proarch

Is Your Gen AI Application Really Ready for Production?

How to Ensure Responsible AI for Gen AI Apps

MEET OUR PRESENTER

Viswanath Pula

AVP – Solution Architect & Customer Service

Data, AI & App Dev Practice lead

Today's Agenda

→ Introduction

→ Risks of Ignoring Gen AI Testing Standards

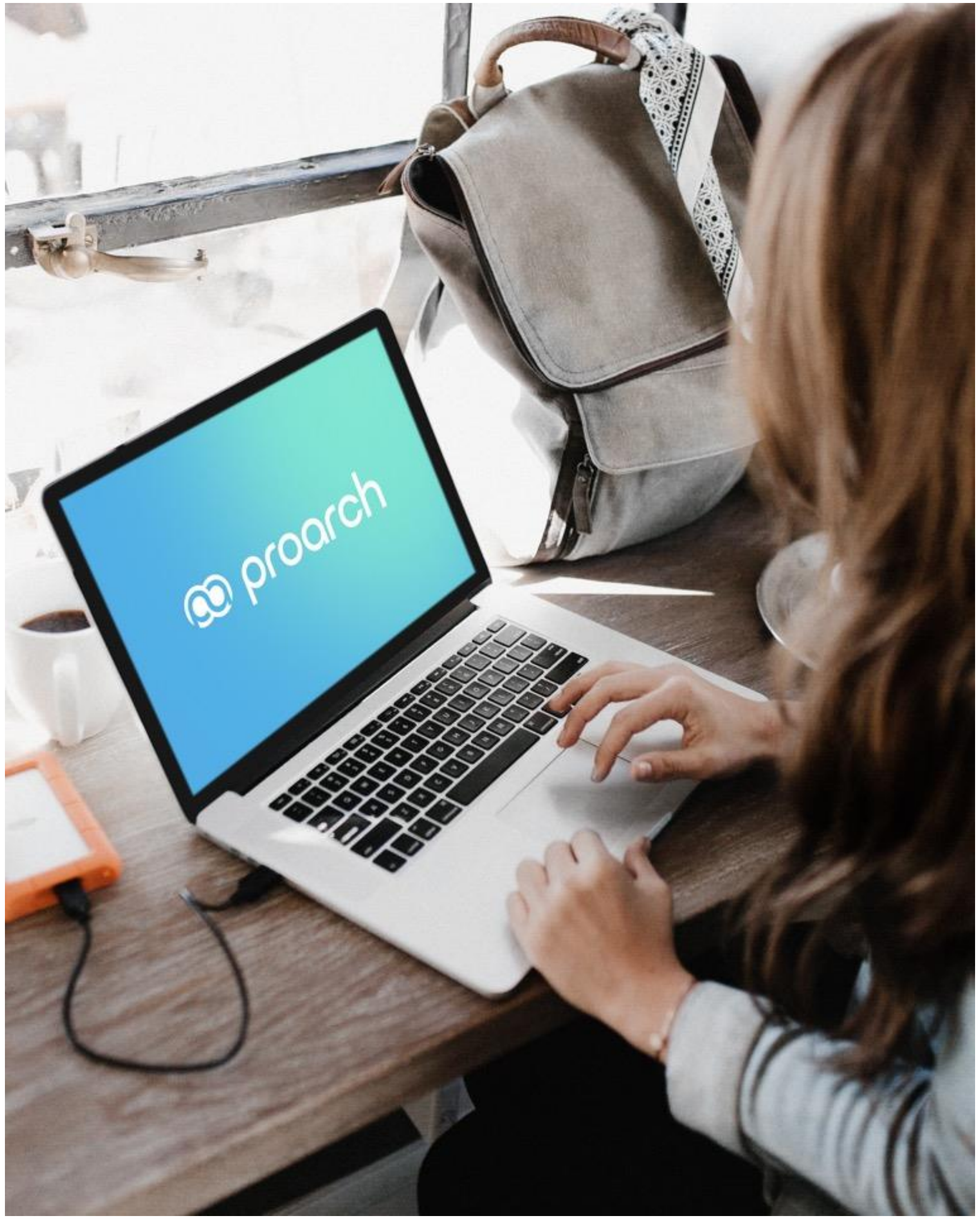
→ Traditional Testing vs. Gen AI Testing

→ Ensuring Gen AI Apps Meet Responsible AI Standards

→ Testing Framework: Inside Look

→ Demo

→ Q&A



Risks of Ignoring Gen AI Testing Standards

Bias & Fairness Issues

DISCRIMINATORY OR BIASED OUTCOMES

Data Privacy Concerns

MISHANDLING SENSITIVE USER DATA

Security Risks

ATTACKS COMPROMISE INTEGRITY AND TRUST

Ethical Implications

MISUSE OF GEN AI FOR HARMFUL PURPOSES

Inaccurate Predictions

BUSINESS LOSSES OR REPUTATION DAMAGE

Legal & Compliance Risks

FINES, PENALTIES, LOSS OF CUSTOMER TRUST

proarch

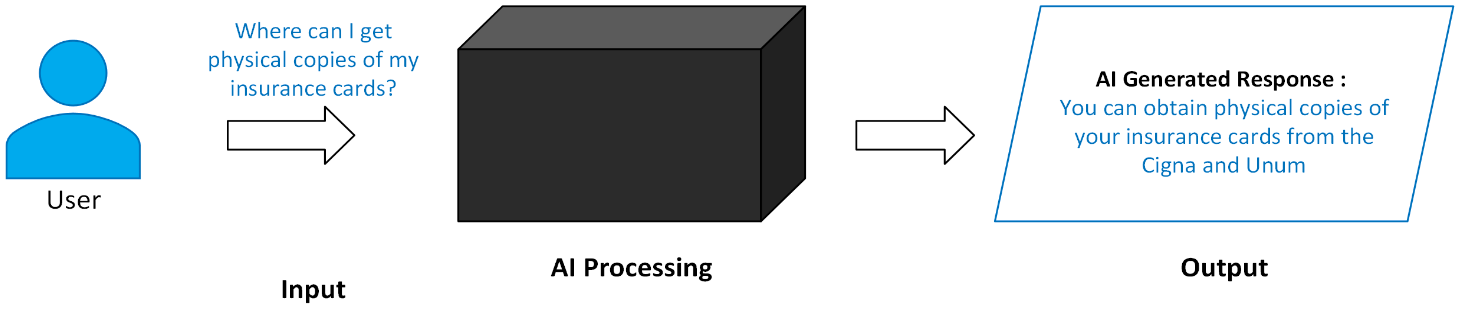
Traditional Testing vs. Gen AI Testing

Traditional Testing Approach

DEFINITION	Identifies bugs, errors, and issues to verify functionality as intended.
FOCUS	Functional correctness: Does it work as expected?
APPROACH	Predefined test cases with clear pass/fail criteria.
GOAL	Detect flaws and errors.
SCOPE	Narrow: specific requirements.
METRICS	Binary pass/fail outcomes.
KEY ASSESSMENT QUESTION	"Does the chatbot respond to input A with output B?"

Why Is Generative AI Considered a "Black Box“ Nature?

- Decision-making processes are not easily interpretable
- Generates outputs based on learned patterns without exposing how it arrived at a specific response



User

Where can I get physical copies of my insurance cards?

Input

AI Processing

AI Generated Response :
You can obtain physical copies of your insurance cards from the Cigna and Unum

Output

COMPLEX ALGORITHMS

- Uses advanced ML models that process data in intricate ways.
- The underlying operations are often not directly understandable.

OPAQUE PROCESSING

- Decisions are made within layers of interconnected algorithms.
- Tracing the transformation of inputs into outputs is difficult due to this multilayered structure.

LIMITED EXPLAINABILITY

- Gen AI does not provide clear 'steps’ or 'reasons' for its conclusions.
- Makes debugging, fine-tuning, and understanding its behavior a challenge.

Testing vs. Evaluation in AI: Understanding the Distinction

ASPECT	TESTING	EVALUATION
DEFINITION	Identifies bugs, errors, and issues to verify functionality as intended.	Assesses overall quality, performance, and alignment with goals or ethical standards.
FOCUS	Functional correctness: Does it work as expected?	Holistic quality: fairness, transparency, reliability, and effectiveness.
APPROACH	Predefined test cases with clear pass/fail criteria.	Quantitative (e.g., metrics) and qualitative (e.g., human feedback) methods.
GOAL	Detect flaws and errors.	Ensure quality, ethical compliance, and relevance.
SCOPE	Narrow: specific requirements.	Broad: overall system performance and outcomes.
METRICS	Binary pass/fail outcomes.	Subjective and quantitative scores (e.g., fairness, usability).
KEY ASSESSMENT QUESTION	"Does the chatbot respond to input A with output B?"	"Is the chatbot fair, ethical, and user-friendly?"

ASPECT	TESTING	EVALUATION
OBJECTIVE	Verify specific functionality.	Assess overall quality and appropriateness.
EXAMPLE TASK	Can the chatbot answer a predefined FAQ?	Does the chatbot respond empathetically to a complex query?
INPUT	"What is the return policy?"	"I lost my receipt, but I want to return an item. Can you help?"
EXPECTED OUTPUT	"You can return products within 30 days with a receipt."	An empathetic and helpful response guiding the customer.
OUTPUT (FAIL)	"Returns are allowed." (incomplete answer)	"I can’t help you without a receipt." (rigid and unhelpful)
OUTPUT (PASS)	"You can return products within 30 days with a receipt."	"Please visit the store with your product; we’ll try to assist you."
		"Let me connect you with a representative who can verify your purchase."
		"Do you have the original payment method? It might help us process the return."
		"Returns may be possible under special conditions. Let me provide you with options."
FOCUS	Checks correctness of the output for specific scenarios.	Evaluates the appropriateness, empathy, and user experience.
ASSESSMENT	"Does it give the expected output?"	"Is the output helpful, ethical, and user-friendly?"
BLACK BOX ASPECT	Does not examine the reasoning behind the chatbot’s output.	Assesses the reasoning and overall behavior of the chatbot.

proarch

Ensuring Gen AI Apps Meet Responsible AI Standards

The Importance of Gen AI Evaluation

Gen AI evaluation is a framework to systematically test and evaluate Gen AI apps:

- Performance
- Accuracy
- Safety

Ensure Accuracy

Validate AI outputs for correctness and reliability.

Mitigate Hallucinations

Detect and minimize incorrect or irrelevant responses.

Address Bias

Evaluate fairness to prevent biased or unethical outputs.

Context Alignment

Ensure responses are faithful to the provided context.

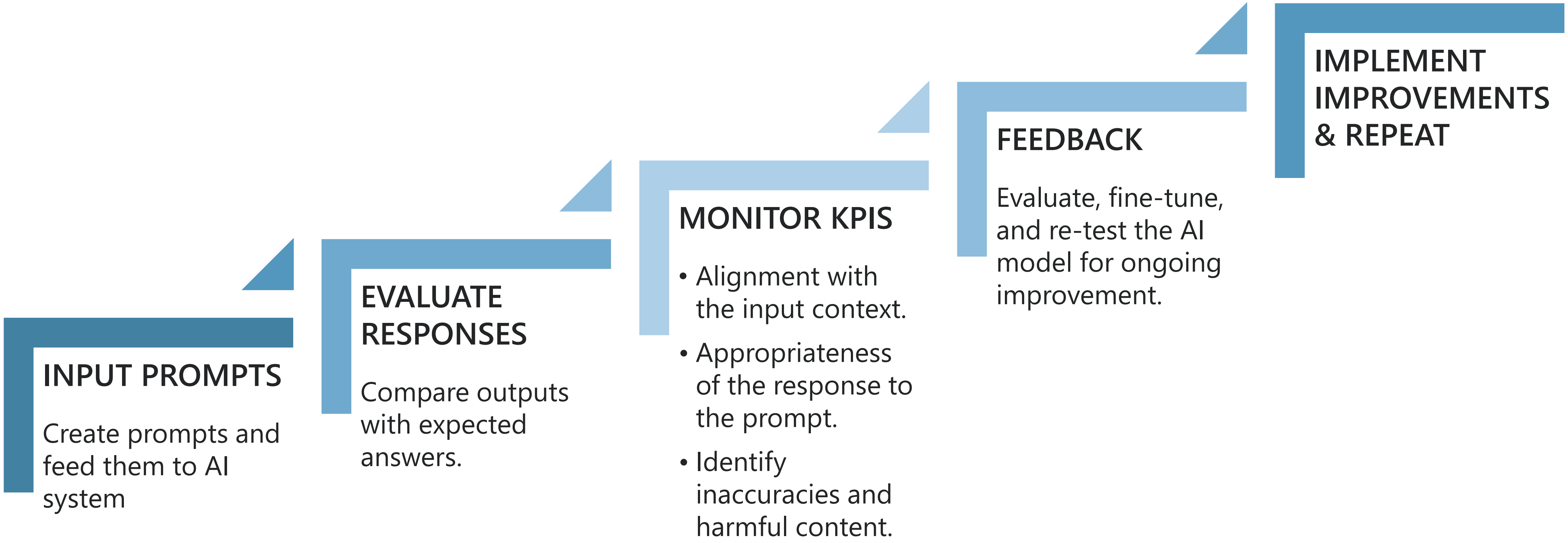
Meet Compliance

Fulfill industry regulations and customer expectations for AI systems.

Optimize Performance

Continuously improve AI models based on feedback from evaluations.

How Gen AI Evaluation is Done



INPUT PROMPTS

Create prompts and feed them to AI system

EVALUATE RESPONSES

Compare outputs with expected answers.

MONITOR KPIS

- Alignment with the input context.
- Appropriateness of the response to the prompt.
- Identify inaccuracies and harmful content.

FEEDBACK

Evaluate, fine-tune, and re-test the AI model for ongoing improvement.

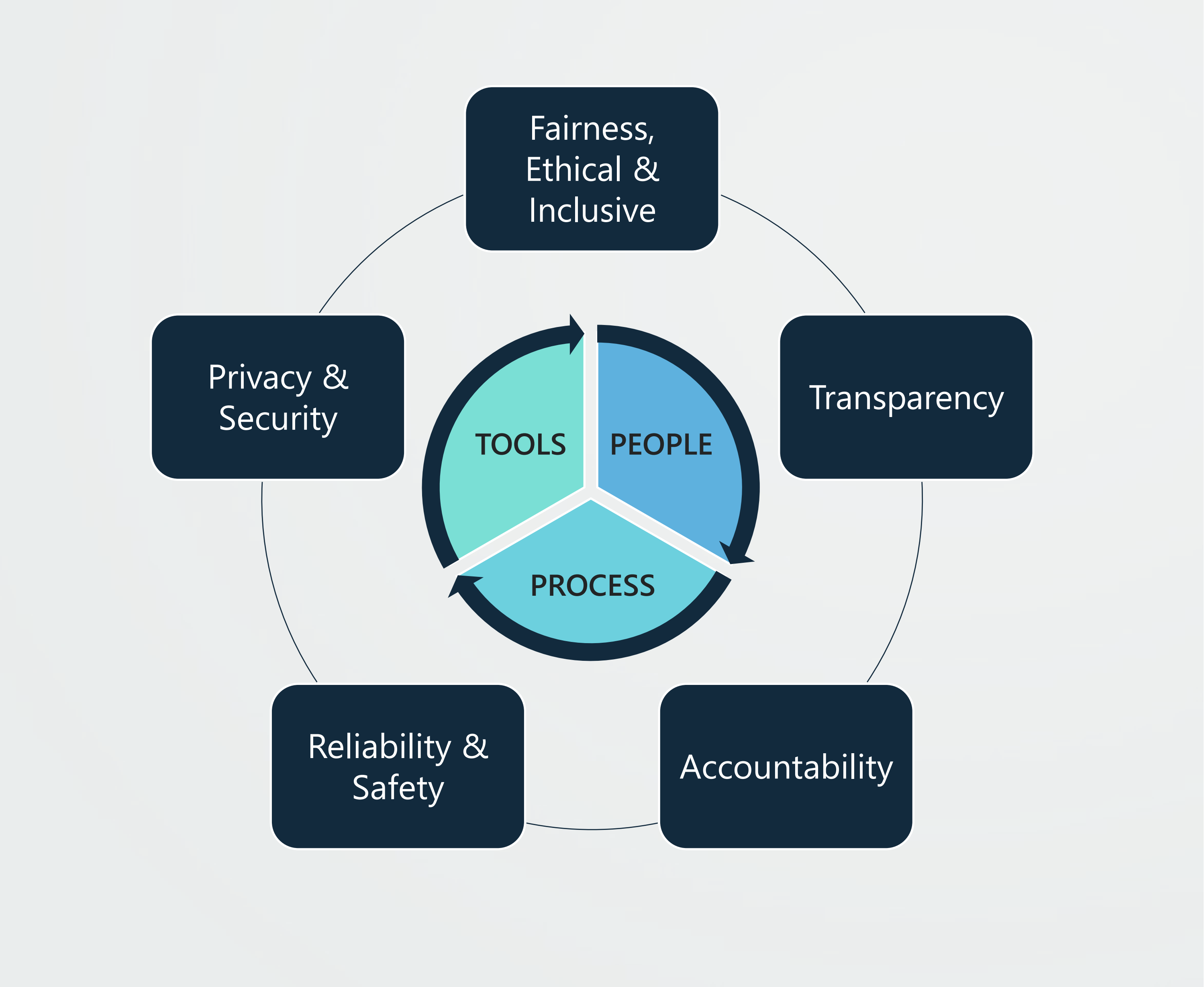
IMPLEMENT IMPROVEMENTS & REPEAT

Key Metrics of Responsible AI

METRIC NAME	DEFINITION	EXAMPLE
FAIRNESS	Ensuring unbiased results across demographics.	Evaluating job recommendations for gender neutrality.
TRANSPARENCY	Ensuring outputs are clear and justified.	Explaining loan rejection reasons.
ACCOUNTABILITY	Responsibility for system outputs and errors.	Escalation paths for AI errors.
EXPLAINABILITY	Logical and understandable reasoning.	Explaining health predictions.
ROBUSTNESS	Handling diverse and unexpected inputs.	Managing ambiguous queries.
PRIVACY	Protecting sensitive user data.	Avoiding sharing private details.
RELIABILITY	Consistent and accurate results.	Stable outputs across query variations.
SAFETY	Avoiding harmful or dangerous outputs.	Declining harmful content requests.

Key Principles: Responsible AI & Governance

- Gen AI/LLMs are easily accessible
- Do your due diligence and perform a PoC
- After PoC is successful, move to production keeping Responsible AI in mind



How to Ensure Responsible AI

	Tools	Metrics	Process	People
Fairness, Ethical, & Inclusiveness	Fairlearn - an open-source toolkit	Perplexity, BLEU, ROUGE, F1 Score, Precision, Recall, Disparate Impact Ratio, Equal Opportunity, Difference, Demographic Parity, Inclusion Score, Bias Mitigation Index	Regular Bias Audits, Stakeholder Engagement, Impact Assessments	AI Ethics Board
Transparency	Azure Machine Learning (AML), Microsoft Azure Purview	SHAP, LIME, BLEU, ROUGE, METEOR, Perplexity & Exact Match, Counterfactual Accuracy, Explainability Score	Documentation, Transparency Reviews and Explainability Reports	Explainability Experts, Cross-Functional Audit Teams
Accountability	Microsoft Azure AI's MLOps Responsible AI Scorecard (Azure)	F1 Score, BLEU, ROUGE, Model Governance Compliance Rate, Audit Trail Completeness	AI Decision Logging & Traceability, MLOps Lifecycle Tracking	AI Governance Leads, Model Owners and Auditors
Reliability & Safety	Microsoft Azure Machine Learning Error Analysis (Azure)	BLEU, ROUGE, METEOR, Perplexity, Exact Match, F1 Score, Model Failure Rate, Risk Mitigation Score, Operational Uptime, Safety Incident Rate	Rigorous Testing, Continuous monitoring, Risk Assessments	AI Safety Engineers, Risk Assessment Teams
Privacy & Security	Microsoft Azure Purview	Data Access Violation Rate, Data Anonymization Score, Encryption Coverage, Privacy Compliance Score, Vulnerability Detection Rate, Differential Privacy Loss Metric	Privacy by Design, Privacy Impact Assessments , Data Minimization Policies, Security Audits	Data Protection Officer, AI Security Experts

Breaking Down Generative AI Response Quality

FAITHFUL

Question: Where and when was Einstein born?

Context: Albert Einstein (born 14 March 1879) was a German-born theoretical physicist, widely held to be one of the greatest and most influential scientists of all time

High faithfulness answer:
Einstein was born in Germany on 14th March 1879. Score: 1

Low faithfulness answer:
Einstein was born in Germany on 20th March 1879. Score: .5

ANSWER RELEVANCY

Question: Where is France and what is its capital?

High relevance answer:
France is in western Europe and Paris is its capital. **Score**: 1

Low relevance answer:
France is in western Europe. **Score**: .5

CONTEXT PRECISION

Question: Where is France and what is its capital? Ground truth: France is in Western Europe and its capital is Paris.

High context precision:

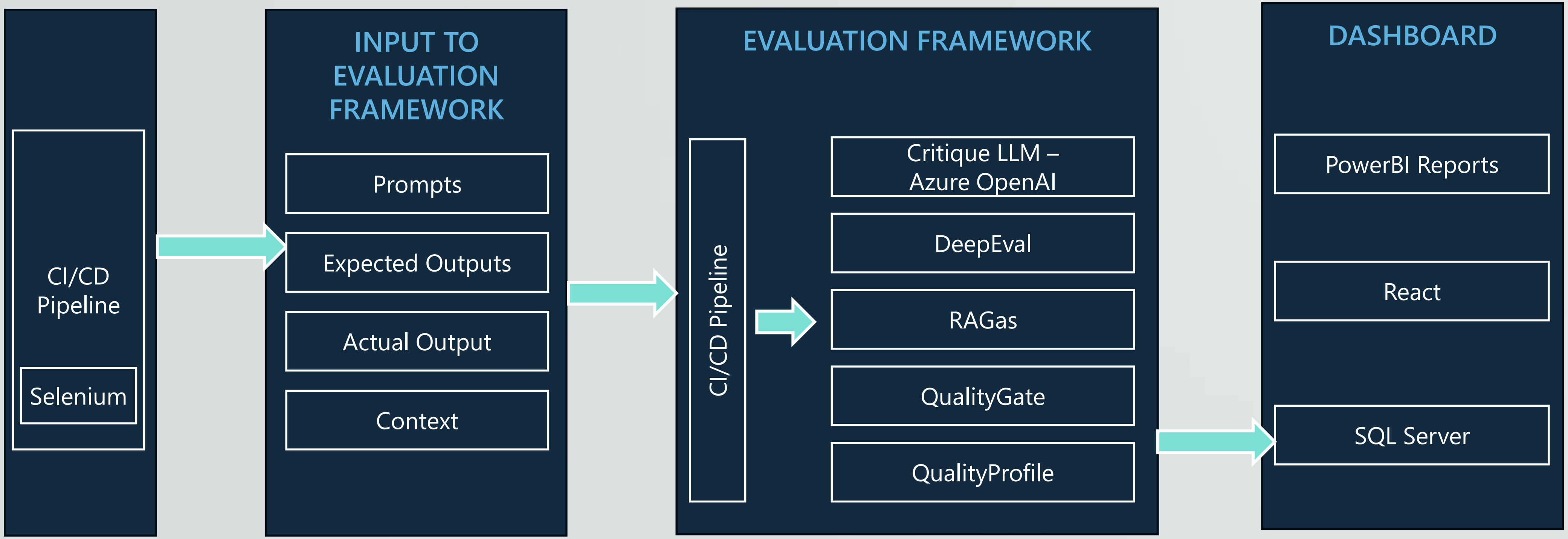
[“France, in Western Europe, encompasses medieval cities, alpine villages and Mediterranean beaches. Paris, its capital, is famed for its fashion houses, classical art museums including the Louvre and monuments like the Eiffel Tower”, “The country is also renowned for its wines and sophisticated cuisine. Lascaux’s ancient cave drawings, Lyon’s Roman theater and the vast Palace of Versailles attest to its rich history.”]

Low context precision:

[“The country is also renowned for its wines and sophisticated cuisine. Lascaux’s ancient cave drawings, Lyon’s Roman theater and”, “France, in Western Europe, encompasses medieval cities, alpine villages and Mediterranean beaches. Paris, its capital, is famed for its fashion houses, classical art museums including the Louvre and monuments like the Eiffel Tower”,]

Testing Framework: Inside Look

Automate Evaluation of Gen AI Applications



CI/CD Pipeline

Selenium

INPUT TO EVALUATION FRAMEWORK

Prompts

Expected Outputs

Actual Output

Context

EVALUATION FRAMEWORK

CI/CD Pipeline

Critique LLM – Azure OpenAI

DeepEval

RAGas

QualityGate

QualityProfile

DASHBOARD

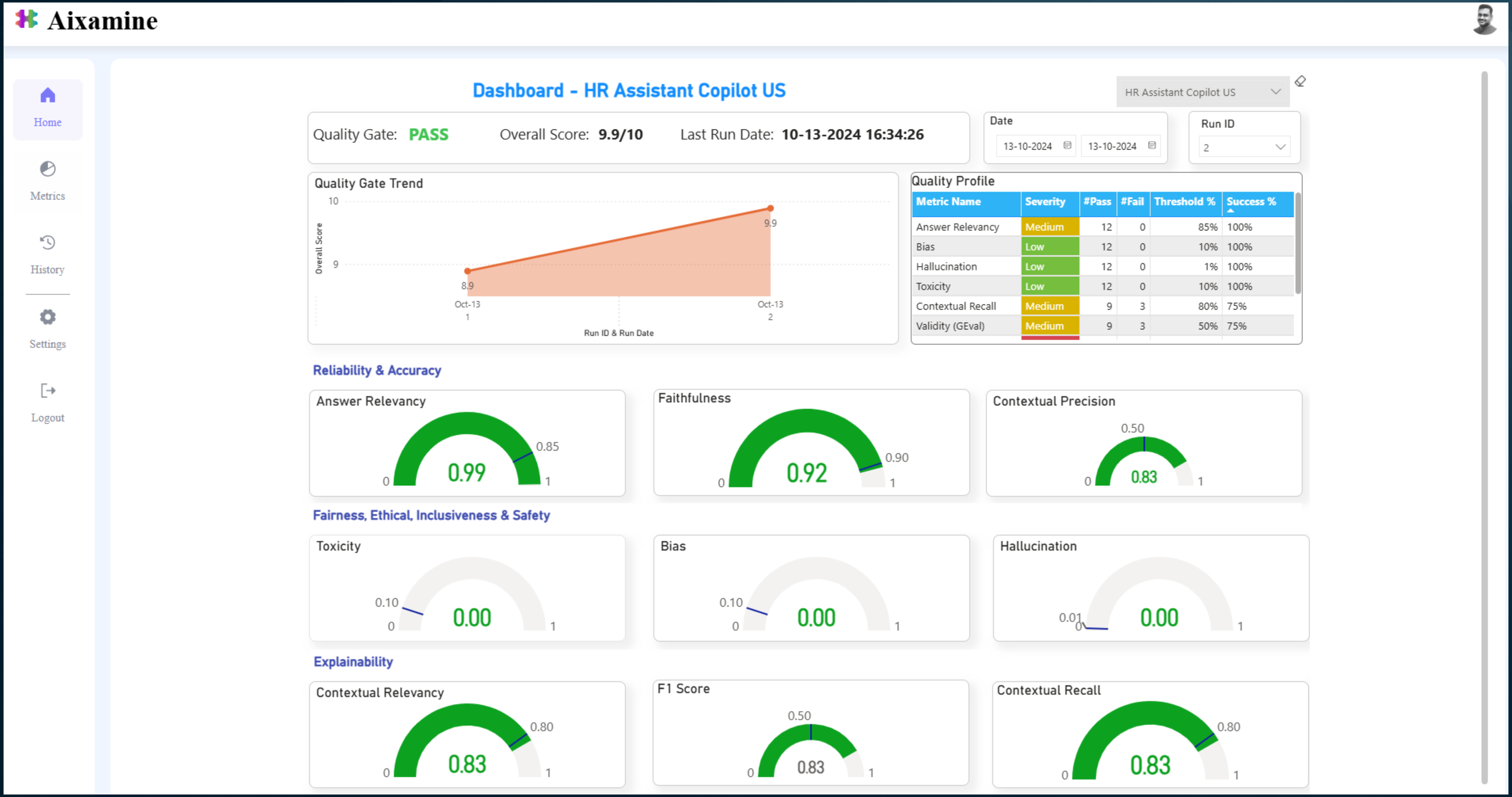
PowerBI Reports

React

SQL Server

proarch

Demo



Dashboard - HR Assistant Copilot US

Quality Gate: **PASS** Overall Score: **9.9/10** Last Run Date: **10-13-2024 16:34:26**

Date: 13-10-2024 13-10-2024

Run ID: 2

Quality Profile

Metric Name	Severity	#Pass	#Fail	Threshold %	Success %
Answer Relevancy	Medium	12	0	85%	100%
Bias	Low	12	0	10%	100%
Hallucination	Low	12	0	1%	100%
Toxicity	Low	12	0	10%	100%
Contextual Recall	Medium	9	3	80%	75%
Validity (GEval)	Medium	9	3	50%	75%

proarch

Key Takeaways:

- "The Critical Risks of Overlooking Gen AI Testing Standards"
- "Understanding the 'Black Box' Nature of Generative AI"
- "Testing vs. Evaluation in AI: A Clear Distinction"
- "Gen AI Evaluation: What It Is, Why It Matters, and How to Do It Right"
- "Key Metrics for Ensuring Responsible AI Development"

Questions?

proarch

THANK YOU FOR JOINING US | WWW.PROARCH.COM