Healthcare AI Model Evaluator: Responsible AI FAQ

- **What is Healthcare AI Model Evaluator?**
- *Healthcare AI Model Evaluator is a comprehensive medical AI model benchmarking tool designed to facilitate evaluation of healthcare models on arbitrary datasets. Built with Microsoft Safe AI principles, it emphasizes transparency, accountability, and reliability. The tool is engineered to support responsible innovation by ensuring that all evaluation processes are auditable, explainable, and compliant with healthcare standards. Its architecture enables organizations to assess AI models in a secure, controlled environment, fostering trust and safety in clinical AI adoption.*
- **What can Healthcare AI Model Evaluator do?**

  Healthcare AI Model Evaluator is a medical AI model benchmarking platform with an integrated evaluation engine to assist multi-disciplinary healthcare teams in building and validating AI systems. It accepts user-curated and defined data as input, utilizes user-defined and controlled endpoints (where applicable) to generate outputs, and enables user-defined evaluators to assess and score model outputs.

  Key capabilities include:
    - *Supporting diverse evaluation scenarios (text, image, multi-image, custom rubrics, and questions).*
    - *Enabling human-in-the-loop review for clinical relevance and safety.*
    - *Providing customizable scoring and reporting to meet specific regulatory and organizational requirements.*
    - *Ensuring all data flows and endpoints are within the control of the customer deployment.*
    - *Maintaining transparency and traceability throughout the evaluation lifecycle.*
- **What is/are Healthcare AI Model Evaluator's intended use(s)?**
  Healthcare AI Model Evaluator is intended to empower customers and end-users to control and monitor all aspects of AI model output evaluation. Its design prioritizes user agency, allowing healthcare organizations to:
    - *Define evaluation criteria tailored to clinical needs.*
    - *Monitor model performance and safety continuously.*
    - *Document and audit evaluation processes for regulatory compliance.*
    - *Integrate responsible AI governance, including risk management and incident reporting.*
    - *Promote responsible use by ensuring that only authorized personnel can configure endpoints and access sensitive data.*
- **How was Healthcare AI Model Evaluator evaluated? What metrics are used to measure performance?**
  Healthcare AI Model Evaluator is tested by the product team and participating Microsoft Research teams to ensure it fulfills its intended purpose. As an AI evaluation tool, internal assessments focus on:
    - *Workflow robustness and coverage of diverse evaluation scenarios (text, image, multi-image, custom rubrics, and questions).*
    - *Accuracy, reliability, and reproducibility of evaluation results.*
    - *Security and compliance with healthcare data standards.*
    - *Usability and clarity of reporting for clinical teams.*

- *Alignment with Microsoft Safe AI principles, including transparency, fairness, and accountability.*

Performance metrics include precision, recall, inter-rater reliability, and ability to export evaluation experiments and results for full transparency. Continuous monitoring and feedback loops are supported to identify and address limitations or emerging risks.

- **What are the limitations of Healthcare AI Model Evaluator? How can users minimize the impact of Healthcare AI Model Evaluator's limitations when using the system?**
  Healthcare AI Model Evaluator cannot currently evaluate pure embedding models or AI models that generate video and/or 3D images. Additionally, the tool relies on model endpoints and input data (including prompts) created and configured by users.
  To minimize the impact of these limitations:
  - *Administrators should adhere to Microsoft Safe AI principles, including responsible data stewardship, robust security controls, and ongoing compliance monitoring.*
  - *Users should ensure that all endpoints and data sources are validated for healthcare use and regularly audited for compliance.*
  - *Evaluation workflows should incorporate human oversight, especially for novel or high-risk model types.*
  - *Organizations should maintain clear documentation of evaluation processes and limitations, and provide training to evaluators on responsible AI use.*
  - *Feedback mechanisms should be established to report issues, monitor performance, and drive continuous improvement.. Please see below.*
- **What operational factors and settings allow for effective and responsible use of Healthcare AI Model Evaluator?**

  Data and Model Endpoint Compliance:

  All data sources and model endpoints integrated with the Healthcare AI Model Evaluator must strictly adhere to healthcare standards such as HIPAA, GDPR, and relevant FDA or ISO regulations. This means ensuring that patient data is protected through robust privacy, security, and audit controls, and that endpoints are validated for safe, ethical use in clinical environments. Regular compliance checks and documentation updates are essential to maintain regulatory alignment.

  Secure, Auditable Environment:

  Deploy the evaluator within a secure, organization-controlled environment (e.g., Azure with healthcare compliance features enabled). This supports data privacy, regulatory compliance, and full auditability of evaluation logic, model performance, and user actions.

  Customizable Evaluation Criteria:

  Empower clinical professionals to define context-specific metrics and benchmarks. The evaluator should allow creation of custom evaluation questions, scoring rubrics, and export of results for further analysis, ensuring that assessments reflect real-world clinical needs and priorities.

  Human-in-the-Loop Review:

Integrate human expert review into the evaluation workflow. Assign qualified clinicians to assess model outputs for accuracy, relevance, and safety, supplementing automated scoring with domain expertise. This mitigates automation bias and ensures that AI recommendations are clinically sound.

Data Quality and Interoperability:

Use high-quality, representative datasets—preferably those curated or approved by clinical experts. Data should be accurate, complete, consistent, and relevant to the intended use case. Interoperability with healthcare data standards (e.g., HL7, FHIR) is critical for seamless integration and reliable analysis

Governance and Policy Frameworks:

Establish formal governance structures for AI deployment, including risk-based management principles, lifecycle oversight, and continuous monitoring. Policies should address selection, implementation, and ongoing evaluation of both internally developed and third-party AI solutions.

Continuous Monitoring and Reporting:

Implement processes to monitor model performance, capture safety or performance issues, and enable voluntary reporting of AI-related incidents. Regular reviews and updates help identify and mitigate risks, biases, or drift in model behavior.

Training and Education:

Provide ongoing training for clinicians and staff on responsible AI use, including understanding model limitations, interpreting outputs, and maintaining vigilance against overreliance on automated recommendations.

Transparency and Explainability:

Ensure that evaluation processes and model decisions are transparent and explainable. Maintain clear documentation of evaluation logic, data provenance, and model behavior to foster trust and accountability

- **How do I provide feedback on Healthcare AI Model Evaluator?**
  We encourage all users and contributors to share their experiences, suggestions, and concerns through the official support channels. Please follow the steps below to ensure your feedback is received and addressed appropriately:
  *• **Use the GitHub Support Mechanism:***
  Healthcare AI Model Evaluator is maintained as an open-source project on GitHub. To provide feedback, report issues, or request new features, please visit the official repository github.
  You can open a new issue describing your feedback, bug report, or feature request.
- *For general questions or discussions, use the repository's Discussions tab if available.*
  Adhere to the Code of Conduct:
  When providing feedback or contributing to the project, please ensure your interactions are respectful, constructive, and aligned with the project's code of conduct.
  All feedback should promote a positive, collaborative environment and support responsible innovation in healthcare AI.

Responsible Feedback Principles:

In line with Microsoft Safe AI principles and the broader healthcare AI community, feedback should:

- *Advance the safety, reliability, and fairness of the evaluator.*
- *Highlight opportunities to improve transparency, explainability, and user control.*
- *Address any concerns about data privacy, security, or ethical use.*
- *Support equitable access and benefit for all users and patient populations.*

Continuous Improvement:

Your feedback helps ensure that Healthcare AI Model Evaluator remains a trustworthy, effective, and human-centered tool. All suggestions are reviewed by the project maintainers and, where appropriate, incorporated into future releases. Please note that feedback may be discussed publicly to foster transparency and collective learning.