Journal of Advanced Robotics, Autonomous Systems and HumanMachine Interaction ISSN:3070-6491

PRIME OPEN ACCESS

**Volume 2, Issue 2**

**Research Article**

**Date of Submission:** 08 May, 2026
**Date of Acceptance:** 08 Jun, 2026
**Date of Publication:** 18 Jun, 2026

# MCP + SKILL: A Deterministic Architecture for Enterprise AI Automation with Selective Intelligence Application

**Debendra Ray***

Independent Researcher, Enterprise AI Architecture Kolkata, India

***Corresponding Author:** Debendra Ray, Independent Researcher, Enterprise AI Architecture Kolkata, India.

**Citation:** Ray, D. (2026). MCP + SKILL: A Deterministic Architecture for Enterprise AI Automation with Selective Intelligence Application. *J Adv Robot Auton Syst Hum Mach Interact, 2*(2), 01-05.

## Abstract

The proliferation of Large Language Model (LLM)-based autonomous agents in enterprise environments has exposed fundamental architectural limitations including unpredictable execution paths, compounding computational costs, non-deterministic outputs, and complex audit trails that impede regulatory compliance. This paper introduces the MCP + SKILL (Model Context Protocol + Structured Knowledge & Instruction Layer for LLM) architecture, a novel framework that addresses these limitations through deterministic workflow execution with selective intelligence application. The architecture comprises four layers: (1) a lightweight orchestration layer utilizing directed acyclic graphs for state management, (2) SKILL files providing human-readable, version-controlled workflow specifications, (3) Model Context Protocol servers offering standardized tool interfaces, and (4) external system integrations. The key contribution lies in the formalization of "decision points"—explicitly defined workflow stages where LLM reasoning provides genuine value—and a conditional routing mechanism that constrains LLM invocation to these points. Empirical evaluation through a B2B customer onboarding case study demonstrates a 97.9% reduction in LLM API costs, 85.3% reduction in execution latency, and 100% output determinism for identical inputs compared to autonomous agent baselines. The framework achieves enterprise-grade auditability while preserving intelligent capability for genuinely ambiguous scenarios. We position MCP + SKILL as complementary to emerging AI governance frameworks, addressing the execution layer while governance frameworks address authorization.

**Keywords:** Deterministic AI Workflows, Model Context Protocol, Enterprise AI Architecture, LLM Cost Optimization, Agentic AI Systems, Workflow Orchestration, AI Auditability, Selective Intelligence

## Introduction

The emergence of Large Language Models (LLMs) with sophisticated reasoning, planning, and tool-use capabilities has fundamentally transformed perspectives on enterprise automation [1-3]. These foundation models demonstrate remarkable ability to understand natural language instructions, decompose complex tasks, and interact with external systems through function calling mechanisms. This capability convergence has catalyzed significant interest in autonomous AI agents—systems that can pursue complex goals with limited direct supervision. However, a persistent gap has emerged between demonstration capabilities and production requirements. While autonomous agents excel in controlled environments with well-defined success criteria, they frequently struggle when deployed against the complexity, variability, and compliance demands of real-world enterprise systems [4-9].

## Problem Statement

Enterprise environments impose constraints that conflict with the core design assumptions of autonomous agents:

**Predictability Requirements:** Financial services, healthcare, and regulated industries mandate consistent outputs from identical inputs. The stochastic nature of LLM reasoning creates compliance and quality assurance challenges [10,11].

**Cost Sensitivity at Scale:** Autonomous agents typically invoke LLM reasoning at multiple stages, resulting in 8-15 LLM calls per task. At enterprise scale—thousands to millions of daily transactions—these costs become prohibitive [12].

**Auditability Demands:** Regulatory frameworks including SOX, GDPR, HIPAA, and Basel III require clear decision trails. The exploratory nature of autonomous agent reasoning produces audit logs difficult to interpret and verify [13].

**Failure Mode Complexity:** Autonomous agents exhibit "graceful degradation pathology"—failing in ways that mask the failure, producing plausible but incorrect outputs [14,15]

## Research Questions

This research addresses three primary questions:
**RQ1:** How can enterprise AI automation architectures achieve deterministic execution while preserving intelligent capability for genuinely ambiguous scenarios?
**RQ2:** What mechanisms enable selective application of LLM reasoning to minimize costs while maintaining decision quality for edge cases?
**RQ3:** How can workflow specifications satisfy both technical implementations need and compliance audit requirements?

## Research Contributions

This paper makes five contributions:
- **Architectural Framework:** A four-layer architecture separating orchestration, workflow definition, tool interfaces, and external systems.
- **Formal Decision Point Model:** Formalization of "decision points" and conditional routing mechanisms constraining LLM invocation.
- **SKILL Specification Language:** Human-readable workflow specifications serving both implementation and compliance needs.
- **Empirical Validation:** Comprehensive evaluation demonstrating 97.9% cost reduction and 100% determinism.
- **Governance Integration:** Positioning within emerging AI governance frameworks including Nomotic AI.

## Literature Review

### Autonomous AI Agents

The ReAct framework established the foundational paradigm for LLM-based agents, introducing the reasoning-action loop where models interleave chain-of-thought reasoning with tool invocation. Subsequent work extended this: Toolformer demonstrated self-taught tool use; HuggingGPT coordinated multiple specialized models; Generative Agents explored emergent social behaviours. The open-source community rapidly operationalized these concepts through AutoGPT , BabyAGI, LangChain, and CrewAI. Despite progress, autonomous agents exhibit systematic limitations impeding enterprise deployment: non-determinism, cost compounding, brittleness to input variation, failure opacity, audit trail complexity , and security vulnerabilities [4,9,11-15-21].

### Workflow Orchestration Systems

Traditional orchestration systems—ApacheAirflow,Temporal,Prefect—provide deterministic execution but lack capability for contextual decisions. LangGraph models' workflows as state machines with optional LLM nodes but provides no principled decision point identification [22-24].

### AI Governance Frameworks

Hood (2025) introduces "Nomotic AI" as a governance counterpart to agentic AI, focusing on pre-action authorization and authority boundaries. OpenAI's practices paper outlines safety considerations. The EU AI Act establishes regulatory requirements. These frameworks address authorization (what AI should do) but provide limited guidance on execution (how to do it reliably). Our work addresses this execution gap [6,25].

## Research Methodology

This research employs Design Science Research (DSR) methodology, appropriate for developing IT artifacts solving organizational problems. Following Peffers et al.'s process model, our research progresses through six activities:
- Problem Identification,
- Objectives Definition,
- Design & Development,
- Demonstration,
- Evaluation, and
- Communication [26,27].

## Case Study Design

We employ an embedded single-case design [28]. The case—B2B customer onboarding with credit pre-qualification—was selected for representativeness, complexity (routine + edge cases), compliance relevance, and measurability. The autonomous agent baseline uses LangChain AgentExecutor with Claude 3.5 Sonnet, standard ReAct prompting, and maximum 15 iterations.

## The Mcp + Skill Architecture

### Design Principles

Five foundational principles guide the architecture:

- **Determinism by Default:** Identical outputs for identical inputs unless explicitly designated as decision points.
- **Surgical Intelligence Application:** LLM reasoning applied only where it provides unique value.
- **Separation of Concerns:** Clean separation of orchestration, specification, interfaces, and systems.
- **Human-Readable Specifications:** Workflow definitions readable by non-technical stakeholders.
- **Complete Audit Trails:** Every execution produces deterministic, reproducible audit documentation.

### Four-Layer Architecture

The architecture comprises four layers:

- **Layer 1 (Orchestration):** Directed acyclic graph execution with immutable state management using LangGraph StateGraph.
- **Layer 2 (SKILL Files):** Human-readable workflow specifications in Markdown with YAML frontmatter.
- **Layer 3 (MCP Servers):** Standardized tool interfaces following Model Context Protocol.
- **Layer 4 (External Systems):** Enterprise applications, APIs, and notification channels.

### Decision Point Formalization

**Definition 1 (Decision Point)**

A workflow step where:

- correct action cannot be determined by codifiable rules,
- contextual judgment adds value, and
- LLM cost is justified by decision complexity.

**Definition 2 (Edge Case Trigger):** A predicate function $t: S \rightarrow \{true, false\}$ evaluating workflow state to determine decision point invocation. The routing function returns "decision point" if any trigger evaluates true, otherwise "proceed".

## Evaluation Results

### Test Scenarios

**Three test scenarios were evaluated:**

**Scenario A:** Established company (15 years), $150K limit—standard flow, no LLM. **Scenario B:** New company (1 year), $500K limit—edge case triggers LLM.

Invalid input—validation failure, no LLM.

### Results Summary

| Metric | MCP + SKILL | Autonomous Agent | Improvement |
|---|---|---|---|
| LLM Calls per Request | 0.2 | 9.8 | 97.9% |
| Cost per 10K Requests | $4.80 | $225.00 | 97.9% |
| Latency (seconds) | 2.67 | 18.15 | 85.3% |
| Output Determinism | 100% | 71-82% | 100% |
| Audit Trail Length | 8 entries | 24 entries | 67% shorter |

**Table 1: Comparison of MCP + SKILL vs Autonomous Agent Baselines**

### Statistical Significance

Two-sample t-tests confirmed significance at $\alpha = 0.05$. Cost reduction: $t = -42.31$, $p < 0.0001$. Latency improvement: $t = -38.76$, $p < 0.0001$. Results validate the core thesis: most enterprise workflow steps do not require AI reasoning.

### Discussion

The evaluation results validate the core thesis of MCP + SKILL: most enterprise workflow steps do not require AI reasoning, and constraining LLM invocation to explicitly defined decision points dramatically improve cost efficiency, latency, and determinism without sacrificing capability for genuinely complex scenarios.

### Governance Integration

The architecture complements authorization-focused governance frameworks. Nomotic AI addresses what AI should be permitted to do. MCP + SKILL addresses how permitted actions should execute reliably. This separation enables independent evolution while maintaining clear responsibility boundaries [29].

### Limitations

The current architecture has limitations:

- Manual decision point identification requires domain expertise.
- Single-case evaluation limits generalizability claims.
- Edge case distribution assumptions (20%) affect cost projections.
- Results depend on specific LLM model (Claude Sonnet).

## Conclusion

This paper introduced MCP + SKILL, an architecture for enterprise AI automation that achieves deterministic workflow execution with selective intelligence application. The key insight is that most workflow steps can execute deterministically, while a minority genuinely benefit from contextual judgment. By formalizing decision points and implementing conditional routing, the architecture achieves 97.9% cost reduction, 85.3% latency improvement, and 100% determinism compared to autonomous agent baselines. The framework positions as complementary to AI governance approaches—governance addresses authorization while MCP + SKILL addresses execution reliability. For organizations seeking production-ready AI automation balancing intelligence with predictability, cost efficiency, and auditability, MCP + SKILL offers a validated path forward.

**Data Availability:** Implementation code and evaluation data are available

## References


1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.
2. OpenAI (2023) 'GPT-4 technical report',
3. Anthropic, A. I. (2024). Claude 3 model card. Technical documentation, Anthropic.
4. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., ... & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. Advances in neural information processing systems, 36, 68539-68551.,
5. Patil, S.G., et al. (2023) 'Gorilla: Large language model connected with massive APIs',
6. Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., Campbell, R., ... & Robinson, D. G. (2023). Practices for governing agentic AI systems. Research Paper, OpenAI.
7. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., ... & Wen, J. (2024). A survey on large language model based autonomous agents. Frontiers of Computer Science, 18(6), 186345.
8. Xi, Z., et al. (2023) 'The rise and potential of large language model-based agents: A survey',
9. Wang, Z., Mao, S., Wu, W., Ge, T., Wei, F., & Ji, H. (2024, June). Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 257-279).
10. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35, 27730-27744.
11. WX, Z. (2023). A survey of large language models.
12. Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). Challenges and applications of large language models. arXiv preprint arXiv:2307.10169.
13. Chan, A., Ezell, C., Kaufmann, M., Wei, K., Hammond, L., Bradley, H., ... & Anderljung, M. (2024, June). Visibility into AI agents. In Proceedings of the 2024 ACM conference on fairness, accountability, and transparency (pp. 958-973).
14. Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023, October). Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th annual acm symposium on user interface software and technology (pp. 1-22).
15. Liu, N.F., et al. (2023a) 'Lost in the middle: How language models use long contexts',
16. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). React: Synergizing reasoning and acting in language models. arXiv preprint arXiv:2210.03629.
17. Shen, Y., Song, K., Tan, X., Li, D., Lu, W., & Zhuang, Y. (2023). Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. Advances in Neural Information Processing Systems, 36, 38154-38180.
18. Significant Gravitas (2023) AutoGPT.
19. Chase, H. (2022) LangChain.
20. Moura, J. (2024) CrewAI.
21. Nakajima, Y. (2023) BabyAGI.
22. Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023, November). Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In Proceedings of the 16th ACM workshop on artificial intelligence and security (pp. 79-90).
23. Apache Software Foundation (2015) Apache Airflow.
24. Temporal Technologies (2020) Temporal.
25. Prefect Technologies (2018) Prefect.
26. Van der Aalst, W. M. (2013). Business process management: a comprehensive survey. International Scholarly Research Notices, 2013(1), 507984.

27. LangChain (2024) LangGraph.
28. European Commission (2024) 'Regulation (EU) 2024/1689 (AI Act)', Official Journal of the European Union.
29. Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research1. MIS quarterly, 28(1), 75-106.
30. Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. Journal of management information systems, 24(3), 45-77.
31. Yin, R. K. (2018). Case study research and applications (Vol. 6). Thousand Oaks, CA: Sage.
32. Anthropic (2024b) 'Model Context Protocol specification'.
33. Hood, C. (2025) 'Nomotic AI: The governance counterpart to agentic AI', PhilArchive.