Market Guide for AI Evaluation and Observability Platforms

2 February 2026 - ID G00842253 - 43 min read

By: Manjunath Bhat, Alex Coqueiro, Wilco van Ginkel

Initiatives: Technologies and Markets

The nondeterminism in GenAI and agentic AI applications makes it difficult to measure and improve reliability and earn user trust. Software engineering leaders must invest in AI evaluation and observability platforms that bring rigor and repeatability to testing business-critical AI applications.

More on This Topic

This is part of an in-depth collection of research. See the collection:

- Journey Guide to AI Application Quality and Testing

This research note is restricted to the personal use of .

Overview

Key Findings

- The 2025 Gartner AI in Software Engineering Survey shows that only 18% of respondents currently use AI evaluation tools to test the outputs and behaviors of custom-built **AI agents**. The low adoption points to opportunities for scaling the testing process through consistent and repeatable evaluations as teams take LLM-based chatbots and AI agents to production.
- The market for **AI evaluation and observability platforms** (AEOPs) is nascent but poised for growth, driven by the need for improved quality of AI applications and agents. The Gartner Software Engineering Survey for 2026 reveals that 57% of engineering leaders consider building AI capabilities in applications a highly important skill for the engineering team to deliver software that meets the goals, yet 67% still view it as a moderate to major pain point.
- AEOPs enable a continuous feedback loop by using live data from production to improve testing in preproduction. Integrating real-world observability (using online evaluations) with offline evaluations helps the team adopt shift-left and shift-right testing practices.
- Vendors differentiate themselves by moving beyond basic evaluations and observability, and offer specialized capabilities across multiple dimensions. These dimensions span openness and interoperability, vertically integrated technology stacks, domain-specific datasets, security guardrails, multimodality, and synthetic test data generation.

This research note is restricted to the personal use of .

Recommendations

- Budget for and prioritize investments in AEOPs to give your teams a systematic way to test and continuously improve GenAI applications. Explore opportunities to leverage AEOP features in existing tools such as **AI application development platforms** and **AI governance platforms** to maximize the return on investment.
- Enable consistent and repeatable evaluations using AEOPs to automate evaluation runs that measure application performance, safety, and reliability against product-specific metrics designed for your unique requirements.
- Utilize AEOPs to adopt shift-left testing practices, such as building evaluation datasets and automating evaluation runs, and complement them with shift-right practices using observability.
- Select AEOPs whose salient differentiators directly address your most critical pain points around domain-specificity, architectural needs, and support for integrating with your technology ecosystem.

Strategic Planning Assumption

By 2028, 60% of software engineering teams will use AI evaluation and observability platforms to build user trust in AI applications, up from 18% in 2025.

Market Definition

Gartner defines AI evaluation and observability platforms (AEOPs) as tools that help manage the challenges of nondeterminism and unpredictability in AI systems. AEOPs automate evaluations (“evals”) to benchmark AI outputs against quality expectations such as performance, fairness and accuracy. These tools create a positive feedback loop by feeding observability data (logs, metrics, traces) back to evals, which helps improve system reliability and alignment. AEOPs can be procured as a stand-alone solution or as part of broader AI application development platforms.

AI applications and agents are inherently nondeterministic and opaque. This makes it difficult to measure and improve their reliability, accuracy and trust using traditional testing and observability methods and tools. In addition, the vast range of possible AI application inputs and outputs makes it impossible to create test cases for every scenario. AEOPs help resolve these challenges by providing a rubric to evaluate how well an AI application performs on relevant metrics (e.g., correctness, accuracy, coherence, fairness, relevance, cost).

This research note is restricted to the personal use of .

Software engineering leaders should, therefore, invest in AEOPs that help developers in two ways. The first way is by pairing observability with evaluations to ensure continuous verification of AI systems. The second is by continuously tracking behavior to prevent application quality from degrading over time.

Software engineering teams can use AEOPs to:

- **Perform multivariate (or A/B) testing:** Testing two or more versions of AI application prototypes, prompts or AI models side-by-side. This process may include comparing a new prompt or models against an older version or evaluating a fine-tuned model against a base model to make data-driven improvement decisions.
- **Prevent regressions:** To ensure that updates or changes to the model, configuration, or prompts do not cause new issues, enabling robust quality control for the AI development pipeline.
- **Ensure safety and alignment:** To check for harmful, biased, or toxic outputs and ensure the model behaves according to responsible AI guidelines.
- **Diagnose root cause of failures or errors:** Using logs and traces to determine if a poor response was due to a hallucination, a failure in a function call, or the retrieval of irrelevant documents in a RAG pipeline.
- **Measure and improve performance:** Measuring the performance of AI applications and agents against intended objectives to identify strengths and limitations.

Mandatory Features

The mandatory features for this market include:

- **Automation of evaluation runs:** The ability to systematically test an AI system against a predefined dataset and score the outputs with custom rubrics, using multiple evaluators — code-based functions, human judgment, or LLM-as-a-judge. The ability to use evals as quality gates and ensure safety and alignment by preventing regressions and unexpected outputs from reaching production.
- **Dataset management and curation:** Curate and manage evaluation datasets at scale. Datasets are a collection of sample prompts with additional context and optional expected outputs. This feature includes capabilities to create new datasets from scratch, upload existing data, manage different versions, and annotate or label data points with ground-truth answers or desired outputs.

This research note is restricted to the personal use of .

- **Prompt life cycle management**: Support the ability to create, parameterize, version, test and replay prompts. Prompt parametrization and versioning promote reusability.
- **Sandbox environments for interactive experiments**: The sandbox environments enable technical and nontechnical stakeholders to iterate on prompts rapidly, experiment with different models and their parameters (e.g., temperature), and visually compare outputs in real time. The environments connect to model provider APIs via API keys and do not need to host the models.
- **AI system observability**: Capture logs, metrics, and traces at various levels of granularity, ranging from multistep agentic workflows to a single request-response interaction with an AI model. Logs and traces provide insights into reliability measures such as latency and error rates; trust measures such as explainability, correctness, relevance, and fairness; and cost measures such as token costs.
- **Support for creating custom metrics to suit application-specific needs**: Support the use of general-purpose metrics frameworks such as Ragas, G-Eval and GPT Estimation Metric-Based Assessment (GEMBA) to quantify subjective measures of faithfulness, coherence, relevance, and precision. Support the creation of application-specific metrics tailored to meet safety and alignment goals.
- **Online and offline evaluations**: Support for both online and offline evaluation capabilities. Offline evaluation includes support for testing the application's performance on curated or external datasets in preproduction environments. Online evaluation includes "live" monitoring of application behavior in production to assess performance and take suitable actions in real time.
- **Model-agnostic nature**: To prevent vendor lock-in and support versatile use cases, AEOPs must be model-agnostic, supporting multiple commercial and open-source models across frontier model providers.

Common Features

The differentiating features for this market include:

Evaluations:

- Running evaluations as "fitness functions" that are part of CI/CD pipelines for benchmarking, comparing AI models, and assessing whether application functionality is fit for use.

This research note is restricted to the personal use of .

- Automated prompt optimization to analyze prompts and generate better-performing alternatives.
- Support for explainability methods to help users understand why an AI model generated a specific response. Sample explainability methods include Shapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME). These techniques help identify biases, enable transparency and build user trust in AI systems.

Datasets:

- Generation of production-quality synthetic data, solving a major testing bottleneck for teams that lack preexisting evaluation datasets.
- Ability to turn production data into evaluation datasets, enabling teams to test their systems against real-world scenarios.
- Prebuilt domain-specific and specialized datasets to test application functionality for adherence to regulations around security, safety, and user privacy (PII detection).

Observability:

- Providing full-stack observability, from the application to the compute layer, to accelerate debugging, scaling, and reliability troubleshooting for AI workloads.
- Continuous monitoring for changes in input data distributions and detection of model drift that may adversely impact output quality.
- Support for integration with OpenTelemetry as an open standard for collecting and routing instrumented telemetry to different back-end systems for analysis, alerting, and visualization.
- Applying evaluations as guardrails at runtime to detect and block hallucinations, PII leakage, and prompt injections in real-time. This capability requires integrating observability and evaluations.

This research note is restricted to the personal use of .

Market Description

A core function of AEOPs is to systematically measure and improve AI application quality, moving beyond subjective human guesswork. AEOPs help design evaluations that test AI applications and **agentic AI** workflows at multiple levels of granularity (e.g., LLM calls, reasoning steps, tools invoked, task executions and workflow completions). This enables developers to make data-driven decisions before rolling out changes to users. A repeatable, data-driven process to quantify "what good looks like" is at the heart of **eval-driven development** (EDD). AEOPs therefore become the primary tools to help operationalize EDD. See How to Improve AI Quality Using Eval-Driven Development.

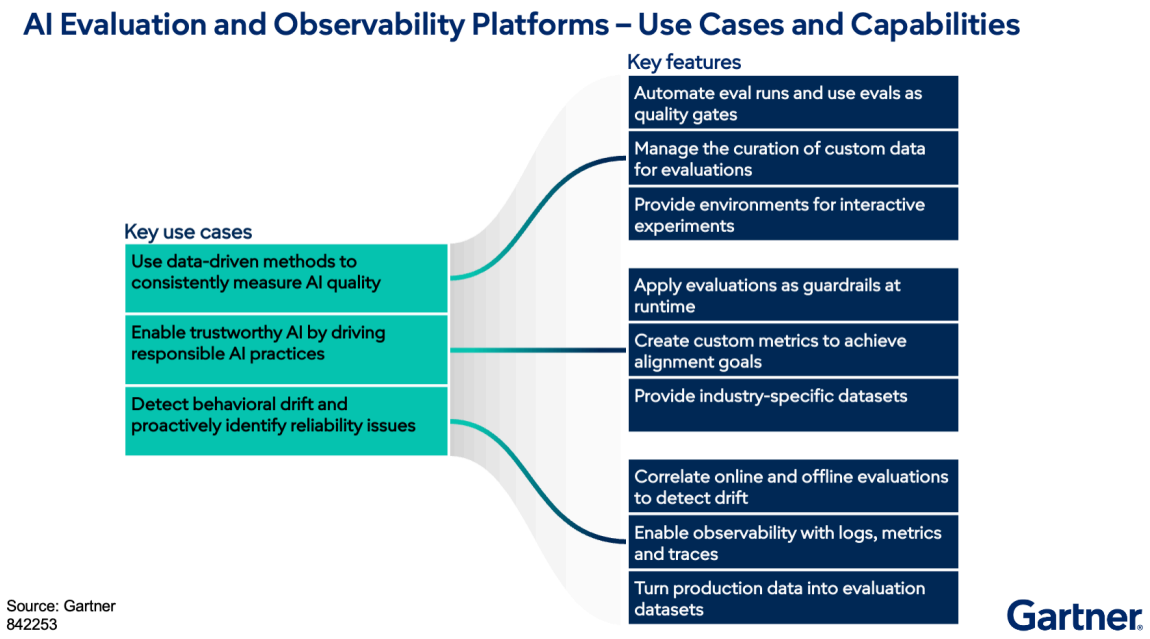
What's the difference between evals (for AI applications) and tests (for traditional applications)? Think of tests like checking if a calculator works. Evals are like grading an essay. The former has a single right answer; the latter requires judgment. Evals bring consistency and rigor to making those judgments.

A key feature of AEOPs is the ability to connect evaluations and observability. Observability data from production, such as logs and traces, helps teams analyze trends, diagnose errors, and identify alignment issues. This data can be converted to offline datasets for running evals in preproduction environments and gauging production readiness. Evals help to verify application behavior on real-world production data as well as offline datasets.

Figure 1 maps the key AEOP capabilities to key use cases.

This research note is restricted to the personal use of .

Figure 1: AI Evaluation and Observability Platform Capabilities



AEOPs provide data-driven answers to questions such as:

- **How do we know if our AI feature is working as expected?Platform feature:** *AEOPs automate evaluations of AI capabilities with real data, scoring, and benchmarking the results. You can determine whether changes improve or hurt performance.*
- **Are bad responses reaching users?Platform feature:** *Drift detection and proactive monitoring tracks live model responses and alerts you when quality drops or incorrect outputs increase.*
- **Can our team improve quality without guesswork?Platform feature:** *AEOPs can show side-by-side diffs, enabling you to compare the performance of different prompts, models and tools and see exactly why one version performs better than another.*

This research note is restricted to the personal use of .

Market Direction

AEOP vendors now offer similar core features, forcing them to differentiate through unique strategic shifts. Capabilities such as automated evaluation runs, curating and managing evaluation datasets, enabling prompt life cycle management (versioning and testing), capturing observability data such as logs and traces to diagnose errors are becoming foundational and "table stakes."

Gartner sees vendors expanding beyond basic capabilities to create distinctive value propositions by exploring market adjacencies and opportunities. This is evident from some vendors featuring across Gartner market categories. Figure 2 illustrates the AEOP market's adjacencies and opportunities for building, governing, securing, and scaling AI applications.

Figure 2: Adjacent Market Opportunities for AEOPs



AEOPs provide differentiated capabilities that overlap with adjacent markets, such as:

- AI application development platforms that use evaluations to build trustworthy AI applications
- AI governance platforms using observability and audit trails for ensuring alignment
- AI security platforms that use evaluations for AI security and compliance

This research note is restricted to the personal use of .

- **AI gateways** that use evaluations to enforce guardrails at runtime

AI Application Development Platforms That Use Evaluations and Observability to Build Trustworthy AI Applications

AEOP capabilities can be bundled as part of broader AI application development platforms. This helps integrate **responsible AI** practices during development to build applications that enhance user trust. Examples include Amazon Bedrock, CoreWeave (Weights & Biases), LangChain (LangSmith), Microsoft Foundry and Orq.ai. See the Magic Quadrant for AI Application Development Platforms.

AI Governance Platforms Using Observability and Audit Trails to Ensure Alignment

Gartner defines AI governance platforms as tools designed to ensure adherence to organizational policies, regulations and industry standards and implement responsible AI practices. AEOPs enable AI governance by transforming ethical requirements such as fairness, transparency, and trust into measurable and enforceable engineering standards using evaluations, observability and guardrails. These features provide visibility, auditability and explainability which are essential tenets of AI governance. See Market Guide for AI Governance Platforms.

AI Security Platforms That Use Evaluations for AI Security and Compliance

Some AEOP vendors offer AI safety and security-focused capabilities to simplify the development of trustworthy applications. This involves providing custom evaluators to mitigate critical risks like sensitive data leakage, prompt injections, toxicity, and bias, and the OWASP Top 10 security risks for LLM applications. These capabilities overlap with AI security platforms that test an AI application's resilience against prompt injection attacks, jailbreak attempts, and enforce security guardrails. See Top Strategic Technology Trends for 2026: AI Security Platforms.

AI Gateways That Use Evaluations to Enforce Policies at Runtime

Some vendors are creating efficient, compact, low-latency evaluation models that can run on affordable hardware and be invoked at runtime. This transforms AI evaluation from a predeployment check into a real-time guardrail that can monitor live traffic and block undesirable outputs before they reach the user. In addition, the integration between AI gateway providers and AEOPs enables their common customers to enforce evaluation checks as runtime guardrails for AI traffic. See Market Guide for AI Gateways.

How AEOP Capabilities Will Evolve

AEOP vendors are advancing their capabilities across three dimensions:

This research note is restricted to the personal use of .

- **Support for building self-healing, self-optimizing and self-improving agentic systems.** This includes features to automatically detect failure points and suggest improvements to prompts, RAG configurations, and tool calls. Observability will become more granular, moving from simple traces to deep behavioral insights with expanded support for step-level root-cause analysis that evaluates planning, tool selection, and verification steps rather than just the final output.
- **Support multimodal, multiturn interactions moving beyond static evaluations.** This includes features that simulate multiturn interactions and "digital twin" scenarios to stress-test agents before they reach production. Although most platforms are currently text-focused, vendors are adding support for multimodal use cases involving images, video, audio.
- **Enable bidirectional feedback loop between offline and online evaluations.** This includes the ability to convert offline evaluations to function as runtime guardrails in production and use production logs and traces to continually improve the quality of evaluation datasets. This bidirectional feedback loop is key to improving both functional and non-functional quality of AI applications and agents.

Market Analysis

Gartner estimates the market size for AEOPs to be approximately between $160 and $200 million in 2025. The estimate is intentionally range-based because this is an emerging market category overlapping with AI application development platforms, AI gateways, AI security platforms and AI governance platforms. Gartner also sees this market as ripe for mergers and acquisitions. For example, Anthropic acquired Humanloop, ClickHouse acquired Langfuse, CoreWeave acquired Weights & Biases and Snowflake acquired Truera.

AEOPs have relatively low market penetration. Gartner estimates that less than 25% of the potential target audience has deployed these solutions to date. However, as organizations accelerate their AI initiatives, the focus on functional quality, reliability and responsible AI practices will drive demand for automating evaluations and implementing observability leading to increased AEOP adoption.

Gartner's 2025 AI in Software Engineering Survey showed that 21% of top-performing organizations in building AI-powered solutions use evaluation tools and frameworks compared to 17% of lower performers. The same survey showed that in 69% of the cases, software engineering teams are responsible for selecting AI application development and tools.

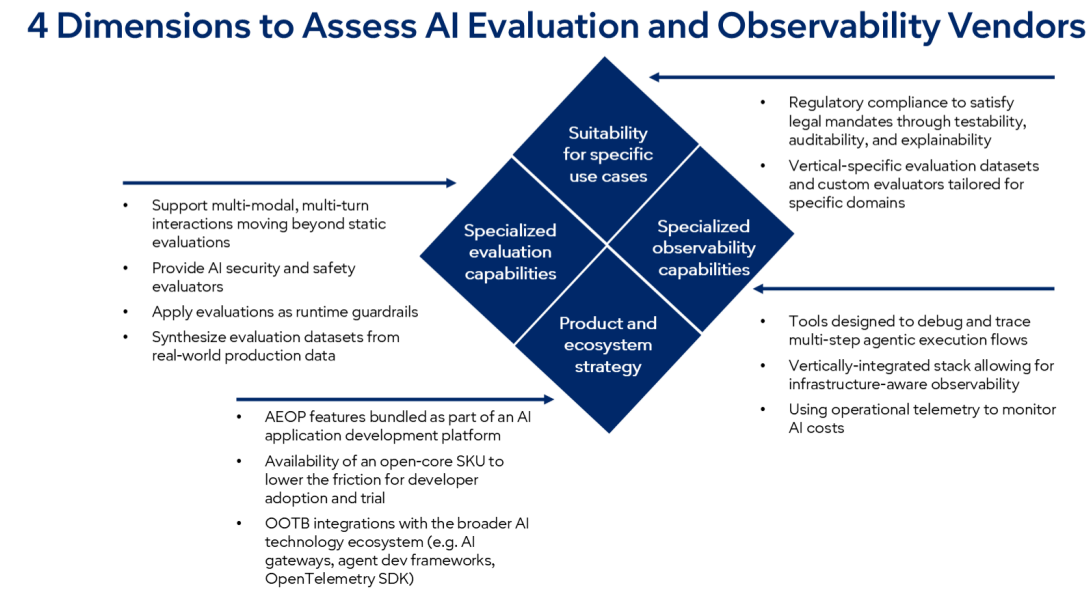
This research note is restricted to the personal use of .

Market Segmentation

The vendors in the AEOP market can be segmented across the following dimensions (see Figure 3):

- Suitability for specific use cases
- Specialized evaluation capabilities
- Specialized observability capabilities
- Product and ecosystem strategy

Figure 3: Four Dimensions to Assess AEOP Vendors



Source: Gartner
842253

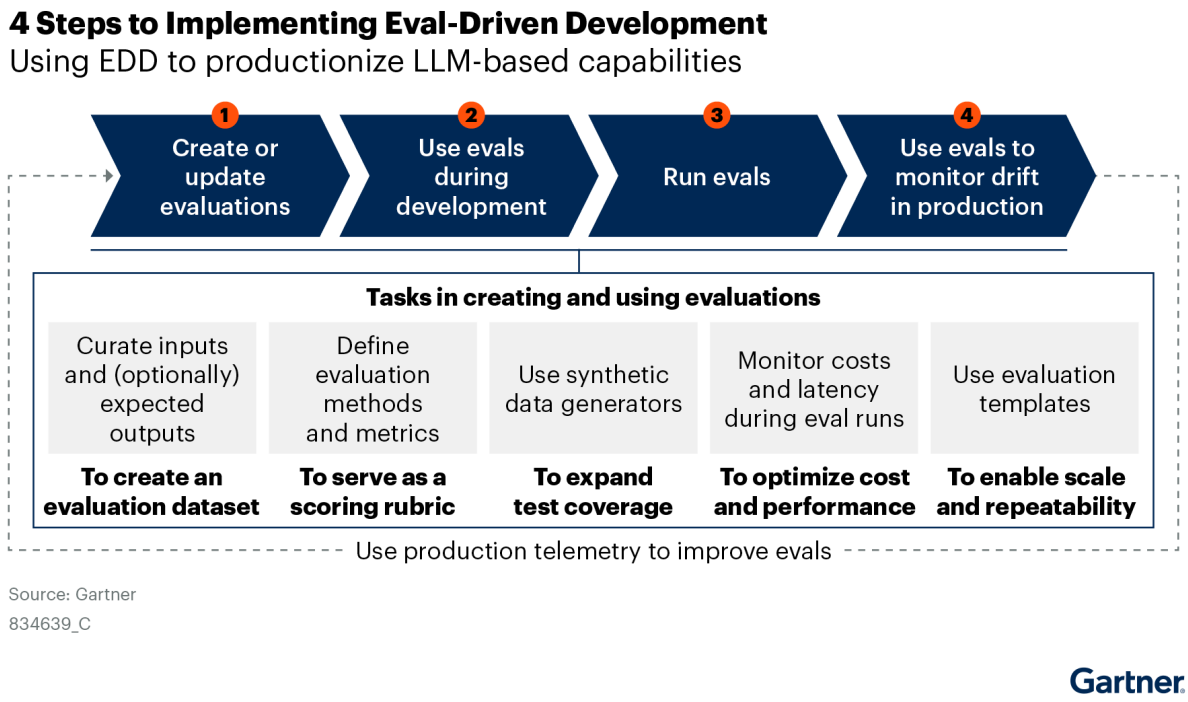
Gartner

Market Drivers Driving the Need for AEOPs

The need for eval-driven development: Traditional testing, which relies on predictable outputs, fails to account for GenAI's non-deterministic responses, resulting in a significant quality gap. An EDD approach is therefore essential for building production-ready AI applications and agents. AEOPs help teams implement EDD to systematically embed quality measures, such as explainability, predictability, and safety, replacing subjective, "gut instinct" evaluations. Figure 4 illustrates how teams can adopt EDD using AEOP capabilities.

This research note is restricted to the personal use of .

Figure 4: AEOPs Enable Adoption of Eval-Driven Development



For more details on testing AI-powered applications, see How to Test AI-Based Applications and How to Test Enterprise AI Agents.

AI agents, multiagent workflows and RAG (retrieval-augmented generation): Software engineering teams are increasingly tasked with building AI agents to address a wide range of business and operational needs. The 2025 Gartner AI in Software Engineering Survey revealed that 68% of respondents are either actively building or have already developed AI agents. However, ensuring the reliability of agentic workflows and RAG architecture in production is difficult as they can fail in unpredictable ways. The observability capabilities in AEOPs for logging and tracing help diagnose the root cause of failures and expedite issue resolution. Evaluations serve as a reliable “acceptance test” to validate that AI agents are behaving as intended, providing the rigor needed to trust these systems.

Rapid pace of innovation and change in AI technologies: The continuous, rapid change in the **generative AI** landscape, including new model capabilities with frequent model updates, requires developers to continuously update their evaluation frameworks. This drives the need for platforms that can handle iteration and experimentation (e.g., comparing model performance side-by-side) quickly. This becomes significantly critical as software engineering teams take responsibility for monitoring and updating models in production.

This research note is restricted to the personal use of .

Sixty-four percent of the respondents to the 2025 Gartner AI in Software Engineering Survey say that they are responsible for monitoring and updating AI models in production.

Need to build user trust and comply with regulatory mandates: Global regulatory efforts, such as the U.S. Executive Order on AI (EO 14179) and the NIST AI Risk Management Framework, mandate critical characteristics like transparency, explainability, and safety (mitigating ideological bias or engineered social agendas). The U.S. executive order (EO) on AI (EO 14179) calls for AI development and use without bias, specifying "ideological bias" or "engineered social agendas." The NIST AI Risk Management Framework outlines transparency, explainability, and fairness as essential characteristics of trustworthy AI. AEOPs enable organizations to comply with these regulatory requirements and standards by identifying and mitigating risk factors using automated techniques.

Representative Vendors

The vendors listed in this Market Guide do not imply an exhaustive list. This section is intended to provide more understanding of the market and its offerings.

Vendor Selection

Gartner sizes the market at about 35 vendors that offer a mix of AI evaluation and observability capabilities. The 20 listed here align with our market definition by focusing on both evaluation and observability functions and are agnostic to agent development frameworks as well as model providers. These vendors include AI evaluation and observability capabilities in their go-to-market strategy and invest in enhancing those capabilities as part of their product roadmaps (see Table 1).

This research note is restricted to the personal use of .

Table 1: Representative Vendors in the AI Evaluation and Observability Platforms Market

(Enlarged table in Appendix)

Vendor Name	Headquarters	Product name	Open-Source Tooling
Airia	Atlanta, Georgia, USA	Airia	N/A
Arize AI	Berkeley, California	Arize AX	Arize-Phoenix
Braintrust	San Francisco, California	Braintrust	Braintrust AI Proxy
Comet	New York, USA	Opik	Opik (open-source edition)
Confident AI	San Francisco, California	Confident AI	DeepEval (LLM Evaluation Framework)
CoreWeave	Livingston, New Jersey, USA	Weights & Biases Weave	Weave Python SDK, Marimo (Python Notebook)
Deepchecks	Tel Aviv, Israel.	Deepchecks LLM Evaluation	Deepchecks (testing, continuous integration and monitoring)
Evidently AI	San Francisco, California	Evidently Platform	Evidently (Python library to test and monitor AI applications)
Fiddler AI	Palo Alto, California, USA	Fiddler AI Observability and Security platform	Fiddler Auditor for auditing and adversarial testing of LLM outputs
Galileo	Burlingame, California	Galileo	N/A
HoneyHive	New York, USA	HoneyHive	N/A
IBM	Armonk, New York	watsonx.governance, watsonx Orchestrate, IBM Instana	N/A
LangChain	San Francisco, California	LangSmith	LangChain and LangGraph — AI application and agent development frameworks
Langfuse	Berlin, Germany	Langfuse	Langfuse (open-source LLM engineering platform)
Maxim	San Francisco, California,	Maxim AI	Bifrost (AI gateway with evals and observability plug-ins)
Microsoft	Redmond, Washington, USA	Microsoft Foundry	N/A
Openlayer	San Francisco, California	Openlayer	N/A
Orq.ai	Amsterdam, The Netherlands	Orq.ai	Evaluatorq (Python evaluation framework for running parallel AI evaluations)
Patronus AI	San Francisco, California	Patronus AI	Lynx (an open language model for detecting hallucinations)
Plurai	Tel Aviv, Israel	Plurai	IntellAgent (multiagent framework for evaluating conversational AI systems)

Source: Gartner

This research note is restricted to the personal use of .

Vendor Profiles

Airia

Evaluations: Airia includes a prototyping studio and an agent builder that enable teams to build custom agents, run tests, and evaluate model and agent behavior in a sandbox environment before deploying to production. Once deployed, Airia continues to evaluate agent behavior in production by capturing execution logs, interaction traces, model outputs, tool calls and user inputs effectively enabling asynchronous evaluation of AI performance, correctness, compliance, safety, and business-level KPIs.

Observability: Airia delivers multilayer observability spanning LLMs, agents, infrastructure, tools, as well as security, compliance and governance. The platform logs model inputs (prompts), outputs (responses), choices of models, token usage, along with metadata (user, request context, model version). For more complex agentic systems where multiple agents or tools collaborate, Airia tracks the orchestration flow: which agent invoked which tool, API calls, retrieval engines, database calls, external integrations, and the subsequent reasoning or action taken. Airia's secure gateway architecture (e.g., the MCP Gateway) acts as a central proxy for agent requests to external tools or APIs.

Industry-specific capabilities with support for regulatory compliance: Airia provides audit trails, data-usage controls, access policies, traceability, human-review workflows, and automated enforcement against compliance violation, aligning to frameworks like GDPR, HIPAA, financial regulations, and emerging AI governance standards.

A salient feature: Airia offers a combination of secure agent orchestration, deep observability, and AI governance delivered as a unified platform.

Arize AI

Evaluations: Arize AX supports both offline (development) and online (production) evaluations. For offline evaluations. Arize AX provides an experiment framework that runs on curated datasets in a controlled, repeatable environment. Teams can test prompts, models, retrieval strategies, and logic changes before deployment using datasets created from production traces, CSV uploads, or programmatic code. The platform supports side-by-side comparison of experiment runs, tracks performance over time.

This research note is restricted to the personal use of .

Observability: Arize AX supports observability for AI applications across multiple layers of the stack. For **LLM observability**, Arize captures full request traces — LLM calls, function calls, retrieved context, chain-of-thought reasoning, and outputs. The platform tracks latency, token usage, cost, and quality metrics in real time, with custom dashboards and alerting for production monitoring. Arize AX uses **OpenTelemetry** and supports OpenTelemetry Collector deployment patterns for distributed systems.

Industry-specific capabilities with support for regulatory compliance: Compliance features (audit logs, data deletion, SSO/RBAC, private connect) are available across industries and certain certifications are tailored to specific verticals (e.g., HIPAA certification for healthcare, PCI-DSS for financial services). The platform also offers specific prebuilt evals and notebooks for many industry-specific use cases, such as evaluating multimodal voice agents in customer service.

A salient feature: Deep visibility into how agents reason, plan, and act through multistep workflows with built-in support for session-level evaluation, tool-calling analysis, and coherence scoring.

Braintrust

Evaluations: Braintrust provides a unified evaluation system that supports offline and online evals using the same scoring functions, datasets, and quality definitions. Offline evals can be authored in code via SDKs or in the UI by PMs and non-developers. Developers can express complex scoring logic programmatically, while PMs define acceptance criteria directly in the UI through structured scoring functions. Online evals measure how your system behaves with real users. Braintrust attaches evaluators to live traffic through asynchronous processing.

Observability: The platform captures structured traces for every model call, tool invocation, agent action, and retrieval step. Traces include prompts, parameters, outputs, latency, cost, metadata, and any attached artifacts (documents, images, etc.). Using OpenTelemetry and the platform's ingestion APIs, Braintrust ties together model behavior with application infrastructure. This enables correlation of events across layers of the technology stack, including distributed tracing scenarios common in multiagent systems.

Industry-specific capabilities with support for regulatory compliance: Braintrust offers a hybrid deployment model where customer data stays in the customer VPC. The solution is HIPAA-compliant with trace-level audit logs to ensure PHI privacy. Financial institutions use Braintrust to document model decisions, enforce quality gates, and meet internal governance requirements.

This research note is restricted to the personal use of .

A salient feature: Braintrust treats every production trace as a potential test case. Teams can capture a misaligned or incorrect output from observability, pull it directly into a dataset, and evaluate new prompts or models against it.

Comet

Evaluations: Comet's Opik platform unifies offline and online evaluations into a single life cycle workflow. It offers a library of out-of-the-box metrics using heuristics and LLM-as-a-judge, and supports custom metrics to assess quality. During development, Opik validates performance by comparing model or prompt versions against datasets, scoring outputs and incorporating human feedback. In production, it continuously monitors traffic using rules and guardrails to detect regressions, hallucinations, and policy violations.

Observability: Opik delivers deep application-level observability, distinct from standard infrastructure monitoring. It provides visibility into LLM and agentic workflows by capturing granular traces—including reasoning steps, multimodal inputs, and tool usage. Opik also supports agent communication protocols like MCP to track tool schemas and selection logic, helping teams debug issues and optimize model performance.

Industry-specific capabilities with support for regulatory compliance: The industry-specific capabilities include multimodal logging for autonomous driving, strict compliance for banking, content verification for e-commerce and code evaluation heuristics for developer tools. Opik's guardrails functionality is designed to mitigate regulatory compliance concerns, with built-in eval metrics such as 'compliance risk judge,' which inspects model responses for regulatory, legal, or policy issues. The Comet platform is certified for SOC 2, ISO 27001, and ISO 9001, and compliant with HIPAA and GDPR needs.

A salient feature: Opik delivers robust multimodal support, combining tracing, auto-optimization, and offline-to-online evaluations for audio, video, and image workflows.

Confident AI

Evaluations: Confident AI's evaluation capabilities are powered by DeepEval, a popular open-source LLM evaluation framework. The platform has 3 components for automating evaluations - first, LLM Metrics which automates manual predeployment AI testing with 40+ LLM-as-a-Judge metrics. Second, Dataset Curation for annotating or generating test datasets. Third, LLM Tracing for Real-time AI execution observability with performance tracking.

This research note is restricted to the personal use of .

Observability: Confident AI provides observability for AI systems across the application stack, spanning LLM behavior, agentic workflows, agent communication protocols (e.g., MCP, A2A), and runtime performance. It captures detailed LLM telemetry including prompts, responses, structured inputs/outputs, token usage, latency, cost, model and prompt versions, and evaluation signals such as correctness, faithfulness, relevancy, safety, and regression over time.

Industry-specific capabilities with support for regulatory compliance: Confident AI's support for multiturn simulations and red teaming in conjunction with evals and observability reduces the time it takes to build and deliver enterprise-grade chatbots. Confident AI is HIPAA compliant and executes Business Associate Agreements (BAAs) with eligible healthcare clients.

A salient feature: Confident AI can run online evaluations on traces in production environments, ingest the traces into the platform and use them for offline evaluations. This flexibility removes the strict need for creating an upfront dataset.

CoreWeave

Evaluations: W&B Weave supports both offline and online evaluations. With Offline evaluations, users can experiment with LLMs, prompts, RAG, and agents and measure improvements across multiple dimensions — reliability, latency, cost, user experience, and safety. With Online evaluations, users can discover failure modes and continuously improve their AI agent or application. Weave online evaluations run on the W&B platform — decoupled from application code — minimizing dependencies and latency for end users.

Observability: W&B Weave offers observability for agentic applications, spanning development to production. It organizes inputs, outputs, and metadata into granular trace trees, aggregating cost and latency metrics at multiple levels for rapid root-cause analysis. In production, Weave uses live traces and online evaluations to identify edge cases and monitor quality. The platform also simplifies integration with auto-logging for MCP and language-agnostic tracing via OpenTelemetry.

Industry-specific capabilities with support for regulatory compliance: W&B Weave is certified under ISO/IEC 27001:2022, ISO/IEC 27017:2015, and ISO/IEC 27018:2019, and are compliant with SOC 2 Type 2 and HIPAA standards making them well-suited to healthcare and financial services verticals. It also offers specialized capabilities for training and evaluating models for use cases in autonomous driving.

This research note is restricted to the personal use of .

A salient feature: W&B Weave is part of a broader AI developer and inference platform that supports model life cycle management and governance as well as sharing and tracking the lineage of training datasets across the AI application development life cycle.

Deepchecks

Evaluations: The Deepchecks LLM Evaluation platform supports offline and online evaluations. During development, it supports offline batch evaluations using prebuilt templates and an SDK to compare different versions of prompts, models and routing mechanisms. In production, the platform performs real-time scoring of every interaction (span), which is then aggregated into application-level scores.

Observability: Deepchecks provides multilayered visibility into AI systems, capturing raw and aggregated data across atomic interactions (spans), full execution paths (traces), and groups of related traces (sessions) to provide a broader view of system behavior. Users can track real-time operational metrics such as latency, token usage and costs. The platform supports auto-instrumentation of agents with development frameworks such as CrewAI, LangGraph, and Google ADK. Deepchecks integrates with monitoring tools such as Datadog, New Relic, and AWS CloudWatch for alerting about drift and anomalies.

Industry-specific capabilities with support for regulatory compliance: Deepchecks enables domain-specific performance tracking using built-in properties developed using data from legal, healthcare, and financial industries. An example of one such property is 'PII Risk' that helps surface personally identifiable information.

A salient feature: Deepchecks provides an analysis tool that helps identify and understand the failure patterns behind poor scores for each LLM-based interaction. The analysis groups related issues into high-level, meaningful categories and presents each category with concrete examples taken from the actual failing cases.

Evidently AI

Evaluations: Evidently AI employs a Hybrid Quality Assurance framework for both structured data and unstructured text, supporting a Shift-Left testing approach that mirrors production monitoring. It utilizes a dual-core engine, comprising a local Python library and centralized platform, to execute deterministic text descriptors and probabilistic LLM-as-a-Judge assessments.

This research note is restricted to the personal use of .

Observability: The platform leverages Tracely, an open-source OTel library, to provide real-time monitoring and Waterfall View visualizations of request life cycles, integrating seamlessly with Grafana and Jaeger. Uniquely, Evidently monitors drift in unstructured data via vector embeddings alongside traditional tabular metrics, ensuring comprehensive coverage for hybrid systems.

Industry-specific capabilities with support for regulatory compliance: Designed for regulated sectors like finance and defense, Evidently AI supports Air-Gapped and on-premises deployments to maintain **data sovereignty**.

A salient feature: Evidently AI positions itself as a solution for Hybrid AI Systems, refusing to bifurcate traditional **MLOps** and GenAI observability. Leverages local-first architecture that ensures superior data control compared to SaaS-only tools.

The vendor did not fact review the content and that the profile is based on publicly available information and other resources.

Fiddler AI

Evaluations: Fiddler Evals is an evaluation platform that includes built-in evaluators, custom evaluation support, and a comparison interface to compare different prompts, test model updates and analyze multiple evaluation runs side-by-side. The platform enables an unified development-to-production workflow by running the same evaluators during development and production monitoring and feeding production insights back into evaluation datasets.

Observability: The observability capabilities are framework-agnostic supporting LangGraph, Strands, ADK or custom implementations (using Fiddler client SDK). Fiddler's LLM monitoring solution tracks the AI application's inputs and outputs, then enriches this data with specialized metrics that measure quality, safety, and performance. Fiddler agentic observability supports multistep agentic workflows through dedicated dashboards, trace visualization and metrics such as success rates, latency and throughput.

Industry-specific capabilities with support for regulatory compliance: Fiddler helps create audit evidence and audit trails to support governance, risk, and compliance requirements(e.g., SR 11-7, EU AI Act). Fiddler supports SaaS, VPC, on-premises, and air-gapped deployment modes that come in handy for highly regulated verticals.

This research note is restricted to the personal use of .

A salient feature: Fiddler provides a control plane that includes evaluations, observability, and real-time guardrails as part of an integrated AI Observability and Security platform. The evaluators use purpose-built 'trust models' that can operate within the customer environment without incurring the cost of external API calls and with low response times.

Galileo

Evaluations: Galileo unifies offline and online evaluations through a continuous 'evaluation-to-guardrail' life cycle. For offline evaluation, users have access to out-of-the-box metrics, **synthetic data** generation, and A/B testing via the 'Experiments' framework. For online evaluations in production, Galileo deploys real-time monitoring and runtime guardrails powered by Luna-2, which are specialized small language models (SLMs). These SLMs help convert offline evaluations into low-latency guardrails that are more cost-effective than using foundation LLMs.

Observability: Galileo supports observability and multiagent orchestration. It captures granular traces including RAG pipelines, tool invocations, and MCP interactions visualized through an interactive 'agent graph' for easy troubleshooting. For debugging, the platform includes an 'Insights Engine' that uses reasoning models and a proprietary knowledge base to identify failure modes and root causes. Galileo supports distributed tracing and provides pretested integrations via OpenTelemetry with frameworks such as LangChain, LangGraph, CrewAI, and the OpenAI Agents SDK.

Industry-specific capabilities with support for regulatory compliance: For companies in regulated verticals, it provides SOC-2 Type 2 compliance, audit trails, and self-hosted deployments. In Healthcare, it supports HIPAA compliance via air-gapped architecture and automated PII masking. For companies in the Telco sector, Galileo protects government-sensitive data with US-only access controls while maintaining high-performance SLAs at scale.

A salient feature: Galileo transforms offline evaluations into runtime guardrails and provides proprietary SLMs to enforce the guardrails in production with minimal impact on user latency.

This research note is restricted to the personal use of .

HoneyHive

Evaluations: HoneyHive integrates offline and online evaluations to ensure quality throughout the life cycle. Before deployment, users can validate performance via curated datasets, historical replays, and scenario tests. In production, evaluations are attached to live traces. Using rule-based, model-based, and human evaluators, HoneyHive links results directly to trace data to detect regressions and enforce quality gates within **CI/CD** workflows.

Observability: HoneyHive provides OpenTelemetry-native tracing for AI systems including LLM calls, multistep agent workflows, tool use, retrieval steps, and agent orchestration frameworks. It unifies logs, metrics, traces, and evaluations into a single consistent schema, allowing for real-time analysis of performance, quality, and drift. Additionally, HoneyHive uses structured span instrumentation to observe agent communication patterns, including MCP and A2A interactions.

Industry-specific capabilities with support for regulatory compliance: HoneyHive supports vertical-specific requirements for financial services, insurance, healthcare, and other regulated industries through capabilities such as on-prem support, PII redaction, audit logging, configurable data residency controls, labeling workflows, and deployment patterns that allow each business unit or region to operate in its own environment while maintaining shared governance.

A salient feature: HoneyHive supports multimodal observability, enabling tracing and analyzing application performance when processing images, audio, video, and documents with embedded media.

IBM

Evaluations: watsonx.governance has capabilities for both build-time (offline) evaluation and runtime/production (online) evaluations for AI applications and agents. watsonx.governance includes out-of-the-box metrics and supports creating custom metrics via an SDK and a GUI. The platform provides time-series charts and the ability to configure metrics with thresholds for alerts.

This research note is restricted to the personal use of .

Observability: watsonx.governance, watsonx Orchestrate and IBM Instana support generative and agentic AI observability. watsonx Orchestrate provides an observability dashboard to track metrics, outputs and latency across AI systems and multiagent workflows. watsonx.governance provides capabilities to monitor behavioral drifts, enforce policies and trigger corrective actions. Instana supports observability across the application and infrastructure stack including models, multiagent systems, agent communication protocols and the deployment infrastructure.

Industry-specific capabilities with support for regulatory compliance: watsonx.governance provides an immutable audit trail and data signals around bias, drift and explainability. These metrics feed into regulatory checks like fairness monitoring for demographic parity, supporting requirements such as New York City Local Law 144. The platform also includes predefined compliance accelerators — covering regulatory frameworks such as the EU AI Act, U.S. State AI laws, Model Risk Management (SR11-7, SS/123), NIST AI RMF and ISO 42001.

A salient feature: watsonx.governance includes purpose-built small models trained in collaboration with IBM Research and IBM Granite Guardian models that are designed to detect risks in prompts and responses and run on edge devices with low computational power.

LangChain

Evaluations: LangSmith utilizes Trajectory Evaluation to assess the sequence of agent reasoning, tool selections, and intermediate steps alongside standard correctness checks. The platform enables comparison of execution trajectories for regression or deviations to convert production traces into regression test datasets for comparative experiments. You can capture production traces and feedback/labels to construct evaluation datasets and then run comparative experiments. The built-in dataset/feedback pipeline supports offline evaluations and A/B testing.

Observability: Acting as a unified OpenTelemetry back end, LangSmith supports native instrumentation and W3C Trace Context for end-to-end distributed tracing across upstream and downstream services. The infrastructure leverages ClickHouse to manage high-volume, high-cardinality data, enabling granular analysis of token costs, latency, and deep-nested hierarchical JSON trees.

Industry-specific capabilities with support for regulatory compliance: LangSmith offers a “Hybrid Deployment” model that decouples the management UI from a self-hosted Data Plane, ensuring sensitive payloads and PII never leave the customer’s infrastructure.

This research note is restricted to the personal use of .

A salient feature: LangGraph enables Time Travel debugging by treating agent state as immutable checkpoints. This allows developers to actively rewind execution, fork processes, and manually edit state or inputs to forward-fix runs and validate logic through interactive counterfactual experiments.

The vendor did not fact review the content and that the profile is based on publicly available information and other resources.

Langfuse

Evaluations: Langfuse treats evaluation as an optimization problem within CI frameworks, using Experiments and versioned datasets as immutable controls to test model or prompt modifications. The platform employs a triad of validation combining deterministic heuristics and LLM-as-a-Judge to calibrate metrics. It also acts as a persistence layer for external libraries enabling visual analysis to compare candidate runs against baselines.

Observability: The platform utilizes a dual-database architecture (PostgreSQL and ClickHouse) to handle massive telemetry volumes with sub-second latency, aligned with OpenTelemetry standards for distributed tracing across polyglot stacks. Key features include **FinOps** tools for granular cost/token attribution and managed prompt infrastructure that supports versioning and hot swapping logic without code redeployment.

Industry-specific capabilities with support for regulatory compliance: Langfuse supports strict data sovereignty via a fully self-hostable version (Docker/**Kubernetes**) for air-gapped deployments where data remains in the customer's VPC. The platform is SOC 2 Type II, ISO 27001, GDPR, and HIPAA aligned.

A salient feature: Langfuse adopts a glass box philosophy with an Open Source (MIT-icensed), framework-agnostic core that ensures low lock-in risk via open standards.

The vendor did not fact review the content and that the profile is based on publicly available information and other resources.

This research note is restricted to the personal use of .

Maxim

Evaluations: Maxim AI can trigger evaluations on production logs and run preproduction experiments using datasets and evaluators (custom or prebuilt). It supports the creation, versioning and management of multimodal datasets by importing CSVs, using APIs or curating directly from production logs or human annotated data. It provides an interactive "playground" to test prompts and agents with multiturn simulations and allows users to experiment with different models and configurations before production. The platform includes an 'Insights layer,' enabling users to query their agent logs and evals using simple natural language queries.

Observability: Maxim AI is an OpenTelemetry-compatible observability platform designed to monitor and operate multiagent systems with the ability to trace the execution path - reasoning steps, tool usage, and retrieval to analyze failures. It also integrates with operational tools like Slack and PagerDuty for alerts and notifications, Snowflake and NewRelic for data connectors/export, Composio/Gumloop for MCP tools.

The platform uses evaluators as real-time checks (via the Bifrost gateway) to block or flag harmful inputs and outputs before they reach the user, ensuring safety during execution.

Industry-specific capabilities with support for regulatory compliance: The platform is compliant with SOC 2 Type 2, HIPAA, and GDPR requirements.

A salient feature: Maxim offers an open-source AI gateway (Bifrost) that unifies access to 15+ providers including OpenAI, Anthropic, AWS Bedrock, Google Vertex through a single OpenAI-compatible API. It supports automatic failover, load balancing and semantic caching with minimal configuration.

Microsoft

Evaluations: Microsoft Foundry supports both offline (development) and online (production) evaluations through a unified evaluation and observability platform. During development, teams can run batch/offline evaluations using built-in or custom evaluators to measure quality, correctness, safety, task adherence, and groundedness across curated test datasets. Integrations with GitHub Actions and Azure DevOps enables offline evaluation of AI models and agents within CI/CD pipelines.

This research note is restricted to the personal use of .

Observability: Microsoft Foundry provides observability across the AI application stack, combining model-specific insights and agent behavior monitoring in one solution. Observability in the Foundry Control Plane enables teams to evaluate, monitor, and optimize the quality, performance, cost and safety of multiagent AI systems across their development life cycle. Developers can trace LLM calls, tool invocations, and agent communication protocols with OpenTelemetry-compatible traces, enabling visibility into latency, cost, failures, and policy compliance.

Industry-specific capabilities with support for regulatory compliance: Microsoft Foundry leverages Azure's built-in regulatory compliance frameworks to address specific concerns such as GDPR, HIPAA, and FedRAMP. Foundry enables audit logging, data classification, and role-based access controls, while native integrations with tools like Azure Policy, Microsoft Purview and Azure Monitor allow automated enforcement of compliance policies and real-time monitoring of data governance.

A salient feature: Microsoft Foundry is an unified platform for building, deploying, and governing AI systems and natively integrates with the broader ecosystem of security and access control tools like Entra ID, Purview, and Defender.

Openlayer

Evaluations: Openlayer unifies governance and observability by bridging offline development and online production evaluations. The platform supports prebuilt behavioral tests covering hallucinations, bias, latency, robustness, and toxicity with integrations into CI/CD pipelines to validate model and prompt updates. Once deployed, the platform supports continuous monitoring to analyze anomalies, regressions, latency, cost and system behavior.

Observability: Openlayer supports observability for AI systems, including LLM calls, multistep agents, retrieval pipelines, and downstream application logic. For multiagent systems, Openlayer records agent-to-agent interactions, tool calls, and message-passing sequences, enabling visibility into coordination failures and protocol deviations. Openlayer also monitors infrastructure-level signals such as throughput, model performance degradation, and cost anomalies through provider APIs and system-level telemetry.

This research note is restricted to the personal use of .

Industry-specific capabilities with support for regulatory compliance: Openlayer prefers customers to extend their library of prebuilt tests to meet industry-specific standards. However, companies in regulated verticals can benefit from real-time security guardrails to avert prompt injection attacks and PII leakage. The platform provides mappings to the EU AI Act, NIST RMF, ISO 42001, TRAIGA, and internal policies, with continuous evidence collection and audit-ready reporting.

A salient feature: Openlayer's evaluation and observability capabilities are part of a broader AI governance platform that unifies oversight, security, and compliance across the full AI life cycle, spanning model and agent evaluation, continuous monitoring, risk scoring, and automated compliance.

Orq.ai

Evaluations: Orq.ai provides a unified evaluator framework that is used consistently for both offline (development) and online (production) evaluations. Orq.ai includes support for built-in and custom evaluators for metadata(e.g., tokens, cost, latency), agent evals (e.g., trajectory evaluations, tool-call quality and task completion), semantic scoring (e.g., quality, tone, reasoning, safety) and RAG-specific metrics (faithfulness, context relevance, precision, recall). The same evaluators can be attached to live deployments for continuous monitoring of quality, safety, compliance, and correctness in production.

Observability: Orq.ai provides observability spanning LLM execution, multiagent orchestration, agent communication protocols, and AI infrastructure. The basic capabilities include OpenTelemetry-support,

and granular insights on costs, token count, pricing per model, latency. The multiagent system observability capabilities include tracing of agent workflows, decisions, tool usage and visibility into agent-to-agent and agent-to-tool message flows. At the infrastructure level, it captures compute utilization, inference latency, and can attribute cost by application, agent, model, and user.

Industry-specific capabilities with support for regulatory compliance: Orq.ai supports role-based access control policies and permissions, audit logging, configurable data retention, and controlled data storage regions. These controls support alignment with regulations such as GDPR, SOC-aligned internal controls, and enterprise data governance requirements. Orq is compliant with GDPR, SOC II, EU AI Act and HIPAA.

This research note is restricted to the personal use of .

A salient feature: Orq.ai offers evaluation and observability capabilities as part of a broader GenAI engineering platform that includes support for an AI Gateway, agent runtime, and RAG-as-a-Service. It supports an airgapped on-premises deployment model.

Patronus

Evaluations: Patronus AI provides an automated framework that replaces human inspection with rigorous pointwise and pairwise assessments using specialized judge models like Lynx and Glider rather than generic LLMs. Users can assess core RAG metrics and enterprise scores while using Patronus Experiments to optimize performance by A/B testing prompts and hyperparameters against evaluation datasets.

Observability: The platform merges observability with evaluation via a managed OpenTelemetry endpoint that automatically converts streamed traces into prompt/response records for scoring. It features deep agentic tracing to log tool invocations and reasoning steps to identify issues like loops, supported by real-time webhooks and alerts for detecting failures in production.

Industry-specific capabilities with support for regulatory compliance: Targeting regulated sectors, the platform offers curated benchmarks like FinanceBench for financial reasoning and EnterprisePII to certify protection against data leakage. Compliance is supported by SimpleSafetyTests for risks like fraud and an auditability feature that generates natural language explanations for test failures to map errors to risk registers.

A salient feature: Patronus AI differentiates itself through a research-first approach centered on Percival, an eval copilot that detects agentic failure modes and generates prompt fixes. It relies on high-fidelity models and valid standards like TRAIL and BLUR to provide a verification layer beyond standard benchmarks.

The vendor did not fact review the content and that the profile is based on publicly available information and other resources.

This research note is restricted to the personal use of .

Plurai

Evaluations: Plurai supports both offline (development) and online (production) evaluations. During development, Plurai can generate realistic test datasets, simulate scenarios, break down product requirements into testable policy graphs for systematic test coverage and run A/B experiments across prompt versions, models, and configurations with detailed performance metrics. In production, Plurai supports custom guardrails tailored for specific use cases (not generic filters) and drift detection by identifying performance degradation and unexpected outputs in real-time.

Observability: Plurai supports observability in three ways. First, it provides visibility for non-engineers including product managers and executives to get a high-level overview of live agent performance, reliability, and regressions. Second, it accelerates root-cause analysis through explainability of failures. Third, it enables an actionable improvement loop by recommending new evaluations and new guardrails for improving accuracy, safety, and reliability.

Industry-specific capabilities with support for regulatory compliance: Plurai provides auditability, version control, and real-time monitoring to ensure agents consistently operate within required regulatory boundaries across sectors like finance, healthcare, and insurance. It is AICPA-SOC-certified.

A salient feature: Plurai supports high-fidelity simulation of real-world scenarios using product specifications, policies, and synthetic datasets. It incorporates an automated evaluation framework (Sentient Agent as a Judge) to simulate human-like emotions and sentiments during multiturn conversations.

Market Recommendations

Software engineering leaders evaluating AEOPs should:

- Guide the procurement of AEOPs based on mandatory and differentiated features for the market as outlined in this research. Use the four dimensions to segment the market while creating a representative list of vendors. However, recognize the nascent stage of the market and the potential for vendor consolidation. For example, the Humanloop platform was sunset as of 8 September 2025 after its acquisition by Anthropic.

This research note is restricted to the personal use of .

- Upskill your teams in AI evaluations and observability as a practice, going beyond building specialized expertise in a specific tool. Invest in continuous training and knowledge sharing via peer-to-peer learning. See How to Upskill Software Engineering Teams in the Age of AI.
- Before selecting an AEOP provider, analyze the failure modes for the AI capabilities in your application to determine the nature of evaluations by performing an error analysis of a small sample of traces (20—50 outputs). This is similar to how you would perform threat modeling to understand the nature of threats before investing in specific application security tools.
- Pair the adoption of AEOP with the adoption of an eval-driven development approach in building AI applications. To maintain error rates within acceptable thresholds, include evals in your "definition of done" and use evaluators as quality gates within your continuous integration (CI) pipelines. Measure application performance, safety, and reliability against product-specific metrics designed for your unique requirements. See How to Build LLM-Powered Applications which provides a 4-step framework to confidently evaluate and deploy LLM-based applications to production.

Evidence

Gartner Software Engineering Survey for 2026. This survey was conducted to provide a comprehensive understanding of the current landscape in software engineering, as well as to determine the priorities and strategic challenges of software engineering leaders. It also aims to identify the demand for various roles and skills within software engineering organizations, and assess their budget expectations, team structures and organizational outcomes. Finally, it explores the integration of AI in software engineering workflows and its impact on engineering organizations. The survey was conducted online from August through November 2025 among 482 respondents from the U.S. (n = 360) and U.K. (n = 122). Qualifying organizations operated in multiple industries and reported enterprisewide revenue for fiscal year 2024 of at least $250 million or equivalent. Qualified participants were highly involved in managing software engineering/application development teams and the activities they perform. Disclaimer: The results of this survey do not represent global findings or the market as a whole, but reflect the sentiments of the respondents and companies surveyed.

This research note is restricted to the personal use of .

2025 Gartner AI in Software Engineering Survey. This study was conducted to explore the adoption of AI within software engineering functions, focusing on two key areas: the use of AI tools (e.g., **AI code assistants**, AI code agents) throughout the software engineering life cycle (SDLC); and the development of AI-powered solutions (or AI engineering) within software engineering functions, along with their contribution to business outcomes. The research was conducted online from 29 April through 25 June 2025 among 299 respondents from North America (n = 150), EMEA (n = 104) and Asia/Pacific (n = 45). Quotas were established for company sizes and for industries to ensure a good representation across the sample. Organizations were required to be either piloting or using AI tools in SDLC for less than four years, and either piloting or having built AI solutions in their software engineering functions. Respondents included both leaders and individual contributors from software engineering functions, each with at least one year of tenure at their current organization. All respondents were involved in decision making or directly engaged in using AI tools or building AI solutions within their software engineering functions. Disclaimer: The results of this survey do not represent global findings or the market as a whole, but reflect the sentiments of the respondents and companies surveyed.

Gartner Software Engineering Survey for 2025. The survey was conducted to provide a comprehensive understanding of the current landscape in software engineering. It aims to identify the demand for various roles, essential skills and upskilling strategies within the software engineering organization. It explores the integration of AI in software engineering workflows, their leadership experiences and prior roles of the current leaders. It also assesses their budget expectations, team structures, organizational outcomes and priorities. The survey was conducted online from October through December 2024 among 400 respondents from the U.S. (n = 320) and U.K. (n = 80). Qualifying organizations operated in multiple industries (excluding the IT software industry involved in the development of commercial software and the education sector) and reported enterprisewide revenue for fiscal year 2023 of at least $250 million or equivalent. Qualified participants were highly involved in managing software engineering/application development teams and the activities they perform. Disclaimer: The results of this study do not represent global findings or the market as a whole, but reflect the sentiments of the respondents and companies surveyed.

Note 1 Gartner's Initial Market Coverage

This Market Guide provides Gartner's initial coverage of the market and focuses on the market definition, rationale for the market and market dynamics.

This research note is restricted to the personal use of .

Recommended by the Authors

Some documents may not be available as part of your current Gartner subscription.

How to Improve AI Quality Using Eval-Driven Development

Embed Responsible AI Into AI Application Development to Build Trust

Magic Quadrant for AI Application Development Platforms

© 2026 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates. This publication may not be reproduced or distributed in any form without Gartner's prior written permission. It consists of the opinions of Gartner's Business and Technology Insights Organization, which should not be construed as statements of fact. While the information contained in this publication has been obtained from sources believed to be reliable, Gartner disclaims all warranties as to the accuracy, completeness or adequacy of such information. Although Gartner insights may address legal and financial issues, Gartner does not provide legal or investment advice and its insights should not be construed or used as such. Your access and use of this publication are governed by Gartner's Usage Policy. Gartner prides itself on its reputation for independence and objectivity. Its insights is produced independently by its Business and Technology Insights Organization without input or influence from any third party. For further information, see " Guiding Principles on Independence and Objectivity." Gartner insights may not be used as input into or for the training or development of generative artificial intelligence, machine learning, algorithms, software, or related technologies.

This research note is restricted to the personal use of .

Table 1: Representative Vendors in the AI Evaluation and Observability Platforms Market

Vendor Name	Headquarters	Product name	Open-Source Tooling
Airia	Atlanta, Georgia, USA	Airia	N/A
Arize AI	Berkeley, California	Arize AX	Arize-Phoenix
Braintrust	San Francisco, California	Braintrust	Braintrust AI Proxy
Comet	New York, USA	Opik	Opik (open-source edition)
Confident AI	San Francisco, California	Confident AI	DeepEval (LLM Evaluation Framework)
CoreWeave	Livingston, New Jersey, USA	Weights & Biases Weave	Weave Python SDK, Marimo (Python Notebook)
Deepchecks	Tel Aviv, Israel.	Deepchecks LLM Evaluation	Deepchecks (testing, continuous integration and monitoring)
Evidently AI	San Francisco, California	Evidently Platform	Evidently (Python library to test and monitor AI applications)
Fiddler AI	Palo Alto, California, USA	Fiddler AI Observability and Security platform	Fiddler Auditor for auditing and adversarial testing of LLM outputs
Galileo	Burlingame, California	Galileo	N/A
HoneyHive	New York, USA	HoneyHive	N/A

This research note is restricted to the personal use of .

Vendor Name	Headquarters	Product name	Open-Source Tooling
IBM	Armonk, New York	watsonx.governance, watsonx Orchestrate, IBM Instana	N/A
LangChain	San Francisco, California	LangSmith	LangChain and LangGraph — AI application and agent development frameworks
Langfuse	Berlin, Germany	Langfuse	Langfuse (open-source LLM engineering platform)
Maxim	San Francisco, California,	Maxim AI	Bifrost (AI gateway with evals and observability plug-ins)
Microsoft	Redmond, Washington, USA	Microsoft Foundry	N/A
Openlayer	San Francisco, California	Openlayer	N/A
Orq.ai	Amsterdam, The Netherlands	Orq.ai	Evaluatorq (Python evaluation framework for running parallel AI evaluations)
Patronus AI	San Francisco, California	Patronus AI	Lynx (an open language model for detecting hallucinations)
Plurai	Tel Aviv, Israel	Plurai	IntellAgent (multiagent framework for evaluating conversational AI systems)

Source: Gartner

This research note is restricted to the personal use of .