

The cognitive leap

How to reimagine work
with AI agents

Content

Key takeaways

- Multiagent AI systems can help transform traditional, rules-based business and IT processes into adaptive, cognitive processes.
- Organizations should leverage key principles of AI agent and multiagent AI system design and management, which borrow from tenets of composable design, microservices architecture, and human resources deployment and teaming.
- The ability to scale AI agents and multiagent frameworks across a range of use cases depends on developing a comprehensive reference architecture populated with reusable core components.
- A systematic approach can make the difference between incremental, isolated improvements and exponential enterprise transformation.

Vaulting ahead on the path to GenAI value	3
How agents deliver a cognitive advantage: Principles of AI agent design and management	5
Adaptive processes for innovative outcomes: Principles of multiagent AI system design and management	8
Expanding and scaling multiagent AI systems: A reference architecture for agent-powered transformation	10
Multiagent AI systems in action: An example use case for transforming traditional IT support processes	12
From generating to innovating: Key considerations on the path to AI agent-enabled transformation	14
Making the cognitive leap	16
Get in touch & Endnotes	17

Vaulting ahead on the path to GenAI value

Everyone remembers that pivotal moment when we first saw what large language models (LLMs) and Generative AI (GenAI) could accomplish. Suddenly, the long-discussed theory of conversational, intuitive, creative AI became a reality, right there at our fingertips. Adoption of GenAI surged across industries: By the end of 2023 most companies had embraced GenAI solutions.² By midyear 2024, 67% of companies using GenAI said they were increasing investments after seeing strong results from the technology.³

But as companies dove into testing GenAI's potential, many came to recognize the limitations of standalone GenAI models. Context and reasoning limitations of typical LLMs can make it difficult to apply GenAI to complex, multistep workflows. As with traditional AI, hallucination and bias can create significant barriers to trust. And the creative outputs for which GenAI is celebrated require continuous human monitoring for quality and accuracy.

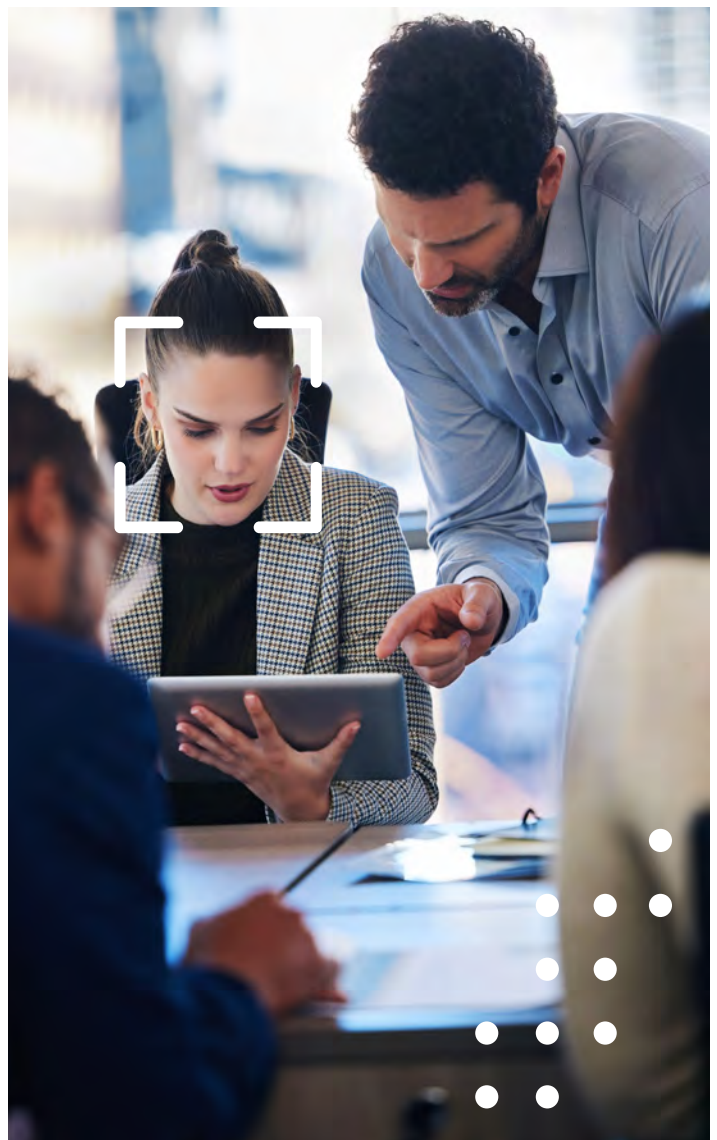
For these and other reasons, early GenAI use cases were mostly limited to isolated or narrowly defined tasks within larger workflows. For example, a wealth management adviser may quickly produce a meeting recap using a standalone, LLM-based solution. But extracting rich post-meeting analytics based on different information categories discussed in the meeting (e.g., client profile, client goals, retirement information, etc.) remained too complex to achieve with a standalone GenAI solution.

AI agents and multiagent AI systems are helping organizations hurdle these limitations and make the cognitive leap into a new paradigm of business process transformation and innovation.

AI agents enable organizations to tackle significantly more complex tasks with GenAI across an expanded range of processes and use cases. When AI agents work together in a system, they can help collaboratively reason, plan, design and execute novel workflows that amplify speed, differentiation and efficiency across the enterprise.

In this paper we outline key design principles and a reference architecture for scaling AI agent use cases that can help your business seize the potential of AI agents now.

Business executives say **deeply embedding GenAI** into business functions and processes is the **No. 1 way to drive value** from the technology.¹





“Each mind is made of many smaller processes. These we’ll call agents. Each mental agent by itself can only do some simple thing that needs no mind or thought at all. Yet when we join these agents in societies—in certain very special ways—this leads to true intelligence.”

—Marvin Minsky, *The Society of Mind*⁴



How agents deliver a cognitive advantage

Determining the most appropriate roles and uses for AI agents begins with adopting a shared, enterprisewide understanding of what they are and how they can fit into your organization.

AI agents are reasoning engines that can understand context, plan workflows, connect to external tools and data, and execute actions to achieve a defined goal. They do so by echoing some of the key qualities and advantages that have helped humans survive and flourish.

As people, we can understand language and creatively articulate responses. By employing specialized tools, we can amplify our physical and mental capabilities. By learning and remembering information, we avoid mistakes and improve on what we've already accomplished.

Language, planning, reasoning, reflection, and the ability to use tools, data and memory: These attributes are central to how AI agents work and demonstrate cognitive abilities as well.

In the realm of business, AI agents and human workers have other broad similarities. Both must be carefully selected, well trained and well equipped to perform their jobs. And both should be smartly deployed and consistently managed in ways that help ensure efficient, value-adding performance.

Not surprisingly then, **our recommended principles of AI agent design and management echo familiar themes from organizational design and human resource management.** *(Please see next page.)*

Principles of AI agent design and management

- **Domain-driven approach:** Every area of expertise and function of your business utilizes different processes, data and tools. While some AI agents may be able to serve multiple domains and processes, most should be sourced and/or designed based on specific domain requirements. To achieve this, each domain of your business should be analyzed, subdomains and processes identified, and agents assigned based on specific roles within the domain.
- **Role-based design:** Agents should be designed to perform roles rather than specific tasks, grouping similar activities to avoid confusion and ensure efficient operation. This approach—which aligns with the “single responsibility principle”⁵—can help your organization reduce AI agent overlap and unnecessary technology complexity. It also can help enable reusability of agents across systems and domains.
- **Right balance:** Related to the principle of role-based design, it is important to find the proper balance between the *number* and the *scope of responsibilities* of individual AI agents. Too many agents with too few responsibilities can result in unnecessary costs as well as challenges related to consistent governance, maintenance, monitoring and upgrades. Too few agents with too many responsibilities can result in bottlenecks and poor performance.
- **Controlled access to data, skills and tools:** You wouldn't give every employee in your enterprise access to every application or data resource in your business. Similarly, the tools, data and skills made available to a given AI agent should be limited to those that are essential to its role. These constraints help reduce risk and improve outputs from the agent. If an agent's role requires more than five tools, consider how you might separate its responsibilities across two or more agents.
- **Reflective cycle:** Agents—like people—get better and better when given an opportunity to reflect on their own performance or receive constructive criticism. That's why it's important to design a self-reflective pattern in which agents critically evaluate their own output by referring to past examples or testing the results of its output. Agents also receive feedback from other agents and humans. This combination of self-assessment and external input creates a continuous loop of learning and improvement that helps ensure compliance with quality, brand and risk standards.



“Synergy (is) the bonus that is achieved when things work together harmoniously.”

—Mark Twain

Adaptive processes for innovative outcomes

The achievements of remarkable individuals—from Aristotle to Simone Biles—are often treated as proof of our boundless human potential. But as any leader today knows, individual strengths are no match for team synergy. Organized and managed well, teamwork leverages and amplifies the strengths of each individual—making it possible to achieve goals that no person could do alone.

As with people, so too with AI agents. Research has shown that AI agents working together are more effective than individual agents.^{6,7} By leveraging an “agency” of role-specific AI agents, multiagent AI systems can understand requests, plan workflows, delegate and coordinate agent responsibilities, streamline actions, collaborate with humans, and ultimately validate and improve outputs. Processes that were considered too complex for typical language models can be automated at scale—securely and efficiently. Projects that once took weeks can be completed in a small fraction of that time. Human workers who previously spent precious hours performing routine, repetitive tasks can instead focus on higher-level, higher-value activities.

So, while standalone AI agents can help accelerate the completion of individual tasks, multiagent AI systems can open new realms of business process automation, speed and reliability. Agents within a system can interact and collaborate in various deployment patterns, depending on the specific needs and complexity of the process.

Multiagent AI systems have the potential to impact every layer of enterprise architecture—not just automating existing processes and tasks, but also reinventing them.

By engaging with users and within workflows semantically rather than syntactically, AI agents can comprehend emerging needs and address them in novel ways that obviate traditional, rules-based processes. By continuously self-monitoring, multiagent AI systems can improve their outputs in near real time. Meantime, the shared persistent state of AI agents in a system enables them to collaborate and coordinate activities in ways that continuously streamline efficiency.

The principles of agent design discussed in the previous section become especially important in this context. For example, dynamic workflow planning and task decomposition in a multiagent AI system are critical to effectively automating and reinventing end-to-end processes—and are dependent on the *right balance of domain-specific, role-based agents* to perform each task. By providing each agent with *controlled access* to data, skills and tools—and by providing *checks and balances* throughout the whole system—redundancies can be avoided and quality improved.

When designing multiagent AI systems, we recommend a set of principles to help ensure that these systems are robust, reliable and trustworthy. *(Please see next page.)*

Principles of multiagent AI system design and management

- **Understandable and explainable systems:** Good business leaders explain and justify their decisions, and AI systems should do the same. The actions of your multiagent AI systems need to be explainable, particularly in tasks related to perception and classification. Systems should be designed to document each agent's chain of thought,⁸ and not just the final output. (Think of it as "showing your work" in math class.) Clarity and interpretability will help minimize biases originating from their design or datasets.
- **Composable design:** Multiagent solutions should be designed with composability in mind. A composable design can allow organizations to bring best-of-breed components together in a microservices architecture to develop optimized and efficient multiagent systems. By orchestrating custom and third-party agents that include different programming languages and agent frameworks, your organization can design more complex agentic patterns that integrate with multiple internal and external systems.
- **Human in the loop:** AI agents shouldn't be solely responsible for critiquing their own or other agents' outputs. Knowledgeable humans must be essential parts of AI systems as a safeguard against potential errors or biases. This isn't just common sense; it's a regulatory mandate in some industries and/or US states. California, for example, recently required that AI-generated health care-related decisions must be reviewed by a human before being shared with consumers.⁹
- **Dynamic data patterns:** In designing multiagent AI systems, data should be able to flow in two distinct patterns: *data to the agent* and *agent to the data*. In the data-to-the-agent pattern, unstructured data is typically captured into a vector or graph database. It's important to include not only the data itself but its hierarchy relevant to the specific use case. This enables agents to apply the data appropriately within various contexts. In the agent-to-the-data pattern, the agent uses suitable tools built into the model (such as search tools or API specifications) to determine how to retrieve relevant structured data for the task at hand.
- **Ecosystem integration:** A multiagent AI system often needs to integrate with various existing applications or processes to achieve its intended goals. Therefore, the design of these systems should consider integration patterns with ecosystem processes and applications. Some integrations may be achieved via application programming interfaces (APIs), while others may be event-driven. For example, a multiagent system for post-meeting analytics may need to integrate with a CRM platform through an API to upload client profiles or other information discussed during the meeting.
- **Continuous improvement and adaptation:** Performance improvement must be built into the "DNA" of multiagent AI systems. Systems should be designed to learn from prior interactions and evolve in response to new data and changing conditions. This capability can be implemented through agent and workflow memory, which stores past interactions and workflow executions. The stored information can later be leveraged to enhance future executions.
- **Ethical considerations:** The same ethical principles you apply to human capital decisions, such as impact, justice and autonomy, should guide the design and deployment of multiagent AI systems. In addition to prioritizing explainability, your organization should regularly assess AI system outputs to ensure they contribute positively to society and avoid causing harm.

Expanding and scaling multiagent AI systems

Imagine you're the chief transformation officer at a global financial services company. You understand the principles of AI agent and multiagent AI system design. You see the potential in this next evolution of Generative AI technology everywhere in your organization.

But where to apply it?

A multiagent AI system could help your HR team identify, recruit and onboard talent by analyzing mountains of resumes against job requirements, intelligently assessing candidates based on skills and experience, even conducting initial screening interviews. The benefits seem obvious: greater scalability and efficiency, improved candidate matching, less bias ...

Then again, AI agents could transform efficiency in your call center by enabling plain-language conversations between clients and chatbots. This could help digital self-service feel more like old-fashioned client service—while your human support reps are freed to focus on more sensitive, higher-value interactions.

Or maybe the place to focus is in improving personalization in financial advisory services? Or in automating financial reports? The list goes on—across every domain of the enterprise.

Thanks to the innate flexibility and scalability of multiagent AI systems, your organization doesn't have to limit its focus. While it is true that no organization possesses the financial, talent or technological resources to design and deploy bespoke multiagent AI systems for every possible domain or use case—no longer are these resources requisite to success.

The key is to treat a multiagent AI system as an ecosystem of capabilities instead of solutions and to develop a reference architecture that can support both business and technical delivery processes. This approach can allow your organization to more rapidly scale, expand and reuse AI agents and multiagent frameworks across a range of use cases—while also streamlining governance, monitoring, operation and improvement of agentic outputs.

The essential layers of a reference architecture are shown in the illustration on the next page. Each layer within the architecture is loosely coupled with—but independent of—other layers. Similarly, each component within a given layer can be leveraged independently. This makes it possible to adapt, connect and apply best-fit solutions for any use case that arises.



A reference architecture for agent-powered transformation

Interaction layer

Purpose: Allow users, processes and existing applications to collaborate with multiagent AI systems.

Actions for success: Develop defensive user interfaces that can anticipate and mitigate potential user errors or misuse, while guiding the multiagent system(s) to respond contextually.

Example elements for a financial services company:



Mobile banking app



CRM system



Conversational IVR system



IT support portal

Workflow layer

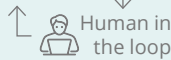
Purpose: Ensure controlled flow engineering to help agents interact with each other efficiently and in a more deterministic manner.

Actions for success: Implement value-stream analysis to monitor efficiency and effectiveness of workflows. Identify governance guardrails and touch points for human monitoring (“human in the loop”) to help reduce risks. Infuse long-term memory into workflows.

Example elements for a financial services company:



Know your customer workflow



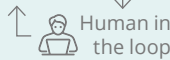
Risk control testing workflow



Financial planning workflow



Software incident support workflow



Agents layer

Purpose: Create, manage, deploy and optimize role-specific AI agents.

Actions for success: Focus on industrializing the creation of role-specific agents to accelerate speed to value.

Each agent should be equipped with:

- A fit-for-use *language model*
- *Tools* that augment language model capabilities with skills to perform specific tasks/roles
- Approved sources of authoritative *data*
- *Memory* of past tasks to help improve performance of new tasks
- Access to effective *prompts* for engaging with other agents and/or humans in a given workflow

Example elements for a financial services company:

<p>MODEL GARDEN</p> <ul style="list-style-type: none"> Multimodal commercial LLM Multimodal open-source LLM Fine-tuned model Domain-skilled SLMs 	<p>AGENT FACTORY</p> <ul style="list-style-type: none"> Data retrieval agent Recommendation agent Incident classification agent Incident analysis agent Incident resolution agent Quality assurance agent Agents from third-party vendors 	<p>TOOLS</p> <ul style="list-style-type: none"> Search engines Financial analysis tool Code interpreter
<p>PROMPT REGISTRY</p> <ul style="list-style-type: none"> Prompt templates Prompt versioning Prompt testing Prompt access management 	<p>DATA SOURCES</p> <ul style="list-style-type: none"> Customer 360 record Financial markets data Incident history 	<p>MEMORY</p> <ul style="list-style-type: none"> Short-term (current session) Long-term (past sessions)

Agent operations layers

Purpose: Monitor outputs and metrics to help ensure agents are functioning as expected.

Actions for success: Implement instrumentation and telemetry, along with logs, traces and metrics, to gather data about system activities. Activate alerts and dashboards to simplify performance monitoring against service-level objectives.

Example elements for a financial services company:



Operational metrics



Qualitative metrics

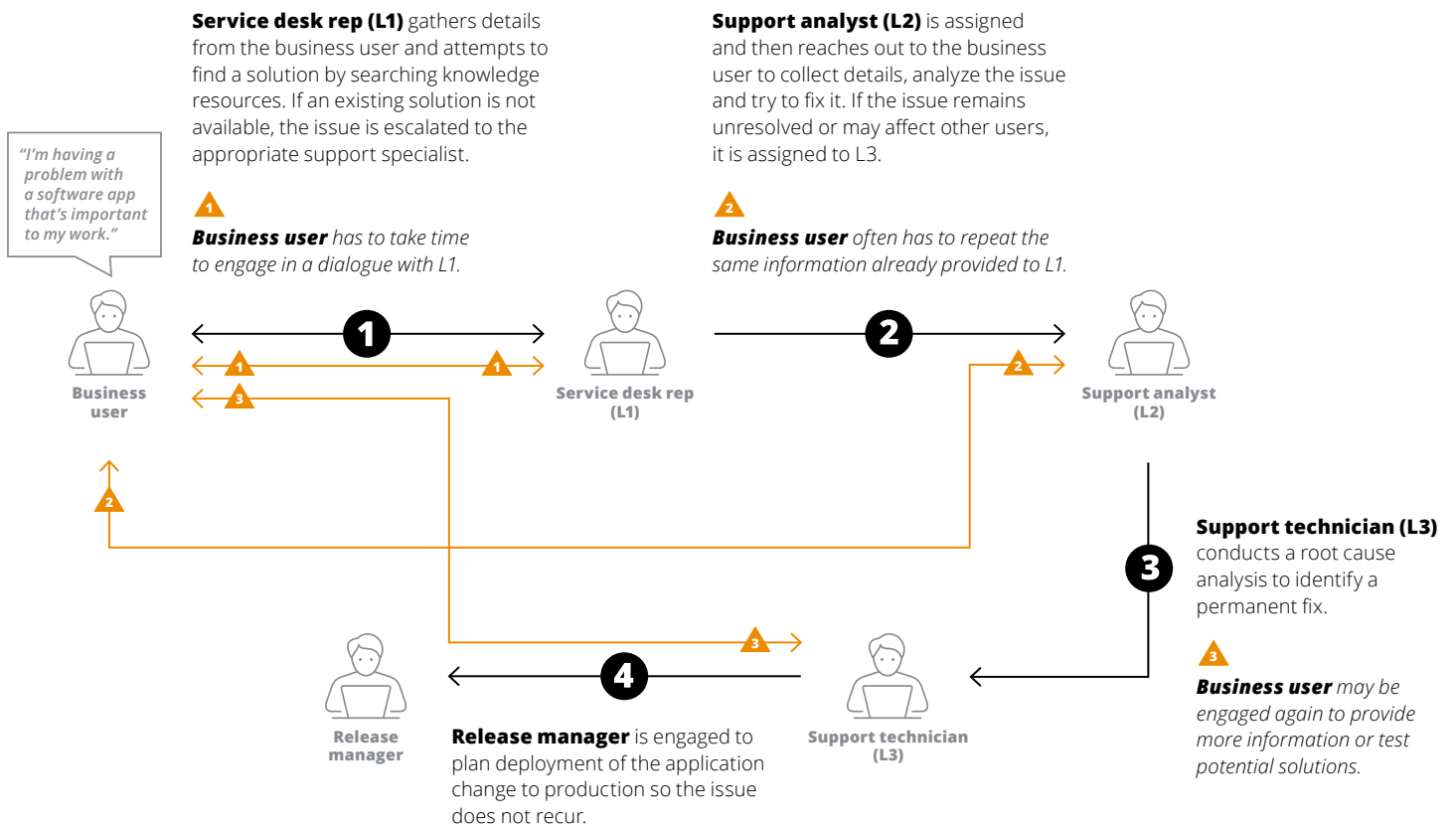


Thought metrics

Multiagent AI systems in action

Continuing our exploration of the reference architecture layers and elements that contribute to effective, efficient and scalable multiagent AI systems, let's look more specifically at an IT operations process—specifically, a support scenario for a business software application.

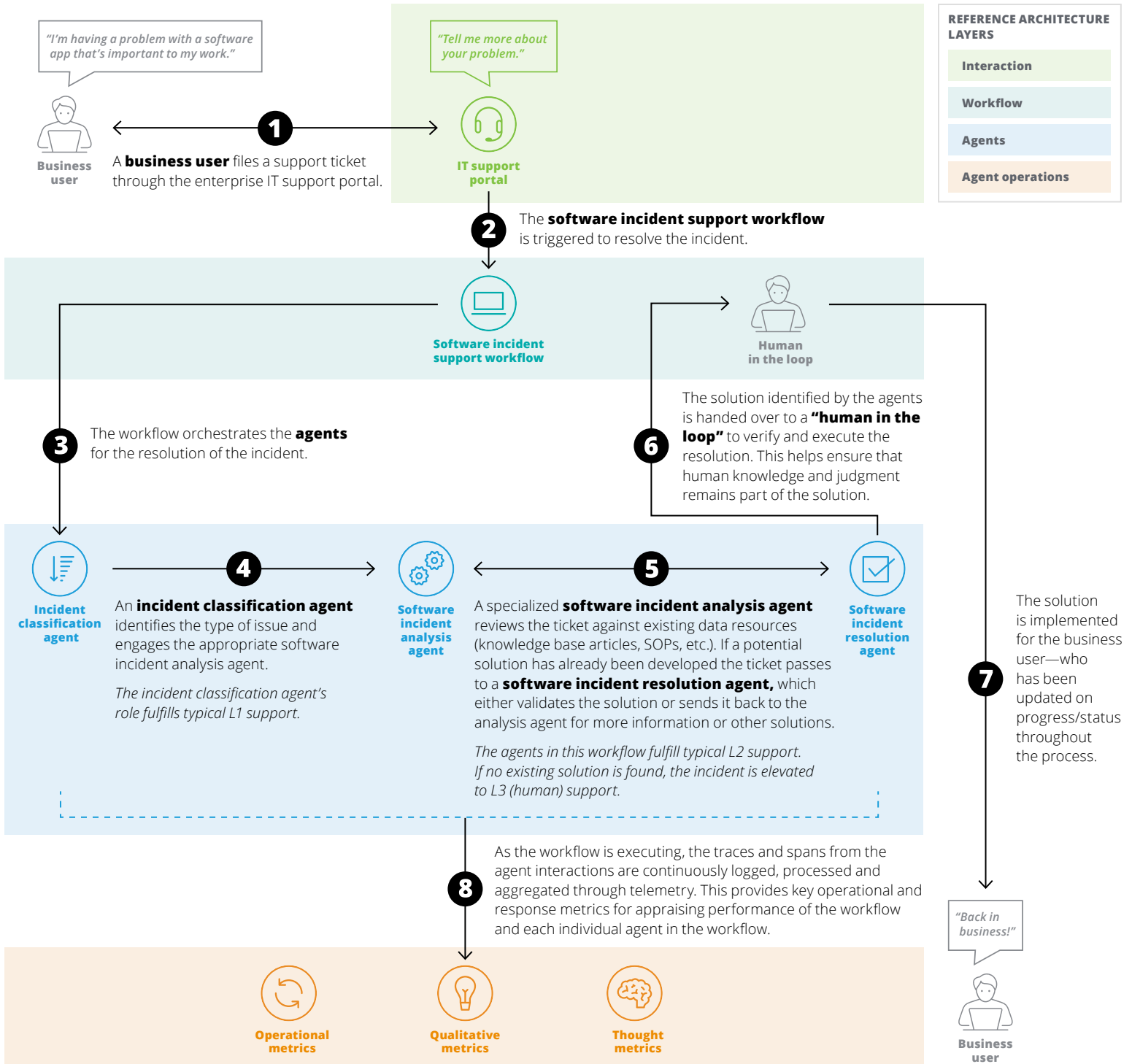
Traditionally, this process involves multiple support team interventions and touch points for the business user. The diagram below illustrates this resource-intensive, inefficient and often time-consuming workflow.



Traditional L1 and L2 IT support workflows are primed for transformation through multiagent AI system solutions. By leveraging an AI agent-enabled process, the user is *continuously updated*—but can be much *less actively engaged*. Support personnel are engaged only to *monitor, review and approve* rather than *find and implement* most solutions. This frees the human support personnel to focus on the most complex and business-critical resolution of select issues. And it frees business users to get back to the important work of generating enterprise value.

Here's how it can work.

(This example shows one variation of IT support for illustrative purposes. The most appropriate solution for your business may differ.)



From generating to innovating: Key considerations on the path to AI agent-enabled transformation

Every promising technology innovation comes with its own set of challenges. Multiagent AI systems are no exception. Strategically, organizations need to identify priority areas and use cases where AI agents can have the most rapid and valuable impact. Implications around change management also come into play, from training employees in new skills to modifying existing processes. At Deloitte we've gleaned valuable lessons that can help you realize the full value potential of this technology innovation.

As you explore the potential for multiagent AI systems for your organization, these considerations can help provide a valuable head start.

1 Starting smartly

With so many potential use cases for multiagent AI systems, it's important to be strategic about where to begin and how to move forward. Executive sponsorship and appetite, rigorous cost/benefit analysis, and a clear understanding of the state of your underlying data fabric form the foundation of use case prioritization and planning. To accelerate return on investment, proactive and thorough change management should be a part of any agent-powered transformation initiative, with an emphasis on building trust across your organization and among your stakeholders as new solutions are rolled out.

2 Pinpointing the right data, in the right context

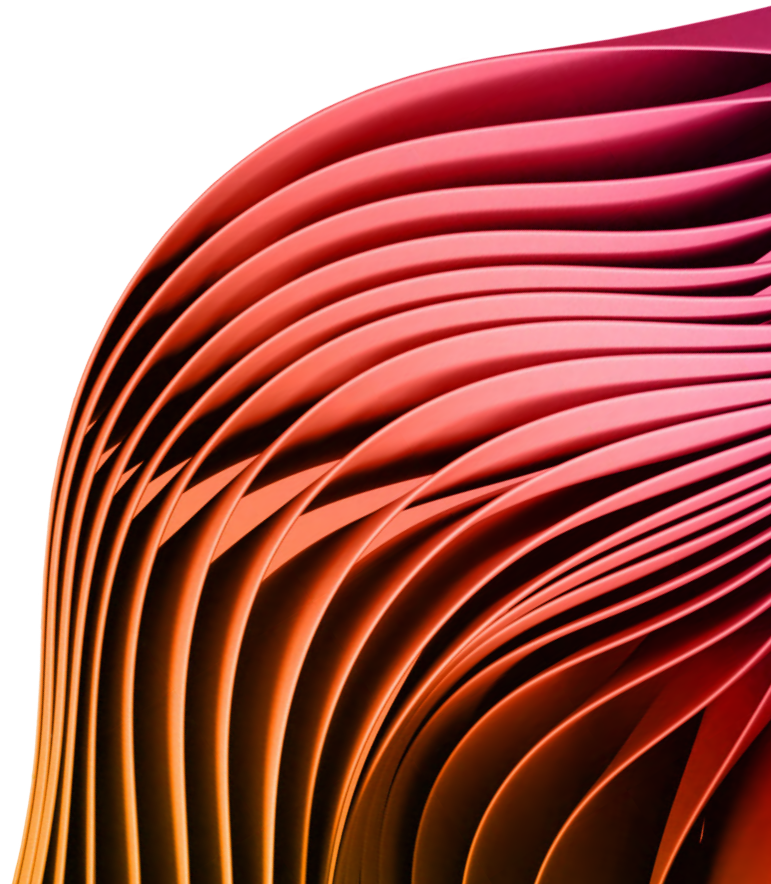
Data forms the backbone of any agentic architecture. For every use case, it's essential to not only identify the authoritative source of data that the agents will use but also ensure that agents can evaluate the appropriate context for that data. This is where knowledge engineering comes into play: By organizing data (i.e., knowledge sources) into a classification system or taxonomy, you make it easier for agents to navigate and retrieve the right data.

3 Tapping talent

Your system's design and development will require data engineering, business process engineering, machine learning and application architecture knowledge—in other words, some of the most high-demand skills in today's talent market. Accessing the necessary human expertise typically involves a combination of workforce upskilling and hiring, combined with strategic outsourcing to fill the roles that will be needed to support agentic AI transformation.

4 Evaluating technologies

There are numerous technology choices related to each layer of the agentic architecture. To simplify the process of selecting the right technology stack and agent development tool kit(s), consider leveraging an evaluation framework that helps to objectively score the choices at each layer to baseline the right-fit technology stack of the agentic architecture.



5 Decomposing processes

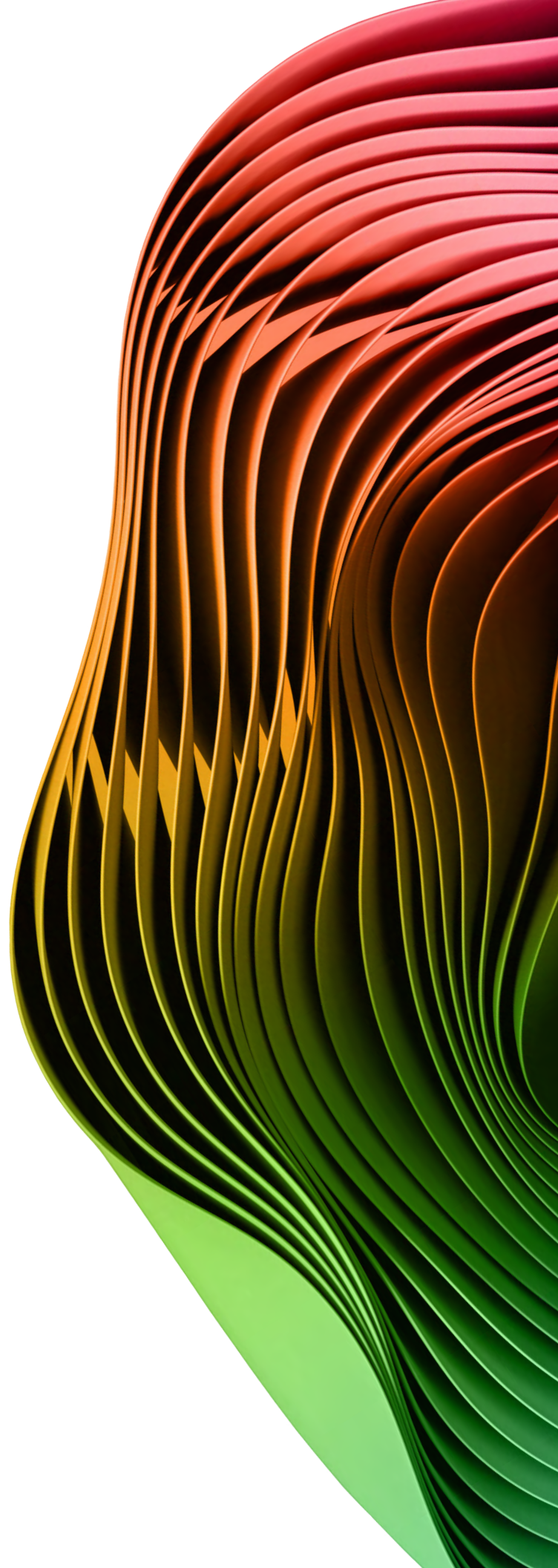
Reimagining an existing process or developing new agent-based workflows means breaking the overall process into smaller, more manageable subprocesses. By decomposing the process based on roles, each agent can specialize in a clear set of tasks, ensuring there are no overlapping responsibilities. To achieve this, consider using domain-driven design principles in which the boundaries for each subprocess are defined by and align with the organization's domain and team structure. This approach not only defines clear task boundaries but helps pinpoint the right number of agents to accomplish the overall process.

6 Scaling multiagent AI system impact with sound reference architecture

A thoughtfully designed reference architecture allows your organization to scale multiagent systems across a wide range of use cases in trustworthy and transparent ways. By embedding best practices and reusable components, this approach establishes a standard and repeatable process for design, deployment and continuous improvement. This not only ensures interoperability and reduces redundancy but also enables rapid adaptation and integration of best-fit agents for any emerging use case. It also provides a solid and ethical foundation for governance and optimization, ensuring that the multiagent AI systems remain aligned with enterprise goals and can evolve in response to changing needs and technological advancements. To design a reference architecture appropriate for your whole organization, we recommend taking into account industry best practices, market and customer expectations, and the technology, process and data realities of your own enterprise.

7 Embedding sound governance

It is very important to ensure that multiagent AI systems, once deployed in production, consistently generate quality outputs that do not introduce enterprise risk. Continuous monitoring and analysis of system outputs is critical to enabling timely identification of any potential anomalies or inaccuracies. It's important therefore to ensure that every multiagent AI system be smartly developed in ways that ensure multiple "checkpoints" within the workflow—and that checks and balances are engineered into each individual agent.



Making the cognitive leap

The rapid evolution of multiagent AI systems is transforming how organizations address challenges and streamline processes. This space is rapidly evolving as commercially available language models, frameworks and agents continue to improve. Organizations that adopt a systematic approach to multiagent AI system design and management will be well positioned to scale these systems effectively. Rather than limiting AI agent deployment to isolated business processes, a comprehensive approach allows for the expansion of AI capabilities across various use cases and domains.

By anchoring in the foundational principles we have outlined—and by leveraging a robust reference architecture that enables reuse and rapid adaptation of core components—organizations can maximize the potential usage and scale of multiagent AI systems. This approach helps empower organizations to derive more value from their AI investments, putting them not just at the forefront of technological advancement but giving them a competitive advantage.



Get in touch



Prakul Sharma
Principal,
AI & Data
Deloitte Consulting LLP
praksharma@deloitte.com



Sanghamitra Pati
Managing Director,
US India AI Leader
Deloitte Consulting LLP
spati@deloitte.com



Abdi Goodarzi
Principal,
GenAI Innovation Leader
Deloitte Consulting LLP
agoodarzi@deloitte.com



Vivek Kulkarni
Managing Director,
AI Transformation
Deloitte LLP
vivkulkarni@deloitte.com



Ed Van Buren
Principal,
GPS Applied AI Leader
Deloitte Consulting LLP
emvanburen@deloitte.com



Rajib Deb
Specialist Leader,
AI & Data
Deloitte Consulting LLP
rajideb@deloitte.com



Parth Patwari
Principal,
AI & Data Offering Leader
Deloitte Consulting LLP
ppatwari@deloitte.com

Contributors to this report:

Jim Rowan, Brijraj Limbad, Pradeep Gorai,
Caroline Ritter, Brendan McElrone, Laura Shact

Endnotes

1. Deborshi Dutt, Beena Ammanath, Costi Perricos and Brenna Sniderman, *Now decides next: Moving from potential to performance*, Deloitte, August 2024, p. 10, <https://www2.deloitte.com/content/dam/Deloitte/us/Documents/consulting/us-state-of-gen-ai-q3.pdf>, accessed December 3, 2024.
2. Benjamin Finzi, Brett Weinberg and Elizabeth Molacek, *Winter 2024, Fortune/Deloitte CEO Survey*, Deloitte, 2024, p. 11, <https://www2.deloitte.com/content/dam/Deloitte/us/Documents/us-winter-2024-fortune-deloitte-ceo-survey.pdf>, accessed December 3, 2024.
3. Dutt et al, *Now decides next: Moving from potential to performance*, p. 8.
4. Marvin Minsky, *The Society of Mind*, New York: Simon & Schuster, March 15, 1988, ISBN 0-671-60740-5.
5. Robert C. Martin, *Agile Software Development: Principles, Patterns, and Practices*, Prentice Hall, 2003, p. 95. ISBN 978-0135974445.
6. KaShun Shum, Shizhe Diao and Tong Zhang, *Automatic Prompt Augmentation and Selection with Chain-of-Thought from Labeled Data*, Cornell University, February 27, 2024, <https://arxiv.org/abs/2302.12822>, accessed September 16, 2024.
7. Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer and Huan Sun, *Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters*, Cornell University, June 1, 2023, <https://arxiv.org/pdf/2212.10001>, accessed September 16, 2024.
8. Wang et al, *Towards Understanding Chain-of-Thought Prompting*.
9. California Legislative Information, "Senate Bill 1120 Health care coverage: utilization review," September 30, 2024, https://leginfo.ca.gov/faces/billTextClient.xhtml?bill_id=202320240SB1120, accessed December 3, 2024.



Deloitte.

About Deloitte

As used in this document, “Deloitte” means Deloitte Consulting LLP, a subsidiary of Deloitte LLP. Please see www.deloitte.com/us/about for a detailed description of our legal structure. Certain services may not be available to attest clients under the rules and regulations of public accounting.

Deloitte provides industry-leading audit and assurance, tax and related services, consulting, financial advisory, and risk advisory services to nearly 90% of the Fortune Global 500® and thousands of private companies. Our people deliver measurable and lasting results that help reinforce public trust in capital markets, enable clients to transform and thrive, and lead the way toward a stronger economy, a more equitable society, and a sustainable world. Building on its 175-plus year history, Deloitte spans more than 150 countries and territories. Learn how Deloitte’s approximately 457,000 people worldwide make an impact that matters at www.deloitte.com.

This publication contains general information only and Deloitte is not, by means of this publication, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This publication is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your business. Before making any decision or taking any action that may affect your business, you should consult a qualified professional advisor. Deloitte shall not be responsible for any loss sustained by any person who relies on this publication.