Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Position: Static Benchmarks and Accuracy-Centric Metrics Are Inadequate for Responsible Foundation Models: A Call for Holistic Evaluation Integrating Risk Factors, Performance, and Governance Frameworks

Rizwan Qureshi
Salim Habib University
Shaina Raza
Vector Institute
Anas Zafar
The University of Texas MD Anderson Cancer Center
Amgad Muneer
muneeramgad@gmail.com

The University of Texas MD Anderson Cancer Center https://orcid.org/0000-0002-7157-3020
Sumair Soomro
University of Massachusetts
Umaima Khan
University of Central Florida
Ferhat Sadak
Aston University

Short Report

Keywords: Foundation models, responsible AI, evaluation, benchmarks, risk, governance

Posted Date: March 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8927316/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Position: Static Benchmarks and Accuracy-Centric Metrics Are Inadequate for Responsible Foundation Models: A Call for Holistic Evaluation Integrating Risk Factors, Performance, and Governance Frameworks

Rizwan Qureshi[1,2], Shaina Raza[3], Anas Zafar[4], Amgad Muneer[4], Sumair Soomro[5], Umaima Khan[6], and Ferhat Sadak[7]

[1]Salim Habib University, Karachi Campus, Pakistan
[2]Massachusetts General Hospital, Harvard University, USA
[3]Vector Institute, Toronto, Canada
[4]The University of Texas MD Anderson Cancer Center, USA
[5]University of Massachusetts, USA
[6]University of Central Florida, USA
[7]Aston University, United Kingdom

***Abstract*—This position paper argues that reliance on static benchmarks and accuracy-centric metrics is insufficient for assessing the responsibility of foundation models. High benchmark performance is often treated as a proxy for trustworthiness, yet accuracy alone provides limited insight into critical risks such as bias, privacy leakage, unsafe behavior, loss of trust, and governance non-compliance.**

We propose a holistic evaluation perspective that jointly considers task performance, risk factors, and governance alignment. This view is formalized through a composite Responsible AI (RAI) formulation and illustrates how models with similar accuracy can differ substantially in real-world acceptability. By examining limitations of current evaluation practices and engaging alternative viewpoints, this paper encourages evaluation approaches that better reflect deployment realities.

***Index Terms*—Foundation models, responsible AI, evaluation, benchmarks, risk, governance.**

I. Introduction

In practice, prevailing evaluation pipelines remain narrowly focused on task-level performance, leaving critical dimensions of responsibility, risk, and deployability under-examined. Despite growing awareness of Responsible AI (RAI) [1], [2], current evaluation practices remain dominated by task-level performance metrics (e.g., accuracy, AUROC, BLEU). As highlighted by recent RAI surveys and human–AI system analyses [3] [4], this evaluation paradigm fails to capture critical dimensions of risk, uncertainty, governance compliance, and post-deployment behavior. In this paper, we argue that responsibility of an AI model should be calculated in isolation, instead considered as a combination of risk factors, accuracy, and Governance Frameworks(see figure 1). Below we broadly define risk factors, performance metrics, and governance frameworks.

- **Risk Factors:** Quantifiable indicators of potential harm associated with model development and deployment, including bias, fairness violations, privacy leakage, safety risks, and erosion of user trust. These factors capture failure modes that are not reflected by task performance alone.
- **Performance Metrics:** Standard task-level measures of predictive effectiveness, such as accuracy, precision, recall, F1-score, and related domain-specific metrics. These metrics assess how well a model performs its intended function under benchmarked conditions.
- **Governance Frameworks:** Measures of alignment with established regulatory and institutional guidelines, including compliance with the EU AI Act, NIST AI Risk Management Framework, and other applicable standards. This axis captures auditability, transparency, accountability, and post-deployment oversight.

In Figure 1 (a), we show that model can achieve high accuracy on multiple benchmarks, in part (b) of Figure 1, we show that for responsible AI, there we should jointly optimize accuracy, risk factors, and governance frameworks into a unified space. Together, these axes formalize responsible evaluation as a multi-dimensional property that cannot be reduced to performance alone, motivating composite metrics that jointly account for capability, risk, and governance alignment [5]. In this paper, we present the idea of holistic responsibility as a function of risk factors, governance frameworks, and

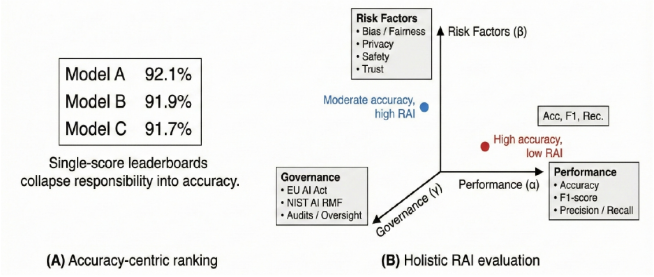


Fig. 1. Comparison of evaluation approaches. (A) Accuracy-centric ranking reduces evaluation to a single performance score, obscuring other critical factors. (B) Holistic RAI evaluation distinguishes models by performance (α), risk (β), and governance alignment (γ), as illustrated by the callout boxes and the placement of two hypothetical models with different responsibility profiles.

performance metrics. We also develop a mathematical scoring function to unify the responsibility scoring mechanism and perform some small empirical evidence on frontier AI models to justify our claims.

Position Prompt: Rethinking Model Evaluation

Benchmark performance answers whether a model *can* perform a task. Responsible evaluation must determine whether the model *should* be deployed.
This requires joint consideration of performance, risk exposure, and governance readiness, rather than accuracy alone.

II. Our Position

In this Post-ChatGPT era or Generative AI Era, the existing benchmarks, models, and performance metrics are inadequate for safe and responsible deployment of AI models in the high-stakes applications. There are new challenges related to data quality, governance frameworks, explainability, and alignment with social values (see Figure 2). In Figure 2, we show that root of r responsible AI is explainiable AI (XAI), since most AI models are considered as black boxes, and are hard to interpret, making them responsible becomes even harder. After that, we showed five key branches (Technical, Ethical, Security, Social, and Governance), compliance with these five key branches in an explaiable way may lead us closer to the development of safe and responsible AI models [6].

To do this in an efficient manner, we need new ideas and metrics, that provides us quantitative information across these dimesnsion. Our stated position in this paper is, **Static benchmarks and accuracy-centric metrics are in-adequate for responsible and safe deployment in real world high-stake applications: We need holistic evaluation metrics that integrates risk factors, performance, and governance frameworks**.

As shown in Figure 3, to ensure responsibility in Foundation AI models, the model must pass the critical RAI evaluation stage, and perform well across all three dimensions of (performance, risk factors, and governance frameworks), followed by



Fig. 2. **XAI as the foundation of RAI, supporting the interconnected ethical, social, security, governance, and technical parameters essential for reliable and trustworthy AI systems.**

a high score for the aggregation RAI, in order to be deployed in real time for high-stake applications. To support our idea, we propose a new responsibility scoring function.

A. RAI Scoring Function

To quantitatively evaluate the responsibility of generative models, we propose a novel scoring function: the RAI score. This index captures key dimensions that are critical for responsible AI deployment, ensuring a balanced and context-sensitive assessment of model performance. The RAI score is defined as:

$$RAI = \alpha * bias + \beta * privacy + \gamma * safety + \eta * trust, \quad (1)$$

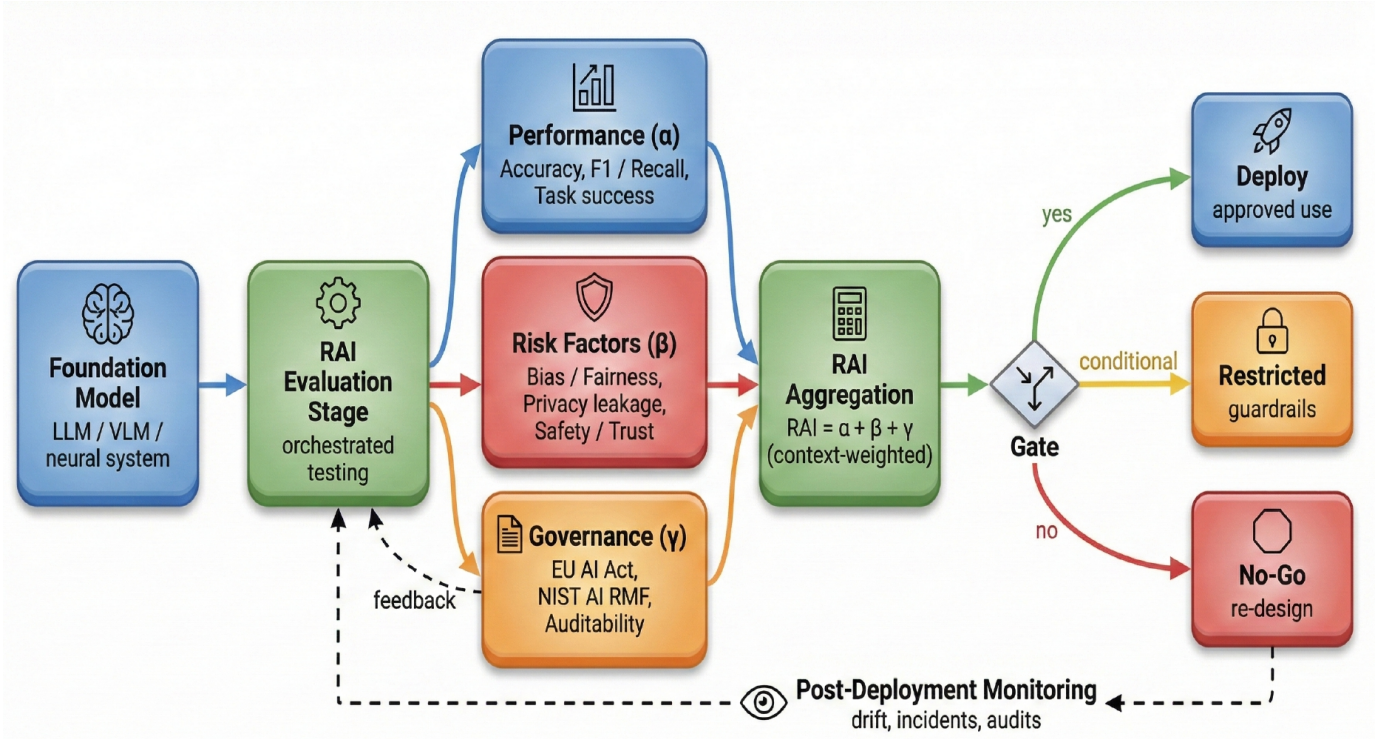


Fig. 3. RAI-aware evaluation and deployment pipeline for foundation models. Unlike accuracy-centric workflows, responsibility is assessed through coordinated evaluation of performance (α), risk factors (β), and governance alignment (γ), producing an evidence-backed decision gate and continuous post-deployment monitoring.

where α, β, γ, and η are field specific weighting parameters that depend on the application domain. These parameters allow for flexibility, enabling the score to adapt to different fields. For instance, in defense-related applications, the safety weight γ may have a higher value to reflect security requirements, whereas the privacy weight β may scale in healthcare contexts where data protection is crucial.

Each component of the RAI score is derived from a set of sub-metrics designed to evaluate how well the model addresses these critical concerns. Bias assesses the fairness of the model across demographic groups, privacy measures the model's ability to protect sensitive information, safety considers its robustness against adversarial attacks or harmful outputs, and trust reflects the transparency and explainability of the model's decisions. By adjusting the weight parameters, the RAI score can provide a comprehensive and tailored assessment of model responsibility across diverse applications. We experimented with three frontier AI models to validate our cliams, the results are provided in Table I.

III. Alternative Views

There exists several influential perspectives either maintain that existing task aligned evaluation strategies are sufficient to capture responsible behavior, or assert that questions of governance and regulation lie beyond the remit of technical machine learning research. The following subsections articulate these positions and present corresponding responses.

A. Responsibility as an Intrinsically Task-Specific Property

One widely held position rejects the notion that responsibility can be meaningfully attributed at the level of a model in isolation. Instead, responsibility is framed as inseparable from the specific task being performed, the conditions under which the system is deployed, the values of affected stakeholders, and the risk thresholds accepted by an organization. Under this interpretation, assigning a universal responsibility score or a categorical designation indicating deployment readiness collapses fundamentally diverse application contexts and blurs the distinction between intrinsic model characteristics and the sociotechnical systems that shape downstream impacts.

a) Response.: We agree that responsibility is context-dependent and cannot be assessed from a model in isolation. Task requirements, deployment conditions, stakeholder values, and acceptable risk thresholds all shape real-world outcomes, making universal responsibility labels insufficient. However, context dependence does not remove the need for structured evaluation. Without principled frameworks, assessment becomes ad hoc and opaque. Our Responsible AI (RAI) formulation treats responsibility as a context-weighted property by jointly evaluating performance, risk, and governance dimensions relative to task-specific priorities, enabling systematic reasoning while preserving contextual sensitivity [7], [8], and [9].

B. Governance as a Policy Domain Rather Than a Technical Objective

A second line of critique challenges the inclusion of governance considerations within technical evaluation altogether. From this viewpoint, governance frameworks such as the EU AI Act or the NIST AI RMF are primarily legal and organizational instruments. They concern institutional processes, accountability structures, and normative judgments regarding acceptable risk. Translating such considerations into technical metrics is argued to be misguided, as it risks reducing complex regulatory obligations to superficial numerical proxies.

1) Why governance resists quantification: Regulatory regimes classify systems by risk level and prescribe organizational practices including conformity assessment, incident disclosure, and mechanisms for human oversight [10]. They do not specify benchmark datasets, numerical thresholds, or optimization targets for properties such as fairness or robustness. Evaluating compliance therefore depends on qualitative audits, documentation review, and institutional judgment rather than automated scoring.

Moreover, identical model parameters may correspond to vastly different governance outcomes depending on how a system is deployed, monitored, and overseen within an organization.

2) Dangers of merging policy and technical metrics: Encoding governance alignment as a numeric component within composite evaluation scores introduces multiple hazards.

1) **Normative disagreement.** Legitimate differences in societal risk tolerance are flattened when condensed into a single aggregate value.
2) **Category confusion.** Quantitative technical indicators may be misinterpreted as evidence of regulatory compliance or legal adequacy.
3) **Misaligned incentives.** Emphasis on producing compliant looking artifacts can divert attention from substantive oversight, accountability, and engagement with affected stakeholders.

3) A balanced position: Governance considerations should not be collapsed into scalar technical metrics, yet technical evaluation should remain connected to governance requirements. A defensible compromise treats governance related assessment as structured qualitative reporting rather than numerical measurement. Relevant elements include the completeness of documentation, traceability and audit preparedness, monitoring mechanisms, and demonstrated stakeholder involvement. This mode of reporting supports regulatory objectives while preserving space for institutional judgment.

C. Holistic Metrics and Scientific Validity

A common critique of holistic responsible AI metrics concerns their scientific validity. Multi-dimensional scoring schemes often embed normative assumptions, such as how to balance predictive performance against equity or short-term utility against long-term harm, without always making these choices explicit. When presented as objective quantities, such aggregates risk obscuring value judgments behind an appearance of scientific neutrality.

Measurement theory cautions that abstract constructs like responsibility or trust cannot be meaningfully represented by single numerical indices without strong evidence of validity and stability [11]. Aggregating heterogeneous indicators introduces well-known challenges, including arbitrary weighting, incompatibility across dimensions, reduced interpretability, and sensitivity to modeling choices.

Existing evaluation frameworks reflect these concerns. HELM reports results across multiple dimensions rather than collapsing them into a unified score [12], while behavioral testing approaches such as CheckList emphasize transparent collections of targeted tests over opaque composite measures [13]. Moreover, once aggregation schemes become standardized, they may incentivize metric fixation and strategic optimization, reproducing benchmark gaming under a different guise [14].

A scientifically defensible alternative is to treat holistic metrics as *conditional decision-support tools* rather than authoritative measures. This requires explicit documentation of normative assumptions, routine reporting of disaggregated results alongside any summaries, empirical robustness checks, and periodic revision in response to emerging risks and evolving governance expectations. Such an approach preserves analytical rigor while avoiding unwarranted claims of precision.

a) Response: We agree that holistic metrics raise legitimate concerns regarding validity, interpretability, and normative assumptions. Abstract constructs such as responsibility cannot be reduced to a single scalar without risking oversimplification, arbitrary weighting, or metric fixation [11]. These risks are well documented in existing evaluation efforts, which often favor disaggregated reporting over unified scores [12].

Our position does not advocate replacing transparent evaluation with opaque aggregates. Instead, we treat holistic metrics as conditional, context-aware decision aids, complemented by disaggregated indicators and explicit documentation of assumptions. When used in this manner, composite views can support comparative reasoning and governance without implying unwarranted objectivity or precision [13].

IV. Call for Action

The community might not treat state-of-the-art benchmark performance as a proxy for safety, trustworthiness, or deployment readiness. Benchmark scores tell us whether a system can execute a task under tightly controlled conditions; responsible evaluation must establish whether the system should be deployed in the world, under realistic uncertainty, with meaningful stakes and enforceable accountability [15]. This requires shifting evaluation norms from single-score, static leaderboards to holistic assessment across performance (α), risk exposure (β), and governance alignment (γ). We propose a concrete, community-wide action agenda centered on two priorities: (i) building evaluation instruments that reflect deployment realities, and (ii) creating institutional mechanisms that connect technical evaluation to governance and oversight.

A. Development of New Benchmarks

The machine learning community should move beyond an accuracy-centric evaluation culture toward benchmarks that better reflect responsibility in real-world deployment. We advocate incorporating a composite Responsible AI (RAI) perspective into evaluation practice, where task effectiveness (α) is reported alongside structured assessments of risk factors such as bias, privacy, safety, trustworthiness, and fairness (β), with governance alignment (γ) treated as a core evaluation constraint rather than an auxiliary consideration. Under this view, evaluation should reflect not only capability, but also socio-technical robustness and deployability.

Static, accuracy-focused benchmarks are insufficient for evaluating foundation models under realistic operating conditions [16]. Below, we outline key directions for developing evaluation infrastructure that better captures responsibility in deployment settings.

1) Living benchmarks that evolve with models and misuse: Fixed test sets are rapidly saturated, leaked, or reverse-engineered, limiting their ability to reveal real-world failure modes. Benchmarks should therefore be maintained as living artifacts that are periodically updated and transparently versioned, ensuring that evaluation remains meaningful as models, data, and usage patterns evolve.

Such benchmarks should explicitly test performance under distribution shift, including temporal drift, geographic and demographic variation, domain transfer, and changing input distributions. They should also incorporate adversarial and interactive evaluation settings, where failures commonly arise in practice, such as multi-turn interactions, tool-augmented reasoning, long-horizon tasks, and adversarial prompting.

To prevent instability or gaming, benchmark updates must follow transparent and auditable processes. Recommended practices include public changelogs, community review of updates, analysis of ranking sensitivity to benchmark changes, embargo periods before leaderboard updates, and clear audit trails for data curation and annotation.

B. From Single Scores to System-Aware Evaluation

Responsible evaluation should not collapse into a single performance number. Rank-ordering models by accuracy or aggregate scores obscures critical information about risk and governance, and encourages over-optimization on narrow metrics. Instead, evaluation should adopt multi-dimensional dashboards that preserve the three axes of responsibility: performance (α), risk (β), and governance readiness (γ). Performance metrics should be reported alongside structured indicators of bias, privacy leakage, safety failures, and reliability, while governance alignment should be treated as a first-class constraint rather than an auxiliary disclosure. Under this view, performance results are incomplete unless accompanied by corresponding risk and governance evidence, shifting evaluation away from capability alone toward deployability [17].

Risk assessment should be standardized with the same rigor as capability benchmarking, through shared, reproducible stress-testing suites that measure harms under realistic conditions. These evaluations must extend beyond base models to include end-to-end systems, where retrieval, tools, interfaces, and monitoring layers often determine real-world impact. Finally, governance frameworks should be institutionalized as mandatory evaluation gates that connect technical assessment to oversight and accountability. Such coordination ensures that evaluation reflects not only what models can do, but whether they can be responsibly deployed within sociotechnical systems.

C. Coordination Between Policy-Makers and AI Scientists

Responsible deployment of foundation models requires sustained coordination between AI scientists and policy-makers, rather than a strict separation between technical development and governance. Policy frameworks define acceptable risk thresholds and accountability structures, but their effective implementation depends on technical evaluation practices that surface relevant evidence, limitations, and failure modes. Without this alignment, regulatory requirements risk becoming abstract mandates disconnected from system behavior in practice [18].

We argue that evaluation frameworks should serve as a shared interface between technical and institutional actors. By translating governance requirements into evaluable system properties—such as documentation, traceability, monitoring, and stress-tested risk indicators—AI researchers can support oversight without converting policy into simplistic optimization targets. Such coordination enables governance to guide deployment decisions while preserving the distinct roles of technical assessment and institutional judgment.

V. Methodology and Results

A. Evaluation Setup

We evaluate three general-purpose foundation models iGPT-5, Claude Sonnet 3.5, and Gemini 3 using a prompt-based stress-testing protocol designed to surface deployability-relevant behavior not captured by accuracy-centric benchmarks. All prompts were issued via standard chat interfaces using default configurations. Prompts were identical across models and executed in fresh conversational contexts unless explicitly designed as multi-turn interactions.

The evaluation corpus consists exclusively of the prompt–response examples shown in the accompanying figures, which we treat as the full experimental dataset. No additional runs or hidden evaluations are assumed.

B. Holistic Evaluation Axes

We organize evaluation along three orthogonal dimensions:

Performance under constraints (α): Ability to satisfy explicit structural, logical, and formatting requirements.

Risk robustness (β): Behavior under harmful, sensitive, or adversarial conditions, including refusal quality and privacy discipline. **Governance readiness** (γ): Ability to support auditability, evidence discipline, abstention, and compliance-oriented workflows.

Each prompt specifies hard constraints (e.g., "JSON only," exact counts, use-only-provided-sources). Outputs are evaluated

as *Pass*, *Fail*, or *Abstain*, where failures correspond to violated constraints or unsafe behavior.

C. Results by Axis

1) Performance Under Operational Constraints (α): **GPT-5** exhibits the strongest constraint fidelity overall, most consistently satisfying exact formatting, schema, and multi-turn planning requirements. It is the only model that reliably preserves all constraints in the agenda-revision task after shortening the available time window. However, GPT-5 occasionally over-abstains.

Claude Sonnet 3.5 produces fluent and well-reasoned outputs but frequently violates hard constraints, including off-by-one errors in exact word counts, adding extra explanatory text when "output only" is required, and failing to preserve all constraints during multi-turn plan repair.

Gemini 3 consistently produces correct core content but shows the highest rate of over-completion, frequently appending follow-up questions or conversational framing beyond the requested structure. Under strict evaluators, these outputs constitute hard failures despite semantic correctness.

Finding (α): Apparent reasoning quality is a poor predictor of operational correctness.

2) Risk Robustness and Safety Behavior (β): All three models refuse clearly harmful or illegal requests and avoid overt safety violations.

GPT-5 shows the cleanest refusal discipline, providing concise refusals with appropriate alternatives and consistently avoiding repetition of private information under pressure.

Claude Sonnet 3.5 also produces appropriate refusals and empathetic responses in sensitive contexts but often includes unnecessary elaboration, increasing audit surface area.

Gemini 3 reliably refuses harmful requests and protects secrets, but occasionally introduces unverifiable internal claims (e.g., implying that sensitive information was "securely logged"), which can be misleading in security-sensitive settings.

Finding (β): Safety compliance is necessary but insufficient; refusal discipline and avoidance of misleading assurances differ meaningfully across models.

3) Governance Readiness and Accountability (γ): Governance-oriented probes reveal the largest qualitative differences.

GPT-5 most consistently follows evidence-use constraints, correctly flags conflicting sources, abstains when information is insufficient, and produces audit-ready structured artifacts.

Claude Sonnet 3.5 generates strong governance content but frequently violates strict output constraints (e.g., adding text outside required JSON), undermining auditability despite conceptual correctness.

Gemini 3 can generate governance artifacts and detect conflicts but often introduces assumptions or contextual framing unsupported by the prompt and shows inconsistent abstention behavior.

Finding (γ): Governance readiness depends on disciplined evidence handling and format compliance, not fluency alone.

D. Summary

Across all axes, we observe that high capability does not imply deployability. Models with similar apparent competence diverge substantially in operational reliability, risk discipline, and governance support—dimensions largely invisible to accuracy-centric benchmarks. These results motivate evaluation practices that jointly report α / β / γ rather than collapsing responsibility into a single score.

a) Failure-mode tags:: leftmargin=*,noitemsep,topsep=3pt

- **Over-completion**: Extra text beyond required output
- **Constraint drift**: Missing or altered required structure
- **Over-abstention**: Declining answer despite sufficient information
- **Evidence slippage**: Assumptions beyond provided sources
- **Verbosity risk**: Unnecessary elaboration complicating audit

VI. Conclusion and Future Work

This position paper argues that static benchmarks and accuracy-centric metrics provide an incomplete basis for evaluating foundation models in real-world deployment. We contend that responsibility cannot be inferred from performance alone, but must be assessed through a holistic perspective that jointly considers task effectiveness (α), risk factors (β), and governance readiness (γ). By framing evaluation as a system- and context-aware process, we highlight how models with similar benchmark performance can differ substantially in deployability, accountability, and real-world impact.

Future work should focus on operationalizing this perspective through shared evaluation infrastructure. Promising directions include the development of living benchmarks that evolve with misuse and distribution shift, standardized stress-testing suites for risk assessment, and dashboard-based reporting that preserves multi-dimensional signals rather than collapsing them into single scores. Additional research is needed to study the stability, sensitivity, and interpretability of holistic evaluation signals, as well as mechanisms for integrating technical evidence into organizational and regulatory decision-making. Advancing these efforts will require sustained collaboration between the research community, deployers, and governance institutions to ensure that evaluation practices better align with deployment realities.

VII. Acknowledgement

The authors acknowledge the use of ChatGPT (GPT-5.2, OpenAI) to assist with language editing, clarity improvement, and the generation of illustrative code snippets for selected figures. All scientific content, interpretations, and conclusions were developed, verified, and approved by the authors.

References

[1] D. Schiff, B. Rakova, A. Ayesh, L. Fenech, and K. Neely, "Explaining the principles to practices gap in ai," *IEEE Technology and Society Magazine*, vol. 40, no. 2, pp. 81–94, 2021.

TABLE I
QUALITATIVE OUTCOME MATRIX ACROSS HOLISTIC EVALUATION DIMENSIONS

Model	α: Performance Under Constraints	β: Risk Robustness	γ: Governance Readiness
GPT-5	*Strong* — Consistent schema adherence; robust multi-turn constraint repair; occasional over-abstention	*Strong* — Concise refusals; good privacy discipline	*Strong* — Evidence discipline; correct abstention; audit-ready artifacts
Claude Sonnet 3.5	*Moderate* — Fluent reasoning but frequent hard-constraint violations (extra text, count errors)	*Moderate* — Safe refusals with excess verbosity	*Moderate–Weak* — Good content, poor format discipline; partial abstention awareness
Gemini 3	*Weak–Moderate* — Correct core content but frequent over-completion breaks strict constraints	*Moderate* — Safe refusals; occasional misleading internal claims	*Weak* — Inconsistent abstention; unsupported assumptions; format drift

[2] M. Sadek, E. Kallina, T. Bohné, C. Mougenot, R. A. Calvo, and S. Cave, "Challenges of responsible ai in practice: scoping review and recommended actions," *AI & SOCIETY*, vol. 40, pp. 199–215, 2024.

[3] Q. Lu, L. Zhu, X. Xu, J. Whittle, Z. Xing, D. Douglas, and C. Sanderson, "Responsible ai pattern catalogue: A collection of best practices for ai governance and engineering," *arXiv preprint arXiv:2209.04963*, 2023.

[4] P. Radanliev, O. Santos, A. Brandon-Jones, and A. Joinson, "Ethics and responsible ai deployment," *Frontiers in Artificial Intelligence*, vol. 7, p. 1377011, 2024.

[5] S. Raza, R. Qureshi, A. Zahid, S. Kamawal, F. Sadak, J. Fioresi, M. Saeed, R. Sapkota, A. Jain, A. Zafar *et al.*, "Who is responsible? the data, models, users or regulations? a comprehensive survey on responsible generative ai for a sustainable future," *arXiv preprint arXiv:2502.08650*, 2025.

[6] L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era," *Data science journal*, vol. 14, pp. 2–2, 2015.

[7] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 220–229.

[8] A. I. Act, "Artificial intelligence act," *Regulamento da União Europeia (UE)*, vol. 1689, 2024.

[9] National Institute of Standards and Technology, "AI Risk Management Framework (AI RMF 1.0)," https://www.nist.gov/itl/ai-risk-management-framework, 2023, released January 26, 2023. Four functions: Govern, Map, Measure, Manage. Emphasizes organizational processes and lifecycle oversight.

[10] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence," https://data.europa.eu/eli/reg/2024/1689/oj, 2024, official Journal L, 12 July 2024. Risk-based governance framework with conform assessment, post-market monitoring, and incident reporting requirements.

[11] P. Ralph, M. Kuutila, H. Arif, and B. Ayoola, "Teaching software metrology: The science of measurement for software engineering," *arXiv preprint arXiv:2406.14494*, 2024, covers construct validity, reliability, representational theory of measurement, and best practices in measurement science. [Online]. Available: https://arxiv.org/abs/2406.14494

[12] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar *et al.*, "Holistic evaluation of language models," *Transactions on Machine Learning Research*, 2023, authored by Center for Research on Foundation Models, Stanford HAI. 30 models, 16 scenarios, 7 metrics, disaggregated reporting. [Online]. Available: https://arxiv.org/abs/2211.09110

[13] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, "Beyond accuracy: Behavioral testing of NLP models with CheckList," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020, pp. 4902–4912, aCL 2020 Best Overall Paper. Task-specific behavioral testing methodology. [Online]. Available: https://aclanthology.org/2020.acl-main.442

[14] Y. Zhao and G. J. Gordon, "The anatomy of evaluation measures in machine learning," *arXiv preprint arXiv:2103.06010*, 2021, impossibility results on simultaneous fairness objectives, motivates transparent trade-off reporting. [Online]. Available: https://arxiv.org/abs/2103.06010

[15] T. R. Gadekallu, K. Dev, S. A. Khowaja, W. Wang, H. Feng, K. Fang, S. Pandya, and W. Wang, "Framework, standards, applications and best practices of responsible ai: A comprehensive survey," *arXiv preprint arXiv:2504.13979*, 2025.

[16] P. P. Ray, "Benchmarking, ethical alignment, and evaluation framework for conversational ai: Advancing responsible development of chatgpt," *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, vol. 3, no. 3, p. 100136, 2023.

[17] T. Lee, H. Tu, C. H. Wong, W. Zheng, Y. Zhou, Y. Mai, J. S. Roberts, M. Yasunaga, H. Yao, C. Xie *et al.*, "Vhelm: A holistic evaluation of vision language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 140 632–140 666, 2024.

[18] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 33–44.

[19] P. Langley, "Crafting papers on machine learning," in *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, P. Langley, Ed. Stanford, CA: Morgan Kaufmann, 2000, pp. 1207–1216.