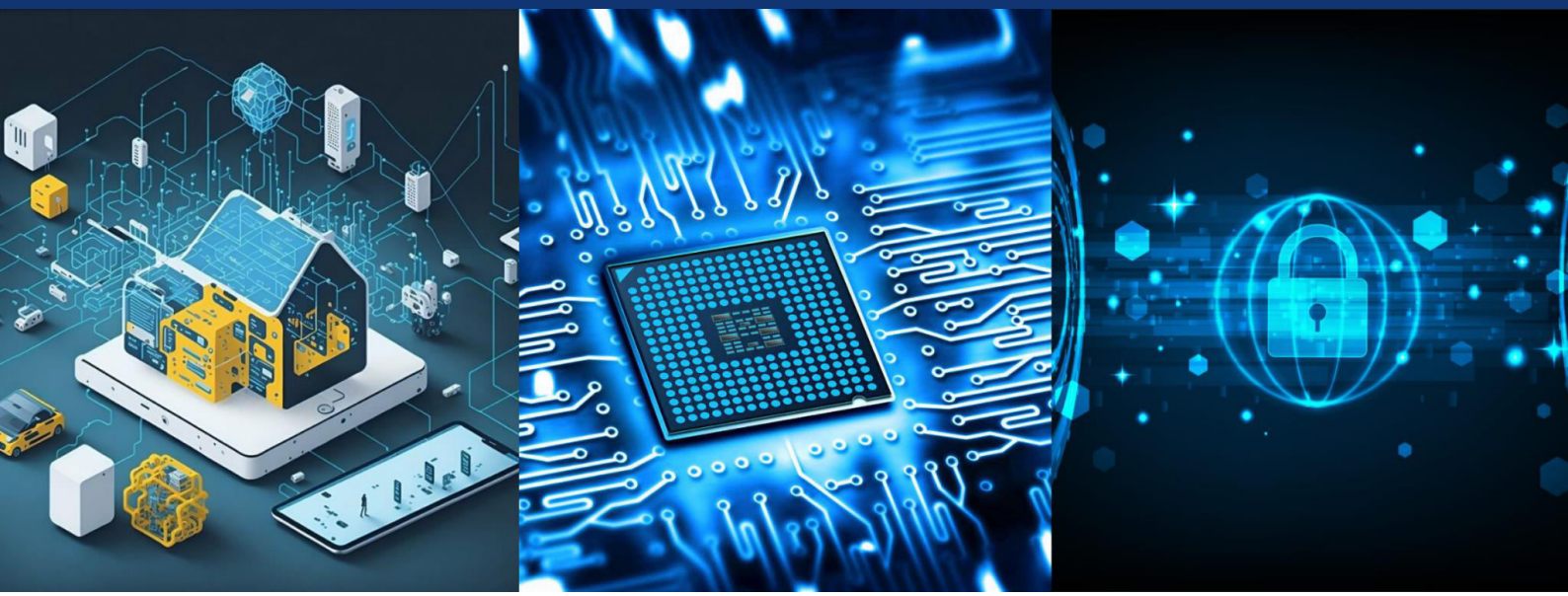
IJIRCCE

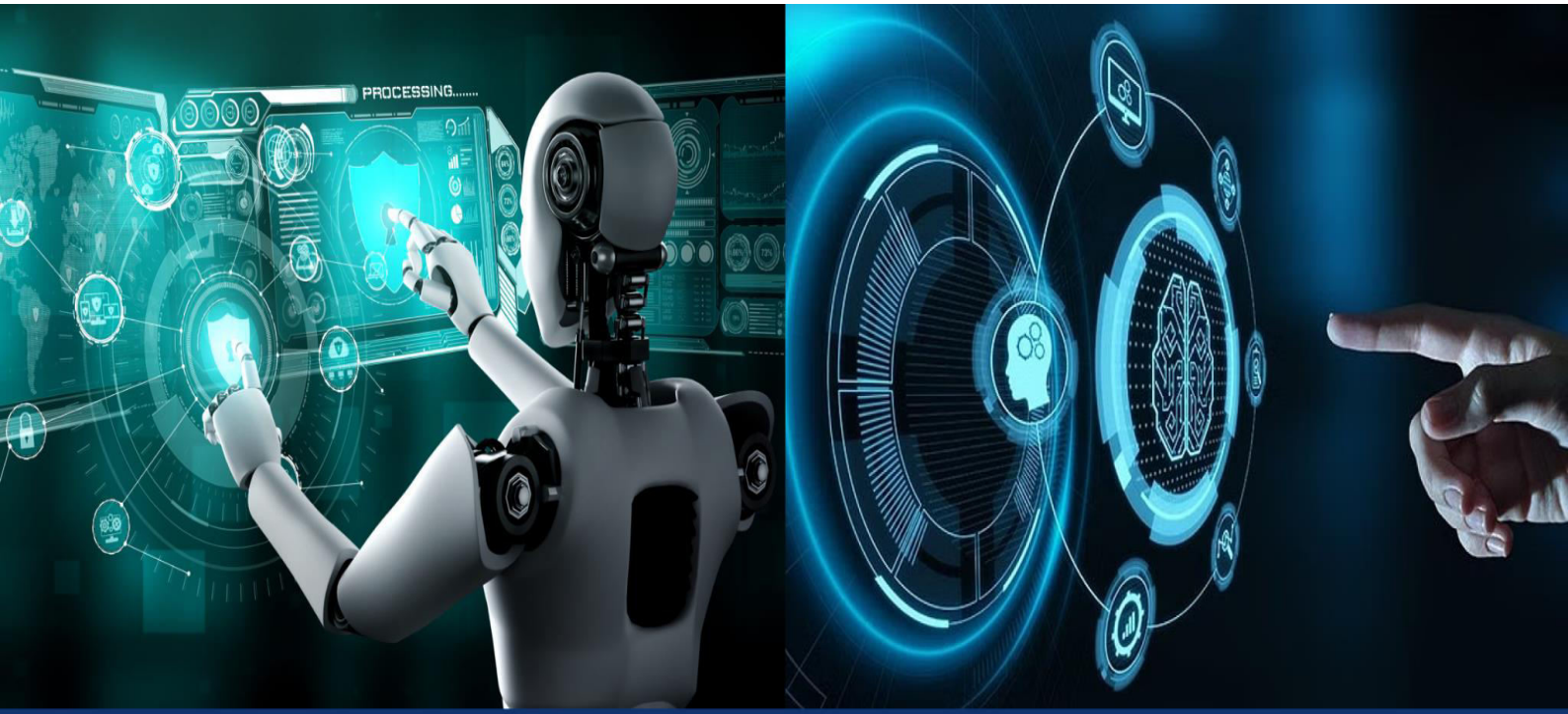
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.927

Volume 14, Issue 10, October 2026

© 2026 IJIRCCE | Volume 14, Issue 10, October 2026 DOI: 10.15680/IJIRCCE.2026.1410035

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

A Survey of Agentic AI and Multi-Agent LLM Systems for Intelligent Business Process Automation

Rakshe Rutuja Navnath[1], Ranpise Vidya Vikas[1], Sandbhor Saloni Sambhaji[1], Prof. Jagdale S.V. [2]

UG Student, Dept. of Computer Engineering, SREIR's Samarth College of Engineering and Management, Belhe, Pune, Maharashtra, India[1]

Assistant Professor, Dept. of Computer Engineering, SREIR's Samarth College of Engineering and Management, Belhe, Pune, Maharashtra, India[2]

ABSTRACT: Business process automation in small and mid-size firms still depends on Robotic Process Automation (RPA) bots and hand-written workflow rules, which stop working when an input format, a web page or a business rule changes. Large Language Model (LLM) agents that plan, call tools and work in teams offer another route. This paper surveys 43 works published between 2023 and 2026 on multi-agent LLM frameworks, reasoning and memory, tool use and agent protocols, LLM-driven workflow automation, agents for supply chain and retail operations, evaluation benchmarks, and governance and security. The works are arranged in a seven-part taxonomy and compared on coordination structure, tool access, human oversight, domain and reported results. Collaboration frameworks are well developed, but results on enterprise-style benchmarks are still modest: the best agent completes about 30% of tasks in a simulated software company, and a strong function-calling agent stays correct across eight repeated trials on fewer than 25% of retail customer-service tasks. Few systems combine multi-agent coordination, policy-based approval and an audit trail in one design, and almost none target small retailers whose staff work from mobile phones. Based on these findings, the paper outlines a reference architecture for retail process automation with a central orchestrator, four domain agents and a governance layer, and lists the open problems that remain.

KEYWORDS: Agentic AI; multi-agent systems; large language models; business process automation; tool use; retrieval-augmented generation; human-in-the-loop; retail automation

I. INTRODUCTION

A retail store runs on connected business processes: stock is counted and refilled, purchase orders go to suppliers, orders are packed and handed to couriers, returns are checked against a policy, and the owner reviews daily sales. In most small and mid-size stores in India these steps are still done with spreadsheets, phone calls and messaging apps. Larger firms use RPA, where a bot replays recorded clicks and keystrokes, but a bot fails as soon as a supplier portal changes its layout or a customer writes a request in free text. Vidgof et al. [1] argued early on that LLMs could fill this gap in business process management because they handle unstructured text and can explain their output.

Since then, research on LLM agents has moved quickly. ReAct [2] interleaves reasoning steps with tool calls, and Toolformer [3] learns when to call an API. Surveys by Wang et al. [4] and Xi et al. [5] describe an agent as a language model core with perception, memory and action modules, and frameworks such as AutoGen [6] and MetaGPT [7] place several such agents in one system, each with a role. The term agentic AI now refers to systems in which several agents split a goal, coordinate and act with limited supervision [8].

Existing surveys cover parts of this field: Guo et al. [9] and Tran et al. [10] classify multi-agent collaboration, Raza et al. [11] review trust and security, and Calvanese et al. [12] set an agenda for agentic business process management. None follows the full path a practitioner needs, from the collaboration framework and tools, to benchmarks that show reliability, to the controls that keep a human in charge of money-related actions. This survey takes that path, with retail operations as the running example.

The paper makes four contributions: a seven-part taxonomy of 43 works from 2023 to 2026, a comparison of the main systems and benchmarks with their reported results, a list of gaps that matter for small retailers, and a reference architecture derived from the review, which is the design of the authors' ongoing project.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

II. BACKGROUND

A. *From RPA to Agentic Process Automation*

Traditional Business Process Automation (BPA) encodes a process as a fixed model or script. Ye et al. [13] introduced Agentic Process Automation (APA) in ProAgent, where an LLM agent builds the workflow from an instruction and a DataAgent and a ControlAgent handle steps that need data understanding or a run-time decision. Vu et al. [14] interviewed 22 practitioners, whose main worries were governance, accountability and unclear human roles. Dumas et al. [15] proposed agentic BPM systems in which agents watch process states, use process mining data and act, across a range from human-driven to fully autonomous processes. Calvanese et al. [12] named four capabilities such agents need: framed autonomy, explainability, conversational actionability and self-modification, and Azarijafari et al. [16] suggested organising processes around goals and business objects instead of fixed task sequences.

B. *Building Blocks of an LLM Agent*

Across the surveyed systems an agent has five parts: perception of inputs such as chat messages or sales tables; reasoning and planning, usually with ReAct [2] or Reflexion [17]; memory of the current task and longer-term facts [18], [19]; action through tool calls to databases and APIs [3], [20]; and feedback that checks each result. Retrieval-Augmented Generation (RAG) [21] adds relevant documents, such as a return policy, to the prompt so that answers stay tied to company data.

III. REVIEW METHOD

Papers were collected from arXiv, OpenReview (ICLR, NeurIPS), the ACL Anthology, ACM Digital Library, IEEE Xplore and Springer for January 2022 to March 2026, using terms such as "LLM multi-agent", "agentic process automation", "LLM agent benchmark" and "LLM supply chain agent". A work was included if it used an LLM as the reasoning core, reported a method or benchmark with results or gave a clear taxonomy, and related to at least one step of business process automation. The final set has 43 works, 14 surveys or position papers and 29 primary studies; 12 are from 2023, 14 from 2024, 14 from 2025 and 3 from 2026.

IV. TAXONOMY AND REVIEW OF EXISTING WORK

Fig. 1 shows the taxonomy used in this survey. The seven groups follow the layers of a working system, from how agents cooperate down to how their actions are controlled.

A. *Multi-Agent Collaboration Frameworks*

CAMEL [22] was one of the first role-playing frameworks, in which an AI user agent and an AI assistant agent cooperate under inception prompting; its authors also report role flipping and endless message loops. AutoGen [6] generalised this into conversable agents backed by an LLM, a tool or a human, and its human proxy agent is a simple way to insert approval into a conversation. MetaGPT [7] encodes Standard Operating Procedures (SOPs) into prompts so that role agents exchange structured documents instead of free chat, which reduced the spread of errors between roles. ChatDev [23] splits software development into phases with a chat chain of two-role dialogues, AgentVerse [24] recruits and adjusts a group of expert agents, and Du et al. [25] showed that debate between model instances improves factual accuracy and reasoning.

Magentic-One [26] is closer to a business setting: a lead Orchestrator keeps a task ledger and a progress ledger, assigns work to web, file, coder and terminal agents, and re-plans when progress stalls. Guo et al. [9] classify such systems by agent profiling, communication structure (layered, centralised, decentralised or shared message pool) and skill acquisition, and Tran et al. [10] add collaboration type and coordination protocol. Among systems that target real tasks, a centralised orchestrator with shared state is the common choice because it is easier to trace and stop.

B. *Reasoning, Planning and Memory*

ReAct [2] alternates Thought, Action and Observation, so the model sees each tool result before choosing the next step. Reflexion [17] adds verbal self-reflection after a failed attempt and stores the reflection for the next try. Generative Agents [18] keep a memory stream of observations and retrieve items by recency, importance and relevance, then reflect and plan. MemGPT [19] treats the context window like main memory and pages older facts in and out of external storage.

© 2026 IJIRCCE | Volume 14, Issue 10, October 2026 DOI: 10.15680/IJIRCCE.2026.1410035

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

In a store, a reorder task benefits from remembering last month's supplier delays, and a customer agent needs the history of an order across chat turns.

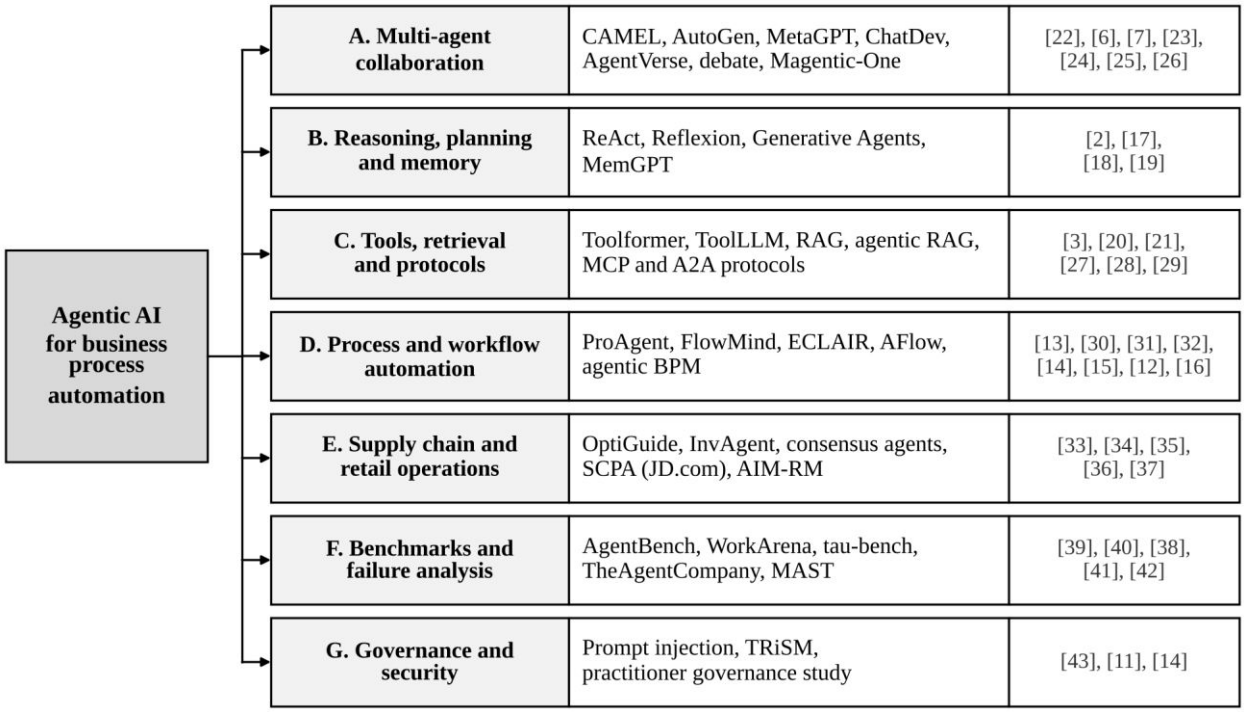


Fig. 1. Taxonomy of the surveyed literature on agentic AI for business process automation

C. *Tool Use, Retrieval and Agent Protocols*

Toolformer [3] trained a model to insert API calls into its own text, and ToolLLM [20] built ToolBench from more than 16,000 real REST APIs. Gao et al. [21] review RAG methods, and Singh et al. [27] describe agentic RAG, where agents decide what to retrieve, check it and retrieve again. Ehtesham et al. [28] compare the Model Context Protocol (MCP), Agent Communication Protocol (ACP), Agent-to-Agent protocol (A2A) and Agent Network Protocol (ANP) and suggest starting with MCP for tool access. Hou et al. [29] list MCP security threats in the creation, operation and update phases of a server.

D. *LLM-Based Workflow and Process Automation*

FlowMind [30] gives the LLM a lecture-style prompt describing a fixed set of trusted APIs and asks it to write a workflow that uses only those APIs; the user reviews the workflow, and the model never touches raw financial data. ECLAIR [31] learns an enterprise workflow from human demonstrations, runs it through the user interface and checks completion, though the authors note that human review is still needed. AFlow [32] searches over workflows written as code with Monte Carlo tree search and reports a 5.7% average gain over prior methods, with smaller models matching GPT-4o on some tasks at 4.55% of its cost.

E. *Agents for Supply Chain and Retail Operations*

OptiGuide [33] lets supply chain planners ask what-if questions in plain language; the LLM writes code against an optimisation model, and the numbers come from the solver. InvAgent [34] uses zero-shot LLM agents for multi-echelon inventory and compares them with heuristic and multi-agent reinforcement learning policies. Jannelli et al. [35] let LLM agents negotiate as supply chain partners towards consensus on orders. Qi et al. [36] deployed a planning agent at JD.com that splits requests into subtasks for data and analysis agents, reporting about 40% less weekly data processing time, 22% more plans within 5% accuracy deviation and a 2 to 3% gain in stock fulfilment. Yoshizato et al. [37] found that a prompt containing a safety-stock formula reached the optimal policy only under constant demand, while an agent retrieving similar past cases from memory performed close to reinforcement learning baselines. The retail domain of tau-bench

© 2026 IJIRCCE | Volume 14, Issue 10, October 2026 DOI: 10.15680/IJIRCCE.2026.1410035

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

[38], covering returns, exchanges and order changes under a written policy, is the nearest benchmark to the customer side of a store.

F. *Benchmarks and Failure Analysis*

AgentBench [39] tests LLMs in eight environments and shows a wide gap between commercial and open models. WorkArena [40] measures web agents on 33 ServiceNow knowledge-work tasks. tau-bench [38] adds a simulated user and a policy document and introduces pass^k, which counts a task as solved only if all k trials succeed; GPT-4o solved about 61% of retail tasks once, but pass^8 fell below 25%. In TheAgentCompany [41], a simulated software firm, the best agent completed 30% of tasks on its own. Cemri et al. [42] built the MAST taxonomy of 14 failure modes in three groups: system design issues, inter-agent misalignment and weak task verification (kappa = 0.88).

G. *Governance, Reliability and Security*

Greshake et al. [43] showed that instructions hidden in retrieved web pages or emails can take over an LLM-integrated application, a risk that grows when agents read supplier documents or customer messages. Raza et al. [11] organise trust, risk and security management for agentic systems around explainability, model operations, security, privacy and governance, and Vu et al. [14] ask for clear accountability before agents enter processes. The reviewed systems use four controls: a restricted tool list [30], argument validation, human approval [6] and step logging, but few combine all four.

V. COMPARATIVE ANALYSIS

Table 1 compares the main collaboration and process automation systems, and Table 2 lists the benchmarks with their headline results.

TABLE 1. COMPARISON OF MULTI-AGENT AND PROCESS AUTOMATION SYSTEMS

Work	Year	Coordination	Tool use	Human oversight	Domain evaluated
CAMEL [22]	2023	Two-agent role play	No	Human sets task	Role-play tasks
AutoGen [6]	2024	Agent group chat	Yes	Human proxy	Maths, coding
MetaGPT [7]	2024	SOP-based roles	Limited	No	Software development
ChatDev [23]	2024	Phase-wise chat chain	Limited	No	Software development
AgentVerse [24]	2024	Dynamic expert group	Yes	No	Reasoning, coding
Magentic-One [26]	2024	Central orchestrator	Yes	Optional	Web and file tasks
ProAgent [13]	2023	Workflow + Data/Control	Yes	No	Business workflows
FlowMind [30]	2023	Single agent	Trusted APIs	User feedback	Finance workflows
ECLAIR [31]	2024	Pipeline	UI actions	Review advised	Enterprise workflows
AFlow [32]	2025	Searched workflow	Code operators	No	QA, code, maths
OptiGuide [33]	2023	Agent + solver	Solver	Planner queries	Supply chain
InvAgent [34]	2024	Agent per echelon	No	No	Inventory
SCPA [36]	2025	Router + sub-agents	Data, analysis	Planner review	E-commerce planning
Yoshizato et al. [37]	2026	One agent per tier	Memory retrieval	No	Inventory

Four patterns stand out. First, practical systems route work through one central orchestrator that holds the plan and state [26], [36], [15]. Second, numbers come from solvers, formulas or code, not from the LLM [33], [36], [37]. Third, human oversight is usually optional, and no reviewed framework ships a policy engine with value limits and an audit log. Fourth, no single system covers the joined retail processes of stock, delivery, customer support and sales analysis.

© 2026 IJIRCCE | Volume 14, Issue 10, October 2026 DOI: 10.15680/IJIRCCE.2026.1410035

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

TABLE 2. AGENT BENCHMARKS AND FAILURE STUDIES WITH HEADLINE RESULTS

Benchmark	Venue	Setting	Headline result
AgentBench [39]	ICLR 2024	8 environments (OS, DB, web)	Large commercial vs open gap
WorkArena [40]	ICML 2024	33 ServiceNow knowledge-work tasks	Web agents far from solved
tau-bench [38]	ICLR 2025	Retail, airline service with policy	GPT-4o retail: pass^1 61%, pass^8 < 25%
TheAgentCompany [41]	NeurIPS 2025	Simulated software company	Best agent completes 30% of tasks
MAST [42]	arXiv 2025	Traces from 7 MAS frameworks	14 failure modes, 3 groups

The benchmark results in Table 2 lead to the same conclusion. Strong models lose accuracy when a task is repeated or a policy must be followed, and MAST [42] traces many failures to weak verification and agent misalignment rather than the base model. For a store, this argues for verification and approval around high-risk actions rather than full autonomy.

VI. REFERENCE ARCHITECTURE FOR RETAIL PROCESS AUTOMATION

Fig. 2 shows a reference architecture that follows from the review. It is the design of the authors' project, in which a React Native Android app serves four roles: store staff, store manager, customer and owner. A Node.js gateway handles JWT login and role checks, a Python FastAPI service on LangGraph hosts the orchestrator and four domain agents, MongoDB Atlas with vector search stores business data and policies, and Redis carries the message queue and shared context.

Each design choice follows the literature. The orchestrator routes a task to the agent whose role description has the highest cosine similarity with the task embedding, with LLM classification as a fallback below a score of 0.55, following the centralised structure in [9], [26]. Each agent runs a ReAct loop [2] capped at five steps to avoid the loops reported in CAMEL [22], and acts only through registered tools with JSON-schema checks, as in FlowMind [30]. Forecasts and reorder points come from exponential smoothing and safety-stock formulas, as in [33], [37], and the customer agent answers from retrieved policy text [21], [27]. A governance layer sends any action above a role limit, such as a refund above Rs. 2,000, to the manager's phone for approval and logs every tool call, which addresses the verification gap in [42] and the framed autonomy in [12]. Implementation and evaluation on sample store data are planned for the next stage.

VII. OPEN CHALLENGES AND RESEARCH DIRECTIONS

Reliability over repeated runs. The pass^k results of tau-bench [38] show that an agent right once is not always right; business use needs consistency metrics, regression sets of real requests and fallbacks.

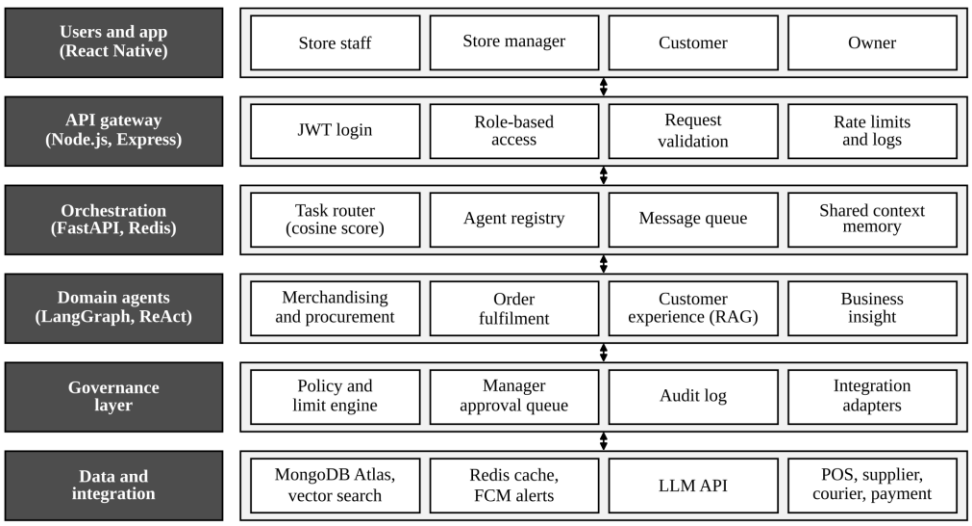


Fig. 2. Reference architecture for agentic retail process automation derived from the review

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Error spread and cost. A wrong value from one agent can move through the whole process [42], [9], and every ReAct step is a paid LLM call. Schema-checked outputs [7], verification agents and searched workflows that let smaller models do the job [32] help, but neither error spread nor cost is usually measured for small businesses.

Benchmarks for small retail. Current benchmarks cover software firms [41], IT service desks [40] and English customer service [38]; none covers stock control, supplier orders and returns in a small store, or requests in Hindi or Marathi.

Security and oversight. Prompt injection [43] and MCP server threats [29] grow more serious when agents can place orders or issue refunds [11], [28]. Approval limits also need calibration: too many requests and managers stop reading them, too few and risk rises [14], [12].

Learning from the process. Process mining [15] and memory-based agents [19], [37] could let a system learn from its own audit log, as long as changes stay inside approved limits.

VIII. CONCLUSION

This survey reviewed 43 works on agentic AI and multi-agent LLM systems from the point of view of business process automation. Collaboration frameworks, reasoning patterns, tool access and protocols are well developed, and work such as ProAgent, FlowMind, AFlow and the agentic BPM proposals shows how agents can replace fixed RPA scripts. Benchmarks, however, show that agents are still far from fully reliable on enterprise tasks, and failure studies point to weak verification and poor coordination. Governance with value limits, approval queues and audit logs is discussed often but implemented rarely, and no reviewed system covers the connected processes of a small retail store. The reference architecture outlined here is one way to close that gap, and building and measuring it on store data is the next step of this work.

REFERENCES

[1] M. Vidgof, S. Bachhofner and J. Mendling, "Large Language Models for Business Process Management: Opportunities and Challenges," in Business Process Management Forum (BPM 2023), Lecture Notes in Business Information Processing, Springer, 2023.
[2] S. Yao, J. Zhao, D. Yu et al., "ReAct: Synergizing Reasoning and Acting in Language Models," in Proc. Int. Conf. on Learning Representations (ICLR), 2023.
[3] T. Schick, J. Dwivedi-Yu, R. Dessi et al., "Toolformer: Language Models Can Teach Themselves to Use Tools," in Advances in Neural Information Processing Systems (NeurIPS), vol. 36, 2023.
[4] L. Wang, C. Ma, X. Feng et al., "A Survey on Large Language Model based Autonomous Agents," Frontiers of Computer Science, vol. 18, no. 6, Art. 186345, 2024.
[5] Z. Xi, W. Chen, X. Guo et al., "The Rise and Potential of Large Language Model Based Agents: A Survey," Science China Information Sciences, vol. 68, no. 2, Art. 121101, 2025.
[6] Q. Wu, G. Bansal, J. Zhang et al., "AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation," in Proc. Conf. on Language Modeling (COLM), 2024.
[7] S. Hong, M. Zhuge, J. Chen, X. Zheng, Y. Cheng et al., "MetaGPT: Meta Programming for a Multi-Agent Collaborative Framework," in Proc. Int. Conf. on Learning Representations (ICLR), 2024.
[8] R. Sapkota, K. I. Roumeliotis and M. Karkee, "AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenges," arXiv preprint arXiv:2505.10468, 2025.
[9] T. Guo, X. Chen, Y. Wang et al., "Large Language Model based Multi-Agents: A Survey of Progress and Challenges," in Proc. 33rd Int. Joint Conf. on Artificial Intelligence (IJCAI), pp. 8048-8057, 2024.
[10] K.-T. Tran, D. Dao, M.-D. Nguyen, Q.-V. Pham, B. O'Sullivan and H. D. Nguyen, "Multi-Agent Collaboration Mechanisms: A Survey of LLMs," arXiv preprint arXiv:2501.06322, 2025.
[11] S. Raza, R. Sapkota, M. Karkee and C. Emmanouilidis, "TRiSM for Agentic AI: A Review of Trust, Risk, and Security Management in LLM-based Agentic Multi-Agent Systems," arXiv preprint arXiv:2506.04133, 2025.
[12] D. Calvanese, A. Casciani, G. De Giacomo, M. Dumas, F. Fournier, T. Kampik et al., "Agentic Business Process Management: A Research Manifesto," arXiv preprint arXiv:2603.18916, 2026.
[13] Y. Ye, X. Cong, S. Tian, J. Cao, H. Wang, Y. Qin et al., "ProAgent: From Robotic Process Automation to Agentic Process Automation," arXiv preprint arXiv:2311.10751, 2023.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

[14] H. Vu, N. Klievtsova, H. Leopold, S. Rinderle-Ma and T. Kampik, "Agentic Business Process Management: Practitioner Perspectives on Agent Governance in Business Processes," arXiv preprint arXiv:2504.03693, 2025.
[15] M. Dumas, F. Milani and D. Chapela-Campa, "Agentic Business Process Management Systems," arXiv preprint arXiv:2601.18833, 2026.
[16] M. Azarijafari, L. Mich and M. Missikoff, "An Agentic AI for a New Paradigm in Business Process Development," arXiv preprint arXiv:2507.21823, 2025.
[17] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan and S. Yao, "Reflexion: Language Agents with Verbal Reinforcement Learning," in Advances in Neural Information Processing Systems (NeurIPS), vol. 36, 2023.
[18] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang and M. S. Bernstein, "Generative Agents: Interactive Simulacra of Human Behavior," in Proc. 36th ACM Symp. on User Interface Software and Technology (UIST), pp. 1-22, 2023.
[19] C. Packer, S. Wooders, K. Lin et al., "MemGPT: Towards LLMs as Operating Systems," arXiv preprint arXiv:2310.08560, 2023.
[20] Y. Qin, S. Liang, Y. Ye, K. Zhu, L. Yan, Y. Lu et al., "ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs," in Proc. Int. Conf. on Learning Representations (ICLR), 2024.
[21] Y. Gao, Y. Xiong, X. Gao et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," arXiv preprint arXiv:2312.10997, 2023.
[22] G. Li, H. A. A. K. Hammoud, H. Itani, D. Khizbullin and B. Ghanem, "CAMEL: Communicative Agents for 'Mind' Exploration of Large Language Model Society," in Advances in Neural Information Processing Systems (NeurIPS), vol. 36, 2023.
[23] C. Qian, W. Liu, H. Liu, N. Chen, Y. Dang, J. Li et al., "ChatDev: Communicative Agents for Software Development," in Proc. 62nd Annual Meeting of the Association for Computational Linguistics (ACL), pp. 15174-15186, 2024.
[24] W. Chen, Y. Su, J. Zuo, C. Yang, C. Yuan, C.-M. Chan et al., "AgentVerse: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors," in Proc. Int. Conf. on Learning Representations (ICLR), 2024.
[25] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum and I. Mordatch, "Improving Factuality and Reasoning in Language Models through Multiagent Debate," in Proc. 41st Int. Conf. on Machine Learning (ICML), 2024.
[26] A. Fourney, G. Bansal, H. Mozannar, C. Tan, E. Salinas, E. Zhu et al., "Magentic-One: A Generalist Multi-Agent System for Solving Complex Tasks," arXiv preprint arXiv:2411.04468, 2024.
[27] A. Singh, A. Ehtesham, S. Kumar and T. T. Khoei, "Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG," arXiv preprint arXiv:2501.09136, 2025.
[28] A. Ehtesham, A. Singh, G. K. Gupta and S. Kumar, "A Survey of Agent Interoperability Protocols: Model Context Protocol (MCP), Agent Communication Protocol (ACP), Agent-to-Agent Protocol (A2A), and Agent Network Protocol (ANP)," arXiv preprint arXiv:2505.02279, 2025.
[29] X. Hou, Y. Zhao, S. Wang and H. Wang, "Model Context Protocol (MCP): Landscape, Security Threats, and Future Research Directions," arXiv preprint arXiv:2503.23278, 2025.
[30] Z. Zeng, W. Watson, N. Cho et al., "FlowMind: Automatic Workflow Generation with LLMs," in Proc. 4th ACM Int. Conf. on AI in Finance (ICAIF), pp. 73-81, 2023.
[31] M. Wornow, A. Narayan, K. Opsahl-Ong, Q. McIntyre, N. H. Shah and C. Re, "Automating the Enterprise with Foundation Models," arXiv preprint arXiv:2405.03710, 2024.
[32] J. Zhang, J. Xiang, Z. Yu, F. Teng, X. Chen, J. Chen et al., "AFlow: Automating Agentic Workflow Generation," in Proc. Int. Conf. on Learning Representations (ICLR), 2025.
[33] B. Li, K. Mellou, B. Zhang, J. Pathuri and I. Menache, "Large Language Models for Supply Chain Optimization," arXiv preprint arXiv:2307.03875, 2023.
[34] Y. Quan and Z. Liu, "InvAgent: A Large Language Model based Multi-Agent System for Inventory Management in Supply Chains," arXiv preprint arXiv:2407.11384, 2024.
[35] V. Jannelli, S. Schoepf, M. Bickel, T. Netland and A. Brintrup, "Agentic LLMs in the Supply Chain: Towards Autonomous Multi-Agent Consensus-Seeking," arXiv preprint arXiv:2411.10184, 2024.
[36] Y. Qi, J. Yin, J. Zhang et al., "Leveraging LLM-Based Agents for Intelligent Supply Chain Planning," arXiv preprint arXiv:2509.03811, 2025.
[37] K. Yoshizato, K. Shimizu, R. Higa and T. Otsuka, "AI Agent Systems for Supply Chains: Structured Decision Prompts and Memory Retrieval," arXiv preprint arXiv:2602.05524, 2026.
[38] S. Yao, N. Shinn, P. Razavi and K. Narasimhan, "tau-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains," in Proc. Int. Conf. on Learning Representations (ICLR), 2025.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

[39] X. Liu, H. Yu, H. Zhang, Y. Xu, X. Lei, H. Lai et al., "AgentBench: Evaluating LLMs as Agents," in Proc. Int. Conf. on Learning Representations (ICLR), 2024.

[40] A. Drouin, M. Gasse, M. Caccia, I. H. Laradji, M. Del Verme, T. Marty et al., "WorkArena: How Capable Are Web Agents at Solving Common Knowledge Work Tasks?," in Proc. 41st Int. Conf. on Machine Learning (ICML), 2024.

[41] F. F. Xu, Y. Song, B. Li, Y. Tang, K. Jain, M. Bao et al., "TheAgentCompany: Benchmarking LLM Agents on Consequential Real World Tasks," in Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, 2025.

[42] M. Cemri, M. Z. Pan, S. Yang et al., "Why Do Multi-Agent LLM Systems Fail?," arXiv preprint arXiv:2503.13657, 2025.

[43] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz and M. Fritz, "Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection," in Proc. 16th ACM Workshop on Artificial Intelligence and Security (AISec), pp. 79-90, 2023.



INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details