opkey

The Responsible Path to Agentic AI for Enterprise Apps

Avinash Tiwari,
Opkey Chief AI Officer

Executive Summary

Enterprise applications have outgrown manual work and inadequate automation, prompting business leaders to adopt large language models (LLMs) and agentic AI that plan, act, and learn within clear guardrails. General-purpose LLMs can perform well on basic reasoning tasks but lack the workflow awareness, relevance, and auditability that ERP environments demand, leading to rework, hidden risks, and stalled programs.

The solution to effectively operationalizing agentic AI is to prioritize specificity over generality, pair autonomy with oversight, and treat responsible adoption as an operating discipline rather than an experiment.

This paper explains this approach in three parts. First, it clarifies why generic AI falls short in production—enterprise apps are “systems of systems” with tight dependencies, compliance boundaries, and data chaos that overwhelm general-purpose models. Second, it lays out a practical path: build a shared understanding of agentic AI, adopt domain-specific capabilities that work natively across deployment, operations, and optimization, and require robust inference accuracy on ERP-relevant tasks. Third, it defines five practices that make agentic AI relevant and safe: select purpose-built, agentic AI; require systems with high inference accuracy; remain stack-agnostic while optimizing end-to-end business workflows; keep humans in the loop with risk-proportional review; and follow responsible AI principles for data handling, explainability, and audit.

The paper closes with concrete guidance on balancing AI innovation and accountability, a realistic roadmap for the first year of deployment, and a clear message about timing: organizations that start narrow, measure results, and expand with evidence will see richer improvements faster.

Introduction

The evolution of enterprise applications is entering a new phase, where large language models (LLMs) and agentic AI are shifting from experimentation to applied use. Business leaders feel increased pressure—often buzzword-driven rather than motivated by core values—to map these technologies to measurable outcomes.

Leaders are caught between streamlining expenses and headcount while keeping pace with innovation and user needs. These requirements vary across industries and become more specific within customer segments. Market and vendor forces require doing more with fewer resources, and agentic AI promises the ideal solution for context-aware actions.

Done right, it can operationalize intelligence in the workflows of enterprise applications. Agents can observe, reason, and act on routine tasks and reduce risks in complex, critical processes. They can validate application configurations, run tests, detect integration issues, and guide users in real time—and hand off more critical cases to humans. The result is faster execution, higher quality, and better control.

Despite their promise, many AI programs struggle to convert interest into sustained operational value. Survey data reinforces both sides of this picture:

[NEED]

52 %

52% of enterprises still rely on manual, in-house ERP testing[1].

84 %

84% of employees anticipate positive agentic AI impacts on productivity, efficiency, and work experience[2].

88 %

88% of business leaders say adopting agentic AI will help their organization be more competitive[3].

1. The 2025 State of Cloud & ERP Operations Report, Opkey
2. EY Agentic AI in the Workplace Survey, EY
3. KPMG Canada survey, KPMG

[REALITY]

25 %

Less than 25% have automated anything[1].

70 %

70% of ERP initiatives will fail to meet business goals by 2027[4].

40 %

Over 40% of agentic AI projects will be cancelled by the end of 2027[5].

4. Enterprise Resource Planning to Optimize Operations, Gartner.
5. Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027, Gartner.

The challenge is easy to explain: Enterprise applications are complex webs of systems, processes, and people—with many operations relying on unquantifiable manual tasks. Even small gaps in agentic AI workflows can compound into frustrations and cost overruns.

> "One of the most common pitfalls teams encounter when deploying AI agents is agentic systems that seem impressive in demos but frustrate users who are actually responsible for the work. It's common to hear users complain about "AI slop" or low-quality outputs."
>
> - One year of agentic AI: Six lessons from the people doing the work
>
> McKinsey

For example, a small update to an enterprise application can ripple through order processing, payroll, and reporting. When automation or agentic AI doesn't account for all the configuration, integration, and testing dependencies, this release can have unforeseen consequences. People then have to step in for unplanned triage, remediation, and potentially reverting the changes.

Generic AI Advice Isn't Enough

An LLM built to answer questions across general-domain tasks may not understand the specific needs of supply chain management or human resources. Similarly, an AI agent that can book a vacation cannot reason and act on the complexity of tasks required for enterprise reconciliation processes or application testing.

Enterprises also need to safeguard their data and customer information. In many cases, this requires AI applications and compute running on private or local deployments – requirements that the popular LLM platforms cannot match.

This paper clears up a common confusion: not all LLMs and agentic frameworks are suited to enterprise applications. It explains the limitations of general-purpose AI, outlines a three-step path to enterprise-specific AI systems, and provides five best practices for effectively deploying a solution.

Choosing specificity over generality now, when the risk impact is low, better positions AI roadmaps for success.

> "Agents must be deeply aligned with the company's logic, data flows, and value creation levers—making them difficult to replicate and uniquely powerful."
>
> - Seizing the agentic AI advantage, McKinsey

Why General-Purpose AI Falls Short for Enterprise Apps

The market uses a single label for many different capabilities and products: “AI.” It can refer to chat-style assistants, retrieval tools, code copilots, workflow bots, agentic frameworks, and more. These systems vary widely in functions, value, and risks. It’s this variance that matters for enterprise applications.

General-purpose LLMs excel at language understanding and generation. They handle knowledge tasks, searching and summarization, and basic reasoning. They predict plausible text sequences to generate likely answers, and this is “good enough” for general public use.

They are not suitable for, nor can they guarantee accurate, auditable actions in enterprise production systems. A general AI model will always provide an answer, but in the absence of domain context, it will improvise, creating risk (e.g., hallucinations and misclassifications). Unreliable answers can lead to issues such as incorrect impact analysis, poor regression coverage, or incomplete configuration changes.

Consider how a general-purpose LLM would account for these real-world forces:

- The unique composition of an enterprise “system of systems,” with customized configurations, integrations, inputs, and outcomes.
- Domain and business knowledge held in the minds of employees and consultants.
- The cascading, often unpredictable, effects between enterprise systems (e.g., a purchase order moving through finance, supply chain, manufacturing, and warehousing operations).
- The tendency toward data chaos, where information is fragmented and inconsistent across systems (e.g., the interrelationships among product hierarchies, customer-vendor relationships, and Chart of Accounts structures).

What is data chaos?

Data chaos is a widespread information management problem where data is inconsistent, ungoverned, and fragmented across enterprise systems. This makes it difficult for LLMs and agentic AI to extract meaningful insights or perform effective actions.

General-purpose AI does not inherently understand enterprise application business logic, workflow interdependencies, or compliance boundaries. It cannot account for organization-specific configurations that fundamentally change system behavior. While heavy prompt engineering and human-added guardrails can support some domain-specific tasks, they require constant supervision and rework—adding cognitive load to users and increasing manual effort rather than reducing it.

Why Domain-Specific AI Excels

Domain-specific AI aims to achieve higher inference accuracy on enterprise application tasks, ensuring that the model's outputs (answers and actions) align with the desired results. This capability must be built into the LLM's design and validated across AI agent outcomes; otherwise, the risk of hallucinations will always remain in question.

What Is Inference Accuracy?

LLM inference is a computational phase where a trained model generates text or makes predictions from an input prompt. Inference accuracy is the proportion of those decisions that accurately match validated truth in a specific domain. "Low" inference accuracy may result in a model returning incorrect data or making a catastrophic mathematical error during financial analysis. "High" inference accuracy would yield factually correct, up-to-date information. Inference accuracy isn't a general benchmark applied across use cases; rather, it must be measured on data and tasks relevant to enterprise applications. As inference accuracy improves, business leaders and users have greater confidence in AI's results.

Domain-specific AI employs structured reasoning that understands and operates within this framework:

Concept

The precise meaning of terms, facts, and rules, such as “invoice,” “accounts payable,” and “warehouse.”

Task

The knowledge to perform specific procedures accurately and consistently, such as “process invoices” or “configure enterprise structure.”

Reasoning

The ability to logically connect concepts and tasks to solve complex, multi-step problems, such as financial analysis and user support.

Instruction-following and chat

Human-centered abilities to interact with users in natural and conversational ways.

Autonomous operation

The capability to perform tasks and reasoning without human input, such as automatic application testing, proactive system monitoring, and workflow orchestration.

Solving The Data Scarcity Challenge

Data scarcity is a significant problem for domain-centric AI, as finding high-quality training data encapsulating application-specific knowledge, dependencies, and problem-solving strategies is not readily available to the public.

Practical solutions mitigate these gaps through an automated pipeline that creates high-quality, in-context datasets by following a process similar to this:

Knowledge retrieval

Targeted web search queries, structured around an enterprise-focused taxonomy (e.g., financial, supply chain management, human capital management), gather relevant information from public sources.

Content filtering

Retrieved content is evaluated across domain-relevant dimensions (e.g., relevance, uniqueness, technical depth), and high-scoring content is collected.

Question generation

The filtered content is processed to generate diverse questions, ensuring an effective mix of interaction modes (one-off queries vs. full conversations), domain patterns (e.g., configuring apps vs. troubleshooting), and experience levels (basic, intermediate, advanced).

Answer generation

For each question, the system repeats the retrieval-filtering-question generation cycle to produce a factually grounded answer. This process is critical for ensuring factual accuracy and reducing hallucinations.

This process ensures comprehensive coverage of the enterprise application landscape and yields highly focused, relevant training data, thereby improving LLM inference accuracy.

3 Steps to Responsible Agentic AI Adoption

STEP 1 : BUILD A SHARED UNDERSTANDING OF AGENTIC AI

Business leaders need a common vocabulary before they can influence their teams. This starts by understanding and explaining how domain-specific LLMs and agentic AI differ from traditional automation and general AI:

- **Rule-based automation and robotic process automation (RPA)** follow predefined rules and scripts. They execute exactly what they are told and tend to fail when unexpected inputs or scenarios arise. These techniques are ideal for stable, deterministic tasks.
- **Retrieval-augmented generation (RAG)** is an AI technique that enriches an LLM model with external documents or data. It can improve the factual grounding of models, but the retriever can still fetch incomplete or irrelevant data, and the “retrieval-augmentation” phase adds latency to the overall process.
- **Model Context Protocol (MCP)** is a standardized protocol that enables LLMs to access external tools and services. Unlike AI agents, it is a connection mechanism only, and does not have the intelligence to reason or orchestrate multi-step tasks.
- **Agentic AI** orchestrates goals, plans, and actions over time by observing the current state of a system and running tools to achieve new, desired states. Each agent in a system focuses on a capability, such as configuration analysis, test creation, or user support, and allows humans to review and control actions at defined points.
- **Domain-specific agentic AI** adds domain depth by training models and agents on processes, test corpora, configuration schemas, and change patterns, thereby increasing confidence in autonomy. It can understand release cadences, cross-module dependencies, and compliance controls, and integrate with existing enterprise systems.

STEP 2 : ADOPT DOMAIN-SPECIFIC AGENTIC AI WITH END-TO-END NATIVE FUNCTIONALITY

Enterprise applications benefit when LLMs and agents participate directly in their lifecycle, from deployment through day-to-day operations to continual optimization. Their functions should be driven by the capabilities above—concepts, tasks, reasoning, instruction-following and chat, and autonomous operation—to improve the accuracy of answers and actions.

End-to-end native functionality matters because customized handoffs between systems require additional development and kill momentum in production. If an agent must export data to another system at each step, its value erodes.

Here is an example of how this works:

- **Deployment:** agents support application and component configuration, data validation, and integration testing autonomously. They detect the impact of changes and automatically propose test coverage, reducing the need to manually craft scripts and success criteria.

- **Operations:** agents monitor integrations, performance, and security posture autonomously, raising deviations against expected conditions before they escalate into issues. They also run regression suites as part of patch cycles and guide users within the context of their workflows, lowering support ticket volume and improving adoption.

- **Optimization:** agents autonomously mine process telemetry and configuration drift to identify inefficiencies and recommend changes. They also quantify their expected impact on reliability and support load, helping users decide which actions to take.

"Domain-specific" does not necessarily imply "large model." Choosing a domain-centric small language model (SLM) provides a compact, focused system with far fewer parameters and far faster response times. They work within a narrower scope and more explicit workflows, which lowers latency, eases resource constraints, and can avoid GPU procurement bottlenecks. It also lets teams deploy more specialized agents in parallel, closer to the systems they supervise.

In practice, a well-scoped SLM often outperforms a large general model on enterprise application-relevant tasks

STEP 3 : FOLLOW BEST PRACTICES FOR RESPONSIBLE IMPLEMENTATION

An AI agent's autonomy is only as effective as its accountability. Leaders should apply a set of consistent practices to ensure security, compliance, and sustained value.

The next section presents five best practices that provide a reliable baseline reflecting lessons from enterprise operations data, domain constraints, and emerging AI governance norms.

5 Best Practices for Adopting Domain-Specific AI Systems

1 : CHOOSE PURPOSE-BUILT AGENTIC AI

Select models and agents that are trained and tuned for enterprise applications, as their workflows, semantics, configuration, and assets form a unique business context necessary for accurate results. Domain-specific training reduces hallucination risk and enables better reasoning over multi-step workflows such as release validation or cross-ledger reconciliation.

2 : REQUIRE HIGH INFERENCE ACCURACY ON RELEVANT TASKS

According to McKinsey, 51 percent of AI users in global enterprises say their organizations “have seen at least one instance of a negative consequence, with nearly one-third of all respondents reporting consequences stemming from AI inaccuracy.”

To avoid these issues, choose models that demonstrate high inference accuracy reflective of business needs. Also consider SLMs that reliably and securely support the addition of business-specific data. This builds trust, boosts adoption, and aligns application workflows more closely with business outcomes. It also enables users to delegate tasks to autonomous agents with more confidence.

3 : MAINTAIN STACK AGNOSTICISM AND OPTIMIZE

For every SAP, Oracle, Workday, and Coupa, there are function-specific vendors occupying real estate in enterprise stacks. Agentic systems that work independently of specific vendors and components lower integration effort and enable cross-system optimization.

This independence comes in the form of broad connector coverage, adaptable data models, and AI that can reason and act across components. For example, a vendor-agnostic agent can integrate with and autonomously run tests between HR and payroll when either system’s configuration changes. Or it can trigger a risk analysis in the finance system when purchasing changes a procurement process.

4 : KEEP HUMANS IN THE LOOP

Much of the discussion around AI pits the “people stack” against the “tech stack,” creating a false choice that erodes trust and slows adoption. A more sustainable approach is to design human oversight into agentic workflows so that expertise and autonomous action aggregate value, rather than compete.

McKinsey research agrees, stating that AI high performers (those using AI to drive growth, innovation, and cost) are “more likely than others are to say their organizations have defined processes to determine how and when model outputs need human validation to ensure accuracy.”

People can set the intent, policies, and thresholds for AI workflows in the following ways:

- Establish controls that connect AI decisions to the need for review, so that low-risk actions execute automatically and higher-risk actions are routed to subject-matter experts for review and approval.

- Define who approves agent policies, who reviews exceptions, and who monitors accuracy, drift, and key performance indicators.

  Train teams to work with LLMs and agents and to interpret their outputs.

- Log evidence for every autonomous decision and action. For higher-risk workflows, this data should include an explanation of why an agent chose a path, the alternatives it evaluated, and the safeguards in place if conditions change.

5 : FOLLOW RESPONSIBLE AI PRINCIPLES

Responsible AI principles shape how models are trained, evaluated, and supervised, ensuring their outputs and autonomous actions are accurate, safe, and consistent with organizational security and regulatory compliance rules.

They also guide AI systems in ways that are ethical, transparent, and aligned with human values. In practice, this means fairness in how systems behave, accountability for the choices they make, clear explanations of how they reached those choices, tight controls that reduce misleading outputs, and steady attention to minimizing harm so people can trust the results in real work.

These principles can be implemented in the following ways:

- Adopt LLMs and agents that comply with data protection, privacy, and security policies. This includes finding platforms that support local or private deployments, such that proprietary and sensitive data can be added without compromising security.

- Treat training data with the same care as live records, checking for completeness, accuracy, and relevance. Remove anything that violates company policies and industry regulations before they're included in a training process.

- Look for and minimize the impact of agentic actions that may make life harder for a specific role, region, or customer segment. For example, agentic training that ignores local employment laws should be reviewed and adapted to fix the root cause in data or reasoning.

- Automate audit and compliance processes, including logs that capture agentic inputs, steps, and results to support reporting, certification, and remediation.

Balancing Innovation with Accountability

AI autonomy can only create business value when paired with guardrails to protect revenue and reputation. Leaders should understand and address these four AI risk categories to minimize any negative impact:

Over-reliance on AI decision-making

Agents can make the wrong decision or take a questionable action. To limit the blast radius, enforce rules that constrain agentic actions to approved domains and keep humans in the loop for material changes and ambiguous cases.

Black box decisioning and opacity

The lack of transparency into AI reasoning and decision-making is often a blocker to trust and adoption. Deploying systems with explainability features for agentic decisions, such as plain-language explanations, helps users validate results. These features do not need to expose model internals; rather, they should show assumptions and decision paths with enough detail for a human to understand.

"By 2027, 80% of Critical AI Decisions Will Require Human Oversight Supported by Visual Explainability Dashboards, Potentially Slowing Processes but Enhancing Accountability."

IDC FutureScape: Worldwide Artificial Intelligence and Automation 2025 Predictions, IDC

Compliance and data governance lapses

Enterprise applications handle sensitive data under strict regulation, so set policies that prohibit using customer data for training without explicit authorization, segregate environments with role-based access limiting what an agent can do, and test these controls regularly.

Workforce concerns and change management

Agentic AI triggers understandable concerns about a person's role and value. To foster trust, explain transformation plans early and demonstrate how agents will eliminate repetitive work, enabling a greater focus on high-value effort. Reinforce this with AI-agent-led training and measure adoption health through satisfaction surveys to quickly address friction points.

What Year One Looks Like

Leaders want to know what to expect out of their domain-specific AI initiatives in the first year. While results vary by company, a realistic roadmap looks like this:

QUARTER 1

Establish AI governance, policies, and target metrics, followed by deploying two use cases in a limited scope. This could be low-risk application integration testing or automated configuration management.

QUARTER 2

Expand to automated regression execution for selected releases and integration verification after patches. Introduce a new training agent to guide end users and boost adoption.

QUARTER 3

Add process discovery and automated impact analysis to assess workflows and streamline system changes across two major components. At the same time, test coverage increases as more modules and low-risk processes are automated.

QUARTER 4

Extend agents across the application stack, connecting analytics and insights to optimized recommendations and autonomous actions. Resource gains are invested into backlog reduction and cross-functional improvements. Close the year with documented results and a plan for the next phase of AI autonomy.

The Agentic AI Enterprise of 2026

What hasn't changed is the need for reliability, efficiency, and clear control over enterprise applications. The difference now is that the growing volume of changes and the tight coupling between systems require a new approach that can learn, plan, and act autonomously within boundaries relevant to the business.

This is where agentic, domain-specific systems prove their worth—connecting what people intend to do with what the technology stack can safely and reliably execute. Responsible adoption tunes models to the semantics of enterprise systems and enables high accuracy proven on tasks tied to business risk. These models and agents must be interoperable across the application estate, support human review at the right points, and handle data in accordance with policies and regulations.

Domain-specific AI creates a foundation that supports change without constant firefighting. The sooner it starts, the sooner organizations shift work from AI triage to real improvements in business outcomes. If your organization is exploring responsible agentic AI for enterprise applications, consider assessing your readiness against the practices in this paper. To help you get started, contact Opkey sales now.

opkey

Implement Faster. Operate Smarter. Optimize at Every Stage.

Opkey delivers an agentic AI-native platform that streamlines every step of the enterprise application lifecycle, designed to accelerate implementations, stabilize operations and maximize adoption. A central orchestration agent coordinates Define, Design, Configure, Test, and Train phases for major transformations and business-as-usual change, reducing cost, risk, and complexity for enterprises.

To learn more or request a demo, visit opkey.com

/company/opkey/ /@Opkey/ /@OpkeyAI