Department of Health Research
Ministry of Health and Family Welfare
Government of India

MATTER: Metrics and Assessment Tools for Trustworthy, Transparent, Explainable, and Reliable AI

Dr. Priyanka Narad

Scientist-C

Indian Council of Medical Research Hqrs,

Ansari Nagar, New Delhi

@ pnarad.hq@icmr.gov.in

The AI Imperative: Why Trust is the New Metric

Context of AI's Expansion:

Artificial Intelligence is rapidly moving from laboratory experiments to critical, real-world deployment. We see profound impacts in domains such as precision healthcare diagnostics, complex financial modeling, and administrative decision-making systems.

The Core Challenge

As AI systems become indispensable, their opaqueness, the "black box" nature, presents significant ethical, operational, and regulatory risks. Untrustworthy AI can lead to misdiagnoses, discriminatory lending, and systemic failures.

The Focus of This Talk:

Today's discussion introduces MATTER, a holistic framework designed to enforce rigor, trust, and accountability. We will explore how MATTER integrates cutting-edge global evaluation.

AI Deployment: The Double-Edged Sword

AI's Transformative Reach

- **Healthcare:** *Early disease detection, personalized treatment plans.*
- **Finance:** *Algorithmic trading, credit scoring, fraud detection.*
- **Administration:** *Resource allocation, automated regulatory compliance.*

The sophistication of AI magnifies potential failure modes

The Corresponding Risks

- **Model Failure:** *Catastrophic decision-making due to brittle models (e.g., misclassifying a tumor).*
- **Ethical Harm:** *Perpetuation and scaling of historical human biases (Fairness).*
- **Security Breaches:** *Data leakage through model inversion or adversarial attacks (Privacy/Robustness).*
- **Accountability Gap:** *Who is responsible when a black-box AI causes harm?*

Moving Beyond Anecdote: Why Systematic Frameworks are Essential

1

2

3

Inconsistent Evaluation

Currently, AI evaluation is often ad-hoc, focusing solely on simple accuracy metrics in controlled settings. This fails to capture real-world performance.

The Requirement for Reproducibility and Auditability

To be trustworthy, AI development must be auditable at every stage, from data selection to post-deployment monitoring. This is a foundational principle of frameworks like STARD-AI for diagnostic models.

From Accuracy to Trustworthiness

Trustworthy AI requires verifiable performance across five pillars:

- ***T**ransparency*
- ***R**eliability/Robustness*
- ***U**nderstandability (Explainability)*
- ***S**ecurity/Privacy*
- ***T**reatability (Fairness/Governance)*

MATTER: Metrics and Assessment Tools for Trustworthy AI

MATTER is an integrated architecture that standardizes the evaluation of complex AI systems, ensuring they meet rigorous standards for performance and ethics.

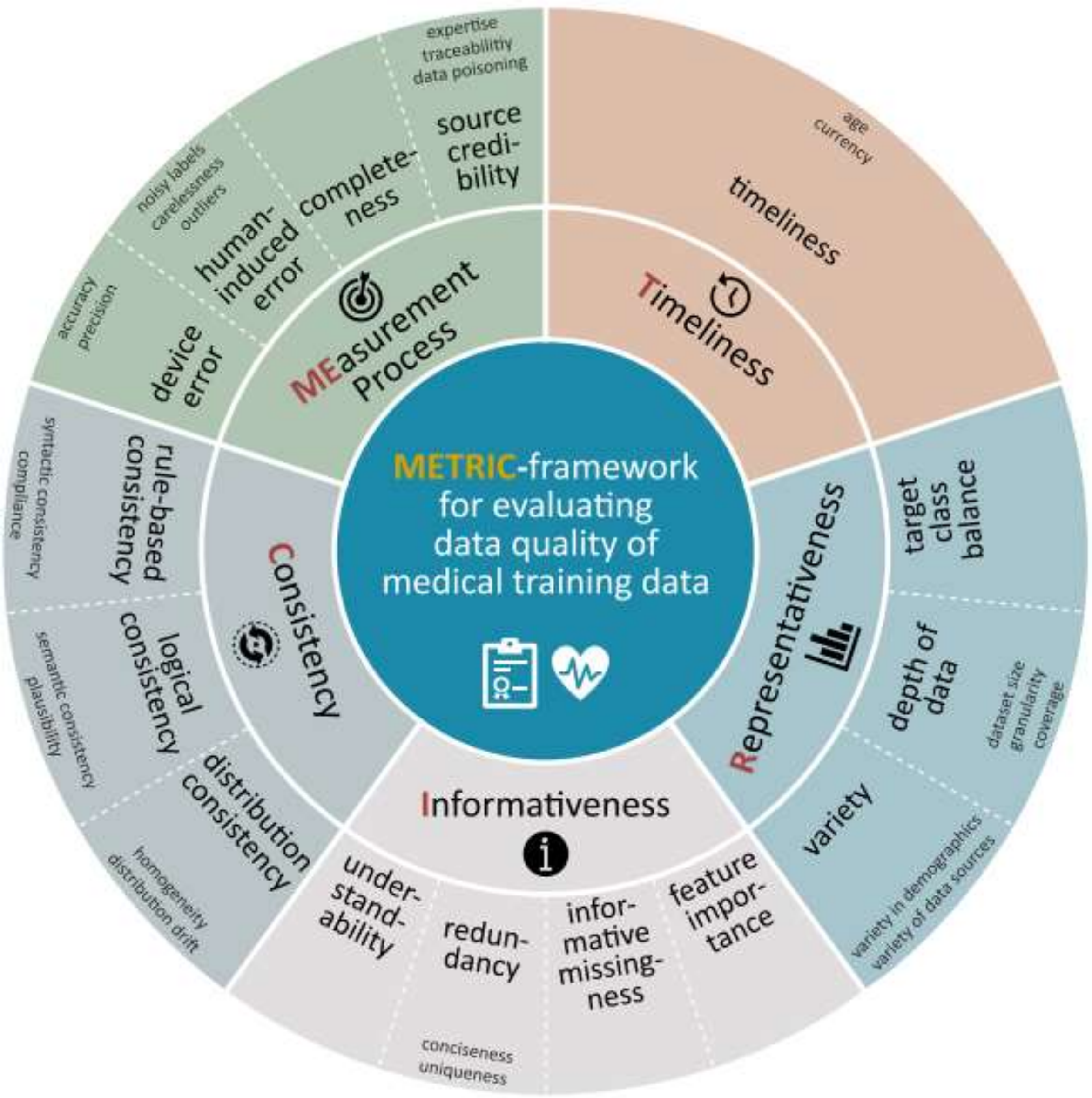
The Acronym Breakdown

- **Metrics**: *Quantitative performance indicators (e.g., F1 Score, Robustness score).*
- **Assessment Tools**: *Software and methodologies (e.g., LIME, IBM AIF360, Benchmarks).*
- **Trustworthy**: *Focusing on model integrity, security, and ethical alignment.*
- **Transparent**: *Clarity regarding data, model behavior, and decision logic.*
- **Explainable**: *Providing human-understandable reasons for outputs.*
- **Reliable**: *Performance consistency over time and across domain shifts.*

Department of Health Research
Ministry of Health and Family Welfare
Government of India

The METRIC-framework

INDIAN COUNCIL OF MEDICAL RESEARCH
Serving the nation since 1911



- This specialised framework for evaluating data quality of the content of medical training data includes a comprehensive set of awareness dimensions.
- The inner circle divides data quality into five clusters.
- These clusters contain a total of 15 data quality dimensions, which are shown on the outer circle.
- The sub dimensions presented in grey on the border of the figure contribute to the superordinate dimension.

The Three Foundational Pillars of the MATTER Framework

MATTER organizes AI validation into three interconnected areas:

Pillar I

1

- ***Benchmarking (External Validation)***
  - *Rigorous, large-scale, and multi-dimensional performance testing against diverse datasets and real-world threats.*

Pillar II

2

- ***Explainability (Internal Transparency)***
- *Providing model-agnostic and human-interpretable insights into how a specific decision was reached.*

Pillar III

3

- ***Governance & Accountability (Societal Alignment)***
  - *Establishing operational processes for data lineage, bias monitoring, and regulatory compliance throughout the model lifecycle.*

Benchmarking AI: Testing Limits and Adversarial Resilience

1

Beyond Test-Set Accuracy

Traditional testing is insufficient. Modern benchmarks must test AI against synthetic and real-world challenges that break simple correlation.

2

Multi-Dimensionality

MATTER necessitates testing across at least 18 dimensions, including:

Robustness

Resistance to adversarial inputs and perturbations.

Hallucination

For LLMs, minimizing fabricated or unsupported outputs.

Bias/Fairness

Equitable outcomes across various demographic subgroups.

3

The Role of Large-Scale Benchmarks

Using standardized, extensive benchmark suites allows for objective, comparative evaluation across different models and research groups.

Department of Health Research
Ministry of Health and Family Welfare
Government of India

TrustLLM: Comprehensive Trustworthiness for Large Language Models

INDIAN COUNCIL OF MEDICAL RESEARCH
Serving the nation since 1911

Key Dimensions: *This benchmark utilizes 30+ datasets to measure performance across critical LLM risks:*

1. **Robustness to Prompt Injection/Jailbreaking:** *Evaluating resistance to malicious inputs designed to override safety controls.*
2. **Safety and Toxicity:** *Quantifying the generation of harmful or biased content.*
3. **Privacy Leakage:** *Assessing the model's propensity to reveal sensitive data it was trained on.*

Focus Area:
TrustLLM specifically targets Large Language Models (LLMs), which present unique challenges due to their emergent and conversational nature.

The TrustLLM Score:
Provides a consolidated, multi-factor score that moves beyond simple fluency to measure an LLM's deployability in sensitive applications.

AIR-Bench 2024: Focus on Reliability and Real-World Domain Shift

AIR-Bench 2024 (AI Robustness) emphasizes reliability and generalization, critical for systems like AI-based Software as a Medical Device (AI-SaMD).

Key Evaluation Dimensions:

1. **Temporal Reliability:** Ensuring performance does not degrade over time (algorithm 'decay' or 'drift') - a key component of the FDA's proposed TPLC (Total Product Life Cycle) approach.
2. **Domain Shift Assessment:** Crucially, evaluating performance when a model trained on one population (e.g., US data) is deployed in another (e.g., Indian context).
3. **Data Quality Robustness:** How well the model handles noisy, corrupted, or incomplete real-world inputs.

Department of Health Research
Ministry of Health and Family Welfare
Government of India

Mandating Global Reporting Standards for AI Evidence

ICMR
INDIAN COUNCIL OF MEDICAL RESEARCH
Serving the nation since 1911

1

The Need for Structured Reporting

Benchmarking results are meaningless without standardized, transparent reporting.

2

STARD-AI
(STAndards for Reporting Diagnostic Accuracy Studies)

An essential reporting guideline for studies evaluating AI-based diagnostic tests.

***Mandates**: Clear reporting on dataset characteristics, including training, tuning, and external validation sets (as seen in WHO guidelines for AI-SaMD).*

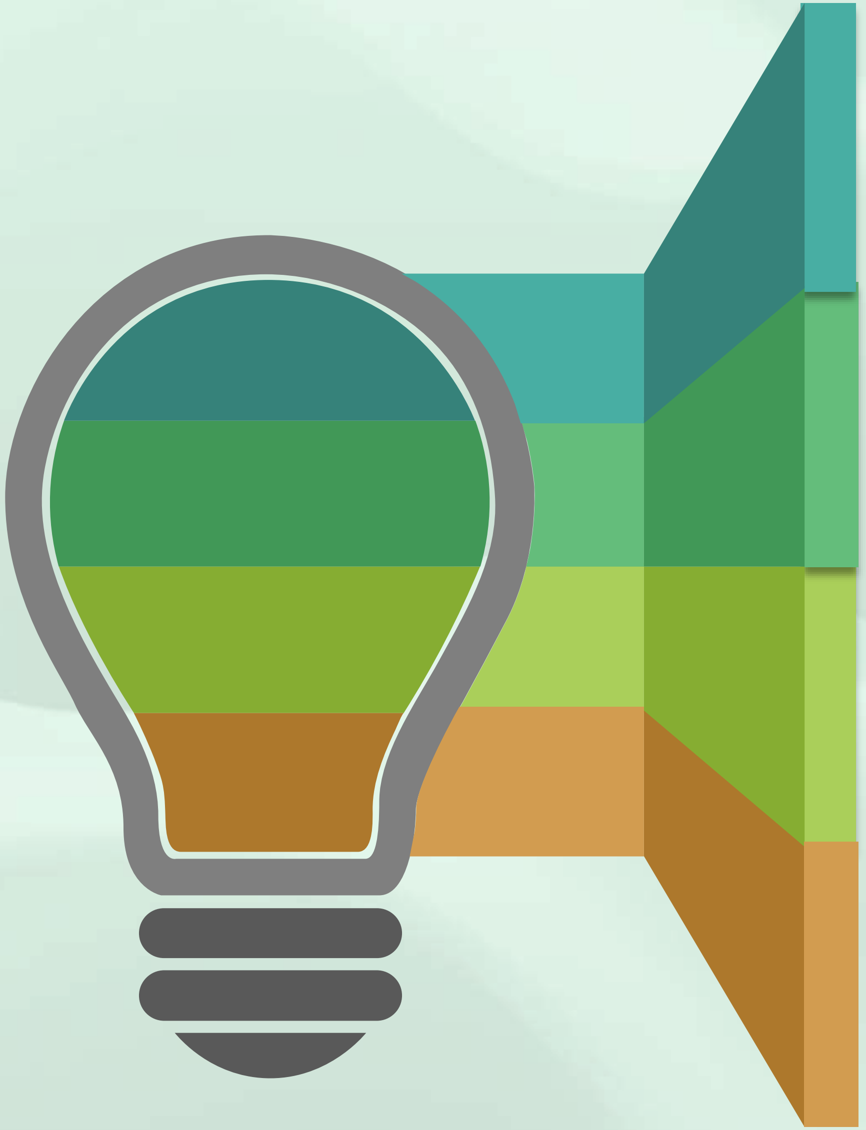
***Focus on Bias**: STARD-AI includes new items (e.g., 23*) for details on algorithmic bias and fairness assessments .*

3

TRIPOD+AI & CONSORT-AI

Guidelines for prediction model development and clinical trial protocols involving AI interventions, ensuring methodological rigor from concept to trial report.

Explainable AI (XAI): From Prediction to Justification



The XAI Goal

- To enable human users to understand, trust, and effectively manage AI-driven decisions. An explanation is not just *what* the model predicted, but *why*.

The Model-Agnostic Advantage

- **MATTER** prioritizes model-agnostic tools that can be applied universally, regardless of the underlying algorithm (deep learning, classical ML).

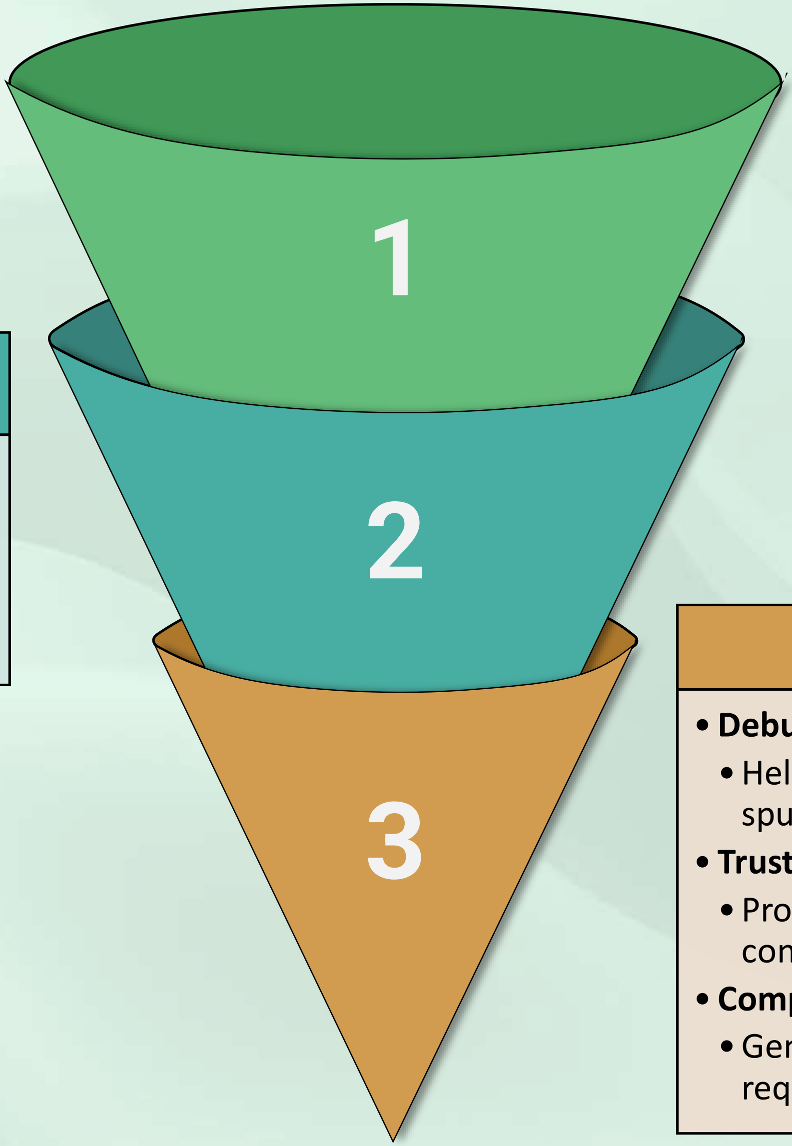
Key Explainability Tools:

- **LIME (Local Interpretable Model-agnostic Explanations):** Explains individual predictions by locally approximating the model with an interpretable model.
- **SHAP (SHapley Additive exPlanations):** Based on game theory, it assigns a contribution value to each feature for a given prediction.
- **Saliency Mapping:** Visually highlights the input regions (e.g., pixels in an X-ray) that were most critical to the model's output.

Department of Health Research
Ministry of Health and Family Welfare
Government of India

The AI Explainability 360 Toolkit: Integrated Interpretation

INDIAN COUNCIL OF MEDICAL RESEARCH
Serving the nation since 1911



Implementation Focus

- MATTER recommends leveraging comprehensive open-source toolkits like IBM's AI Explainability 360 (AIX360).

Toolkit Features

- AIX360 embeds multiple XAI algorithms (LIME, SHAP, etc.) into a single platform, facilitating seamless integration into development pipelines.

The Value Proposition

- **Debugging**
  - Helps developers identify and fix faulty model logic or spurious correlations.
- **Trust**
  - Provides clinicians and domain experts the necessary context for human-in-the-loop validation.
- **Compliance**
  - Generates artifacts (e.g., local feature importance reports) required for regulatory submission.

XAI and Bias: Uncovering Algorithmic Inequity in Diagnosis

The Scenario

An AI model is trained to detect pneumonia from chest X-rays. While overall accuracy is high, it systematically underperforms for specific subpopulations.

1

2

MATTER's Action

This XAI insight forces data scientists to re-examine the training data for selection or annotation bias, leading to model retraining and improvement in fairness metrics before deployment.

3

Diagnosis with XAI:

Applying SHAP reveals that for a certain demographic group, the model relies on irrelevant features (e.g., hospital ID watermark, image texture) instead of clinical signs.

Saliency maps show the model looking at the wrong part of the image for certain skin tones or body types due to data imbalance.

Governance in MATTER: Bridging Technical Evaluation and Societal Accountability

The Governance Mandate

Governance moves beyond technical metrics to establish organizational and procedural rigor for responsible deployment.

Key Governance Metrics and Practices

Data Lineage Tracking: *Mandatory tracking of all data sources, transformations, and biases introduced at each stage - from collection to deployment.*

Bias Monitoring Pipelines: *Continuous, real-time auditing of model outputs for disparate impact across defined groups in the production environment.*

Adaptive Trust Calibration: *Dynamic adjustment of the human-in-the-loop threshold based on the model's performance stability and real-world conditions (e.g., reducing human override when confidence is high, and vice versa).*

Ensuring Durability: The Total Product Life Cycle (TPLC) Approach

WHO and FDA Alignment:

The TPLC approach (adopted by the FDA and referenced in WHO guidelines) is central to MATTER's reliability focus

- ***Phase IV (Durability/Monitoring):*** *Performance monitoring must be ongoing.*
- ***Predetermined Change Control Plan (PCCP):*** *Developers must pre-specify what types of model changes can be automatically implemented and which require regulatory re-review.*

MATTER's Requirement: *Continuous logging and tracking of deployed model performance against pre-specified safety and effectiveness targets.*

The Challenge of Algorithm Drift

- *Unlike traditional software, AI models degrade over time as real-world data distributions change (e.g., new strains of disease, changing population demographics).*

ICMR's Ethics Guidelines 2023

Context

The Indian Council of Medical Research (ICMR) has taken a leading role, releasing the 2023 Guidelines on AI in Biomedical Research and Healthcare.

Key ICMR Principles (Deeply Aligned with MATTER)

Algorithmic Transparency:

Requiring detailed reporting on model logic and decision paths (Pillar II: Explainability).

Human-in-the-Loop Validation:

Mandating human oversight, especially in high-risk tasks.

India-Specific Datasets:

Crucially, addressing the bias of global models by emphasizing the use of diverse, Indian-contextualized data for training and validation.

Task-Based Risk Evaluation

Categorizing AI applications based on potential harm and tailoring evaluation rigor, accordingly, focusing resources on high-stakes clinical decisions.

Driving Context-Sensitive Bias Assessments

Utilizing AI itself to enhance the efficiency, integrity, and regulatory compliance of clinical trials (e.g., patient recruitment, data quality checks).

AI-Driven Clinical Trial Management:

Digital Health Registries

ICMR is investing in standardized, high-quality digital health registries to aggregate large, diverse, and well-annotated Indian data. This directly mitigates global model bias.

Context-Sensitive Bias

The key differentiator: *Evaluating fairness based on local social, cultural, and genetic diversity. A model fair in one region of India may not be fair in another due to epidemiological or input data differences.*

A Matter of Logic: The Challenge of Chain-of-Thought Reliability

What is CoT?

- *Chain-of-Thought (CoT) prompting or reasoning refers to the intermediate steps an LLM generates to arrive at a final answer (e.g., the steps in a medical diagnosis or a financial calculation).*

The Reliability Gap

- *While CoT output looks logical and enhances transparency, the steps themselves can be unreliable or fabricated (hallucinated reasoning) even if the final answer is coincidentally correct.*

MATTER's Stance

- *The framework requires new metrics that validate the internal consistency and factual accuracy of each CoT step, not just the final output. The explanation must be reliable, not just plausible.*

The Moving Target: Dynamic Bias and Continuous Fairness Monitoring

Static vs. Dynamic Bias

Bias isn't a one-time check. Static bias is found in the training data; dynamic bias emerges when a deployed model interacts with a changing user population or when an intervention changes user behavior itself.

Fairness Under Domain Shift

This is the challenge of ensuring fairness metrics (e.g., equal opportunity, demographic parity) hold true when the model is applied to an entirely new clinical or socioeconomic context.

The FairDomain Approach (Conceptual)

A model for continuous retraining and bias mitigation that uses active learning to detect domain shift and trigger recalibration specifically for sensitive subgroups - a necessity for MATTER compliance.

Safeguarding Sensitive Data: Privacy Leakage Analysis

Data Sensitivity

In healthcare and finance, data is often protected (e.g., GDPR, HIPAA, India's DPDPA). AI models inherently learn from this data.

Privacy Leakage

Benchmarks like TrustLLM specifically test for methods like model inversion and membership inference attacks, where an attacker can deduce private training data from the model's outputs or parameters.

MATTER's Defense Requirements

Differential Privacy

Requiring models to be trained with techniques that mathematically bound the information an attacker can gain about any single individual in the training set.

Securing Explanations

Ensuring XAI outputs (e.g., saliency maps) do not inadvertently reveal private patient data.

Ethical Reasoning: Evaluating AI's Adherence to Principles

Beyond Technical Compliance

MATTER mandates evaluation of how AI aligns with core ethical principles, particularly for high-stakes decisions.

Scenario Testing

Using specialized datasets to test the model's response to ethical dilemmas (e.g., resource allocation, triage decisions).

Does the model's output adhere to ICMR's core ethical guidelines?

The Alignment Challenge

Ensuring that the technical objective function (what the model is trying to optimize) is perfectly aligned with the desired human ethical or societal outcome (e.g., optimizing for maximum life-years saved rather than just total lives saved).

MATTER as the Unifying Architecture: Global Best Practices, Local Relevance

Synthesis of Standards

- *MATTER is not just a collection of tools; it is the glue that connects:*
  - ***Global Benchmarks (TrustLLM, AIR-Bench):*** *Providing internationally recognized rigor and technical security checks.*
  - ***International Reporting Guidelines (STARD-AI, CONSORT-AI):*** *Ensuring technical results are transparently communicated.*
  - ***Local Policy (ICMR Guidelines):*** *Tailoring the final assessment for India-specific needs - human-in-the-loop, context-sensitive bias, and India-specific datasets.*

Outcome

- *This unified architecture enables the development of AI systems that are:*
  - *Technically Robust (via Benchmarking).*
  - *Ethically Sound (via Governance & ICMR principles).*
  - *Societally Aligned (via XAI and Context-Sensitive Bias monitoring).*

The Future of Trust: AI Certification and Auditing

Formalizing Trust

- *The ultimate vision for MATTER is to serve as the foundation for formal AI certification - similar to ISO standards or medical device approvals.*

The AI Audit Trail

- *Establishing a robust, verifiable audit trail that demonstrates compliance across all three pillars:*
  - *Benchmarking*
  - *Explainability*
  - *Governance*

Collaboration with Regulators

- *Working alongside bodies like the Central Drugs Standard Control Organisation (CDSCO) to integrate MATTER metrics into regulatory clearance pathways for AI-SaMDs.*

Continuous Improvement

- *The framework itself must be adaptive, continuously incorporating new benchmarks and governance practices as AI technology evolves.*

MATTER: Final Key Takeaways

AI Trust is Multi-Dimensional

- *Trustworthiness requires rigorous evaluation across robustness, fairness, privacy, and reliability, not just accuracy.*

The Three Pillars

- *MATTER provides structure through three focus areas: Benchmarking, Explainability (XAI), and Governance.*

XAI is Essential for Debugging and Trust

- *Tools like LIME and SHAP are vital for transforming 'black-box' predictions into human-understandable justifications.*

India is Setting the Pace

- *The ICMR guidelines provide a critical model for operationalizing AI ethics through local data focus and human oversight, providing a blueprint for MATTER's implementation in the Indian context.*

Challenges Remain

- *We must overcome issues in CoT reliability, dynamic bias detection, and domain shift to achieve truly reliable deployment.*

References

1. **US FDA.** (2021). *Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan.* U.S. Food and Drug Administration.

2. **Indian Council of Medical Research (ICMR).** (2023). *Ethical Guidelines for Application of Artificial Intelligence in Biomedical Research and Healthcare.* New Delhi: ICMR.

3. **European Union.** (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act).* Official Journal of the European Union.

4. **Sounderajah, V., et al. (STARD-AI Steering Committee).** (2025). The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nature Medicine*.

5. **CLAIM Working Group.** (2024). CLAIM: A Reporting Checklist for Artificial Intelligence in Medical Imaging. *Nature Medicine*.

6. **Giraud, R., et al. (DECIDE-AI Consensus Group).** (2024). DECIDE-AI: Consensus Reporting Guideline for Early Stage AI Models. *The Lancet Digital Health*.

7. **Huang, Y., et al.** (2024). TrustLLM: Trustworthiness in Large Language Models. *arXiv:2401.05561*.

8. **Arya, V., et al. (IBM)** (2019). AI Explainability 360: An Extensible Toolkit for Understanding Data and Machine Learning Models. *JMLR*.

Thank you for your attention!