

AI Safety Evaluation – Benchmarking Framework

Maha Vishnu Sura
PhD - Student

The rapid evolution of artificial intelligence (AI) has brought immense benefits to society but also significant safety risks, including unintended behaviors, bias, and vulnerabilities to misuse. Our startup proposes the development of a comprehensive 'AI Safety Protocol and Benchmarking Test' to evaluate AI systems' safety, security, and ethical alignment before deployment. This protocol will combine 'benchmark evaluations', 'red-teaming exercises', and 'real-world simulations' to identify potential risks and ensure compliance with industry standards. What makes our approach unique is its adaptability to different AI models, integration of automated and human-in-the-loop testing, and focus on creating a standardized framework that can be adopted across industries. This solution is commercially viable as it addresses the increasing demand for trustworthy AI systems while aligning with global regulatory initiatives like the NIST AI Risk Management Framework.

1. Problem overview

AI systems are increasingly deployed in sensitive areas such as healthcare, finance, and critical infrastructure, where errors or misuse can have catastrophic consequences. Current safety evaluations often fall short due to their limited scope, lack of standardization, and inability to adapt to evolving threats. For example:

- Many benchmarks focus on performance metrics rather than safety (e.g., bias or adversarial robustness).
- Red-teaming exercises are inconsistent across organizations, leading to gaps in identifying vulnerabilities.
- Existing frameworks lack scalability and fail to address real-world complexities like dynamic threats or diverse user interactions.

Without robust safety protocols, AI systems risk causing harm through unintended actions, biased decision-making, or exploitation by malicious actors.

2. Your solution

The proposed solution is to develop a **Framework for AI Safety Evaluations and Benchmark Testing** to address critical AI safety challenges. This framework will evaluate AI models across key dimensions such as robustness, fairness, interpretability, and ethical alignment. It includes capability-specific tests like adversarial robustness assessments, fairness evaluations, and simulations for detecting deceptive behaviors (e.g., sandbagging or "person washing"). By standardizing safety protocols, the framework ensures AI systems are trustworthy

and aligned with societal values while mitigating risks of misuse in high-stakes applications.

The framework's technical foundation integrates state-of-the-art tools like Clever Hans for adversarial testing, IBM's AI Fairness 360 for bias detection, and SHAP for interpretability. It also incorporates ethical alignment methods such as reinforcement learning from human feedback (RLHF). A cloud-based benchmarking platform will enable large-scale evaluations, with a public leaderboard to incentivize improvement among developers. This modular design ensures adaptability across industries like healthcare, finance, and autonomous systems, while continuous monitoring tools ensure ongoing compliance with safety standards.

This solution is innovative due to its comprehensive scope and scalability. It combines cutting-edge research in adversarial testing and ethical AI into a unified evaluation framework. By fostering transparency and accountability, the framework addresses critical gaps in AI safety while positioning itself as a potential industry standard. Its modularity and ability to adapt to emerging threats make it a vital tool for ensuring safe and ethical AI deployment across diverse applications.

3. Pilot experiment or demo

During the hackathon, we outlined a step-by-step strategy to develop a Minimum Viable Product (MVP) for an AI safety evaluation framework. The first step involves simulating deepfake-based KYC scenarios to detect vulnerabilities such as identity theft and "person washing." Next, adversarial testing will be conducted using techniques like Wild Teaming to evaluate models' robustness against malicious inputs. Following this, deceptive alignment evaluations will focus on detecting latent planning behaviors encoded in non-human-readable formats, such as high-dimensional embeddings. Finally, a modular design will be integrated to ensure scalability across various industries and adaptability to emerging threats.

If implemented, this strategy is anticipated to produce a comprehensive framework capable of identifying vulnerabilities in AI systems with high accuracy. It will address critical safety challenges such as adversarial misuse and deceptive alignment while demonstrating scalability through its modular design. This approach ensures that the framework can be applied across diverse domains, including finance, healthcare, and autonomous systems, providing a robust foundation for safe and ethical AI deployment.

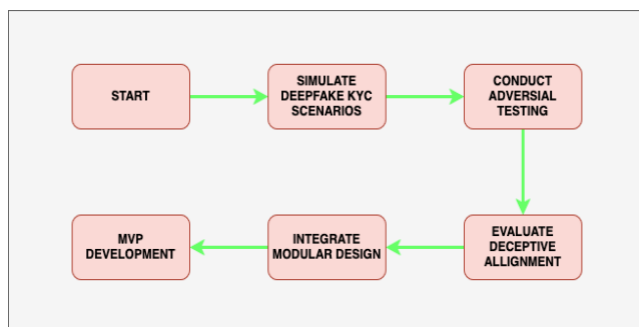


Figure 1 – MVP development Roadmap

4. Process

If given unlimited resources and one year,

Timeframe	What will you do?
Next 3 months	Develop and integrate advanced evaluation modules for detecting emerging threats, such as steganographic planning and situational awareness misuse. Begin forming partnerships with regulatory bodies like NIST and ISO to align the framework with industry standards.
2025	Launch a cloud-based benchmarking platform for global accessibility, enabling automated safety evaluations. Expand the protocol's scope to include domain-specific benchmarks tailored for industries like healthcare and finance.
2026	Strengthen collaborations with regulatory bodies to establish the framework as an industry standard. Enhance the platform's scalability and modularity to adapt to evolving AI safety challenges across diverse applications.
2027	Achieve widespread adoption of the framework globally as a trusted certification tool for AI safety. Continuously update evaluation modules to address new threats and maintain relevance in rapidly advancing AI technologies.

5. Impact on AI safety & key risks

Our solution enhances AI safety by

1. Positive Impact:

- Reducing risks of unintended harm through rigorous pre-deployment testing.
- Identifying vulnerabilities related to adversarial misuse (e.g., deepfakes or deceptive planning).
- Promoting ethical alignment by addressing bias and transparency issues.
- Mitigating misuse risks via continuous monitoring frameworks.

2. Key Risks & Mitigation Strategies:

- Potential risks include overfitting benchmarks or resistance from stakeholders due to implementation costs. These will be mitigated by regularly updating benchmarks based on real-world feedback.
- Resistance from developers due to increased compliance costs → Mitigation: Demonstrate ROI through reduced liability risks.

- Rapidly evolving threats → Mitigation: Continuous updates based on emerging research.
3. Success Metrics:
- Adoption rate by organizations across industries.
 - Reduction in reported incidents involving unsafe AI deployments.
 - Number of submissions on the public benchmarking leaderboard.
 - Positive feedback from regulatory audits and industry partners

6. Appendix

By addressing critical gaps in AI safety evaluations, our protocol aims to set a new standard for responsible AI deployment while fostering societal trust in these transformative technologies.

Citations:

1. <https://www.frontiermodelforum.org/updates/issue-brief-preliminary-taxonomy-of-pre-deployment-frontier-ai-safety-evaluations/>
2. <https://www.lesswrong.com/posts/gaKar8ga6ycJASCe5/which-ai-safety-benchmark-do-we-need-most-in-2025>
3. <https://testpros.com/artificial-intelligence/understanding-upcoming-ai-testing-standards-vet-ai-act/>
4. <https://techcrunch.com/2024/08/04/many-safety-evaluations-for-ai-models-have-significant-limitations/>
5. <https://aiethicslab.rutgers.edu/e-floating-buttons/responsible-ai-safety-and-effectiveness-raise-benchmarks/>
6. <https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/>
7. <https://partnershiponai.org/ai-safety-in-practice-insights-from-a-joint-workshop-with-pai-and-mlcommons/>
8. <https://allenai.org/evaluation-frameworks>
9. <https://www.techpolicy.press/no-safety-without-standards-defining-protocols-for-ai-redteaming-disclosures/>
10. <https://www.aisi.gov.uk/work/early-lessons-from-evaluating-frontier-ai-systems>
11. https://downloads.regulations.gov/NIST-2023-0009-0120/attachment_1.pdf
12. <https://campustechnology.com/articles/2024/04/03/us-uk-ai-testing-frameworks.aspx>
13. <https://arxiv.org/html/2404.05388v3>
14. <https://www.edsafeai.org/safe>
15. <https://www.tigera.io/learn/guides/llm-security/ai-safety/>
16. <https://research.google/blog/supporting-benchmarks-for-ai-safety-with-mlcommons/>
17. <https://cloudsecurityalliance.org/artifacts/ai-resilience-a-revolutionary-benchmarking-model-for-ai-safety-japanese-translation>

18. <https://www.industry.gov.au/publications/voluntary-ai-safety-standard>
19. <https://www.holisticai.com/blog/generative-ai-evaluations-audits>
20. <https://dl.acm.org/doi/10.1145/3664646.3664766>
21. <https://www.nist.gov/artificial-intelligence/ai-standards>
22. <https://www.nist.gov/artificial-intelligence/executive-order-safe-secure-and-trustworthy-artificial-intelligence/test>
23. <https://www.nist.gov/aisi>
24. <https://aaai.org/aaai-publications/aaai-publication-policies-guidelines/>
25. <https://www.aisi.gov.uk/work/inspect-evals>