Evolution of Agentic Artificial Intelligence: From Classical Intelligent Agents to LLM-Based Autonomous Systems

Abhishek Sharma[1], Surjeet Sah[2], Mohammad Sayeed[3]

[1]*Professor, Department of Electronics & Communication Engineering, Shivalik College of Engineering, Dehradun*
[2]*Assistant Professor, Department of Electronics and Communication Engineering, JBIT College, Dehradun*
[3]*Assistant Professor, Department of Civil Engineering, Shivalik College of Engineering, Dehradun*

Emails: *abhishek.kaushik1@gmail.com, surjeetsah2526@gmail.com, sayeedraza666@gmail.com*
Corresponding Author Email: *sayeedraza666@gmail.com*

Abstract: Agentic Artificial Intelligence (Agentic AI) refers to AI systems that can plan, take actions and adapt their behavior to achieve goals with limited human supervision. This review traces the evolution of agentic AI from classical intelligent agents and multi-agent systems to today's Large Language Model (LLM)-based agents that use tools, memory and multi-step reasoning. The chapter synthesizes key ideas, architectures, enabling technologies, evaluation methods, real-world applications and emerging safety and governance practices. It highlights how modern agentic systems combine planning, tool use, retrieval, reflection and coordination across multiple agents, while also introducing new risks such as unsafe autonomy, prompt injection and reliability failures. Finally, the review proposes research directions toward more robust, transparent and accountable agentic AI suitable for high-stakes deployment.

Keywords: *Agentic AI, autonomous agents, large language models, tool use, planning, multi-agent systems, memory, safety, governance.*

1. Introduction

Artificial intelligence has long aimed to build systems that do more than produce predictions or classifications. A central ambition has been to create *agents*: entities that perceive an environment, decide what to do, and act to achieve goals (Russell & Norvig, 2020). Early work in intelligent agents focused on symbolic planning, decision-theoretic reasoning, and multi-agent coordination (Wooldridge, 2009).

In 2023–2025, "agentic AI" became widely discussed due to LLMs being embedded inside control loops that allow them to plan, call tools (search, code execution, databases) and iteratively improve outputs based on feedback (Wang et al., 2023; Yao et al., 2022). Recent surveys formalize *agentic LLMs* as systems that reason, act and interact, often with memory and tool access (Wang et al., 2023; "Agentic Large Language Models, a survey," 2025). Industrial and public-facing explanations similarly describe agentic AI as autonomous systems that plan and act to complete tasks with minimal oversight (IBM, 2025; University of Cincinnati, 2025).

This review paper provides a structured account of how agentic AI evolved, what modern agent architectures look like, how they are evaluated, where they are applied, and what safety/governance mechanisms are increasingly required (NIST, 2023; European Union, 2024).

2. Conceptual Foundations of Agents and Agency

2.1 What is an agent?

In classical AI, an agent is commonly defined as a system that maps perceptions to actions to maximize performance in an environment (Russell & Norvig, 2020). In multi-agent systems (MAS), many agents interact, cooperate, or compete, requiring communication and coordination mechanisms (Wooldridge, 2009).

2.2 Human planning and the inspiration for agentic behavior

A long-standing theme in cognitive science and philosophy is that purposeful behavior depends on intentions, plans, and the ability to revise plans when circumstances change (Bratman, 1987). This idea strongly aligns with modern agentic AI, where systems maintain goals, generate plans and update them based on new evidence or tool outputs.

2.3 Reinforcement learning and sequential decision-making

Reinforcement learning (RL) formalizes learning to act through reward-driven interaction (Sutton & Barto, 2018). While many LLM agents are not trained via classical RL in deployment, the *interaction loop* which observe, decide, act, receive feedback closely resembles RL settings and hybrid approaches increasingly blend LLM reasoning with RL-like feedback signals (Shinn et al., 2023).

3. Historical Evolution of Agentic AI

3.1 Era I: Symbolic AI and planning-based agents

Early agent systems relied on symbolic representations, rule-based reasoning, and explicit planning. These approaches were interpretable but brittle in complex real-world settings (Russell & Norvig, 2020). Multi-agent research developed protocols for coordination and negotiation, but often assumed structured environments (Wooldridge, 2009).

3.2 Era II: Statistical AI and learning-driven agents

With machine learning, agents became more adaptive. RL achieved major milestones in games and control, but training was expensive, data-hungry and often domain-specific (Sutton & Barto, 2018).

3.3 Era III: Foundation models and language-centric agency

LLMs introduced a major shift: natural language became a general interface for goals, plans, tools and environments. LLMs can interpret instructions, generate step-by-step strategies, and manage task decomposition (Achiam et al., 2023). This accelerated the development of LLM-based autonomous agents (Wang et al., 2023).

Table 1. Comparison Between Traditional AI Systems and Agentic AI

Feature	Traditional AI Systems	Agentic AI Systems
Output Type	Single prediction/classification	Multi-step goal-oriented actions
Autonomy	Minimal	Moderate to high
Memory	No persistent memory	External structured memory
Tool Usage	Not integrated	Built-in tool calling capability
Planning	Not explicit	Explicit multi-step planning
Risk Profile	Output-level error	Action-level system risk

4. The Modern Agentic AI Stack (LLM-Based Agents)

A practical way to understand modern agentic AI is to view it as an architecture, not just a model. Many systems follow a pattern:

1. Goal interpretation (convert user intent into objectives)
2. Planning (create or refine a multi-step plan)
3. Tool use (call APIs/tools for search, computation, code, databases)
4. Memory and retrieval (store and fetch relevant context)
5. Reflection and self-correction (learn from feedback without retraining)
6. Execution monitoring (detect errors, revise plan, continue)

Flow chart structure of the modern Agentic AI stack system is shown in figure 1.

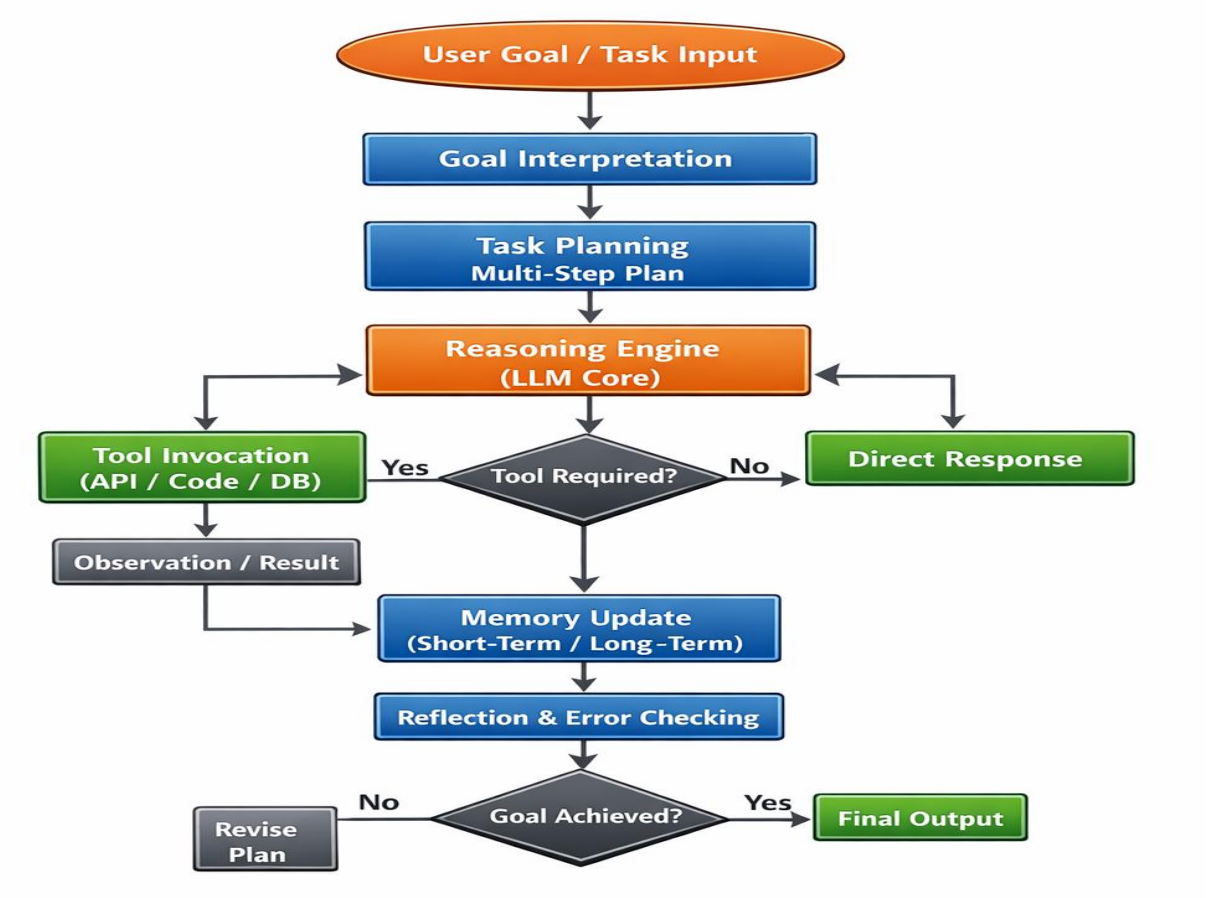

Figure 1. Unified Workflow of an LLM-Based Agentic AI System

4.1 Reasoning, Tool Use, and Memory Integration

Modern agentic systems combine reasoning with structured external interaction.

• Interleaved Reasoning and Action

- Agents alternate between internal reasoning and external actions.
- Improves factual grounding and robustness.
- Demonstrated in ReAct (Yao et al., 2022).

• Tool Calling and Functional Interfaces

- Agents invoke APIs, code execution, search engines, and databases.
- Reduces hallucination and increases precision.
- Toolformer showed models can learn tool usage (Schick et al., 2023).
- Industrial systems formalize schema-based function calling (OpenAI, 2025b).

• Retrieval-Augmented Generation and Agent Memory

- External retrieval supports up-to-date, evidence-based outputs (Lewis et al., 2020).
- Agent memory stores task history, intermediate results, and contextual state (Weng, 2023).
- Enables long-horizon, goal-oriented execution.

Core Insight:

Reasoning + tools + retrieval + memory transform LLMs into structured decision systems.

4.2 Reflection, Self-Correction, and Adaptive Control

Beyond execution, modern agents incorporate improvement mechanisms.

• Reflection-Based Learning

- Agents generate self-reflections after feedback.
- Stored reflections guide future attempts.
- Introduced in Reflexion (Shinn et al., 2023).
- Enables improvement without retraining.

• Iterative Monitoring and Plan Revision

- Detect failed tool calls or inconsistencies.
- Revise plans dynamically.
- Reduce long-horizon error propagation.

• Controlled Autonomy

- Execution monitoring ensures reliability.
- Supports bounded, goal-driven autonomy.

Architectural Shift:

From single-response generation → to iterative think–act–observe–reflect cycles.

5. Agent Architectures: Design Patterns and Taxonomy

Modern agentic AI systems follow diverse architectural patterns depending on task complexity, autonomy requirements, and domain specialization. These architectures can be broadly

categorized into centralized, collaborative, and adaptive autonomous structures.

5.1 Single-Agent and Modular Architectures

Single-agent architectures represent the most direct implementation of agentic AI, where a primary LLM-based controller is responsible for interpreting goals, generating plans, invoking tools, and managing execution. These systems typically integrate retrieval mechanisms and structured memory while maintaining centralized decision-making authority. Their relative simplicity enables efficient deployment and clear accountability; however, they may encounter limitations when handling highly complex, multi-domain workflows (Wang et al., 2023). To enhance reliability and compositional reasoning, modular neuro-symbolic approaches extend this paradigm by routing specific subtasks to specialized components such as symbolic solvers, calculators, or retrievers. The MRKL framework (Karpas et al., 2022) exemplifies this hybrid design, combining language understanding with discrete reasoning modules to improve precision and interpretability. Overall, this architectural class emphasizes centralized coordination supported by modular functional specialization.

5.2 Multi-Agent and Role-Based Collaborative Architectures

As task complexity increases, distributed cognitive architectures become advantageous. Multi-agent systems divide responsibilities across specialized agents that collaborate through structured communication protocols. In such systems, roles may include planner, coder, critic, or reviewer, allowing for division of expertise and iterative cross-validation. Frameworks such as AutoGen (Wu et al., 2023) demonstrate how coordinated agent conversations can enhance problem-solving performance and reduce errors through structured dialogue. Beyond conversational collaboration, agent–computer interface models further extend this paradigm by enabling agents to interact directly with software repositories, file systems, and development tools. The SWE-agent system (Yang et al., 2024) illustrates how structured interaction with digital environments can automate software engineering tasks under controlled operational constraints. Collectively, collaborative architectures introduce distributed intelligence, role specialization, and enhanced robustness through coordinated execution.

5.3 Generative and Embodied Autonomous Agents

• Generative Agents for Simulation

- Maintain memory, reflect, and plan activities over time.
- Produce believable social and emergent behaviors.

- Example: Generative Agents (Park et al., 2023).
- Applicable to simulation, education, and behavioral research.

• Embodied and Open-Ended Skill Learning Agents

- Operate within interactive environments.
- Build skill libraries and improve iteratively.
- Example: Voyager in Minecraft (Wang et al., 2023).
- Represent a step toward persistent, cumulative autonomy.

Key Characteristic:

Long-horizon memory, reflection, and skill accumulation.

6. Evaluation of Agentic AI

Evaluating agentic AI is harder than evaluating single-turn models because success depends on multi-step execution, tool interactions, and safety constraints (Wang et al., 2023).

Table 2. Core Evaluation Dimensions for Agentic AI

Dimension	Description	Example Measure
Task Success	Goal completion accuracy	Success rate
Efficiency	Resource utilization	Steps / API calls
Robustness	Error recovery capability	Recovery ratio
Groundedness	Factual reliability	Citation support
Safety	Policy compliance	Violation rate
Human Oversight	Intervention needs	Approval frequency

6.1 Core evaluation dimensions

Common dimensions include:

- Task success (did it achieve the goal?)
- Efficiency (steps, time, tool calls, cost)
- Robustness (does it recover from errors?)
- Groundedness (are outputs supported by retrieved evidence?)
- Safety and policy compliance (does it avoid unsafe actions?)
- Human satisfaction and oversight burden (how much intervention is needed?)

These dimensions are emphasized across surveys and practice-oriented guides (Wang et al., 2023; Weng, 2023; OpenAI, 2025a).

7. Applications of Agentic AI

Agentic AI applications can be broadly grouped into two major domains:

(1) Cognitive–Digital Task Automation and

(2) Interactive–Autonomous Environments.

7.1 Cognitive and Enterprise Task Automation

This category includes applications where agents operate within digital ecosystems to support knowledge work, engineering, and organizational processes.

Agentic systems:

- Automate literature review, summarization, and analytical reporting
- Support research synthesis and structured decision-making (Lewis et al., 2020; Weng, 2023)
- Navigate repositories, generate code, test patches, and refine outputs (Yang et al., 2024)
- Coordinate planner–reviewer workflows in multi-agent settings (Wu et al., 2023)
- Orchestrate enterprise operations across databases and internal tools (OpenAI, 2025a; IBM, 2025)

Primary Contribution:

Improved efficiency, structured automation, and reduced manual coordination.

Primary Risk:

Incorrect actions at scale without adequate verification or governance.

7.2 Simulation, Training, and Embodied Autonomy

This category includes applications where agents operate in interactive or simulated environments requiring long-horizon reasoning and adaptive behavior.

Agentic systems:

- Maintain persistent memory for simulated populations (Park et al., 2023)
- Model social and behavioral dynamics for training and education
- Operate in open-ended environments with cumulative skill learning (Wang et al., 2023)
- Demonstrate long-horizon planning and iterative adaptation

Primary Contribution:

Realistic simulation, training support, and progressive autonomous capability.

Primary Challenge:

Safety assurance, bias control, and environmental uncertainty in real-world deployment.

APPLICATION SPECTRUM OF AGENTIC AI

Cognitive & Enterprise Automation	Simulation and Autonomous systems
• Knowledge synthesis	• Social simulation
• Report drafting	• Training environments
• Software engineering	• Embodied agents

- Workflow orchestration
  Digital tool interaction
  Structured workflow
- Open-ended exploration
  Interactive environment
  Adaptive behavior

Increasing Autonomy and Environmental Interaction → This merged structure improves clarity by:

• Reducing sectional fragmentation

• Grouping related domains logically

• Highlighting contributions and risks

• Providing a clean visual conceptual understanding

If needed, I can now similarly condense Section 8 (Safety and Governance) while preserving regulatory depth.

8. Safety, Security, and Governance

Agentic AI introduces risks beyond typical chatbots because it can take actions, trigger workflows and interact with external systems.

8.1 Key risks

- **Hallucinated actions**: incorrect tool calls or unsafe commands (Yao et al., 2022).
- **Prompt injection and tool manipulation**: malicious content can steer an agent to reveal data or perform unwanted actions (Weng, 2023).
- **Over-autonomy**: agents may act beyond intended scope, especially when goals are ambiguous (University of Cincinnati, 2025).
- **Accountability gaps**: unclear responsibility when an autonomous workflow cause harm (NIST, 2023).

8.3 Practical controls for safer agentic systems

Common controls recommended across industry guides and technical practice include:

- Tool permissions and least privilege (limit what an agent can do) (OpenAI, 2025a).
- Human-in-the-loop approval for high-impact actions (NIST, 2023).
- Auditing and tracing of tool calls and decisions (OpenAI, 2025a).
- Grounding and verification using retrieval, citations, and test execution where possible (Lewis et al., 2020; Yang et al., 2024).
- Sandboxing for code execution and system actions (Yang et al., 2024).

9. Open Challenges and Research Directions

9.1 Technical and Operational Challenges

Agentic AI systems continue to face challenges related to reliability, long-horizon planning, and real-world robustness. Overconfident errors and limited calibration hinder deployment in high-stakes contexts (Achiam et al., 2023; Wang et al., 2023). Sustained multi-step tasks introduce memory instability and goal drift, particularly when contextual information accumulates over time (Park et al., 2023; Weng, 2023). Existing benchmarks also fail to reflect practical complexities such as tool failures, ambiguous objectives, and adversarial inputs (Wang et al., 2023). Strengthening uncertainty estimation, memory management, and secure tool interfaces remains essential for dependable operation.

9.2 Governance and Responsible Autonomy

As agentic systems gain greater autonomy, governance alignment becomes critical. Frameworks such as the NIST AI Risk Management Framework (NIST, 2023) and the EU AI Act (European Union, 2024) emphasize transparency, accountability, and structured oversight. Agentic AI must therefore incorporate auditable decision processes, controlled tool access, and bounded autonomy, where high-impact or uncertain actions are escalated to human supervision (OpenAI, 2025a; NIST, 2023). Future progress depends on integrating technical safeguards with regulatory and institutional controls to ensure responsible deployment.

10. Conclusion

Agentic Artificial Intelligence has progressed from classical intelligent agent theory and multi-agent systems to advanced LLM-based architectures capable of planning, tool invocation, retrieval grounding, memory persistence, and reflection-driven improvement. This evolution marks a transition from static predictive models to iterative, goal-oriented systems that integrate reasoning and action within structured control loops.

Modern agentic systems demonstrate significant practical value across knowledge work, software engineering, enterprise automation, simulation, and embodied environments. Architectural developments such as modular design, multi-agent collaboration, and retrieval-augmented memory have expanded autonomy while improving task robustness and adaptability.

However, increased autonomy introduces critical challenges. Reliability, long-horizon stability, secure tool interaction, and realistic evaluation remain open technical issues. Equally important are governance integration, transparency, bounded autonomy, and alignment with emerging regulatory frameworks.

Future advancement depends on achieving responsible autonomy systems that are not only capable and adaptive but also auditable, controllable, and aligned with institutional oversight. Sustainable

progress in agentic AI will require embedding safety, evaluation rigor, and accountability directly within system architecture.

References

1 **Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Alembic Team, … OpenAI. (2023).** *GPT-4 technical report* (arXiv:2303.08774). arXiv.

2 **Bratman, M. E. (1987).** *Intention, plans, and practical reason*. Harvard University Press.

3 **European Union. (2024).** *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Official Journal of the European Union.

4 **IBM. (2025).** *What are AI agents?* IBM.

5 **Karpas, E., Abend, O., Belinkov, Y., Lenz, B., Lieber, O., Ratner, N., … Tenenholtz, M. (2022).** *MRKL Systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning* (arXiv:2205.00445). arXiv.

6 **Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., … Kiela, D. (2020).** *Retrieval-augmented generation for knowledge-intensive NLP tasks* (arXiv:2005.11401). arXiv.

7 **National Institute of Standards and Technology. (2023).** *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). NIST.

8 **OpenAI. (2025a).** *A practical guide to building agents*. OpenAI.

9 **OpenAI. (2025b).** *Function calling (tool calling) guide*. OpenAI Platform Documentation.

10 **Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023).** *Generative agents: Interactive simulacra of human behavior* (arXiv:2304.03442). arXiv.

11 **Piccialli, F., Giampaolo, F., Cuomo, S., & others. (2025).** AgentAI: A comprehensive survey on autonomous agents in distributed AI for Industry 4.0. *Expert Systems with Applications*.

12 **Russell, S. J., & Norvig, P. (2020).** *Artificial intelligence: A modern approach* (4th ed.). Pearson.

13 **Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023).** *Toolformer: Language models can teach themselves to use tools* (arXiv:2302.04761). arXiv.

14 **Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., & Yao, S. (2023).** *Reflexion: Language agents with verbal reinforcement learning* (arXiv:2303.11366). arXiv.

15 **Sutton, R. S., & Barto, A. G. (2018).** *Reinforcement learning: An introduction* (2nd ed.). MIT Press.

16 **University of Cincinnati. (2025).** *What is agentic AI? (Definition and 2025 guide).*

17 **Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., & Anandkumar, A. (2023).** *Voyager: An open-ended embodied agent with large language models* (arXiv:2305.16291). arXiv.

18 **Wang, L., Ma, Z., Zhang, R., Ni, X., Zhang, J., Zhang, Y., … Liu, X. (2023).** *A survey on large language model based autonomous agents* (arXiv:2308.11432). arXiv.

19 **Weng, L. (2023).** *LLM powered autonomous agents*. Lil'Log.

20 **Wooldridge, M. (2009).** *An introduction to multiagent systems* (2nd ed.). Wiley.

21 **Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., & Wang, C. (2023).** *AutoGen: Enabling next-gen LLM applications via multi-agent conversation* (arXiv:2308.08155). arXiv.

22 **Yang, J., Jimenez, C., Wettig, A., Yao, S., Narasimhan, K., & others. (2024).** *SWE-agent: Agent-computer interfaces enable automated software engineering* (arXiv:2405.15793). arXiv.

23 **Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022).** *ReAct: Synergizing reasoning and acting in language models* (arXiv:2210.03629). arXiv.

24 **Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023).** *Tree of thoughts: Deliberate problem solving with large language models* (arXiv:2305.10601). arXiv.

25 "**Agentic Large Language Models, a survey." (2025).** *arXiv* (arXiv:2503.23037).