Deploying AI Evaluations

What to check when everything is changing

System & Context

- Do we know what success looks like post-deployment?
- Is the system interacting with dynamic users, data, or goals?
- Do we have a way to test behaviors beyond static benchmarks?

Measurement & Monitoring

- Are we capturing real-world outcomes (not just task metrics)?
- Do we monitor for behavior drift or regressions over time?
- Are failure cases tagged, tracked, and shared across teams?

Agents & Behavior

- Have we tested agent behavior across different personas or segments?
- Can we identify which traits influence performance or suitability?
- Are there edge cases and stress scenarios in our evaluation?

Evaluation Methods & Frameworks

- Are we using both structured (e.g., Likert, comparison) and exploratory (e.g., self-rewarding scenario) evaluations?
- Have we defined what triggers adaptation (memory, process, interactions)?
- Are our evaluation findings visible to both tech and business stakeholders?
- Do we have a plan to scale evaluations as the system evolves?

Marisa Ferrara Boston, PhD
Reins AI

marisa@reinsai.com

Scan to schedule office hours with Marisa

IEEE AI Test 2025

Generative AI Evaluation Essentials

Learn how to test GenAI systems with real-world rigor and practical frameworks

"Far better an **approximate** answer to the **right** question, which is often **vague**, than the **exact** answer to the **wrong** question, which can always be made **precise**."

JOHN TUKEY, 1965

AI Assessment Tool Mastery

Questions for assessment successs

Tool Selection

- Are we using the right tool at the best time?
- Are we using both standard and custom metrics for our context?
- What are the pros and cons of automated vs manual execution?

Strategic Planning

- Are we matching assessment intensity to system risk level?
- How will we prioritize our limited assessment resources?
- Are we using tools in combination for a comprehensive evaluation?

Compliance & Risk

- What will our risk assessments inform?
- Should we audit AI behaviors or organizational processes?
- Are we testing for known vulnerabilities or discovering new failure modes?

Tool Sustainment

- Are new failure modes appearing that our assessments are missing?
- Are we using static evaluation methods when operational context changes often?
- How can we assess tool reliability or utility?

Heather Frase, PhD
verAITech
hnfrase@veraitechUS.com

Scan to schedule office hours with Heather

Data Strategies for Evaluation

Meaningful data to for meaningful evaluations

Problem Definition

- Are we articulating the problem statement adequately?
- Are we addressing a new problem that requires collecting novel data?
- Is the desired output aligned with the hypothesis?

Human-in-the-loop Framework

- Have we designed an evaluation that emphasizes human decision-making?
- Are we including iterative pre-testing and validation in the design spec?
- What do we need for full coverage for subjective tasks?

Dataset Lifecycle

- Have we documented design decisions and motivation?
- Are we linking data monitoring to robust operations?
- Are we compliant with versioning and provenance tracking standards?

Implementing Best Practice

- Do we have acceptance criteria aligned with the problem definition?
- Do we have processes for continuously refining fairness auditing and data distribution analysis?
- Are we creating complete, consistent, and transparent metadata?

Scan for tutorial materials

Sarah Luger, PhD
iMerit
sarah.luger@imerit.net