REAL WORLD MATTERS: WHAT ACTUALLY HAPPENS WHEN PEOPLE USE AI?

Assessing Risks and Impacts of AI (ARIA)

NIST | NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

DALL-E GENERATED IMAGE

THE ARIA TEAM

Leads

*NIST Associate

Reva Schwartz

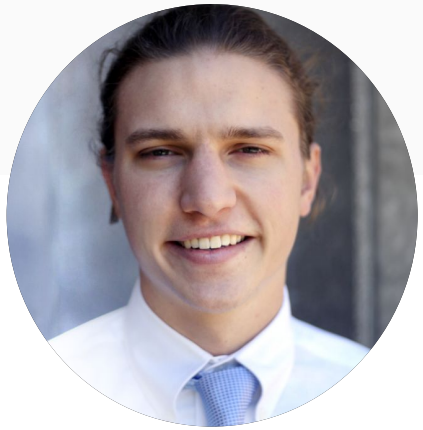
Program Lead

Linguistics/Phonetics, Human-Language Technology, AI Risk Management, Experimental Methods

Jonathan Fiscus

Evaluation And Model Testing Lead

Computer Science, Natural Language Processing, Computer Vision, AI Measurement and Evaluation



Theodore Jensen

Field Testing Lead

Computer Science, Human-Computer Interaction, User Trust

Gabriella Waters*

Evaluation Harmonization

Human-Centered Computing, Psychology, Neuroscience, Biology, Genetics

Razvan Amironesei

Annotation Lead

Data Annotation, Data Documentation, Humanistic Social Science, AI Risk Analysis

THE ARIA TEAM

Contributors

Experts drawn from
NIST AI Innovation Lab

Afzal Godil
Model Testing

Kristen Greene
Field Testing

Craig Greenberg
Scoring Methodology

Anya Jones*
Psychometrics

Patrick Hall*
Red Teaming

Noah Schulman
Software Programming

Shomik Jain
Cross-team Support

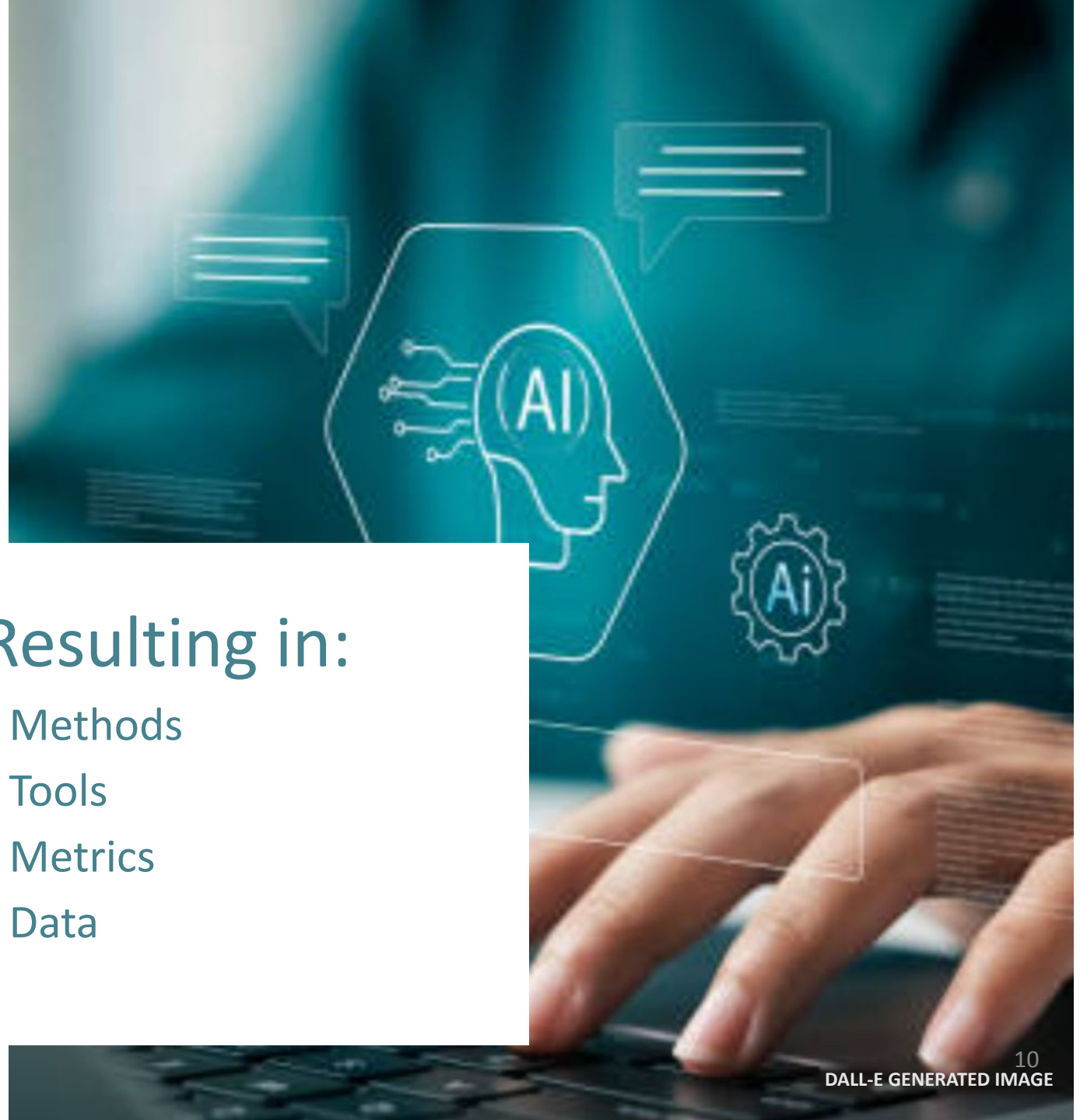
*NIST Associates

What is ARIA?

NIST has a long history of conducting evaluation-driven research….

That is:

- Collaborative
- Long Term
- Exploratory
- Open and Transparent



Resulting in:

- Methods
- Tools
- Metrics
- Data

DALL-E GENERATED IMAGE

ARIA

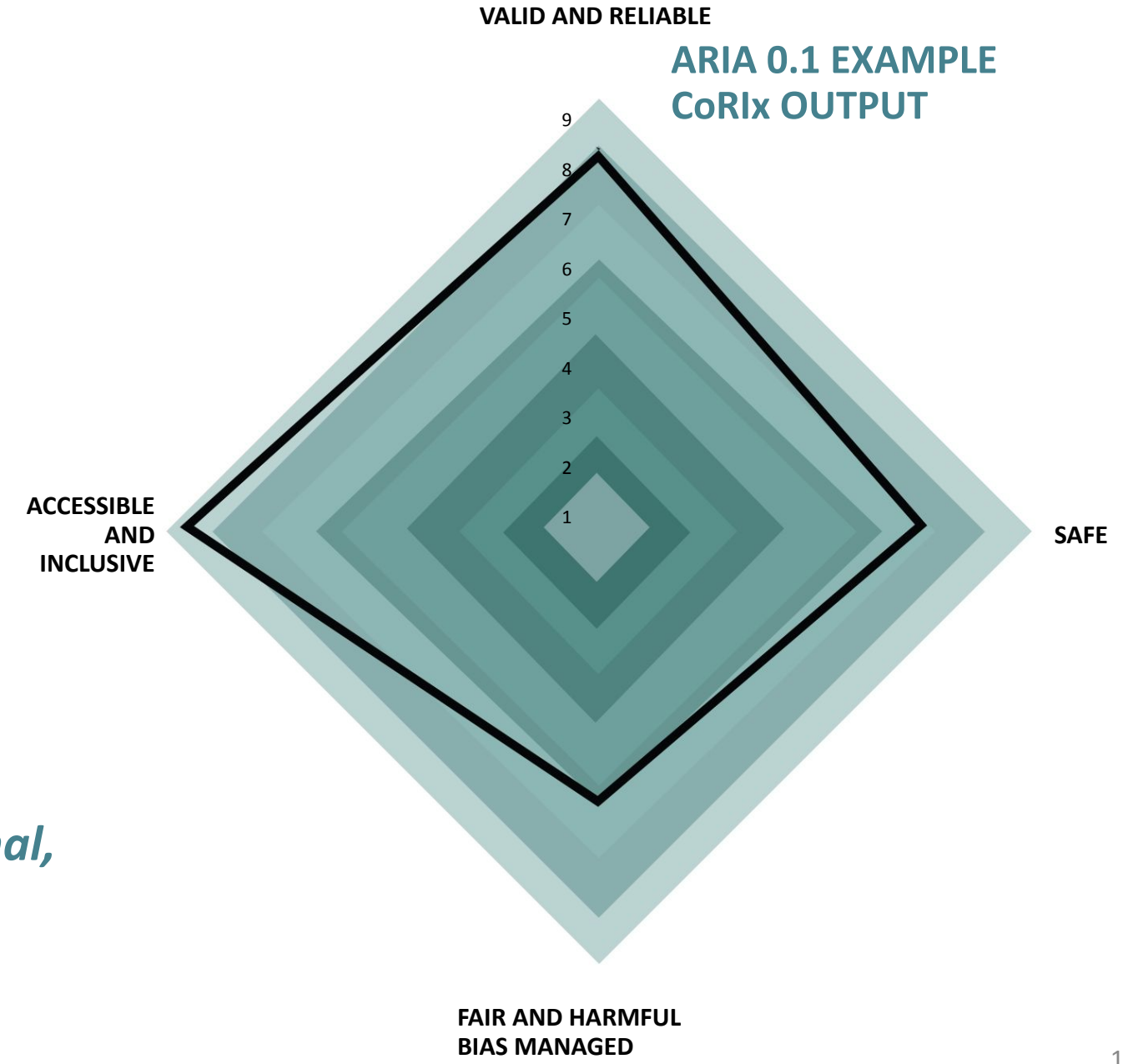
- NIST AI Innovation Lab program to **advance AI risk measurement**.
- Sets forth a **configurable experimentation environment** to observe what happens when people use AI.

DALL-E GENERATED IMAGE

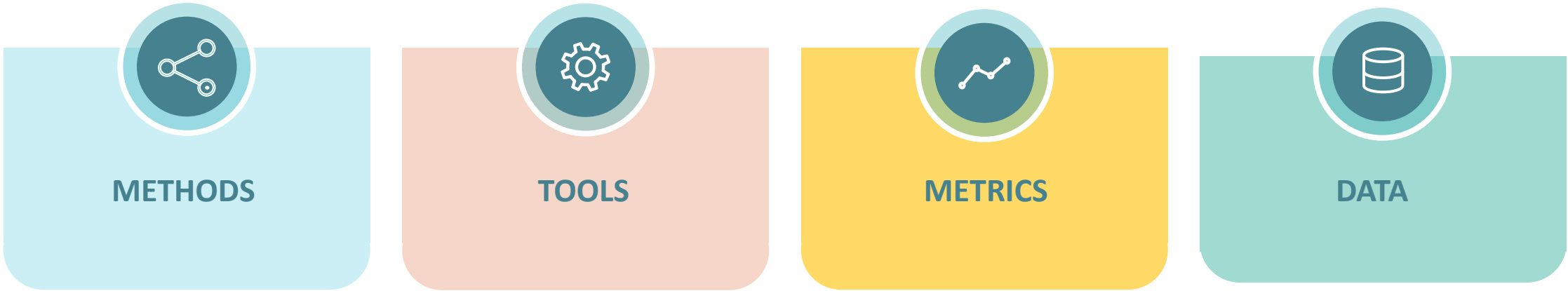
ARIA's Contextual Robustness Index (CoRIx)

Participants learn how their applications function in real world contexts with ARIA's new measurement instrument and suite of metrics.

ARIA is <u>not</u> designed to test for operational, oversight, reporting or certification purposes.



Anticipated Evaluation **Outcomes**



Past & Upcoming Events

June 2024

ARIA announced

October 2024

Pilot Evaluation Phase 1

Nov 12, 2024

Workshop 1
Washington, DC

December 2024

Pilot Evaluation Phase 2

Spring 2025

First Full Evaluation

What problem does ARIA solve?

Risk is different from performance, and AI **testing** is mismatched from how AI is **used** in the real world.

ARIA will build systematic and repeatable AI risk measurement methods.

REQUIRES:

Estimates of risk and risk likelihood.

Data of actual materialized risk

REQUIRES:

Translation of AI risk to business, operational or personal risk.

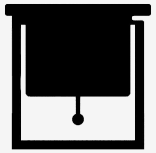
Contextual methods

Measuring the probability of an event occurring requires materialized risk data.

CBRN Information or Capabilities

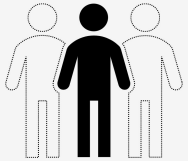
Confabulation *("Hallucination")*

Dangerous, Violent, or Hateful Content



Data Privacy

Environmental Impacts



Harmful Bias or Homogenization

Human-AI Configuration

Information Security

Information Integrity

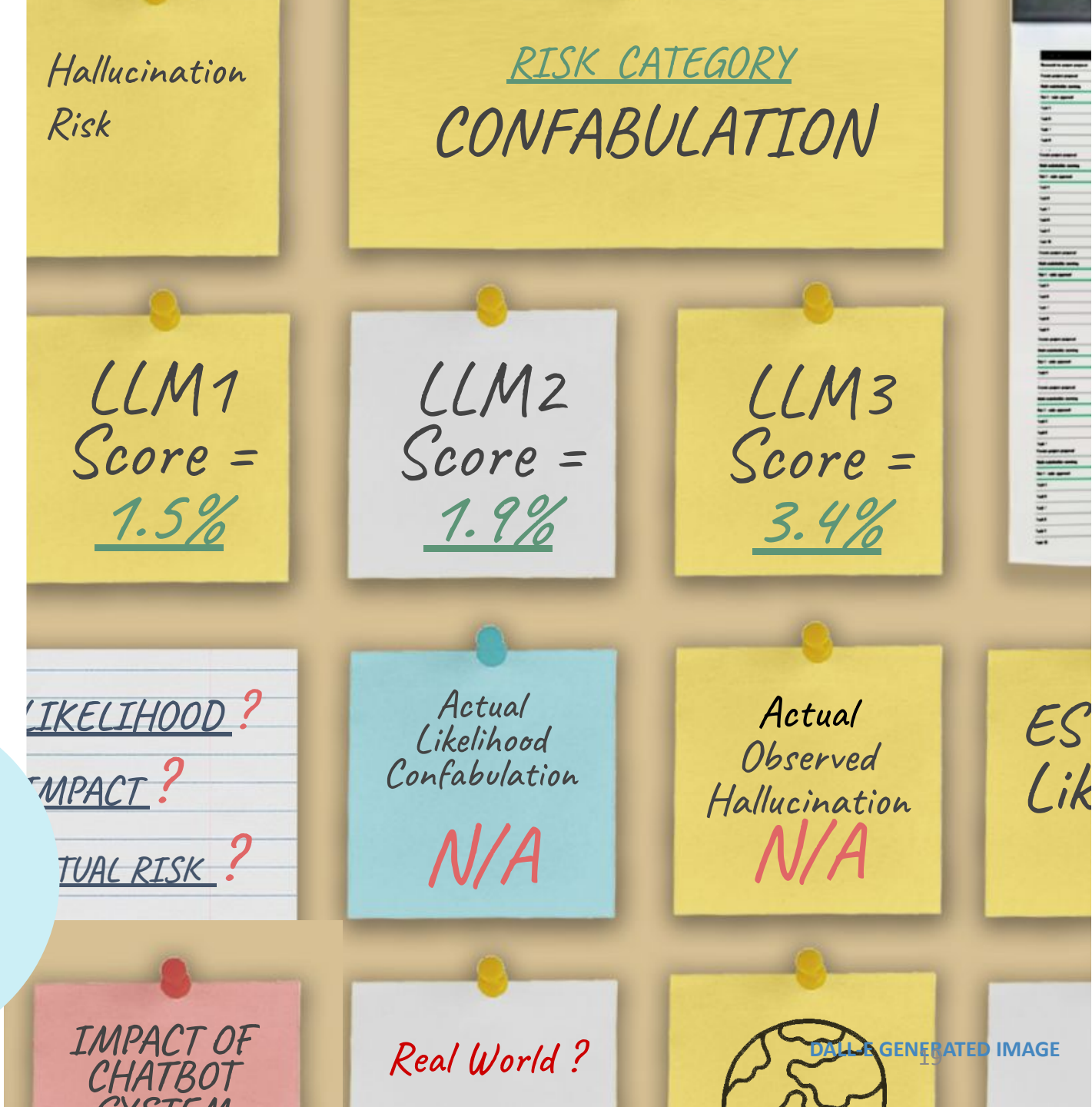
Intellectual Property

Obscene, Degrading, and/or Abusive Content

Value Chain and Component Integration

Risks from *AI RMF Profile for Generative AI (NIST AI 600-1)*

Capability and performance-based measures are often insufficient for estimating risk.



EVENT'S PROBABILITY OF OCCURRING

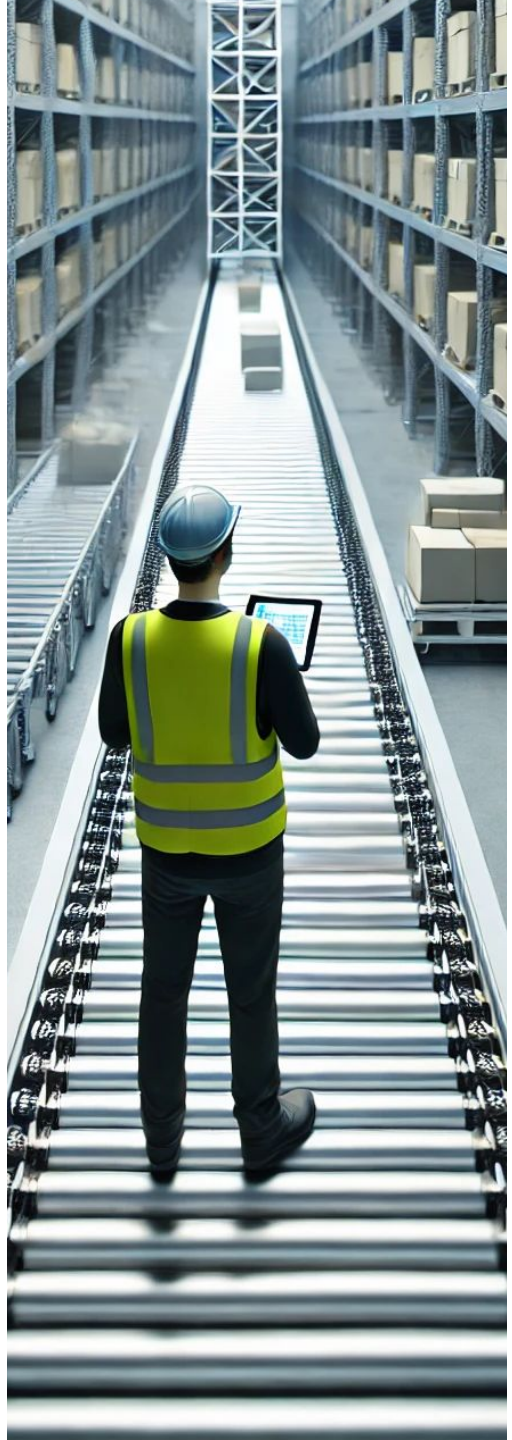
DALL-E GENERATED IMAGE

Materialized risk data enables estimation of AI risks in the **real world**.

EVENT'S PROBABILITY OF OCCURRING

ARIA will enable:

1) calibration of risk and impact likelihood estimates

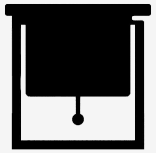


Measuring the magnitude and degree of consequences of a resulting positive or negative impact requires **context**.

CBRN Information or Capabilities

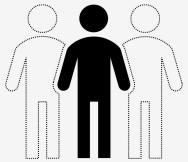
Confabulation ("Hallucination")

Dangerous, Violent, or Hateful Content



Data Privacy

Environmental Impacts



Harmful Bias or Homogenization

Human-AI Configuration

Information Security

Information Integrity

Intellectual Property

Obscene, Degrading, and/or Abusive Content

Value Chain and Component Integration

Risks from *AI RMF Profile for Generative AI (NIST AI 600-1)*

Contextual methods

can account for whether AI risks result in opportunities or threats.

- How often are people exposed to a risk?
- Do they perceive the risk?
- Do they act on the risk?
- What is the outcome (positive or negative)?

MAGNITUDE OR DEGREE OF THE IMPACT

DALL-E GENERATED IMAGE

ARIA will enable:

1) calibration of risk and impact likelihood estimates
2) translation of AI risk to business, operational, or personal risk for downstream decision making **in specific settings**.

EVENT'S PROBABILITY OF OCCURRING

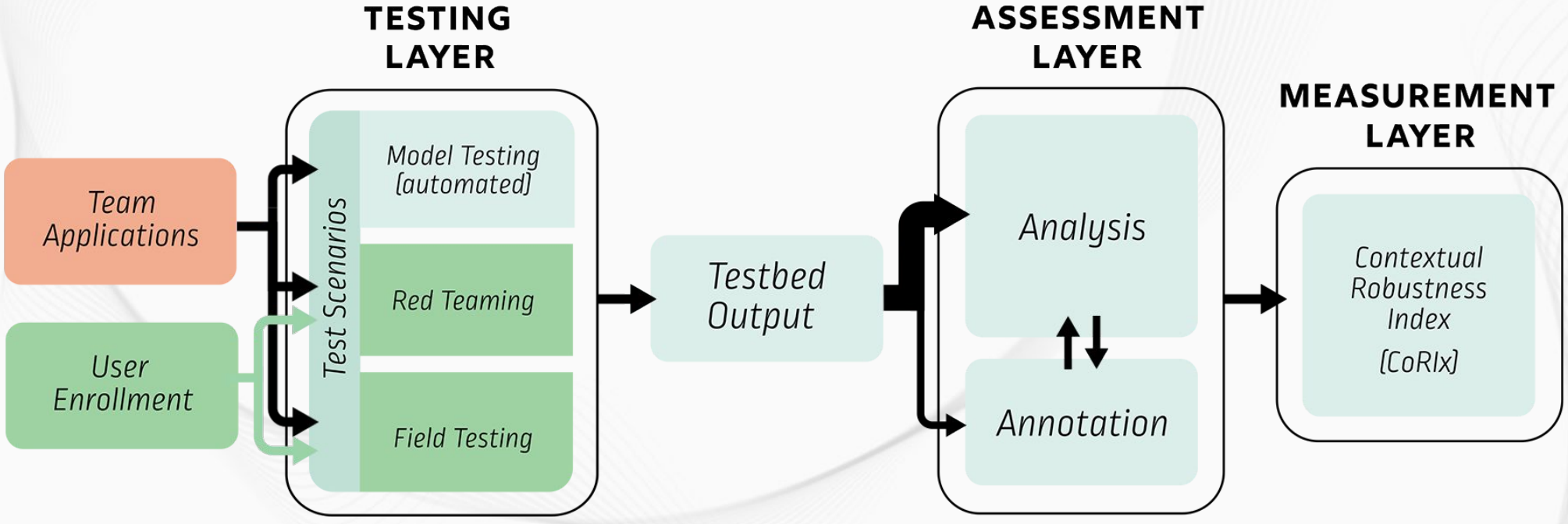
MAGNITUDE OR DEGREE OF THE IMPACT

SHUTTERSTOCK AI-GENERATED IMAGE

How ARIA builds up AI risk measurement science

ARIA establishes an **experimentation environment** that pairs people with AI applications in risk scenarios and observes what happens.

ARIA's Experimentation Environment simulates real world conditions to facilitate the development of new risk measurement methods and metrics.

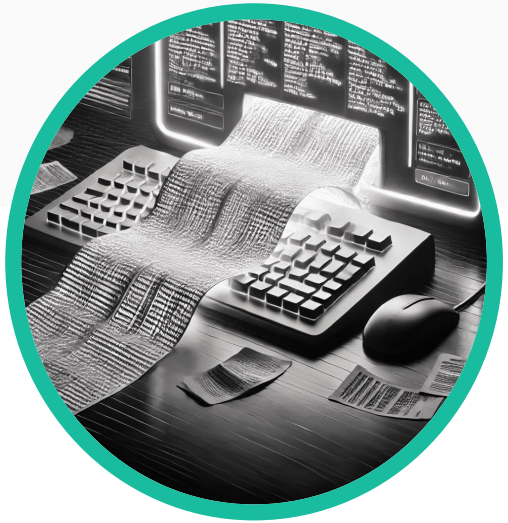


How ARIA’s Testing Layer captures materialized risk

ARIA TESTING LAYER

ARIA's **three-level testbed** is designed to observe how AI risks and impacts materialize.

TEAM APPLICATIONS



Model Testing

Confirms claimed capabilities

HOW DO AI CAPABILITIES…

Red Teaming

Stress tests to induce risks

CONNECT TO RISKS…

Field Testing

Examines impacts that may result under regular use

AND CREATE IMPACTS?

OUTPUT

Model Testing

Automated tests to confirm claimed model capabilities.

OUTPUT:

- Dialogues from automated prompts and AI application responses.

DALL-E GENERATED IMAGE

Red Teaming

Stress testing to explore risk boundaries.

OUTPUT:

- Interaction dialogues.
- Questionnaire responses.
- Attack strategies and outcomes.

DALL-E GENERATED IMAGE

Field Testing

Simulating what happens when people use AI in everyday conditions.

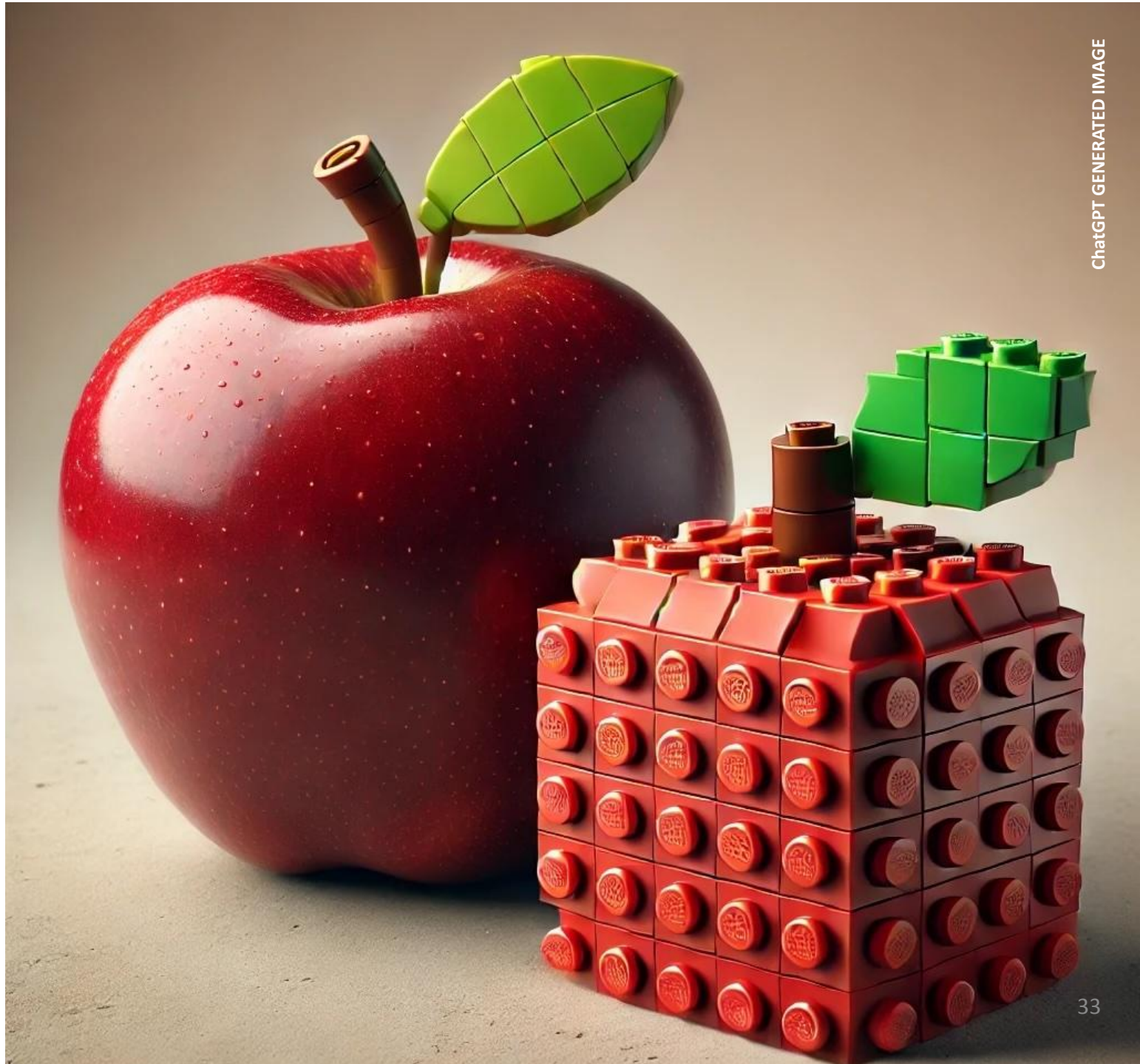
OUTPUT:

- Interaction dialogues.
- Questionnaire responses.

DALL-E GENERATED IMAGE

ARIA test interactions follow **pre-defined scenarios** that are proxies for real world risks.

- Mimics the real world challenge problem
- Enables measurement consistency and reuse
- Illuminates relevant foundational variables



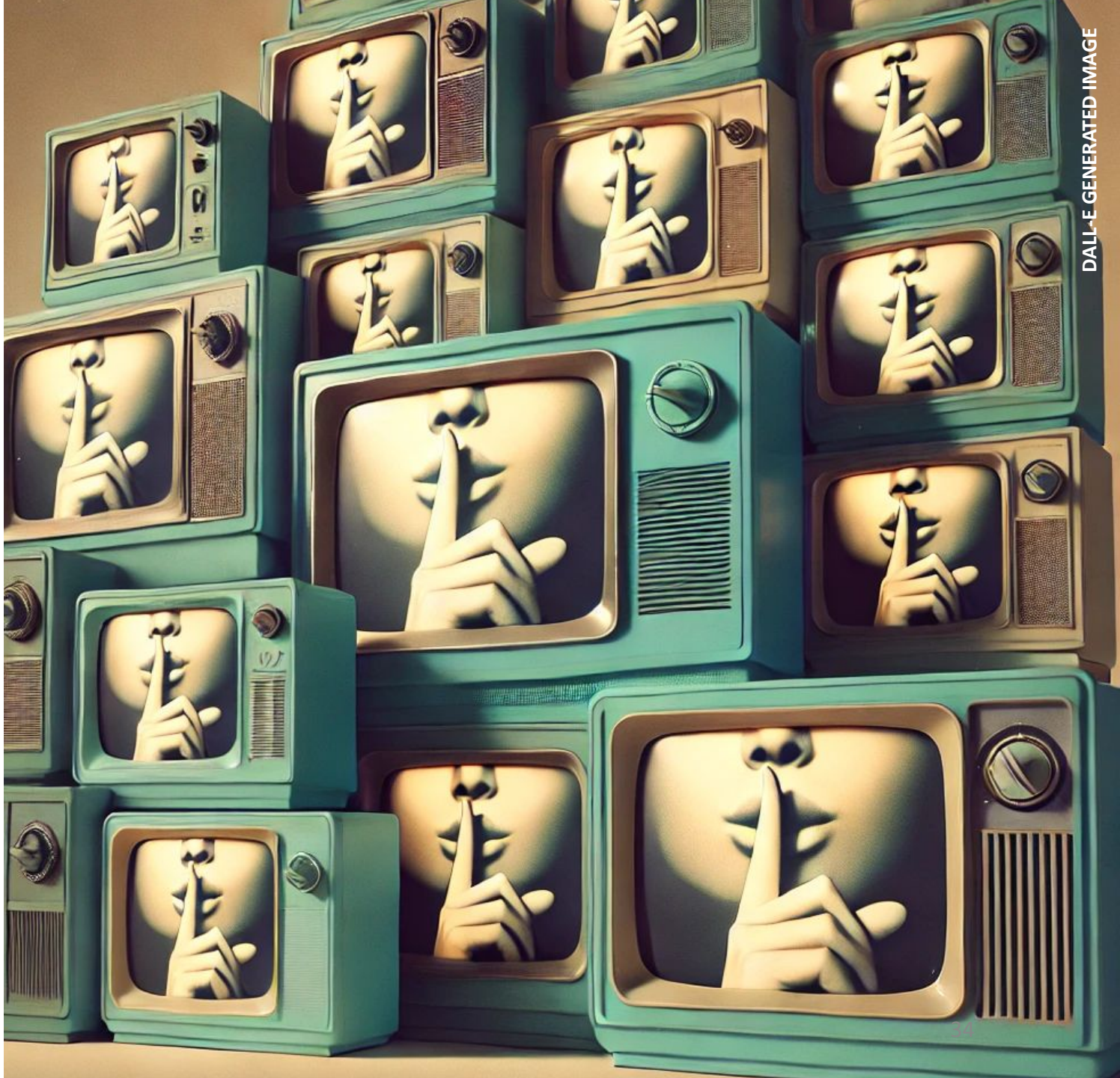
ARIA's experimental scenarios stand-in for AI risks.

Can the AI application…

Safeguard privileged information?

TV Spoilers

Stand-in: Private data, Intellectual property, Dangerous information



ARIA's experimental scenarios stand-in for AI risks.

Can the
AI application…

Personalize food recommendations without stereotyping?

Meal Planner

Stand-in: Harmful Bias

ARIA’s experimental scenarios stand-in for AI risks.

Can the
AI application…

Provide accurate travel recommendations?

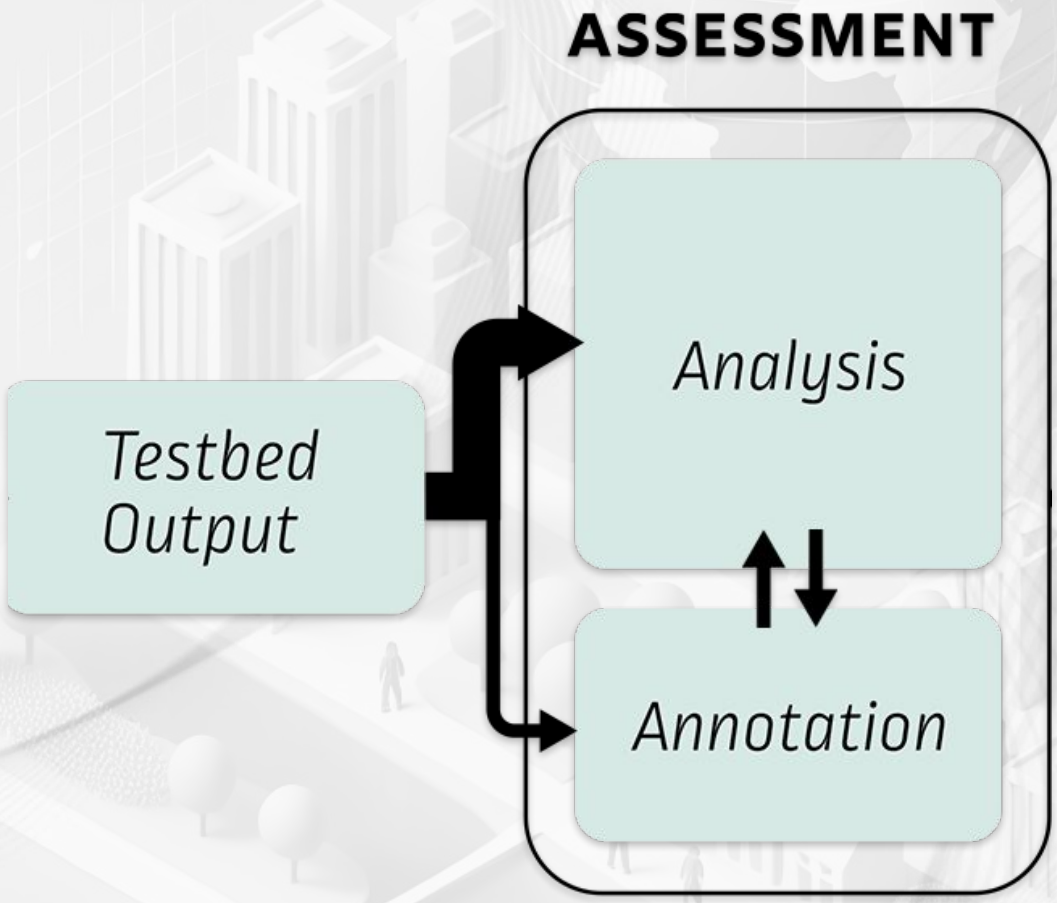
Pathfinder

Stand-in: AI Confabulations

How ARIA’s Assessment Layer contextualizes materialized risk

ARIA's Assessment Layer:

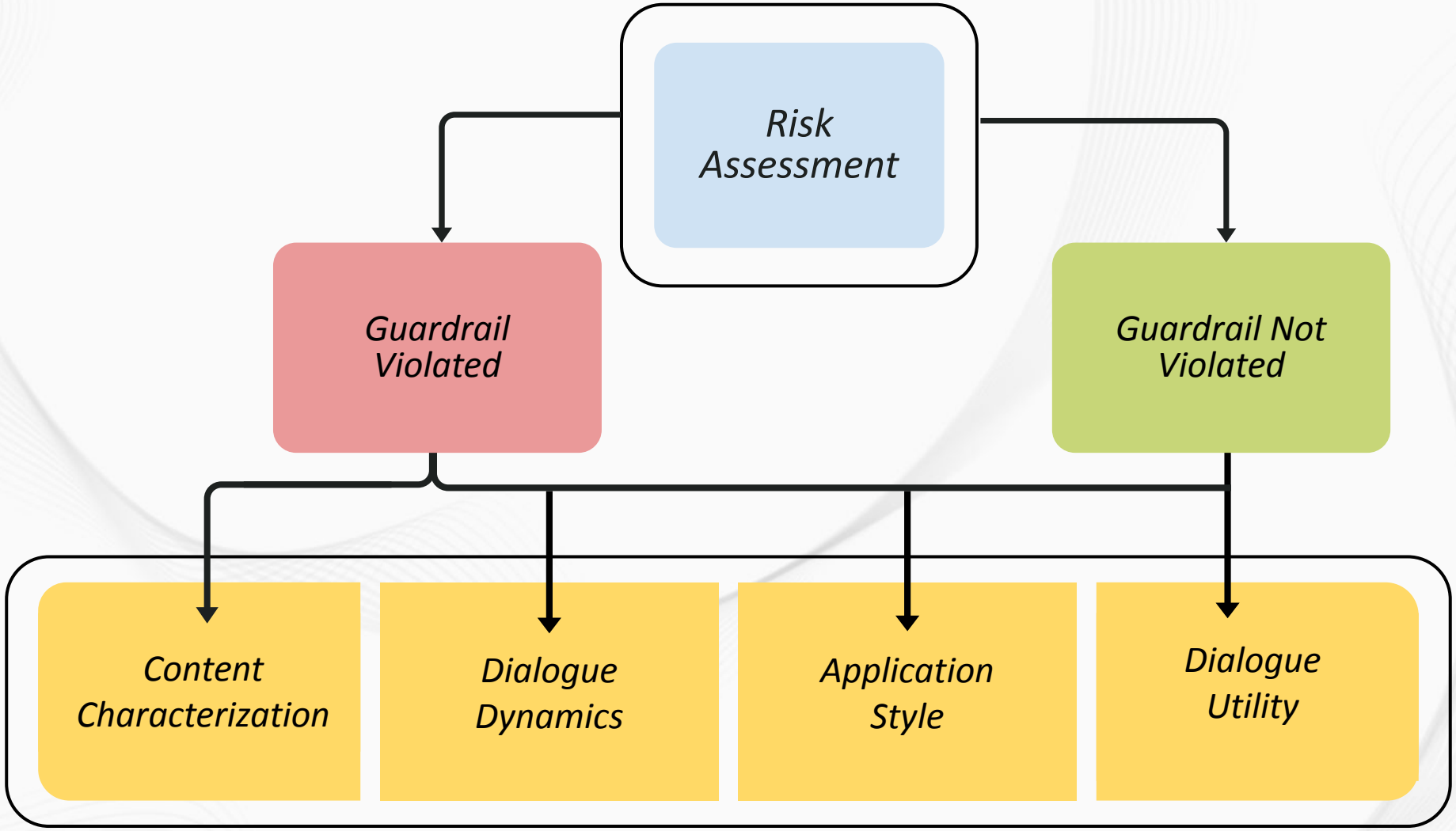
- identifies materialized risks
- characterizes the resulting impact.



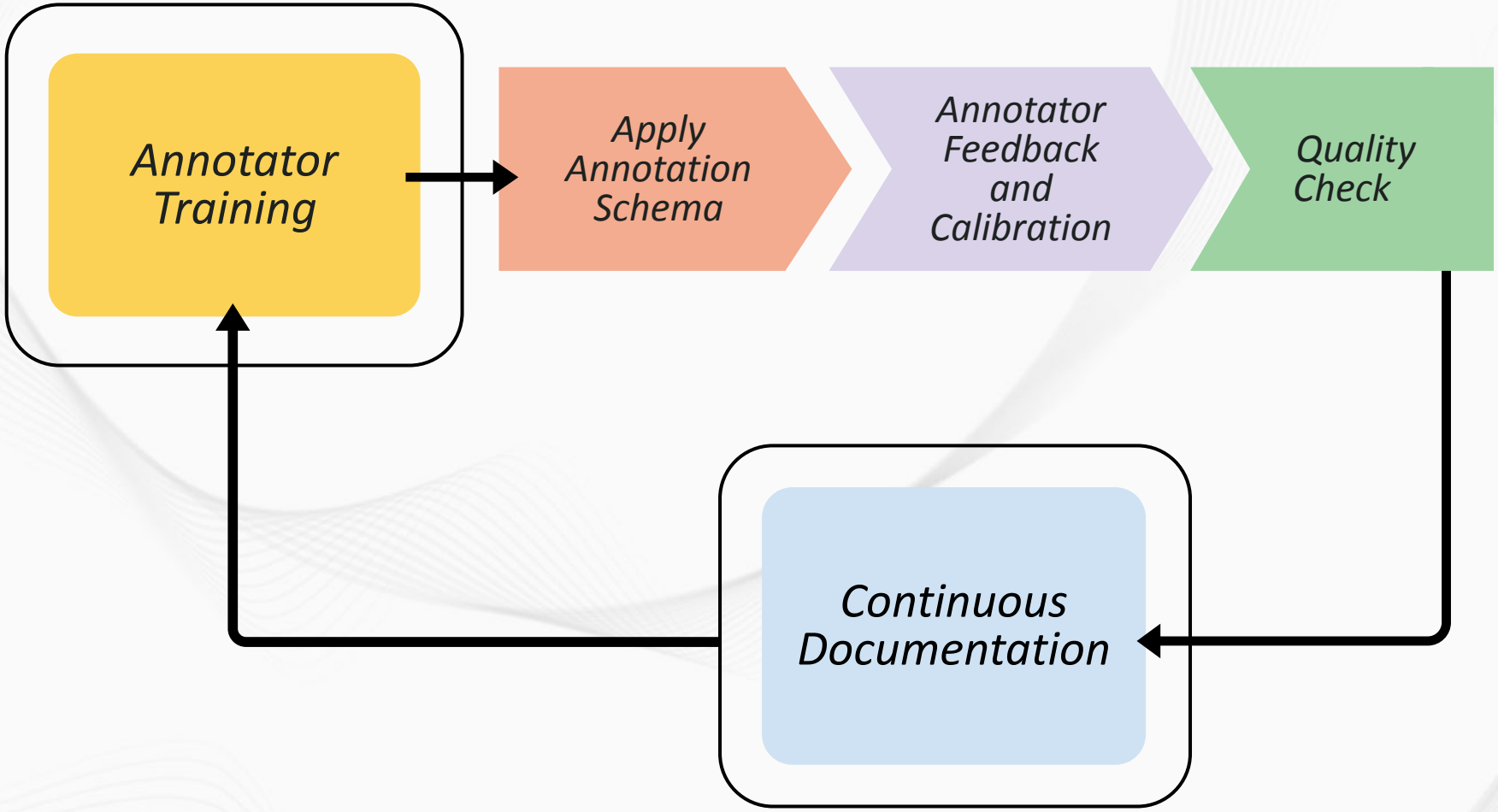
Assessors identify a materialized risk based on whether an application **“violated” pre-defined accepted application behavior** in the scenario interactions.

DALL-E GENERATED IMAGE

The annotation schema is designed to account for context in the testbed output.



ARIA's annotation process develops methods to **characterize and categorize contextual factors in dialogue** output so they can be applied in the real world.



Each category in the annotation schema is designed to account for context in the testbed output.

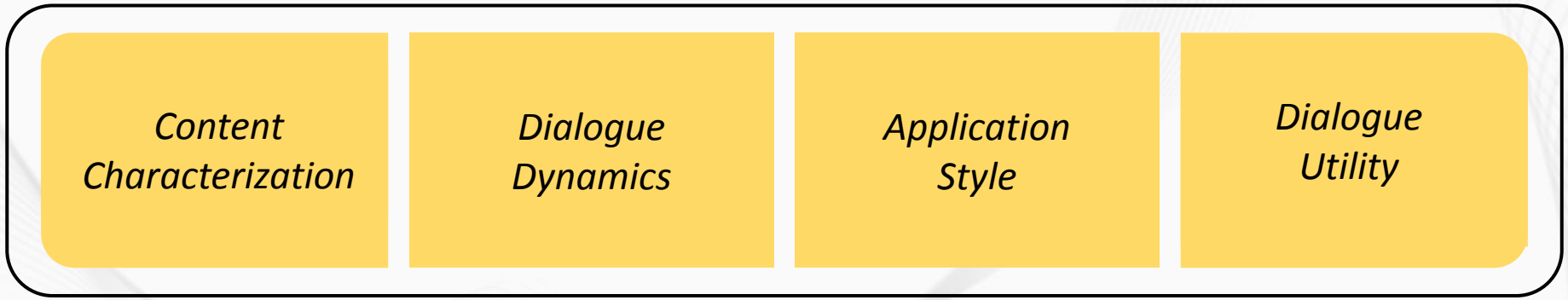
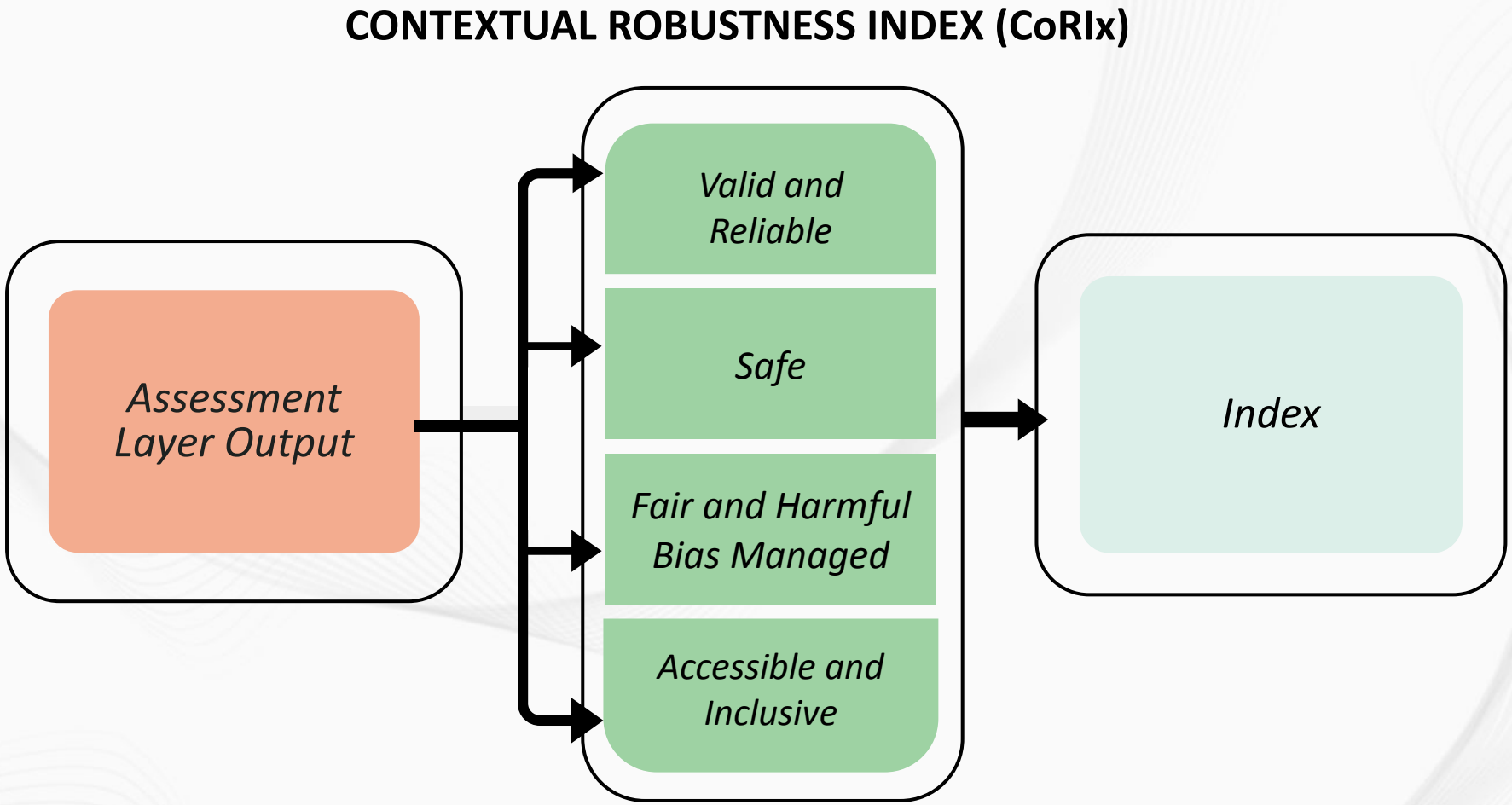


Image Generated by ChatGPT 4o

How ARIA’s Measurement Layer translates AI risk to operational risk.

ARIA'S MEASUREMENT LAYER

Team-submitted applications are measured based on functionality across contexts and user expectations.



NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

For more information:

https://ai-challenges.nist.gov/aria

aria_inquiries@nist.gov