JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer−Reviewed

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 37–59
DOI: 10.63646/jaiaa.2025.030103

The Age of Autonomous Agents: A Bibliometric Review of Agentic AI Architectures, Applications, and Emerging Challenges

ARTICLE INFO
Received
October 18, 2024
Revised
December 21, 2024
Accepted
February 12, 2025
Available Online
March 30, 2025
DOI
10.63646/jaiaa.2025.030103

Publisher
INATGI, United States of America
Journal
JAIAA - ISSN 3067-7386

Abstract
The rapid evolution of large language models (LLMs) has catalyzed a shift from passive AI systems toward autonomous agentic architectures capable of reasoning, memory, tool use, and multi-agent collaboration. This bibliometric review characterizes the emerging field through 810 publications retrieved from the Web of Science Core Collection for the period 2023–2025. Annual output rose sharply over this window—from 4 publications in 2023 to 96 in 2024 and 710 in 2025—accompanied by a parallel rise in citations, indicating rapid mainstream adoption. Author-keyword analysis reveals a landscape dominated by large language models, artificial intelligence, and multi-agent systems, with agentic AI, generative AI, and retrieval-augmented generation (RAG) emerging as core themes. Research output is geographically concentrated, led by China and the United States, and is distributed across a broad range of engineering, applied-science, and domain-specific journals rather than a single specialist venue, reflecting the field's cross-disciplinary uptake. Synthesizing this corpus, we organize the technical landscape around reasoning, memory, tool integration and RAG, and multi-agent orchestration; survey application domains spanning healthcare, scientific discovery, education, and software engineering, with emerging activity in finance and law; and analyze the principal challenges—hallucination, trust and robustness, inter-agent coordination, scalability, and governance. The review provides a structured, evidence-based map of agentic AI research to orient researchers and practitioners navigating this rapidly evolving field.

Keywords: Large Language Models, Agentic AI, Multi-Agent Systems, AI Agents, Generative AI, Retrieval-Augmented Generation, Bibliometric Analysis

1. Introduction

The artificial intelligence landscape has undergone a profound transformation since the emergence of transformer-based large language models. While early LLMs excelled at static text generation, the contemporary frontier has shifted toward agentic systems—autonomous entities that perceive environments, reason about goals, execute actions through tool use, and collaborate with other agents or humans. This transition from "model-centric" to "agent-centric" AI represents not merely an incremental improvement but a fundamental reconceptualization of how intelligent systems operate in complex, dynamic environments.

The concept of agency in artificial intelligence has roots extending to classical planning and reinforcement learning, yet the contemporary agentic AI paradigm is distinguished by the integration of LLM-based reasoning with modular tool ecosystems, persistent memory architectures, and multi-agent coordination protocols. Unlike traditional software agents that operated on rigid, pre-programmed rules, modern agentic systems leverage the

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 37–59
DOI: 10.63646/jaiaa.2025.030103

emergent reasoning capabilities of foundation models to handle novel situations, adapt strategies, and engage in open-ended problem-solving.

This review is organized as follows. Section 2 describes the methodology; Section 3 presents a bibliometric overview of the corpus; Section 4 examines the core architectures of agentic AI; Section 5 surveys the application domains; Section 6 analyzes the emerging challenges and risks; and Section 7 discusses future directions and concludes.

2. Methods

2.1 Search Strategy and Data Source

The literature was retrieved from the Web of Science Core Collection on 20 May 2026. A topic-field (TS) query was constructed to combine terms denoting agentic systems with terms denoting large language models and generative AI, so that retrieved records concerned LLM-driven agents specifically rather than the older and much broader literature on software or reinforcement-learning agents. The complete query was:

```
TS=(("agentic AI" OR "agentic artificial intelligence" OR "agentic workflow*" OR "agentic system*" OR "agentic framework*" OR "agentic reasoning") OR ("AI agent*" OR "LLM agent*" OR "LLM-based agent*" OR "language model agent*" OR "large language model agent*" OR "generative agent*" OR "autonomous agent*" OR "tool-using agent*" OR "multi-agent system*" OR "multi-agent framework*" OR "agent orchestration" OR "role-playing agent*") AND ("large language model*" OR LLM OR LLMs OR "foundation model*" OR "generative AI" OR "generative artificial intelligence" OR GPT* OR "transformer model*")))
```

2.2 Corpus Construction and Eligibility Criteria

The initial query returned 2,048 records. These were restricted to English-language publications (removing 5 records), to the publication years 2023–2025 (removing 1,163 records), and to the document types Article and Review (removing 70 records), yielding a final corpus of 810 publications comprising 733 articles and 77 reviews. Because all exclusions were applied through the database's built-in filters and no additional manual title/abstract or full-text screening was performed, the present work is best characterized as a bibliometric review rather than a systematic review; the selection process is summarized in Figure 1. Notably, 133 records carry a 2026 final publication date owing to early-access (online-first) indexing, even though the search was restricted to the publication years 2023–2025.

JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer−Reviewed

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 37–59
DOI: 10.63646/jaiaa.2025.030103

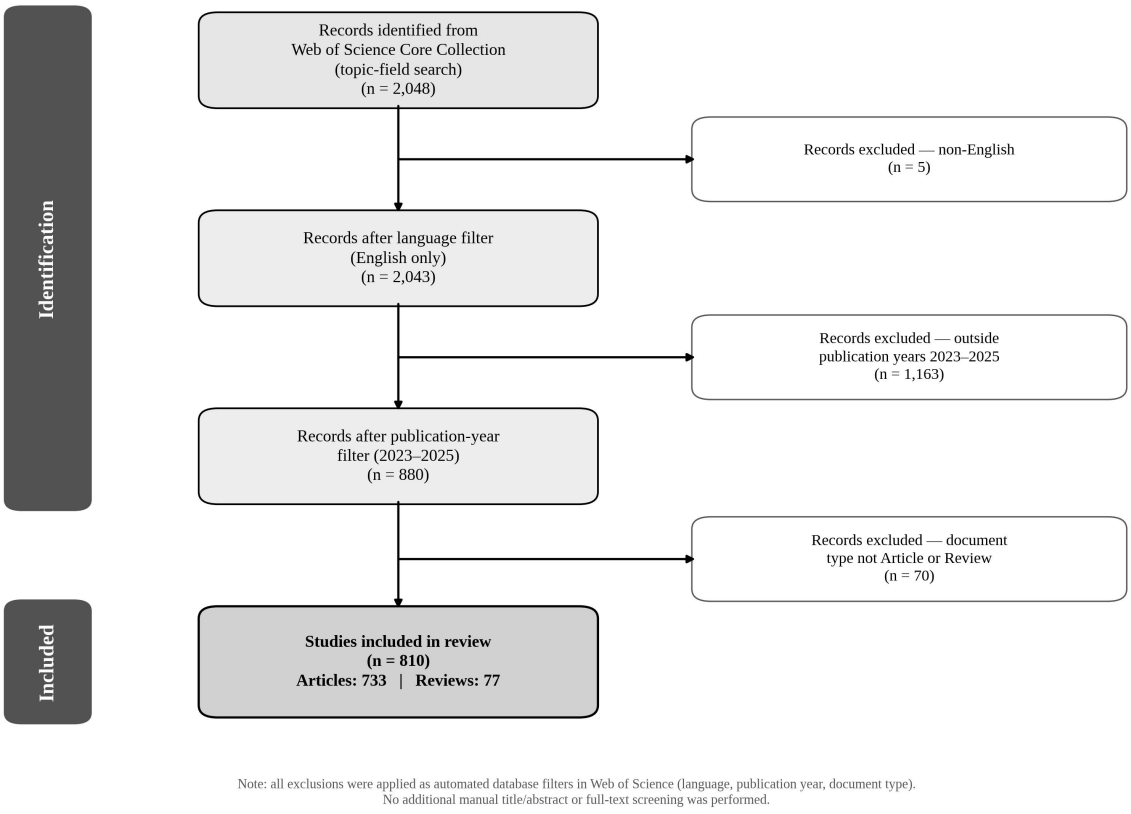

Figure 1. PRISMA flow diagram of the study identification and selection process.

2.3 Bibliometric Analysis

The exported records were analyzed along several dimensions using the Python pandas library. Annual publication counts and their corresponding Web of Science citation totals were tabulated to trace the field's growth (Figure 2). Author keywords were aggregated across all records, with spelling and abbreviation variants merged (for example, “large language model,” “large language models,” and “LLM” were treated as one term) and each keyword counted once per publication, to identify the most frequent research topics (Figure 3). Geographic output was derived from author affiliation addresses, counting each country once per publication and consolidating United States sub-national entries (Figure 4). The most productive source journals (Figure 5) and institutions (Figure 6) were ranked by publication count, and the most influential works were identified by Web of Science Core Collection citation count (Figure 7).

3. Bibliometric Overview

This section characterizes the 810-publication corpus along its temporal, thematic, geographic, and productivity dimensions, providing the empirical basis for the qualitative synthesis in the sections that follow.

Our analysis of 810 publications reveals a striking inflection point in research activity. Prior to 2023, agentic AI remained a niche concern within the broader AI community, with annual publications in the single digits. The release of ChatGPT in late 2022 and subsequent advances in GPT-4, Claude, and open-source alternatives triggered an explosive growth in research output. As illustrated in Figure 2, annual output rose from just 4 publications in 2023 to 96 in 2024 and then surged to 710 in 2025, while citations to this body of work grew in parallel from 2 to 3,702 over the same period. This trajectory reflects both the maturation of underlying LLM technologies and the growing recognition that standalone language models are insufficient for complex real-world tasks requiring interaction with external systems, persistent state management, and collaborative reasoning.

JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer−Reviewed

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 37–59
DOI: 10.63646/jaiaa.2025.030103

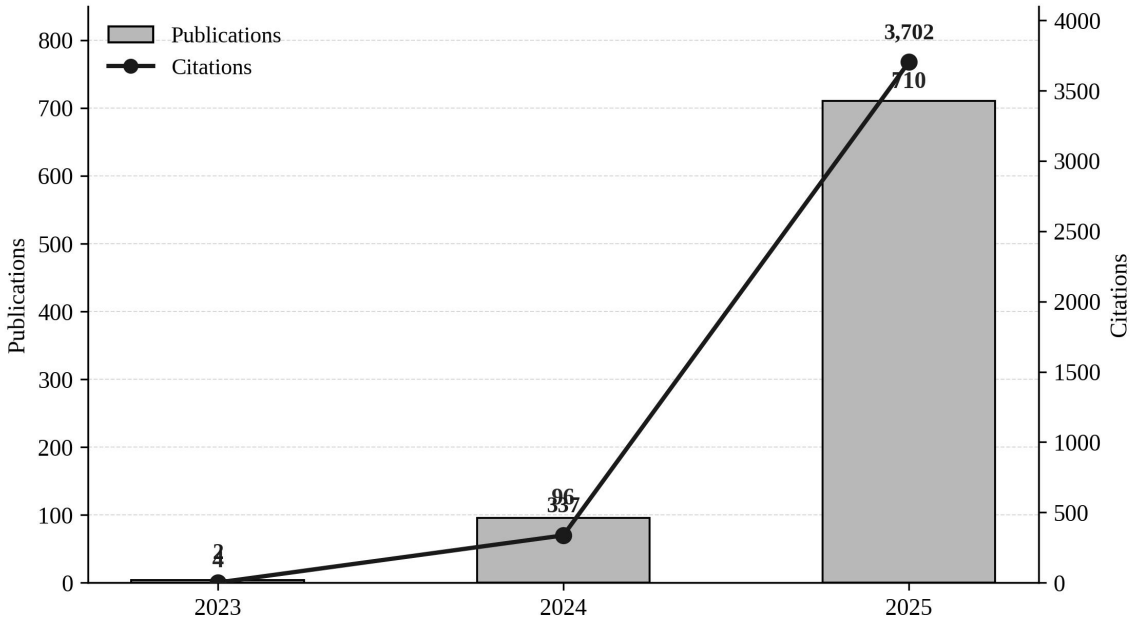

Figure 2. Annual Publications and Citations of the Included Corpus (2023–2025).

The steep trajectory depicted in Figure 2—an almost twenty-four-fold increase in annual output between 2023 and 2024, followed by a further sevenfold rise in 2025—signals the rapid mainstream adoption of agentic paradigms. The parallel surge in citations, exceeding 3,700 by 2025, indicates that this work is not merely accumulating but actively shaping subsequent research. This growth trajectory substantially outpaces the broader AI field, indicating that agentic AI has emerged as a distinct and vibrant research frontier rather than merely an application area within LLM research.

JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer−Reviewed

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 37–59
DOI: 10.63646/jaiaa.2025.030103

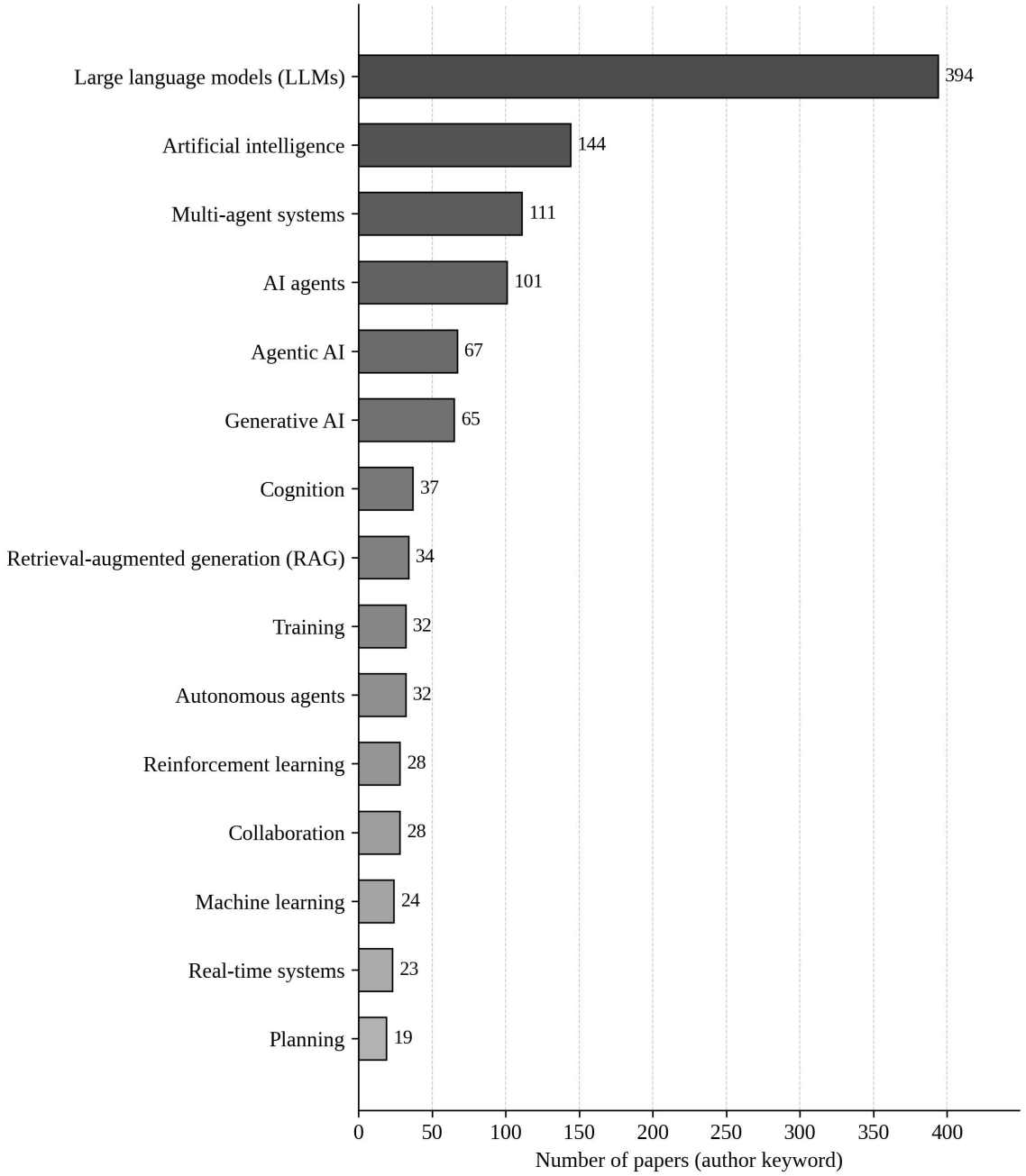

Figure 3. Most Frequent Author Keywords in the Corpus.

Author-keyword analysis confirms that large language models dominate the corpus, appearing in 394 of the 810 papers, followed by artificial intelligence (144) and multi-agent systems (111). The prominence of AI agents (101), agentic AI (67), and generative AI (65) reflects the field's consolidation around LLM-driven autonomous agents, while retrieval-augmented generation (34), reinforcement learning (28), and planning (19) recur as the core technical building blocks. Variant spellings of each term (e.g., “large language model,” “LLM”) were merged, and each keyword is counted once per paper.

JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer−Reviewed

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 37–59
DOI: 10.63646/jaiaa.2025.030103

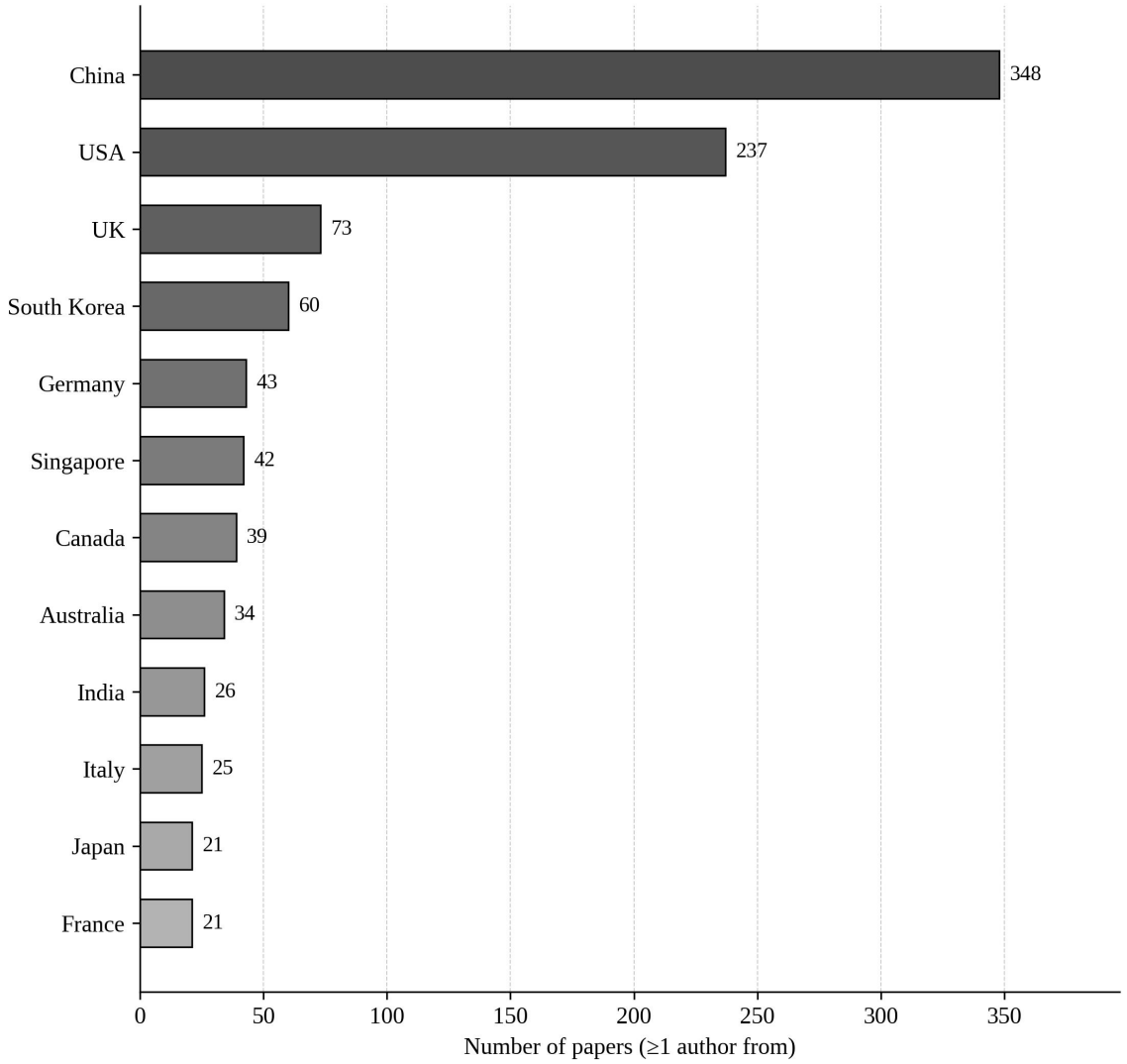

Figure 4. Leading Countries and Regions by Publication Output.

Research output is geographically concentrated. China leads by a wide margin, contributing to 348 of the 810 papers, followed by the United States (237); together they account for most of the corpus. The United Kingdom (73), South Korea (60), Germany (43), and Singapore (42) form a second tier, with Canada, Australia, India, Italy, France, and Japan also active. A paper is counted once for each contributing country.

JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer−Reviewed

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 37–59
DOI: 10.63646/jaiaa.2025.030103

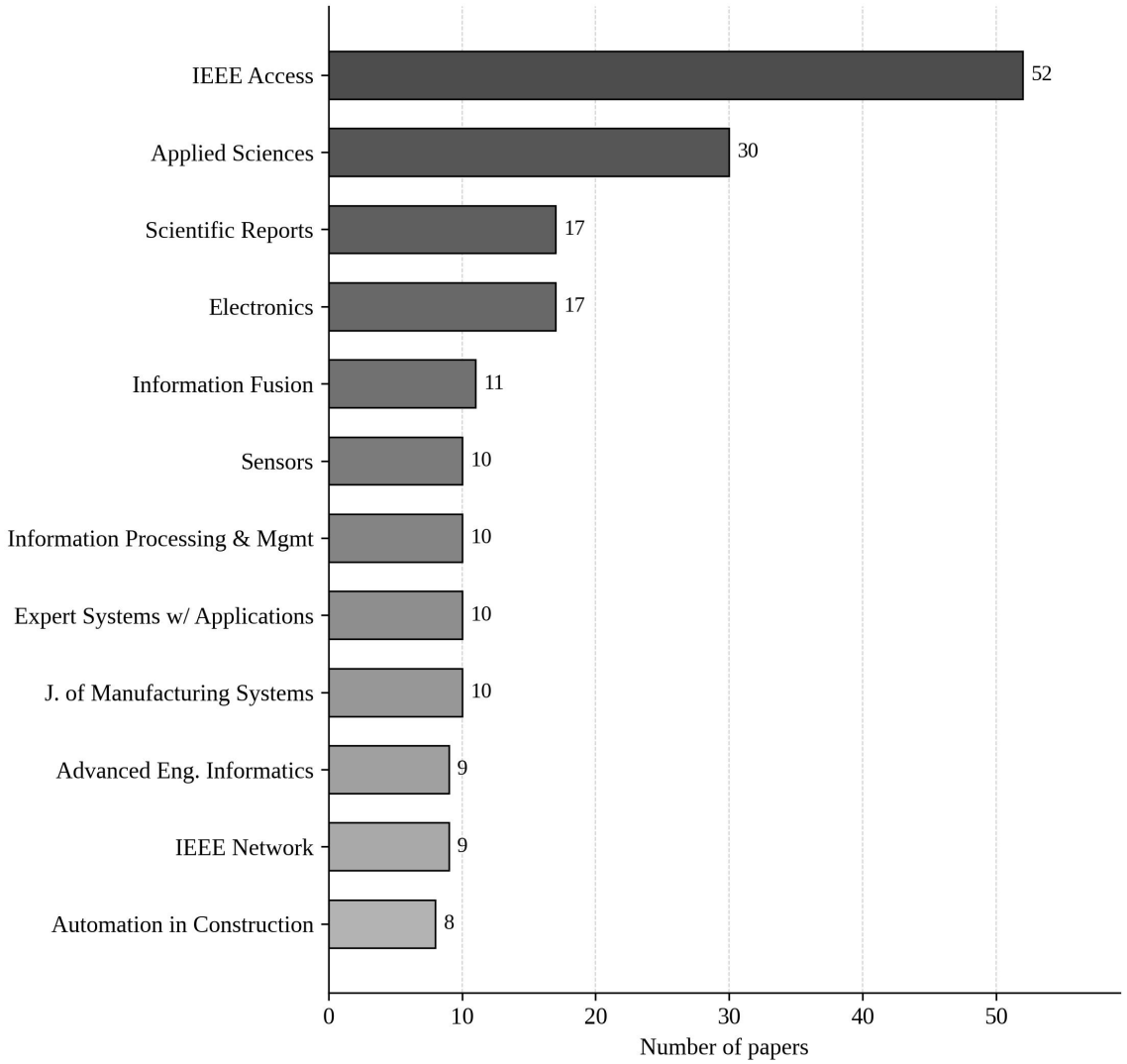

Figure 5. Most Productive Source Journals.

The corpus is spread across a wide range of venues rather than concentrated in a single specialist outlet. IEEE Access is the most frequent source (52 papers), followed by Applied Sciences (30) and Electronics and Scientific Reports (17 each). The mix of engineering, applied-science, medical, and information-systems journals—including Information Fusion, Expert Systems with Applications, the Journal of Manufacturing Systems, and npj Digital Medicine—reflects the cross-disciplinary uptake of agentic AI.

JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer−Reviewed

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 37–59
DOI: 10.63646/jaiaa.2025.030103

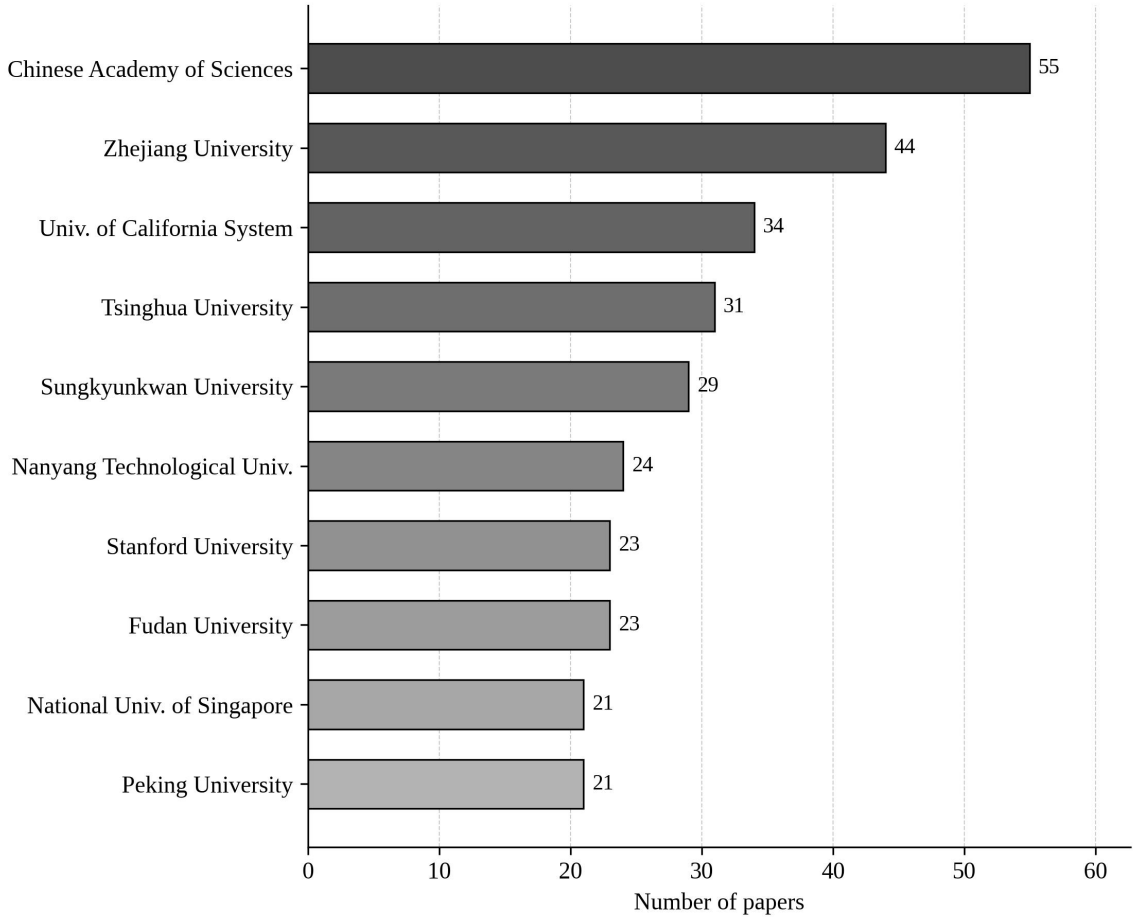

Figure 6. Most Productive Institutions.

Institutional output mirrors the geographic pattern. The Chinese Academy of Sciences (55 papers), Zhejiang University (44), and Tsinghua, Fudan, and Peking Universities lead the field, alongside the University of California System (34) and Stanford University (23) in the United States, Sungkyunkwan University in South Korea, and Nanyang Technological University and the National University of Singapore.

JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer−Reviewed

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 37–59
DOI: 10.63646/jaiaa.2025.030103

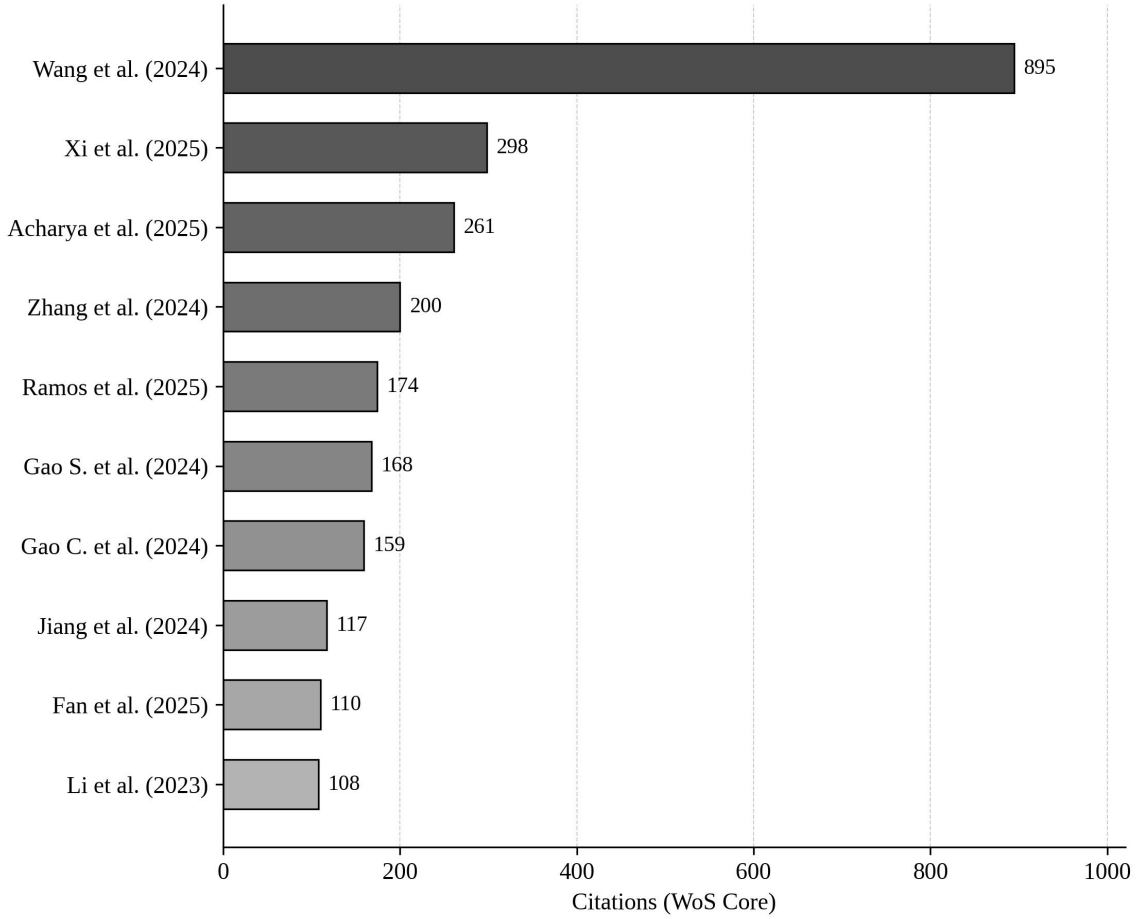

Figure 7. Ten Most-Cited Publications in the Corpus.

The single most-cited work is Wang et al. (2024), "A survey on large language model based autonomous agents" (895 citations), followed by Xi et al. (2025) on the rise of LLM-based agents (298) and Acharya et al. (2025) on agentic AI for complex goals (261). The remaining highly-cited papers span biomedicine (Gao S. et al., 2024, in Cell), chemistry (Ramos et al., 2025), satellite communications (Zhang et al., 2024), and manufacturing (Fan et al., 2025), showing that the field's most influential contributions are both broad surveys and domain-specific applications.

4. Core Architectures of Agentic AI

4.1 From Reasoning to Acting: The ReAct Paradigm

The integration of reasoning and action constitutes a foundational advance in agentic architecture. In the reasoning-and-acting paradigm popularized by the ReAct framework, agents interleave explicit reasoning traces with environmental interactions, grounding each decision in observable outcomes rather than generating reasoning in isolation as in chain-of-thought prompting. Across the reviewed corpus this reasoning–action loop recurs as the default control structure for LLM-based agents, and the major surveys treat it as a defining capability of the paradigm (Wang et al., 2024; Xi et al., 2025; Acharya et al., 2025).

Reasoning-driven agents appear across a wide range of domains in the corpus, including automated chemical synthesis and discovery (Ruan et al., 2024; Ramos et al., 2025), biomedical research (Gao, S. et al., 2024), geospatial analysis (Li et al., 2023), and software engineering (He et al., 2025). A common finding across these applications is that grounding

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 37–59
DOI: 10.63646/jaiaa.2025.030103

reasoning in tool feedback and intermediate observations improves reliability on multi-step tasks relative to reasoning without environmental feedback.

The paradigm also carries recognized limitations. Interleaved reasoning traces consume substantial context-window capacity, and the sequential nature of the reasoning–action loop constrains parallelization in multi-step tasks—trade-offs that recur in the reviewed applications and that motivate much of the memory and orchestration work discussed below.

4.2 Memory Architectures: From Ephemeral to Persistent

The transition from stateless inference to stateful operation requires memory architectures that persist information across an agent's reasoning steps and interactions. A dedicated survey of memory mechanisms in LLM-based agents (Zhang et al., 2025) organizes these into a multi-tiered hierarchy that recurs throughout the corpus: short-term memory held in the model's context window, working-memory extensions such as scratchpads and thought buffers, and long-term memory implemented through external stores.

Long-term memory is most commonly realized with vector databases that support retrieval by semantic similarity, supplemented in some systems by graph-structured stores that capture relational information between concepts (Zhang et al., 2025). These mechanisms allow agents to retain and recall knowledge across sessions rather than being confined to a single context window.

Memory-rich designs are especially prominent where agents must maintain coherent behavior over long horizons. Autonomous scientific agents that accumulate experimental context across many steps (Zou et al., 2025) and digital-twin and monitoring agents that integrate streaming sensor data (Tasmurzayev et al., 2025) are representative examples in the reviewed literature, where persistent memory is essential to sustained, context-aware operation.

Table 1 summarizes the recurring architectural components of agentic systems, together with representative implementations reported across the corpus.

Table 1. Core Architectural Components of Agentic AI Systems

Component	Function	Representative Implementations
Reasoning Engine	Interleaves thought and action	ReAct, Chain-of-Thought, Tree of Thoughts
Short-term Memory	Maintains context window state	In-context learning, scratchpads
Long-term Memory	Persistent knowledge storage	Vector DBs (Pinecone, Weaviate), knowledge graphs
Tool Interface	External capability integration	Function calling, API wrappers, MCP
Planning Module	Task decomposition and scheduling	Hierarchical planning, task graphs, PDDL
Reflection Mechanism	Self-evaluation and correction	Self-reflection, critic agents, reward models
Agent Orchestration	Multi-agent coordination	AutoGen, CrewAI, LangGraph, MetaGPT

Table 1 groups these components by function and lists representative implementations reported in the literature, spanning reasoning frameworks such as ReAct and Tree of Thoughts, memory mechanisms built on vector databases and knowledge graphs, tool interfaces including function calling and the Model Context Protocol, and orchestration frameworks such as AutoGen and LangGraph. The recurring distinction between reasoning, memory, tool use, and coordination provides the organizing lens for the architectural discussion that follows.

JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer−Reviewed

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 37–59
DOI: 10.63646/jaiaa.2025.030103

4.3 Tool Use and Retrieval-Augmented Generation

The capacity to invoke external tools transforms language models from closed knowledge stores into systems that can act on current information. Two capabilities dominate the reviewed architectures: retrieval-augmented generation (RAG), which grounds outputs in external documents and data, and structured tool or function calling, through which agents query APIs, databases, and simulators.

RAG is among the most widely adopted techniques in the corpus, serving both as a grounding mechanism and as a basis for dynamic information access. Reviewed systems report that retrieval augmentation improves factual accuracy relative to unaugmented models, particularly in specialized domains (Woo et al., 2025), and that domain-tuned retrieval pipelines outperform generic prompting for technical question answering and monitoring tasks (Qian et al., 2025; Jeon et al., 2025). A recurring trend is the shift from single-shot retrieval toward adaptive, multi-step retrieval that refines queries based on intermediate results.

Beyond retrieval, tool-using agents in the corpus act on the world through domain-specific interfaces, from electronic-design automation (Wu et al., 2024) to clinical decision support (Ferber et al., 2025) and scientific-discovery pipelines that chain retrieval, computation, and simulation (Ghafarollahi et al., 2025; Gao, S. et al., 2024). As these tool ecosystems grow, standardizing how agents discover and invoke tools has become an active concern; emerging interface standards such as the Model Context Protocol aim to replace the custom adapters that individual systems previously required, though the reviewed literature indicates this standardization is still maturing.

Tool selection itself is emerging as a distinct problem: as the number of available tools grows, agents must identify the relevant ones efficiently, and reviewed systems approach this through semantic matching, model-based relevance scoring, and hierarchical organization of tools into domain categories.

4.4 Multi-Agent Architectures and Orchestration

The progression from single agents to multi-agent systems is one of the most active directions in the corpus; author-keyword analysis places multi-agent systems among the most frequent terms in the reviewed literature (Figure 3). Multi-agent designs decompose tasks across specialized agents, enable parallel execution of independent subtasks, and support cross-verification through agent disagreement.

Reviewed systems illustrate a range of coordination patterns. Centralized, controller-driven orchestration—in which a manager agent decomposes a task and integrates the outputs of worker agents—appears in code-generation pipelines built on role specialization (Dong et al., 2024; He et al., 2025) and in scientific-discovery frameworks that assemble teams of expert agents (Ghafarollahi et al., 2025; Swanson et al., 2025). The same pattern extends to engineering mechanics (Ni et al., 2024) and to next-generation communication networks (Jiang et al., 2024).

Broad analyses of the paradigm (Hughes et al., 2025; He et al., 2025) emphasize that although multi-agent collaboration can improve solution quality, it also introduces failure modes absent from single-agent systems: communication overhead that grows with agent count, goal misalignment across independently developed agents, and cascading errors in which an early mistake propagates through the system. These reliability concerns—rather than raw capability—increasingly define the multi-agent research frontier.

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 37–59
DOI: 10.63646/jaiaa.2025.030103

Table 2 summarizes the design characteristics of several widely used open-source orchestration frameworks.

Table 2. Comparison of Multi-Agent Orchestration Frameworks

Framework	Coordination Model	State Management	Key Strength	Primary Limitation
AutoGen	Conversational, hierarchical	Message history, shared context	Flexible agent roles, human-in-the-loop	Complex configuration, debugging difficulty
CrewAI	Role-based, task-driven	Task outputs, agent memories	Intuitive role assignment, process templates	Less flexible for non-standard workflows
LangGraph	Graph-based state machines	Explicit state graph, checkpoints	Deterministic execution, visual debugging	Steeper learning curve, verbose setup
MetaGPT	SOP-based, structured	Document pool, code repositories	Software engineering optimization	Narrow domain applicability
DSPy	Programmatic, declarative	Compiled execution graphs	Systematic optimization, modular design	Requires programming expertise

The comparative analysis in Table 2 reveals a fundamental tension in multi-agent framework design: flexibility versus determinism. Frameworks like AutoGen prioritize adaptability and ease of human intervention, while LangGraph emphasizes reproducible execution through explicit state management. This tension reflects broader questions about the appropriate role of human oversight in agentic systems, with implications for safety, accountability, and practical deployment.

5. Application Domains

5.1 Healthcare and Biomedicine

Healthcare and biomedicine are among the most intensively studied application domains for agentic AI in the corpus, reflected in a substantial cluster of medical and clinical papers. The complexity of medical decision-making—integrating diverse information sources, reasoning under uncertainty, and meeting strict safety requirements—makes it both a demanding testbed and a high-stakes deployment setting for agentic systems (Gao, S. et al., 2024; Mehandru et al., 2024).

Reviewed systems span several roles. Agentic systems have been developed and validated for autonomous clinical tasks (Ferber et al., 2025), evaluated as clinical agents against realistic case workflows (Mehandru et al., 2024), and benchmarked on standardized medical examinations such as USMLE items (Yaneva et al., 2024). Reported accuracy on such benchmarks can be high, but the reviewed studies consistently caution that real-world performance depends on integration with clinical workflows and electronic health records rather than benchmark scores alone.

Biomedical discovery is a second high-impact area. Multi-agent systems have been proposed to accelerate stages of biomedical research, from hypothesis generation to experimental design (Gao, S. et al., 2024), and in structural biology a coordinated team of agents was used to design candidate SARS-CoV-2 nanobodies (Swanson et al., 2025). The reviewed work indicates such systems can compress early-stage discovery timelines, while emphasizing that experimental and biological validation remains an essential bottleneck that agents cannot remove.

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 37–59
DOI: 10.63646/jaiaa.2025.030103

A recurring theme across these applications is the tension between automation and human oversight. The reviewed literature consistently favors human-in-the-loop designs in which agents propose but clinicians approve consequential decisions, reflecting both regulatory requirements and the recognition that current systems lack the contextual and ethical judgment required for fully autonomous practice (Mehandru et al., 2024; Ferber et al., 2025).

5.2 Scientific Research and Knowledge Discovery

Scientific research is a prominent application frontier in the corpus, spanning chemistry, materials science, and the broader automation of research workflows. A review of LLMs and autonomous agents in chemistry (Ramos et al., 2025) documents how such systems automate multi-step tasks that traditionally demand extensive human expertise.

Chemistry and materials science illustrate the pattern. Autonomous systems have been built for reticular-chemistry design (Rampal et al., 2025), end-to-end chemical synthesis (Ruan et al., 2024), and quantum-chemistry workflows (Zou et al., 2025), while multi-agent frameworks such as SciAgents automate aspects of scientific discovery through coordinated specialist agents (Ghafarollahi et al., 2025). The reviewed studies report substantial reductions in workflow time, while stressing that expert oversight remains essential for interpreting and validating results.

A consistent insight is that scientific agents deliver the most value when they augment rather than replace human expertise. The most effective implementations position agents as assistants that handle routine computation, suggest hypotheses, and manage data integration, while reserving strategic and creative judgment for human researchers (Ramos et al., 2025; Swanson et al., 2025).

5.3 Education and Personalized Learning

Education and learning form another active application area in the corpus, with growing attention to how generative and agentic systems reshape teaching and learning. Much of this work examines the changing roles of educators and learners as these tools are adopted (Lan et al., 2024; Huo et al., 2024).

Reviewed studies address both opportunities and risks. Work on teachers' agency emphasizes that the value of these systems depends on how they are integrated into pedagogy rather than on raw capability (Lan et al., 2024), and studies of generative AI in higher education highlight the need to cultivate critical digital literacies alongside the technology (Huo et al., 2024; Liu et al., 2025). Where agentic assistants are used for feedback and assessment, the reviewed literature stresses transparency and human review of uncertain cases.

A notable emphasis in educational settings is explainability: unlike domains where agent outputs may be accepted opaquely, learning contexts require that systems articulate their reasoning to support student understanding, which has motivated designs that generate step-by-step explanations and scaffolded hints (Lan et al., 2024).

5.4 Software Engineering and Code Generation

Software engineering is a well-represented application domain in the corpus, favored because code offers execution feedback and clear success criteria—compilation, passing tests—that give agents relatively well-defined action spaces (Dong et al., 2024; He et al., 2025).

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 37–59
DOI: 10.63646/jaiaa.2025.030103

Multi-agent approaches are especially prominent here. Self-collaboration frameworks assign complementary roles—such as analyst, coder, and tester—to cooperating agents and report improved performance on standard code-generation benchmarks over single-agent baselines (Dong et al., 2024). A literature review of LLM-based multi-agent systems for software engineering (He et al., 2025) catalogs this rapid growth across tasks ranging from code generation to debugging and testing.

A recurring limitation is the management of long-horizon tasks: while agents perform well on isolated functions or small programs, they struggle to sustain coherence across large codebases with many modules and evolving requirements—a limitation attributed to context constraints and the difficulty of hierarchical planning in complex design spaces (He et al., 2025).

5.5 Emerging Domains: Finance, Law, and Creative Industries

Beyond these dominant areas, the corpus shows growing activity in specialized domains. In finance and business, agentic and generative systems are applied to decision support (Albashrawi et al., 2025), supply-chain process automation (Aylak et al., 2025), structured extraction of ESG information (Zou et al., 2025), and the reshaping of business models (Teng et al., 2025).

Legal applications remain earlier-stage but are emerging: recent work simulates judicial judgment with LLMs to study how models weigh competing considerations (Loru et al., 2025), while cautioning that such systems can misinterpret nuanced legal reasoning. Decision-support studies similarly note both promise and the need for professional oversight in high-stakes domains (Albashrawi et al., 2025).

Creative and design-oriented applications appear as well, typically positioning agents as collaborative partners rather than autonomous creators—leveraging the complementary strengths of human contextual judgment and rapid AI iteration. Broad surveys of the field note this human–AI co-creation pattern across design and content-generation tasks (Acharya et al., 2025).

6. Emerging Challenges and Risks

6.1 Hallucination and Factual Grounding

Despite advances in retrieval augmentation and tool use, hallucination—plausible but incorrect output—remains a persistent challenge, and the reviewed literature identifies it as a central barrier to trustworthy deployment (Murugesan et al., 2025). In multi-agent settings the problem takes on new dimensions: errors introduced by one agent can be amplified rather than corrected as they propagate through the system.

Reviewed systems pursue several mitigation strategies with varying success. Retrieval augmentation reduces but does not eliminate hallucination, since retrieved sources may themselves be wrong or misapplied, and multi-agent verification—independent agents cross-checking one another—shows promise but adds computational overhead and can fail when agents share the same biases (Murugesan et al., 2025).

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 37–59
DOI: 10.63646/jaiaa.2025.030103

A critical distinction is that hallucination in agentic systems differs qualitatively from that in standalone models: when agents act autonomously, incorrect outputs can trigger real-world consequences—executing an erroneous query, generating unsafe code, or issuing a flawed recommendation—before human oversight can intervene, elevating hallucination from a quality issue to a safety concern (Murugesan et al., 2025).

6.2 Trust, Reliability, and Robustness

Establishing trustworthy agentic systems is a cross-cutting concern in the corpus, encompassing correctness (reliability), predictable behavior under varying conditions (robustness), and transparency to human overseers (explainability) (Murugesan et al., 2025).

Agentic systems also present an expanded attack surface relative to standalone models—through their tool interfaces, memory, and inter-agent communication channels. Reviewed work documents privacy and security risks specific to tool-using agents and proposes safeguards for protecting user data during tool invocation (Zhang et al., 2024). Adversarial perturbations and prompt injection are recurring concerns, and the added cost and latency of verification further complicate trust in practice.

In scientific applications, the reviewed literature highlights a subtler issue: foundation models blur the line between tools that we rely upon and collaborators that we trust, a categorical ambiguity with consequences for how responsibility is assigned when agent actions cause harm (Ramos et al., 2025).

6.3 Coordination Failures and Emergent Dynamics

Multi-agent systems introduce failure modes absent from single-agent deployments, including communication breakdowns, goal misalignment, and emergent dynamics that subvert intended behavior. Generative agent-based modeling has become a common lens for studying these dynamics (Gao, C. et al., 2024; Lu et al., 2024).

Documented failure patterns include goal misalignment, in which agents optimize local objectives that conflict with system goals; information cascades, in which early incorrect decisions by influential agents propagate as later agents defer to apparent consensus; and oscillation, in which agents enter cycles of contradictory actions, particularly in competitive or mixed-motive settings (Lu et al., 2024; Ghaffarzadegan et al., 2024).

Studies of LLM agents in game-theoretic and social settings show that behaviors such as cooperation, defection, and conditional strategies emerge from interaction dynamics rather than being explicitly programmed, varying with prompt framing and model choice (de Zarzà et al., 2023; Xu et al., 2024). This emergent complexity enables rich social simulation but complicates the prediction and control of system behavior.

6.4 Scalability and Resource Efficiency

Scalability is a prominent concern in the corpus, encompassing the computational cost, latency, and environmental footprint of agentic deployment. Multi-agent systems in particular incur substantial overhead relative to single-model approaches, with token consumption growing non-linearly in the number of agents and interaction rounds (Gao, C. et al., 2024).

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 37–59
DOI: 10.63646/jaiaa.2025.030103

The cost–performance trade-off is not straightforward: providing more information to more agents in parallel can improve results but quickly becomes expensive, and reviewed work suggests that how information flow is structured—what to share, when, and with which agents—can matter as much as model choice for overall efficiency (Lu et al., 2024).

Environmental considerations, though less frequently addressed, are gaining attention: the energy demands of systems that issue many model calls per task raise sustainability questions that the reviewed literature expects to grow more salient as deployment scales (Murugesan et al., 2025).

7. Future Directions and Conclusion

7.1 Toward Standardized Evaluation and Benchmarking

A recurring gap identified across the corpus is the lack of standardized evaluation for agentic systems. While individual studies report strong results on bespoke tasks, the absence of common benchmarks complicates comparison and progress tracking, and the reviewed surveys call for evaluation practices tailored to the distinctive demands of agentic systems (Acharya et al., 2025; Hughes et al., 2025).

Future benchmarks should capture what makes agentic systems distinctive: long-horizon task completion, multi-agent coordination under partial observability, robustness to distribution shift, and graceful degradation under resource limits. Evaluation must also extend beyond accuracy to efficiency, reliability, and safety indicators that reflect real deployment conditions (Murugesan et al., 2025).

7.2 Human-Agent Collaboration and Social Integration

A consistent theme across the reviewed literature is that the most promising near-term trajectory lies not in full autonomy but in effective human–agent collaboration, with human-in-the-loop designs recurring across high-stakes domains (Mehandru et al., 2024; Ferber et al., 2025). Future work should clarify optimal task allocation between humans and agents and the interaction protocols that keep human overseers informed.

The social integration of agentic systems also raises broader questions. Studies of populations of interacting LLM agents report emergent collective behaviors analogous to human social dynamics (Gao, C. et al., 2024; Xu et al., 2024), suggesting that understanding and shaping these dynamics will be important for beneficial deployment.

7.3 Governance, Safety, and Societal Impact

The pace of agentic-AI deployment is outrunning regulatory development, creating governance challenges that the reviewed literature increasingly acknowledges (Murugesan et al., 2025). Frameworks designed for human-intermediated processes are poorly suited to autonomous agent behavior in financial, legal, and safety-critical settings.

Reviewed work points to several governance priorities: explainability mechanisms that make agent decisions interpretable to non-technical stakeholders, security architectures that guard against prompt injection and data leakage

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 37–59
DOI: 10.63646/jaiaa.2025.030103

(Zhang et al., 2024), and accountability structures that clarify responsibility when agent actions cause harm (Murugesan et al., 2025).

7.4 Conclusion

This comprehensive review of 810 publications reveals agentic AI as a rapidly maturing field with transformative potential across diverse domains. The exponential growth in research output, the diversification of application areas, and the increasing sophistication of architectural approaches collectively indicate that agentic systems are transitioning from research curiosity to practical infrastructure.

Yet significant challenges persist. The field must address hallucination in safety-critical contexts, establish robust coordination protocols for multi-agent systems, and develop governance frameworks appropriate for autonomous decision-making. The tension between flexibility and determinism, between automation and human oversight, and between innovation and safety will shape the trajectory of agentic AI development.

The age of autonomous agents has arrived, but its ultimate character—whether empowering or destabilizing, collaborative or displacing—remains to be determined by the collective choices of researchers, developers, policymakers, and society. This review provides a foundation for informed engagement with these choices, mapping the current landscape and identifying the critical paths forward.

References

Acharya, D. B., Kuppan, K., & Divya, B. (2025). Agentic ai: Autonomous intelligence for complex goals-a comprehensive survey. IEEE ACCESS, 13, 18912-18936. https://doi.org/10.1109/ACCESS.2025.3532853.

Albashrawi, M. (2025). Generative ai for decision-making: A multidisciplinary perspective. JOURNAL OF INNOVATION & KNOWLEDGE, 10(4). https://doi.org/10.1016/j.jik.2025.100751.

Aylak, B. L. (2025). Sustai-scm: Intelligent supply chain process automation with agentic ai for sustainability and cost efficiency. SUSTAINABILITY, 17(6). https://doi.org/10.3390/su17062453.

de Zarzà, I., de Curtò, J., Roig, G., Manzoni, P., & Calafate, C. T. (2023). Emergent cooperation and strategy adaptation in multi-agent systems: An extended coevolutionary theory with llms. ELECTRONICS, 12(12). https://doi.org/10.3390/electronics12122722.

Dong, Y. H., Jiang, X., Jin, Z., & Li, G. (2024). Self-collaboration code generation via chatgpt. ACM TRANSACTIONS ON SOFTWARE ENGINEERING AND METHODOLOGY, 33(7). https://doi.org/10.1145/3672459.

Fan, H. L., Liu, X., Fuh, J. Y. H., Lu, W. F., & Li, B. B. (2025). Embodied intelligence in manufacturing: Leveraging large language models for autonomous industrial robotics. JOURNAL OF INTELLIGENT MANUFACTURING, 36(2), 1141-1157. https://doi.org/10.1007/s10845-023-02294-y.

Ferber, D., El Nahhas, O. S. M., Wölflein, G., Wiest, I. C., Clusmann, J., Lessmann, M. E., et al. (2025). Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. NATURE CANCER, 6(8). https://doi.org/10.1038/s43018-025-00991-6.

Gao, C., Lan, X. C., Li, N., Yuan, Y., Ding, J. T., Zhou, Z. L., et al. (2024). Large language models empowered agent-based modeling and simulation: A survey and perspectives. HUMANITIES & SOCIAL SCIENCES COMMUNICATIONS, 11(1). https://doi.org/10.1057/s41599-024-03611-3.

Gao, S. H., Fang, A., Huang, Y. P., Giunchiglia, V., Noori, A., Schwarz, J. R., et al. (2024). Empowering biomedical discovery with ai agents. CELL, 187(22), 6125-6151. https://doi.org/10.1016/j.cell.2024.09.022.

Ghafarollahi, A., & Buehler, M. J. (2025). Sciagents: Automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. ADVANCED MATERIALS, 37(22). https://doi.org/10.1002/adma.202413523.

Ghaffarzadegan, N., Majumdar, A., Williams, R., & Hosseinichimeh, N. (2024). Generative agent-based modeling: An introduction and tutorial. SYSTEM DYNAMICS REVIEW, 40(1). https://doi.org/10.1002/sdr.1761.

He, J. D., Treude, C., & Lo, D. (2025). Llm-based multi-agent systems for software engineering: Literature review, vision, and the road ahead. ACM TRANSACTIONS ON SOFTWARE ENGINEERING AND METHODOLOGY, 34(5). https://doi.org/10.1145/3712003.

Hughes, L., Dwivedi, Y. K., Malik, T., Shawosh, M., Albashrawi, M. A., Jeon, I., et al. (2025). Ai agents and agentic systems: A multi-expert analysis. JOURNAL OF COMPUTER INFORMATION SYSTEMS. https://doi.org/10.1080/08874417.2025.2483832.

Huo, X. N., & Siau, K. L. (2024). Generative artificial intelligence in business higher education: A focus group study. JOURNAL OF GLOBAL INFORMATION MANAGEMENT, 32(1). https://doi.org/10.4018/JGIM.364093.

Jeon, J., Sim, Y., Lee, H., Han, C., Yun, D., Kim, E., et al. (2025). Chatcnc: Conversational machine monitoring via large language model and real-time data retrieval augmented generation. JOURNAL OF MANUFACTURING SYSTEMS, 79, 504-514. https://doi.org/10.1016/j.jmsy.2025.01.018.

Jiang, F. B., Peng, Y. B., Dong, L., Wang, K. Z., Yang, K., Pan, C. H., et al. (2024). Large language model enhanced multi-agent systems for 6g communications. IEEE WIRELESS COMMUNICATIONS, 31(6), 48-55. https://doi.org/10.1109/MWC.016.2300600.

Lan, Y. J., & Chen, N. S. (2024). Teachers" agency in the era agency of llm and generative ai: Designing pedagogical ai agents. EDUCATIONAL TECHNOLOGY & SOCIETY, 27(1). https://doi.org/10.30191/ETS.202401_27(1).PP01.

Li, Z. L., & Ning, H. (2023). Autonomous gis: The next-generation ai-powered gis. INTERNATIONAL JOURNAL OF DIGITAL EARTH, 16(2), 4668-4686. https://doi.org/10.1080/17538947.2023.2278895.

Liu, G. L., Lee, J. S., & Zhao, X. (2025). Critical digital literacies, agentic practices, and ai-mediated informal digital learning of english. SYSTEM, 134. https://doi.org/10.1016/j.system.2025.103797.

Loru, E., Nudo, J., Di Marco, N., Santirocchi, A., Atzeni, R., Cinelli, M., et al. (2025). The simulation of judgment in llms. PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA, 122(42). https://doi.org/10.1073/pnas.2518443122.

Lu, Y. K., Aleta, A., Du, C. P., Shi, L., & Moreno, Y. (2024). Llms and generative agent-based models for complex systems research. PHYSICS OF LIFE REVIEWS, 51, 283-293. https://doi.org/10.1016/j.plrev.2024.10.013.

Mehandru, N., Miao, B. Y., Almaraz, E. R., Sushil, M., Butte, A. J., & Alaa, A. (2024). Evaluating large language models as agents in the clinic. NPJ DIGITAL MEDICINE, 7(1). https://doi.org/10.1038/s41746-024-01083-y.

Murugesan, S. (2025). The rise of agentic ai: Implications, concerns, and the path forward. IEEE INTELLIGENT SYSTEMS, 40(2), 8-14. https://doi.org/10.1109/MIS.2025.3544940.

Ni, B., & Buehler, M. J. (2024). Mechagents: Large language model multi-agent collaborations can solve mechanics problems, generate new data, and integrate knowledge. EXTREME MECHANICS LETTERS, 67. https://doi.org/10.1016/j.eml.2024.102131.

Qian, Z. H., & Shi, C. (2025). Large language model-empowered paradigm for automated geotechnical site planning and geological characterization. AUTOMATION IN CONSTRUCTION, 173. https://doi.org/10.1016/j.autcon.2025.106103.

Ramos, M. C., Collison, C. J., & White, A. D. (2025). A review of large language models and autonomous agents in chemistry. CHEMICAL SCIENCE, 16(6), 2514-2572. https://doi.org/10.1039/d4sc03921a.

Rampal, N., Inizan, T. J., Borgs, C., Chayes, J. T., & Yaghi, O. M. (2025). Large language models for reticular chemistry. NATURE REVIEWS MATERIALS, 10(5), 369-381. https://doi.org/10.1038/s41578-025-00772-8.

Ruan, Y. X., Lu, C. Y., Xu, N., He, Y. C., Chen, Y. X., Zhang, J., et al. (2024). An automatic end-to-end chemical synthesis development platform powered by large language models. NATURE COMMUNICATIONS, 15(1). https://doi.org/10.1038/s41467-024-54457-x.

Swanson, K., Wu, W., Bulaong, N. L., Pak, J. E., & Zou, J. (2025). The virtual lab of ai agents designs new sars-cov-2 nanobodies. NATURE, 646(8085). https://doi.org/10.1038/s41586-025-09442-9.

Tasmurzayev, N., Amangeldy, B., Imanbek, B., Baigarayeva, Z., Imankulov, T., Dikhanbayeva, G., et al. (2025). Digital cardiovascular twins, ai agents, and sensor data: A narrative review from system architecture to proactive heart health. SENSORS, 25(17). https://doi.org/10.3390/s25175272.

Teng, D. Q., Ye, C., & Martinez, V. (2025). Gen-ai's effects on new value propositions in business model innovation: Evidence from information technology industry. TECHNOVATION, 143. https://doi.org/10.1016/j.technovation.2025.103191.

Wang, L., Ma, C., Feng, X. Y., Zhang, Z. Y., Yang, H., Zhang, J. S., et al. (2024). A survey on large language model based autonomous agents. FRONTIERS OF COMPUTER SCIENCE, 18(6). https://doi.org/10.1007/s11704-024-40231-1.

Woo, J. J., Yang, A. J., Olsen, R. J., Hasan, S. S., Nawabi, D. H., Nwachukwu, B. U., et al. (2025). Custom large language models improve accuracy: Comparing retrieval augmented generation and artificial intelligence agents to noncustom models for evidence-based medicine. ARTHROSCOPY-THE JOURNAL OF ARTHROSCOPIC AND RELATED SURGERY, 41(3). https://doi.org/10.1016/j.arthro.2024.10.042.

Wu, H. Y., He, Z. L., Zhang, X. Y., Yao, X. F., Zheng, S., Zheng, H. S., et al. (2024). Chateda: A large language model powered autonomous agent for eda. IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS, 43(10), 3184-3197. https://doi.org/10.1109/TCAD.2024.3383347.

Xi, Z. H., Chen, W. X., Guo, X., He, W., Ding, Y. W., Hong, B. Y., et al. (2025). The rise and potential of large language model based agents: A survey. SCIENCE CHINA-INFORMATION SCIENCES, 68(2). https://doi.org/10.1007/s11432-024-4222-0.

Xu, M. R., Niyato, D., Kang, J. W., Xiong, Z. H., Mao, S. W., Han, Z., et al. (2024). When large language model agents meet 6g networks: Perception, grounding, and alignment. IEEE WIRELESS COMMUNICATIONS, 31(6), 63-71. https://doi.org/10.1109/MWC.005.2400019.

Xu, R. X., Sun, Y. F., Ren, M. J., Guo, S. G., Pan, R. T., Lin, H. Y., et al. (2024). Ai for social science and social science of ai: A survey. INFORMATION PROCESSING & MANAGEMENT, 61(3). https://doi.org/10.1016/j.ipm.2024.103665.

Yaneva, V., Baldwin, P., Jurich, D. P., Swygert, K., & Clauser, B. E. (2024). Examining chatgpt performance on usmle sample items and implications for assessment. ACADEMIC MEDICINE, 99(2), 192-197. https://doi.org/10.1097/ACM.0000000000005549.

Zhang, R. C., Du, H. Y., Liu, Y. Q., Niyato, D., Kang, J. W., Xiong, Z. H., et al. (2024). Generative ai agents with large language model for satellite networks via a mixture of experts transmission. IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, 42(12), 3581-3596. https://doi.org/10.1109/JSAC.2024.3459037.

Zhang, X. Y., Xu, H. Y., Ba, Z. J., Wang, Z. B., Hong, Y., Liu, J., et al. (2024). Privacyasst: Safeguarding user privacy in tool-using large language model agents. IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, 21(6), 5242-5258. https://doi.org/10.1109/TDSC.2024.3372777.

Zhang, Z. Y., Dai, Q. Y., Bo, X. H., Ma, C., Li, R., Chen, X., et al. (2025). A survey on the memory mechanism of large language model-based agents. ACM TRANSACTIONS ON INFORMATION SYSTEMS, 43(6). https://doi.org/10.1145/3748302.

Zou, Y., Shi, M. Y., Chen, Z. J., Deng, Z., Lei, Z. X., Zeng, Z. H., et al. (2025). Esgreveal: An llm-based approach for extracting structured data from esg reports. JOURNAL OF CLEANER PRODUCTION, 489. https://doi.org/10.1016/j.jclepro.2024.144572.

Zou, Y. H., Cheng, A. H., Aldossary, A., Bai, J. R., Leong, S. X., Campos-Gonzalez-Angulo, J. A., et al. (2025). El agente: An autonomous agent for quantum chemistry. MATTER, 8(7). https://doi.org/10.1016/j.matt.2025.102263.