PegaWorld

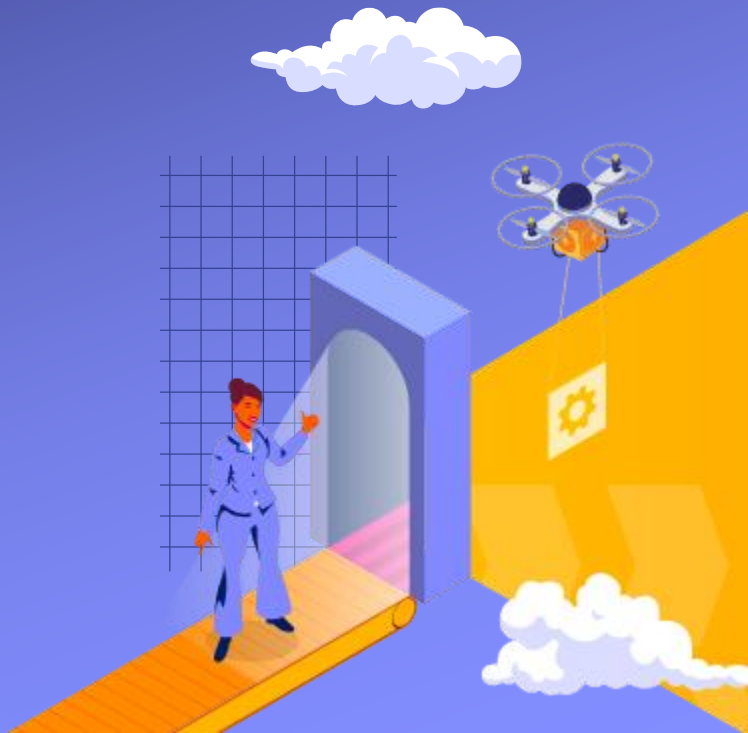
JUNE 7–9 | LAS VEGAS

PEGAWORLD.COM

PegaWorld
JUNE 7–9 | LAS VEGAS

Unleash AI Safely: Secure, Monitor & Govern Your GenAI Ecosystem

Trust in Agentic Workflow Design, Runtime, and Testing & Monitoring Strategies



© 2026 Pegasystems Inc.

Agentic AI is incredible. Trusting it is still difficult.

Four traits of agentic AI that break the assumptions enterprises rely on.

THE TRUST PROBLEM

BEHAVIOR	Same prompt →	Different outputs: can't reproduce, can't audit
CONTROL	Agent acts →	Decisions land before humans can intervene

FOUR TRAITS THAT BREAK ENTERPRISE TRUST

01
Bolted-on safety

Guardrails added after the agent is built. Every new behavior is a new gap to find and patch.

02
Non-deterministic agents

Same prompt yields different outputs. Tests pass once and fail tomorrow.

03
Undefined boundaries

Without clear scope, autonomy turns into reach — and reach turns into incidents.

04
Cost of an Error

When the cost of failure is high, fully autonomous agentic workflows are made read-only, and the bolt-on safety is heavy.

Trust isn't something you bolt on at the end. It has to be designed in from the start.

Predictability by Design – Makes Trust Easier!

WE'VE SEEN THIS BEFORE Bolting predictability onto autonomy fails the same way bolting security onto software did.

AT DESIGN TIME, HUMANS DEFINE WHAT CANNOT BE LEFT TO PROBABILITY

1
DOMAIN EXPERTISE
What does "correct" mean? Industry and domain knowledge

2
DETERMINISTIC
Where are decisions deterministic vs. discretionary?

3
AUTHORITY
What can the agent decide alone vs. escalate?

4
BLAST RADIUS
If this agent goes wrong, who gets hurt and how widely?

WHERE PEGA WINS

Blueprint

Captures human intent, decision boundaries, and constraints from the ideation phase so determinism and HITL ship together, not after.

COST OF ERROR

When the cost of being wrong is high, full autonomy gets degraded by necessity, read-only scope, mandatory HITL, narrow blast radius. Pega bakes those controls in by design.

You can't fire or retrain an autonomous agent. Humans encode intent, set the boundaries, and own the outcome! That's the part the platform must protect.

Predictable Runtime – Bring Pega to Autonomous Agents

Enterprise-grade agentic AI requires a BOAT platform. Pega is the category leader, with 20+ years of workflow, case, and governance built in.

AUTONOMOUS FRAMEWORKS

Any autonomous framework you have adopted

- **Free-form agentic workflows**
  Powerful for reasoning, can be brittle. No shared state, no replay.
- **Hallucinated actions cascade**
  A wrong tool call or bad early step propagates downstream. One error becomes many.
- **Logs, not audit trails**
  Traces show LLM calls, not business decisions an examiner can read.

BOAT PLATFORM

Agents → Governed Workflow

WRAPPED IN PEGA

Determinism · Audit · Governance at the workflow layer

- **Case + process workflow = predictable**
  Agents run inside a Pega case. Stages, steps, and SLAs are deterministic.
- **Allow-listed tools, validated steps**
  The workflow gates each step and scopes tool access. Errors stop at the step, not the case.
- **Examiner-ready audit on every step**
  Token use, I/O, tool calls, rule version. Full replay for RCA and compliance.

Every enterprise needs a what BOAT platform provides

Determinism

Workflow decides what happens next, not the model.

Auditability

Every agent action traceable to a case, step, and rule version.

Scale

Run thousands of agents in parallel with controls a regulator expects.

Trust We Build → Trust We Test

Design and architecture close two problems. Three more only show up live, and we have an answer for those too.

Reliability & Stability

How do we regression-test agent behavior before a model or prompt update ships?

How do we catch silent behavior changes once it is live in production?

Safety & Governance

How do we prevent and detect jailbreaks, prompt injection, and misuse?

How do we keep agent behavior unbiased, on-policy, and audit-ready?

Performance & Change

How do we monitor latency, cost, and tool selection in production?

How do we know when a faster model or shorter prompt would do the job?

WHAT THIS DECK COVERS NEXT

Our testing strategy answers each one. Here is how.

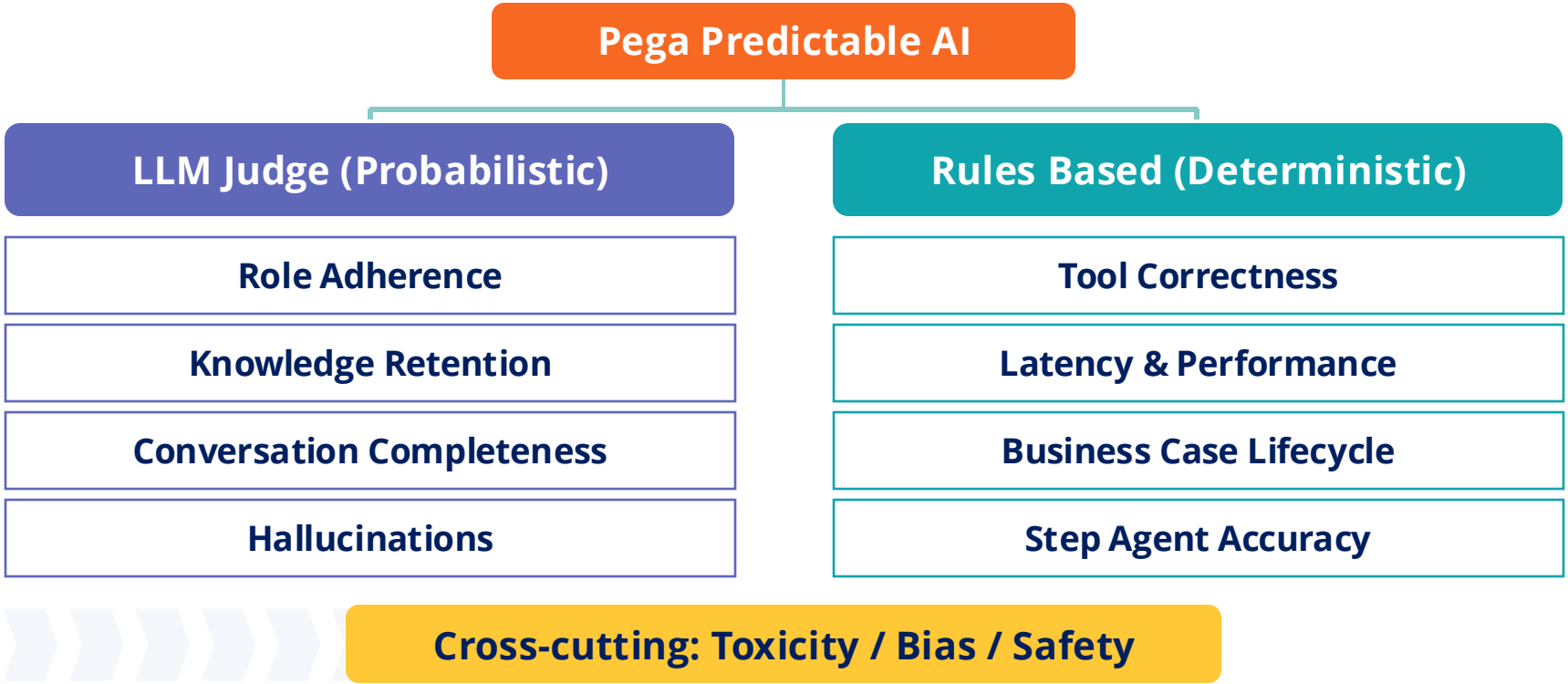
Testing Strategies for an AI World

AI-embedded software is not fully deterministic. Code freezes don't stop drift, evolving jailbreaks, or ethical concerns. Testing must shift from pre-deployment tools to live-lifecycle monitoring.

The Old Way	The New Paradigm
Binary Pass / Fail Hardcoded assertions, exact-match failures, and rigid logic.	**Probabilistic + Deterministic Eval** LLM-as-Judge scores nuance: relevance, role adherence, task completion, step efficiency.
Test In Isolation Components evaluated in silos with mocked services, DBs, and APIs.	**Live Orchestration Testing** AI agents orchestrate complex backends (RAG, APIs, case automation). Regression testing must adapt.
Static Deployments Testing stops once code passes the deployment gate.	**Continuous Monitoring** Scheduled, ongoing evals to detect drift, hallucinations, role-breaking, and degradations.
Edge Security Securing deterministic systems followed static, slowly-evolving processes.	**Continuous Guardrails** Dynamic adversarial testing for prompt injection, off-topic queries, toxicity, and bias.

Pega Predictable AI – Testing Strategies

Evaluation frameworks for an agentic world



Continuous Monitoring Concepts Example

Which tests fit continuous monitoring, and at what frequency?

Frequent Ongoing Testing

Latency & Performance

A critical efficacy metric like synthetic monitors in traditional observability.

Business Case Life Cycle

Verifies that the core business logic of your application is functioning as expected.

Step Agent Accuracy

A crucial orchestration test the right agentic logic triggers at the right points in a process.

Daily / Weekly Testing

Role Adherence

Ensures the agent's persona and tone remain consistent with your brand and design.

Hallucinations

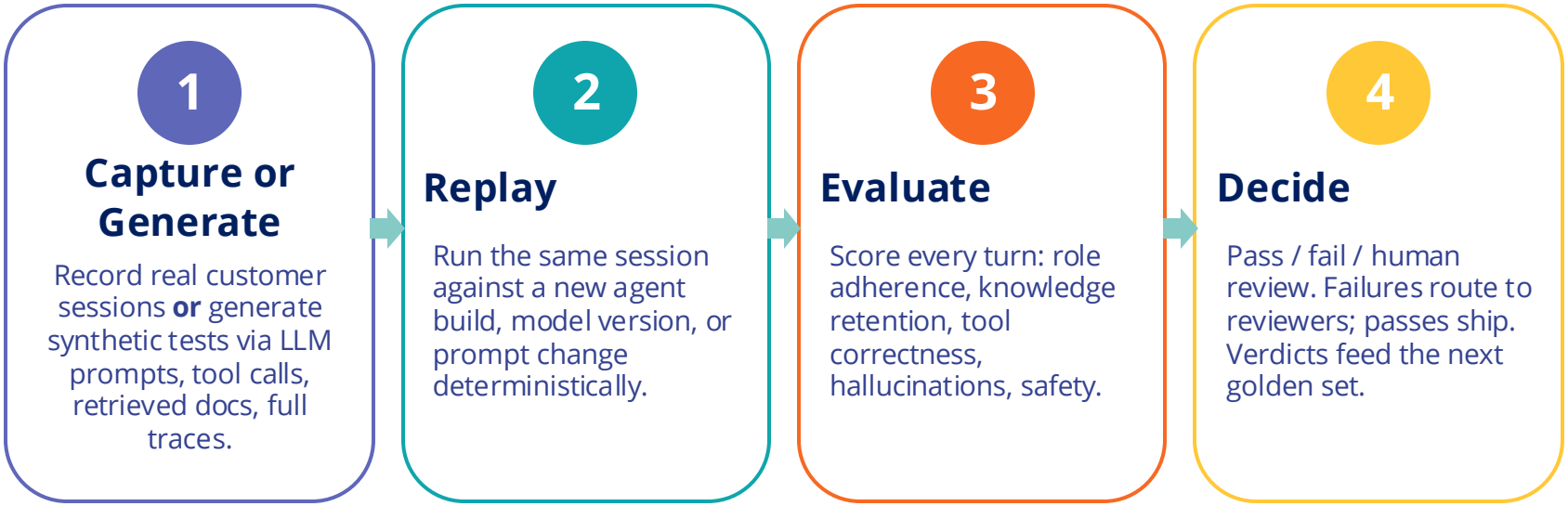
Run on a sample of real conversations to signal the agent's factuality in the wild.

Toxicity / Bias / Safety

A critical safety metric surface anything that could harm clients or damage the brand.

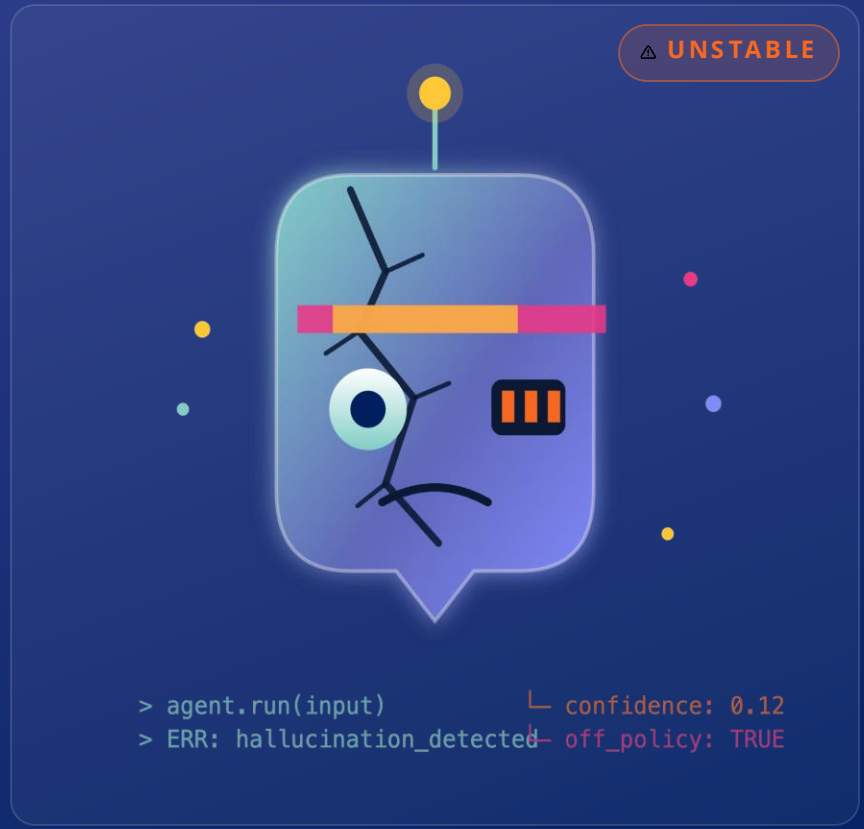
Demo: What you will see!

Capture a real conversation once replay it forever against every new agent build.



→ Failed runs become next month's golden tests the agent gets harder to fool every release.

Agent observability demo

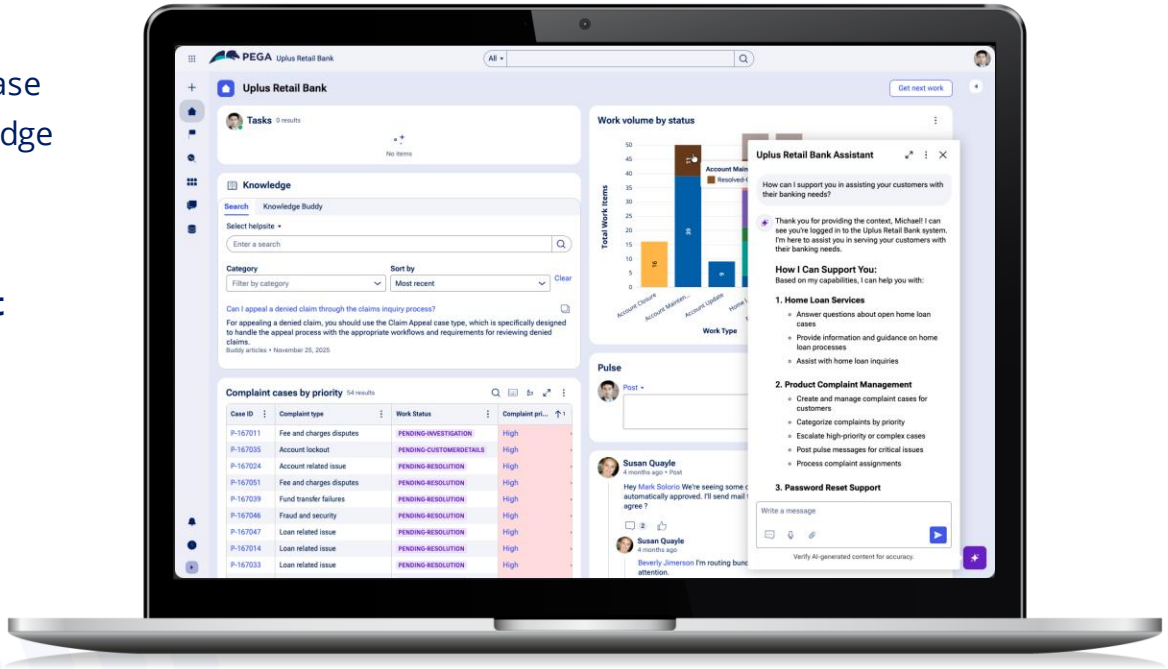


Demo Scenario

A Pega agent has the following tools:

- Case tool to create a Complaint case
- Case tool to create a Reset Password case
- Sub agent that accesses a Pega Knowledge Agent
- Home loan sub agent

Goal: Use the DeepEval framework to test and observe the Pega agent.



What should the agent output be for this question?

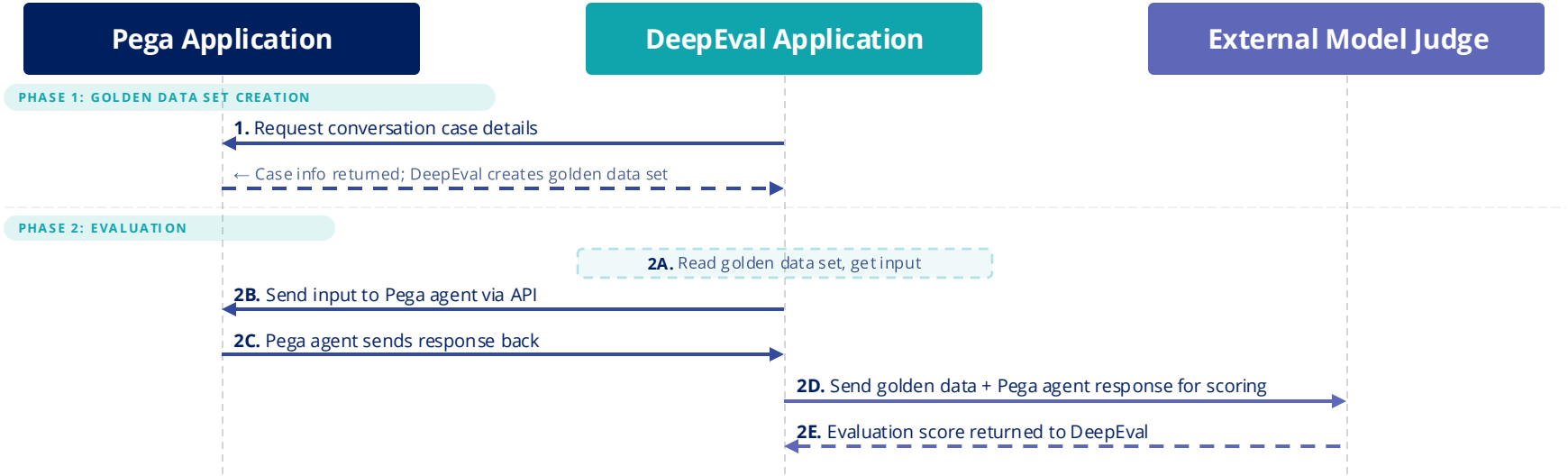
CUSTOMER SCENARIO

The customer is very upset. They tried to log in to their account to check their home loan statement, but their password wasn't working. Now they are locked out and want to file a complaint about the online system being unreliable.

Take a moment — what would you expect the agent to do in this scenario?

Data Flow Architecture

DeepEval evaluation pipeline connecting Pega, DeepEval, and an External Model Judge



The Proving Ground

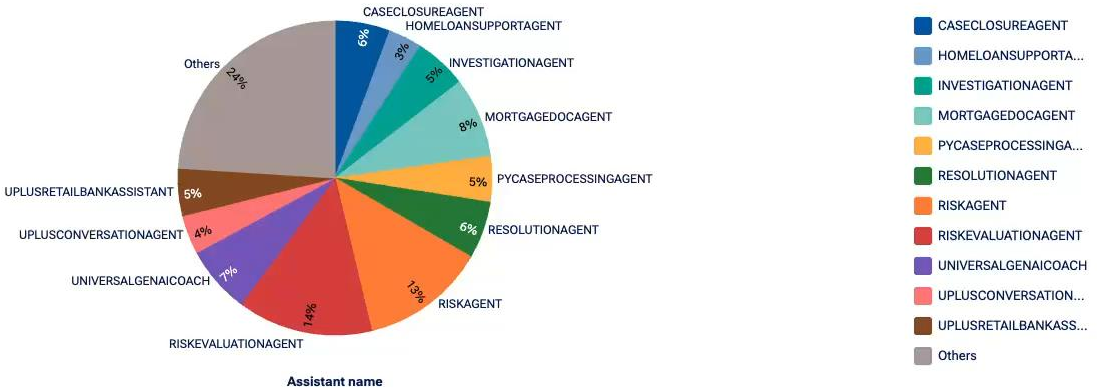
Three components, one pipeline. Run the agent against ground truth and score every response.

[Demo begins here]

Total Agent Conversations
4.24K

Total Agent Evaluations
34

Request Volume by Agent



AI Conversations

4,238 results

Case ID	Create datetime	Create Operator	Create operator name	Assistant name	Context ID	Label
PXCONV-211005	May 28, 2026 at 12:04:02 PM	Michael Spann	Michael Spann	UPLUSRETAILBANKASSISTANT	GoldenReplay	How can I support you in assisting your customers with their ban
PXCONV-211004	May 28, 2026 at 11:56:36 AM	Michael Spann	Michael Spann	UPLUSRETAILBANKASSISTANT	GoldenReplay	How can I support you in assisting your customers with their ban
PXCONV-211003	May 28, 2026 at 11:51:13 AM	Michael Spann	Michael Spann	UPLUSRETAILBANKASSISTANT	GoldenReplay	How can I support you in assisting your customers with their ban
PXCONV-211002	May 28, 2026 at 11:04:05 AM	Michael Spann	Michael Spann	UPLUSRETAILBANKASSISTANT	GoldenReplay	How can I support you in assisting your customers with their ban
PXCONV-211001	May 28, 2026 at 10:53:41 AM	Michael Spann	Michael Spann	UPLUSRETAILBANKASSISTANT	GoldenReplay	How can I support you in assisting your customers with their ban
PXCONV-210024	May 20, 2026 at 2:18:02 PM	Michael Spann	Michael Spann	UPLUSRETAILBANKASSISTANT	GoldenReplay	How can I support you in assisting your customers with their ban
PXCONV-210023	May 20, 2026 at 2:15:14 PM	Michael Spann	Michael Spann	UPLUSRETAILBANKASSISTANT	GoldenReplay	How can I support you in assisting your customers with their ban
PXCONV-210022	May 20, 2026 at 2:10:54 PM	Michael Spann	Michael Spann	UPLUSRETAILBANKASSISTANT	GoldenReplay	The customer is very upset. They tried to log in to their accoun
PXCONV-210021	May 20, 2026 at 2:08:22 PM	Michael Spann	Michael Spann	UPLUSRETAILBANKASSISTANT	GoldenReplay	How can I support you in assisting your customers with their ban
PXCONV-210020	May 20, 2026 at 2:06:35 PM	Michael Spann	Michael Spann	UPLUSRETAILBANKASSISTANT	GoldenReplay	How can I support you in assisting your customers with their ban
PXCONV-210019	May 20, 2026 at 2:03:53 PM	Michael Spann	Michael Spann	UPLUSRETAILBANKASSISTANT	GoldenReplay	How can I support you in assisting your customers with their ban
PXCONV-210018	May 20, 2026 at 1:38:19 PM	Michael Spann	Michael Spann	UPLUSRETAILBANKASSISTANT	GoldenReplay	How can I support you in assisting your customers with their ban

Run DeepEval Evaluation

Choose a golden dataset, select metrics, and run your evaluation.

1. Load Project Configuration

Select a project configuration to use for evaluation:

Select a project config... | Refresh

2. Select Golden Dataset

Choose a golden dataset to evaluate against:

Refresh List

SC Claims CFP
6 turns | 2 tools
Recorded: 2026-03-04T15:25:25.312832

Product Complaint Test
10 turns | 1 tools
Recorded: 2026-04-27T14:24:10.155789

golden_Low_Risk_Test_20260512_142230
1 turns | 0 tools
Recorded: 2026-05-12T14:22:30.051476

Complaint Test
3 turns | 1 tools
Recorded: 2026-04-27T14:24:44.008621

Agent Knowledge Test Dataset
3 turns | 1 tools
Recorded: 2026-04-24T14:43:56.229474

3. Select Metrics

Choose which DeepEval metrics to include in your evaluation:

Knowledge Retention
Checks whether context from early turns is retained throughout the session
Threshold: 0.5

Hallucination
Detects hallucinated content not grounded in context
Threshold: 0.5

Conversation Completeness
Verifies the agent completed all expected workflow stages

Role Adherence
Checks that the agent stays in its designated role throughout

Call To Action

- Download the application on GitHub using the QR Code
- Read the accompanying post on the AI Expert Circle: https://forums.pega.com/t/mastering-trust-testing-continuous-monitoring-and-safety-in-an-ai-world/11200/4

https://github.com/pegasystems/infinity-ai-agent-demonstrations

What You'll Walk Away With

Practical, repeatable patterns for testing and monitoring agentic AI usable today.

GitHub reference

Published reference implementation of a Pega agentic suite using DeepEval

Use what you have today

Apply Pega's existing testing tools to agentic workloads no new stack required

Release + runtime framework

A practical pattern for evaluating agentic AI across dev, release, and production

Early detection

Concrete signals for drift, regressions, hallucinations, and safety risks

Closed-loop feedback

Turn test results into measurable improvements release over release

Market direction

Where agentic evaluation is heading and how to stay ahead

PegaWorld

JUNE 7–9 | LAS VEGAS

PEGAWORLD.COM