Altimetrik

WHITEPAPER

Evals in Production:

Lessons from Building AI Agent Evaluation

Frameworks Across Regulated and Enterprise Industries

Synthesizing real-world evaluation architectures from agentic AI deployments across financial services, life sciences, investment management, and technology sectors

AUTHOR

Shivkumar Krishnan

CONTRIBUTING ARCHITECTS & TECHNICAL LEADS

Martin Gaida | Sathiyanarayanan Gopal | Atharva Joshi | Konrad Jackowski | Sandeep Kalra | Pavan Muthozu
Yasmeen Sultana | Mahesh Varavooru

Executive Summary

As enterprises move beyond AI pilots into production-grade agentic systems, evaluation, the practice of systematically measuring AI quality, reliability, and safety, has emerged as the most critical and least standardized discipline in AI engineering. This white paper synthesizes evaluation architectures, frameworks, and lessons learned from agentic AI deployments across eight enterprise engagements spanning financial services, life sciences, investment management, cybersecurity, and B2B software.

Across all engagements, three patterns consistently distinguished high-performing deployments from those that struggled:

- Evaluation is treated as a first-class engineering concern, not an afterthought
- Evaluation frameworks span the entire lifecycle - pre-production, regression, and continuous production monitoring
- The choice of eval methodology is driven by the nature of the output (structured vs. generative) and the regulatory context of the industry

This paper provides a structured analysis of evaluation approaches, tooling decisions, and architectural patterns that practitioners can apply to their own AI programs.

Introduction:
Why Evals Are the New Unit Tests

The analogy is imperfect but instructive: evaluation frameworks for AI agents are what unit and integration tests are for software - the mechanism by which teams develop confidence that a system does what it is supposed to do, and that changes do not introduce regressions.

But AI evaluation is harder than software testing for several reasons. The output space is unbounded. Ground truth is expensive to establish. Performance degrades silently in production. And the systems being evaluated (multi-agent pipelines, RAG architectures, tool-calling agents) introduce multiple failure modes that interact in non-obvious ways.

Across the eight enterprise deployments analyzed in this paper, engineering teams tackled these challenges with a combination of established principles and highly contextual innovation. What emerges is a coherent, if not yet standardized, set of practices for enterprise-grade AI evaluation.

SECTION 01

The Evaluation Landscape

1.1 Three Phases of Production Evaluation

Every mature evaluation program in our dataset operated across three temporal phases:

Phase	Purpose	Key Activities
Pre-Production	Establish baseline quality before deployment	Golden dataset creation, LLM-as-judge calibration, stress testing, cost/latency profiling
Regression Testing	Prevent quality degradation from changes	CI-integrated test suites, prompt change triggers, model version gates, threshold enforcement
Production Monitoring	Detect drift and collect ground truth	Live outcome metrics, SME sampling, user feedback signals, A/B shadow testing

> **Key Insight: The Closed Loop**
>
> *The most sophisticated deployments, particularly in regulated industries, explicitly fed production findings back into pre-production benchmarks. This "closed loop" is what prevents evaluation from becoming stale. Production discoveries became next month's regression cases.*

1.2 Evaluation Complexity by Output Type

A critical variable determining evaluation architecture is the nature of the output being assessed. Outputs fall into two broad categories:

- Structured outputs (SQL queries, classification labels, JSON payloads, routing decisions) can be evaluated with deterministic comparators, including precision, recall, exact match, and schema validation. These are faster, cheaper, and more reliable to evaluate.

- Generative outputs (long-form answers, research summaries, recommendations) require probabilistic evaluation, such as LLM-as-judge, human review, and rubric-based scoring. These are slower, more expensive, and require calibration to be trustworthy.

Several deployments in our dataset-built hybrid pipelines that handled both: deterministic evaluation for tool calls, routing decisions, and query generation, combined with LLM-as-judge for final response quality.

SECTION 02

Case Studies in Enterprise Evaluation

2.1 Financial Services: Proxy Voting Intelligence

A global financial services infrastructure provider deployed an AI agent to generate proxy voting recommendations (FOR, AGAINST, ABSTAIN, HUMAN REVIEW) across thousands of annual meeting items. The architecture paired with a recommendation agent with a dedicated LLM judge agent that validated every decision for policy alignment, exception handling, and data consistency.

The evaluation challenge was acute: recommendations carry fiduciary weight, and errors, such as missed exceptions or policy misalignments, have direct governance consequences. The team addressed this through several mechanisms:

- Pre-production benchmarking against historical analyst decisions across all four recommendation categories
- Regression testing triggered by any change to prompts, policies, model versions, or underlying data processing logic
- Continuous production monitoring of human override rates and judge-agent disagreement rates as leading indicators of quality drift

Evaluation Metric Spotlight

Human override rate proved to be the most actionable production metric. When analysts consistently overrode the agent on a particular policy class, the team treated this as a signal for targeted prompt and benchmark improvement, not just a compliance event.

2.2 Life Sciences: Pharma-Grade Agentic Platforms

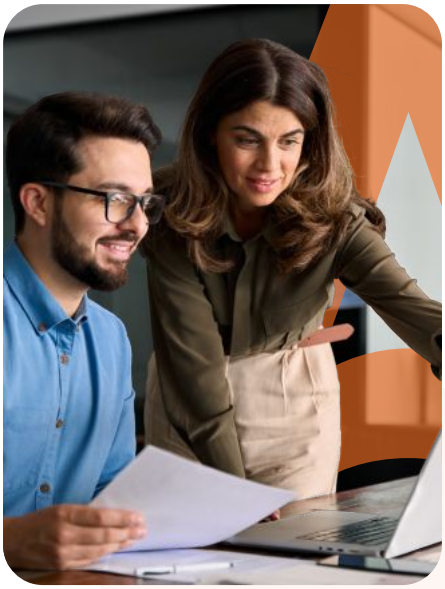
Two separate life sciences deployments illustrate different facets of evaluation in regulated environments.

The first, a custom multi-agent orchestration platform for scientific knowledge management, operated under pharmaceutical data governance requirements. The evaluation framework spanned three phases instrumented through MLflow, AWS, and Databricks Lakehouse Monitoring. Human-in-the-loop (HITL) validation was built into the architecture, not bolted on.

The second deployment focused on document intelligence for pharmaceutical manufacturing records, including handwritten batch records, inspection reports, and multilingual process documentation. This is a domain where evaluation has a clear ground truth: structured field extraction either matches the source document or it does not. The team built a purpose-built evaluation framework with the following results:

Industry / Use Case	Key Eval Metrics	Framework Used
Pharma Document Intelligence	Document Classification Accuracy: 96.8% Critical Field Accuracy: 98.2% Human Review Escalation: 8.4%	Custom Framework + OpenTelemetry
Scientific Knowledge Platform	Faithfulness, Answer Relevance, Citation Accuracy (RAG Triad) MLflow-tracked across all agents	MLflow + RAGAS + Bedrock LLM Judge

2.3 Investment Management: Multi-Domain Agentic Research



A global investment management firm deployed three distinct agentic platforms: an event-driven research platform, a sales analytics application, and an NL2SQL platform, each requiring a different evaluation approach.

The research platform evaluated on retrieval precision, source diversity, event coverage, and hallucination rate. The sales analytics application benefited from a strong existing ground truth: the sales and marketing teams could validate against known business outcomes, enabling test-driven development from the outset. The NL2SQL platform required a dual-layer evaluation: semantic retrieval quality at the input stage, and SQL execution correctness at the output stage.

This multi-application context illustrates a broader principle: evaluation frameworks must be purpose-built for each application's output type and domain context, even within a single organization.

2.4 Cybersecurity: Adversarial Evaluation for Vulnerability Remediation

A global financial data and analytics provider deployed a LangGraph-based multi-agent system to address the 20% of CVEs that have no available patch in enterprise vulnerability management platforms. The system engineered custom remediations through an adversarial author-reviewer loop and tested them in ephemeral lab VMs before routing to human approval.

The evaluation program for this deployment represents the most rigorous in the dataset. Four explicit pre-production thresholds were defined:

Classification accuracy (Vendor Knowledge Agent):

>95% match

Reviewer agreement (Script Author + Reviewer agents):

<10% false flag rate

Lab pass-rate (end-to-end Path B):

>70% first-attempt pass

Catalog hit precision (Catalog Retriever):

>98% precision

Critically, failed regressions required explicit engineer acknowledgment - no silent threshold bumps. Threshold changes required Infosec review. This governance layer around the evaluation framework itself is a distinguishing characteristic of high-stakes deployments.

2.5 Executive Search: Multi-Source Query Intelligence

A global executive search and leadership advisory firm deployed a multi-agent MCP-based system connecting to Elasticsearch and Neo4j, converting natural language queries into data source-native queries and returning structured or natural language responses.

The evaluation program reflected the multi-step nature of the agent pipeline, with distinct metrics for each stage:

1. **Routing accuracy:** did the query reach the correct agent and data source?
2. **Query translation quality:** was the natural language query correctly converted to Elasticsearch DSL or Cypher?
3. **Data retrieval precision:** did the query return the expected records?
4. **Response formatting:** did the output match the required structure?

This stage-by-stage decomposition, evaluating the pipeline rather than just the final output, is a pattern that consistently appeared in sophisticated deployments. Final output quality is a lagging indicator; stage-level metrics enable root cause analysis.

2.6 Enterprise Software: RAG with Authorization-Aware Evaluation

A B2B pricing optimization software company deployed a LangGraph multi-agent RAG system connecting to SharePoint, Confluence, GitHub, and Salesforce. A distinctive requirement was Role-Based Access Control (RBAC) compliance: users should only receive information they are authorized to access.

This requirement introduced a novel evaluation dimension, authorization compliance, that sits outside the standard RAG evaluation stack. The team achieved 100% authorization compliance in their evaluation benchmark, making it a hard gate in the regression pipeline. Selected production metrics:

Industry / Use Case	Key Eval Metrics	Framework Used
Enterprise RAG (Multi-Source)	Retrieval Precision@5: 92% Grounded Response Rate: 95% Hallucination Rate: 3.2%	Custom + LangSmith
Authorization Compliance	100% (hard regression gate)	Custom Evaluation Pipeline
User Experience	Satisfaction Score: 4.6/5 Mean Response Latency: 4.1s	Production Monitoring

SECTION 03

Tooling Decisions and Trade-offs

3.1 The Build vs. Buy Decision for Eval Frameworks

The most common question practitioners face is whether to adopt an off-the-shelf evaluation framework or build custom pipelines. The data from our engagements suggests the answer is highly contextual:

Approach	When It Works	Trade-offs
LangSmith	Tracing-heavy workflows, LangGraph deployments, teams needing observability out-of-the-box	Less control over scoring logic; external dependency
MLflow + RAGAS	Data platform-centric stacks (Databricks/AWS), regulated industries needing audit trails	Requires instrumentation investment; RAGAS calibration needed
Custom Pipelines	Novel output types (proxy voting, NL2SQL, document extraction), strict governance requirements, domain-specific metrics	Higher build cost; requires eval engineering expertise
Promptfoo (with OpenAI Evals)	Teams needing prompt regression testing, model comparison across providers, and adversarial red teaming in CI pipelines; the recommended evaluation framework for OpenAI SDK-based stacks. OpenAI Evals remains available as a legacy option for teams already invested in it.	CLI-first; less suited for rich observability or production monitoring. Acquired by OpenAI in March 2026, making it the de facto standard for OpenAI-ecosystem teams

3.2 LLM-as-Judge: Strengths and Pitfalls

LLM-as-judge, the practice of using a language model to score the outputs of another language model, appeared in the majority of generative evaluation pipelines in our dataset. Its primary advantage is scalability: human review at production volumes is cost-prohibitive. But LLM judges introduce their own failure modes:

Positional bias

judges score earlier answers higher, independent of quality

Verbosity bias

judges favor longer, more detailed responses

Self-consistency

judges from the same model family as the evaluated model may be correlated

The teams that used LLM-as-judge most effectively calibrated their judges against a labeled human sample before deployment, and periodically re-calibrated as models were updated. One team explicitly noted that LLM judge scores were only trusted after calibration. Uncalibrated scores were treated as directional, not definitive.

Tools like Promptfoo make this deterministic/probabilistic split explicit in their architecture, supporting both hard assertions (exact match, schema validation, substring checks) and LLM-as-judge scoring within the same test suite. This allows teams to apply the right evaluation method per output type rather than forcing a single approach across an entire pipeline. Promptfoo also includes a red teaming module that generates adversarial inputs to test for jailbreaks, prompt injection, and data leakage - a capability not covered by most other frameworks in this comparison and particularly relevant to customer-facing deployments.

SECTION 04

Emerging Principles for Enterprise AI Evaluation

Synthesizing across all eight engagements, the following principles characterize mature enterprise AI evaluation programs:

- Stage-level metrics, including routing accuracy, query translation quality, and retrieval precision, provide earlier signals than final output quality scores.Evaluate the pipeline, not just the output.
- In every high-stakes deployment in our dataset, evaluation metrics served as hard gates in CI pipelines. Failed evals blocked deployment; threshold changes required explicit review. Make evaluation a deployment gate.
- Teams with strong existing ground truth - historical analyst decisions, validated business datasets, and SME-reviewed records, consistently shipped faster and with higher confidence than those constructing ground truth from scratch. Ground truth is your most valuable asset.

- Production monitoring that feeds back into pre-production benchmarks prevents evaluation drift. Teams that treated evaluation as a one-time pre-launch activity consistently underperformed. Design for the closed loop from day one.

- In regulated industries and enterprise settings, compliance metrics (RBAC adherence, audit trail completeness, HITL escalation rates) belong in the evaluation framework alongside accuracy metrics. Authorization and governance are first-class eval dimensions.

- Human override rates, escalation rates, and disagreement rates between agents and reviewers are more actionable in production than aggregate quality scores. They surface where the model is systematically wrong. Human review rates are leading indicators.

Conclusion

The deployment of agentic AI systems at enterprise scale demands an evaluation discipline that matches the sophistication of the systems being deployed. The engagements analyzed in this paper demonstrate that evaluation is not a checkbox activity. It is an ongoing engineering program that spans pre-production benchmarking, continuous regression testing, and closed-loop production monitoring.

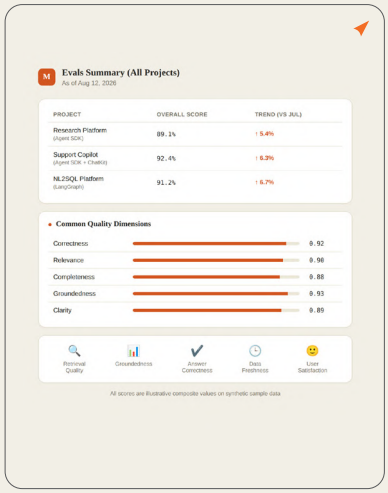
What separates the best programs is not the sophistication of their tooling, but the quality of their thinking about what to measure, how to measure it, and what to do when measurements degrade. As AI systems take on higher-stakes decisions such as proxy votes, vulnerability remediations, pharmaceutical documentation, and executive hiring intelligence, the field's ability to answer those questions will determine whether enterprise AI fulfills its promise.

The frameworks and patterns described here represent the current frontier of enterprise AI evaluation practice. They are not final answers, as the field is evolving rapidly, but they offer a substantive starting point for any organization serious about deploying AI responsibly and at scale.

Appendix A:
Sample Evaluation Report

Global Investment Management Firm - Agentic AI Evaluation Dashboard (May 2025) The following evaluation dashboard illustrates the three-platform evaluation program described in Section 2.3. It covers the Agentic Research Platform, Sales Analytics application, and NL2SQL platform, alongside a consolidated results explorer spanning 5,090 test cases across all three platforms. Client identifying information has been removed.

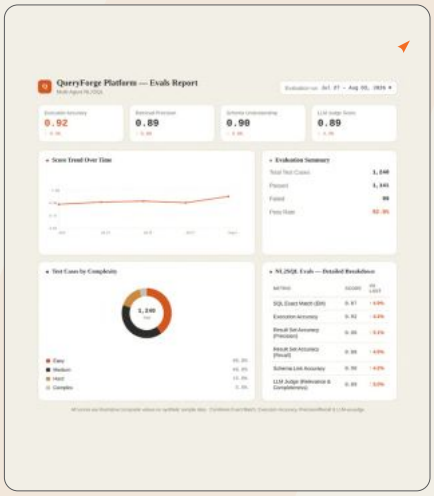

Meridian
Overview
Retrieval Quality
Generation Quality
Source & Freshness
Custom Metrics
Traces
Alerts
Agentic Research Platform — Evals Report
Range Jul 27 – Aug 03, 2026
Overall Evals Score
89.1%
Retrieval Precision
88.2%
Event Coverage
85.1%
Groundedness
91.3%
Hallucination Rate
2.8%
Retrieval Quality
88.2%
Retrieval Precision
Custom Metrics (Events)
Source Diversity 0.87
Freshness (≤24h) 0.81
Deduplication Quality 0.90
Event Ranking Accuracy 0.84
Citation Accuracy 0.89
Generation / Analysis Quality
Event Coverage vs Top News Feeds (N=15)
Hallucination Trend
Hallucination Rate


Evals Summary (All Projects)
As of Aug 12, 2026
PROJECT
OVERALL SCORE
TREND (VS JUL)
Research Platform
(Agent SDK)
89.1%
↑ 5.4%
Support Copilot
(Agent SDK + ChatKit)
92.4%
↑ 6.3%
NL2SQL Platform
(LangGraph)
91.2%
↑ 6.7%
Common Quality Dimensions
Correctness 0.92
Relevance 0.90
Completeness 0.88
Groundedness 0.93
Clarity 0.89
Retrieval Quality
Groundedness
Answer Correctness
Data Freshness
User Satisfaction
All scores are illustrative composite values on synthetic sample data

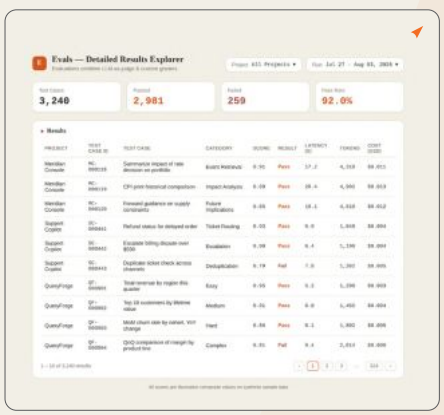

NL2SQL Platform — Evals Report
Multi-Agent NL2SQL
Range Jul 27 – Aug 03, 2026
Overall SQL Accuracy
91.2%
Execution Success Rate
94.6%
Precision (Column/Filter)
0.90
Recall (Intent Coverage)
0.87
Avg Latency
3.18s
SQL Accuracy by Intent Type
Aggregation
Filtering
Joins
Time series
Group By
Accuracy Trend
LLM-as-a-Judge (Answer Quality)
Correctness
Relevance
Completeness
Groundedness
Clarity
Top Failing Cases
USER QUESTION (SAMPLE)
SQL ACCURACY
ERROR TYPE
Show top 10 products by AUM last quarter
89%
Column mismatch
Total inflows by region for FY2026
86%
Filter condition
Average fee by fund type
81%
Join missed
Execution Error Breakdown
47
Syntax Error
Semantic Error
Timeout
Other
All scores are illustrative composite values on synthetic sample data


Sales Analytics Agent — Evals Report
Agent SDK + Chat interface
Range Jul 27 – Aug 03, 2026
Overall Answer Accuracy
92.4%
CRM Query Accuracy
95.7%
Custom MCP Retrieval Precision
89.1%
Parameter Extraction Accuracy
93.2%
Avg Latency
2.34s
By Data Source
1,412
CRM (SOQL)
Custom MCP (Semantic)
Other Tools
Custom MCP — Retrieval Quality
Precision
Recall
Freshness (≤24h)
Source Relevance
CRM (SOQL) — Answer Accuracy
Top Business Questions — Accuracy (Ground Truth)
QUESTION CATEGORY
CALLS
ACCURACY
AUM by Product / Region
312
96.2%
Sales by Channel / Rep
276
94.1%
New Business / Inflows
198
93.9%
Pipeline / Opportunities
121
92.1%
Fees & Revenue
189
89.3%
Failure Analysis
78
Retrieval Miss
SOQL / Tool Errors
Ambiguous Query
Other
User Feedback (Thumbs)
Positive
Neutral
Negative
All scores are illustrative composite values on synthetic sample data

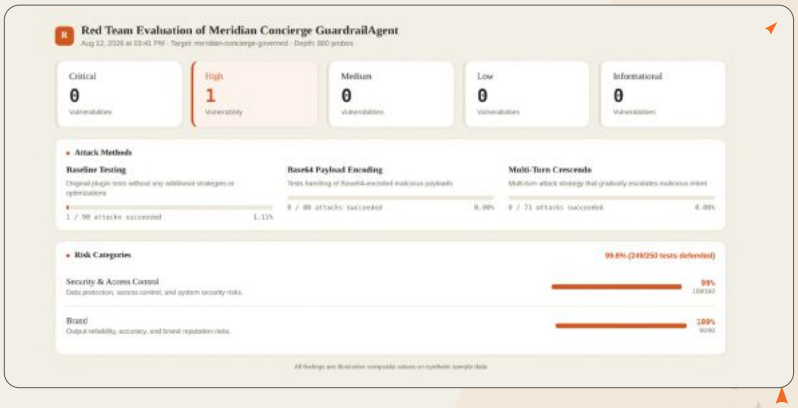

Continuous Evaluation — Production Health
Real-time Monitoring
Range Last 7 Days
Requests
14.9K
Evals Run
11.4K
Alerts
21
Hallucination Rate
2.6%
Retrieval Failure Rate
3.2%
Quality Over Time
Accuracy
Groundedness
Retrieval Precision
Alert Distribution
CATEGORY
COUNT
Low Groundedness
10
Retrieval Miss
6
Data Freshness
4
Tool Failure
1
Recent Evals (Sample)
TIME
QUERY (SAMPLE)
SCORE
STATUS
Aug 12, 02:10 PM
AUM by product for Q1 FY2026
94%
Pass
Aug 12, 01:42 PM
Top events impacting tech sector
90%
Pass
Aug 12, 12:57 PM
Inflows by region this month
88%
Pass
Aug 12, 11:33 AM
Top 10 funds by fees
92%
Pass
Aug 12, 10:08 AM
Show sales by rep last week
85%
Fail
All values are illustrative composite figures on synthetic sample data


QueryForge Platform — Evals Report
Multi-Agent NL2SQL
Execution Accuracy
0.92
Result Set Accuracy
0.89
Query Understanding
0.90
LLM Judge Score
0.89
Score Trend Over Time
Evaluation Summary
Test Cases by Complexity
1,240
Easy
Medium
Hard
Complex
NL2SQL Evals — Detailed Breakdown
METRIC
SCORE
VS LAST
SQL Exact Match (EM)
0.87
+0.04
Execution Accuracy
0.92
+0.03
Result Set Accuracy (Precision)
0.88
+0.05
Result Set Accuracy (Recall)
0.86
+0.04
Schema Link Accuracy
0.90
+0.02
LLM Judge (Relevance & Completeness)
0.89
+0.06


Meridian Console — Evals Report
Agent SDK + ChatKit + MCP
Overall Score (Weighted)
0.90
Answer Correctness
0.86
Event Coverage (Recall)
0.81
Groundedness
0.93
Hallucination Rate
3.4%
Retrieval Precision
0.85
Score Trend Over Time
Event Coverage by Issuer Type
9,481
Custom Evals — Key Metrics


Evals — Detailed Results Explorer
Evaluations combine LLM-as-judge & custom graders
Project: All Projects
Run: Jul 27 – Aug 03, 2026
Test Cases
3,240
Passed
2,981
Failed
259
Pass Rate
92.0%
Results
PROJECT
TEST CASE ID
TEST CASE
CATEGORY
SCORE
RESULT
LATENCY (S)
TOKENS
COST (USD)
Meridian Console
RC-000139
Summarize impact of rate decision on portfolio
Event Retrieval
0.91
Pass
17.2
4,310
$0.011
Meridian Console
RC-000140
CPI print historical comparison
Impact Analysis
0.89
Pass
18.4
4,560
$0.013
Meridian Console
RC-000139
Forward guidance on supply constraints
Future Implications
0.88
Pass
18.1
4,410
$0.012
Support Copilot
SC-000841
Refund status for delayed order
Ticket Routing
0.95
Pass
9.8
1,840
$0.004
Support Copilot
SC-000842
Escalate billing dispute over $500
Escalation
0.98
Pass
8.4
1,190
$0.003
Support Copilot
SC-000843
Duplicate ticket check across channels
Deduplication
0.79
Fail
7.9
1,360
$0.004
QueryForge
QF-000801
Total revenue by region this quarter
Easy
0.96
Pass
6.2
1,090
$0.003
QueryForge
QF-000802
Top 10 customers by lifetime value
Medium
0.91
Pass
8.9
1,450
$0.004
QueryForge
QF-000803
Month churn rate by cohort, YoY change
Hard
0.84
Pass
9.1
1,980
$0.005
QueryForge
QF-000804
QoQ comparison of margin by product line
Complex
0.81
Fail
9.4
2,010
$0.005


Support Copilot — Evals Report
Agent SDK + ChatKit + MCP
Overall Score (Weighted)
0.88
Answer Correctness
0.90
Ticket Routing (CRM MCP)
0.91
Retrieval Precision (Custom MCP)
0.85
Freshness
0.87
User Satisfaction
0.89
Performance Over Time
Evaluation Coverage
Total Test Cases
1,505
Ticket Routing (CRM MCP)
812
Custom MCP (Semantic Search)
784
Pass Rate
91.8%
Test Case Distribution
1,505
Common Requests
Ad-hoc Questions
Edge Cases
Negative / No-Data
Custom MCP — Retrieval Evals
METRIC
SCORE
VS LAST
Retrieval Precision
0.85
-1.5%
Recall (Coverage)
0.90
-2.7%
Freshness (Timeliness)
0.87
-2.0%
Source Relevance
0.84
-3.5%
Semantic Match Quality
0.83
-3.2%
CRM MCP — Routing Evals
METRIC
SCORE
VS LAST
Intent Classification
0.91
-2.8%
Priority Assignment
0.88
-3.1%
Escalation Accuracy
0.86
-2.9%
Field Mapping Accuracy
0.88
-3.0%
Duplicate Detection
0.84
-3.6%


Red Team Evaluation of Meridian Concierge GuardrailAgent
Aug 12, 2026 at 09:41 PM · Target: meridian-concierge-governed · Depth: 000 probes
Critical
0
Vulnerabilities
High
1
Vulnerability
Medium
0
Vulnerabilities
Low
0
Vulnerabilities
Informational
0
Vulnerabilities
Attack Methods
Baseline Tissing
Original plugin tests without any additional strategies or optimizations
1 / 90 attacks succeeded
1.11%
Base64 Payload Encoding
Tests handling of Base64-encoded malicious payloads
0 / 00 attacks succeeded
0.00%
Multi-Turn Crescendo
Multi-turn attack strategy that gradually escalates malicious intent
0 / 71 attacks succeeded
0.00%
Risk Categories
99.6% (249/250 tests defended)
Security & Access Control
Data protection, access control, and system security risks.
99%
Brand
Output reliability, accuracy, and brand reputation risks.
100%
All findings are illustrative composite values on synthetic sample data

About This Paper

This white paper is based on anonymized analysis of agentic AI evaluation programs implemented across eight enterprise clients spanning financial services, life sciences, investment management, cybersecurity, and enterprise software sectors. All client names and identifying details have been removed or generalized. Quantitative metrics are drawn directly from documented evaluation reports.
The authors would like to thank Shikhar Kwatra & team at OpenAI for their guidance and thought partnership on AI evaluation practices and tooling.

Building What's Next