

Building enterprise-ready AI agents on synthetic evaluation data

How Moolai uses Rockfish to generate evaluation datasets and ground-truth answers for testing agents that have no production data to learn from

Customer	Moolai - early-stage AI agent platform, San Ramon, CA
Industry / focus	Enterprise AI agents; focus on EPM (enterprise performance management), the category occupied by Workday and SAP
Use case	Synthetic evaluation datasets and expected-outcome (ground-truth) generation for testing AI agents before production
Rockfish products	SchemaFuel (schema-based generation), Scenario Studio (incident injection), AgentFuel (expected outcomes)
Status	Active POC; Rockfish is embedded as the data-generation backend of Moolai's platform.

Summary

Moolai is building a platform for enterprises to create, test, and deploy AI agents, with current focus on EPM (enterprise performance management) systems. Before an agent goes to production, it has to be evaluated - but Moolai's enterprise customers can't share sensitive production data, and Moolai has none of its own to test against. **Rockfish closes that gap:** Moolai uses Rockfish to generate realistic, domain-specific evaluation datasets, inject edge-case incidents, and produce the expected (ground-truth) answers needed to score an agent's responses. Rockfish now runs as the data-generation backend inside Moolai's own platform.

About Moolai

Moolai is an early-stage startup based in San Ramon, California, building a platform for enterprises to build and ship AI agents, the software category occupied by vendors like Workday, Pigment, Anaplan and SAP. The team's background is deep in this space: the founder previously led machine learning at Workday.

Moolai's product lets an agent designer define an agent conversationally, generate data to test it, run experiments, evaluate the results, and deploy the agent as a self-contained container - think of it as "WordPress for agents." Learn more at <https://moolai.ai>

The Challenge

Agents must be proven before they reach production - but the data to prove them doesn't exist. Enterprise customers' data is sensitive and unavailable for testing, and as a young company Moolai has no production data of its own to fall back on. To evaluate an agent's accuracy, the team needed three things they couldn't easily get:

Realistic, domain-specific datasets (e.g., EPM or sales-performance data) to run an agent against.

Edge cases, not just the happy path - anomalies and incidents that expose where an agent hallucinates or fails.

Trustworthy ground-truth answers - a reliable "correct answer" for each test question so agent responses can be scored objectively.

The Rockfish Approach

Rockfish gave Moolai a single backbone for evaluation data - and the expected answers needed to score against it. Working on Rockfish's SchemaFuel, Scenario Studio, and AgentFuel tooling, the team stood up a complete generate-to-score pipeline for a representative domain:

1. **Schema-based generation.** Moolai can stand up a complete domain specific dataset from a schema alone, without needing real customer records to start from. (SchemaFuel)
2. **Breadth across data types.** Generate tabular, time-series, relational/JSON, financial, and network data through one backbone.
3. **Incident injection.** Overlay the edge cases that break agents - spike, outage, sustained shift, and ramp (Scenario Studio).
4. **Expected-outcome generation.** Produce the correct answer alongside each question, giving a scoring oracle for every test case (AgentFuel).

How Moolai uses Rockfish

Moolai embedded Rockfish as the data-generation backend of its own platform - supporting both manual generation (the user configures each column) and automated generation (driven by the agent designer's configuration through the Rockfish API), with an analyst- and evaluation-agent layer scoring whether agents follow tool-use instructions correctly.

Results

Running the loop on Rockfish, Moolai proved out domain-specific agent evaluation end to end - with no production, seed, or customer data:

- **Evaluation data for a new domain, from a schema alone.** Moolai stood up complete datasets without seed records, data-access agreements, or waiting on data it had not yet collected - removing the data bottleneck that normally blocks evaluation.
- **One evaluation backbone across five data types.** Tabular, time-series, relational/JSON, financial, and network data all generated cleanly end to end - enough for Moolai to standardize on a single evaluation tool rather than stitching together several specialized ones.
- **Agent behavior characterized under real failure modes.** Four injected incident patterns - spike, outage, sustained shift, and ramp - exposed how agents respond to the conditions that drive hallucination and incorrect tool use, not just the happy path.
- **Every test case scored against ground truth.** AgentFuel's generated expected answers replaced manual grading with an automated scoring oracle, making objective, repeatable scoring possible.

The result is a complete evaluation loop that runs from nothing to a scored agent - turning “we believe the agent is ready” into “we have evidence it is ready,” within a customer engagement window.

The Impact

5

data types validated end to end -
tabular, time-series,
relational/JSON, financial, network

0

seed, production, or customer data
required to start

1

generate-to-score loop - from a
schema to a scored agent

- **Embedded, recurring usage.** Rockfish is now a standing step in Moolai's agent lifecycle (define → generate → experiment → evaluate → deploy), firing on every agent build and every new test version - architectural use rather than a one-time test pass.
- **Joint engineering collaboration.** The teams created a shared public repository (“tacklebox”) of reusable scripts for the Rockfish SDK to avoid duplicate work and speed adoption.

What's next

- **Graded scoring.** Moving from boolean pass/fail to graded scores so Moolai can rank agent configurations and track improvement over time.
- **Customer-supplied ground truth.** Letting domain experts provide the expected answers for cases where generation shouldn't infer them - important for financial/EPM data.
- **Data lineage tracking.** Tracing each incident-injected dataset back to its baseline, identified jointly as a valuable future capability.
- **Beyond flat time-series.** Support for multi-dimensional hierarchies and period-over-period comparisons common in business-metrics data.
- **Design-partner rollout.** Moolai is targeting a rollout to friendly customers, with Rockfish embedded underneath.

About Rockfish

Rockfish Data, built on foundational research from Carnegie Mellon University, generates high-coverage, labeled datasets purpose-built for evaluating AI models and data analytics agents in production. Unlike generic benchmarks, Rockfish systematically covers edge cases, rare scenarios and domain-specific failure modes - giving teams the evaluation infrastructure to measure accuracy and reliability before deployment. Learn more at <https://rockfish.ai>