
NTT DATA
Agentic AI:
Instilling the right balance
Calibrating performance, controllability and flexibility in AI agents

Contents

Delivering transformative business value

The rapid rise of AI agents makes the promised impact of GenAI increasingly tangible

We're at a point in the generative AI (GenAI) hype cycle where questions about enterprise viability and return on investment are at the forefront. The frenzied activity of the last 18 months — ideation sessions, proofs of concept (POCs), roadmapping and so on — is giving way to some palpable doubt and frustration. IT and business leaders, and the boards to which they are accountable, are wondering whether GenAI will deliver transformative business value.

It's a common pattern in the history of technology innovations: from the PC to the internet to the cloud, every breakthrough has its dips and doubts, and GenAI will be no different. What is different about GenAI is the speed at which innovation is taking place and the scope of its transformative impact.

This means the first wave of doubt is likely to be a short one, and what lies beyond it — the rapid rise of agents — will make the promised impact of GenAI increasingly tangible.

What are GenAI agents?

Simply put, agents can act on a user's behalf. Most of the GenAI POCs and solutions we've seen to date focus on retrieving knowledge from unstructured data. They use a retrieval augmented generation (RAG) structure and are highly effective at delivering quick, accurate answers. But their value tends to end there.

How are GenAI agents different from traditional approaches?

Agents enhance traditional large language model (LLM) approaches, making them more usable and valuable. Rather than relying on simple point-to-point communication between the LLM and the user, they use reasoning capabilities to carry out specific tasks. These agents use the LLM to perform a series of activities and adjust in real time to optimize performance. As conditions change, they respond with new approaches to achieve the desired outcome.

It's this ability to carry out tasks autonomously without waiting to be instructed that makes AI agents so powerful. They can drive execution, allowing organizations to create value and increase productivity even as the environment changes.

Areas where agentic solutions deliver transformative business value

Here are some of the specific capabilities that agentic solutions bring to the table — and which make them more likely to deliver greater business value.

Reduce time to market

Reasoning capabilities in agents can **speed up the implementation of new operating and business models**, and get products to market faster.

Reduce human error

The **automation of routine tasks** such as data entry and processing **reduces the risk** of human error.

Improve customer satisfaction

Automated customer-service solutions and interactions enable **AI agents** to **handle** inquiries and provide **personalized assistance, 24x7**, allowing human agents to tackle more complex issues.

Improve employee satisfaction

When routine tasks are automated, employees are freed up to focus on strategic initiatives, **innovation** and **creative problem-solving.**

Inform decisions

By analyzing vast amounts of data swiftly and automating decisions based on this data, agents can identify **trends** and **opportunities** that might otherwise remain hidden.

Decrease costs and increase performance

By analyzing supply chain and operational data in **real time**, agents enable businesses to **optimize resource** allocation, leading to cost savings and improved overall performance.

Enable greater efficiency and productivity

Agentic AI is revolutionizing business operations through intelligent automation that **enhances decision-making** capabilities and transforms departments across the organization.

Characteristics of agentic AI

Agentic AI refers to autonomous or semiautonomous software entities that use AI techniques to perceive, reason, plan, make decisions, take actions and achieve goals in digital or physical environments.

Flexibility and adaptability*

- Don’t require explicit inputs.
- Don’t produce predetermined outputs.
- Receive instructions and generate dynamic, goal-oriented outputs.

Autonomy

- Have agency and the ability to choose which action(s) to take to achieve a particular outcome.
- Agents could have different levels of autonomy, depending on the degree of freedom permitted by the specific use case.

Goal-oriented*

Execute actions to achieve or enhance the predefined outcome.

LLMs*

Allow agents to understand problems, interpret the specific context of each situation, and handle structured and unstructured data.

Reasoning*

Capable of planning and breaking down large goals into smaller, manageable tasks

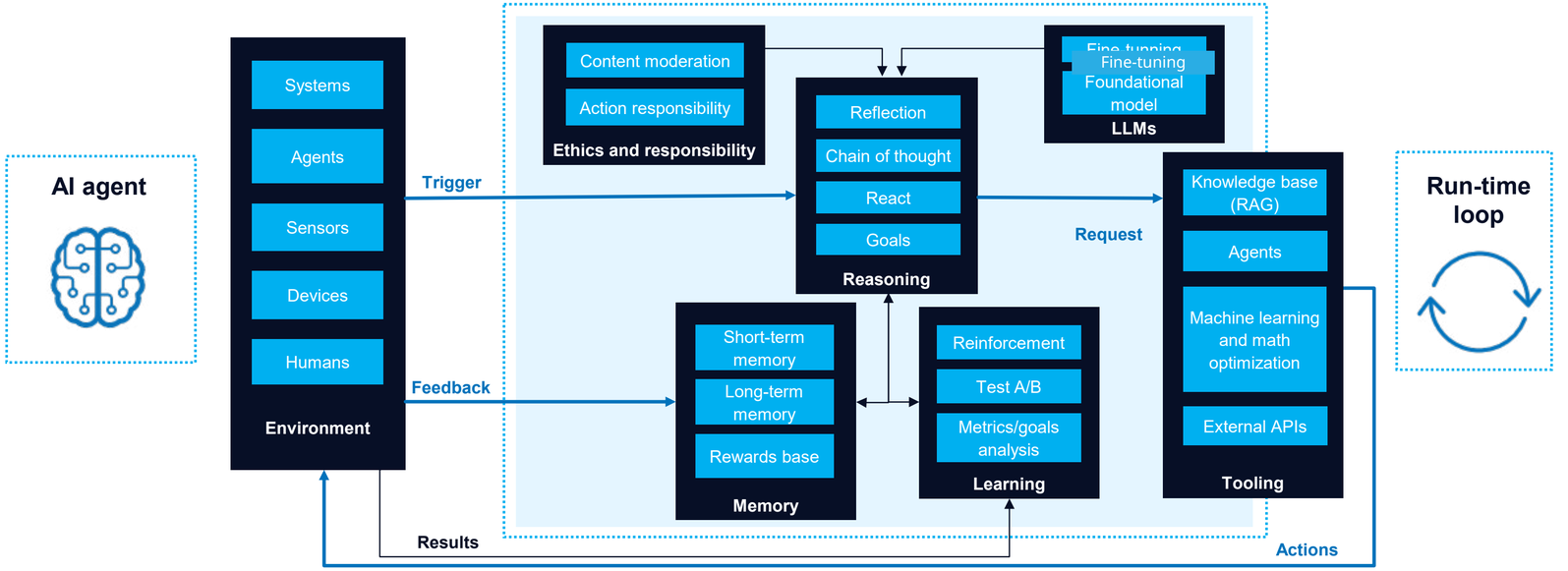
Learning

Optimize decisions based on past experiences or user behavior received back in feedback loop.

**Standard characteristics*

How agents work

Agents are designed to extend LLM capabilities to solving problems. Robust problem-solving capabilities enable agents to perform novel tasks. To solve real-world problems effectively, agents require the ability to reason. They also require call tools that interact with the external environment. This tight synergy between "acting" and "reasoning" allows agents to perform new tasks quickly, even in previously unseen circumstances.



The agentic system

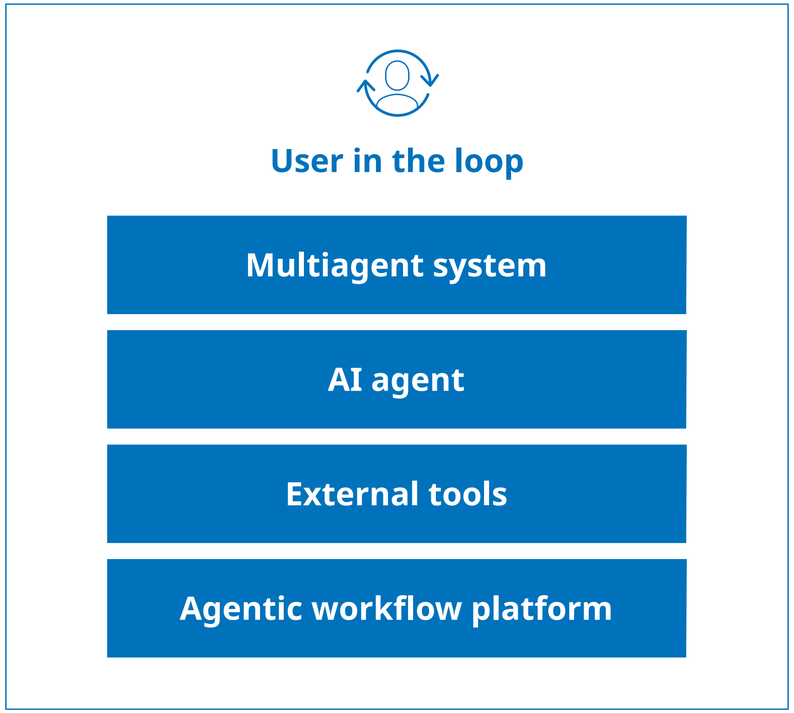
An agentic system is designed to accomplish a specific goal by executing a series of tasks. It comprises a group of AI agents working together to accomplish these tasks, leveraging their collective capabilities to maximize efficiency and effectiveness. Designed for adaptability, agentic systems can learn from experience to enhance their performance. They are pivotal in automating complex decision-making, problem-solving and task-coordination across domains.

Multiagent systems

Multiagent systems are composed of **collaboration** agents, each with **specialized** capabilities, **working together** to achieve a defined goal by sharing information and dividing tasks in an adaptative and flexible way in a changing environment.

External tools

External tools are **corporate resources** or capabilities that an agent **dynamically** invokes to **extend** its functionality beyond internal capabilities — for example, corporate APIs and retrieval solutions. They provide the **foundation for AI agents** to go beyond basic tasks and **tackle complex challenges** effectively.



AI agent

The agent serves as the fundamental building block, or atomic unit, designed to support specific stages of a business process or application functionality. These agents are built to handle a range of **tasks on their own, adjusting to different needs with efficiency.**

Agentic workflow platform

The platform is a stack of **integrated components** capable of executing, orchestrating and enabling communication between **multiple running agents** while seamlessly integrating them with the surrounding environment. It serves as a **central hub for coordinating the activities of autonomous agents**, ensuring collaboration, efficient resource allocation, and real-time interaction with external systems, data sources and APIs.

New technology brings new challenges

Agentic systems have a long list of high-value capabilities, but they're not without challenges. As you design and deploy these systems in your organization, consider these seven key issues — and how to address them.

1. Potential for added complexity

Agentic solutions are new, so not many of them have been used in large-scale enterprises for a significant period of time. This means there isn't a body of knowledge or best practices you can draw on as you chart your course with agents. You're innovating and improving on the fly.

Mitigation tips

- Build your agentic solutions using a modular approach that delivers interoperability. Think about your architecture with the goal of making it as flexible as possible so you won't have to throw away your hard work as technology evolves.
- Treat agent development projects more like innovation initiatives than traditional IT projects so you can iterate along the way. Use a sandbox approach to experiment with your emerging technologies. Include technologies you've vetted at an initial stage in your development environment.
- Engage third-party partners for some vetting and experimenting processes. Top firms are deeply connected with the tech startup community and better equipped to stay on top of key developments in this incredibly fast-moving GenAI environment.

2. Reliance on high-quality data

Agents rely on data to take action or complete a task. As with most GenAI solutions, the quality of their outputs relies heavily on the quality of data provided. Organizations have been struggling with this challenge for many years. Can we realistically overcome it? We would argue that enterprise data will never be perfect, and trying to make it so could significantly stall your progress. However, there are ways to get your enterprise data to the extent that you are able to deploy agentic AI and realize tangible value from it.

Mitigation tips

- Use a capability roadmap that's grounded on the incremental value you can deliver, not on the technology or data quality. For example, the first phase would deliver 30% of the total value of your GenAI initiative (and only require a portion of your data to be in prime condition). The next phase unlocks 60% of the total value, and so on. In this model, you work with data workstreams to get them ready for the next stage of value — you don't prepare everything upfront, which would delay the delivery of incremental business benefits.
- Data governance is a key element in ensuring high-quality data. Launch a data-governance program, if you don't already have one, and update existing programs to align with evolving needs and standards.

3. Mature AgentOps are required for success

Like the ML and AI models that came before GenAI, it takes a team, a process and a platform to keep these models on track. If you have an existing MLOps program in place, you may think you can use the same or a similar approach for agents. Unfortunately, that's not the case. You need robust AIOps — also called LLMOps — to safeguard against more than model drift. Effectively managing agent workflows is a more involved process. It includes special guardrails, continuous monitoring and a robust user-feedback loop.

Mitigation tips

- Assemble an open-minded team who can identify the specific challenges that come with managing LLMs. This will enable you to assign the appropriate resources, either internal or external, to address any issues that arise.
- Create a "tiger team" to track and evaluate new tools, datasets and frameworks that may help. It's a new area, but it's evolving rapidly. Staying on top of new options is likely to pay off.
- Consider implementing an LLM-as-a-judge construct which leverages the capabilities of LLMs to automate some of the monitoring and management of GenAI responses. LLMs are scalable and cost-effective and can adapt to judging different types of responses without extensive retraining. That said, take this approach with caution: LLM training data can have inherent bias and the reasoning behind responses is not always clear.

4. Data access and privacy concerns

Certain tasks require agents to have access to internal data to act on a person's behalf. This could raise concerns about governance vulnerability.

Mitigation tips

- For especially sensitive situations, consider instituting access controls, limiting the actions agents can take and creating walled-off environments for agents, thereby limiting their access to specific data sources and systems.
- Implement real-time agent-activity monitoring that generates alerts for suspicious behavior. Compliance checks and audits on a consistent schedule can also be effective in tracking and minimizing potential data-security risks.
- Consider private GenAI models that are designed to operate in secure, controlled environments and that maintain data privacy and confidentiality. These models process and generate content entirely within predefined boundaries, such as on-premises servers or private cloud infrastructure, without exposing sensitive information to external entities.

5. Cost management

Although the cost of GenAI systems is decreasing, unnecessarily or improperly applied agents can be expensive and inefficient. Carefully evaluate the value within a workflow.

Mitigation tips

- Use a multimodel approach. Because some models are cheaper than others, using a hierarchy of multiple models allows you to cycle through them from least to most expensive. It's an effective method for weighing the performance you need against the cost involved.

6. Creativity and “hallucinations”

“Hallucinations” are the opposite side of creativity. For certain use cases, the tolerance for hallucinations may be high. However, in areas like finance or operations where accuracy is critical, minimizing hallucinations is essential.

Mitigation tips

Evaluating LLMs is a challenging and evolving domain. Traditional AI metrics such as accuracy and the receiver operating characteristic curve are inadequate in this area. The dynamic nature of LLMs and their vast potential for applications make validation difficult. So, consider the following tactics:

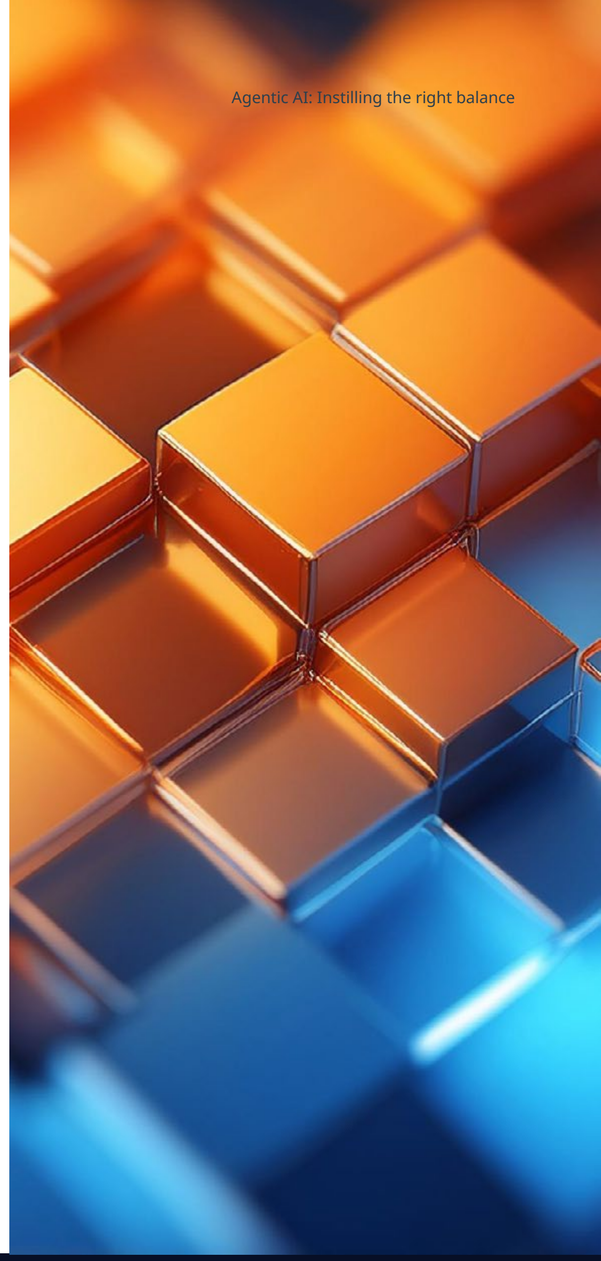
- Include users in the project from the very beginning.
- Create a list of questions and expected answers for the initial stage of the project.
- Create a list of non-tolerated answers to build the first definition of the guardrails layer.
- Train users throughout the organization so they know that it is very difficult to evaluate the outputs using traditional metrics (for example, two summaries of the same document can be equally good despite using different words and sentence structures).
- Establish a comprehensive LLMOps practice that can evaluate the performance and results of the model in production.

7. Risk and trust

Normal security risks are not the only consideration for AI agents. Because they have agency — the ability to choose which action(s) to take in order to achieve a particular outcome — it is critical to understand what can go wrong as the agent takes such actions. Is the agent using accurate data? Is the agent making good decisions, and are the adaptations it makes over time accurate? Can it be hacked?

Mitigation tips

- Start with controlled initiatives where the level of autonomy is low. Always employ a user-in-the-loop strategy to validate the final step of the process before committing to any action.
- Increase the agency of the initiatives in small steps, gaining the trust and confidence of both the company and users.
- Include cybersecurity profiles in the initiatives, and define an ethics and compliance guide aligned with the company’s strategy and values.



Where to leverage autonomous agents: Understanding the tradeoffs

Not all applications benefit from autonomous AI agents. Although these agents offer flexibility and can handle complex tasks, they may not always be the best choice. It's essential to weigh reliability and consistency against flexibility and adaptability, and to understand the tradeoffs. Controllable workflows ensure consistent outcomes, but autonomous agents provide the flexibility needed for complex tasks. For complex use cases, autonomous agents must be tested thoroughly before deployment to ensure they provide the desired outcome.

Controllable workflows

For applications with a clear, predictable workflow (such as data processing and task automation with well-defined steps), it's better to use systems with a **controllable workflow**. These systems provide **reliability and consistency**, ensuring that every task is performed in a predictable and cost-effective way.

Autonomous agents

When the task is broad and requires **numerous complex actions**, or the environment is highly dynamic, autonomous agents may be a good option. These agents bring **flexibility** and the ability to adapt to changing circumstances, but they can introduce **uncertainty and variability** in results.

Main domain-specific use cases

HR

- Training and development recommendations
- Employee onboarding automation
- Employee wellbeing monitoring
- Résumé screening and shortlisting
- Employee performance report summarization
- Corporate presentation generation
- Diversity and inclusion analytics
- Automated job-description generation

Marketing

- Personalized marketing and service recommendations
- Social media monitoring and engagement
- Campaign automation
- Market research
- Content creation and copywriting
- Predictive analytics for campaign performance
- Competitor analysis
- SEO optimization

Customer support

- Automated FAQs and troubleshooting guides
- Chatbots for instant customer assistance
- Conversation summaries
- Sentiment analysis in customer interactions
- Smart ticket-sorting
- Predictive ticket-resolution
- Knowledge-base enhancement and maintenance
- Customer feedback analysis
- Multilingual support through translation services

IT

- Predictive maintenance
- Security threat detection
- Capacity planning
- Automated code review
- Test-data generation
- Automated IT help-desk support
- Natural language processing for data analysis
- IT asset management
- Automated report generation

Legal

- Legal document categorization, summarization and analysis
- Legal case synthesis and brief generation
- Legal drafting assistance
- Automated contract review
- Legal compliance monitoring
- Due diligence automation
- Litigation prediction

Main industry-specific use cases

Banking

- Loan processing system
- Personalized financial planning
- Virtual financial adviser
- Credit-scoring assessment
- Fraud detection and prevention
- Loan application document classification
- Antimoney-laundering (AML) compliance

Insurance

- Claims processing optimization
- Insurance underwriting automation
- Automated regulatory compliance
- Dynamic pricing strategies
- Personalized insurance-plan recommendations
- Fraud detection in insurance claims

Telecom

- Information extraction from installation documentation
- Infrastructure maintenance agents
- Natural language processing for voice assistants
- Music for call waiting
- Conversation summarization

Healthcare

- Medical image analysis
- Personalized treatment plans
- Drug discovery and development
- Research-paper summarization and synthesis
- Healthcare chatbots and virtual-health assistants
- Genomic data interpretation

Public sector

- Citizen query-resolution chatbot
- Automated compliance monitoring
- Emergency response systems
- Legislative communication
- Automated document management
- Predictive policing and crime prevention
- Smart-city planning and traffic optimization

Main industry-specific use cases

Energy and utilities

- Audit processes verification
- Processing and validation of electrical certificates
- Energy consumption and market forecasting
- Predictive maintenance for equipment
- Grid optimization and management
- Smart-meter analytics for customer insights

Retail

- Virtual try-on and augmented-reality experiences
- Personalized product recommendations
- Inventory management and demand forecasting
- Virtual shopping assistants
- Visual search and image recognition
- Dynamic pricing optimization

Manufacturing

- Product design and optimization
- Predictive maintenance for machinery
- Quality control and defect detection
- Supply chain optimization
- Customized production planning
- Natural language processing in manufacturing documentation

Successful multiagent use cases

Commercial smart adviser

Business need

An insurance company wanted to improve **customer service** and optimize their **policy issuance** and **underwriting** processes. Their main challenges were:

- **Inefficient customer service:** Customers experienced delays and difficulties in getting quick and accurate responses.
- **Slow manual processes:** Policy issuance and underwriting were time-consuming and therefore had a negative effect on customer satisfaction.
- **Untapped cross-sell opportunities:** Lack of integration hindered personalized recommendations for additional products that may be of interest to customers.

Technology

- Microsoft Azure
- OpenAI
- LangGraph
- LangChain

Solution

To maximize the sales experience and improve efficiency, a **hyperpersonalized conversational digital application** was developed, based on **several agents**, to optimize processes. The application enabled:

- **Customer recommendations and query resolutions** by offering accurate and relevant answers to customer queries in an automated and efficient manner.
- **Online policy underwriting and issuance**, simplifying and speeding up these critical processes to significantly reduce response times and improve the customer experience.
- **Integration with a virtual assistant**, allowing the company to maintain a fluid and efficient interaction with customers, improving both the user experience and operational efficiency.
- **Conversation moderation** by implementing effective controls to ensure each interaction remains within the boundaries of the defined function and content.

Outcomes

Speed to respond

- Customers now get quick and accurate responses to their queries, significantly improving customer satisfaction.

Customer satisfaction

- 24x7 customer service and ongoing support.

Income through cross-selling

- Opportunities are taken to recommend additional products, increasing sales across business units and improving personalization in recommendations.

Potential

- Transform the sector’s value chain by eliminating intermediaries.
- Future evolutions can include using photos to validate insurance terms.

Successful multiagent use cases

User profiling for targeted advertising

Business need

- Improve ad relevance for the 8 million users of a television streaming service by **creating psychographic profiles** based on content preferences.
- Identify detailed user preferences across demographics, behavior and viewing patterns to **optimize ad targeting**.
- Address the complexity of handling large volumes of user data to **develop meaningful insights efficiently**.

Technology

- AWS
- Mistral AI
- CrewAI

Solution

- Used a **multiagent autonomous system** (CrewAI and LangChain) integrated within AWS to automatically analyze and categorize user profiles.
- **AI-driven agents** (Profiler, Chief Director, Advertising Specialist) combine to build unique psychographic profiles by processing demographic, behavioral and transactional data.
- Insights allow for **hyperpersonalized ad recommendations**, enhancing user engagement and ad ROI.

Outcomes

- Generated psychographic profiles for millions of users based on their unique viewing habits.
- Automated profiling reduced time and resource costs.
- Tailored ad recommendations increased user engagement and ad relevance, delivering a measurable impact on advertising effectiveness.
- Replacing the Amazon Titan model with the Mistral 7B model significantly reduced the pipeline execution time.

Relevance

- Generative AI enables autonomous and dynamic profiling and ad targeting based on users' preferences, without prior data categorization.
- The multiagent framework identifies and categorizes behavioral patterns, and refines ad suggestions based on psychographic insights.
- The system continually learns and adapts, allowing for real-time refinement in ad-personalization strategies.

Successful multiagent use cases

Insurance assistant

Business need

- An insurance company faced significant challenges with **high churn rates** among life-insurance policyholders, primarily because policyholders did not understand their policy benefits and encountered difficulties in obtaining answers to policy-related queries.
- The goal was to reduce churn rates by providing a tool that **enhances the user experience and answers customers' questions.**

Technology

- Microsoft Azure
- OpenAI
- LangGraph
- LangChain

Solution

NTT DATA developed a **GenAI conversational multiagent system** that's accessible via the insurance company's existing WhatsApp channel. The system includes the following agents:

- **Supervisor agent:** Interprets customers' questions and directs them to the appropriate agent.
- **Policy details agent:** Answers specific questions about the customer's life insurance policy, such as coverage and beneficiaries.
- **Retrieval agent:** Answers general procedural questions, like how to change beneficiaries.
- **Health and wellness agent**: Provides recommendations about the company's health and wellness program and tracks the customer's progress in using it.
- **Dumpster agent:** Handles unrelated queries by providing generic responses.

Outcomes

- **Enhanced user experience:** Customers receive immediate and accurate responses, significantly improving satisfaction and reducing churn rates.
- **Increased engagement:** Personalized and informative interactions foster greater loyalty and customer retention.

Successful multiagent use cases

Expert agents for a restaurant wholesale business

Business need

- **Improve the end-to-end sales experience** for customers in the web-based B2B company marketplace.
- **Increase customer engagement** by providing a personalized and expert advisory process with product recommendations based on past orders, current needs and company knowledge.
- **Increase sales** by improving the buyer experience.

Technology

- Microsoft
- OpenAI
- LangGraph
- LangChain

Solution

NTT DATA developed a **GenAI multiagent system** embedded in the customers' marketplace. Several agents were developed and deployed to be integrated seamlessly in the user's buying journey:

- **Products and prices agent:** Provides information about products, including product descriptions, prices and potential use.
- **Profiler agent:** Understands user profiles based on historical orders.
- **Commercial offers agent:** Specializes in current commercial offers and campaigns that need to be prioritized.
- **Recommender agent (main agent):** Recommends products based on current needs, and explains the recommendation in natural language.

Outcomes

- Personalized recommendations and tailor-made product suggestions.
- Optimized order and pricing procedures for efficiency.
- Better engagement and faster transactions.

Relevance

- Customers can explain their current need using natural language, for instance, "I have a restaurant, and I have a reservation for tonight for 500 pizzas, what do I need?"
- The agents are orchestrated to ask more questions, for example regarding size and ingredients. The system then creates a proposed order with the products needed.
- Recommendations are fully aligned with sales objectives and targets.

Successful multiagent use cases

Autonomous agent for a corporate investment bank

Business need

- Consultants in the corporate investment bank areas carry out an **intensive data-collection** process in response to new demands for an increase or definition of a credit limit.
- The process accesses **multiple sources**, with or without defined interfaces, to collect information and produce executive **summaries and insights.**
- The bank wanted to determine whether an **agentic** solution would simplify this process and implemented a POC to **validate** it.

Technology

- Microsoft
- OpenAI
- LangGraph
- LangChain

Solution

A digital application was designed using several agents, each with its own set of features and functionalities:

- Agent to gather all **public information** related to the company, including **news** and any other data, using web-crawling tools.
- Agent to gather structural economic data from **data brokers**.
- Agent gather data from the **Yahoo Finance** API and **SEC EDGAR API**.
- Agent to gather **internal financial data** from the bank's systems.
- Agent to **process** all data and **generate insights** with some **predefined** questions.
- Agent to **answer questions,** on demand.

The POC involved implementing the solution on an Azure platform using Azure Open API.

Outcomes

- Validated the hypothesis with the results of the POC.
- Defined the technology stack and platform.
- Delivered total cost of ownership of the solution.
- Demonstrated the applications to internal stakeholders.

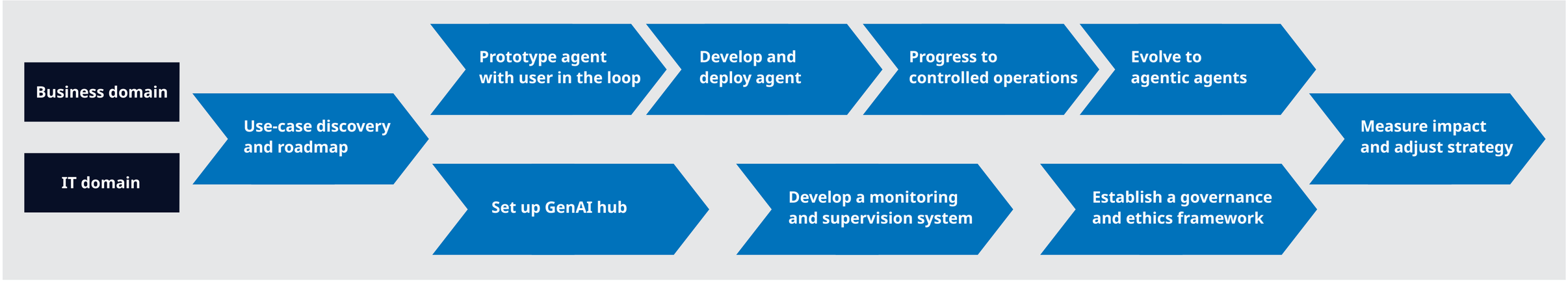
Framework components

- Azure data platform and data applications.
- The tooling layer comprises a set of custom tools which can be invoked by the AgentREST when required. Some of these are dedicated to extracting information from third-party data sources.
- The agent layer is where the Agent REST API obtains the user's query and operates with the LLM to perform the requested task.
- The Cross Tech Tools team built an LLM-based autonomous agent framework. The AI framework will be distributed to other teams to create AI use cases with all the capabilities that it brings.
- Implementation of LangChain and Confluent Kafka (sync process flow).

Partner with NTT DATA

Getting started with fully intelligent and autonomous agentic agents is ambitious, but it can also be risky, especially if you don't have experience with simpler agents or agents with lower levels of autonomy.

Testing first with agents that are enabled by GenAI but have a limited degree of autonomy can be an excellent way to overcome this. Follow a user-in-the-loop strategy to reduce risks and facilitate organizational learning. Taking this kind of phased approach allows you to experiment with advanced GenAI capabilities while at the same time building the foundation for a progressive transition to more complete agentic agents.



NTT DATA offers a practical, step-by-step approach to help organizations move forward with their AI initiatives. Our engagement models are designed to align with your organization's unique goals, focusing on phased implementations that make it easier to scale and measure real impact along the way.


NTT DATA