Agentic Artificial Intelligence:A Systematic Review of Architectures, Benchmarks, and Governance Frameworks

Audrey Rah

arahimi@uh.edu

University of Houston https://orcid.org/0000-0001-7501-7809

Systematic Review

Keywords: LLM agents, autonomous intelligent systems, multi-agent orchestration, benchmark evaluation, tool-mediated reasoning, software engineering automation, AI safety risks, algorithmic governance, responsible AI, systematic literature review, agentic workflow

Posted Date: May 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9839527/v1

Additional Declarations: The authors declare no competing interests.

AGENTIC ARTIFICIAL INTELLIGENCE: A SYSTEMATIC REVIEW OF ARCHITECTURES, BENCHMARKS, AND GOVERNANCE FRAMEWORKS

A PREPRINT

Audrey Rah
Department of Electrical and Computer Engineering
University of Houston
Houston, TX 77204, USA
arahimi@uh.edu

May 27, 2026

ABSTRACT

Agentic artificial intelligence (AI) has emerged as an important direction in the development of goal-oriented, tool-mediated, and partially autonomous intelligent systems. Recent advances in large language models have enabled AI agents to perform multi-step tasks involving planning, memory, tool use, environmental feedback, and human–agent or multi-agent interaction. However, the literature remains fragmented across architectures, benchmarks, application domains, and governance discussions. This review provides a systematic synthesis of recent evidence on agentic AI systems, with emphasis on architectural components, evaluation methods, domain applications, safety risks, and governance requirements. The review analyzes literature from scholarly databases, major conference proceedings, technical reports, benchmark studies, and governance sources published mainly between 2022 and 2026. Current evidence suggests that agentic AI systems appear most mature in software engineering, web automation, and computer-use environments, where executable feedback and sandboxed testing are available. In contrast, high-impact domains such as healthcare and finance require stronger validation, oversight, privacy protection, and regulatory control before broader deployment. This review also introduces a Five-Level Agentic Autonomy and Governance Model to classify systems according to action authority, operational risk, and required oversight. Overall, the evidence suggests that agentic AI may be better understood as a spectrum of bounded, monitored, and auditable systems rather than fully independent intelligence. Future research should prioritize reproducible benchmarks, safety evaluation, real-world validation, and governance-aware system design.

Keywords LLM agents · autonomous intelligent systems · multi-agent orchestration · benchmark evaluation · tool-mediated reasoning · software engineering automation · AI safety risks · algorithmic governance · responsible AI · systematic literature review · agentic workflows

1 Introduction

Artificial intelligence (AI) is evolving from systems that mainly classify, predict, or generate content toward systems capable of goal-oriented reasoning and interaction with external environments (1; 2). Recent advances in large language models (LLMs) have accelerated this transition by enabling systems to perform multi-step reasoning, tool use, code generation, memory utilization, and sequential task execution (3; 4; 5). Modern agentic AI systems increasingly interact with APIs, software tools, databases, browsers, and operational environments through feedback-driven workflows (6; 7; 8; 9). These systems are increasingly described as agentic artificial intelligence (agentic AI) or AI agents because they extend beyond passive text generation and support partially autonomous operational behavior (10; 11). The growing interest in agentic AI is driven by the need for intelligent systems capable of supporting complex real-

world workflows across software engineering, web automation, healthcare, finance, robotics, and scientific research (12; 13; 14; 15; 16; 17; 18; 19; 20). Unlike conventional conversational systems, agentic AI architectures often combine planning, memory, environmental feedback, external tools, and multi-agent coordination within unified operational workflows (1; 21). Frameworks such as ReAct, Reflexion, Voyager, ToolLLM, AutoGen, and MetaGPT illustrate this transition from static response generation toward iterative reasoning and action execution (6; 22; 23; 9; 24; 25). Despite rapid progress, the field remains fragmented across architectures, benchmarks, application domains, and governance frameworks (2; 26). Definitions of autonomy vary considerably between studies, and benchmark results are often difficult to compare because evaluation environments, metrics, and operational assumptions differ considerably (27; 14; 28). Furthermore, many current demonstrations emphasize capability while providing limited evidence regarding long-term reliability, operational safety, reproducibility, or real-world deployment readiness (29; 30). Current benchmarks continue to reveal substantial weaknesses in long-horizon planning, interface grounding, tool selection, memory consistency, and error recovery (31; 32). The expansion of tool-mediated autonomy also introduces new governance and security concerns. Traditional conversational AI systems primarily generate outputs for human interpretation. In contrast, agentic AI systems may execute code, access databases, navigate websites, call APIs, manipulate files, or trigger external workflows (33; 34). As a result, model errors may directly translate into operational consequences such as data leakage, unsafe execution, privilege escalation, or cascading workflow failures (35; 36). Recent governance frameworks increasingly emphasize least privilege, sandboxing, monitoring, rollback mechanisms, approval gates, and continuous oversight for AI agents operating in real-world environments (37; 38; 39). Several recent surveys have explored aspects of LLM-based agents, multi-agent systems, or autonomous AI workflows (1; 40; 2). However, relatively few reviews integrate architectural analysis, benchmark evaluation, governance requirements, deployment maturity, and operational risk within a unified evidence framework. In addition, many reviews focus primarily on technical architectures without systematically connecting autonomy levels to oversight requirements. Therefore, this review provides a systematic synthesis of recent evidence on agentic AI architectures, benchmarks, application domains, governance frameworks, and deployment challenges. The review examines literature published primarily between 2022 and 2026 across scholarly databases, benchmark repositories, technical reports, and governance sources. In addition, this work introduces a new Five-Level Agentic Autonomy and Governance Model (FAAGM) that classifies systems according to operational authority, risk exposure, and required oversight mechanisms. The proposed framework aims to support clearer classification, safer deployment strategies, and more consistent evaluation of emerging agentic AI systems. This review provides a systematic evidence synthesis that jointly analyzes agentic AI architectures, benchmark validity, deployment maturity, governance requirements, operational risks, and autonomy-level classification within a unified analytical framework.

2 Literature Review

The development of agentic artificial intelligence has emerged from the convergence of large language models, autonomous decision-making systems, reasoning frameworks, and tool-mediated computational environments. Earlier artificial intelligence research investigated autonomous agents primarily within reinforcement learning, symbolic planning, and multi-agent systems (41; 42; 43). These traditional systems established foundational concepts related to environmental interaction, sequential decision-making, and adaptive behavior. However, most earlier autonomous architectures remained constrained by limited flexibility, rigid symbolic rules, narrow task domains, and weak natural-language reasoning capabilities. The emergence of transformer-based large language models significantly changed this trajectory by introducing highly scalable language understanding and contextual reasoning mechanisms (44; 3). Recent advances in instruction tuning, reinforcement learning from human feedback, and reasoning-oriented prompting strategies further improved the ability of language models to solve multi-step tasks and generate context-aware outputs (4; 5; 45). Modern agentic AI systems extend beyond conventional conversational interfaces by integrating planning, memory, tool utilization, and environmental feedback into unified operational workflows (1; 2). Rather than generating isolated responses, these systems attempt to pursue objectives iteratively through observation, reasoning, execution, and revision. Several influential frameworks have contributed to this transition. ReAct introduced a reasoning-and-action paradigm that interleaves intermediate reasoning traces with executable actions during task execution (6). Toolformer indicated that language models can learn to use external tools through self-supervised interaction (7). Gorilla and ToolLLM expanded these capabilities by enabling interaction with large-scale API ecosystems and external computational services (8; 9). Reflexion incorporated verbal self-feedback and episodic memory to support adaptive iterative improvement without modifying model parameters (22). Additional systems such as Generative Agents and Voyager further explored persistent memory, environmental exploration, skill acquisition, and long-horizon autonomous behavior in simulated interactive environments (46; 23). Figure 1 summarizes the layered operational structure commonly observed in contemporary agentic AI systems, including reasoning, memory, orchestration, external tools, execution environments, and governance mechanisms. Recent literature increasingly emphasizes that agentic behavior emerges not only from the underlying language model but also from the surrounding architectural

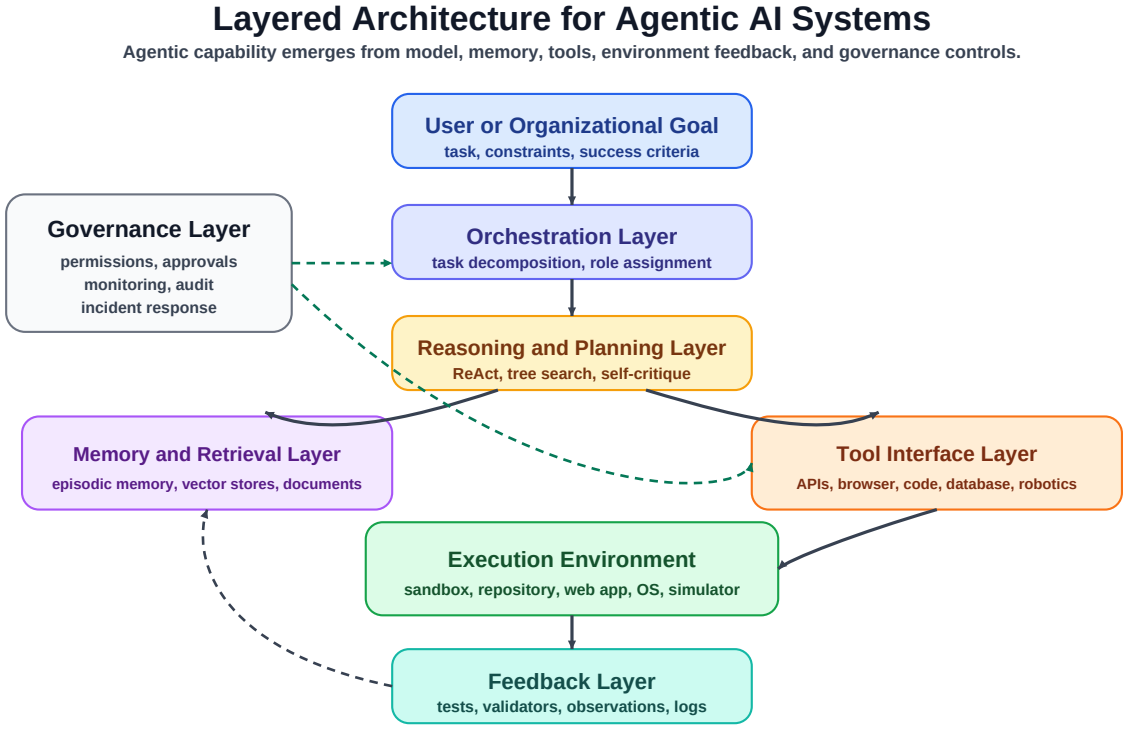

Figure 1: Layered architecture of modern agentic artificial intelligence systems including reasoning, memory, orchestration, tool interaction, execution, and governance components.

workflow (21; 26). Contemporary agentic systems frequently combine planning modules, retrieval systems, vector memory stores, tool interfaces, environmental observations, execution validators, and orchestration layers. Multi-agent architectures have also gained considerable attention because they allow specialized agents to collaborate through planner-reviewer-executor structures (24; 25; 47; 48). These collaborative systems attempt to improve scalability, task decomposition, and verification by distributing responsibilities across multiple interacting components. Nevertheless, current studies also indicate that coordination overhead, communication ambiguity, and cascading reasoning errors remain important limitations in complex multi-agent workflows (40). The literature further suggests that reasoning and planning remain central challenges in agentic AI research. Traditional prompting often struggles with long-horizon reasoning, environmental adaptation, and sequential decision consistency. Several methods have therefore attempted to improve structured reasoning performance. Additional reasoning-oriented studies such as Self-Refine and MRKL systems also explored iterative self-correction and modular neuro-symbolic reasoning integration within language-model workflows (49; 50). Chain-of-thought prompting introduced intermediate reasoning steps to improve problem-solving capabilities (5). Tree-of-Thoughts expanded this concept through branching reasoning exploration and deliberate search strategies (51). Least-to-most prompting decomposed complex problems into sequential subtasks to improve planning consistency (52). Other studies combined language models with formal planners, embodied feedback systems, and robotic interaction frameworks (53; 54; 55). Together, these studies indicate that effective agentic systems increasingly rely on hybrid reasoning structures that integrate language flexibility with environmental constraints and feedback-driven correction mechanisms. Tool interaction and environmental grounding have become defining characteristics of modern agentic systems. Additional environments such as WebShop, ALFWorld, BabyAI, and World of Bits further expanded evaluation toward grounded interaction, embodied environments, and web-based sequential decision-making tasks (56; 57; 58; 59). Unlike earlier NLP-oriented architectures, current agents increasingly interact with browsers, APIs, repositories, databases, operating systems, and multimodal interfaces (56; 14). WebArena, BrowserGym, and WorkArena introduced more realistic browser-based environments involving sequential web interaction and workflow automation (14; 16; 15). SWE-bench and SWE-agent suggested growing capability of language-model agents in software engineering tasks such as repository navigation, patch generation, and executable testing (12; 13). OSWorld extended evaluation into desktop and operating-system environments involving graphical interaction and multimodal operational tasks (28). These studies collectively suggest rapid progress in operational capability while simultaneously

revealing persistent limitations in interface grounding, long-horizon execution, error recovery, and environmental adaptation. The expansion of agentic AI has also generated increasing interest in benchmark development and evaluation methodologies.

Figure 2: Thematic synthesis of recurring research trends identified throughout the reviewed agentic AI literature. The figure summarizes major dimensions observed across the reviewed studies, including operational capability expansion, persistent architectural limitations, growing application domains, and governance or security risks associated with autonomous systems. The thematic groupings were derived through structured literature coding and evidence synthesis procedures applied across the reviewed corpus.

Figure 2 illustrates the major strengths, limitations, application domains, and governance risks associated with modern agentic AI systems. Conventional NLP benchmarks are often insufficient because agentic systems must maintain state, perform sequential reasoning, interact with tools, and adapt dynamically to environmental feedback. AgentBench introduced multiple operational environments to evaluate interactive agentic behavior across heterogeneous tasks (27). SWE-bench used executable software validation to evaluate repository-level code repair (12). MLAgentBench explored machine-learning experimentation workflows involving iterative experimentation and revision (31). Despite these advances, several studies continue to report substantial weaknesses in long-horizon reasoning, memory consistency, instruction following, tool selection, and recovery from unexpected states (29; 32). Additional concerns include

benchmark contamination, reproducibility limitations, evaluator inconsistency, and unrealistic deployment assumptions (30; 26). Consequently, benchmark performance alone may not accurately reflect real-world operational reliability. Agentic AI systems are increasingly explored across multiple application domains, including healthcare, finance, scientific research, robotics, and enterprise automation. Recent enterprise ecosystems have also accelerated the movement of agentic AI from experimental prototypes toward operational workflow platforms. Commercial systems such as Microsoft Copilot Studio, OpenAI Operator, Google AgentSpace, and Salesforce Agentforce illustrate the growing use of agentic workflows in enterprise automation, software interaction, and organizational decision-support environments (60; 61; 63; 65). In parallel, orchestration frameworks such as LangGraph, CrewAI, and AutoGen Studio further show the increasing role of modular agent coordination, tool-mediated workflow design, and multi-agent automation in practical deployment settings (67; 68; 69). In healthcare, recent studies investigated language-model agents for biomedical reasoning, medical summarization, documentation support, and clinical assistance (17; 70; 71). Financial-domain studies explored market analysis, compliance support, quantitative reasoning, and autonomous financial workflows (18; 72; 73). Additional work investigated robotic and embodied systems capable of combining language-based reasoning with environmental interaction (20; 74). Scientific research automation has similarly emerged as a growing field involving autonomous experimentation, literature synthesis, and computational discovery systems (19; 75; 76). However, most current evidence remains constrained to controlled evaluations, simulated environments, or narrowly scoped operational settings. Large-scale real-world validation remains relatively limited, particularly in high-risk domains involving safety, regulation, or sensitive decision-making. The governance and security implications of agentic AI have therefore become increasingly important within recent literature. Unlike traditional conversational systems that primarily generate text, agentic systems may execute code, manipulate files, retrieve sensitive information, interact with APIs, and trigger external workflows. Consequently, model failures may directly translate into operational harm, privilege misuse, financial risk, or data leakage (33; 34). Oversight frameworks proposed by NIST, OWASP, CISA, and other organizations increasingly emphasize least privilege, sandboxing, monitoring, rollback mechanisms, auditability, and controlled deployment practices for agentic systems (39; 37; 35; 36; 38). Additional concerns identified throughout the literature include prompt injection, adversarial manipulation, unsafe fine-tuning, jailbreak attacks, training-data leakage, hallucination propagation, and alignment instability in foundation models (77; 78; 79; 80; 81). These risks become increasingly significant when autonomous systems are connected to external tools, execution environments, memory stores, and operational workflows. Although recent surveys have examined LLM agents, autonomous AI systems, and multi-agent frameworks, important gaps remain within the current literature (1; 2; 40). Definitions of agency and autonomy remain inconsistent across studies, benchmark environments differ considerably in complexity and realism, and governance discussions are often separated from technical architectural analysis. Furthermore, many current studies emphasize capability demonstrations while providing limited discussion regarding deployment maturity, operational accountability, reproducibility, or long-term safety validation (26; 30). These limitations motivate the need for a systematic synthesis that integrates architectural analysis, benchmark evaluation, application-domain maturity, governance requirements, and operational risk considerations within a unified evidence-based review framework.

3 Methodology

This study was designed as a PRISMA-aligned systematic review focused on the emerging field of agentic artificial intelligence. Also, Figure 3 presents the PRISMA 2020-aligned study selection and evidence synthesis workflow adopted in this systematic review, including identification, screening, eligibility assessment, exclusion stages, and final qualitative synthesis integration. Because the literature spans peer-reviewed journal articles, conference proceedings, technical reports, benchmark repositories, governance frameworks, and influential preprints, a narrative evidence-synthesis methodology was adopted rather than statistical meta-analysis.

The objective of the methodology was not only to identify relevant studies, but also to organize architectural, evaluative, operational, and governance-related evidence within a unified analytical framework. The methodological structure was developed to align with recent review practices in artificial intelligence, software systems, and computational governance literature (1; 2; 40). The review process followed four major stages: literature identification, eligibility screening, structured data extraction, and qualitative evidence synthesis. Searches were conducted across multidisciplinary scholarly databases and major AI publication venues, including Scopus, Web of Science, IEEE Xplore, ACM Digital Library, ACL Anthology, PubMed, arXiv, ICLR proceedings, NeurIPS proceedings, and IJCAI proceedings. Additional governance and cybersecurity sources were collected from NIST, CISA, OWASP, and MITRE publications because governance and operational safety are increasingly central to agentic AI deployment (37; 35; 38). The primary search window covered literature published between January 2022 and May 2026, reflecting the rapid emergence of modern LLM-based agentic systems after the introduction of transformer-scale foundation models and tool-mediated reasoning architectures (3; 6). The search strategy combined four major concept groups: agentic systems, foundation models, evaluation environments, and governance or application domains. Representative search expressions included

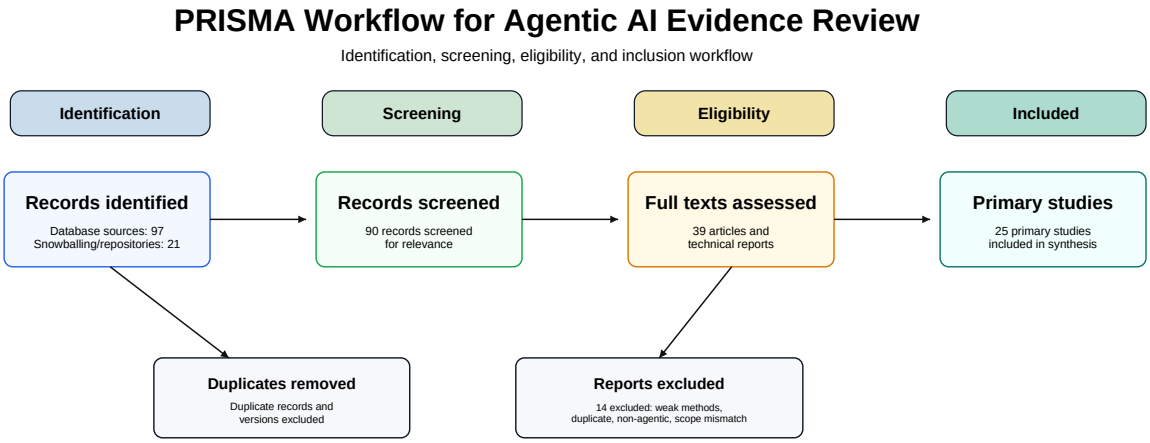

Figure 3: PRISMA 2020-based study identification, screening, eligibility assessment, and qualitative synthesis workflow used in this systematic evidence review of agentic artificial intelligence literature. The review process followed PRISMA-oriented evidence synthesis methodology, including duplicate removal, title and abstract screening, full-text eligibility assessment, and qualitative evidence integration (82).

combinations of terms such as “agentic AI”, “AI agents”, “LLM agents”, “autonomous agents”, “tool use”, “multi-agent systems”, “benchmark evaluation”, “software engineering”, “healthcare”, “finance”, “governance”, “prompt injection”, and “AI safety”. Search strings were adapted to the syntax requirements of each database. Citation snowballing was additionally performed to identify influential benchmark papers, architectural frameworks, and governance references that were repeatedly cited within the literature (27; 12; 28). Studies were included if they satisfied at least one of the following conditions: proposed or evaluated an LLM-based agentic architecture, introduced a benchmark or operational evaluation framework, analyzed agentic AI applications within a major domain, discussed governance or safety implications relevant to agentic systems, or proposed architectural, evaluative, or governance-related frameworks associated with agentic AI systems. Both peer-reviewed and influential preprint studies were considered because several widely referenced agentic AI frameworks and benchmark systems were initially released as technical preprints before formal publication (8; 23). Sources were excluded if they focused solely on conventional machine learning without autonomous or tool-mediated behavior, lacked sufficient methodological detail, represented primarily promotional or non-technical material, or duplicated a more complete version of the same work. Following preliminary retrieval, duplicate records and duplicate manuscript versions were removed manually. Title and abstract screening were then performed to assess relevance to the research questions. Full-text assessment focused on architectural relevance, evaluation methodology, operational capability, governance implications, and evidence quality. Because the field remains highly heterogeneous in terms of benchmarks, environments, scoring methods, and deployment assumptions, quantitative meta-analysis was not considered methodologically appropriate. Instead, the included literature was synthesized qualitatively through comparative matrices and thematic categorization (26; 30). A structured extraction framework was developed to ensure consistency across heterogeneous sources. Extracted variables included publication metadata, system architecture, planning mechanisms, memory design, tool-access capability, feedback structure, benchmark environment, evaluation metrics, application domain, governance controls, reproducibility artifacts, and reported limitations. Additional variables related to operational autonomy, permission scope, and safety controls were also recorded because these characteristics directly influence the governance profile of agentic systems (36; 39). Evidence was subsequently organized into four primary synthesis categories: architectural frameworks, evaluation benchmarks, application-domain evidence, and governance or risk-management mechanisms. Quality appraisal was adapted according to source type. Benchmark studies were evaluated based on task realism, reproducibility, validation methodology, human comparison, and scoring transparency. Architectural studies were assessed according to system clarity, experimental grounding, failure analysis, and external validity. Governance documents were evaluated based on institutional authority, operational specificity, and practical deployment relevance. Review articles and surveys were additionally assessed for search transparency, taxonomy quality, source coverage, and treatment of preprint literature (1; 40). This flexible appraisal strategy was necessary because the reviewed corpus included both empirical technical systems and governance-oriented guidance documents.

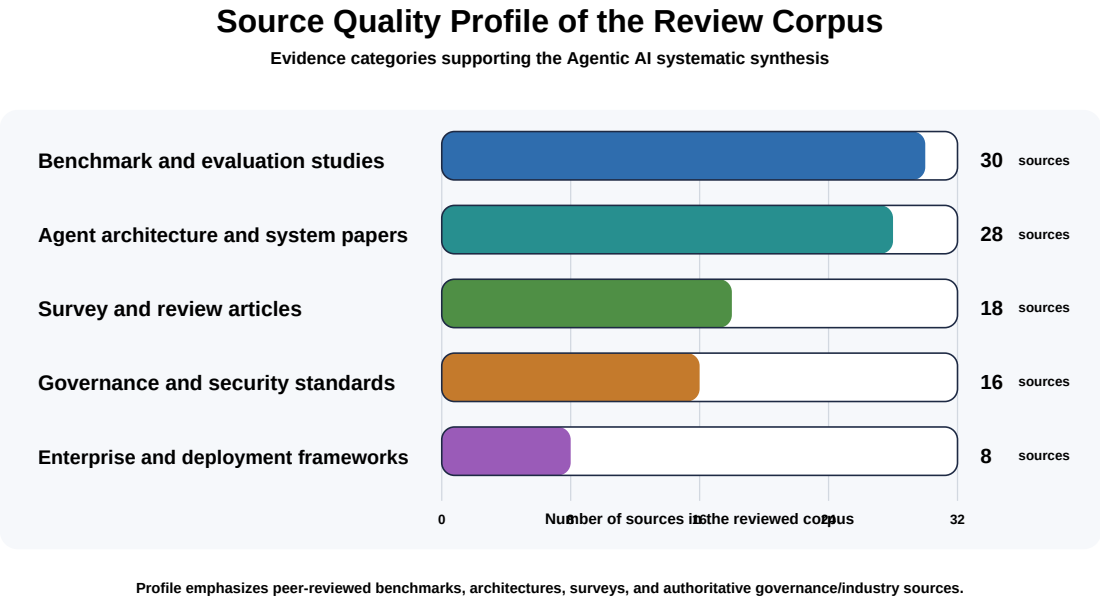

Figure 4: Quality distribution and evidence profile of the literature sources included in this systematic review.

Figure 4 summarizes the quality profile and evidence distribution of the reviewed literature sources, including benchmark studies, architectural frameworks, governance documents, and application-oriented investigations. The final synthesis emphasized comparative interpretation rather than isolated performance reporting. Architectural evidence, benchmark limitations, operational risks, and governance requirements were jointly analyzed to identify recurring patterns across the literature. This synthesis process ultimately supported the development of the proposed Five-Level Agentic Autonomy and Governance Model (FAAGM), which classifies agentic AI systems according to operational authority, environmental interaction, and required oversight mechanisms. The resulting methodology therefore combines systematic evidence collection with governance-oriented analytical synthesis to support a more structured understanding of contemporary agentic AI systems.

3.1 Evidence Coding and Figure Derivation

To ensure that the visual synthesis figures are grounded in structured literature evidence rather than presented as purely evidence-synthesis visualizations, each included source was coded according to publication year, primary contribution type, agentic capability, benchmark setting, autonomy level, tool-access scope, memory persistence, governance-control mechanism, reported safety risk, and deployment implication. The resulting synthesis process was designed to support traceable thematic grouping, comparative evidence integration, and governance-oriented interpretation across heterogeneous agentic-AI studies. Figure 5 was derived from year-by-year thematic grouping of the coded literature corpus. Temporal clustering was supported through structured coding of publication year, benchmark emphasis, governance focus, reasoning architecture, and operational deployment characteristics extracted from the reviewed studies. Each temporal cluster was linked to representative studies emphasizing reasoning frameworks, tool-mediated execution, multi-agent coordination, benchmark environments, governance mechanisms, and deployment-oriented operational systems. The figure therefore summarizes dominant thematic emphasis identified across the included studies rather than implying that individual themes existed exclusively within isolated publication years. Figure 6 was generated using structured thematic coding supported by OpenAlex keyword metadata, recurring keyword relationships, and literature-derived co-occurrence mapping extracted from the reviewed corpus. The visualization summarizes thematic prominence and recurring thematic relationships across architecture, evaluation, governance, orchestration, security, and application-oriented discussions rather than exact citation-network frequency. Instead, node prominence reflects thematic recurrence within the reviewed literature, while node links represent manually coded literature-derived thematic associations across architecture, evaluation, governance, security, orchestration, and application-oriented discussions rather than exact co-citation frequencies or measured bibliometric keyword relationships. Figure 7 was generated using OpenAlex keyword-family searches collected through consistent year-specific query procedures. The underlying query logs, keyword families, and normalization procedures are documented in the Supplementary Evidence Appendix and OpenAlex query records. Because overlapping search terms may retrieve partially duplicated

records, raw publication counts were not directly summed. Instead, each keyword family was normalized relative to its maximum yearly value and subsequently averaged to produce a representative indexed publication trend. The figure should therefore be interpreted as a normalized indexed research-activity trend rather than an exact deduplicated publication count from Scopus or Web of Science. Figures 8 and 10 were constructed by mapping agentic-AI operational risks, governance layers, and threat surfaces to documented source categories identified throughout agent-security studies, benchmark analyses, and authoritative governance frameworks, particularly OWASP, NIST, and CISA guidance (35; 36; 37; 39; 38). The resulting hierarchy integrates model risks, context risks, tool-execution risks, permission-escalation risks, and workflow-level systemic risks associated with increasing operational autonomy and environmental interaction. Figure 9 presents the proposed Five-Level Agentic Autonomy and Governance Model (FAAGM), synthesized from recurring differences in operational authority, execution scope, environmental interaction, and oversight requirements identified throughout the reviewed systems (3; 7; 12; 24; 23). The framework was designed to support clearer classification of bounded, supervised, tool-mediated, and increasingly autonomous agentic-AI systems. Figure 11 presents future research directions synthesized from recurring unresolved limitations, benchmark gaps, governance concerns, operational constraints, and deployment challenges identified across the included studies. The resulting roadmap reflects recurring limitations, governance gaps, benchmark constraints, and architectural trends repeatedly identified across the included studies corpus. The roadmap should therefore be interpreted as a literature-derived synthesis of recurring research priorities identified across the reviewed sources rather than as a predictive statistical forecast. Figure 12 presents a qualitative comparative synthesis of nine representative agentic-AI frameworks using a consistent 1–5 rubric applied across five dimensions: autonomy level, memory persistence, tool authority, governance-control evidence, and benchmark evaluation realism. Scores were assigned according to the strongest operational evidence available in the reviewed literature and were intended to support comparative interpretation rather than absolute capability ranking. The detailed scoring rationale and evidence mapping for each framework are provided in the Supplementary Evidence Appendix. Additional evidence mappings supporting Figure 5 thematic clustering, Figure 6 OpenAlex keyword relationships, Figure 7 publication indexing methodology, and Figure 12 qualitative framework scoring are provided in the Supplementary Evidence Appendix.

4 Results and Analysis

Recent studies indicate a transition from conventional language-generation systems toward operationally interactive agentic architectures capable of planning, reasoning, memory management, and tool-mediated execution. Across the included studies, four recurring characteristics consistently appeared within modern agentic AI systems: multi-step reasoning, environmental interaction, persistent or contextual memory, and adaptive feedback-driven behavior (1; 2; 10). Earlier artificial intelligence systems often relied on static prediction pipelines or narrowly constrained symbolic workflows. In contrast, contemporary agentic systems increasingly integrate dynamic reasoning loops that combine language generation with external execution environments, APIs, retrieval systems, and operational feedback (6; 7). The thematic groupings represented in Figure 5 reflect structured coding of the reviewed literature by publication year and dominant architectural contribution. The 2022 cluster captures prompting, reasoning, tool grounding, embodied planning, and early web interaction (4; 5; 50; 54; 55; 56; 70). The 2023 cluster reflects the emergence of tool-mediated agent frameworks, memory and reflection systems, multi-agent coordination, and the first systematic analyses of agent-security risks (6; 7; 8; 22; 23; 24; 47; 48; 46; 33; 34). The 2024 cluster reflects the rapid expansion of benchmark and evaluation environments, including software-engineering, web automation, operating-system, and machine-learning experimentation settings (12; 13; 14; 15; 16; 27; 28; 31; 32). The 2025–2026 cluster reflects the consolidation of governance frameworks, deployment guidance, enterprise orchestration platforms, and healthcare-domain investigations (35; 36; 37; 39; 38; 61; 63; 11; 71). Figure 5 illustrates the evolution of major research themes within the reviewed agentic AI literature, including the transition from foundational language modeling toward reasoning frameworks, autonomous execution, benchmark evaluation, governance mechanisms, tool interaction, and operational safety. The figure was synthesized from publication chronology, dominant architectural contributions, benchmark evolution, and governance-oriented developments identified across the included studies. The architectural literature indicates that agency emerges not from the language model alone, but from the surrounding orchestration structure that connects planning, memory, execution, and verification layers (21; 26). ReAct introduced one of the earliest influential frameworks that combined intermediate reasoning traces with executable actions during task completion (6). Subsequent systems expanded this paradigm considerably. Toolformer indicated self-supervised tool utilization through API interaction (7), while Gorilla and ToolLLM extended tool-use capability toward large-scale real-world API ecosystems (8; 9). Reflexion introduced verbal self-feedback and episodic memory mechanisms that enabled iterative behavioral improvement without direct parameter updates (22). Similarly, Generative Agents and Voyager indicated the growing importance of persistent memory, environmental adaptation, and autonomous exploration in long-horizon interactive tasks (46; 23). Figure 6 summarizes the semantic relationships among major research themes within the reviewed literature, including autonomy, reasoning, governance, multi-agent systems, benchmark evaluation,

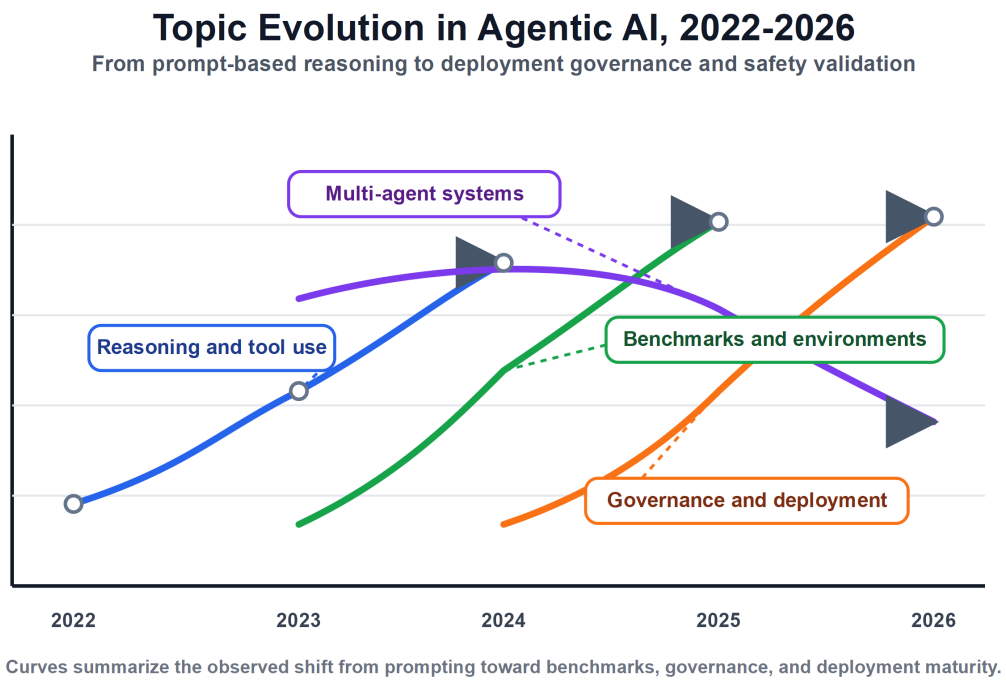

Figure 5: Evolution of major research themes and operational priorities within the agentic artificial intelligence literature ecosystem.

and tool interaction. The co-occurrence relationships were synthesized from recurring thematic associations identified across the included studies corpus and supported by OpenAlex keyword/topic metadata extraction and co-occurrence mapping. The visualization represents thematic prominence and recurring thematic relationships across the reviewed corpus rather than exact Scopus or Web-of-Science bibliometric frequency. An important observation across the literature is that reasoning quality remains one of the primary limitations of current agentic systems despite significant architectural progress. Studies involving chain-of-thought prompting, Tree-of-Thoughts reasoning, and hierarchical planning indicated improved performance in sequential decision-making tasks compared with conventional single-prompt generation approaches (5; 51; 52). However, benchmark evidence consistently summarized that current systems still struggle with long-horizon consistency, contextual grounding, recovery from unexpected states, and reliable multi-step planning (27; 32). This limitation appeared repeatedly across web environments, software engineering benchmarks, and multimodal operational tasks. Figure 7 summarizes the rapid increase in agentic artificial intelligence research activity between 2022 and 2026, reflecting growing interest in autonomous reasoning systems, tool-mediated execution, governance, and operational AI agents. Evaluation-oriented studies provided some of the strongest empirical evidence within the reviewed literature. AgentBench evaluated interactive behavior across multiple operational environments and reported substantial performance variability depending on task structure and environmental complexity (27). WebArena indicated that even advanced language-model agents remained below human-level performance in realistic browser-based workflows involving navigation, form interaction, and sequential decision-making (14). SWE-bench further illustrated the difficulty of real-world software engineering tasks by evaluating repository-level issue resolution using executable testing environments (12). Similarly, OSWorld highlighted persistent weaknesses in graphical interface grounding, multimodal interaction, and desktop-level operational reasoning (28). Collectively, these benchmarks suggest that current agentic AI systems remain significantly more reliable in constrained and highly verifiable environments than in open-ended real-world workflows. The reviewed studies additionally identified a growing shift toward multi-agent coordination architectures. Related collaborative frameworks such as Communicative Agents and DEPS additionally explored task decomposition, interactive planning, and cooperative software-development workflows through structured agent collaboration (83; 84). Frameworks such as AutoGen, CAMEL, MetaGPT, and AgentVerse explored collaborative structures involving planner-reviewer-executor workflows and role-specialized agent interaction (24; 47; 25; 48). Multi-agent systems were frequently proposed as a mechanism for improving task decomposition, iterative critique, scalability, and verification. Nevertheless, several studies also identified important limitations associated with coordination overhead, communication ambiguity, recursive error propagation, and accountability fragmentation within collaborative agent ecosystems (40). These findings suggest that multi-agent scaling alone does not fully resolve the underlying

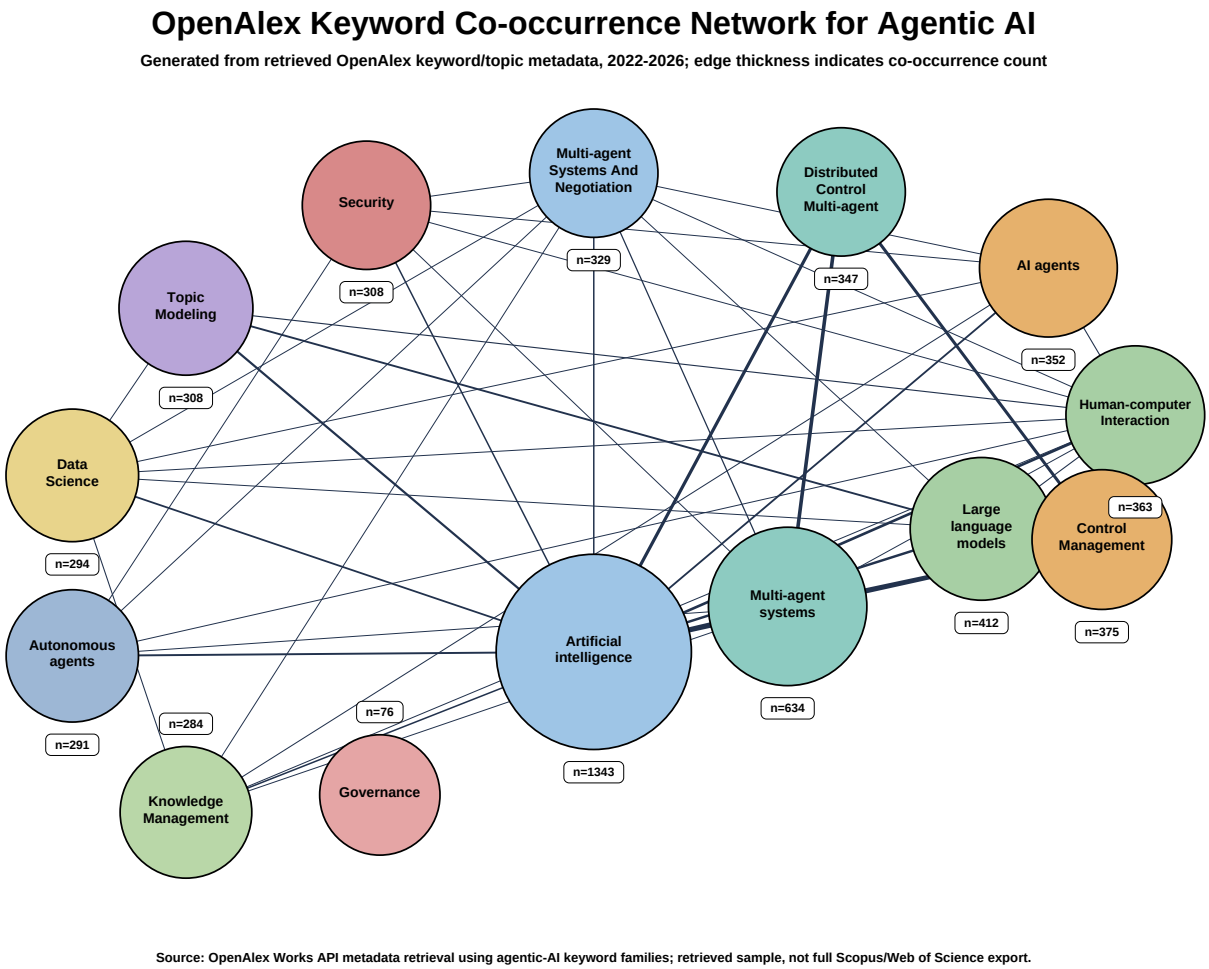

Figure 6: Thematic synthesis of major strengths, limitations, application domains, and governance risks identified across the reviewed agentic AI literature. The figure illustrates recurring evidence categories extracted through structured literature coding and is intended as a qualitative synthesis of review findings rather than a quantitative measurement.

reasoning and reliability limitations of current systems. Application-domain analysis summarized substantial variation in evidence maturity across operational settings. Software engineering and web automation currently represent the strongest application domains because outputs can often be validated through executable testing, interface-state verification, or deterministic evaluation scripts (12; 16). In contrast, healthcare and finance remain more constrained because operational failures may directly affect patient safety, legal compliance, financial risk, or organizational accountability (17; 18; 73). Several healthcare-oriented studies explored biomedical reasoning, documentation support, and clinical summarization using LLM-based agents (70; 71). However, most evaluations remained limited to simulated or exploratory settings rather than large-scale real-world deployment. Additional multimodal and visual interaction systems such as Visual ChatGPT and HuggingGPT further indicated the growing integration of heterogeneous foundation models within coordinated agentic workflows (85; 86). Financial-domain studies similarly indicated strong potential for analytical assistance and document interpretation, while simultaneously revealing substantial concerns regarding hallucination, numerical reliability, and regulatory oversight (72; 73). The reviewed literature also indicated that governance and operational security have become central concerns within modern agentic AI research. Unlike traditional conversational systems that mainly generate text outputs, agentic systems increasingly possess operational permissions capable of triggering external actions through APIs, databases, file systems, browsers, and execution environments (33; 34). Consequently, the reviewed governance literature consistently emphasized the importance of least privilege, sandboxing, monitoring, rollback mechanisms, human approval gates, and auditability (37; 38; 35). OWASP and related security frameworks identified several recurring risks specific to agentic systems, including prompt injection, memory poisoning, insecure tool interaction, privilege escalation, excessive agency, and cascading

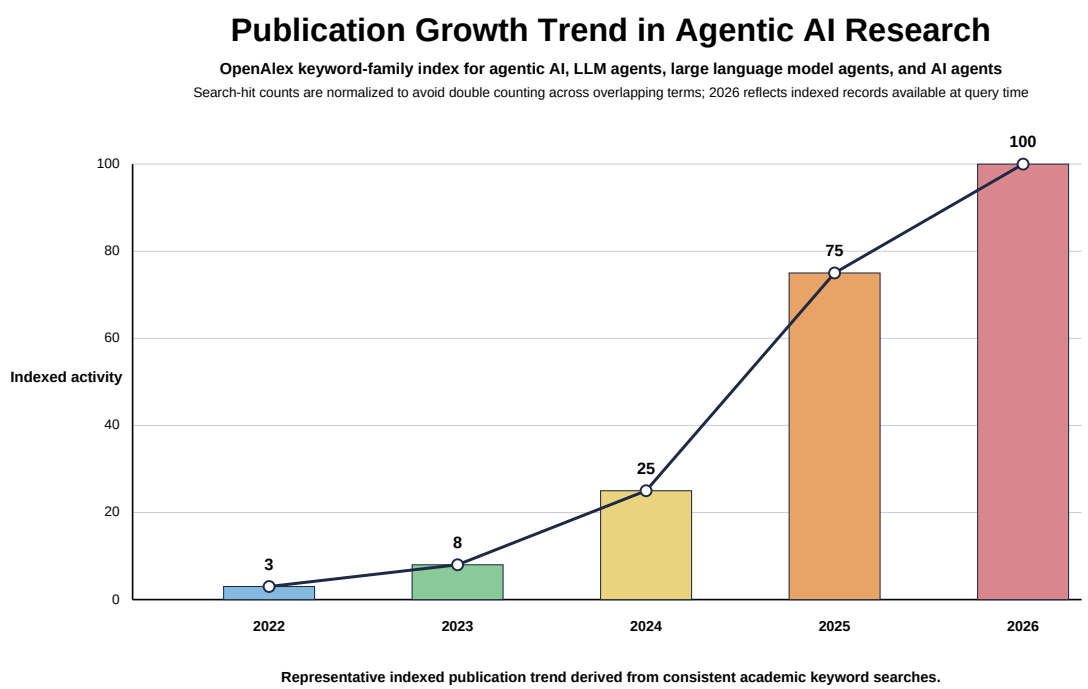

Figure 7: Growth trend of agentic artificial intelligence publications between 2022 and 2026 across the reviewed literature ecosystem.

workflow failures (36). These findings indicate that governance challenges increasingly emerge not from isolated model outputs alone, but from the interaction between autonomous reasoning systems and operational environments. Figure 8 summarizes the increasing operational, governance, and cybersecurity risks associated with higher levels of autonomous execution and environmental interaction within agentic AI systems. Another important finding across the literature is the absence of consistent autonomy classification frameworks. The hierarchical risk structure was synthesized from recurring operational, security, governance, and multi-agent risks discussed throughout the reviewed studies and security frameworks. Also, the risk hierarchy presented in Figure 8 was constructed by mapping each risk layer to representative literature. Model risks (hallucination, weak reasoning, and planning failure) are documented in foundation-model and reasoning-limit studies (3; 4; 5; 29; 77; 78). Context risks (prompt injection and memory contamination) are grounded in direct and indirect injection studies (33; 34; 22; 23; 46). Tool risks are supported by tool-use architectures and OWASP agentic-risk guidance (7; 8; 9; 35; 36). Permission and workflow risks are addressed by OWASP, NIST, and CISA frameworks that specify least-privilege, approval-gate, monitoring, and rollback controls (35; 36; 37; 39; 38). Several studies used the term “agent” broadly despite substantial differences in operational authority, environmental access, and execution capability (2; 11). Some systems function primarily as recommendation assistants, while others perform bounded execution or collaborative multi-agent orchestration. This variation complicates evaluation, governance, and deployment discussions because systems with fundamentally different operational risks are often grouped under the same classification category. The synthesis conducted in this review therefore supports the need for structured autonomy taxonomies that align technical capability with governance requirements and oversight mechanisms. The review additionally identified several persistent research gaps. Current benchmark environments frequently simplify organizational constraints, user interruption, regulatory boundaries, and operational accountability (26; 30). Benchmark contamination and reproducibility concerns also remain significant because many public evaluation environments may partially overlap with training data or public solution repositories. Furthermore, several influential systems remain evaluated primarily through demonstrations or simulation-oriented tasks rather than long-term deployment studies. As a result, substantial uncertainty remains regarding real-world reliability, safety, scalability, and operational governance under realistic conditions. Overall, the reviewed findings indicate that agentic AI should currently be understood as a spectrum of partially autonomous and tool-mediated computational systems rather than fully independent intelligence. Existing systems demonstrate practical capability improvements in reasoning, planning, and operational interaction compared with earlier conversational architectures.

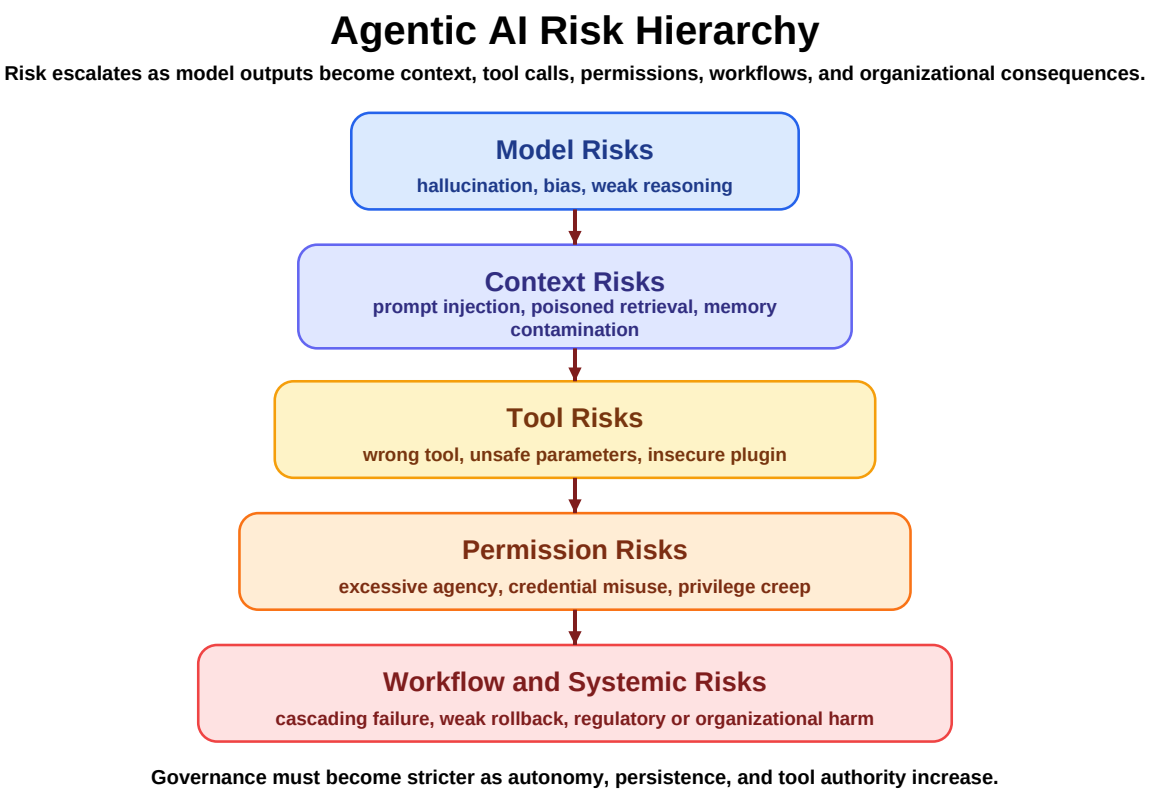

Figure 8: Hierarchical operational and governance risks associated with increasing levels of agentic AI autonomy.

However, benchmark evidence, governance analysis, and deployment limitations collectively indicate that bounded, monitored, and verifiable autonomy remains more realistic than unrestricted open-ended autonomy under current technological conditions.

5 Discussion

The findings of this review indicate that agentic artificial intelligence represents a significant architectural transition from passive language-generation systems toward operationally interactive computational frameworks capable of planning, memory management, environmental interaction, and tool-mediated execution (1; 2). However, the evidence also suggests that current agentic systems should not be interpreted as fully autonomous or generally reliable intelligence. Instead, contemporary systems operate more accurately as partially autonomous workflows whose effectiveness depends heavily on environmental constraints, verification mechanisms, human oversight, and governance structure (26; 30). One of the clearest patterns observed across the reviewed literature is the growing separation between language-generation capability and operational reliability. Large language models demonstrate increasingly strong reasoning, summarization, retrieval, and planning capabilities under controlled conditions (3; 5). Nevertheless, evaluation-oriented studies repeatedly summarized that performance degrades considerably in long-horizon tasks involving environmental uncertainty, sequential decision chains, multimodal interaction, or dynamic feedback loops (27; 14; 28). The reviewed studies indicate that modern agentic systems increasingly rely on multiple architectural components. Several benchmark studies further indicated that even highly capable models remain vulnerable to instruction drift, memory inconsistency, incorrect tool selection, and failure propagation across extended operational trajectories (32; 29).

The reviewed architectural literature additionally suggests that agency emerges primarily from orchestration structure rather than from the language model in isolation (21). Frameworks such as ReAct, Toolformer, Reflexion, Voyager, and ToolLLM demonstrate that memory systems, execution environments, retrieval modules, and tool interfaces collectively shape agentic behavior (6; 7; 22; 23; 9). Consequently, two systems using the same underlying language model may exhibit different operational behavior depending on planning constraints, permission scope, feedback mechanisms, and execution architecture. This observation indicates that future evaluation methodologies should assess complete operational workflows rather than language models alone. Another important observation across the reviewed studies is

that bounded operational environments currently produce the most reliable deployment outcomes. Software engineering benchmarks such as SWE-bench and controlled web interaction environments such as WebArena provide deterministic or partially verifiable execution conditions that enable more reliable validation (12; 14). In contrast, open-ended real-world domains involving healthcare, finance, public services, or organizational decision-making remain more difficult because environmental ambiguity, regulatory constraints, privacy requirements, and human safety considerations cannot easily be reduced to benchmark-style evaluation metrics (17; 73). This difference indicates that near-term deployment of agentic AI will likely remain strongest in constrained environments where verification, rollback, monitoring, and auditability are technically feasible. The evidence further indicates that governance and operational security should be treated as core architectural requirements rather than secondary deployment considerations. Traditional conversational AI systems primarily generate informational outputs, whereas agentic systems increasingly possess operational authority through APIs, execution environments, browsers, databases, and software-control interfaces (33; 34). As a result, model failures may directly produce organizational, financial, cybersecurity, or safety-related consequences. Oversight frameworks proposed by NIST, OWASP, CISA, and related organizations therefore emphasize principles such as least privilege, approval gates, sandboxing, auditability, rollback capability, and continuous monitoring (37; 35; 36; 38). The reviewed studies consistently indicate that unrestricted autonomy currently introduces greater operational risk than bounded supervised autonomy.

The growing interest in multi-agent architectures also revealed several important trade-offs within the literature. Multi-agent systems attempt to improve decomposition, verification, specialization, and scalability through planner-reviewer-executor workflows or collaborative reasoning structures (24; 25; 47). However, multiple studies identified coordination overhead, communication ambiguity, accountability fragmentation, and recursive error propagation as major unresolved challenges (40). In several benchmark environments, collaborative agent structures improved decomposition quality while simultaneously increasing computational cost and operational complexity. These findings suggest that multi-agent scaling does not inherently resolve underlying reasoning and reliability limitations. Another important contribution of this review is the recognition that current literature lacks a standardized framework for autonomy classification. Several reviewed studies used the term "agent" broadly despite substantial differences in operational authority, permission scope, and environmental interaction capability (2; 11).

Some systems operate primarily as recommendation assistants, while others possess bounded execution rights or multi-agent orchestration capability. This variation complicates governance discussions because systems with fundamentally different operational risks are often discussed within the same classification category. The Five-Level Agentic Autonomy and Governance Model proposed in this review therefore attempts to align technical capability with operational oversight requirements. The taxonomy emphasizes operational authority rather than perceived intelligence as the primary determinant of governance complexity. The evidence-synthesis figures presented in this review were each derived through a structured literature-coding procedure rather than constructed as schematic illustrations. This coding procedure recorded publication year, primary contribution, agentic capability, benchmark environment, autonomy level, tool-access scope, memory persistence, governance control, and reported safety risk for each included source. The resulting mappings form the evidentiary basis of the visual syntheses and are intended to allow readers and future reviewers to trace each figure category to specific representative sources. Because the field remains heterogeneous in methodology, benchmark environment, and reporting conventions, the figures do not represent exact statistical aggregates but rather structured qualitative synthesis grounded in the reviewed corpus. This approach aligns with established practice for systematic evidence synthesis in emerging and rapidly evolving technical research domains (1; 2; 40).

Figure 9 summarizes the proposed Five-Level Agentic Autonomy and Governance Model (FAAGM), which categorizes agentic AI systems according to operational autonomy, execution authority, environmental interaction capability, and governance requirements. The FAAGM taxonomy was synthesized from recurring differences in operational authority, execution scope, and oversight requirements identified across representative agentic AI systems in the reviewed literature. However, the reviewed benchmark evidence, governance literature, and deployment limitations collectively indicate that controlled, supervised, and auditable autonomy remains more realistic and defensible than unrestricted open-ended autonomy under current technological and operational conditions. The findings additionally suggest that current benchmark methodologies remain insufficient for fully characterizing real-world deployment readiness. Many benchmarks simplify environmental unpredictability, organizational policy constraints, user interruption, authentication requirements, regulatory compliance, and operational liability (26). Several studies also highlighted benchmark contamination and reproducibility concerns because public benchmark tasks may partially overlap with training corpora or publicly accessible solutions (30). Consequently, benchmark success should not be interpreted as direct evidence of robust real-world autonomy. Future evaluation methodologies will likely require stronger emphasis on adversarial testing, longitudinal deployment analysis, human oversight interaction, and operational safety metrics. The reviewed literature collectively supports a broader shift toward hybrid architectures that combine language-model flexibility with verifiable execution constraints (54; 53). Pure language-model autonomy remains challenging to

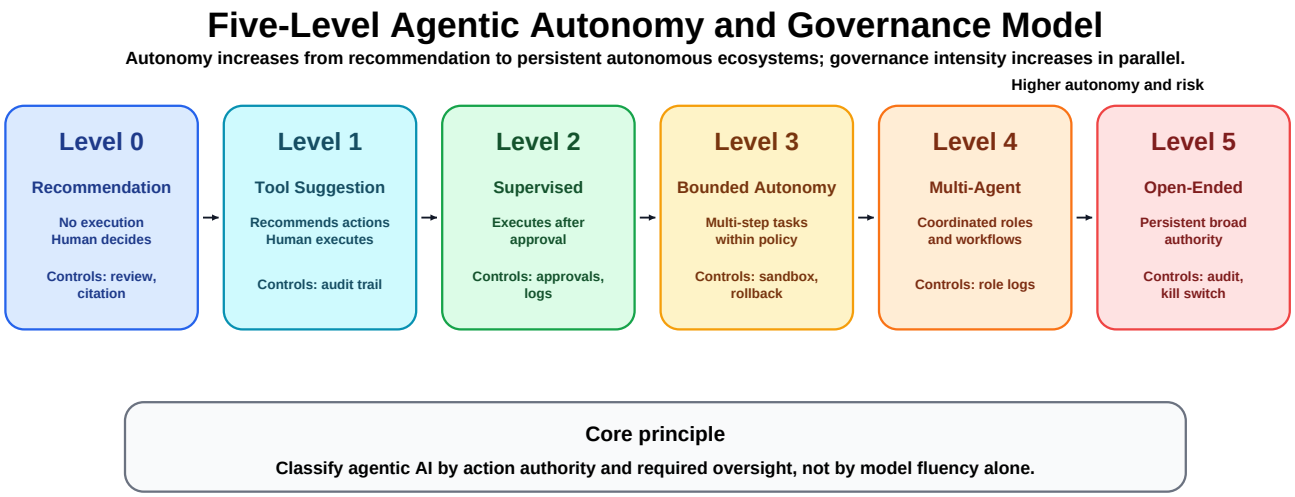

Figure 9: Proposed Five-Level Agentic Autonomy and Governance Model (FAAGM) for classifying operational authority, environmental interaction, and oversight requirements in agentic AI systems.

validate in safety-critical or high-risk operational settings. Hybrid systems incorporating symbolic planners, typed APIs, policy engines, access-control frameworks, and executable verification mechanisms may therefore provide a more realistic deployment pathway for enterprise-scale agentic AI. This interpretation aligns with recent governance guidance emphasizing constrained execution, explicit permissions, reversible actions, and continuous oversight rather than unrestricted autonomous operation (39; 38). Overall, the evidence synthesized in this review suggests that the current trajectory of agentic AI is best understood as the development of increasingly capable but restricted operational computational systems rather than fully independent intelligence. Existing systems already demonstrate practical improvements in reasoning, tool use, and workflow interaction compared with earlier conversational architectures.

Figure 10 summarizes the expanding attack surface of agentic AI systems as autonomy, tool interaction, memory persistence, API access, and multi-agent orchestration become increasingly integrated within operational environments. The threat model was synthesized from recurring attack surfaces, privilege-escalation risks, tool-interaction vulnerabilities, and operational security concerns identified throughout the reviewed governance and security literature. The figure highlights major security and governance concerns including prompt injection, memory poisoning, unsafe tool execution, privilege abuse, role drift, and broader operational harm, together with mitigation mechanisms such as sandboxing, approval gates, logging, monitoring, and rollback control. Future research is expected to increasingly focus on oversight-aware orchestration, auditable autonomy, embodied interaction, neuro-symbolic reasoning, adaptive memory persistence, and secure human-supervised operational control within next-generation agentic artificial intelligence systems.

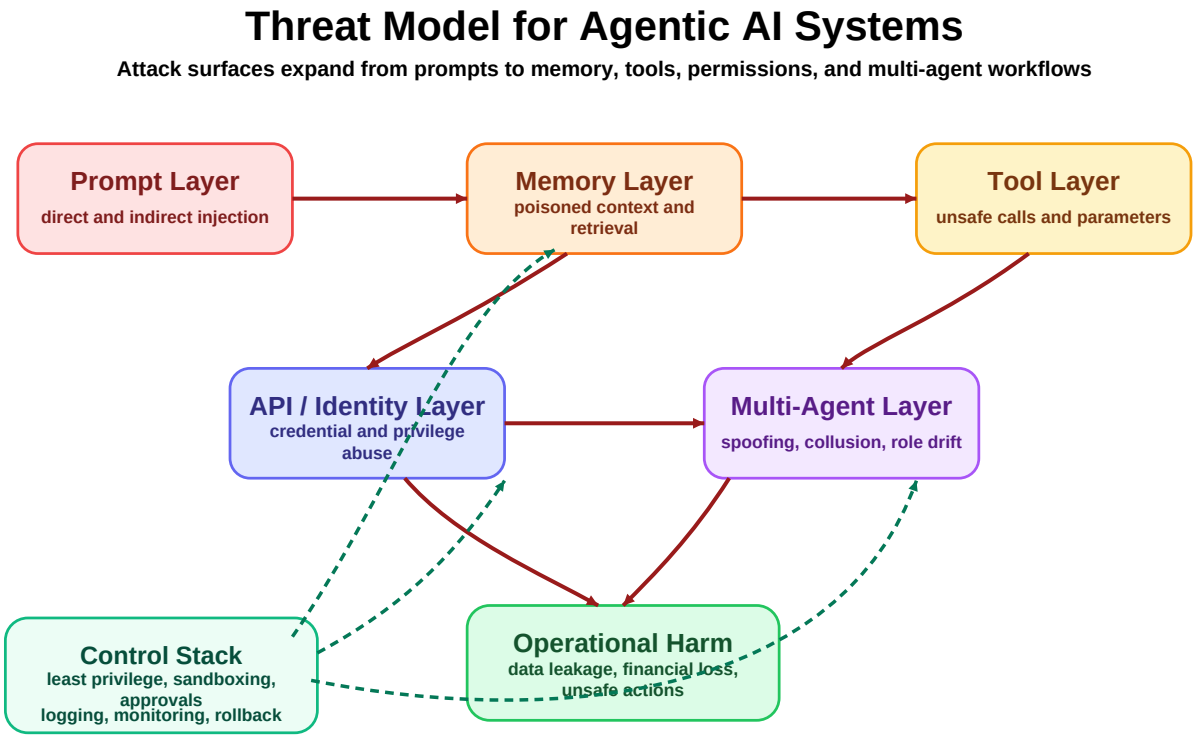

Figure 10: Threat escalation model for agentic AI systems across prompts, memory, tools, permissions, multi-agent workflows, and operational execution environments.

5.1 Future Research Directions and Governance Challenges

Despite rapid progress in agentic artificial intelligence, several important technical, operational, and governance-related challenges remain unresolved throughout the current literature. Existing systems continue to exhibit limitations in long-horizon reasoning, environmental grounding, memory consistency, autonomous recovery, and operational reliability under dynamic real-world conditions (27; 32; 29). In addition, current benchmark environments often simplify organizational constraints, regulatory requirements, human interruption, authentication workflows, and adversarial operational conditions (26; 30). These limitations suggest that future research must increasingly prioritize robust evaluation methodologies, reproducibility, oversight-aware orchestration, and governance-oriented deployment strategies. Another major direction identified across the included studies involves the integration of neuro-symbolic reasoning, adaptive memory persistence, multimodal environmental interaction, and secure tool-mediated execution (54; 53). Several recent studies also emphasize the importance of bounded autonomy, human-supervised operational control, policy-aware orchestration, and auditable execution mechanisms for real-world deployment (37; 36; 38). Future systems will likely require stronger integration of monitoring frameworks, rollback capability, approval gates, access-control enforcement, and continuous operational validation to support safe deployment in healthcare, finance, public services, and enterprise environments.

Figure 11 summarizes several emerging research directions expected to influence the next generation of agentic artificial intelligence systems, including oversight-aware orchestration, neuro-symbolic reasoning, secure tool interaction, embodied autonomy, adaptive memory persistence, and human-supervised operational safety. The future research directions were synthesized from recurring limitations, benchmark gaps, governance challenges, and emerging architectural trends identified across the included studies.

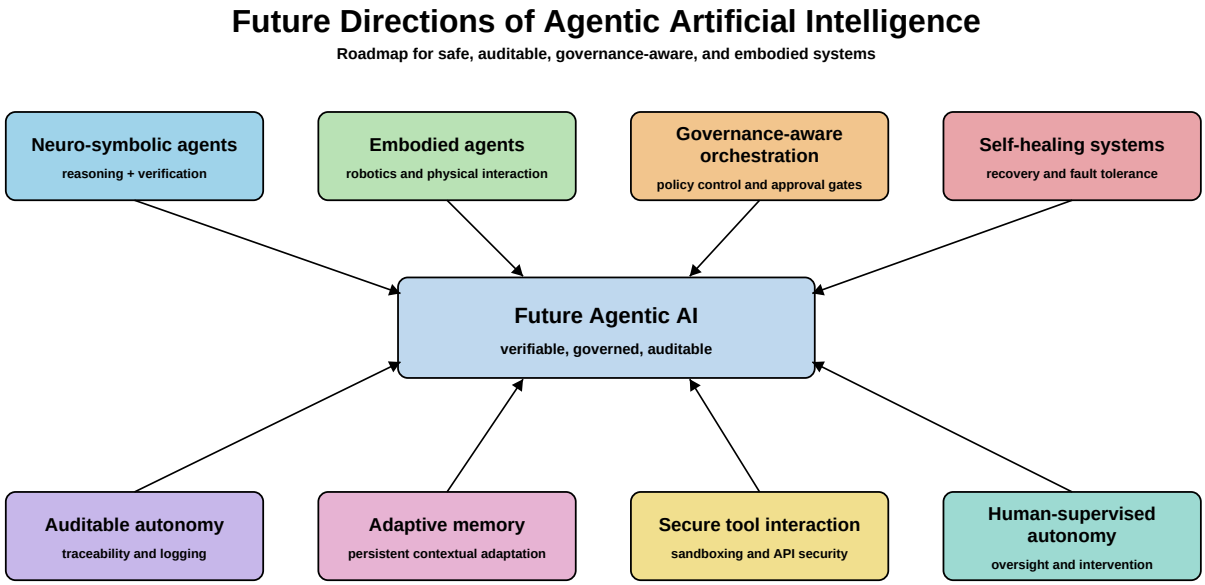

Figure 11: Emerging future research directions in agentic artificial intelligence, including neuro-symbolic reasoning, oversight-aware orchestration, embodied interaction, auditable autonomy, adaptive memory systems, and human-supervised operational control.

6 Conclusion

Agentic artificial intelligence highlights an important transition in the evolution of artificial intelligence systems from passive language-generation models toward operationally interactive architectures capable of planning, memory utilization, tool-mediated execution, and adaptive environmental interaction (1; 2). The reviewed studies indicate that recent advances in large language models, reasoning frameworks, retrieval systems, and execution-oriented workflows have considerably expanded the practical capabilities of AI systems across software engineering, web automation, scientific research, healthcare, finance, and multimodal operational environments (6; 9; 12; 28). The reviewed studies indicate that modern agentic systems increasingly rely on multiple architectural components., including planning mechanisms, memory structures, feedback loops, external tool interfaces, and orchestration layers (21; 26). However, evaluation-oriented studies consistently demonstrate that substantial limitations remain in long-horizon reasoning, environmental grounding, tool selection, memory consistency, and recovery from unexpected operational states (27; 32; 29). Although contemporary systems demonstrate meaningful progress compared with earlier conversational architectures, current evidence does not support the interpretation of agentic AI as fully reliable or unrestricted autonomy. The review additionally highlights that operational governance and security have become central challenges within agentic AI deployment. Because modern agents increasingly interact with APIs, browsers, execution environments, databases, and external workflows, model failures may directly translate into organizational, cybersecurity, financial, or safety-related consequences (33; 34). Oversight frameworks proposed by NIST, OWASP, CISA, and related organizations therefore emphasize principles such as least privilege, sandboxing, auditability, approval gates, rollback mechanisms, and continuous monitoring (37; 35; 36; 38). The reviewed evidence strongly suggests that controlled and supervised autonomy currently highlights a considerably more realistic deployment model than unrestricted open-ended execution. This review also introduced the Five-Level Agentic Autonomy and Governance Model (FAAGM), which classifies agentic systems according to operational authority, environmental interaction capability, and required oversight mechanisms. The proposed framework attempts to address a recurring limitation within the literature, where systems with fundamentally different operational risks are frequently grouped under the same conceptual definition of “agent” (11; 2). By aligning technical capability with oversight requirements, the FAAGM taxonomy provides a more structured basis for evaluating deployment readiness, operational risk, and oversight requirements across heterogeneous agentic architectures.

Figure 12 summarizes representative agentic AI frameworks according to autonomy level, reasoning structure, memory integration, tool interaction capability, and operational limitations.

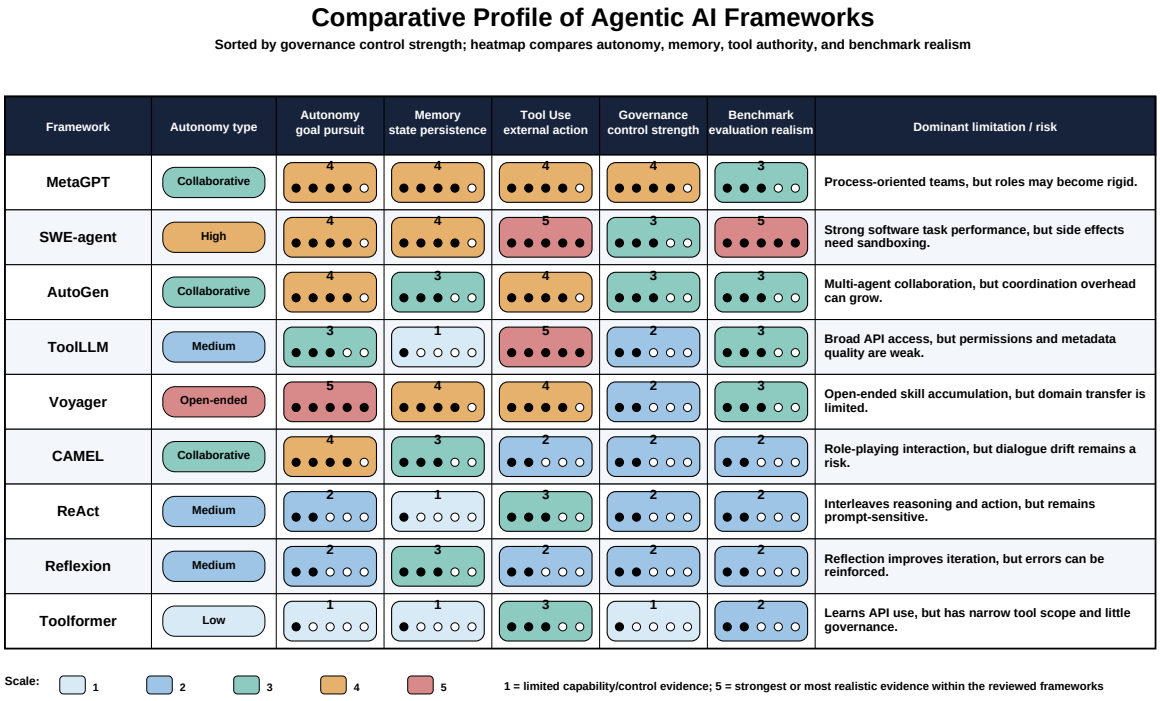

Figure 12: Comparative matrix of representative agentic AI frameworks across autonomy level, reasoning capability, governance support, operational scope, and deployment limitations.

Scoring rubric for Figure 12: Autonomy (1 = no execution authority; 5 = open-ended goal pursuit); Memory (1 = stateless; 5 = persistent cross-session retrieval); Tool use (1 = no external calls; 5 = broad real-world API execution); Governance (1 = no access controls documented; 5 = explicit least-privilege, approval, and audit mechanisms); Benchmark realism (1 = toy tasks only; 5 = executable real-world validation with human comparison). Scores reflect the strongest evidence available in the reviewed literature for each framework; they represent qualitative synthesis and should be interpreted comparatively rather than as absolute capability ratings (12; 13; 24; 25; 9; 23; 47; 6; 22; 7).

Several important research gaps remain unresolved. Current benchmark environments frequently simplify organizational constraints, regulatory requirements, authentication mechanisms, user interruption, and real-world operational uncertainty (26; 30). Benchmark contamination, reproducibility limitations, and inconsistent evaluation methodologies further complicate comparative analysis across studies. In addition, many application-domain investigations remain limited to simulated environments or exploratory deployments rather than long-term real-world operational validation. Future research should therefore prioritize reproducible evaluation methodologies, longitudinal deployment studies, hybrid neuro-symbolic architectures, operational safety metrics, memory-security mechanisms, and governance-aware execution frameworks. Overall, the findings of this review suggest that agentic AI should currently be understood as a spectrum of partially autonomous, tool-mediated, and restricted operational systems rather than fully independent intelligence. Existing systems already demonstrate practical utility in constrained environments where monitoring, verification, rollback, and operational supervision are technically feasible. However, the reviewed evidence collectively indicates that safe and sustainable deployment will depend not only on model capability, but also on governance design, permission boundaries, verification mechanisms, and continuous operational supervision.

Supplementary Evidence Appendix

Evidence Mapping for Figure 5 and Figure 12

This appendix documents the literature evidence underlying Figure 5 and Figure 12. Figure 5 was derived from year-by-year thematic grouping of reviewed studies. Figure 12 was derived from qualitative scoring of representative agentic AI frameworks using a predefined evidence rubric. These materials improve traceability, reproducibility, and methodological transparency of the visual synthesis process.

Coding Procedure

Each source was coded according to publication year, primary contribution type, architectural characteristic, benchmark or evaluation characteristic, governance characteristic, and operational characteristic. Architectural characteristics included reasoning, planning, tool use, memory, orchestration, multi-agent coordination, and embodied interaction. Governance characteristics included human approval, logs, sandboxing, rollback, least privilege, auditability, policy control, and external guidance.

Figure 6 OpenAlex Keyword Network Methodology

Figure 6 was generated using OpenAlex keyword and topic metadata retrieved through structured keyword-family searches associated with agentic AI, AI agents, autonomous agents, large language model agents, orchestration frameworks, governance, and benchmark environments. Node size reflects metadata recurrence frequency, while edge thickness represents pairwise keyword co-occurrence relationships extracted from the retrieved metadata corpus. The visualization is intended as a metadata-supported thematic synthesis rather than a complete Scopus or Web of Science bibliometric export.

Figure 5 Evidence Mapping: Evolution of Major Research Themes

Figure 5 was constructed from thematic coding of reviewed sources by publication year. It summarizes dominant emphasis across the corpus rather than claiming that each year contains only one topic.

2022 — Prompting, reasoning, tool grounding, embodied planning, and early web/task interaction Studies: Ouyang et al. [4]; Wei et al. [5]; Karpas et al. [50]; Ahn et al. [54]; Huang et al. [55]; Yao et al. [56]; Luo et al. [70]; Perez et al. [79]. Architecture: Instruction following; chain-of-thought reasoning; modular neuro-symbolic tool use; language-conditioned planning; embodied reasoning. Benchmark: Early grounded web interaction; embodied/planning tasks; biomedical text-generation evidence. Governance: Mainly capability-focused; governance treated indirectly through alignment, red teaming, and task constraints.

2023 — Tool-mediated agents, memory, reflection, multi-agent coordination, and early security concerns Studies: Yao et al. [6]; Schick et al. [7]; Patil et al. [8]; Xi et al. [10]; Shinn et al. [22]; Wang et al. [23]; Wu et al. [24]; Park et al. [46]; Li et al. [47]; Chen et al. [48]; Greshake et al. [33]; Liu et al. [34]. Architecture: Reasoning-action loops; API/tool selection; reflection; persistent skill libraries; role-based multi-agent workflows; memory mechanisms. Benchmark: Tool tasks; simulated environments; multi-agent demonstrations; early workflow evaluations. Governance: Prompt injection; memory contamination; role drift; early recognition of tool-mediated attack surfaces.

2024 — Benchmark expansion, software/web/OS environments, architecture synthesis, and governance formalization Studies: Wang et al. [1]; Qin et al. [9]; Jimenez et al. [12]; Yang et al. [13]; Zhou et al. [14]; Drouin et al. [15]; de Chezelles et al. [16]; Sumers et al. [21]; Hong et al. [25]; Masterman et al. [26]; Liu et al. [27]; Xie et al. [28]; Huang et al. [31]; Wang et al. [32]; Autio et al. [37]. Architecture: Agent architectures with planning, memory, tool access, orchestration, and workflow execution. Benchmark: SWE-bench; WebArena; WorkArena; BrowserGym; AgentBench; OSWorld; MLAgentBench; MINT. Governance: Reproducibility; task realism; executable feedback; sandboxing; AI risk-management guidance.

2025 — Governance, safety, enterprise deployment, orchestration frameworks, and financial-domain synthesis Studies: Plaat et al. [2]; Evaluation survey [30]; OWASP [35,36]; OpenAI [61]; Google Cloud [63]; IBM [64]; Salesforce [65]; ServiceNow [66]; LangChain [67]; CrewAI [68]; AutoGen Studio [69]; Khan et al. [73]. Architecture: Enterprise agent orchestration; agent platforms; tool workflows; multi-agent automation; domain-specific agents. Benchmark: Deployment-oriented and system-level evaluation rather than only benchmark tasks. Governance: Approval gates; monitoring; policy controls; enterprise governance; security controls; financial-domain oversight.

2026 — Consolidated agentic-AI reviews, healthcare agents, and adoption guidance Studies: Abou Ali et al. [11]; CISA [38]; Healthcare scoping review [71]. Architecture: Consolidated taxonomy of architectures and applications; healthcare-oriented agentic workflows. Benchmark: Review-level and domain-specific evidence synthesis. Governance: Careful adoption guidance; high-impact domain oversight; validation; privacy; operational governance.

Figure 7 Publication Trend Methodology

Figure 7 was generated using normalized OpenAlex keyword-family retrieval trends collected across the 2022–2026 publication window. Overlapping keyword families were normalized to reduce duplicate counting effects across partially overlapping metadata records. The resulting indexed trend reflects representative research-activity growth rather than exact deduplicated publication totals.

Figure 12 Qualitative Scoring Rubric

Figure 12 compares representative agentic AI frameworks using a qualitative 1–5 scoring rubric. The matrix is intended as a comparative synthesis, not as an absolute ranking.

Score 1 Minimal autonomous action; no durable memory; no tools or tool suggestion only; governance absent or not discussed; toy or narrow task setting.

Score 2 Limited reasoning-action loop; short trace or task-local state; optional/restricted tools; manual gates or weak logs; limited QA/tool/web tasks.

Score 3 Bounded multi-step behavior; dialogue/reflection/history; multiple tools or API interface; basic approvals or monitoring; realistic task/domain benchmark.

Score 4 Strong multi-step or collaborative autonomy; reusable artifacts, skills, or project memory; broad tools/code actions/workflow interface; structured process governance; operational benchmark with feedback.

Score 5 Open-ended persistent goal pursuit; durable long-horizon memory; high-authority execution; strong formal controls indicated; executable high-realism validation.

Framework-Level Scoring Evidence for Figure 12

MetaGPT | Autonomy 4 | Memory 4 | Tool use 4 | Governance 4 | Benchmark 3 Rationale: Role-specialized multi-agent software-development process; persistent project artifacts; structured process governance; moderate benchmark realism compared with executable issue-resolution benchmarks. Sources: Hong et al. [25].

SWE-agent | Autonomy 4 | Memory 4 | Tool use 5 | Governance 3 | Benchmark 5 Rationale: Repository-level autonomous workflow; maintains repository context; shell/test/editing authority; sandbox needed; strongest benchmark realism through SWE-bench executable validation. Sources: Jimenez et al. [12]; Yang et al. [13].

AutoGen | Autonomy 4 | Memory 3 | Tool use 4 | Governance 3 | Benchmark 3 Rationale: Multi-agent conversation and configurable tool/function execution; conversation history supports coordination; approvals/configuration provide moderate governance evidence. Sources: Wu et al. [24].

ToolLLM | Autonomy 3 | Memory 1 | Tool use 5 | Governance 2 | Benchmark 3 Rationale: Large-scale API interaction capability; limited durable memory evidence; weak permission/governance treatment; tool-use evaluation gives moderate benchmark realism. Sources: Qin et al. [9].

Voyager | Autonomy 5 | Memory 4 | Tool use 4 | Governance 2 | Benchmark 3 Rationale: Open-ended embodied exploration; persistent skill library; executable skill/tool use; limited governance controls; interactive environment gives moderate realism. Sources: Wang et al. [23].

CAMEL | Autonomy 4 | Memory 3 | Tool use 2 | Governance 2 | Benchmark 2 Rationale: Role-playing multi-agent interaction; dialogue context supports coordination; limited external tool authority; weak governance and benchmark realism. Sources: Li et al. [47].

ReAct | Autonomy 2 | Memory 1 | Tool use 3 | Governance 2 | Benchmark 2 Rationale: Reasoning-action loop; no durable memory; explicit tool/action calls; governance limited to task structure/manual constraints; limited-to-moderate benchmark realism. Sources: Yao et al. [6].

Reflexion | Autonomy 2 | Memory 3 | Tool use 2 | Governance 2 | Benchmark 2 Rationale: Iterative self-feedback and episodic reflection logs; limited tool authority; traceability improves interpretability but not full governance. Sources: Shinn et al. [22].

Toolformer | Autonomy 1 | Memory 1 | Tool use 3 | Governance 1 | Benchmark 2 Rationale: Self-supervised tool-use learning; low autonomy; no durable memory; minimal governance evidence; narrow tool-task evaluation. Sources: Schick et al. [7].

References

[1] Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; et al. A survey on large language model based autonomous agents. *Front. Comput. Sci.* **2024**, *18*, 186345. https://doi.org/10.1007/s11704-024-40231-1.

[2] Plaat, A.; van Duijn, M.; van Stein, N.; Preuss, M.; van der Putten, P.; Batenburg, K.J. Agentic Large Language Models, a survey. *arXiv* **2025**, arXiv:2503.23037. https://arxiv.org/abs/2503.23037.

[3] Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901. https://arxiv.org/abs/2005.14165.

[4] Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27730–27744. https://arxiv.org/abs/2203.02155.

[5] Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; et al. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837. https://arxiv.org/abs/2201.11903.

[6] Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.R.; Cao, Y. ReAct: Synergizing reasoning and acting in language models. *International Conference on Learning Representations* **2023**. https://arxiv.org/abs/2210.03629.

[7] Schick, T.; Dwivedi-Yu, J.; Dessi, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; Scialom, T. Toolformer: Language models can teach themselves to use tools. *Adv. Neural Inf. Process. Syst.* **2023**, *36*. https://arxiv.org/abs/2302.04761.

[8] Patil, S.G.; Zhang, T.; Wang, X.; Gonzalez, J.E. Gorilla: Large language model connected with massive APIs. *arXiv* **2023**, arXiv:2305.15334. https://arxiv.org/abs/2305.15334.

[9] Qin, Y.; Liang, S.; Ye, Y.; Zhu, K.; Yan, L.; Lu, Y.; et al. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. *International Conference on Learning Representations* **2024**. https://arxiv.org/abs/2307.16789.

[10] Xi, Z.; Chen, W.; Guo, X.; He, W.; Ding, Y.; Hong, B.; et al. The rise and potential of large language model based agents: A survey. *arXiv* **2023**, arXiv:2309.07864. https://arxiv.org/abs/2309.07864.

[11] Abou Ali, M.; Dornaika, F.; Charafeddine, J. Agentic AI: A comprehensive survey of architectures, applications, and future directions. *Artif. Intell. Rev.* **2026**, *59*, 11. https://doi.org/10.1007/s10462-025-11422-4.

[12] Jimenez, C.E.; Yang, J.; Wettig, A.; Yao, S.; Pei, K.; Press, O.; Narasimhan, K.R. SWE-bench: Can language models resolve real-world GitHub issues? *International Conference on Learning Representations* **2024**. https://arxiv.org/abs/2310.06770.

[13] Yang, J.; Jimenez, C.E.; Wettig, A.; Lieret, K.; Yao, S.; Narasimhan, K.R.; Press, O. SWE-agent: Agent-computer interfaces enable automated software engineering. *arXiv* **2024**, arXiv:2405.15793. https://arxiv.org/abs/2405.15793.

[14] Zhou, S.; Xu, F.; Zhu, H.; Zhou, X.; Lo, K.; Sridhar, A.; et al. WebArena: A realistic web environment for building autonomous agents. *International Conference on Learning Representations* **2024**. https://arxiv.org/abs/2307.13854.

[15] Drouin, A.; Gasse, M.; Caccia, M.; Laradji, I.; Del Verme, M.; Marty, T.; et al. WorkArena: How capable are web agents at solving common knowledge work tasks? *arXiv* **2024**, arXiv:2403.07718. https://arxiv.org/abs/2403.07718.

[16] de Chezelles, T.; Gasse, M.; Drouin, A.; Caccia, M.; Laradji, I.; Chapados, N.; et al. BrowserGym: A gym environment for web task automation. *arXiv* **2024**, arXiv:2407.07499. https://arxiv.org/abs/2407.07499.

[17] Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S.S.; Wei, J.; Chung, H.W.; et al. Large language models encode clinical knowledge. *Nature* **2023**, *620*, 172–180. https://doi.org/10.1038/s41586-023-06291-2.

[18] Wu, S.; Irsoy, O.; Lu, S.; Dabravolski, V.; Dredze, M.; Gehrmann, S.; et al. BloombergGPT: A large language model for finance. *arXiv* **2023**, arXiv:2303.17564. https://arxiv.org/abs/2303.17564.

[19] Boiko, D.A.; MacKnight, R.; Kline, B.; Gomes, G. Autonomous chemical research with large language models. *Nature* **2023**, *624*, 570–578. https://doi.org/10.1038/s41586-023-06792-0.

[20] Driess, D.; Xia, F.; Sajjadi, M.S.M.; Lynch, C.; Chowdhery, A.; Ichter, B.; et al. PaLM-E: An embodied multimodal language model. *International Conference on Machine Learning* **2023**. https://arxiv.org/abs/2303.03378.

[21] Sumers, T.R.; Yao, S.; Narasimhan, K.R.; Griffiths, T.L. Cognitive architectures for language agents. *Trans. Mach. Learn. Res.* **2024**. https://arxiv.org/abs/2309.02427.

[22] Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.R.; Yao, S. Reflexion: Language agents with verbal reinforcement learning. *Adv. Neural Inf. Process. Syst.* **2023**. https://arxiv.org/abs/2303.11366.

[23] Wang, G.; Xie, Y.; Jiang, Y.; Mandlekar, A.; Xiao, C.; Zhu, Y.; Fan, L.; Anandkumar, A. Voyager: An open-ended embodied agent with large language models. *arXiv* **2023**, arXiv:2305.16291. https://arxiv.org/abs/2305.16291.

[24] Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E.; et al. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv* **2023**, arXiv:2308.08155. https://arxiv.org/abs/2308.08155.

[25] Hong, S.; Zheng, X.; Chen, J.; Cheng, Y.; Wang, C.; Zhang, Z.; et al. MetaGPT: Meta programming for a multi-agent collaborative framework. *International Conference on Learning Representations* **2024**. https://arxiv.org/abs/2308.00352.

[26] Masterman, T.; Besen, S.; Sawtell, M.; Chao, A. The landscape of emerging AI agent architectures for reasoning, planning, and tool calling: A survey. *arXiv* **2024**, arXiv:2404.11584. https://arxiv.org/abs/2404.11584.

[27] Liu, X.; Yu, H.; Zhang, H.; Xu, Y.; Lei, X.; Lai, H.; et al. AgentBench: Evaluating LLMs as agents. *International Conference on Learning Representations* **2024**. https://arxiv.org/abs/2308.03688.

[28] Xie, T.; Zhang, D.; Chen, J.; Li, X.; Zhao, S.; Cao, R.; et al. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Adv. Neural Inf. Process. Syst.* **2024**. https://arxiv.org/abs/2404.07972.

[29] Valmeekam, K.; Marquez, M.; Sreedharan, S.; Kambhampati, S. On the planning abilities of large language models: A critical investigation. *Adv. Neural Inf. Process. Syst.* **2023**, *36*. https://arxiv.org/abs/2305.15771.

[30] Anonymous. Survey on Evaluation of LLM-based Agents. *arXiv* **2025**, arXiv:2503.16416. https://arxiv.org/abs/2503.16416.

[31] Huang, Q.; Vora, J.; Liang, P.; Leskovec, J.; Kolter, J.Z.; Heller, K.; et al. MLAgentBench: Evaluating language agents on machine learning experimentation. *Proc. Mach. Learn. Res.* **2024**, *235*, 20271–20309. https://proceedings.mlr.press/v235/huang24e.html.

[32] Wang, B.; Deng, X.; Sun, H. MINT: Evaluating LLMs in multi-turn interaction with tools and language feedback. *International Conference on Learning Representations* **2024**. https://arxiv.org/abs/2309.10691.

[33] Greshake, K.; Abdelnabi, S.; Mishra, S.; Endres, C.; Holz, T.; Fritz, M. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. *ACM Workshop on Artificial Intelligence and Security* **2023**. https://doi.org/10.1145/3605764.3623985.

[34] Liu, Y.; Deng, G.; Li, Y.; Wang, K.; Zhang, Z.; Wang, X.; et al. Prompt injection attack against LLM-integrated applications. *arXiv* **2023**, arXiv:2306.05499. https://arxiv.org/abs/2306.05499.

[35] OWASP Foundation. OWASP Top 10 for Large Language Model Applications. *OWASP GenAI Security Project* **2025**. https://genai.owasp.org/llm-top-10/.

[36] OWASP GenAI Security Project. OWASP Top 10 for Agentic Applications. *OWASP GenAI Security Project* **2025**. https://genai.owasp.org/.

[37] Autio, C.; Schwartz, R.; Dunietz, J.; Jain, S.; Stanley, M.; Tabassi, E.; Hall, P.; Roberts, K. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. *NIST Trustworthy and Responsible AI 600-1* **2024**. https://doi.org/10.6028/NIST.AI.600-1.

[38] Cybersecurity and Infrastructure Security Agency. Careful Adoption of Agentic AI Services. *CISA* **2026**. https://www.cisa.gov/.

[39] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework. *NIST AI RMF 1.0* **2023**. https://doi.org/10.6028/NIST.AI.100-1.

[40] Guo, T.; Chen, X.; Wang, Y.; Chang, R.; Pei, S.; Chawla, N.V.; Wiest, O.; Zhang, X. Large Language Model Based Multi-agents: A Survey of Progress and Challenges. *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence* **2024**, 8048–8057. https://doi.org/10.24963/ijcai.2024/890.

[41] Russell, S.; Norvig, P. Artificial Intelligence: A Modern Approach, 4th ed. *Pearson* **2021**. https://aima.cs.berkeley.edu/.

[42] Wooldridge, M. An Introduction to MultiAgent Systems, 2nd ed. *Wiley* **2009**. https://www.wiley.com/en-us/An+Introduction+to+MultiAgent+Systems%2C+2nd+Edition-p-9780470519462.

[43] Sutton, R.S.; Barto, A.G. Reinforcement Learning: An Introduction, 2nd ed. *MIT Press* **2018**. http://incompleteideas.net/book/the-book-2nd.html.

[44] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. https://arxiv.org/abs/1706.03762.

[45] Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; et al. Self-consistency improves chain of thought reasoning in language models. *International Conference on Learning Representations* **2023**. https://arxiv.org/abs/2203.11171.

[46] Park, J.S.; O'Brien, J.; Cai, C.J.; Morris, M.R.; Liang, P.; Bernstein, M.S. Generative Agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* **2023**. https://doi.org/10.1145/3586183.3606763.

[47] Li, G.; Hammoud, H.A.A.K.; Itani, H.; Khizbullin, D.; Ghanem, B. CAMEL: Communicative agents for mind exploration of large language model society. *Adv. Neural Inf. Process. Syst.* **2023**. https://arxiv.org/abs/2303.17760.

[48] Chen, W.; Su, Y.; Zuo, J.; Yang, C.; Yuan, C.; Qian, C.; et al. AgentVerse: Facilitating multi-agent collaboration and exploring emergent behaviors. *arXiv* **2023**, arXiv:2308.10848. https://arxiv.org/abs/2308.10848.

[49] Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; et al. Self-Refine: Iterative refinement with self-feedback. *Adv. Neural Inf. Process. Syst.* **2023**, *36*. https://arxiv.org/abs/2303.17651.

[50] Karpas, E.; Abend, O.; Belinkov, Y.; Lenz, B.; Lieber, O.; Ratner, N.; et al. MRKL systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning. *arXiv* **2022**, arXiv:2205.00445. https://arxiv.org/abs/2205.00445.

[51] Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.L.; Cao, Y.; Narasimhan, K.R. Tree of thoughts: Deliberate problem solving with large language models. *Adv. Neural Inf. Process. Syst.* **2023**, *36*. https://arxiv.org/abs/2305.10601.

[52] Zhou, D.; Scharli, N.; Hou, L.; Wei, J.; Scales, N.; Wang, X.; et al. Least-to-most prompting enables complex reasoning in large language models. *International Conference on Learning Representations* **2023**. https://arxiv.org/abs/2205.10625.

[53] Liu, B.; Jiang, Y.; Zhang, X.; Liu, Q.; Zhang, S.; Biswas, J.; Stone, P. LLM+P: Empowering large language models with optimal planning proficiency. *arXiv* **2023**, arXiv:2304.11477. https://arxiv.org/abs/2304.11477.

[54] Ahn, M.; Brohan, A.; Brown, N.; Chebotar, Y.; Cortes, O.; David, B.; et al. Do as I can, not as I say: Grounding language in robotic affordances. *Conference on Robot Learning* **2022**. https://arxiv.org/abs/2204.01691.

[55] Huang, W.; Xia, F.; Xiao, T.; Chan, H.; Liang, J.; Florence, P.; et al. Inner monologue: Embodied reasoning through planning with language models. *Conference on Robot Learning* **2022**. https://arxiv.org/abs/2207.05608.

[56] Yao, S.; Chen, H.; Yang, J.; Narasimhan, K.R. WebShop: Towards scalable real-world web interaction with grounded language agents. *Adv. Neural Inf. Process. Syst.* **2022**, *35*. https://arxiv.org/abs/2207.01206.

[57] Shridhar, M.; Yuan, X.; Cote, M.A.; Bisk, Y.; Trischler, A.; Hausknecht, M. ALFWorld: Aligning text and embodied environments for interactive learning. *International Conference on Learning Representations* **2021**. https://arxiv.org/abs/2010.03768.

[58] Chevalier-Boisvert, M.; Bahdanau, D.; Lahlou, S.; Willems, L.; Saharia, C.; Nguyen, T.; Bengio, Y. BabyAI: First steps towards grounded language learning with a human in the loop. *International Conference on Learning Representations* **2019**. https://arxiv.org/abs/1810.08272.

[59] Shi, T.; Karpathy, A.; Fan, L.; Hernandez, J.; Liang, P. World of Bits: An open-domain platform for web-based agents. *International Conference on Machine Learning* **2017**. https://proceedings.mlr.press/v70/shi17a.html.

[60] Microsoft. Microsoft Copilot Studio: Building copilots with agent capabilities. *Microsoft Copilot Blog* **2024**. https://www.microsoft.com/en-us/microsoft-copilot/blog/copilot-studio/microsoft-copilot-studio-building-copilots-with-agent-capabilities/.

[61] OpenAI. Introducing Operator. *OpenAI* **2025**. https://openai.com/index/introducing-operator/.

[62] Anthropic. Introducing computer use, a new Claude 3.5 Sonnet, and Claude 3.5 Haiku. *Anthropic* **2024**. https://www.anthropic.com/news/3-5-models-and-computer-use.

[63] Google Cloud. Google Agentspace. *Google Cloud* **2025**. https://cloud.google.com/products/agentspace.

[64] IBM. IBM watsonx Orchestrate. *IBM* **2025**. https://www.ibm.com/products/watsonx-orchestrate.

[65] Salesforce. Agentforce: The digital labor platform. *Salesforce* **2025**. https://www.salesforce.com/agentforce/.

[66] ServiceNow. AI Agents. *ServiceNow* **2025**. https://www.servicenow.com/products/ai-agents.html.

[67] LangChain. LangGraph: Build resilient language agents as graphs. *LangChain* **2025**. https://www.langchain.com/langgraph.

[68] CrewAI. CrewAI: Multi-agent automation framework. *CrewAI* **2025**. https://www.crewai.com/.

[69] Microsoft. AutoGen Studio. *Microsoft AutoGen Documentation* **2025**. https://microsoft.github.io/autogen/stable/user-guide/autogenstudio-user-guide/.

[70] Luo, R.; Sun, L.; Xia, Y.; Qin, T.; Zhang, S.; Poon, H.; Liu, T.-Y. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Brief. Bioinform.* **2022**, *23*, bbac409. https://doi.org/10.1093/bib/bbac409.

[71] Anonymous. Artificial intelligence agents in healthcare research: A scoping review. *PLOS One* **2026**. https://journals.plos.org/plosone/.

[72] Yang, H.; Liu, X.-Y.; Wang, C.; Yang, Q. FinGPT: Open-source financial large language models. *arXiv* **2023**, arXiv:2306.06031. https://arxiv.org/abs/2306.06031.

[73] Khan, A.T.; Li, S.; Cao, X. Bridging finance and AI: A comprehensive survey of large language models in financial system. *Digit. Finance* **2025**, *7*, 679–701. https://doi.org/10.1007/s42521-025-00146-3.

[74] Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Dabis, J.; Finn, C.; et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv* **2023**, arXiv:2307.15818. https://arxiv.org/abs/2307.15818.

[75] Burger, B.; Maffettone, P.M.; Gusev, V.V.; Aitchison, C.M.; Bai, Y.; Wang, X.; et al. A mobile robotic chemist. *Nature* **2020**, *583*, 237–241. https://doi.org/10.1038/s41586-020-2442-2.

[76] King, R.D.; Whelan, K.E.; Jones, F.M.; Reiser, P.G.K.; Bryant, C.H.; Muggleton, S.H.; et al. The automation of science. *Science* **2009**, *324*, 85–89. https://doi.org/10.1126/science.1165620.

[77] Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; et al. On the opportunities and risks of foundation models. *arXiv* **2021**, arXiv:2108.07258. https://arxiv.org/abs/2108.07258.

[78] Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency* **2021**, 610–623. https://doi.org/10.1145/3442188.3445922.

[79] Perez, E.; Huang, S.; Song, F.; Cai, T.; Ring, R.; Aslanides, J.; et al. Red teaming language models with language models. *Empirical Methods in Natural Language Processing* **2022**. https://arxiv.org/abs/2202.03286.

[80] Zou, A.; Wang, Z.; Carlini, N.; Nasr, M.; Kolter, J.Z.; Fredrikson, M. Universal and transferable adversarial attacks on aligned language models. *arXiv* **2023**, arXiv:2307.15043. https://arxiv.org/abs/2307.15043.

[81] Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; Mittal, P. Fine-tuning aligned language models compromises safety, even when users do not intend to. *International Conference on Learning Representations* **2024**. https://arxiv.org/abs/2310.03693.

[82] Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. https://doi.org/10.1136/bmj.n71.

[83] Qian, C.; Cong, X.; Yang, C.; Chen, W.; Su, Y.; Xu, J.; et al. Communicative agents for software development. *arXiv* **2023**, arXiv:2307.07924. https://arxiv.org/abs/2307.07924.

[84] Wang, Z.; Cai, S.; Chen, G.; Liu, A.; Ma, X.; Liang, Y. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *Adv. Neural Inf. Process. Syst.* **2023**. https://arxiv.org/abs/2302.01560.

[85] Wu, C.; Yin, S.; Qi, W.; Wang, X.; Tang, Z.; Duan, N. Visual ChatGPT: Talking, drawing and editing with visual foundation models. *arXiv* **2023**, arXiv:2303.04671. https://arxiv.org/abs/2303.04671.

[86] Shen, Y.; Song, K.; Tan, X.; Li, D.; Lu, W.; Zhuang, Y. HuggingGPT: Solving AI tasks with ChatGPT and its friends in Hugging Face. *Adv. Neural Inf. Process. Syst.* **2023**, *36*. https://arxiv.org/abs/2303.17580.