URL https://arxiv.org/abs/2009.03300.

Jennifer Hsia, Danish Pruthi, Aarti Singh, and Zachary C Lipton. Goodhart's law applies to nlp's explanation benchmarks. *arXiv preprint arXiv:2308.14272*, 2023.

Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation (TEAC)*, 8(4):1–23, 2020.

Bernard Koch, Emily Denton, Alex Hanna, and Jacob G. Foster. Reduced, reused and recycled: The life of a dataset in machine learning research, 2021. URL https://arxiv.org/abs/2112.01716.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models, 2023. URL https://arxiv.org/abs/2211.09110.

Jessica P. Marini, Paul A. Westrick, Linda Young, and Emily J. Shaw. Sat® score relationships with college gpa: First-year through fourth-year cumulative gpa, October 2024. URL https://research.collegeboard.org/media/pdf/SAT%28R%29%20Score%20Relationships%20with%20College%20GPA.pdf. Research Report.

Arvind Narayanan and Sayash Kapoor. Gpt-4 and professional benchmarks: the wrong answer to the wrong question. *AI Snake Oil*, 20:2023, 2023.

National Institute of Standards and Technology. Face recognition vendor test (frvt) 2006. https://www.nist.gov/itl/iad/image-group/face-recognition-vendor-test-frvt-2006, 2006.

NIST. Face recognition vendor test (frvt). https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt, 2010.

NIST. Nist study evaluates effects of race, age, sex on face recognition software, 2019. URL https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software.

NIST. Face recognition vendor test (frvt) participation agreement. Online, 2023. URL https://pages.nist.gov/frvt/agreements/frvt_participation_agreement.pdf.

OpenAI. evals. GitHub repository, 2023. URL https://github.com/openai/evals.

OpenAI. metrics.py file from openai evals repository. GitHub repository, 2025. URL https://github.com/openai/evals/blob/main/evals/metrics.py. Accessed: 2025-05-12.

Paravision. Introduction to NIST FRVT. https://www.paravision.ai/news/introduction-to-nist-frvt/, 2023.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.

Phan et. al. Humanity's last exam, 2025. URL https://arxiv.org/abs/2501.14249.

P. Jonathon Phillips, W. Todd Scruggs, Alice J. O'Toole, Patrick J. Flynn, Kevin W. Bowyer, Cathy L. Schott, and Matthew Sharpe. Frvt 2006 and ice 2006 large-scale experimental results. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):831–846, 2010. doi: 10.1109/TPAMI.2009.59.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. Ai and the everything in the whole wide world benchmark, 2021. URL https://arxiv.org/abs/2111.15366.

Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. Outsider oversight: Designing a third party audit ecosystem for ai governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 557–571, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392471. doi: 10.1145/3514094.3534181. URL https://doi.org/10.1145/3514094.3534181.

Tanmay Rajore, Nishanth Chandran, Sunayana Sitaram, Divya Gupta, Rahul Sharma, Kashish Mittal, and Manohar Swaminathan. Truce: Private benchmarking to prevent contamination and improve comparative evaluation of llms, 2024. URL https://arxiv.org/abs/2403.00393.

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J Kochenderfer. Betterbench: Assessing ai benchmarks, uncovering issues, and establishing best practices. *Advances in Neural Information Processing Systems(NeurIPS)*, 2024.

Scale AI. Scale's seal research lab launches expert-evaluated llm leaderboards, May 2024a. URL https://scale.com/blog/leaderboard.

Scale AI. Seal leaderboards announcement, 2024b. URL https://scale.com/leaderboard.

Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'Souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah Smith, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. The leaderboard illusion, 2025. URL https://arxiv.org/abs/2504.20879.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi, editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446/.

Angelina Wang, Aaron Hertzmann, and Olga Russakovsky. Benchmark suites instead of leaderboards for evaluating ai fairness. *Patterns*, 2024a.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models, 2024b. URL https://arxiv.org/abs/2306.11698.

Jason Wei. Successful language model evals, 2023. URL https://www.jasonwei.net/blog/evals.

Kyle Wiggers. With evals, openai hopes to crowdsource ai model testing. *TechCrunch*, 3 2023. URL https://techcrunch.com/2023/03/14/with-evals-openai-hopes-to-crowdsource-ai-model-testing/.

Kyle Wiggers. Ai benchmarking organization criticized for waiting to disclose funding from openai, 2025a. URL https://techcrunch.com/2025/01/19/ai-benchmarking-organization-criticized-for-waiting-to-disclose-funding-from-openai/.

Kyle Wiggers. Study accuses LM Arena of helping top AI labs game its benchmark. TechCrunch, April 2025b. URL https://techcrunch.com/2025/04/30/study-accuses-lm-arena-of-helping-top-ai-labs-game-its-benchmark/.

A The Stakeholders' Incentives in AI Benchmarking

Benchmarks are not neutral artifacts - they are created, adopted, and interpreted by communities with diverging goals and values. These divergent stakeholder interests directly shape whether benchmarks trend toward public or private designs. Some care about transparency, others about competitive advantage or regulatory compliance. Their incentives often shape what is measured, how success is defined, and which benchmarks gain traction. In this section, we identify three major stakeholder groups in the AI benchmarking ecosystem, in line with [Reuel et al., 2024], each comprising multiple distinct personas with varied incentives: (1) end-users and deployers, (2) model developers, and (3) benchmark creators.

End-Users and Deployers End-users and deployers focus on real-world utility, reliability, and domain-specific performance [Hardy et al., 2025], valuing metrics like factuality, latency, and user satisfaction over leaderboard rankings. They prefer independent evaluations that reflect practical use cases [Hardy et al., 2025], with recent efforts like Scale AI's SEAL addressing this need.

Academic Researchers Academic researchers continue to be significant contributors to benchmark creation, especially in the LLM era. They value transparency and reproducibility, with venues like NeurIPS Datasets and Benchmarks track facilitating academic contributions. Researchers often face institutional pressure to demonstrate improvements on existing benchmarks rather than questioning fundamental evaluation paradigms. While industry players increasingly influence benchmark creation through access to proprietary data and computing, academic-industry collaborations have emerged as a productive middle ground for developing more comprehensive evaluations [Wang et al., 2024b].

These groups share a demand for transparent, trustworthy benchmarks resistant to gaming.

Model Developers (e.g., OpenAI, Google) often have incentives to achieve state-of-the-art results on benchmarks [Hardy et al., 2025]. High leaderboard rankings are a marketing tool and credibility signal, attracting customers and investors [Hardy et al., 2025]. Internally, these companies also use evaluations to guide model development, but there can be tension between thorough, critical assessment and the rush to publish impressive numbers [Hardy et al., 2025]. Model developers value benchmarks as a competitive battleground, which can lead to rapid progress and potential overfitting to metrics [Hsia et al., 2023].

Benchmark Creators includes academic teams, non-profits, and startups that design and maintain benchmarks. Their goals range from driving research and spotlighting underserved capabilities (e.g., GLUE, SuperGLUE, MMLU) to building businesses around evaluation (e.g., Scale AI's SEAL) [Biderman et al., 2024, AI, 2025]. While academic efforts focus on setting research targets, transparency and community relevance [Hardy et al., 2025], private providers – both non-profit and for-profit entities – aim to offer evaluation-as-a-service and promote themselves as neutral referees [Chiang et al., 2024, Chollet et al., 2024, AI, 2025, Ghosh et al., 2025]. Their credibility depends on fairness and the evaluation quality, but they face tensions around transparency, commercialization, and potential conflicts of interest [Wiggers, 2025a]. Ultimately, benchmark creators seek wide adoption and trust, balancing openness with the need to prevent test gaming [Reuel et al., 2024]. The rise of dedicated tracks like NeurIPS Datasets and Benchmarks has further incentivized benchmark creation as a scholarly contribution, helping shape community standards around evaluations.

Differing incentives across stakeholders can collide and manifest in trade-offs in design choices. One of the most consequential outcomes of these incentive tensions is the privatization of benchmarking itself. As the field matures and competition intensifies, we have seen at least a partial shift from traditional public benchmarks toward private or proprietary evaluation frameworks. This move is often justified by concerns about benchmark contamination and metric gaming [Hardy et al., 2025], but it also reflects the strategic interests of key stakeholders, particularly industry actors [Raji et al., 2022]. In the following section, we examine how this privatization reshapes the benchmarking landscape and what trade-offs it entails for the broader AI community. Understanding these stakeholder dynamics is essential for evaluating the trade-offs between public and private benchmarks.

B Incentive-theoretic Advantages of Benchmark Privatization

A known issue with public benchmarks is the gaming and saturation effects that render the benchmarks useless. Here we formalize the argument for privatizing parts of the benchmarking process through

referencing existing work in theoretical modeling in economics and mechanism design. Theoretical modeling in microeconomics and mechanism design has studied the role of randomness and opacity in evaluation and whether this can curb strategic gaming. Here, we review several such models of opacity in evaluation, which differ in subtle ways but ultimately share a similar bottom line: benchmark privatization can have advantages from an incentive-theoretic standpoint.

Analogies and insights from strategic classification A recent literature that has addressed the interaction between the opacity of a reward mechanism and gaming behavior is strategic classification [Hardt et al., 2016, Perdomo et al., 2020, Kleinberg and Raghavan, 2020]. While these works are primarily motivated by strategic interactions between an end-user and a predictive model, there are analogous insights that extend to AI evaluation. Braverman and Garg [2020] allude to the tradeoff between transparency in machine learning and strategic manipulation by saying, "One way to prevent such manipulation is by keeping the classification algorithms a secret, but this is not a practical solution to the problem, as some information is bound to leak over time and the transparency of these algorithms is a growing social concern." Below we explore connections between this literature and our setting of AI evaluation.

In essence, we may think of the "classifier" as the evaluation benchmark or a function $f : \mathcal{X} \to \mathbb{R}$ that maps an agent's (or model developer's) effort $x \in \mathcal{X}$ to a reward. In a game theoretic sense, we may think of the interaction in LLM evaluation as a Stackelberg game where the benchmark creator leads as the principal and the model developer follows as the agent. In strategic classification, the agent is allowed to change its features x at some cost to try to perform well on the classifier f. "Changing features" is analogous in our setting to a model developer adapting a model to the evaluation benchmark, via e.g., post-training, and we can think of x as analogous to an instantiation of model parameters. The analogous question, then, is whether the evaluation benchmark induces gaming or improvement behavior on the part of the model developer.

Braverman and Garg [2020] show that adding randomness to the evaluation function f is a necessary condition to maximize the efficiency of the classification process (accuracy of the classifier minus the cost of changing features). The randomness that they consider is thinking of f as a "probabilistic classifier," where $f(x)$ is a random variable for a given x. In LLM evaluation, this means that a model developer would get a random score from the benchmark for a given submitted model x. This would be true of an evaluation benchmark with a dynamic, hidden test set, where submitting the same model twice might yield different results. The test set being hidden is important, as if it is publicly known to the model developer, there would no longer be randomness from the perspective of the model developer.

In addition to randomized classifiers, the strategic classification literature also explores settings where a fixed classifier is hidden [Ghalme et al., 2021, Bechavod et al., 2022]. Both of these works indicate that opacity can sometimes have disadvantages. Bechavod et al. [2022] quantify disparities in improvement behavior when multiple agents can band together to infer a private classifier's parameters. The analogy to our setting is if there is a private benchmark with a fixed test set, multiple model developers could band together to try to infer test set characteristics from the outcomes that they achieve for different submitted models x. Bechavod et al. [2022] show that different subpopulations of agents might share different information, and thus have disparate abilities to improve. Ghalme et al. [2021] quantify a "price of opacity," and show that opacity can actually backfire and lead to a less accurate classifier, particularly when the principal developing the classifier does not know the agent's estimate for the classifier. This is analogous to the benchmark creator not knowing how the model developers *think* their evaluation benchmark behaves, which can come around to hurt the validity of the evaluation.

Strategic ambiguity and strategic opacity in contract theory. Prior to strategic classification, there is a long history of contract theory in economics, which studies models where a principal defines rewards to an agent based on the agent's performance. In our setting, we think of the principal as an entity who either is the benchmark creator, or works closely together with the benchmark creator (e.g. a government organization), to define benchmarks and rewards for the agent based on benchmark performance. The agent, or model developer, decides how to invest their model building effort and receives rewards as a function of benchmark performance as defined by the principal. An example of rewards coming from benchmark performance include the boost in attention and investment that can come from being at the top of the leaderboard, or a penalty (or lack thereof) for receiving a poor

grade on a safety benchmark [Ghosh et al., 2025]. In contract theory, we review two concepts that are related to private benchmarks: *strategic ambiguity* and *strategic opacity*.

Strategic ambiguity refers to purposely defining an incomplete contract, which is a contract that does not define the reward for the agent in all states of the world, instead leaving some aspects open for renegotiation later on. This is often due to some important aspects of the agent's performance being unverifiable. For example, the benchmark creator could tell the model developer that they will be evaluated on some dimensions for safety, but not commit to a threshold to yield a letter grade of A. Bernheim and Whinston [1998] explore the value of strategic ambiguity in contracting, and show that "once some aspects of performance are unverifiable, it is often optimal to leave other verifiable aspects of performance unspecified."

A subtlety here is that *unspecified* is not necessarily the same as *hidden*. An aspect of the reward could be unspecified and not hidden: the benchmark creator could tell the model developer that *some* threshold will be used to determine their safety grade, but simply say that this threshold has not been determined yet. Conversely, an aspect of the reward could be both specified and hidden: the threshold is specified ahead of time, but not revealed to the model developer. These two settings can lead to different equilibria and incentives for the model developer. Intuitively, if the model developer knows that the threshold is specified ahead of time, they can try to guess what that threshold is likely to be. Bernheim and Whinston [1998] show that sometimes it can be advantageous to purposely leave some aspects of the reward unspecified.

Considering the hidden (but specified) setting, strategic opacity has also been explored by Ederer et al. [2018], who consider a model where "'opacity' corresponds to a lack of transparency about the weights on performance indicators that are used to determine rewards." In our setting, this would be analogous to the benchmark creator publishing a leaderboard with weights for individual metrics, but not revealing to the model developer what those weights are. Ederer et al. [2018] focus on an "ex-ante randomization" mechanism, where the principal tells the agent that they will randomize between two possible sets of weights. This causes the agent to have to diversify their effort across multiple metrics. They show that this ex-ante randomization mechanism can dominate the best transparent scheme when the agent has more information than the principal about their comparative advantages between metrics and the agent is also risk-averse.

Summary. Many models have explored the consequences of (lack of) transparency in evaluation. Each model subtly differs in how a "private metric" is defined, and the conclusions can be susceptible to modeling assumptions such as the game sequence, information asymmetries, and action spaces, i.e., *who knows what and when*. Still, across these many different models, a common theme is that strategic advantages exist to hide parts of the evaluation process. Of course, depending on the modeling assumptions, there can be advantages to transparency as well, but this depends on whether one assumes the agents are gaming or "improving" [Bechavod et al., 2022]. To curb unwanted *gaming* behavior from the model developer, most models point to opacity and randomness being a useful tool.

C Supplementary Case Studies: OpenAI Evals

In contrast to private test sets, OpenAI's Evals project represents a contemporary move toward open and crowdsourced benchmarking. Launched in 2023 alongside GPT-4, OpenAI Evals is an open-source framework that invites anyone to create and contribute evaluation tasks. [Wiggers, 2023]

Public vs. Private Benchmarking: OpenAI explicitly chose a public, collaborative approach to evaluate its models. They open-sourced the entire Evals tool on GitHub, with the stated goal of allowing "anyone to report shortcomings in [OpenAI's] models to help guide improvements" [OpenAI, 2023]. Rather than keeping evaluation data secret, OpenAI is banking on the breadth and diversity of community-contributed tests to cover many failure modes. As OpenAI wrote, "We are hoping Evals becomes a vehicle to share and crowdsource benchmarks, representing a maximally wide set of failure modes and difficult tasks." In practice, this means developers and researchers worldwide can submit evaluation scripts or datasets – from logic puzzles to factual QA to ethical dilemmas – and these tests become part of a public registry of benchmarks. It's a philosophy of transparency and inclusivity, aligning with open science: the test cases are public, the process is public, and improvements to the benchmark suite come from the community. However, public benchmarks like OpenAI Evals risk becoming optimization targets rather than evaluation tools [Dickson, 2021], leading to models that perform well on known test cases but fail on novel challenges [Dickson, 2023]. Additionally,

community contributions may introduce sampling biases, with certain domains or problem types receiving disproportionate attention while others remain underrepresented [Koch et al., 2021].

Validation of Test: One challenge of open benchmarking is maintaining quality control and coverage. OpenAI has addressed this by providing a framework that supports standard metrics and custom logic [OpenAI, 2025] and seeding it with examples (e.g., they included a sample "logic puzzles" eval where GPT-4 fails). They also made Evals compatible with popular benchmarks [Wiggers, 2023], so known high-quality tests (like HumanEval for code, etc.) can be integrated. The crowdsourced tests are reviewed. OpenAI incentivized high-quality contributions by offering GPT-4 API access to users who submit good evaluations [Wiggers, 2023]. This helped bootstrap a wide range of assessments quickly. While anyone can contribute, the promise of access to the latest model encouraged people to design rigorous, challenging tests rather than trivial ones. Over time, the idea is that the community will naturally curate a set of trusted benchmarks (through reuse and discussion on the open-source repo).

Organizational Trust: OpenAI flips the traditional secrecy model by open-sourcing its evaluations and inviting the public to identify model weaknesses. Anyone can run or create evals, fostering both accountability and collaboration. While this openness risks models being trained to the test, OpenAI counters it with constant updates and a focus on dynamic, adversarial testing - drawing inspiration from projects like Dynabench [Wiggers, 2023]. By integrating Evals into their development process and encouraging community contributions, OpenAI promotes a transparent, participatory approach to benchmarking.