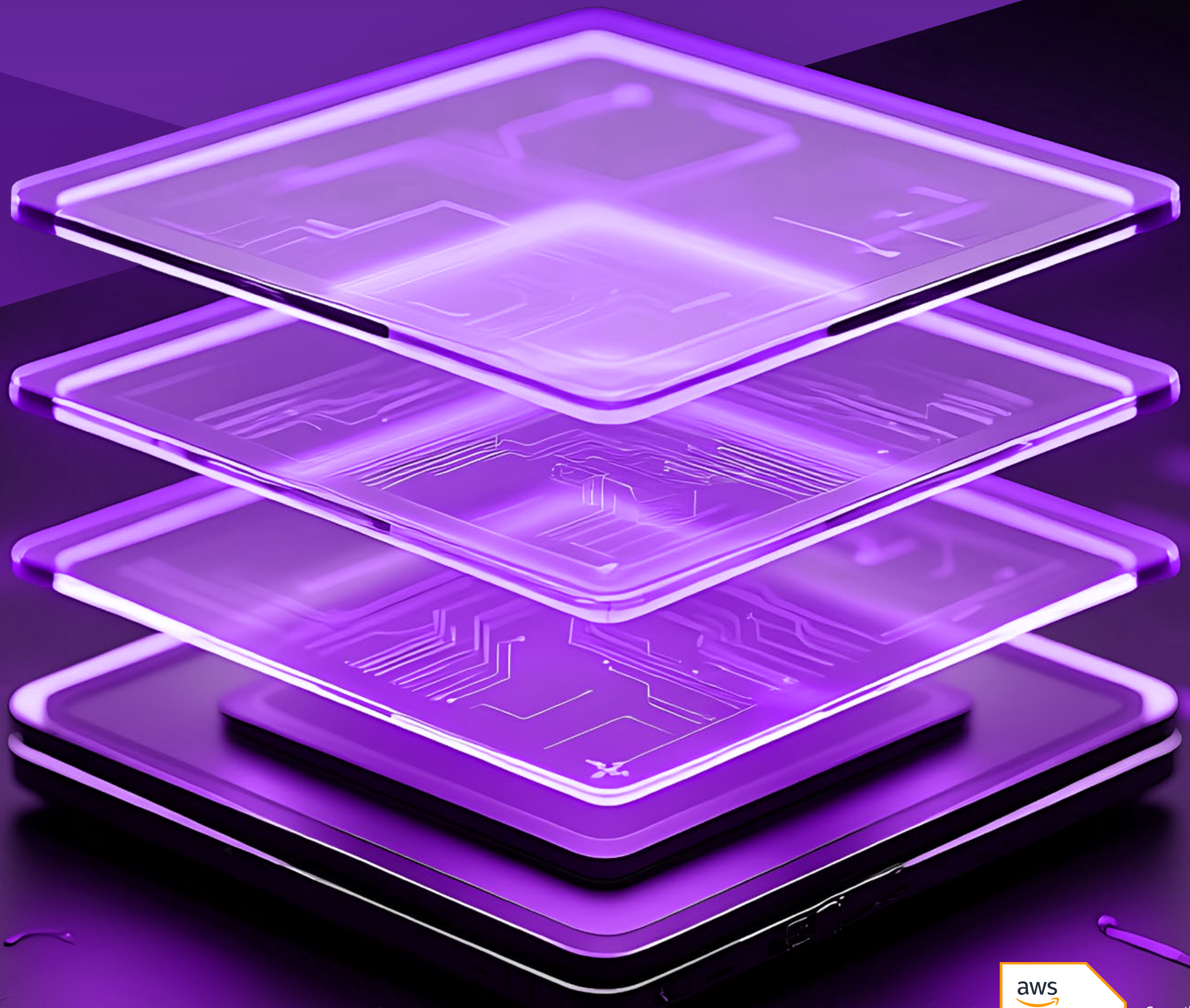
deepchecks.

Why AI agents fail and how to build them right

Best practices for building, evaluating, and productionizing agentic AI systems, and how to avoid the hidden traps that derail enterprise adoption.



aws PARTNER Generative AI Software Competency

• Generative AI Applications

Executive summary

AI is moving from experimentation to enterprise-scale deployment. For CXOs, the challenge is no longer whether to adopt agentic AI, but how to do so safely, reliably, and profitably. This ebook outlines the hidden traps that derail adoption, the evaluation strategies that prevent failure, and the organizational practices that turn AI into measurable business value.

At Deepchecks, we specialize in one thing: systematic evaluation. Our platform ensures that enterprises can scale AI confidently – with guardrails for compliance, continuous monitoring for drift, and the ability to link AI outcomes directly to business value.

This guide distills what we've learned from helping leading enterprises move from proof-of-concept to production. The goal is to equip leaders with the knowledge to act decisively, while showing how Deepchecks can accelerate the journey.

Contents

Part 1: Intro

Part 2: The strategic foundations for AI success

Part 3: The Future of Agentic AI

Introduction

Why now and why AI agents matter

Agentic AI systems promise more than efficiency gains. Done right, they can rewire workflows, reduce compliance risk, and unlock entirely new opportunities. But why now? What makes this moment different?

Two forces are converging. On the technology side, generative AI has matured to the point where production-grade deployments are possible. Foundation models are not only more powerful, but also more accessible through cloud-native platforms. At the same time, capital inflows are staggering — with funding in generative AI exceeding $56 billion in 2024 alone, according to data from S&P Global Market Intelligence[1].

Executives across industries expect massive productivity gains. The investment flows reflect this conviction: while generative AI funding topped $56 billion in 2024, the scale is accelerating even further. The StarGate project alone – a $500 billion, four-year initiative backed by OpenAI, Oracle, and Nvidia[2] – signals that capital commitments are now orders of magnitude greater. Similar mega-projects are emerging globally, underscoring the expectation that AI will reshape economies, not just individual enterprises.

The result is urgency, even FOMO (the fear of missing out). Startups are already building products on AI-first foundations, threatening to outpace incumbents weighed down by legacy systems. Enterprises that delay risk being outcompeted not just by peers, but by lean, next-gen challengers.

The potential is evident in early ROI stories. Klarna, for instance, reported sales and marketing spend savings, with 37% of the savings – around $10 million on an annualized basis – attributed to AI[3]. Another example comes from Base44, a company that scaled to an $80 million acquisition with fewer than ten employees, powered by AI-first operations[4]. Across US enterprises, a massive 62% have embedded generative AI into at least one core business process as of Q1 2025, according to SQ Magazine[5].

The message is clear: join the AI revolution, or risk being left behind.

Still, the volatility is real. A mortgage chatbot might behave like an unlicensed broker, a clinical trial assistant could hallucinate statistics, or an airline bot might authorize discounts it was never meant to. These aren't theoretical risks — they've already happened in various forms. As Philip Tannor, CEO of Deepchecks, puts it:

> **“Good evaluation isn't just a safety net; it's a feedback loop. Without it, you're flying blind.”**

The business value of agentic AI

For enterprises that get it right, agentic AI delivers transformative results. Financial services firms report up to a 60% reduction in model-risk review time. Pharma teams have cleared HIPAA sign-off two months faster. Across industries, early adopters are seeing development cycles shrink by a factor of two to three.

But the value extends far beyond cost savings. Agentic AI can reshape business models. Banks are automating complex customer queries while remaining compliant. Pharmaceutical companies are accelerating trial design. SaaS providers are deploying copilots that actively reduce churn and improve satisfaction.

One enterprise saw customer satisfaction climb from under 50% to nearly 90% after introducing continuous evaluation into their AI workflows. Another, a leading U.S. grocery chain, saw a 3.5% increase in sales, an 11% boost in on-shelf availability, and a $200 million uplift in incremental profit by using AI[6].

Yet the upside depends on rigor. As Tannor warns, "If your test set doesn't look like the real world, your metrics are just theatre." True business value is tied not only to what agents can do, but how reliably they perform under real-world conditions.

> **"If your test set doesn't look like the real world, your metrics are just theatre."**
>
> Philip Tannor,
> CEO, Deepchecks

Mind the PoC-to-production gap

Despite today's mature AI ecosystem, most enterprises remain stuck in proofs of concept that never graduate to production. The traps are consistent:

- **Misalignment between business and technical teams.** Without shared metrics, a PoC may look like a success in isolation but fails to translate into business impact. Teams often test the last ten things that went wrong instead of using representative datasets.
- **Poor use-case selection.** Some projects are too trivial to matter; others are too complex to manage within risk tolerances. Even the right use case often lacks production-ready governance and compliance.
- **Drift.** Models degrade silently as vendors change versions, prompts are edited, or retrieval layers shift. Confidential data can slip into vector databases unnoticed until it surfaces in outputs.

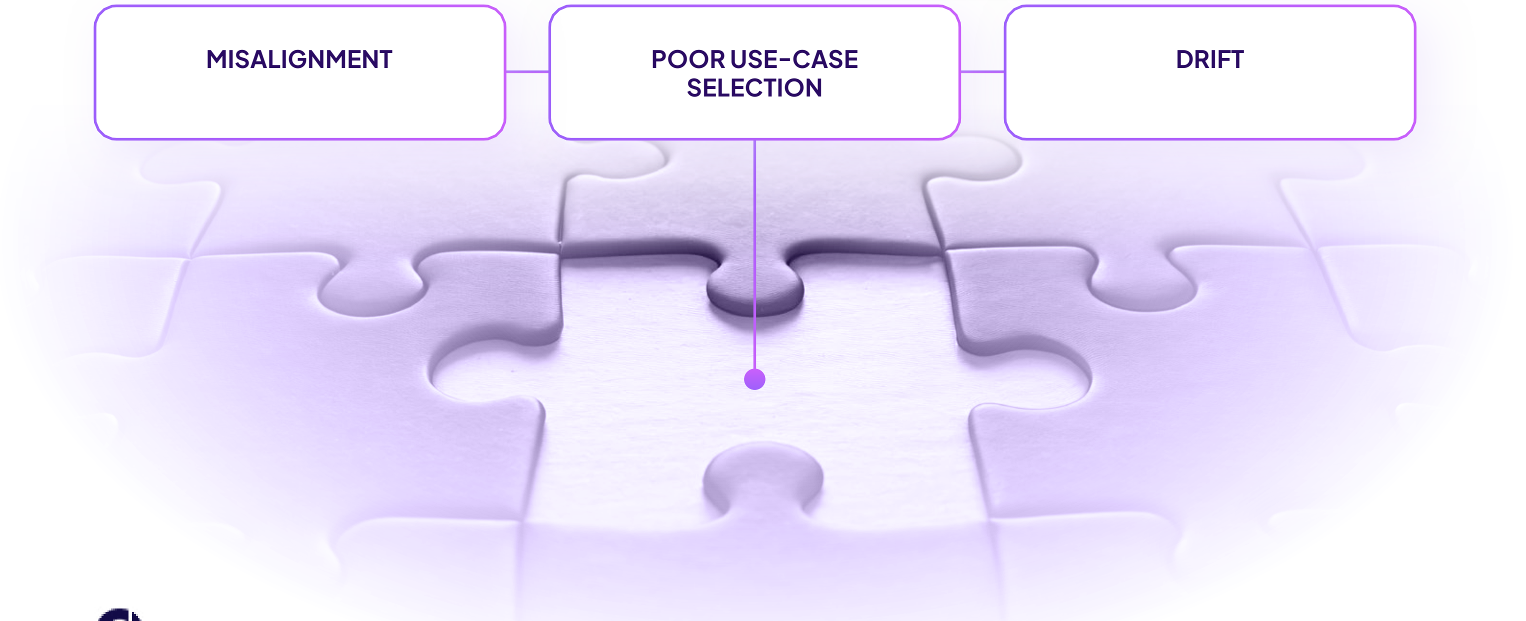
As Tannor cautions: "A model can be perfect on Monday and unsafe by Friday."

Manual reviews only catch cherry-picked examples. LLM-as-a-judge methods are partial at best. And without continuous integration, evaluations are one-and-done, quickly abandoned.

Doing nothing is not neutral — it is choosing to accept unmeasured drift. The result: blocked launches and mounting compliance debt.

The short blanket problem fits here: cover your head and your feet are exposed. Optimizing one stage of an AI workflow without systemic evaluation destabilizes another. This is why robust production planning is essential.

> "A model can be perfect on Monday and unsafe by Friday."
>
> Philip Tannor,
> CEO, Deepchecks


MISALIGNMENT

POOR USE-CASE SELECTION

DRIFT

The challenge of agentic AI

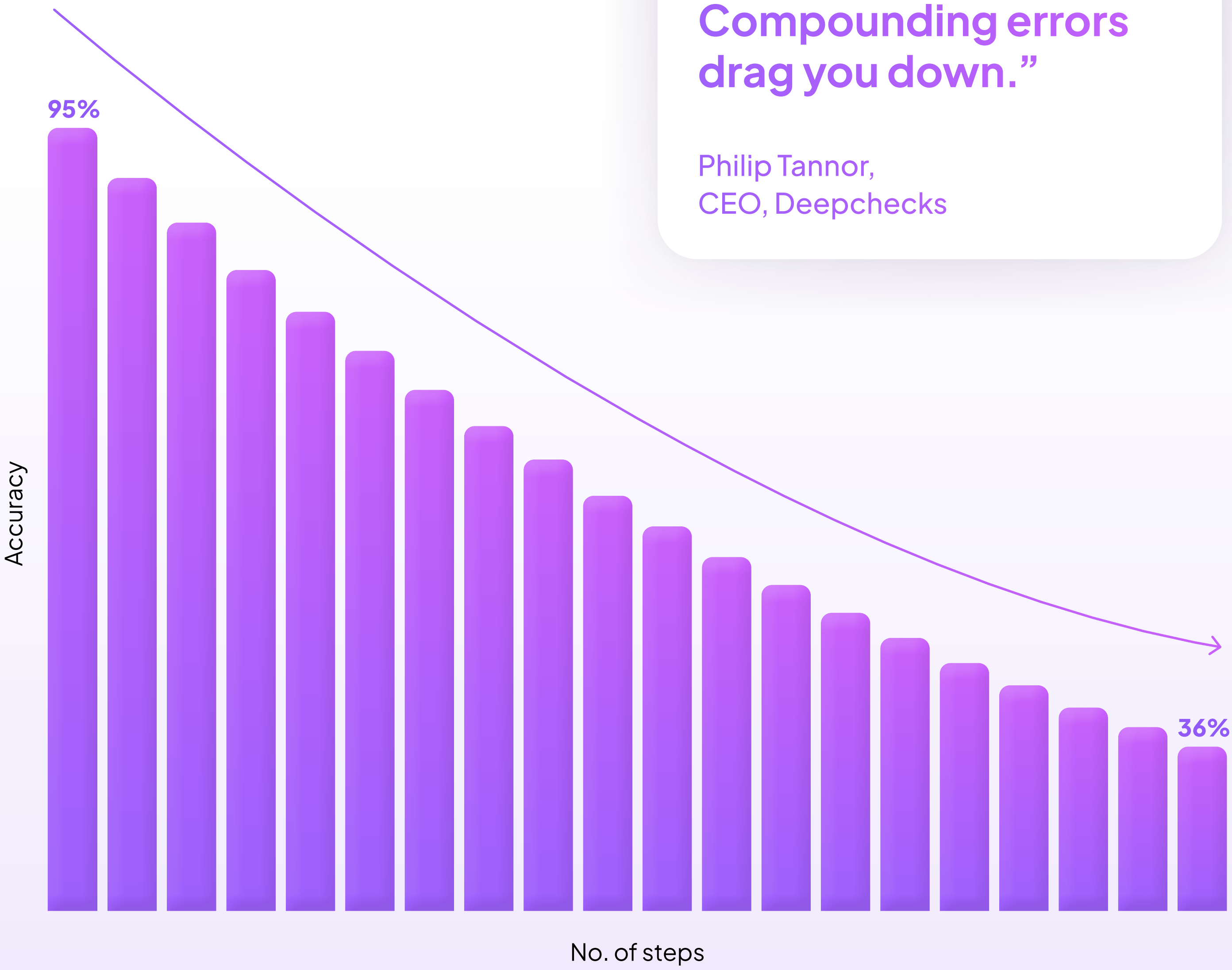
If single-turn LLMs can be fragile, multi-step agents are exponentially more so. Every decision, every external call, every chain introduces another point of failure.

“A step in the process handled by AI with a 95% accuracy score sounds impressive,” Tannor explains. “But string 20 steps together, and your overall accuracy collapses to barely 36%. Compounding errors drag you down.”

The butterfly effect applies: a tiny prompt tweak can ripple unpredictably through a workflow, producing outsized impacts. This is what makes agentic AI fundamentally different from traditional automation — errors multiply, cascade, and accelerate.

Evaluation is not optional. For agents, rigorous, automated, continuous testing is the only defense against fragility. Without it, drift will surface in production, often with costly consequences.

> **“A step in the process handled by AI with a 95% accuracy score sounds impressive, but string 20 steps together, and your overall accuracy collapses to barely 36%. Compounding errors drag you down.”**
>
> Philip Tannor,
> CEO, Deepchecks



The strategic foundations for AI success

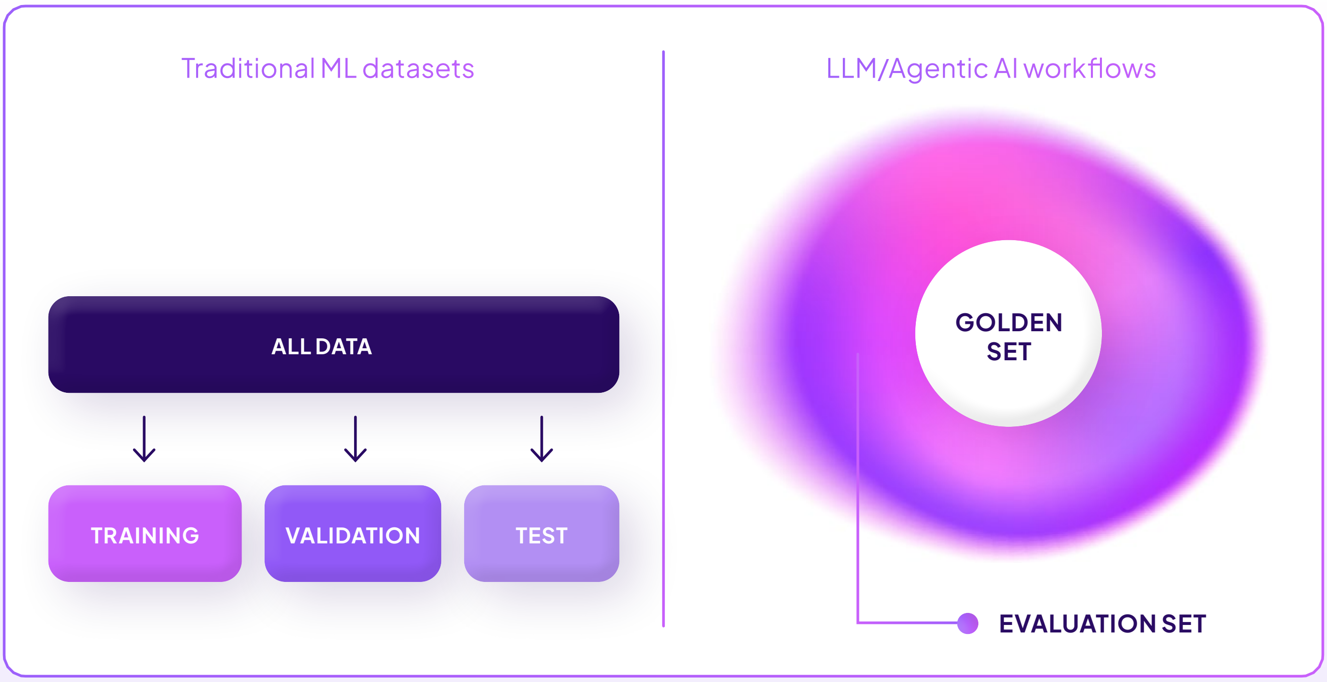
Defining success before you build

The first step in scaling agentic AI is agreeing on what success means. Alignment across business, product, compliance, and technical teams is essential.

Success isn't just about a model passing a test set. It's about meeting real-world outcomes while adhering to security and operational standards. This begins with the evaluation set — a carefully constructed dataset used to measure system performance. Unlike the training set, which teaches the model, or a golden set, which reflects known-correct outputs, an evaluation set is designed to mirror production realities. It becomes the benchmark against which success or failure is judged.

Teams must define thresholds for performance and compliance, identify edge cases that expose risk, and set escalation rules for drift or unexpected behavior. For multi-step agents, these edge cases multiply quickly, making manual evaluation impossible at scale. Automation — regression checks, routing validation, and continuous monitoring — becomes the backbone of reliability.

As agents grow more complex, decision frameworks such as visual trees or mixture-of-experts diagrams help teams navigate trade-offs and maintain alignment. **"You can't manage what you haven't measured,"** says Tannor. Defining success upfront is not bureaucracy; it is the foundation for resilience, trust, and ROI.



From proof-of-concept to reliable production

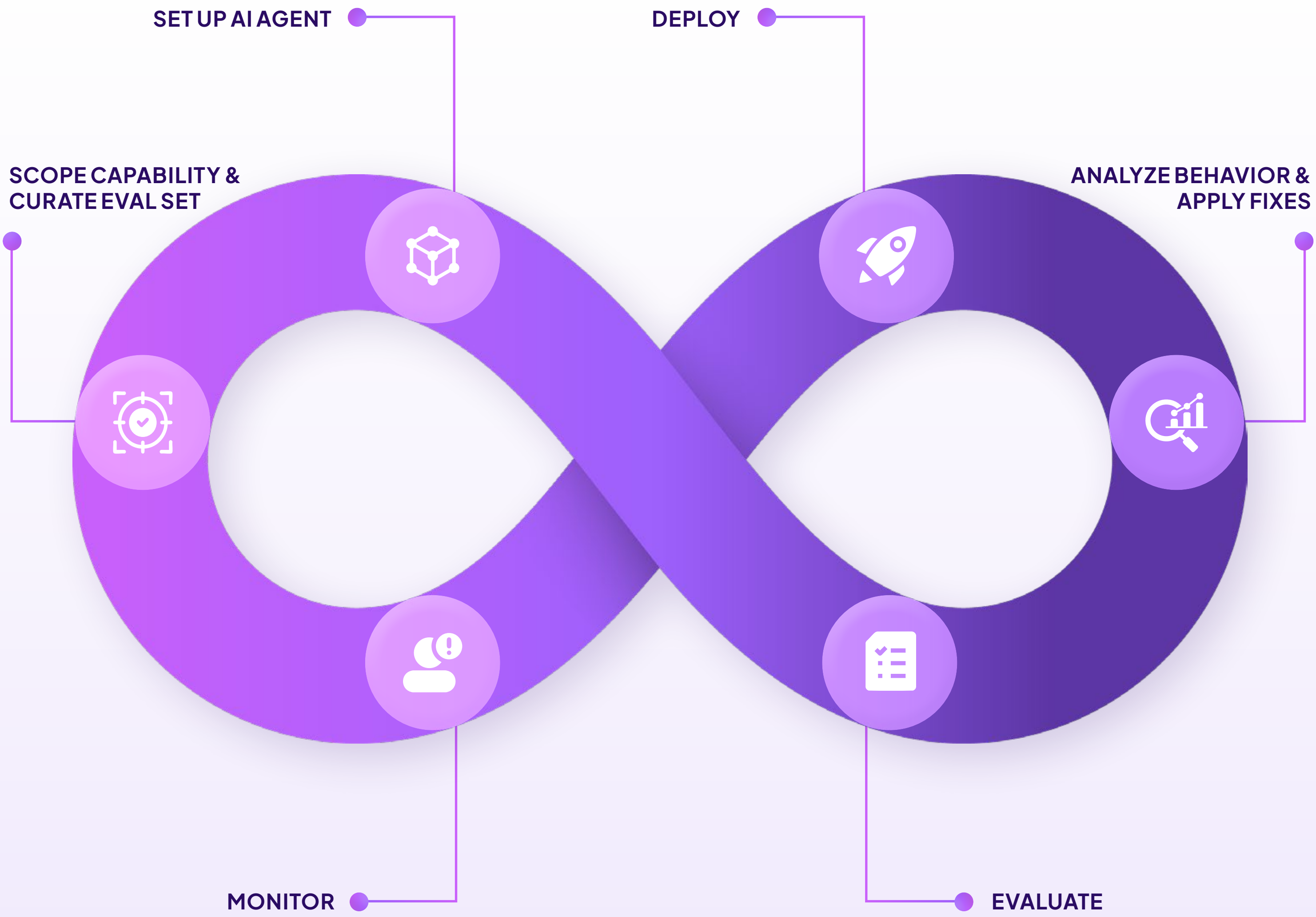
Moving from a promising prototype to a stable production system is rarely straightforward. Even when a PoC shows value, scaling requires evaluation embedded across the lifecycle. Without it, risks accumulate invisibly.

Continuous evaluation helps teams catch regressions early, compare agent versions, and ensure new features don't break existing ones. In high-stakes cases, human oversight must remain in the loop. Hybrid evaluation — combining automation, rules, and expert review — balances innovation with discipline.

Governance and organizational alignment

AI agents touch business processes, customer experience, compliance, and security. Effective governance begins with accountability. Cross-functional boards — spanning product, legal, technical, and compliance — ensure that deployment reflects organizational priorities.

Ownership matters. Every agent should have custodians responsible for monitoring, incident response, and evaluation. Governance also extends to data handling, security protocols, and audit trails. Clear playbooks prevent minor issues from snowballing. Governance isn't a hurdle; it's a shield.


SET UP AI AGENT
DEPLOY
SCOPE CAPABILITY & CURATE EVAL SET
ANALYZE BEHAVIOR & APPLY FIXES
MONITOR
EVALUATE

How to build an effective, performant LLMOps tech stack

Scaling agentic AI requires more than powerful models. Infrastructure must handle concurrency, orchestrate complex workflows, and enforce compliance. Cloud-native platforms such as AWS Bedrock provide these foundations: scalability, integrated security, and access to multiple foundation models. Orchestration layers such as LangGraph or Crew AI build on top of this, managing multi-agent workflows efficiently.

But infrastructure is only half the equation. Agents must be monitored continuously. With AWS-native logging, monitoring, and identity management, safeguards can be embedded into existing operations. Evaluation tools integrated into deployment pipelines ensure reliability is baked in from day one.

In this way, infrastructure becomes not just a platform for speed, but a framework for resilience, auditability, and continuous improvement.

Measuring what matters

As AI evolves — from deterministic software, to machine learning, to LLMs, and now to multi-step agents — evaluation paradigms have had to evolve too.

Traditional software testing focused on deterministic correctness.ML introduced metrics like accuracy and F1, anchored to reference datasets.

With LLMs, those metrics break down: BLEU and F1 fail because they assume one “right” answer, when in fact there may be many acceptable outputs.

Agentic AI compounds the challenge further, layering reasoning, retrieval, and tool use.

Success must therefore be measured across multiple axes: correctness, relevance, safety, compliance, and efficiency. Automated checks capture most errors, but high-stakes edge cases demand human review. Hybrid approaches provide the rigor needed to prevent drift.

It’s similar in many ways to the conveyor belt in manufacturing. If quality isn’t checked at every stage, flaws in inputs cascade; even catastrophically. Only by combining multiple evaluation lenses can teams understand true system performance.

Experimentation is critical. Iterating across prompts, architectures, processes, models, and data sources helps teams converge on performance thresholds. The question becomes: when do you trust an agent enough to ship it?

Regression testing provides confidence. By comparing versions against baselines, teams can detect whether fixes in one area destabilize another — revisiting the short blanket metaphor. Evaluation is not static; it is a continuous process of iteration and validation.

LLMOps tech stack power players

The agentic foundation

LLM Providers	Amazon Bedrock ANTHROP\C OpenAI NVIDIA Hugging Face
Agent Orchestration Layer	crewai LangChain LlamaIndex Haystack by deepset ADEPT
Agent Builder	CLOUDFLARE data iku Agentforce sas IBM UiPath
Tracing	deepchecks. LLM EVALUATION Langfuse LangSmith
Evaluation	deepchecks. LLM EVALUATION ragas Galileo PHOENIX

Memory & Context

Graph & Vector DB	Pinecone Weaviate Chroma Milvus LanceDB neo4j ArangoDB elastic
Data Platform/ Lakehouse	databricks snowflake amazon REDSHIFT NUTANIX Qlik
Memory & Caching	redis HAZELCAST DynamoDB
Tool Integrations	zapier apify n8n Voiceflow

Infrastructure and deployment

NeoClouds	NEBIUS CoreWeave Highrise Voiceflow together.ai Crusoe
Hyperscalers	aws
Other	NVIDIA NEMO intel Anyscale servicenow

MLOps, governance and continual improvement

Data Labelling	Toloka SuperAnnotate Labelbox Snorkel deepchecks. LLM EVALUATION
Security & Guardrails	NVIDIA NEMO deepchecks. LLM EVALUATION AIM aim security LAKERA Prompt:
Monitoring & Observability	deepchecks. LLM EVALUATION arize DATADOG new relic Grafana dynatrace

When to go live?

The decision to move from controlled testing to production is as strategic as it is technical. Best practice involves phased rollout: expose new features to a subset of users, monitor patterns, and refine before expanding.

Agents may also be adaptive — tailored architectures for different domains, e.g. one tuned for financial scenarios, another for medical. This allows risk calibration without compromising speed.

Once beta users engage, teams must track both user behavior and agent quality over time. Issues should inform iterative releases, with regression tests ensuring new builds don't reintroduce old errors. Slow, monitored rollout reduces risk while maintaining velocity.

Linking AI to business value

The economic case for agentic AI extends well beyond development costs. Maintenance structures matter. By designing systems that can be managed day-to-day by analysts or product managers — with clear escalation paths for engineers — enterprises free scarce technical talent to work on the next big project.

Unit economics are equally critical. Every token, evaluation check, and safeguard has a cost. Teams must measure, track, and optimize at scale. This means experimenting with smaller or newer models, benchmarking performance against latency and cost, and periodically reevaluating workflows to tighten efficiency.

In this way, AI becomes not just a tool, but a driver of measurable and sustainable business advantage.

Security and safeguards

AI systems introduce new risks: data leakage, prompt injection, jailbreaks, and hidden vulnerabilities. Enterprises must enforce guardrails, traceability, and continuous security monitoring.

Security is not separate from evaluation — it is part of it. Every deployment should be auditable, with mechanisms to detect and respond to anomalies quickly. Security failures are business failures; they must be treated as first-order concerns.

The Future of Agentic AI

Future-proofing your investment

Deployment is only the beginning. Continuous evaluation and monitoring determine whether enterprises realize long-term value or accumulate risk.

Future-ready strategies will be hybrid: smaller specialized models for niche tasks, orchestration layers to manage complexity, and integrated evaluation pipelines to catch drift early.

Deepchecks' Orion engine exemplifies systematized evaluation — monitoring multiple agents, catching subtle errors, and ensuring reliability scales with complexity. The principle applies broadly: evaluation is never one-and-done.

Action plan for leaders

For executives, progress can be mapped through milestones. Within 90 days, align stakeholders, define success metrics, and establish pipelines. Within six months, deploy hybrid evaluation strategies and governance boards. Within a year, tie outcomes directly to efficiency, compliance, and trust.

Leadership requires foresight and adaptability. Investments in talent, platforms, and processes must anticipate not only today's agents, but the next generation. Evaluation and monitoring must become cultural norms, embedded in operational excellence.

The enterprises that succeed will treat AI as a strategic capability, maturing over time. The upside is enormous: efficiency, innovation, competitive advantage. But only if fragility is acknowledged and systematically managed.

Taking the next step

AI is no longer futuristic — it is an operational reality. The stakes are high, but so are the rewards. By combining foresight, rigorous evaluation, and disciplined governance, enterprises can unlock transformative potential while minimizing risk.

Those that treat AI as a living, evolving capability will define the next wave of innovation.

Deepchecks LLM Evaluation Your AI Trust Layer

The themes in this ebook are not abstract. They reflect thousands of real-world AI projects where teams finally realized meaningful ROI.

Deepchecks is a turnkey solution for coders and clickers to define, measure, and validate AI progress. Our swarm of evaluation agents streamlines validation for every LLM app, from simple prompts to complex agents.

We help enterprises scale AI agents with confidence, reducing compliance risk, accelerating time-to-market, and ensuring that every AI initiative ties back to measurable business outcomes.

The key to accelerated GenAI adoption is LLM Evaluation done right.

deepchecks.

aws PARTNER Generative AI Software Competency

- Generative AI Applications

Deepchecks provides up to three weeks of trial access, along with a complimentary audit of your agentic AI system. You will receive a clear, actionable report that highlights risks, blind spots, and opportunities for improvement.

[1] https://www.spglobal.com/market-intelligence/en/news-insights/articles/2025/1/genai-funding-hits-record-in-2024-boosted-by-infrastructure-interest-87132257

[2] https://openai.com/index/announcing-the-stargate-project/

[3] https://www.klarna.com/international/press/ai-helps-klarna-cut-marketing-agency-spend-by-25-and-run-more-campaigns/

[4] https://techcrunch.com/2025/06/18/6-month-old-solo-owned-vibe-coder-base44-sells-to-wix-for-80m-cash/

[5] https://sqmagazine.co.uk/generative-ai-statistics/

[6] https://www.symphonyai.com/resources/case-study/retail-cpg/national-supermarket-chain-drives-200m-profit-vertical-ai/