SPARK BEYOND

Introducing Agentune

OPEN-SOURCE AI AGENT OPTIMIZATION LIBRARY

Open-Source AI Agent Optimization Framework

Structured Cycle for Continuous Improvement

Simulated Environments for Risk-Free Testing

Insight-Driven Performance Enhancements

From Generic LLMs to Enterprise-Ready Agents

TL;DR

Despite powerful LLMs and slick frameworks, real-world agents often underperform—hallucinating policies, misusing tools, or mishandling edge cases—because they haven't been optimized for messy, domain-specific realities. Moreover, agents underperform when they don't learn from their own success and failure.

Agentune is an open-source engine that brings structure to agent performance through a disciplined Analyze → Improve → Evaluate cycle. It treats agents like teammates: scoring real and simulated interactions, mining transcripts for root causes, and iteratively shipping targeted improvements.

Last week we released the first Agentune module, Agentune-Simulate which addresses the essential challenge in Evaluate for agents: it lets teams evaluate agents safely in the lab using synthetic customers and edge-case stress tests. Coming soon, Agentune-Analyze and Insight-Eval will apply SparkBeyond's proven insight discovery methods to uncover and validate the true drivers of agent behavior. The mission: transform raw LLM output into finely tuned, high-performing agents - at machine speed, with open-source transparency.

Why Great LLMs Still Ship Mediocre Agents?

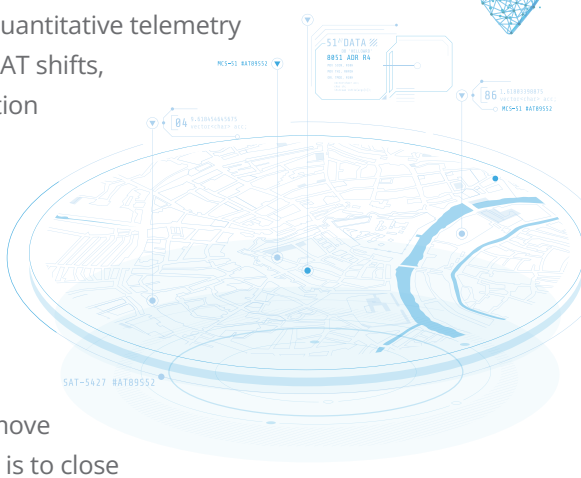
Despite powerful LLMs and slick agent frameworks (LangGraph, AutoGen, DSPy, Guardrails, etc.), customer-facing agents start far from optimal because real performance hinges on messy, domain-specific realities they haven't seen or been tuned for. Prompts that ace sandbox evals crumble on live edge cases (billing disputes, partial refunds, regional regulations). Tool use is brittle—agents call the CRM API with the wrong ID schema, or never call the pricing calculator when discounts matter. They hallucinate policy ("We can waive that fee") or over-escalate to humans to stay safe. Tone control and compliance drift across long chats; a sales bot pushes the flagship plan to a student on a budget, while a support bot skips authentication steps. Small upstream shifts—LLM version updates, KB changes, new product SKUs—quietly degrade behavior. Until you measure, dissect transcripts, and iteratively coach them, these gaps stay invisible—and costly.

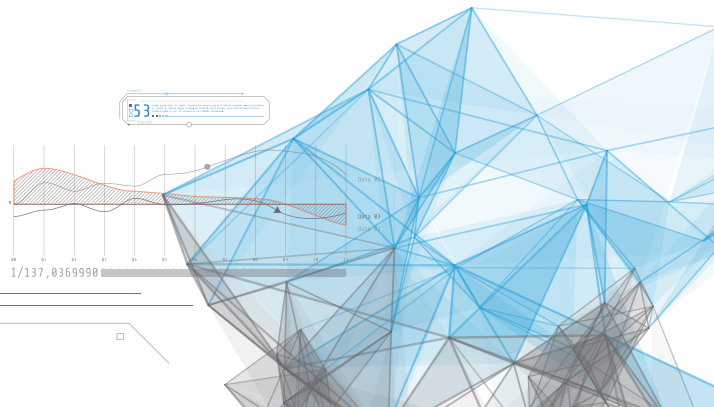
Where Mediocre Outputs Meet Metrics: Enter Agent Optimization

The gap between powerful models and mediocre frontline behavior is exactly where **agent optimization** lives. Once you see how often transcripts expose missed tools, policy slips, and tone misfires, it's clear you need an explicit discipline to close that gap—instrumenting agents like products, not prompts. Optimization turns raw interaction data (metrics + conversations/recordings) into hypotheses about what to tweak, then tests those tweaks fast. In other words: we don't just accept "good enough LLM output"; we continuously coach the agent, just as we would a human rep—only now with tighter loops and far more data.

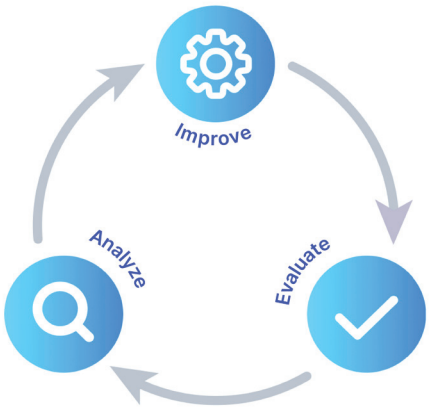
What's Agent Optimization?

AI agent optimization is the disciplined practice of managing customer-facing bots—sales reps, support agents—like performance-tracked teammates. Every interaction emits two rich data streams: quantitative telemetry (conversion rate, CSAT shifts, handle time, escalation paths) and the full conversation text or recording itself. Together, these let you see which prompts, tools, skills, and behaviors actually move the metrics. The job is to close the loop: instrument the agent, mine transcripts for patterns and root causes, run controlled changes (new skills, guardrails, reasoning styles, data access), and ship improvements fast while preventing regressions. In short, you analyze what drives outcomes, teach the agent new tricks, prune bad habits, and continuously coach it—exactly as you would a human team, just at machine speed and scale.





Analyze → Improve → Evaluate



Analyze → **Improve** → **Evaluate** is a tight feedback loop for agent performance. Analyze takes hard KPIs (conversion, CSAT, FCR, AHT) and soft signals from transcripts/recordings (tone, compliance, hallucination flags), clusters and tags those conversations, correlating behaviors and tool usage with outcomes, and isolates root causes—what prompts, skills, or routing choices helped or hurt. Improve by shipping targeted fixes: refine prompts and guardrails, add or revoke tools/knowledge, retrain skills, adjust policies or handoff logic—then Evaluate by running simulations, A/B testing and monitoring for regressions. As the loop is now complete, feed the new data back into the next analysis pass. Repeat until the curve flattens, then raise the bar.

The Problem with Intuition-driven Tuning

- "Gut feel" changes introduce biases and blind spots, not solutions.
- Anecdotal improvements rarely scale—data, not intuition, defines success.
- Unmeasured adjustments risk false confidence—hidden failures remain unchecked.
- Effective AI optimization demands disciplined, systematic experimentation.

As Heraclitus observed, "No man ever steps in the same river twice." In today's fast-paced business environment, no optimization solution remains optimal for long. Continuous adaptation and learning are now essential for success.

The transition from static efficiency to dynamic optimization marks a pivotal moment in business history. By embracing AI-driven continuous optimization, organizations can achieve unprecedented levels of agility and resilience—ensuring they remain competitive in an ever-changing world.

Adopting a Scientific Optimization Mindset

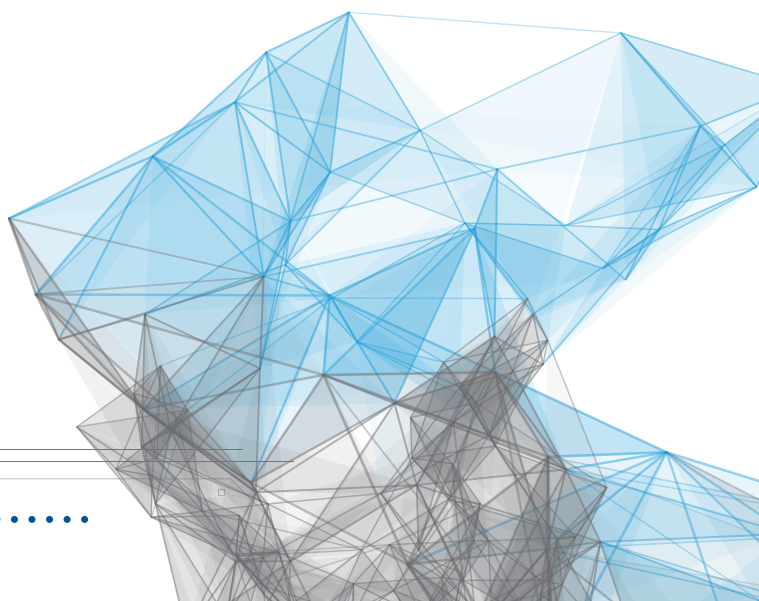
In an increasingly complex and data-rich world, objective decisions, structured experimentation, rapid iteration, and data-backed results are paramount for sustained growth and efficiency.

Data-driven decisions eliminate bias and assumptions, transforming decision-making into a science based on verifiable facts and quantifiable metrics. This leads to more accurate predictions and effective strategies.

Structured experimentation, using methodologies like A/B testing, turns random improvements into predictable progress by isolating variables and measuring the precise impact of interventions. This systematic approach ensures sustainable improvement.

Rapid iteration through controlled testing accelerates meaningful improvements by enabling quick deployment of new features, real-time feedback, and immediate adjustments. This fosters continuous learning and adaptation to evolving market demands.

Data-backed results create organizational buy-in and measurable ROI. Quantifiable evidence of increased revenue, reduced costs, or enhanced customer satisfaction builds trust and confidence, paving the way for future investment and expansion of data-driven practices.



Synthetic Customers, Real Metrics: Simulation-Driven Agent Evaluation

Evaluating customer-facing agents starts with rigorous, always-on measurement: hard KPIs (conversion, CSAT proxy scores, FCR, AHT, cost/interaction) plus qualitative rubrics for tone, compliance, hallucinations, and tool usage. You score each turn and whole conversations, ideally with a blend of human review and LLM "judges" calibrated against humans. Instrumentation should capture not just outcomes but decision traces—what the agent knew, which tools it called, why it escalated—so you can attribute success or failure to specific behaviors. Batch replays of historical transcripts against new policies or prompts let you estimate uplift before risking production traffic.

A simulated world—especially one that models customer behavior—lets you run those evaluations safely in the lab. You spin up synthetic customers with goals, constraints, emotions, and randomness (impatient churn risk, budget shoppers, policy abusers), backed by a product/catalog/policy "world model" that enforces realities like inventory or refund rules. Multi-turn scenarios, noise injections (typos, contradictory info), and edge-case generators stress-test reasoning, tool orchestration, and guardrails. Because you control the distribution of scenarios, you can oversample rare but costly failures, do A/B/C testing at scale, and iterate fast—then graduate only the best variants to real traffic.

From Optimizing Companies to Optimizing Agents

The **Analyze** step is basically large-scale hypothesis generation and driver discovery—the exact craft SparkBeyond has honed for over a decade. Since 2013, the company has built engines that automatically propose and test millions of candidate signals, features, and explanations, surfacing the real levers behind KPI movement and letting teams prioritize fixes with evidence.

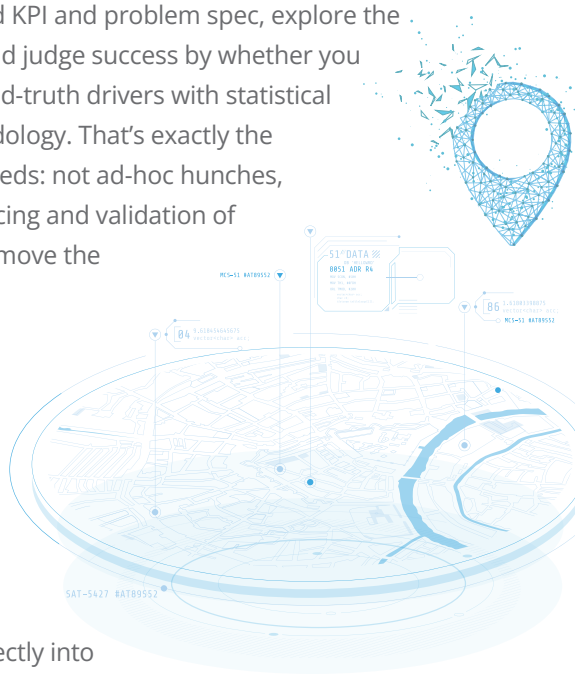
The kind of improvements we helped our partners drive:

- Cut churn ~30% in three months for a European media company by mining "thousands of clues" about why readers leave
- 600 new stores targeted by Zabka; SparkBeyond generated millions of hypotheses to optimize location & revenue per store

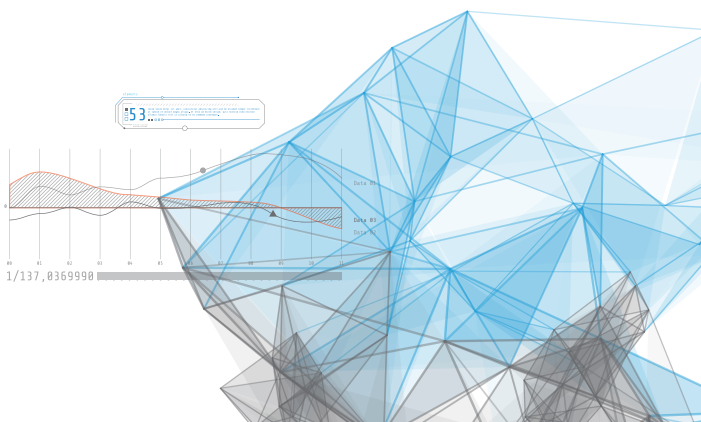
We believe in the world of agents, closing the loop can be much faster as the change management process is smoother.

Benchmarking Agent KPI driver discovery

In the Analyze step, we move from "how did the agent perform?" to "why did it perform that way?"—and the SparkBeyond benchmark gives us a ready-made template for answering that why. SparkBeyond's "insight discovery" benchmark formalizes what good analysis looks like: start with a clearly defined KPI and problem spec, explore the underlying tables, and judge success by whether you rediscover the ground-truth drivers with statistical lift and clean methodology. That's exactly the discipline Analyze needs: not ad-hoc hunches, but systematic surfacing and validation of factors that actually move the needle.



We can wire those same metrics—coverage of true insights, predictive power of the features built from them, and data hygiene checks—directly into our Analyze scorecard. Every time our agent proposes a hypothesis ("customers on plan X churn more after event Y"), we score it the way the benchmark does, creating an objective bar the Improve phase must beat on the next iteration. In short: the SparkBeyond framework becomes our unit test suite for Analyze, ensuring we're optimizing against verified insight quality, not just prettier dashboards.



Introducing Agentune

Agentune is our end-to-end engine for the Analyze → Improve → Evaluate cycle, but the first module you'll get hands-on with is Agentune-Simulate which we've just released. It lets you replay and perturb realistic customer behavior, so you can grade agents in a safe lab before they ever touch production. Think A/B tests, edge cases, and counterfactual "what ifs" on tap—so Evaluate is reproducible, and Analyze has rich, labeled traces to mine.

Agentune

- Simulate real-world customer conversations
- Highlight weak spots and failure points
- See clear metrics and intent analysis
- Run locally in seconds.

Whether you need a realistic virtual customer or full conversation simulations, **Agentune Simulate** has you covered.

How to use Agentune:

1. Upload 100+ real conversations from your AI or human agents
2. Agentune builds a simulated user model
3. Connect your AI Agent and simulate full dialogues
4. You get insights on what worked, what failed, and how to improve

Try it now: Agentune-Simulate is live!

Repo: github.com/SparkBeyond/agentune

Install: pip install agentune-simulate

What's next?

Over the next weeks and months, we will extend Agentune with a range of useful tools, starting with:

- Agentune-Analyze: a toolkit to uncover what drives agent performance and suggest actionable improvements.
- Benchmark for Agent Insight Discovery – a standard for evaluating how well systems identify and explain what's working (or not).
- The first release will tackle generic insight discovery challenges. Up next: analysis of agent conversations and structured data like CRMs.

Sergey Davidovich
Co-Founder & President

Dr. Ron Karidi
CTO & Co-founder
ron@sparkbeyond.com

About SparkBeyond

SparkBeyond is powering a new breed of market leaders, leveraging AI to accelerate the process of turning data into impact. By augmenting internal data with external sources and massively scaling the interrogation of data, SparkBeyond amplifies the discovery of hidden insights and drivers of positive outcomes. From risk scoring and fraud detection to demand forecasting and churn reduction, SparkBeyond helps global organisations drive tangible and lasting impact across a broad range of use cases. Learn more at www.sparkbeyond.com.