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

Emails: abhishek.kaushik1@gmail.com, sayeedraza666@gmail.com, surjeetsah2526@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

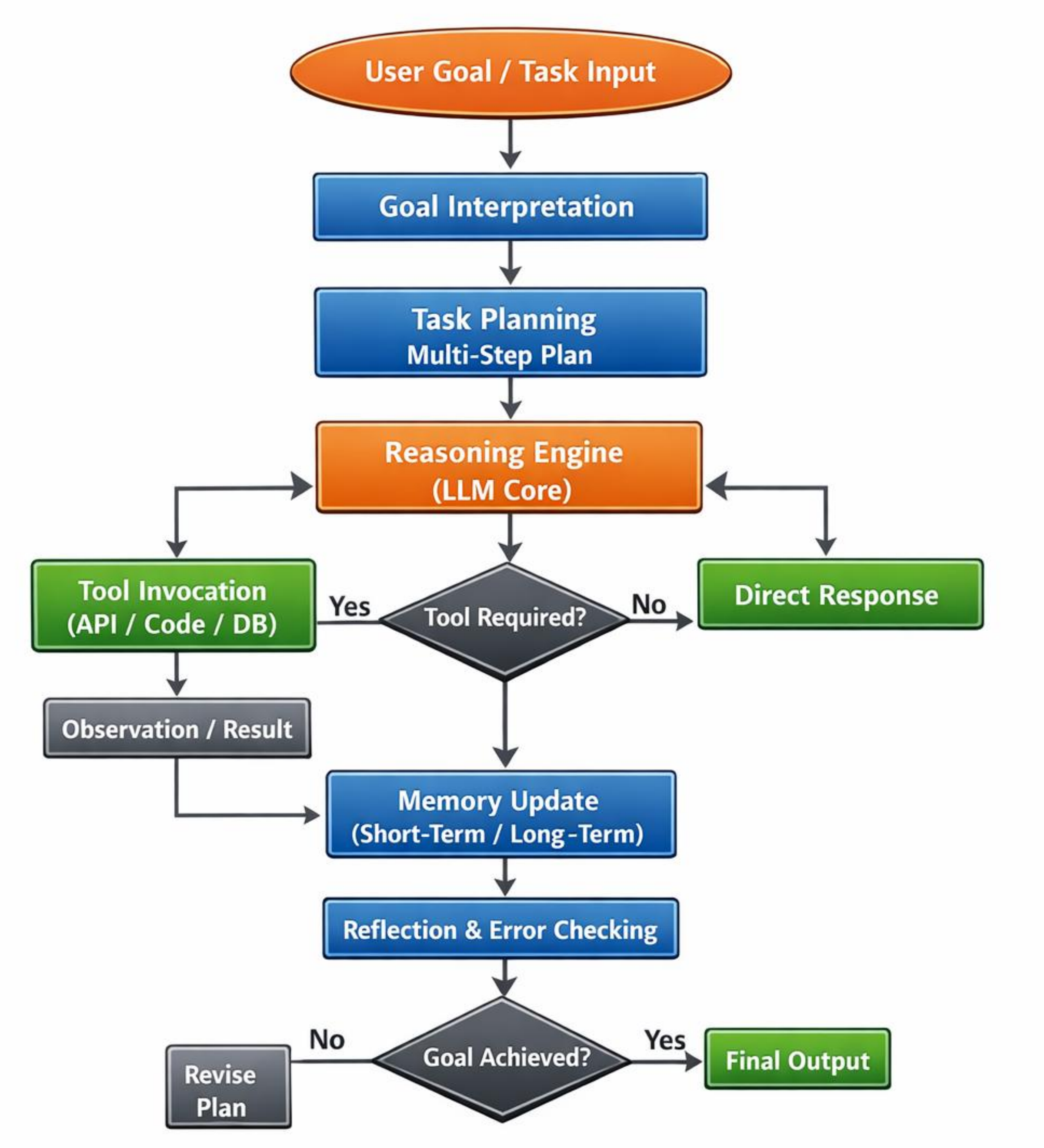

Figure 1. Unified Workflow of an LLM-Based Agentic AI System

4.1 Reasoning + acting: interleaving thought and tool use

A landmark idea is that reasoning and action should be interleaved rather than separated. ReAct demonstrates that agents can alternate between reasoning traces and actions (e.g., searching a knowledge source), improving factuality and robustness on interactive tasks (Yao et al., 2022).

4.2 Tool learning and tool calling

Tool use addresses a core limitation: models can hallucinate or fail at precise operations (e.g., calculation, fresh facts). Toolformer showed that language models can learn to decide when and how to call tools using self-supervision (Schick et al., 2023). Industrial tool-calling systems formalize tool interfaces so models can reliably call functions defined by schemas (OpenAI, 2025b).

4.3 Retrieval and external memory

Retrieval-Augmented Generation (RAG) improved knowledge-intensive tasks by retrieving documents from an external index and conditioning generation on them, supporting more up-to-date and source-grounded responses (Lewis et al., 2020). Many agents extend this idea into *agent memory*: storing task history, intermediate results, and reflections for future steps (Weng, 2023).

4.4 Reflection and self-improvement without weight updates

Reflexion introduced a practical method: agents store "reflections" in language after receiving feedback, then use those reflections as memory to improve subsequent attempts (Shinn et al., 2023). This is important because it reduces the need for costly model retraining while still enabling iterative improvement.

5. Agent Architectures: Design Patterns and Taxonomy

5.1 Single-agent planners

These systems use one primary agent that plans and executes tasks, often using tools and memory. They are simpler to build but can struggle with complex workflows or domain breadth (Wang et al., 2023).

5.2 Multi-agent collaboration and role-based systems

Multi-agent frameworks assign roles (e.g., planner, coder, reviewer) and allow structured conversation among agents. AutoGen is a well-known framework enabling multi-agent conversation patterns for problem solving and tool-based workflows (Wu et al., 2023).

5.3 Neuro-symbolic modular systems

MRKL systems proposed a modular approach: route tasks to specialized modules (symbolic solvers, retrievers, calculators) while using an LLM for language understanding and orchestration (Karpas et al., 2022). This supports compositionality and can improve reliability for tasks requiring precise reasoning.

5.4 "Generative agents" for social simulation

Generative Agents demonstrated agents that maintain memories, reflect, and plan daily activities, producing believable individual and emergent group behavior in simulated environments (Park et al., 2023). This opened new directions in simulation, games, and social science prototyping ("Agentic Large Language Models, a survey," 2025).

5.5 Embodied agents and open-ended skill learning

Voyager showed an LLM-based agent that explores an open-ended environment (Minecraft), builds a library of skills, and improves through feedback and iterative prompting (Wang et al., 2023). This is a key step toward agents that can accumulate competencies over time.

5.6 Agent-computer interfaces for software engineering

SWE-agent demonstrated how agent-computer interfaces can automate software engineering tasks by letting an agent operate with structured interaction on repositories, files and tools (Yang et al., 2024). This illustrates a broader class of agents that operate on digital systems in a controlled manner.

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).

6.2 Benchmarks and realistic testing

Agentic systems often require *scenario-based evaluation* (e.g., shopping environments, tool-based QA, coding tasks). ReAct reported improvements on interactive environments like ALFWorld and WebShop (Yao et al., 2022). SWE-agent evaluates real software engineering tasks (Yang et al., 2024).

A persistent challenge is that lab benchmarks may not capture messy real-world conditions, such as partial tool failures, ambiguous goals, or adversarial inputs (Wang et al., 2023).

7. Applications of Agentic AI

7.1 Knowledge work and decision support

Agents can automate literature search, summarize evidence, draft reports and manage multi-step workflows using retrieval and tools (Lewis et al., 2020; Weng, 2023). However, they require strong grounding and verification to be used responsibly.

7.2 Software engineering and DevOps

Agentic systems can navigate repositories, propose patches, run tests and iterate—illustrated by SWE-agent (Yang et al., 2024). Multi-agent setups can separate responsibilities (coding vs. reviewing) to reduce errors (Wu et al., 2023).

7.3 Operations, business processes, and customer support

Organizations increasingly adopt “agent” patterns to orchestrate tasks across tools, databases, and internal systems (OpenAI, 2025a; IBM, 2025). The benefit is reduced manual coordination; the risk is incorrect actions at scale.

7.4 Simulation, training, and education

Generative Agents enable simulated populations for training scenarios and prototyping social dynamics (Park et al., 2023). This can support education and behavioral research, but raises ethical questions about realism, bias and consent.

7.5 Embodied robotics and digital autonomy

Embodied agents like Voyager show promise for open-ended learning (Wang et al., 2023). Translating these ideas to real robotics requires addressing safety, sensing uncertainty and physical-world constraints.

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 causes harm (NIST, 2023).

8.2 Risk management frameworks

The NIST AI Risk Management Framework provides structured guidance to identify and manage AI risks across the lifecycle (NIST, 2023). For regulatory governance, the EU AI Act (Regulation (EU) 2024/1689) lays down harmonized rules for AI systems, including obligations for certain risk categories (European Union, 2024). These frameworks are increasingly relevant as agentic systems move into regulated and high-impact domains.

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 Reliability and calibration

Agents need to know when they are uncertain and request help. Overconfident failures remain a major barrier to deployment (Achiam et al., 2023; Wang et al., 2023).

9.2 Long-horizon planning and memory

Long tasks require stable goals, durable memory and the ability to revise plans without drifting. Memory systems can also introduce privacy and security issues if not managed properly (Park et al., 2023; Weng, 2023).

9.3 Evaluation for real-world conditions

More realistic evaluations are needed, including adversarial settings, tool outages and ambiguous objectives (Wang et al., 2023).

9.4 Governance for autonomous action

Regulatory expectations and organizational accountability will shape the safe use of agentic AI (European Union, 2024; NIST, 2023). Technical work should align with governance by supporting transparency, auditability and controllability (OpenAI, 2025a).

9.5 Toward "responsible autonomy"

A likely direction is *bounded agency*: agents that can act autonomously within explicit constraints, with strong monitoring and escalation to humans for uncertain or high-impact actions (NIST, 2023; OpenAI, 2025a).

10. Conclusion

Agentic AI has evolved from classical intelligent agent concepts and multi-agent coordination into modern LLM-based systems capable of planning, tool use, memory, reflection and multi-agent collaboration. Research from ReAct, Toolformer, Reflexion, Tree of Thoughts and major frameworks such as AutoGen, SWE-agent and Voyager shows rapid progress in practical autonomy. At the same time, agentic AI amplifies safety and governance concerns because it can take actions and affect real systems. Future progress toward publishable, real-world-grade agentic AI will depend on stronger evaluation, robust grounding, principled memory, secure tool interfaces and alignment with risk management and regulatory frameworks.

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).