International Journal of Sustainable Development Through AI, ML and IoT
Volume 4 | Issue2 | 2025
https://ijsdai.com/index.php/IJSDAI/index
ISSN (Online): 2584-0827

Agentic Artificial Intelligence: A Systematic Review of Architectures, Capabilities, Applications, and Open Challenges

Pavika Sharma [1]
BPIT, India
Corresponding author: ***Dr. Pavika Sharma***

ARTICLE INFO
Received: 31 May 2025
Revised: 20 Nov 2025
Accepted: 31 Dec 2025

ABSTRACT

Agentic artificial intelligence (AI) denotes a class of systems, typically built on large language models (LLMs), that autonomously perceive context, reason over multi-step goals, invoke external tools, retain and update memory, and act within digital or physical environments with limited human intervention. The rapid transition from single-turn generative models to goal-directed, tool-using, and often multi-agent systems has produced a large and fast-growing body of research spanning reasoning and planning strategies, memory architectures, multi-agent coordination protocols, evaluation benchmarks, and domain-specific deployments. This paper presents a structured review of the agentic AI literature, synthesizing contributions from foundational reasoning frameworks such as Chain-of-Thought and ReAct, tool-augmented approaches such as Toolformer and ToolLLM, reflective and memory-augmented agents such as Reflexion and MemGPT, and multi-agent orchestration frameworks such as AutoGen and MetaGPT. A consolidated taxonomy of core agentic capabilities is proposed, covering reasoning and planning, tool use, memory management, self-reflection, and multi-agent collaboration.

Keywords: *Agentic AI; large language model agents; multi-agent systems; reasoning and planning; tool use; autonomous agents; agent evaluation benchmarks; AI safety and governance*.

1. Introduction

Artificial intelligence research has progressed through several paradigm shifts, from symbolic rule-based expert systems to statistical machine learning, and more recently to large-scale pre-trained language models capable of open-ended text generation. Over the past three years, a further shift has taken shape: the transition from passive, prompt-response generative models toward agentic systems that autonomously plan, invoke external tools, retain memory across interactions, and pursue multi-step objectives with reduced human supervision. This paradigm,

commonly termed agentic AI, extends the capabilities of large language models (LLMs) beyond text generation into dynamic interaction with software environments, application programming interfaces (APIs), other agents, and, in embodied settings, physical surroundings.

The motivation for this shift is twofold. First, many real-world tasks, such as software debugging, multi-step research synthesis, financial analysis, or clinical triage support, require iterative reasoning, intermediate verification, and interaction with external systems that a single forward pass of a language model cannot provide. Second, empirical evidence from frameworks such as ReAct and Reflexion demonstrated that interleaving reasoning traces with actions and feedback substantially improves task success rates relative to static prompting, motivating a wave of research into agent architectures, memory systems, and multi-agent coordination.

Despite the volume of recent publications, the agentic AI literature remains fragmented across several largely disconnected sub-communities: prompting and reasoning research, tool-augmented language modeling, multi-agent systems research with roots in distributed AI, and, more recently, AI safety and security research addressing the risks of autonomous agent deployment. Existing surveys tend to specialize in one of these threads, for example reasoning strategies, multi-agent coordination, or security, without offering a single consolidated account that connects the technical foundations of agentic AI to its practical applications and governance challenges. This fragmentation makes it difficult for newcomers to the field, as well as practitioners evaluating agentic AI for deployment, to obtain a coherent picture of the state of the art.

This review addresses that gap by synthesizing the literature across four interconnected layers: (i) the core technical capabilities that constitute an agentic system, namely reasoning and planning, tool use, memory, reflection, and multi-agent collaboration; (ii) the architectural patterns used to compose these capabilities into functioning agents and multi-agent systems; (iii) the domains in which agentic AI has been applied, including healthcare, finance, software engineering, cybersecurity, and robotics; and (iv) the technical, ethical, and security challenges that recur across this literature, together with the research directions proposed to address them.

The specific objectives of this paper are to: (1) trace the conceptual evolution of agentic AI from earlier rule-based and reinforcement-learning agents to contemporary LLM-based systems; (2) synthesize and organize the literature on reasoning, tool use, memory, and multi-agent coordination into a coherent taxonomy; (3) review representative evaluation benchmarks and application domains reported in the literature; (4) consolidate the challenges associated with agentic AI, spanning reliability, cost, privacy, bias, and security; and (5) identify priority directions for future research.

The remainder of this paper is organized as follows. Section 2 presents background concepts and a taxonomy of agentic AI. Section 3 reviews the literature thematically and presents a comparative summary table of representative studies. Section 4 discusses architectural and methodological patterns used to build agentic systems. Section 5 surveys application domains. Section 6 reviews evaluation benchmarks. Section 7 discusses challenges, Section 8 outlines future research directions, and Section 9 concludes the paper.

2. Background and Taxonomy of Agentic AI

The notion of an intelligent agent, an entity that perceives its environment through sensors and acts upon it through actuators to achieve goals, has a long history in artificial intelligence, with formal treatments dating to early work on agent theory and practice in distributed systems research. Classical agents were typically rule-based or relied on hand-crafted planning algorithms, followed by a generation of reinforcement-learning agents that learned policies through trial-and-error interaction with simulated environments.

The emergence of large pre-trained language models such as GPT-3 marked a turning point, as these models demonstrated strong few-shot reasoning and instruction-following ability without task-specific training. Prompting techniques such as Chain-of-Thought prompting showed that eliciting intermediate reasoning steps substantially improved performance on multi-step reasoning tasks, and follow-up work such as Self-Consistency and Tree-of-Thoughts extended this idea to structured search over multiple reasoning paths. These reasoning-oriented advances laid the groundwork for treating LLMs not merely as text generators but as controllers capable of directing multi-step problem-solving processes.

Building on this foundation, the ReAct framework proposed interleaving reasoning traces with discrete actions and environment observations, allowing a language model to update its plan based on real-time feedback from external tools or environments. This combination of reasoning and acting is widely regarded as a foundational step toward what is now termed agentic AI. Shortly afterward, complementary capabilities were introduced: Toolformer and related tool-learning approaches such as ToolLLM and Gorilla enabled models to autonomously decide when and how to invoke external APIs; Reflexion introduced verbal self-reflection as a substitute for gradient-based

reinforcement learning, allowing agents to learn from prior failures within the context window; and memory-augmented architectures such as MemGPT proposed operating-system-inspired memory hierarchies to extend an agent's effective context beyond a single prompt window.

Figure 1 summarizes this trajectory at a coarse granularity, situating the emergence of LLM-based reasoning and tool-use agents, and subsequently multi-agent systems, within the broader history of agent research in artificial intelligence.

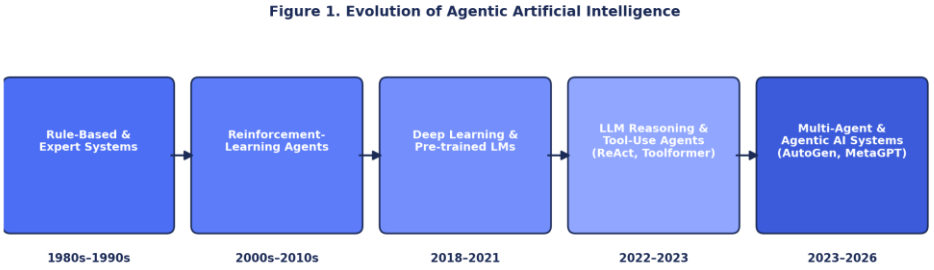

Figure 1. Evolution of Agentic Artificial Intelligence.

2.1 Defining Agentic AI

While definitions vary across the literature, a recurring characterization treats an agentic AI system as one that autonomously decomposes a high-level goal into sub-tasks, selects and executes actions (including invoking external tools), maintains state across steps, and adapts its plan based on feedback, with a degree of autonomy and persistence that distinguishes it from single-turn generative use of an LLM.

2.2 Core Capability Taxonomy

Building on this history, the literature converges on five interrelated capabilities that characterize a contemporary agentic AI system, summarized in Figure 5 and elaborated in Section 3: reasoning and planning, tool use and API grounding, memory management, reflection and self-correction, and multi-agent collaboration. An agentic system typically combines several of these capabilities within an orchestration loop in which the language model core repeatedly perceives context, reasons about the next step, optionally invokes a tool or consults memory, and updates its plan based on the resulting observation, as illustrated conceptually in Figure 2.

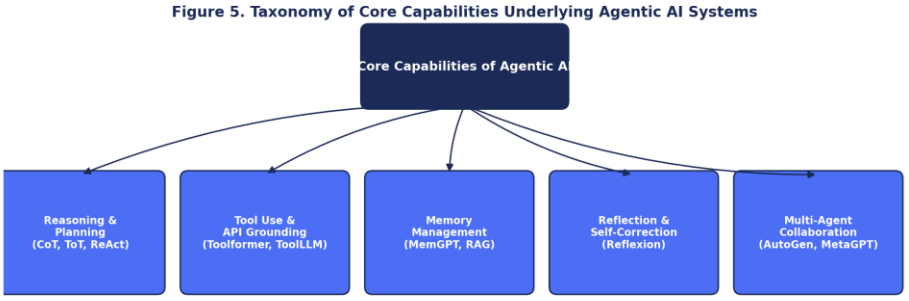

Figure 5. Taxonomy of Core Capabilities Underlying Agentic AI Systems.

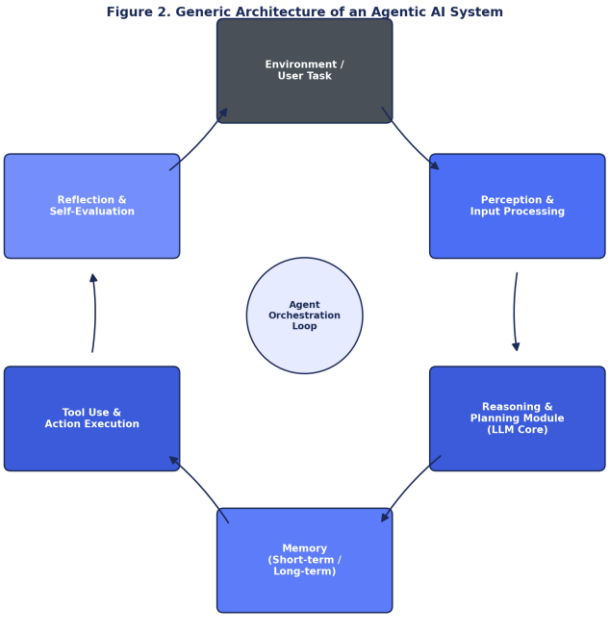

Figure 2. Generic Architecture of an Agentic AI System.

3. Literature Review

This section synthesizes the literature thematically across five capability areas, followed by a review of multi-agent systems research and cross-cutting survey work. Table 1 consolidates representative studies, their methodological focus, and reported contributions and limitations.

3.1 Reasoning and Planning

Reasoning and planning form the cognitive core of agentic systems. Chain-of-Thought prompting demonstrated that language models produce more accurate answers on arithmetic, commonsense, and symbolic reasoning tasks when prompted to generate intermediate reasoning steps rather than direct answers. Self-Consistency extended this by sampling multiple reasoning paths and selecting the majority answer, improving robustness to individual reasoning errors. Tree-of-Thoughts further generalized Chain-of-Thought into a deliberate search process over a tree of intermediate reasoning states, enabling backtracking and lookahead for problems that a single linear reasoning chain solves poorly. ReAct unified reasoning with acting by interleaving thought generation, action selection, and observation of environment feedback within a single prompting loop, and was shown to outperform reasoning-only and acting-only baselines on interactive decision-making benchmarks such as ALFWorld and WebShop, while also improving the interpretability of agent behavior.

3.2 Tool Use and API Grounding

A second thread of research addresses grounding language models in external tools and structured knowledge. Toolformer showed that a language model could be trained to decide autonomously when to call external APIs, such as calculators or search engines, and how to incorporate the results into its output. Subsequent work scaled this idea considerably: ToolLLM introduced a framework for enabling models to use over sixteen thousand real-world APIs, while Gorilla focused on accurately generating API calls by retrieving relevant API documentation at inference time. HuggingGPT proposed using an LLM as a central controller that delegates subtasks to specialized models hosted on a model-sharing platform, illustrating an early pattern of tool orchestration across heterogeneous AI systems. Collectively, this body of work establishes tool use as a mechanism for extending an LLM's effective capability beyond its parametric knowledge and computation.

3.3 Memory Management

Because LLMs operate within a bounded context window, sustaining coherent behavior over long horizons requires explicit memory mechanisms. MemGPT proposed an operating-system-inspired architecture in which the language model manages a hierarchy of memory tiers, paging information in and out of the active context analogous to virtual memory management, allowing an agent to maintain persistent state across extended interactions. Related work on cognitive architectures for language agents draws on cognitive science to propose structured working and long-term

memory components, procedural knowledge, and action selection mechanisms as organizing principles for agent design. Generative Agents demonstrated an applied instance of memory-augmented agency, in which simulated characters in an interactive sandbox environment retrieved and synthesized past experiences from a memory stream to produce believable, contextually consistent behavior over simulated days.

3.4 Reflection and Self-Correction

Reflection mechanisms allow agents to improve performance without updating model weights. Reflexion introduced a framework in which an agent verbally reflects on the outcome of a failed attempt, storing the resulting self-critique in an episodic memory buffer that is then used to inform subsequent attempts at the same or similar task, reporting substantial gains on decision-making, reasoning, and coding benchmarks relative to non-reflective baselines. This line of work positions natural-language self-critique as a lightweight substitute for gradient-based reinforcement learning, particularly relevant given the cost of fine-tuning large models.

3.5 Multi-Agent Collaboration

As single-agent capabilities matured, research attention shifted toward multi-agent systems in which multiple LLM-based agents, often with distinct roles, communicate to solve tasks collaboratively. AutoGen proposed a general-purpose framework for building applications through configurable, conversable agents that can be composed into flexible multi-agent workflows. MetaGPT introduced standardized operating procedures inspired by software engineering practice, assigning agents specialized roles such as product manager, architect, and engineer to structure collaborative software development. Voyager demonstrated an open-ended embodied agent operating within the Minecraft environment, using a growing skill library and iterative curriculum to acquire new capabilities autonomously. Comprehensive surveys of this space, including the IJCAI 2024 survey on LLM-based multi-agent systems and related work cataloguing communication protocols, coordination strategies, and application domains, report that LLM-based multi-agent systems have rapidly expanded from simple debate and simulation settings toward complex, tool-integrated collaborative pipelines, while also noting open challenges in coordination overhead, role specialization, and emergent misbehavior.

3.6 Broader Surveys and Field Positioning

Several broader surveys contextualize this literature. Xi et al. trace the rise of LLM-based agents across reasoning, planning, and application dimensions; a Frontiers of Computer Science survey by Wang et al. organizes the literature around agent construction, application, and evaluation; and a 2025 Artificial Intelligence Review survey explicitly frames agentic AI as a distinct paradigm from earlier LLM-agent conceptions, arguing that many existing surveys retrofit agentic AI onto architectures that do not share its degree of autonomy and persistence. Domain-specific reviews have also emerged, including surveys on agentic AI for Industry 4.0 distributed automation, radiology decision support, and financial services, indicating that the field is simultaneously deepening theoretically and broadening across application verticals.

Table 1. Comparative Summary of Representative Studies in Agentic AI

Author(s)	Year	Method / Focus	Environment / Dataset	Key Findings	Limitations
Yao et al. [1]	2023	Interleaved reasoning-and-acting prompting (ReAct)	ALFWorld, WebShop, HotpotQA	Improved success rate and interpretability over reasoning-only or acting-only baselines	Relies on carefully designed few-shot exemplars; limited to text-based environments
Shinn et al. [2]	2023	Verbal self-reflection with episodic memory (Reflexion)	AlfWorld, HotPotQA, HumanEval	Learns from failure without weight updates; gains on decision-making and coding tasks	Reflection quality bounded by base model's self-assessment ability
Schick et al. [4]	2023	Self-supervised tool-use training (Toolformer)	Multiple QA and reasoning datasets	Model learns when and how to call external APIs autonomously	Limited to tools seen during self-supervised data generation
Wu et al. [5]	2024	Conversable multi-agent framework (AutoGen)	Coding, QA, decision-making tasks	General-purpose, composable multi-agent orchestration	Coordination overhead grows with agent count and role complexity
Hong et al. [6]	2024	Role-based multi-agent software workflow (MetaGPT)	Software engineering benchmarks	Standardized operating procedures reduce cascading hallucination in collaboration	Domain-specific to structured software development tasks
Wang et al. [7]	2023	Open-ended embodied agent with skill library (Voyager)	Minecraft environment	Continual skill acquisition without human intervention	Embodied simulation setting; limited transfer evidence to real-world robotics
Liu et al. [8]	2024	Multi-environment LLM agent benchmark (AgentBench)	8 environments, 29 LLMs	First systematic cross-environment agent evaluation	Benchmark saturates quickly as newer LLMs improve
Guo et al. [9]	2024	Survey of LLM-based multi-	Literature synthesis	Consolidates domains,	Survey scope precedes rapid

Author(s)	Year	Method / Focus	Environment / Dataset	Key Findings	Limitations
		agent systems		communication, and capability acquisition in LLM-MA systems	2024-2026 protocol developments (e.g., MCP, A2A)
Qin et al. [17]	2023	Large-scale tool-use benchmark (ToolLLM)	16,000+ real-world APIs	Demonstrates scalable tool-use training beyond narrow API sets	API coverage and realism dependent on public API documentation quality
Packer et al. [15]	2024	OS-inspired hierarchical memory (MemGPT)	Conversational and document tasks	Extends effective context via memory paging analogy	Added architectural complexity and latency from memory management overhead
Park et al. [3]	2023	Memory-stream-based generative agents	Simulated sandbox society	Believable long-horizon agent behavior from retrieval-augmented memory	Evaluated in simulated, not real-world, decision-making settings
Chhabra et al. [24]	2025	Taxonomy of agentic AI security threats and defenses	Literature and benchmark synthesis	Consolidated threat taxonomy distinct from traditional LLM safety	Field evolving rapidly; taxonomy requires continuous revision

4. Architectures and Methodological Patterns

Because this paper is a literature review rather than an empirical study, its methodology concerns literature identification, thematic synthesis, and taxonomy construction rather than a predictive model pipeline. Studies were identified through targeted search of arXiv, IEEE Xplore, ACM Digital Library, and ScienceDirect using terms combining agentic AI, LLM agents, tool use, and multi-agent systems, with priority given to peer-reviewed venues (ICLR, NeurIPS, ICML, IJCAI, ACL family) and widely cited preprints that introduced foundational frameworks later adopted across the field. Retrieved studies were organized according to the five-capability taxonomy in Section 2.2, compared along methodological focus, evaluation environment, contribution, and limitation to form Table 1, while cross-cutting themes such as benchmarks, applications, and safety were synthesized separately in Sections 5 through 7.

Figure 2 formalizes the generic architectural pattern recurring across the reviewed single-agent frameworks: an orchestration loop in which the language model core perceives context, consults memory, reasons about the next step, optionally invokes a tool, and reflects on the outcome before proceeding. Figure 3 extends this pattern to the multi-agent case, in which an orchestrator or planner agent decomposes a task among specialized worker agents that communicate through a shared memory and tool layer, consistent with the designs reported in AutoGen and MetaGPT.

Figure 3. Representative Multi-Agent Collaboration Architecture

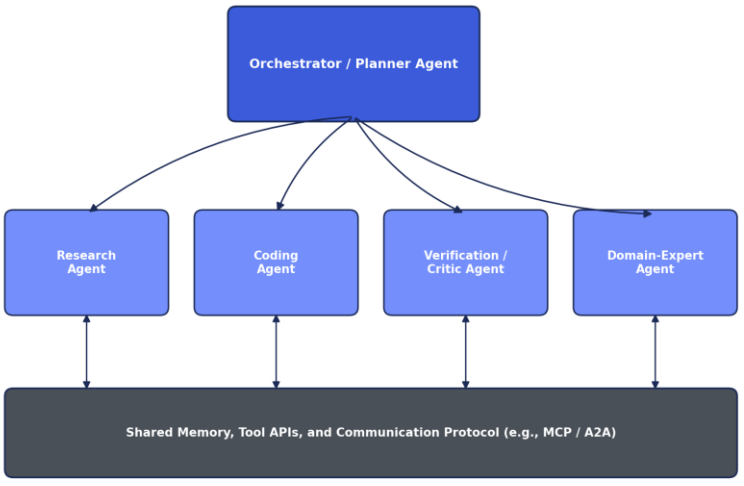

Figure 3. Representative Multi-Agent Collaboration Architecture.

5. Application Domains

The capabilities reviewed in Section 3 have been applied across a broad range of domains, summarized in Table 2 and Figure 4. Reported applications range from consumer-facing conversational automation to safety-critical decision support, with the depth of real-world validation varying considerably across domains.

Figure 4. Principal Application Domains of Agentic AI Systems.

Table 2. Application Domains of Agentic AI Reported in the Literature

Domain	Representative Use Cases	Reported Architectural Approach
Healthcare	Clinical decision support, radiology triage assistance, patient monitoring and treatment-protocol drafting	Multi-agent role specialization for cross-validation; retrieval-augmented grounding to reduce hallucination [8][30]
Finance	Autonomous trading support, compliance and regulatory monitoring, financial analysis and reporting	Multi-agent frameworks simulating specialized analyst roles; tool-grounded retrieval of market data [29]
Software Engineering	Automated debugging, code generation, issue resolution, test generation	Role-based multi-agent pipelines (e.g., product manager, architect, engineer) with iterative self-verification [6][20]
Cybersecurity	Threat detection, automated incident response, security-workflow orchestration	Defensive and offensive agent taxonomies; benchmark-driven evaluation of agent-executed exploits [37]
Customer Service & Enterprise Workflows	Conversational support automation, workflow orchestration across enterprise software	Tool-integrated conversational agents connected to enterprise APIs and databases
Robotics & Embodied AI	Open-ended skill acquisition, task planning for physical or simulated agents	Curriculum-driven skill libraries combined with language-grounded planning [7]

Across domains, a common pattern reported in the literature is the use of multi-agent role specialization, for example separating a proposing agent from a verifying or critic agent, to mitigate the reliability limitations discussed in Section 7, combined with retrieval-augmented grounding to reduce reliance on parametric knowledge alone.

6. Evaluation Benchmarks and Metrics

Evaluating agentic AI systems is more complex than evaluating static language generation, because agent behavior unfolds over multiple steps and interacts with external state. AgentBench addresses this by evaluating language models across eight distinct environments spanning code, game, and web-grounded tasks, reporting substantial performance gaps between proprietary and open-source models on multi-step agentic tasks even when both perform comparably on standard language benchmarks. SWE-bench evaluates agents on their ability to resolve authentic GitHub issues by generating patches that pass held-out test suites. Earlier interactive benchmarks such as ALFWorld and WebShop remain widely used for grounded decision-making in text-based embodied and web-navigation settings respectively, while multi-hop question-answering datasets such as HotpotQA continue to serve reasoning-oriented agents. Table 3 summarizes representative benchmarks; a recurring observation is that agent performance depends jointly on the base model's reasoning capability and on surrounding scaffolding, including prompt design, tool interfaces, and memory management, making cross-study comparisons sensitive to implementation details.

Table 3. Representative Evaluation Benchmarks for Agentic AI Systems

Benchmark	Year	Scope / Environment	Primary Evaluation Focus
AgentBench [8]	2023/2024	Multi-environment (code, game, web)	Cross-environment task completion via CoT-prompted evaluation of 29 LLMs
SWE-bench [20]	2024	Real-world GitHub issue resolution	Whether an agent-generated patch resolves an authentic

Benchmark	Year	Scope / Environment	Primary Evaluation Focus
			software issue
ALFWorld	2021	Text-based embodied household tasks	Success rate on multi-step household instruction-following
WebShop	2022	Simulated e-commerce web navigation	Task completion and product-matching accuracy
HotpotQA	2018	Multi-hop question answering	Answer accuracy requiring multi-document reasoning
ToolBench / ToolLLM [17]	2023	Large-scale API tool invocation	Correctness and executability of generated API call sequences

7. Challenges

Across the reviewed literature, a consistent set of technical, ethical, and operational challenges recurs regardless of application domain. Table 4 consolidates these challenges; the following discussion elaborates on the most frequently reported categories.

Table 4. Technical, Ethical, and Security Challenges in Agentic AI

Challenge Category	Description	Representative Source(s)
Reliability & Hallucination	Agents may generate plausible but factually incorrect reasoning or actions, particularly consequential in multi-step tool-use chains	[8], radiology review
Autonomy & Unbounded Action	Increased operational autonomy raises the risk of unintended or harmful actions executed without adequate human oversight	[38], AURA framework
Privacy & Data Exposure	Agents interacting with external tools, browsers, and enterprise systems may inadvertently expose sensitive data	Agentic AI security survey [24]
Security Vulnerabilities	Prompt injection, memory poisoning, and tool-hijacking attacks exploit the expanded attack surface of agentic systems	[24], layered attack-surface survey
Computational Cost	Multi-step reasoning, reflection loops, and multi-agent communication substantially increase inference cost relative to single-turn prompting	General multi-agent literature [9]
Coordination Complexity	Multi-agent systems face overhead from role negotiation, message-passing, and emergent miscoordination as agent count grows	[9], LLM multi-agent challenges survey
Governance & Accountability	Regulatory and organizational governance frameworks lag behind the pace of agentic AI deployment	State of AI security and governance reports

Reliability remains the most cited limitation of agentic systems. Because agents chain multiple reasoning and action steps, an early error can propagate and compound across subsequent steps, a failure mode particularly consequential in high-stakes domains such as clinical decision support, where reviewed literature emphasizes that multi-agent cross-validation and retrieval grounding reduce but do not eliminate hallucination risk. Computational cost is a related concern: reflection loops, multi-agent communication, and long reasoning chains multiply inference cost relative to single-turn prompting, raising questions about the economic viability of agentic deployment at scale.

The expanded autonomy that defines agentic AI also expands its attack surface and risk profile relative to conventional LLM deployments. Recent security-focused surveys document threat categories specific to agentic systems, including prompt injection that hijacks an agent's tool-use decisions, memory poisoning that corrupts an agent's persistent state, and cascading failures arising from multi-agent miscoordination. Documented incidents cited in this literature, including unauthorized data exposure by browser-integrated agents and unintended destructive actions by coding agents, illustrate that these risks are not purely theoretical. Governance research further notes that organizational oversight mechanisms and regulatory frameworks have not kept pace with the speed of agentic AI adoption, with several industry surveys reporting that a majority of deployed agents lack adequate monitoring coverage.

Ethical considerations extend beyond security to questions of accountability, bias propagation, and appropriate human oversight. Because agentic systems can execute multi-step plans with limited human-in-the-loop checkpoints, responsibility for downstream harms becomes more difficult to attribute than in single-turn generative use, a concern raised explicitly in recent work on agent autonomy risk assessment and on responsibility allocation in multi-agent systems.

8. Future Research Directions

Building on the challenges identified above, the reviewed literature converges on several priority directions for future research, summarized in Figure 6 and Table 5.

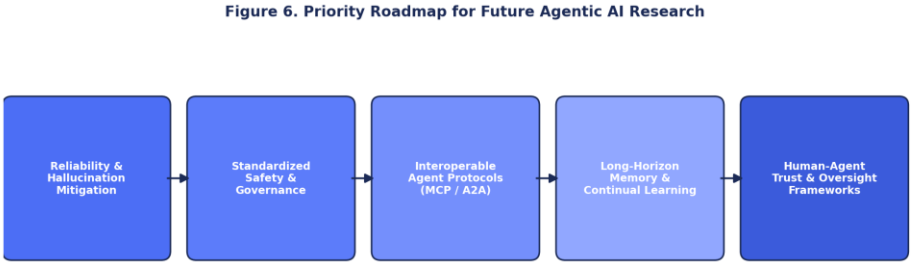

Figure 6. Priority Roadmap for Future Agentic AI Research.

Table 5. Priority Directions for Future Agentic AI Research

Direction	Description
Standardized Evaluation	Development of benchmarks that jointly assess reasoning quality, tool-use correctness, and safety rather than task completion alone
Interoperable Communication Protocols	Maturation of open standards such as Model Context Protocol and Agent-to-Agent protocol for cross-vendor agent interoperability
Long-Horizon Memory	More efficient and reliable memory architectures capable of sustaining coherent agent behavior across extended, multi-session tasks
Human-Agent Trust Frameworks	Mechanisms for calibrated human oversight, intervention, and explainability proportional to agent autonomy level
Robust Security-by-Design	Defenses against prompt injection, memory poisoning, and tool-hijacking integrated into agent architectures rather than added post hoc
Domain-Specific Validation	Rigorous clinical, financial, and safety-critical validation studies to move agentic AI from pilots to accountable production deployment

First, the field requires more standardized and holistic evaluation methodology that jointly measures task success, reasoning faithfulness, tool-use correctness, and safety behavior, rather than treating these as separate evaluation tracks. Second, the emergence of interoperability protocols such as the Model Context Protocol and Agent-to-Agent communication standards suggests a trajectory toward vendor-agnostic agent ecosystems, though their security and governance implications remain under active investigation. Third, long-horizon memory remains an open architectural problem: while approaches such as MemGPT demonstrate feasibility, efficient, reliable, and privacy-preserving memory management across extended multi-session deployments is not yet solved. Fourth, human-agent trust frameworks that calibrate the degree of autonomy granted to an agent against the stakes of its task, with proportional human oversight and intervention mechanisms, are identified across multiple governance-oriented papers as a prerequisite for responsible scaling. Finally, domain-specific validation, particularly in healthcare and finance, is repeatedly flagged as necessary before agentic AI can move from promising pilots to accountable, auditable production systems.

9. Conclusion

This review has synthesized the rapidly expanding literature on agentic AI, organizing contributions around a five-part taxonomy of reasoning and planning, tool use, memory, reflection, and multi-agent collaboration. The synthesis shows that agentic AI has progressed from isolated reasoning and tool-use techniques, such as Chain-of-Thought prompting and Toolformer, toward integrated single-agent architectures exemplified by ReAct and Reflexion, and subsequently toward coordinated multi-agent systems such as AutoGen and MetaGPT that decompose complex tasks across specialized roles.

Despite substantial technical progress, the literature consistently identifies unresolved challenges in reliability, computational cost, security, and governance that must be addressed before agentic AI can be deployed safely at scale in high-stakes domains. The consolidated taxonomy, comparative literature summary, and challenge and future-direction tables presented in this paper are intended to provide researchers, practitioners, and policymakers with a coherent orientation to a field that, despite its recency, already spans foundational reasoning research, systems engineering, domain-specific application studies, and AI safety and governance scholarship. Continued progress will depend on evaluation standards, interoperability protocols, and governance mechanisms that mature at a pace commensurate with the technical capabilities of these systems.

References

[1] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "ReAct: Synergizing reasoning and acting in language models," in Proc. 11th Int. Conf. Learning Representations (ICLR), Kigali, Rwanda, 2023.

[2] N. Shinn, F. Cassano, E. Berman, 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.

[3] 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," arXiv preprint arXiv:2304.03442, 2023.

[4] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, and T. Scialom, "Toolformer: Language models can teach themselves to use tools," arXiv preprint arXiv:2302.04761, 2023.

[5] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, et al., "AutoGen: Enabling next-gen LLM applications via multi-agent conversation," in Proc. Conf. Language Modeling (COLM), 2024.

[6] S. Hong, M. Zhuge, J. Chen, X. Zheng, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin, et al., "MetaGPT: Meta programming for a multi-agent collaborative framework," in Proc. 12th Int. Conf. Learning Representations (ICLR), 2024.

[7] G. Wang, Y. Xie, Y. Jiang, A. Mandlekar, C. Xiao, Y. Zhu, L. Fan, and A. Anandkumar, "Voyager: An open-ended embodied agent with large language models," in Advances in Neural Information Processing Systems (NeurIPS), vol. 36, 2023.

[8] X. Liu, H. Yu, H. Zhang, Y. Xu, X. Lei, H. Lai, Y. Gu, H. Ding, K. Men, K. Yang, et al., "AgentBench: Evaluating LLMs as agents," in Proc. 12th Int. Conf. Learning Representations (ICLR), 2024.

[9] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, and X. Zhang, "Large language model based multi-agents: A survey of progress and challenges," in Proc. 33rd Int. Joint Conf. Artificial Intelligence (IJCAI), 2024, pp. 8048-8057.

[10] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, et al., "The rise and potential of large language model based agents: A survey," arXiv preprint arXiv:2309.07864, 2023.

[11] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, et al., "A survey on large language model based autonomous agents," Frontiers of Computer Science, vol. 18, no. 6, art. 186345, 2024.

[12] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in Advances in Neural Information Processing Systems (NeurIPS), vol. 35, 2022, pp. 24824-24837.

[13] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," arXiv preprint arXiv:2203.11171, 2022.

[14] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," in Advances in Neural Information Processing Systems (NeurIPS), vol. 36, 2023.

[15] C. Packer, S. Wooders, K. Lin, V. Fang, S. G. Patil, I. Stoica, and J. E. Gonzalez, "MemGPT: Towards LLMs as operating systems," arXiv preprint arXiv:2310.08560, 2024.

[16] T. R. Sumers, S. Yao, K. Narasimhan, and T. L. Griffiths, "Cognitive architectures for language agents," arXiv preprint arXiv:2309.02427, 2024.

[17] Y. Qin, S. Liang, Y. Ye, K. Zhu, L. Yan, Y. Lu, Y. Lin, X. Cong, X. Tang, B. Qian, et al., "ToolLLM: Facilitating large language models to master 16000+ real-world APIs," arXiv preprint arXiv:2307.16789, 2023.

[18] S. G. Patil, T. Zhang, X. Wang, and J. E. Gonzalez, "Gorilla: Large language model connected with massive APIs," arXiv preprint arXiv:2305.15334, 2023.

[19] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, "HuggingGPT: Solving AI tasks with ChatGPT and its friends in Hugging Face," in Advances in Neural Information Processing Systems (NeurIPS), vol. 36, 2023.

[20] C. E. Jimenez, J. Yang, A. Wettig, S. Yao, K. Pei, O. Press, and K. R. Narasimhan, "SWE-bench: Can language models resolve real-world GitHub issues?" in Proc. 12th Int. Conf. Learning Representations (ICLR), 2024.

[21] M. Shridhar, X. Yuan, M.-A. Côté, Y. Bisk, A. Trischler, and M. J. Hausknecht, "ALFWorld: Aligning text and embodied environments for interactive learning," in Proc. 9th Int. Conf. Learning Representations (ICLR), 2021.

[22] S. Yao, H. Chen, J. Yang, and K. Narasimhan, "WebShop: Towards scalable real-world web interaction with grounded language agents," in Advances in Neural Information Processing Systems (NeurIPS), vol. 35, 2022.

[23] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning, "HotpotQA: A dataset for diverse, explainable multi-hop question answering," in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), 2018, pp. 2369-2380.

[24] A. Chhabra, S. Datta, S. K. Nahin, and P. Mohapatra, "Agentic AI security: Threats, defenses, evaluation, and open challenges," arXiv preprint arXiv:2510.23883, 2025.

[25] A. Acharya, D. Kuppan, and S. Divya, "Agentic AI: A comprehensive survey of architectures, applications, and future directions," Artificial Intelligence Review, Springer, 2025.

[26] Y. Cheng, C. Zhang, Z. Zhang, X. Meng, S. Hong, W. Li, Z. Wang, Z. Wang, F. Yin, J. Zhao, et al., "Exploring large language model based intelligent agents: Definitions, methods, and prospects," arXiv preprint arXiv:2401.03428, 2024.

[27] S. Han, Q. Zhang, Y. Yao, W. Jin, Z. Xu, and C. He, "LLM multi-agent systems: Challenges and open problems," arXiv preprint arXiv:2402.03578, 2024.

[28] M. Wooldridge and N. R. Jennings, "Intelligent agents: Theory and practice," The Knowledge Engineering Review, vol. 10, no. 2, pp. 115-152, 1995.

[29] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., "Language models are few-shot learners," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 1877-1901.

[30] "AI agents in finance and fintech: A scientific review of agent-based systems, applications, and future horizons," ScienceDirect, 2025.

[31] "Agentic AI governance and lifecycle management in healthcare," arXiv preprint arXiv:2601.15630, 2026.

[32] "AURA: An agent autonomy risk assessment framework," arXiv preprint arXiv:2510.15739, 2025.

[33] "Agentic AI and large language models in radiology: Opportunities and hallucination challenges," PMC, National Library of Medicine, 2025.

[34] "AgentAI: A comprehensive survey on autonomous agents in distributed AI for industry 4.0," ScienceDirect, 2025.

[35] "From multi-agent systems and the semantic web to agentic AI: A unified narrative of the web of agents," arXiv preprint arXiv:2507.10644, 2025.

[36] "A systematic survey of security threats and defenses in LLM-based AI agents: A layered attack surface framework," arXiv preprint arXiv:2604.23338, 2026.

[37] "A survey of agentic AI and cybersecurity: Challenges, opportunities and use-case prototypes," arXiv preprint arXiv:2601.05293, 2026.

[38] R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders, et al., "WebGPT: Browser-assisted question-answering with human feedback," arXiv preprint arXiv:2112.09332, 2021.

[39] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, et al., "Do as I can, not as I say: Grounding language in robotic affordances," arXiv preprint arXiv:2204.01691, 2022.