MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 35 Issue 2, 2025, Pages: 1889-1897

# Enterprise-Scale Gen AI Orchestration Using Small LMs and LLM Agents for Intelligent ITSM and HRSD Automation in Enterprise Ecosystems

**Siva Hemanth Kolla, Gen AI Research Scientist, siva.kolla.hemanth@gmail.com, ORCID ID: 0009-0009-2644-5298**

## Abstract

Enterprise-scale orchestration of Generative AI (Gen AI) provides an intelligent integration of ecosystem actors within processes and workflows. The deployment of small enterprise-specific language models (LMs) combined with autonomous Behavioural LLM Agent architectures supports sophisticated orchestration of IT Service Management (ITSM) and Human Resource Service Delivery (HRSD) automation Use Cases, Including Digital Incident Management, Employee Onboarding, Offboarding, and Activity Allocation. Unlike traditional orchestration techniques, which typically rely on hardcoded workflows and limited logic, Gen AI techniques allow for enhanced coordination among Process-Performer Agents with minimal human involvement during execution.

An orchestrating agent guides execution, managing Process Performer communication, triggering actions based on input messages, keeping track of execution state, and rolling back processes when required. Agents can collaborate or operate independently based on trigger conditions and a behaviour set associated with each Use Case. During execution, Gen AI supports the automated generation of documentation, audit logging, and telemetry collection for the orchestration infrastructure. In this way, the Gen AI approach enables the intelligent, autonomous orchestration of processes and workflows across enterprise ecosystems—a goal that is exceedingly complex to achieve using traditional techniques.

**Keywords:** Generative AI Orchestration, Enterprise Gen AI Architectures, Autonomous LLM Agents, Behavioral LLM Agent Frameworks, Enterprise Language Models, Intelligent Process Orchestration, AI-Driven Workflow Automation, IT Service Management Automation (ITSM), Human Resource Service Delivery (HRSD), Digital Incident Management, Employee Onboarding and Offboarding Automation, Agent-Based Systems, Multi-Agent Collaboration, Process-Performer Agents, Orchestrating Agent Design, Autonomous Enterprise Workflows, AI-Based Decision Logic, Event-Driven Agent Triggers, Execution State Management, Intelligent Rollback Mechanisms, AI-Generated Documentation, Audit Logging Automation, Telemetry Collection with AI, Enterprise Ecosystem Integration, Next-Generation Service Orchestration.

## 1. Introduction

Enterprise-scale Generative AI orchestration integrates enterprise ecosystem planning, scheduling, management, and control tasks, providing significant new capabilities for Information Technology Service Management (ITSM) governance and Human Resource Service Delivery (HRSD) processes.

The next generation of enterprise ITSM and HRSD automation leverages Generative AI within enterprise ecosystems orchestrated by Smart MPCs and Smart Controllers. The Smart MPCs and Smart Controllers cover the management of IaaS services and the governance of HSML Services, respectively. Generative AI adds meaningful automation capabilities that Enterprise Resource Planning (ERP) Systems cannot provide due to the multifactorial nature of the triggers, the impact of the changes being considered, and the requirements for a fluid employee experience in HRSD processes. In ITSM, the automation of incident and problem management represents the main driver of investment, with changes responding to a more limited cycle and requiring solution implementation approval from the management layer.

### 1.1. Overview of Generative AI in Orchestration

Generative AI empowers natural-language-based communication with machines. But Gen AI's writing, dialogue, image, and code generation capabilities — already harnessed in numerous applications — lack a key ingredient in traditionally engineered solutions: Intelligence. Tasks and automation flows therefore still need to be designed and engineered, and the labor, complexity, and cost of this orchestration work often outweigh the labor, complexity, and cost of the automation itself. With Global 2000 companies investing billions of dollars in Gen AI, it is timely and important to discuss its orchestration value specifically for Intelligent IT Service Management (ITSM) and Human Resources Service Delivery (HRSD).

Solutions to these orchestration challenges are sought in the Concordia research program, (1) which uses Generative AI responsible deployment principles (2) to investigate not only the business value but also business risk and Governance implications of Enterprise-scale Gen AI Orchestration. Concordia analysis shows that Gen AI agents are an enabling technology for Autonomous Orchestration, a specific form of orchestration that uses available data and Gen AI agent capabilities to transform information-hungry workflows into low-touch workflows that require little or no human input during execution.

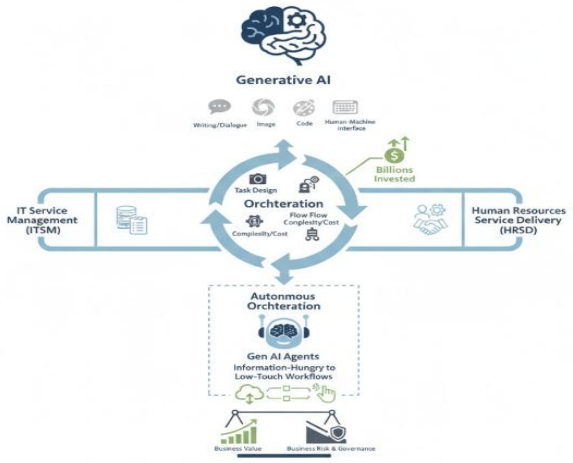


**Fig 1: Autonomous Orchestration: Advancing Enterprise Gen AI from Generative Output to Low-Touch Service Delivery**

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 35 Issue 2, 2025, Pages: 1889-1897

## 2. Conceptual Foundations of Gen AI Orchestration

The foundational concepts of enterprise-scale Generative AI Orchestration are delineated. At the core lies the notion of orchestration itself, which is understood as managing interdependence and cooperation among heterogeneous entities. Orchestration is further characterized by decision-making authority and, where such authority is absent or insufficient, by governance—the formalization of protectors' intents, purposes, and actions relevant to the interests of entities that lack the strength, courage, resources, or other prerequisites to fulfill these interests themselves or in combination. For effective enterprise-scale Generative AI orchestration, both direction (intent) and resources are crucial. Therefore, a set of orchestrating and resource-providing actors, together with the intents they express, are defined. Finally, orchestration workflows—sets of interrelated tasks performed by agents in response to an intent or collaborative agreement among agents—are identified, together with an assessment framework.

The idea of orchestration is borrowed from organizational behavior and design, where it refers to managing interdependence among actors in a way that assures cooperation, fulfillment of an overarching purpose, and (ideally) synergy. Contrast the function of orchestration with that of a traditional controller of a large army who seeks to direct all action. An effective enterprise orchestrator does not seek to control everything but invests in establishing an organizational ecology characterized by healthy relationships, collaboration, trust, and a shared sense of purpose; directing only where absolutely necessary. Orchestration is also different from leadership, which is concerned with setting visionary directions; at its best, orchestration can contribute fundamentally to realizing a vision and making it a reality.

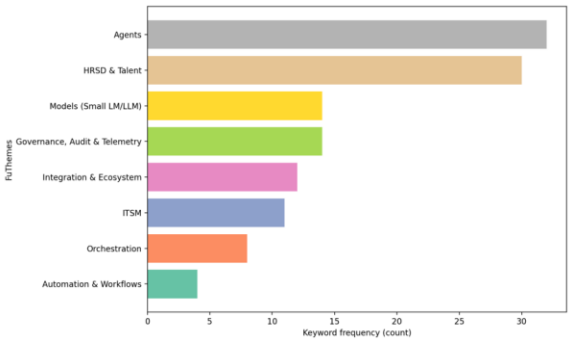


**Fig 2: Graph-Based Representation of Orchestration Workflows with Trigger-Driven Task Dependencies**

**Equation 1) Orchestration workflows as graphs (tasks + trigger edges)**

**Step 1 — Define the graph**

$$G = (V, E)$$

- $V$: elementary tasks
- $E \subseteq V \times V$: directed edges (i → j) meaning "i can trigger j"

**Step 2 — Define trigger predicates**
For each edge $e = (i \to j)$, define a predicate based on runtime context $x(t)$ (telemetry/messages):

$$g_e\big(x(t)\big) \in \{0,1\}$$

**Step 3 — Define enablement condition**
Task $j$ becomes enabled if some predecessor $i$ is complete and its trigger is true:

$$j \text{ enabled at } t \Leftrightarrow \exists i\colon (i \to j) \in E,\ \text{completed}(i,t) = 1,\ g_{(i\to j)}\big(x(t)\big) = 1$$

### 2.1. Small Language Models in Enterprise Contexts

Small language models (LMs) 1 to 100 billion parameters may be more cost-effective than large language models (LLMs) for enterprise use cases involving confidential data, latency-sensitive applications, or highly governed domains. Like LLMs, trained from the beginning, LMs may be trained by beginning with general data and following up with domain-specific data. LMs may also be safety-tuned to minimize hallucination and toxicity. A systematic classification provides guidance on these aspects and suggests that small language models outperform large language models for reasonable tasks.

Most applications involve only a small subset of possible intents, and the model is likely to see only a small fraction of the vocabulary during inference; hence, a smaller model that is focused on the use case can yield similar or even better performance than LLMs. A recent evaluation test InstructGPT versus FLAN-T5-Small models with 60 million parameters—a small model versus 137 billion parameters—found that the smaller model performs significantly better on multi-choice reasoning tasks that assess knowledge of the world and reasoning ability, and similarly on closed-book question answering tasks with knowledge of the world, while reliability and long-form answer tasks are much worse.

### 2.2. LLM Agents and Autonomous Orchestration

LLM Agents and Autonomous Orchestration

LLM agents implement the agentic decision-making method, characterized by LLMs that incorporate action-oriented, task-oriented capabilities into an agentic architecture. These agents decompose high-level intents into concrete tasks through (conditional) planning, enabling the launch of subordinate specialty LLM instances with explicit, low-level intents. Principled negotiation principles govern task assignments among co-located autonomous agents to harmonize business-as-usual functions and ensure coordinated reactions to large-scale changes. Monitoring for supplier reliability, including LLM health alerts, initiates escalation to assess any deteriorating performance conditions. Failing supplier capabilities trigger

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 35 Issue 2, 2025, Pages: 1889-1897

transformation processes to LLM-supported regimes, where supervision and consolidation guarantee outcome correctness. A safety-oriented, stepwise testing-and-approval method extends multi-agent tasking and agent-based, autonomous process initiation to other businesses.

Orchestration automation leverages decomposition and negotiation for co-located LLM agents, where action-oriented prompts enable negotiation among individual LLMs responsible for assigned tasks, drawing on reflexive knowledge and capabilities. Autonomous LLM agents have been introduced for business-as-usual management of repetitive tasks that demand little behavioral variability. The next level of complexity allows multitasking among an ensemble of LLMs, such as during onboarding cycles when education, training, and introduction to future changes concurrently require attention. Yet untrained LLM agents cannot be connected without supervision and audit capabilities.

## 3. Architectural Paradigms for ITSM and HRSD Automation

Enterprise-scale Gen AI orchestration leverages Generative AI for enterprise IT Service Management (ITSM) and Human Resource Shared Services Delivery (HRSD). Enterprise ITSM and HRSD environments typically comprise large numbers of smaller tasks that require a vast domain knowledge and are labour-intensive to complete. These tasks appear to be repeatable yet are difficult to express in simple rule-based systems. Generative AI techniques offer powerful solutions but, until now, have not been applied to these specific orchestration functions.

Three types of architecture design activity are important for Enterprise-scale Gen AI orchestration to succeed: architectural paradigms, advanced support techniques, and process or scenario flows. The former define mechanisms for practical, safe, rapid, and complete deployment of Generative AI components in enterprise ITSM and HRSD workflows. Small Language Models (LMs), because of their low cost and enterprise data safety, are proposed for functions such as ITSM change approval notifications, while Large Language Model (LLM) agents can play roles for user onboarding and offboarding. LLM agents support much greater autonomy. More complex functions in the ITSM and HRSD workflows involve multiple, interdependent, or concurrent tasks needing collaboration among LLM agents.

### 3.1. Data Governance and Compliance

Data governance defines how an organization manages the availability, usability, integrity, and security of the data employed in the fulfillment of its business objectives. It encompasses data lineage, privacy, access control, retention, and regulatory mapping. Data governance must therefore ensure that data movement is properly monitored and tracked throughout its entire lifecycle, enabling the analysis of data patterns and the assessment of any potential compliance issues. The risk of breaches or violations must be controlled, supported by proper auditing and validated incident response processes.

Data-age and access-control policies must be enforced, with enterprise custodians maintaining full access rights. Data must be held only for the time necessary to accomplish the intended purpose, aligned to the relevant legal and internal policies. An owner must be designated for each piece of sensitive data, mapping privacy regulations and responsibilities related to the management of each type. Sensitive data should only be accessible to authorized individuals and associated to specific compute resources judged appropriate for hosting the sensitive data required for the operation of the workload.

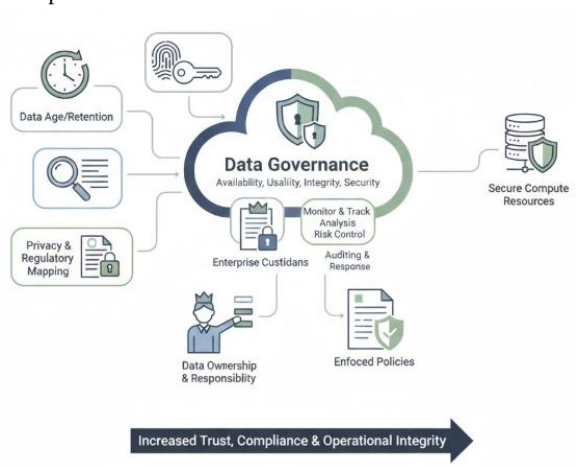


**Fig 3: A Lifecycle-Centric Framework for Enterprise Data Governance: Integrating Regulatory Mapping, Access Control, and Proactive Risk Mitigation**

### 3.2. Integration Patterns within Enterprise Ecosystems

Within an enterprise context, Gen AI orchestration encompasses a range of processes that may operate across the organization's external boundaries. Integration with external parties, products, and services (e.g., Supply Chain Management (SCM), Customer Relationship Management (CRM), Industry Cloud, talent marketplaces) form an essential component of many Gen AI workflows. These integrations can adopt an API-based paradigm, where LLM Agents invoke services exposed by connected systems, or be driven by data/event-publishing from the LLM Agents that others are subscribed to and consume. More specifically, Azure-supported enterprise integrations with SAP, ServiceNow and Salesforce are common, allowing transactions to be processed directly within the respective system, and notifications/states to be published along Azure Service Bus topics.

Although both approaches are valid, they represent different patterns and have varying strengths. If supported, the event-driven pattern allows Gen AI workflows to passively await arrival of telemetry direct from other systems or chat interfaces, potentially reducing average decision latency by a factor proportional to the time taken by the LLM Agent(s) to process and negotiate a response to the request. On the other hand, API-based services allow Gen AI calls to actively initiate cross-ecosystem processes. Residing within

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 35 Issue 2, 2025, Pages: 1889-1897

a message-bus integration are systems that follow neither of the prior patterns.expose neither consuming service nor subscribing to any data but are nevertheless important to the enterprise ecosystem. Other systems, such as an Enterprise Data Warehouse (EDW) accessing all enterprise data from a single common layer, are also outside the enterprise perimeter but nevertheless belong to its ecosystem. In the Gen AI context, they are typically used for auditing purposes and consume telemetry from within the wants to all ongoing processes (for the AI community), for retention of conversations beyond what the Gen AI data retention policy allows (for records management), or for service quality assessment (for the service delivery teams). An Enterprise Security Management (ESM) service can provide long-term data retention to securely capture these events.

## 4. Intelligent ITSM Automation with Small LMs

Incident Management and Response
Small language models have a proven ability to comprehend and generate comprehensible human-like narratives. These abilities can be harnessed in incident management scenarios to simplify user interactions, allowing users to submit IT requests in natural language without violating information security policies. Internal knowledge bases can also be leveraged to draft appropriate responses for both common and more sophisticated IT issues based on the nature of the incident request. Being inherently safer and cheaper to operate, small language models are the preferred choice here.
Supported by sufficient deliberation and testing data, reliable automated enabling capabilities can be provided. Real-time dashboards and suitable analytics can be configured in the underlying systems to observe the overall performance levels and ensure that no declining trends are overlooked. LLMs being capable of expressing their uncertainties during decision-making, monitoring systems can be set up to watch for cases where confidence levels are low, and appropriate action thereof can be triggered, thereby still ensuring an effective safety net even when relying upon small models.
Change and problem management implications, automation boundaries, escalation protocols
However, automated responses to incident requests must be managed as per standard operating procedures. Mission-critical incidents cannot be handled entirely through automation; rather, such cases must also be routed to the appropriate resolution teams for attention. A suitable arbitration mechanism based on a knowledge base can be set up to classify the incoming incident requests into the aforementioned categories. Any direct decision or response must also be verified against the knowledge sources through strict checks by an appropriate human expert in a test environment before being released to the production environment.

**Equation 2) Multi-agent orchestration as a state machine**

**Step 1 — Define the state and inputs**

- Let the workflow state at time $t$ be: $s(t) \in S$ (finite set of states)
- Let incoming message/event be: $m(t)$
- Let "secondary states / condition variables" mentioned in the paper be: $\kappa(t)$

**Step 2 — Write the state transition equation**
A state machine updates via a transition function $\delta$:

$$s(t+1) = \delta\big(s(t), m(t), \kappa(t)\big)$$

This directly formalizes "active state specifying the current operation and topic context" plus condition monitoring.

**Step 3 — Model the router/subroutine selection**
The paper says the router selects a subroutine based on keyword matching.
So define a policy/router:

$$a(t) = \pi\big(s(t), m(t)\big)$$

where $a(t)$ is the chosen subroutine/action.

**Step 4 — Model subroutine execution (message + guards update)**
Each action transforms the message/context and updates condition variables:

$$\big(m_{\text{out}}, \kappa(t+1)\big) = f_{a(t)}\big(m(t), \kappa(t)\big)$$

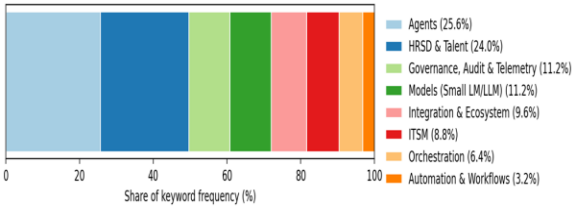


**Fig 4: State-Machine Model for Multi-Agent Orchestration in Intelligent ITSM Automation**

### 4.1. Incident Management and Response

Numerous internal APIs and cloud-based GenAI services are already available in many enterprises for handling recurring incidents. However, the following decision tree can be used when there is no specific implementation in place.
The initial automation boundary is set by whether the query is sufficiently detailed for first-level resolution; otherwise, the incident is routed to first-level support. When the query contains enough information for first-level resolution, the next step is to validate the resolution with an internal API or cloud-based GenAI service. Only if this validation fails does the incident get routed to first-level support. This approach helps identify incorrect resolutions and improves knowledge-base curation. The next question determines whether the incident falls into a common category (for instance, password resets or email phishing) and can therefore be addressed using trigger-word-driven templates.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 35 Issue 2, 2025, Pages: 1889-1897

Successful application of such templates completes the incident resolution.

As an additional validation step, events in this category can be reported/flagged for monitoring by a supervisory GenAI agent, informing it of the similarity in context and importance. If the agent determines that offending data is being used to initiate a fraudulent course of action, it can take the appropriate action (for example, disabling the sender's account) autonomously. If the issue cannot be resolved by internal API or cloud-based GenAI service, or if it requires vendor escalation, relations with the vendor are leveraged automatically to escalate the query intelligently and provide status updates to the reporting employee.

## 5. HRSD Automation with LLM Agents

HRSD Automation with LLM Agents

Enterprise Human Resource Service Delivery (HRSD) automations – being employee-centric in nature – must have an engagement lens throughout their design. While the true user experiences can only be realized through an actual deployment, it is feasible to assess the various user-facing aspects of the envisioned automations by modelling the use case workflows, designing the agentisation and enabling cross-agent collaboration. Two of the most complex and high-volume HR tasks – employee onboarding and offboarding – are considered for illustration.

Generative LLM agents present a potential to significantly automate and speed up the fulfilment of employee onboarding and offboarding requests in a highly effective manner. However, as with any automation powered by LLMs, fulfillment quality will directly depend on careful prompt engineering. Although the requests can be fulfilled primarily via LLM-based conversations, integration with other systems is necessary for retrieving user-specific details, fulfilling compliance requirements, and accessing organizational and role-specific knowledge stores. The management of the various integrations needs to be proper too. Since employee onboarding and offboarding involves data pertaining to the entire organization and user experience, data privacy and protection also need to be given paramount consideration, particularly while working with 3rd party cloud-based LLMs and other LLMs powering the automation.

### 5.1. Employee Onboarding and Offboarding

Streamlined employee onboarding and offboarding is crucial for talent acquisition and reducing resource drain. Secure interaction with enterprise knowledge bases is essential to prevent data exposure. In addition to these routines, other operational domains of Human Resources Services Management with the potential for higher value orchestration are employee and skill demand planning, user monitoring and signalling, work and role assignment, and Talent Knowledge Management and Task Sharing. Specialized agents are required to support complex automated scenarios involving these domains.

Routing requests to suitable agents requires coordination and communication mechanisms enabling agents to commit to Tasks in segments. These mechanisms must provide status and validation indicators that a higher-level orchestration or incident-management system can monitor. The automation of these processes cannot compromise compliance or user experience, including reliability, auditability, and fail-safes. Human-in-the-loop guidelines must also be defined for scenario areas that demand higher levels of intervention. Information security and data-privacy policies must be enforced to the necessary extent in integrated enterprise deployments.

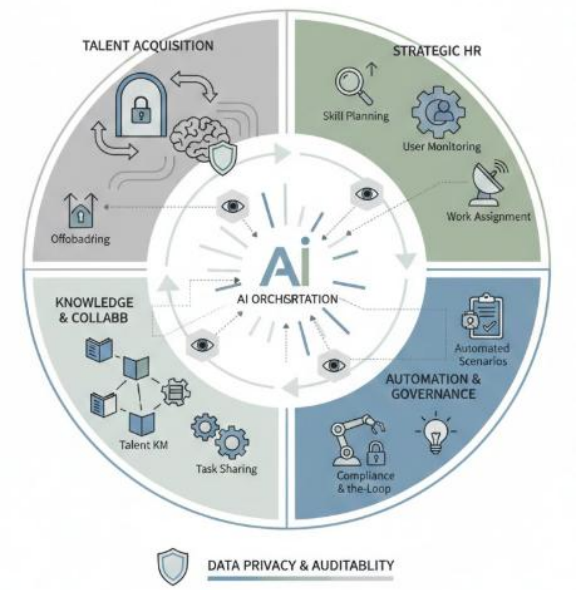


**Fig 5: Orchestrating Secure Talent Lifecycles: A Multi-Agent Framework for Intelligent Human Resources Services Management (HRSM)**

### 5.2. Talent Management and Resource Allocation

Intelligent Talent Management & Resource Allocation

Human resource service delivery can provide enterprises with an opportunity to create a digital workplace environment that fosters productivity through world-class services, increased employee engagement, and improved organizational efficiency. Automating entire workflows will help create a more personalized employee experience, often powered by conversational interfaces, where employees and managers can easily get answers to their HR questions and also carry out common HR actions in a quick and efficient manner.

Specifically, LLM agents can be employed as self-help and execution assistants for employees. They can help employees manage simple requests, such as country leave requests, change of bank details, benefits, advertisements on the intranet, user access requests, etc. Employee onboarding and offboarding workflows can be completely automated, and all employee travel-related requests and processing can be simplified via an LLM agent interfacing with various systems. Other areas of talent management and resource allocation, such as lateral movement, promotions, reassignment, internal job posting, and employee switching, can also be automated to some degree using LLM agents. The service can be enriched by integrating semantic text

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 35 Issue 2, 2025, Pages: 1889-1897

mining-based knowledge platforms for specific expertise in support of sensitive creation of Professional Development Records. User experience, data privacy, and governance considerations must be taken into account.

| Component | API-driven | Event-driven |
|---|---|---|
| Agent reasoning | 0.8 | 0.8 |
| Integration/API call | 1.2 | 0 |
| Queue/Bus overhead | 0.3 | 0.3 |
| Waiting for request | 0 | 0.1 |

**Table 1: Comparative Latency Characteristics of API-Based vs. Event-Driven Gen AI Integration**

## 6. Orchestration Techniques and Workflows

Agents can collaborate and coordinate with one another to deliver more complex responses to specific intents via various orchestration techniques that manage multi-agent interactions across independent locations. Techniques govern how agents communicate and synchronize at a high level, while the planner, reasoning, and decision-support speeches primarily enable internal agent operations.

Multi-agent orchestration is organized as a state machine controlled by LLM prompts, with the active state specifying the current operation and topic context. Each subroutine of the state machine uses a pointer that activates an LLM-syntactically formatted prompt router. The router identifies the correct subroutine based on keyword matching. The specific operation involves recognizing and executing the required prompt-orchestration task.

Because many processes of interest exhibit various non-linear operation modes at different times, LLM agent orchestration techniques support execution in scenarios that resemble a state machine with primary and secondary states. Primary states express the scenario's triggering conditions (e.g., onboarding a new joiner), while secondary states describe the secondary paths related to other agents, globally shared conditions, or resources not strictly bound to the main path. The overlay to the core prompt-control concept captures these states, allowing instantiation on the fly for selected process instances and contributing to monitoring and alert notifications. The latter monitors condition variables and alerts administrators to situations that require human intervention.

Orchestration of procedures that process a limited number of trigger conditions requires dedicated LLM-control prompts. Such procedures do not respond correctly to high-level navigation state variables but rather require specific paths across facilities owned by different agents or specialized components not entirely governed by LLM prompts. All such process-control patterns leverage the distribution of instantiation code, skill repositories, and shared high-level resources.

**Equation 3) Latency reduction: event-driven vs API-driven integration**

**Step 1 — Define latency components**

- $L_{\text{agent}}$: agent reasoning + negotiation time
- $L_{\text{api}}$: synchronous API/service call time
- $L_{\text{bus}}$: message bus overhead
- $L_{\text{wait}}$: waiting time until event arrives (often near 0 in push systems)

**Step 2 — API-driven latency**

$$L_{\text{API}} = L_{\text{agent}} + L_{\text{api}} + L_{\text{bus}}$$

**Step 3 — Event-driven latency**

$$L_{\text{EVT}} = L_{\text{wait}} + L_{\text{agent}} + L_{\text{bus}}$$

**Step 4 — Speedup factor**
If pushed events make $L_{\text{wait}} \approx 0$:

$$\text{Speedup} = \frac{L_{\text{API}}}{L_{\text{EVT}}} \approx \frac{L_{\text{agent}} + L_{\text{api}} + L_{\text{bus}}}{L_{\text{agent}} + L_{\text{bus}}}$$

### 6.1. Multi-Agent Collaboration and Coordination

Multi-agent collaboration and coordination techniques facilitate non-trivial task orchestration through extensible workflows comprising elementary tasks. Multiple Gen AI agents collaborate to accomplish complex tasks through logical decomposition of goals. Such collaboration is crucial in large-scale scenarios, where a number of business intents are fulfilled by an eco-system of Gen AI agents from multiple organization units or enterprises. An agent is responsible for a specific role, function, or specialized domain within the orchestration. Command control mechanisms are useful to centralize decision-making for controlling LLM agents and when the consequences of an agent's decisions are critical to business and society. Agents can also negotiate with one another to de-conflict on resource utilization or power during multi-agent collaboration protocols.

A scenario-based design approach is proposed for orchestrating such complex multi-agent collaborative workflows. These workflows are represented as orchestration graphs, where the nodes are elementary tasks and the edges represent the trigger conditions for firing a specific elementary task. The execution state of the overall orchestration process is externalized and exposed to the individual workflows, either through a single data store or a multitude of state stores, to support dynamic creation of such processes. A generic token-based state machine is defined for representing the execution states in a structured manner, facilitating State machines for multi-agent workflows.

## 7. Conclusion

Research findings indicate that enterprise-scale generative artificial intelligence (Gen AI) orchestration can be achieved

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 35 Issue 2, 2025, Pages: 1889-1897

by harnessing small language models (LMs) for intelligent automation of information technology service management (ITSM) and human resources service delivery (HRSD) within an enterprise ecosystem. Such orchestration integrates information and operational technology functions, minimizing or eliminating human intervention and maximizing transparency and security. Small LMs support automation of a wide range of incidents across business functions and offer a compelling value proposition relative to other candidates. Large language model agents acting as orchestration managers safely automate iterative multi-agent scenarios with interdependencies, including onboarding and offboarding of employees and talent management.

Orchestration boundaries are stabilized using a three-partite design framework—telemetry, auditing, and risk management—aligned with the economics and technical maturity of available resources. Multi-agent collaboration and coordination techniques applied to natural-science domains provide guidance for successful implementation in social-science domains. With the expansion of Gen AI services across business functions, enterprises are rapidly establishing Gen AI control towers to monitor and govern external risks, with special emphasis on data security and privacy. The proposed small-LM and LLM-agent solutions contribute to meeting these control objectives.


**Fig 6: SLM & LLM-Agent Functional Maturity**

### 7.1. Future Directions and Implications of Generative AI in Orchestration

Considerable attention must be devoted to introducing adequate safeguards for the autonomous deployment of Generative AI technologies as part of Organizational Orchestration. The development of suitable Governance Models capable of addressing all aspects of Governance in the Orchestration context and related policy and regulatory definitions and controls supported by organizations in the private and public sectors will guide respective practical developments. The definition of open standards and common frameworks – both technical and organizational – covering service provision technologies, such as APIs, Data Interoperability Layers, Message Bus Systems, and the business capabilities delivered through Workflow Engines is also fundamental for successful Research and Innovation efforts.

Many opportunities presently being examined and experimented with in the generation of content, images, code, and 3D models, for instance, will have a major impact on the organizational ecosystem in the Automation context. Organizations whose size and sensitivity to Data Controllers Governance Assistances afford deploying LLM technologies at a reduced operational, safety, sensitivity concerns, and Data Privacy Governance risk will benefit greatly from employing these services in Workflow Automation Areas. Considerably reduced Latency – an important factor for Customer Experience – and Governance processes built on Enterprise Business Procedures and Policies constitute attractive reasons to favor in-house Data Excel for Data Generation Processes where Data Control is of utmost importance.

## 8. References

[1] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., … OpenAI. (2023). *GPT-4 technical report*. arXiv.

[2] Garapati, R. S. (2025). Real-Time Monitoring and AI-Based Control of Industrial Robots Using Cloud-Hosted Web Applications. Available at SSRN 5612491.

[3] Kwon, H., & Park, J. (2021). API gateway patterns for enterprise microservices integration: A systematic mapping study. Journal of Systems Architecture, 117, 102103.

[4] Ahmad, W., Chakraborty, S., Ray, B., & Chang, K. W. (2021). Unified pre-training for program understanding and generation. In *Proceedings of NAACL-HLT 2021* (pp. 2655–2668). Association for Computational Linguistics.

[5] Kummari, D. N., Challa, S. R., Pamisetty, V., Motamary, S., & Meda, R. (2025). Unifying Temporal Reasoning and Agentic Machine Learning: A Framework for Proactive Fault Detection in Dynamic, Data-Intensive Environments. Metallurgical and Materials Engineering, 31(4), 552-568.

[6] Ajzen, I. (1991). The theory of planned behavior. Organizational Behavior and Human Decision Processes, 50(2), 179–211.

[7] Aitha, A. R., & Jyothi Babu, D. A. (2025). Agentic AI-Powered Claims Intelligence: A Deep Learning Framework for Automating Workers Compensation Claim Processing Using Generative AI. Available at SSRN 5505223.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 35 Issue 2, 2025, Pages: 1889-1897

[8] Allen, F., Carletti, E., & Marquez, R. (2023). Financial system resilience, regulation, and digital transformation. *Journal of Financial Stability, 66, 101122.

[9] Anagnostopoulos, I. (2022). Artificial intelligence in financial services: A critical review of applications and challenges. *Journal of Financial Regulation and Compliance, 30*(2), 195–210.

[10] Arora, S., Ge, R., Liang, Y., Ma, T., & Zhang, Y. (2018). A theoretical analysis of contrastive unsupervised representation learning. arXiv.

[11] Sheelam, G. K., & Komaragiri, V. B. (2025). Self-Adaptive Wireless Communication: Leveraging ML And Agentic AI In Smart Telecommunication Networks. Metallurgical and Materials Engineering, 1381-1401.

[12] Kratzke, N., & Quint, P. (2017). Understanding cloud-native applications after 10 years of cloud computing. Journal of Systems and Software, 126, 1–16.

[13] Bansal, G., & Weld, D. S. (2018). A coverage-based utility model for identifying unknown unknowns. In *Proceedings of AAAI 2018* (pp. 706–713). AAAI Press.

[14] Nagabhyru, K. C. (2025). Beyond Automation: The 2025 Role of Agentic AI in Autonomous Data Engineering and Adaptive Enterprise Systems.

[15] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of FAccT 2021* (pp. 610–623). ACM.

[16] Meda, R. (2025). AI-Driven Demand and Supply Forecasting Models for Enhanced Sales Performance Management: A Case Study of a Four-Zone Structure in the United States. Metallurgical and Materials Engineering, 1480-1500.

[17] Bholat, D., Gharbawi, M., & Thew, O. (2023). Machine learning, big data, and financial stability. Financial Stability Review, 27, 33–49.

[18] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., … Liang, P. (2021). On the opportunities and risks of foundation models. arXiv.

[19] Vajpayee, A., Khan, S., Gottimukkala, V. R. R., Sharma, D., & Seshasai, S. J. (2025). Digital Financial Literacy 4.0: Consumer Readiness for AI-Driven Fintech and Blockchain Ecosystems. International Insurance Law Review, 33(S5), 963-973.

[20] Camarinha-Matos, L. M., & Afsarmanesh, H. (2008). Collaborative networks: Reference modeling. Springer Series in Computer Science. Springer.

[21] Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J., Dhariwal, P., … Zaremba, W. (2021). Evaluating large language models trained on code. arXiv.

[22] Inala, R. (2025). A Unified Framework for Agentic AI and Data Products: Enhancing Cloud, Big Data, and Machine Learning in Supply Chain, Insurance, Retail, and Manufacturing. EKSPLORIUM-BULETIN PUSAT TEKNOLOGI BAHAN GALIAN NUKLIR, 46(1), 1614-1628.

[23] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems, 30*, 4299–4307.

[24] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186). ACL.

[25] Segireddy, A. R. (2025). GENERATIVE AI FOR SECURE RELEASE ENGINEERING IN GLOBAL PAYMENT NETWORK. Lex Localis: Journal of Local Self-Government, 23.

[26] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv.

[27] Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9*(3–4), 211–407.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 35 Issue 2, 2025, Pages: 1889-1897

[28] Amistapuram, K. (2025). Agentic AI for Next-Generation Insurance Platforms: Autonomous Decision-Making in Claims and Policy Servicing. Journal of Marketing & Social Research, 2, 88-103.

[29] Fagan, M. H., Neill, S., & Wooldridge, B. (2008). Exploring the intention to use computers. Journal of Computer Information Systems, 49(1), 94–102.

[30] Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review, 1*(1).

[31] Varri, D. B. S. V. (2025). Human-AI collaboration in healthcare security.

[32] Kelley, P. G., Cranor, L. F., & Sadeh, N. (2013). Privacy as part of the app decision-making process. In *Proceedings of CHI 2013 (pp. 3393–3402). ACM.

[33] Nagubandi, A. R. (2024). Breakthrough Real-Time AI-Driven Regulatory Intelligence for Multi-Counterparty Derivatives and Collateral Platforms: Autonomous Compliance for IFRS, EMIR, NAIC, SOX & Emerging Regulations. Journal of Information Systems Engineering and Management, 9.

[34] Gentry, C. (2009). Fully homomorphic encryption using ideal lattices. In Proceedings of STOC 2009* (pp. 169–178). ACM.

[35] Yandamuri, U. S. AI-Driven Decision Support Systems for Operational Optimization in Hospitality Technology.

[36] Hasselt, H. V., Guez, A., & Silver, D. (2016). Deep reinforcement learning with double Q-learning. AAAI Conference on Artificial Intelligence, 2094–2100.

[37] Guntupalli, R. (2025). Federated Deep Learning for Predictive Healthcare: A Privacy-Preserving AI Framework on Cloud-Native Infrastructure. Vascular and Endovascular Review, 8(16s), 200-210.

[38] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. In *Proceedings of ICLR 2021*.

[39] Rongali, S. K. (2025, August). Deep Learning for Cybersecurity in Healthcare: A Mulesoft-Enabled Approach. In 2025 International Conference on Artificial Intelligence and Machine Vision (AIMV) (pp. 1-6). IEEE.

[40] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33, 6840–6851.

[41] Jiang, Y., Zhang, S., Ma, X., & Chen, Y. (2023). A survey on observability for large language model systems. arXiv.