Who Evaluates the Evaluators? Governance Challenges in AI Safety and Alignment Evaluations: A Review and Future Agenda

Zhuo Wang[1], **Xiliang Liu**[1,*], **Ligang Sun**[1]

[1]College of Computer Science, Beijing University of Technology, China
[*]Corresponding author: liuxl@bjut.edu.cn

Abstract

AI safety and alignment evaluations (benchmarks) have become de facto governance tools in the rapidly evolving landscape of artificial intelligence. However, these critical instruments themselves lack systematic governance oversight, raising the central question: “Who evaluates the evaluators?” Through a systematic mapping of 298 papers and in-depth thematic analysis of 50 core studies, this paper identifies fundamental governance challenges in current AI evaluation practices. The bibliometric analysis reveals explosive growth in this field since 2023, with persistent gaps in transparency, accountability, and institutional independence. This position paper synthesizes current knowledge on evaluation governance and proposes a future agenda grounded in three core principles: participatory design with adversarial scrutiny, radical transparency through “Evaluation Cards,” and institutional pluralism to ensure independent oversight. The paper argues for a paradigm shift from “evaluating AI” to “governing evaluations,” offering actionable recommendations for researchers, policymakers, and practitioners to build a trustworthy evaluation ecosystem.

Introduction

The rapid advancement of artificial intelligence systems, particularly large language models (LLMs) and agentic AI, has catalyzed an urgent need for robust safety and alignment evaluations. In response, a proliferation of benchmarks has emerged to assess AI capabilities, risks, and compliance with ethical and regulatory standards (Zeng et al. 2024; Li et al. 2024). These evaluation tools have evolved beyond mere technical assessments to become **de facto governance instruments**, shaping deployment decisions, informing policy frameworks, and establishing industry norms (Daly et al. 2024; Guldimann et al. 2024).

However, a critical paradox persists: while these benchmarks govern AI systems, **the benchmarks themselves remain largely ungoverned**. This governance gap raises fundamental questions about the validity, transparency, and accountability of evaluation practices. Who designs these evaluators? What interests do they serve? How can they accurately measure safety rather than merely general capabilities? (Ren et al. 2024) Most critically: **Who evaluates the evaluators?**

Recent regulatory developments, particularly the EU AI Act, have intensified scrutiny on AI safety evaluation practices (Guldimann et al. 2024; Zhong 2024). Yet empirical evidence reveals substantial misalignments between existing benchmarks and regulatory requirements (Prandi et al. 2025). Studies demonstrate that many safety benchmarks correlate strongly with general model capabilities, risking “safetywashing” where capability improvements are mistaken for genuine safety progress (Ren et al. 2024). Furthermore, the concentration of benchmark development within a small number of industry actors raises concerns about conflicts of interest and insufficient coverage of societal risks (Gruetzemacher et al. 2023).

This position paper addresses these governance challenges through a comprehensive review and future agenda. The paper presents three key contributions: First, it conducts a systematic mapping of 298 papers and thematic analysis of 50 core studies, revealing the landscape of AI safety evaluation research and its governance deficits. Second, it synthesizes evidence of fundamental gaps in transparency, institutional independence, and multi-stakeholder participation. Third, it proposes a future agenda grounded in three principles: (1) **participatory design with adversarial scrutiny**, (2) **radical transparency through “Evaluation Cards,”** and (3) **institutional pluralism and independence**. The analysis calls for a paradigm shift from “evaluating AI” to “governing evaluations,” ensuring that the instruments of AI governance are themselves trustworthy, accountable, and aligned with public interest.

Methodology

This paper employs a three-step research workflow to systematically map the landscape of AI safety and alignment evaluation research and identify core governance challenges, as illustrated in Figure 1.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

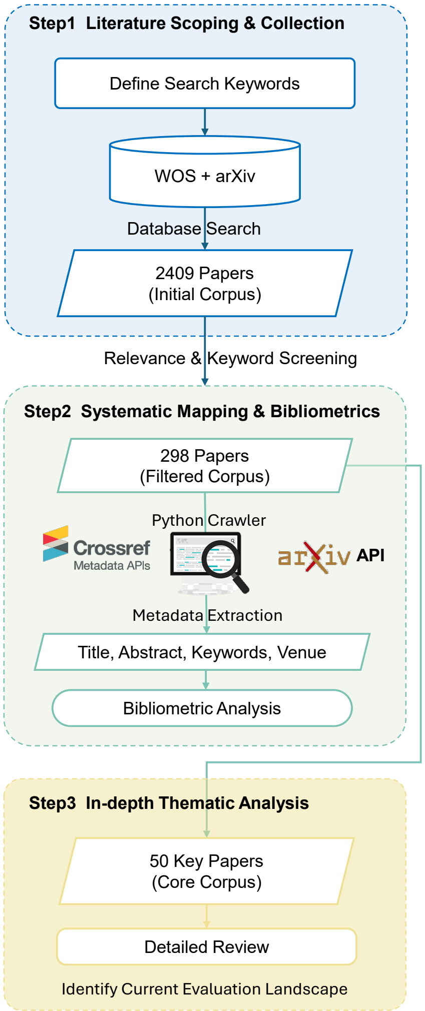


Figure 1: Three-Step research workflow: (1) Literature Scoping and Collection, (2) Systematic Mapping and Bibliometric Analysis, and (3) In-depth Thematic Analysis.

Step 1: Literature Scoping and Collection. The research began by defining a comprehensive set of search keywords related to AI safety, alignment, evaluation, and governance. Broad searches were conducted across two primary sources: the Web of Science (WOS) Core Collection and the arXiv preprint server. This initial search yielded an **Initial Corpus** of 2,409 papers spanning publications from 2021 to 2025.

Step 2: Systematic Mapping and Bibliometric Analysis. Rigorous relevance and keyword screening criteria were applied to filter the initial corpus. Papers were included if they addressed AI safety evaluation, benchmarking methodologies, governance frameworks, or regulatory compliance. Papers were excluded if they were purely technical without governance implications, duplicate publications, or outside the temporal scope. This process refined the dataset to a **Filtered Corpus** of 298 highly relevant papers. To enable macro-level analysis, Python-based web scrapers utilizing the Crossref and arXiv APIs were developed to automatically extract comprehensive metadata, including titles, abstracts, keywords, author information, and publication venues. This structured dataset enabled bibliometric analysis to reveal research trends, thematic clusters, and key publication outlets (detailed in Section 3.1).

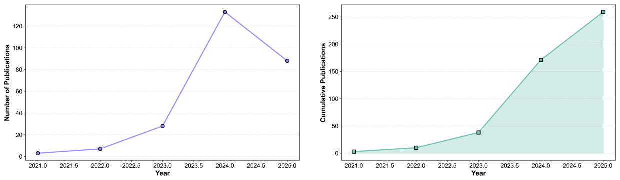
Step 3: In-depth Thematic Analysis. From the 298 papers, 50 papers were purposively selected for detailed qualitative review, forming the **Core Corpus**. Selection criteria prioritized: (1) papers introducing influential benchmarks (e.g., MMLU, TruthfulQA, AIR-Bench), (2) highly cited works shaping governance discourse, (3) papers proposing evaluation frameworks aligned with regulations, and (4) studies examining meta-evaluation and benchmark validity. Systematic thematic coding was conducted to identify governance challenges across five dimensions: regulatory alignment, evaluator reliability, transparency practices, ethical integration, and institutional models. This analysis reveals critical gaps and informed the future agenda (Sections 3.2 and 4).

Current State of Evaluation Research

Bibliometric Analysis

The bibliometric analysis of 298 papers reveals the rapidly evolving landscape of AI safety and alignment evaluation research, characterized by explosive growth, thematic convergence, and emerging governance concerns.

Explosive Growth in Research Output. As shown in Figure 2(a), a remarkable surge in publications has occurred, rising from merely 3 papers in 2021 to a peak of 133 in 2024. The cumulative growth curve in Figure 2(b) demonstrates an inflection point in 2023, with publications accelerating from 38 to 171 papers within a single year—a 350% increase. This explosive growth coincides with the widespread adoption of LLMs and heightened regulatory attention, particularly the development of the EU AI Act. As of mid-2025, the field has already produced 88 papers, suggesting sustained momentum despite platform maturation.



(a) Publication trend in 2021-2025

(b) Cumulative publication count in 2021-2025

Figure 2: Publication trends: (a) Annual publication counts showing peak in 2024, (b) Cumulative growth revealing inflection point in 2023.

Thematic Landscape and Research Priorities. The keyword analysis presented in Figure 3 identifies *safety* (233 occurrences) and *evaluation* (230 occurrences) as dominant research foci, followed by *compliance* (134), *ethics* (128), and *governance* (127). Thematic clustering reveals five major research streams: **Safety & Evaluation** (254 papers, 32.8%), **Governance & Regulation** (194 papers), **Technical Frameworks** (153 papers), **Ethics & Fairness** (152 papers), and **Risk Management** (22 papers). The hierarchical sunburst visualization in Figure 3(c) demonstrates rich sub-structures within each theme, with Safety encompassing evaluation, benchmarking, auditing, and risk assessment dimensions.

Notably, the *ethics* keyword exhibits the most dramatic evolution, growing from 0 mentions in 2021 to 68 in 2024, as shown in Figure 4. This surge reflects increasing recognition that evaluation governance requires ethical frameworks alongside technical rigor. Conversely, *Risk Management* remains underrepresented (22 papers), suggesting an important gap in addressing systemic and catastrophic risks.

Knowledge Flow and Interdisciplinary Connections. Figure 5 presents a Sankey diagram tracing knowledge flows from main themes through sub-themes to specific keywords. AI Safety generates the strongest flows toward Evaluation & Testing (351 papers) and Safety Assurance (309 papers), while AI Governance primarily channels into Policy & Regulation (288 papers). Critically, limited flow exists between AI Governance and Evaluation & Testing domains (only 45 papers address both), highlighting the disconnect between governance frameworks and actual evaluation practices—the core problem motivating this paper.

Research Clustering and Temporal Evolution. Figure 6 further visualizes the structural and temporal dimensions of this research landscape. The keyword clustering map in Figure 6(a) reveals ten major research clusters, with *#0 evaluation*, *#3 governance*, *#4 benchmark*, and *#8 ethics* forming a tightly interconnected core. This empirically confirms this paper's central premise: evaluation is inherently a sociotechnical problem inseparable from governance and ethics. The timeline visualization in Figure 6(b) traces the evolution of research foci. Early attention (2021-2022) concentrated on macro-concepts such as *#7 compliance* and *#2 safety*. However, since 2023, the research frontier has rapidly shifted toward *#0 evaluation*, *#4 benchmark*, and *#1 genai* (generative AI). This pattern demonstrates that evaluation has become the epicenter of contemporary AI governance debates, underscoring the urgency of this research.

The bibliometric analysis reveals rapid yet fragmented growth, with limited interdisciplinary flow between governance and evaluation domains (Figure 5: only 45 papers bridge both). This fragmentation creates systemic governance deficits examined through thematic analysis of 50 core papers below.

Literature Review

The thematic analysis of 50 core papers reveals five critical governance challenges in AI safety evaluation practices: regulatory misalignment, evaluator validity deficits, transparency gaps, ethical fragmentation, and institutional concentration.

Regulatory Misalignment and Coverage Gaps. Multiple studies document substantial gaps between existing benchmarks and regulatory requirements. Guldimann et al. developed COMPL-AI to translate the EU AI Act into technical benchmarks, revealing widespread non-compliance in model documentation and training data transparency (Guldimann et al. 2024). Prandi et al.'s quantitative analysis found that current benchmarks neglect critical functional capabilities related to loss-of-control scenarios, particularly evading human oversight and autonomous AI development (Prandi et al. 2025). While frameworks like AIR-Bench 2024 attempt to bridge regulatory categories and safety benchmarks (Zeng et al. 2024), significant coverage gaps persist. This regulatory-technical disconnect aligns with the limited knowledge flow observed in Figure 5, where only 45 papers bridge AI Governance and Evaluation & Testing domains. Such misalignment creates compliance blind spots where AI systems may meet benchmark criteria while violating regulatory intent.

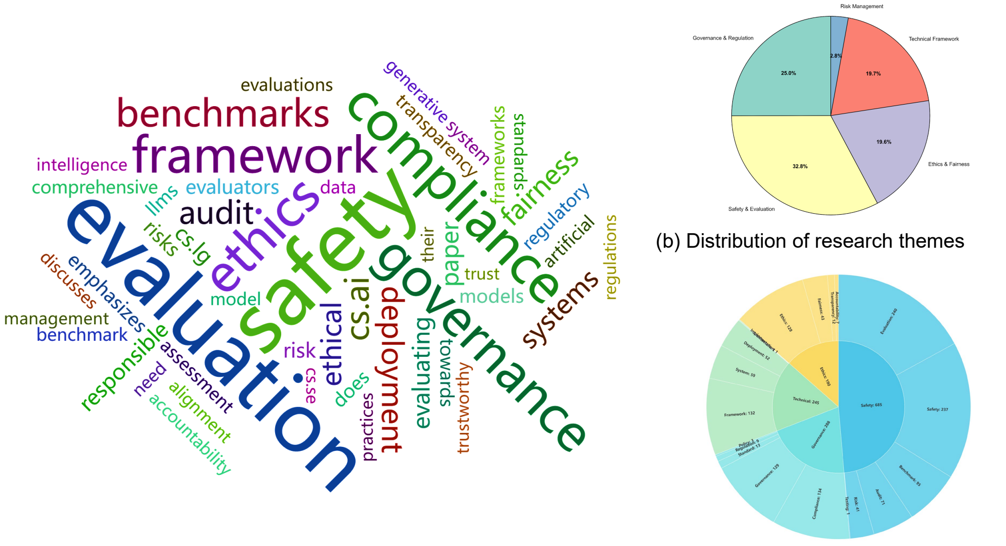
Table 1: Research Growth and Distribution Metrics (2021-2025)*

Metric	Value	Percentage
Total Papers (Filtered)	298	100%
Core Analysis Papers	50	16.8%
Thematic Distribution[1]		
Safety & Evaluation	254	85.2%
Governance & Regulation	194	65.1%
Technical Frameworks	153	51.3%
Ethics & Fairness	152	51.0%
Risk Management	22	7.4%
Top Publication Venues		
arXiv	124	41.6%
AAAI/NeurIPS/ICML/ICLR	87	29.2%
Journals	54	18.1%
Others	33	11.1%
Annual Growth		
2021	3	–
2022	18	500%
2023	38	111%
2024	133	250%
2025	176	32%

Evaluator Validity and the Safetywashing Problem. A meta-analysis by Ren et al. exposed a critical flaw: many ”safety” benchmarks correlate strongly (r ¿ 0.9) with general model capabilities, enabling ”safetywashing” where capability improvements masquerade as safety progress (Ren et al. 2024). Tan et al.'s evaluation of LLM-as-a-Judge systems reveals low reliability on complex tasks, with agreement rates below 60% (Tan et al. 2024). Perlitz et al. introduced Benchmark Agreement Testing (BAT) to assess evaluator consistency, finding substantial inter-evaluator disagreement even on standardized tasks (Perlitz et al. 2024). These validity deficits undermine the trustworthiness of evaluation outcomes and compromise governance effectiveness. Amram et al.'s simulation-based meta-evaluation framework offers promising methods to validate evaluator quality (Amram et al. 2025), while Saxon et al. call for model metrology to enable dynamic capability measurement (Saxon et al. 2024). Binette and Reiter's estimands framework provides statistical rigor for evaluation validity (Binette and Reiter 2024), but standardized meta-evaluation practices remain absent across the field.

Transparency Deficits and Documentation Gaps. De-

[1] * Papers may be classified into multiple themes

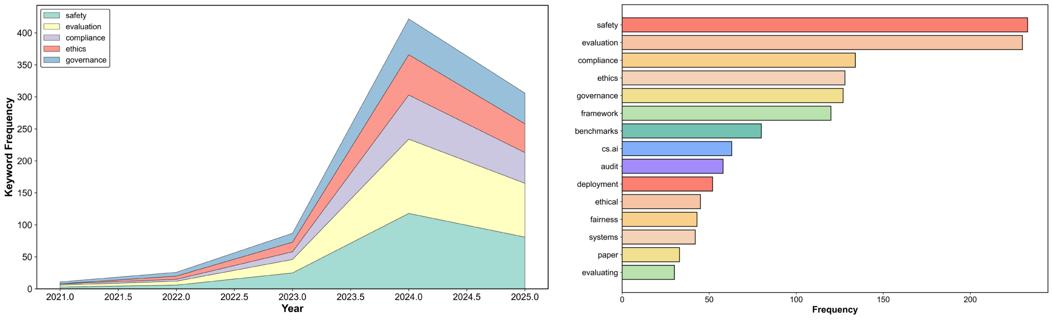


(b) Distribution of research themes

(a) Word Cloud of literature within the scope of the research

(c) Sunburst chart of research themes

Figure 3: Research themes and keywords: (a) Word cloud of top 15 keywords, (b) Distribution across five major research themes, (c) Hierarchical theme structure showing sub-categories.



(a) Evolution of top 5 keywords in 2021-2025

(b) Top 15 most frequent keywords in AI safety and governance literature

Figure 4: Keyword evolution over time showing the rise of ethics discourse and sustained dominance of safety and evaluation themes.

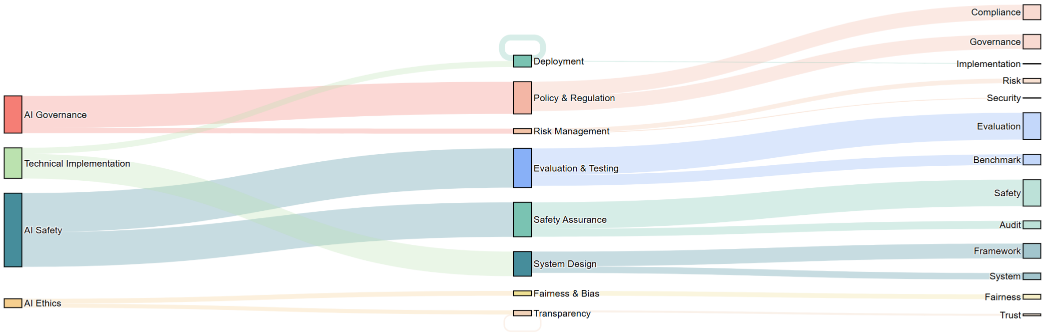
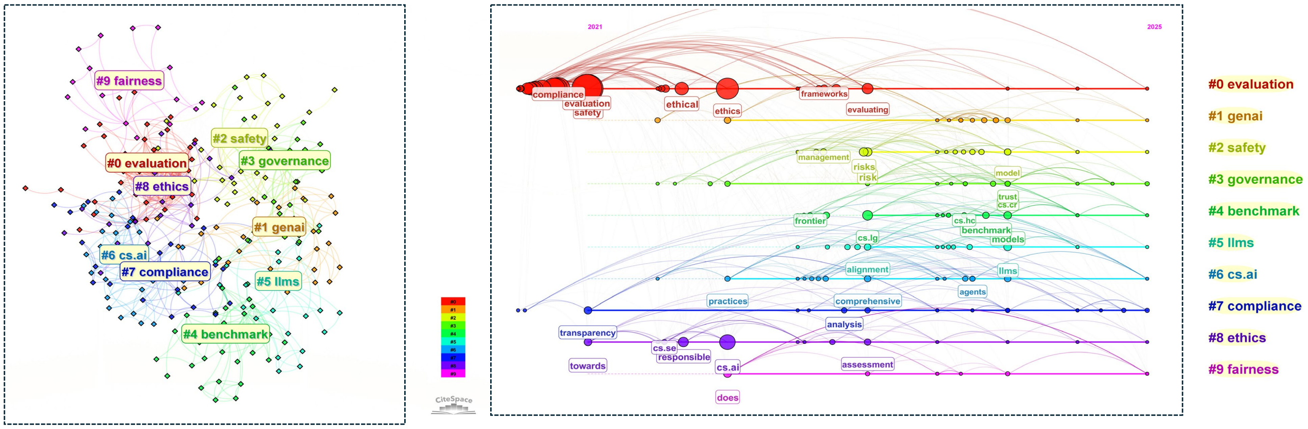


Figure 5: Three-layer Sankey diagram illustrating knowledge flows from main themes to sub-themes to keywords, revealing fragmentation between governance and evaluation research.



(a) The Keyword Co-occurrence Cluster Network of AI Safety and Alignment Evaluation Research (Top 10 Clusters)

(b) The Keyword Evolution Timezone View of AI Safety and Alignment Evaluation (2021-2025)

Figure 6: Knowledge mapping and temporal evolution: (a) Keyword clustering map showing 10 major research clusters with evaluation, governance, benchmark, and ethics forming the interconnected core; (b) Timeline visualization revealing the shift from early compliance/safety focus (2021-2022) to contemporary emphasis on evaluation and benchmarking (2023-2025).

spite widespread calls for transparency, implementation remains fragmented. Sokol et al.'s BenchmarkCards framework proposes structured documentation of ethical and legal considerations, yet analysis reveals most benchmarks lack comprehensive demographic data, privacy disclosures, and bias assessments (Sokol et al. 2024). Reuel et al.'s systematic review found that fewer than 30% of benchmarks meet basic reproducibility standards, with many lacking proper version control and statistical validation (Reuel et al. 2024). Golpayegani et al.'s AI Cards framework attempts machine-readable risk documentation (Golpayegani et al. 2024), while Sharma's blockchain-based compliance system offers immutable audit trails (Sharma 2025). Barnett and Thiergart emphasize explicit assumption disclosure for regulatory effectiveness (Barnett and Thiergart 2024), and Pintz et al. develop platform architectures for automated trustworthiness assessments (Pintz et al. 2024). However, these innovations remain nascent, and the absence of mandatory transparency requirements allows proprietary benchmarks to operate as ”black boxes.”

Ethical Fragmentation and Evolving Risks. While ethical considerations feature prominently in governance discourse, integration remains inconsistent. Sam and Vavekanand's comparative analysis of ethical benchmarks reveals substantial variation in evaluation criteria and moral frameworks (Sam and Vavekanand 2024). Abishek et al.'s data governance framework addresses bias and fairness across the AI lifecycle (Abishek et al. 2025), while Peckham develops comprehensive harm assessment frameworks (Peckham 2024). Yet many benchmarks focus narrowly on technical metrics while neglecting broader societal impacts. The rapid evolution of AI capabilities—particularly agentic and multi-modal systems—outpaces ethical framework development. Weidinger et al. argue for holistic safety approaches integrating diverse risk domains (Weidinger et al. 2024), while Kasirzadeh analyzes measurement challenges in catastrophic risk governance (Kasirzadeh 2024). Dey and Bhaumik examine governance pressures in AI innovation (Dey and Bhaumik 2024), but achieving consensus on ethical priorities across stakeholders remains elusive. This fragmentation enables selective compliance where organizations prioritize measurable metrics over substantive ethical commitments.

Institutional Concentration and Conflicts of Interest. Benchmark development exhibits troubling concentration, consistent with the venue distribution patterns observed in Table 1 where 41.6% of publications originate from arXiv and 29.2% from major industry-affiliated conferences. Gruetzemacher et al. propose international consortia to diversify evaluation capacity (Gruetzemacher et al. 2023), recognizing that industry-dominated benchmarks may reflect commercial interests over public safety. Stein et al.'s analysis of public versus private auditing roles highlights resource constraints limiting government oversight capacity (Stein et al. 2024). Fort's examination of AI Safety Institutes identifies their potential as independent evaluators but notes nascent institutional capacity (Fort 2024). The concentration of benchmark development within a handful of technology companies creates conflicts of interest where evaluators assess their own products. Manheim et al.'s proposal for an AI Audit Standards Board aims to establish independent oversight (Manheim et al. 2024), yet implementing such governance structures faces political and financial obstacles.

Emerging Solutions: Multi-Agent and Hybrid Governance. Recent research explores innovative governance models. Zhuge et al.'s Agent-as-a-Judge framework demonstrates how agentic evaluation can provide nuanced, intermediate feedback surpassing human baselines on certain tasks (Zhuge et al. 2024). Joshi's Joint Evaluation (Jo.E) framework integrates humans, LLMs, and AI agents, reducing expert time requirements by 70% while improving vulnerability detection (Joshi 2025). Pervez et al.'s Governance-as-a-Service enables modular policy enforcement with dynamic trust scoring (Pervez et al. 2025). Zhu et al. develop automated agent-flow systems for safety benchmarking (Zhu et al. 2025a) and rigorous agentic benchmark checklists (Zhu et al. 2025b). Vijayvargiya et al. propose comprehensive real-world agent safety evaluation frameworks (Vijayvargiya et al. 2025), while Liu et al. introduce jailbreak judge benchmarks with multi-agent explanation (Liu et al. 2024). Dorn et al. develop structured tests for emerging failure modes (Dorn et al. 2024). While these multi-agent approaches show promise for scalability and robustness, they introduce new challenges in explainability, bias propagation, and coordination complexity. Yu's comprehensive review cautions against over-reliance on automated judges without rigorous meta-evaluation (Yu 2025). Shukla proposes balanced evaluation frameworks for agentic systems integrating capability, safety, and ethics (Shukla 2025).

Future Agenda: Building a Trustworthy Evaluation Ecosystem

Addressing the governance deficits identified above requires a paradigm shift from "evaluating AI" to "governing evaluations." This paper proposes a future agenda grounded in three interconnected principles that collectively establish a trustworthy evaluation ecosystem.

Principle 1: Participatory Design and Adversarial Scrutiny

Current benchmark development remains concentrated within technical communities, lacking meaningful participation from affected stakeholders. This paper calls for **multi-stakeholder design processes** that include civil society organizations, domain experts from diverse cultural contexts, ethicists, policy makers, and representatives of communities disproportionately affected by AI systems. Participatory design ensures benchmarks reflect diverse values, contextual knowledge, and societal priorities beyond narrow technical metrics (Weidinger et al. 2024).

Complementing participatory design, this paper advocates establishing **"Evaluation Red Teams"**—dedicated groups tasked with adversarially probing benchmarks to expose blind spots, gaming vulnerabilities, and validity threats. Drawing inspiration from cybersecurity practices, these red teams would systematically challenge evaluation methodologies, test edge cases, and identify scenarios where benchmarks fail to capture genuine risks. Red team findings should be publicly disclosed to drive continuous improvement. This adversarial scrutiny prevents complacency and counters the tendency for benchmarks to become optimized proxies disconnected from underlying safety properties (Ren et al. 2024).

Institutional implementation requires funding mechanisms supporting independent red teams with protected whistleblower status, ensuring they can critique powerful actors without retaliation. Regulatory frameworks should mandate red team assessments for benchmarks used in high-risk applications, with results informing certification decisions.

Principle 2: Radical Transparency through "Evaluation Cards"

Echoing the model card paradigm (Mitchell et al. 2019), this paper proposes mandatory **"Evaluation Cards"** for all benchmarks deployed in governance contexts. Evaluation Cards would comprehensively document: (1) *funding sources and conflicts of interest*, disclosing any commercial relationships that might bias design choices; (2) *design rationale and theoretical foundations*, explicating what safety properties the benchmark purports to measure and why; (3) *dataset provenance, composition, and known limitations*, including demographic representation, consent protocols, and identified failure modes; (4) *validation evidence*, reporting meta-evaluation results, inter-evaluator agreement, and correlation analyses with general capabilities; and (5) *appropriate use cases and exclusions*, specifying contexts where the benchmark should not be applied (Sokol et al. 2024; Golpayegani et al. 2024).

Evaluation Cards must be machine-readable to enable automated compliance checking and comparative analysis. Regulatory bodies should establish repositories of certified Evaluation Cards, refusing to recognize benchmarks lacking proper documentation. Transparency extends to evaluation processes themselves: audit trails documenting who conducted evaluations, under what conditions, and with what results should be maintained and, where appropriate, publicly accessible (Sharma 2025).

Critically, radical transparency includes admitting uncertainty and limitation. Evaluation Cards should prominently feature sections on *known failure modes* and *open questions*, resisting the false precision that undermines stakeholder trust. This epistemic humility positions evaluation as an ongoing process rather than definitive judgment.

Principle 3: Institutional Pluralism and Independence

Breaking the concentration of evaluation power requires deliberate cultivation of **institutional pluralism**. Governments and philanthropic organizations must fund **independent third-party evaluation institutes** with sufficient resources to conduct rigorous assessments free from commercial pressures (Fort 2024; Gruetzemacher et al. 2023). These institutes should operate as public utilities, obligated to serve public interest rather than private clients.

International coordination mechanisms—such as the proposed AI Audit Standards Board (Manheim et al. 2024)—can harmonize evaluation methodologies while respecting jurisdictional differences. Multi-track governance models enabling collaboration between public, private, and civil society evaluators distribute power and leverage diverse expertise (Stein et al. 2024). However, clear delineation of roles prevents regulatory capture: public bodies should retain authority over high-risk assessments, with private auditors operating under strict licensing and oversight.

Institutional independence extends to benchmark development. Public funding for benchmark creation should prioritize projects led by academic institutions, non-profits, and international consortia rather than defaulting to industry-developed tools (Gruetzemacher et al. 2023). Competitive grant programs can spur innovation in evaluation methodologies while ensuring results remain open-source and freely accessible.

Finally, governance structures must be **adaptive**, incorporating sunset clauses for benchmarks, regular meta-evaluations, and mechanisms for rapid response to emergent risks. The AI Governance-as-a-Service model (Pervez et al. 2025) demonstrates how modular policy enforcement can enable dynamic adaptation without sacrificing accountability. Combining institutional stability with methodological agility positions evaluation systems to remain effective amid rapid technological change.

Toward Trustworthy AI Through Trustworthy Evaluation. These three principles—participatory design with adversarial scrutiny, radical transparency through Evaluation Cards, and institutional pluralism—are mutually reinforcing. Participation without transparency enables capture; transparency without institutional independence cannot be enforced; independence without participatory legitimacy fosters technocratic isolation. Collectively, they establish an evaluation ecosystem worthy of the governance responsibilities it has assumed. The urgent task for researchers, policymakers, and practitioners is translating these principles into concrete practices, regulations, and institutions. Only by governing the evaluators can society govern AI.

References

Abishek, K.; et al. 2025. Data and AI Governance Framework for Bias and Fairness. *IEEE Transactions on AI*.

Amram, M.; et al. 2025. LaaJMeter: Simulation-Based Meta-Evaluation for LLM-as-a-Judge. *arXiv preprint*.

Barnett, S.; and Thiergart, Z. 2024. Framework for Explicit Assumptions in AI Evaluations. *Proceedings of NeurIPS*.

Binette, O.; and Reiter, J. P. 2024. Estimands Framework to Improve Validity of AI/ML Evaluations. *Journal of Machine Learning Research*.

Daly, A.; et al. 2024. Governance Frameworks for AI Safety Benchmarks. *AI and Society*.

Dey, S.; and Bhaumik, R. 2024. Governance Framework Balancing Pressures in AI Innovation. *Technology Innovation Management Review*.

Dorn, J.; et al. 2024. Structured Tests for LLM Safeguards and Emerging Failure Modes. *arXiv preprint*.

Fort, H. 2024. AI Safety Institutes and International Standards. *Journal of AI Governance*.

Golpayegani, D.; et al. 2024. AI Cards: Machine-Readable AI Risk Documentation. *Semantic Web Journal*.

Gruetzemacher, R.; et al. 2023. Proposal for an International Consortium for AI Risk Evaluations. *AI and Ethics*.

Guldimann, T.; et al. 2024. COMPL-AI: Evaluating LLMs for Compliance with the EU AI Act. *arXiv preprint*.

Joshi, M. 2025. Joint Evaluation: Integrating Humans, LLMs, and Agents. *arXiv preprint*.

Kasirzadeh, A. 2024. Analysis of Measurement Challenges in Catastrophic AI Risk Governance. *AI Ethics*.

Li, X.; et al. 2024. Balanced Safety Benchmarks for Large Language Models. *Proceedings of NeurIPS*.

Liu, H.; et al. 2024. Comprehensive Jailbreak Judge Benchmark with Multi-Agent Explanation. *Proceedings of AAAI*.

Manheim, D.; et al. 2024. Proposal for an AI Audit Standards Board. *AI Policy Journal*.

Mitchell, M.; et al. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229.

Peckham, J. 2024. AI Harms and Governance Framework for Trustworthy AI. *AI and Society*.

Perlitz, J.; et al. 2024. Benchmark Agreement Testing for Evaluator Consistency. *Proceedings of ICML*.

Pervez, H.; et al. 2025. Governance-as-a-Service: Modular Runtime Policy Enforcement. *arXiv preprint*.

Pintz, M.; et al. 2024. Platform Architecture for Automated AI Trustworthiness Assessments. *Proceedings of CHI*.

Prandi, C.; et al. 2025. Bench-2-CoP: Quantifying Benchmark-Regulation Coverage Gaps. *arXiv preprint*.

Ren, A. Z.; et al. 2024. Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress? *arXiv preprint*.

Reuel, A.; et al. 2024. Best Practices for Foundation Model Benchmarking. *arXiv preprint*.

Sam, D. B.; and Vavekanand, V. 2024. Comparative Analysis of Ethical Benchmarking in Large Language Models. *International Journal of AI Ethics*.

Saxon, M.; et al. 2024. Call for Model Metrology to Improve Benchmarking Science. *Proceedings of ICLR*.

Sharma, A. 2025. AI Ethics Compliance System using Blockchain and Smart Contracts. *arXiv preprint*.

Shukla, P. 2025. Balanced Evaluation Framework for Agentic AI Systems. *arXiv preprint*.

Sokol, K.; et al. 2024. BenchmarkCards: Framework for Structured LLM Risk Reporting. *arXiv preprint*.

Stein, D.; et al. 2024. Public versus Private Auditing in AI Governance. *Regulation & Governance*.

Tan, Y.; et al. 2024. Large Language Models as Judges: Challenges and Opportunities. *arXiv preprint*.

Vijayvargiya, A.; et al. 2025. Comprehensive Framework for Real-World AI Agent Safety Evaluation. *arXiv preprint*.

Weidinger, L.; et al. 2024. Holistic Safety and Responsibility Evaluation Approaches. *Nature Machine Intelligence*.

Yu, H. 2025. Agent-as-a-Judge: A Comprehensive Review. *arXiv preprint*.

Zeng, Y.; et al. 2024. AIR-Bench 2024: A Safety Benchmark Aligning with Regulations. *arXiv preprint*.

Zhong, R. 2024. Interdisciplinary Governance for EU AI Act Implementation. *European Journal of Risk Regulation*.

Zhu, Y.; et al. 2025a. Automated Agent-Flow System for LLM Safety Benchmarking. *arXiv preprint*.

Zhu, Y.; et al. 2025b. Checklist for Rigorous Agentic Benchmark Construction. *arXiv preprint*.

Zhuge, M.; et al. 2024. Agent-as-a-Judge: Evaluating Agentic AI Systems. *arXiv preprint*.