

Evaluating and Regulating Agentic AI: A Study of Benchmarks, Metrics, and Regulation

Azib Farooq¹, Shaina Raza¹, Nazmul Karim¹, Hasan Iqbal¹, Athanasios V Vasilakos¹, and Christos Emmanouilidis¹

¹Affiliation not available

November 06, 2025

Abstract

Agentic AI represents a new generation of Artificial Intelligence (AI) systems capable of perceiving, reasoning, planning, and acting toward achieving goals with a degree of autonomy. Unlike traditional AI models that merely generate outputs, these systems maintain memory, interact with their environment, and adapt over time. However, evaluating such interactive and evolving behavior remains a significant challenge. While several recent surveys have examined agentic AI architectures, components, and applications, few have systematically reviewed their evaluation, particularly regarding performance, reliability, and governance across an evolving agentic AI ecosystem. This paper addresses that gap by reviewing recent progress in the development and assessment of agentic AI, focusing on three core dimensions: benchmarks, metrics, and governance. We analyze how current evaluation frameworks capture reasoning, planning, collaboration, and ethical alignment across single- and multi-agent systems. Ultimately, this study aims to establish a unified foundation for building trustworthy, auditable, and human-aligned AI agents. The project webpage is available at [project link](#).

Evaluating and Regulating Agentic AI: A Study of Benchmarks, Metrics, and Regulation

Azib Farooq^{a,1}, Shaina Raza^{b,1}, Nazmul Karim^c, Hasan Iqbal^d, Athanasios V. Vasilakos^{e,2} and Christos Emmanouilidis^{f,2}

^aUniversity of Cincinnati, Cincinnati, OH, USA

^bVector Institute, Toronto, ON, Canada

^cUniversity of Central Florida, Orlando, FL, USA

^dRocket Companies, Detroit, MI, USA

^eLulea University of Technology, Sweden

^fUniversity of Groningen, Netherlands

ABSTRACT

Agentic AI represents a new generation of Artificial Intelligence (AI) systems capable of perceiving, reasoning, planning, and acting toward achieving goals with a degree of autonomy. Unlike traditional AI models that merely generate outputs, these systems maintain memory, interact with their environment, and adapt over time. However, evaluating such interactive and evolving behavior remains a significant challenge. While several recent surveys have examined agentic AI architectures, components, and applications, few have systematically reviewed their evaluation, particularly regarding performance, reliability, and governance across an evolving agentic AI ecosystem. This paper addresses that gap by reviewing recent progress in the development and assessment of agentic AI, focusing on three core dimensions: benchmarks, metrics, and governance. We analyze how current evaluation frameworks capture reasoning, planning, collaboration, and ethical alignment across single- and multi-agent systems. Ultimately, this study aims to establish a unified foundation for building trustworthy, auditable, and human-aligned AI agents. The project webpage is available at [project link](#).

1. Introduction




Traditionally, artificial intelligence (AI) in agents and robotics has referred to systems designed to perceive their environment, make autonomous decisions, and act toward achieving defined goals [39]. With the advent of large language models (LLMs), the notion of agency has expanded, enhancing agents' ability to reason, plan, and interact through natural language. Earlier forms of agents already incorporated learning and decision-making components, but LLMs have introduced a new level of generality and adaptability that enables richer, more context-aware behaviors. Prior research in cognitive architectures and goal reasoning has long demonstrated active, deliberative forms of agency. Early work connected dynamical systems and information-theoretic principles to cognitive behavior, showing that agents can engage in adaptive and self-organizing activity rather than fixed responses [67]. Psychological models further highlighted the role of motivation and affect in shaping agent behavior [134], later extending to the design of ethical cognitive agents capable of moral reasoning and decision-making [14]. In parallel, the goal-reasoning community explored meta-cognitive mechanisms that allow agents to generate, pursue,

and revise goals autonomously, as exemplified by the MIDCA architecture[29].

Building on this progression, agentic AI represents a broader paradigm that goes beyond single-step task prediction to operate as persistent, goal-directed agents [182]. Unlike traditional models that take an input and return an output, agentic AI maintains state, plans over multiple steps, leverages tools, and adapts its behavior dynamically in open-ended environments. These capabilities allow agentic AI systems to function more like collaborators or decision-makers than passive models.

Open source LLMs, like DeepSeek R1[136], phi 3 [138], gemma 7B [37], mistral 7B [137] and Meta LLaMa 3 [143] models, have further accelerated this shift. With natural language as their primary interface, LLM-based agents can chain reasoning steps, call external tools and APIs, browse information sources, and interact with humans [101]. Early agentic AI systems such as AutoGPT [175] and BabyAGI [86] demonstrated the potential of LLMs to achieve autonomy by decomposing goals into subtasks, while more recent frameworks integrate memory, multi-agent collaboration, and real-world interaction [32, 46]. The result is a rapid proliferation of LLM-based agentic AI systems across domains ranging from software development to scientific discovery.³

As agentic AI research progresses toward richer multi-agent ecosystems, the growing complexity of these systems has also motivated a shift toward modular and interpretable designs. The notion of decomposing an agentic system into

 farooqai@mail.uc.edu (A. Farooq); shaina.raza@torontomu.ca (S. Raza ); nazmul.karim170@gmail.com (N. Karim ); hasaniqbal@rocket.com (H. Iqbal); athanasios.vasilakos@ltu.se (A.V. Vasilakos); c.emmanouilidis@rug.nl (C. Emmanouilidis)

ORCID(s):

¹Equal contribution.

²Senior authors.

³For the scope of this paper, the term "agentic AI systems" specifically refers to LLM-based agents.

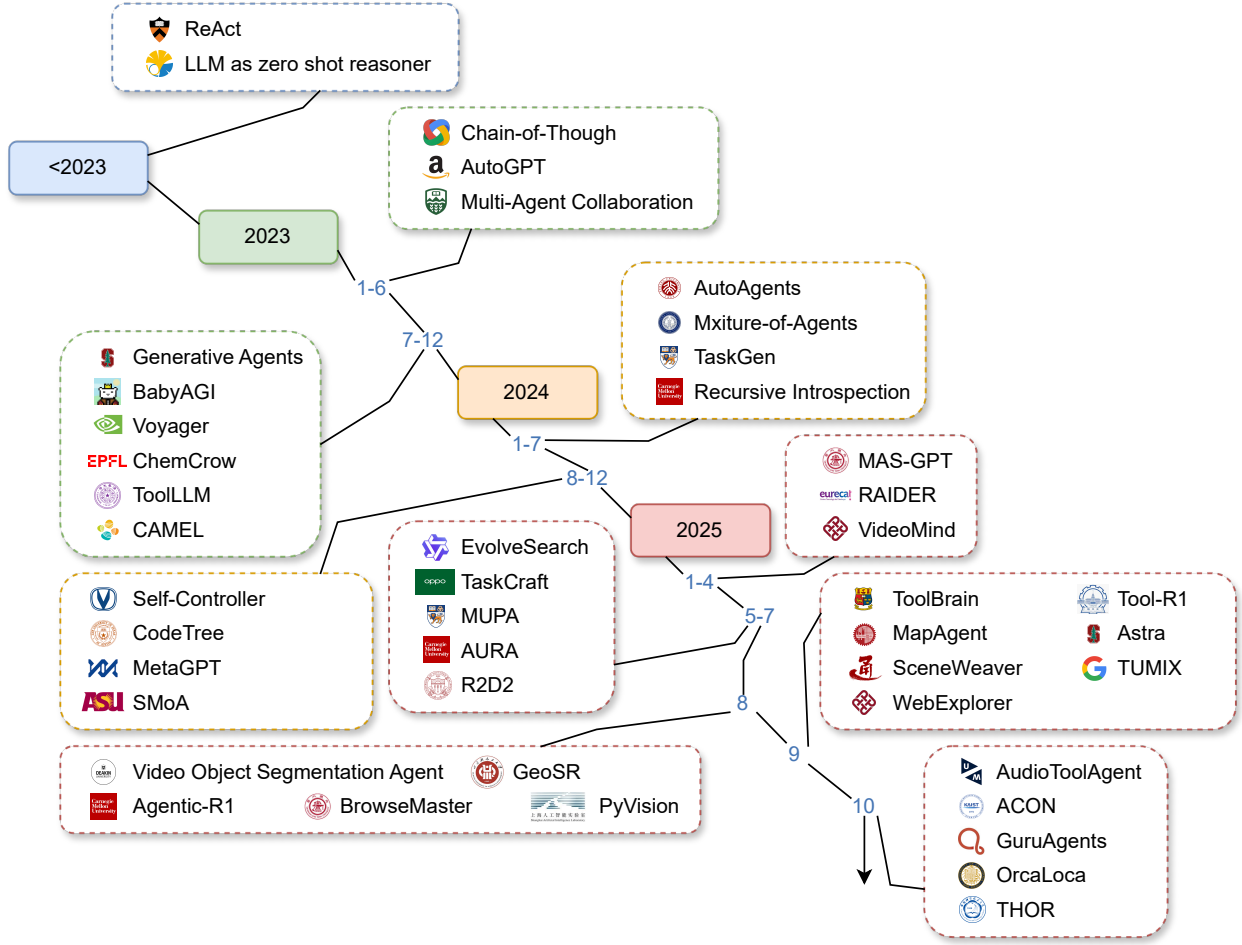


Figure 1: Timeline of Agentic AI papers which are released from 2023 to 2025

smaller and functional cognitive blocks is referred to as cognitive modularity [41]. For example, an agentic AI system can be partitioned into (i) a planner that produces hierarchical goals, (ii) a reasoning module that generates explanations or chains of thought, (iii) tool-use wrappers that mediate external APIs, and (iv) a memory subsystem that stores episodic and semantic traces. This modularization has been proposed as a design principle in recent cognitive-architecture work [21, 188, 174, 168] to improve interpretability, enable targeted fine-tuning and evaluation, and support incremental learning.

In parallel, the integration of retrieval-augmented generation (RAG) and explicit memory architectures with human-in-the-loop (HITL) oversight has become a standard design pattern to reduce hallucinations [117], ground decisions in external knowledge [69], and permit safe intervention [183]. Recent RAG extensions and agent-memory studies emphasize retrieving not only facts but also structured exemplars and past action trajectories [161] for improved plan supervision. Task-specific context and human feedback are driving a transition from static, single-turn LLM interactions toward continuously learning, socially aware systems that (i) retain and retrieve past interactions, (ii) solicit clarifications when

intentions are ambiguous, and (iii) use modular reflection to improve collective behavior over time [168, 153].

1.1. Motivation

Despite this rapid evolution, evaluating the effectiveness of agentic AI systems remains a major challenge. For instance, consider an agentic AI pipeline specialized in multi platform e-commerce recommendation, and the user want to buy a laptop for less than \$1000 for graphic designing application with long battery life. In Figure 3, all the necessary steps are shown along with their potential failure cases and the metrics for each individual step. Success on all these metrics would ensure the overall task success for the AI agent.

Traditional LLM evaluation methods [108, 110], such as static accuracy or correctness on held-out benchmarks, single-turn prompt-response metrics (e.g., BLEU, ROUGE, or exact match), and tool-invocation precision, fall short of capturing the dynamic, interactive, and goal-oriented nature of agentic AI behavior [84, 46]. Static benchmarks rarely account for temporal coherence, task decomposition, coordination, or how agents recover from mistakes and adapt to shifting goals. Emerging research explores new paradigms of agentic AI evaluation, emphasizing metrics such as task completion rate,

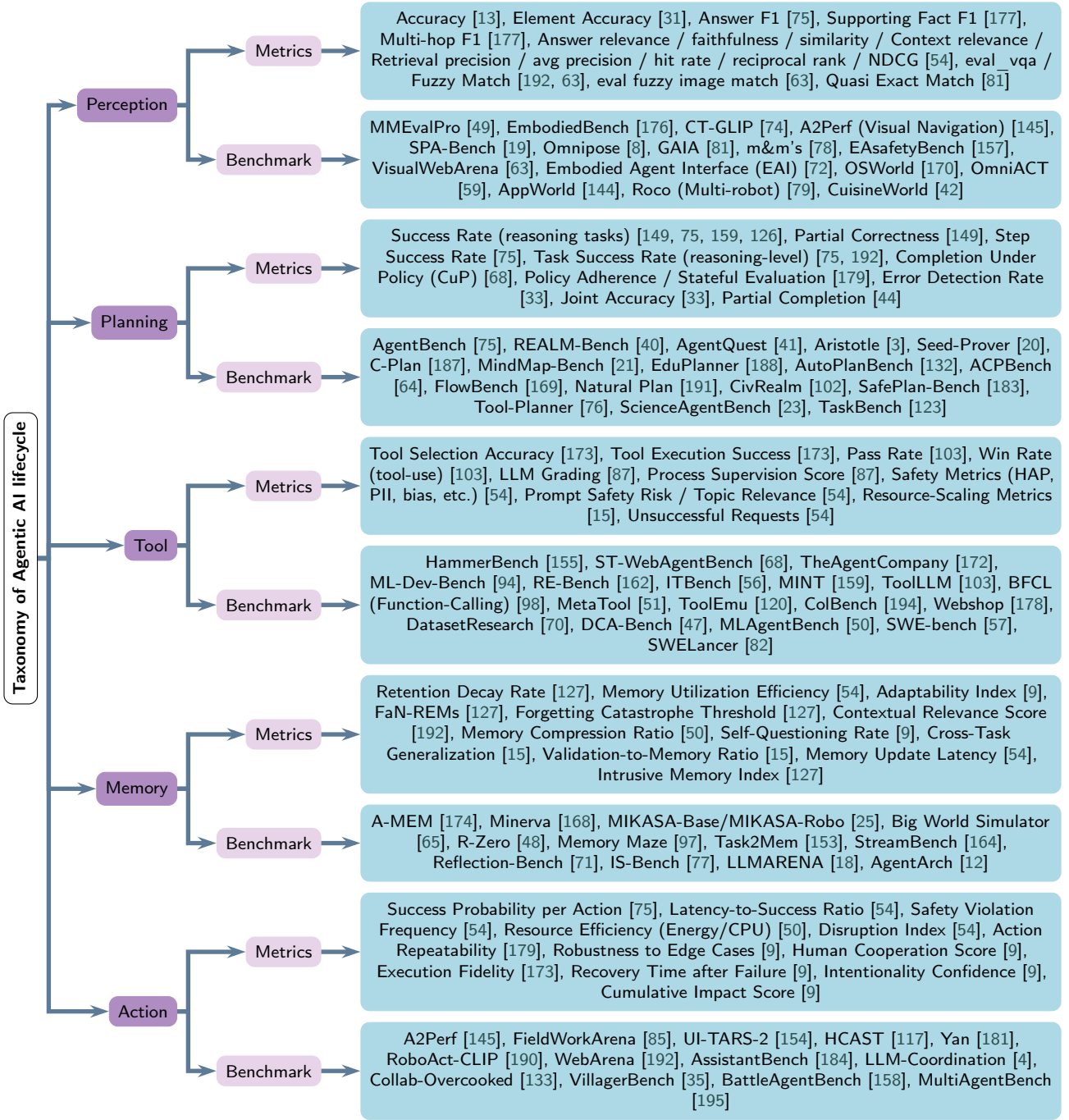


Figure 2: Taxonomy of Benchmarks and Evaluation Metrics for Agentic Lifecycle.

reasoning depth, cooperation efficiency, and ethical alignment [123, 127, 4, 183].

Current surveys on agentic AI largely emphasize taxonomies of components (e.g., planning, memory, perception, tool use), architectures (single-agent vs. multi-agent), or application domains (education, healthcare, software engineering), as shown in Table 1. While these reviews provide valuable conceptual overviews, they lack systematic analyses of benchmarks, metrics, and governance frameworks. There is limited clarity on how to measure agentic AI performance across dimensions such as robustness, reliability, efficiency,

safety, alignment, and governance. This gap underscores the need for a dedicated study on evaluation and regulatory alignment to ensure that agentic AI systems are not only powerful but also trustworthy, auditable, and aligned with societal values.

1.2. Main Contributions

To address this gap, this paper presents a dedicated review of evaluation methodologies for agentic AI systems. Our specific contributions are as follows:

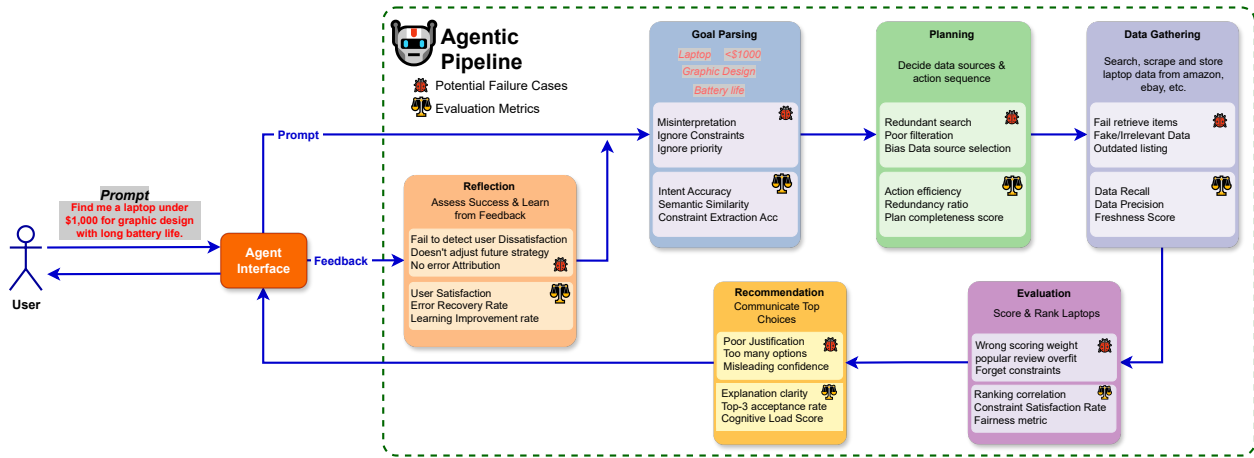


Figure 3: Working Instance Diagram, showing the workflow for buying a laptop using agentic pipeline. The bugs show potential failure cases that can happen at each pipeline step, and the evaluation metrics which could be used to cope up with those failure cases.

Table 1

Symbolic comparison of prior survey papers and our work on agentic AI.

Survey	Focus	Timeline	Benchmarks	Metrics	Governance
Yehudai et al. [182]	Evaluation of LLM-based agents	Post	✓	✓	⦿
Plaat et al. [101]	Agentic LLMs: reason-act-interact	Post	⦿	⦿	⦿
Acharya et al. [2]	Agentic AI foundations & applications	Pre+Post	⦿	⦿	✓
Bandi et al. [9]	Taxonomy: AI agents vs. agentic AI	Pre+Post	✗	✗	⦿
Hughes et al. [52]	Multi-expert industry analysis	Pre+Post	✗	✗	✓
Piccialli et al. [100]	Distributed AgentAI for Industry 4.0	Pre+Post	⦿	⦿	✓
Mohammadi et al. [84]	Evaluation/benchmarking of LLM agents	Post	✓	✓	✗
Nisa et al. [90]	Agentic AI overview (org transformation)	Pre+Post	⦿	⦿	✗
Yu et al. [185]	Trustworthy & Security evaluation of agents	Post	✗	⦿	✗
Ours	Benchmarks, metrics, and governance	Post	✓	✓	✓

Legend: ✓ = Present / Strong, ⦿ = Partial / Limited, ✗ = Absent, Pre = Pre-LLM era, Post = Post-LLM era.

- We provide a comprehensive taxonomy (Figure 2) and a structured review of existing benchmarks and evaluation metrics for agentic AI systems, organized by lifecycle stages. This includes a consolidated mapping of benchmark types, evaluation dimensions, and performance indicators.
- We present an analysis of the literature (Sections 2, 5, 6) highlighting emerging trends, methodological gaps, and open challenges in the evaluation of agentic AI.
- We examine the governance dimensions and regulatory alignment of current evaluation frameworks (Section 7) with respect to trust, accountability, and compliance standards.
- We introduce a novel evaluation framework (Section 4) specifically designed for agentic AI systems. Unlike static benchmark approaches, it integrates lifecycle-aware assessment, multi-step reasoning, robustness, and ethical alignment into a unified evaluation paradigm.

1.3. Paper Organization

The remainder of this paper is organized as follows. Section 2 outlines the literature review methodology, detailing the search strategy, inclusion-exclusion criteria, and thematic synthesis used to map existing studies on agentic AI evaluation. Section 3 provides background on LLMs, VLMs, RAG, LLM-based agents and multi-agent systems. Section 4 provides a novel evaluation framework for any general purpose agent. Followed by section 5 discusses the major benchmarks and datasets for assessing autonomous LLM agents, highlighting their domains, design principles, and comparative scope. Section 6 presents the taxonomy of evaluation metrics, describing both quantitative and qualitative dimensions for measuring performance, reliability, and human alignment. Section 7 examines governance, policy, and audit frameworks for agentic AI, identifying emerging standards and best practices for oversight and safety. Penultimate section 8 of the paper provide open challenges and future directions for the agentic AI. Finally, Section 9 concludes with insights on open challenges, gaps, and future directions toward unified evaluation and governance frameworks for agentic AI systems.

2. Literature Review Methodology

This review followed a structured process to ensure balanced coverage of existing work on agentic AI evaluation methods and benchmarks. Relevant studies were identified through database searches using keywords such as “Agentic AI”, “LLM-based agents”, “evaluation frameworks”, and “benchmarks”. Papers were included if they discussed evaluation methodologies, benchmark design, or assessment metrics relevant to LLM-driven or autonomous agents. The collected literature was then analyzed and grouped thematically to capture current trends, limitations, and open challenges.

The following search terms are used to search material across major databases such as IEEE Xplore, ACM Digital Library, SpringerLink, and arXiv. “Agentic AI”, “autonomous agents”, “LLM-based agents”, “AI evaluation”, “benchmarking”, “agentic frameworks”, “multi-agent systems”, “reflection in agents”, “tool use”, and “reasoning assessment”. Boolean operators (AND, OR) were applied to refine queries (e.g., “Agentic AI” AND “evaluation frameworks”). The search covered publications from **2023–2025** to capture the latest post-LLM developments.

Inclusion Criteria

- Studies focusing on evaluation frameworks, benchmarks, or assessment metrics for LLM-based or autonomous agent systems.
- Works discussing planning, reflection, memory, or collaborative reasoning within agentic settings.
- Peer-reviewed articles, high-quality preprints, or technical reports from reputable research groups.

Exclusion Criteria

- Opinion pieces, editorials, or non-technical blog posts without methodological detail.
- Studies unrelated to evaluation, benchmarking, or agentic systems.
- Duplicate or early workshop versions of later-published papers.

A total of 150 papers were initially retrieved, of which 120 were carefully screened and assessed by the authors for inclusion based on relevance to agentic AI evaluation and benchmark design. Few foundational papers on the concepts like agents, LLMs, trustworthy AI are also include. A comprehensive release timeline of the papers is shown in figure 1.

Results The thematic analysis, table 2, reveals that the literature on agentic AI spans five major research trajectories. Foundational works focus on defining agentic AI, its conceptual boundaries, and system architectures, establishing theoretical and applied perspectives. Governance and safety studies highlight growing concerns over autonomy, alignment, and regulation, supported by specialized benchmarks

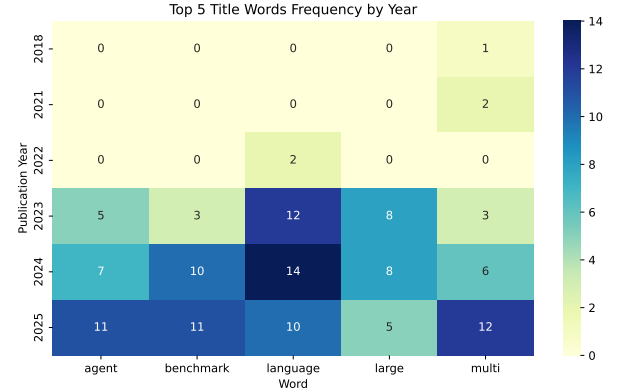


Figure 4: Top five most frequent keywords in the paper titles.

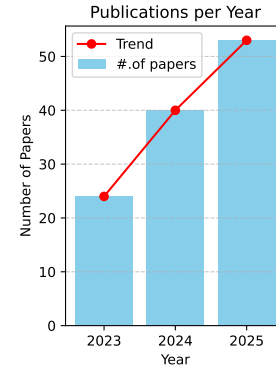


Figure 5: Barplot showing frequency of published papers.

assessing robustness and ethical compliance. The results on the literature review analysis also show that benchmarking research advances the systematic evaluation across reasoning, planning, and adaptive performance. Multi-agent studies emphasize emergent collaboration, competition, and social dynamics among agents. Finally, embodied and tool-use research grounds agentic AI in real-world and multimodal environments, showcasing progress toward autonomous, context-aware, and interactive decision-making systems. Figure 4 and 5 depicts the frequency trend of the top five keywords and the number of paper published each year in the domain of agentic AI. Both figures shows rising trend in multi-agent LLM publications.

3. Background

In this work, we use *agentic AI* to mean LLM/VLM systems that (i) maintain state across steps, (ii) plan and select tools/actions, and (iii) operate in interactive environments (web, OS, APIs, or embodied simulators). Out of scope are single-shot LLM prompts without tool use or state. We distinguish *working memory* (in-context state), *long-term memory* (external stores such as vector DBs or key-value logs), and *provenance/audit logs* (who/what/when/why for each action), which we later evaluate for fidelity and traceability (§6). We also assume optional *HITL gates* (approve/override), policy

Table 2
Thematic Analysis of Agentic AI Literature

Theme	Representative Works and Focus
Foundations and Conceptual Frameworks	Core conceptualizations of agentic AI, its definitions, architecture, and taxonomy. [9] provides a comprehensive synthesis of agentic AI evolution, while [101], [2], establish conceptual distinctions between AI agents and agentic AI. [32] and [52] outline system architectures and interdisciplinary perspectives, whereas [90], [100], and [46] extend the discussion toward industrial and social applications. Classical theoretical grounding is supported by [121] and [163].
Autonomy, Governance, and Safety	Ethical, regulatory, and safety implications of autonomous systems. [39] and [124] examine levels of autonomy and governance frameworks, while [5] and [80] raise alignment and auditing concerns. Benchmarks like [6], [30], [183], and [77] evaluate safety, robustness, and defense mechanisms. [157] and [145] explore embodied agent security and real-world performance monitoring.
Benchmarks and Evaluation Frameworks	Evaluation protocols and benchmarking suites for assessing LLM-based agents. Key foundational works include [75], [84], and [182]. Comprehensive multi-domain benchmarks are represented by [123], [12], [102], [40], and [82]. [170], [191], and [169] benchmark general reasoning and planning ability. Continuous evaluation approaches such as [164], [71], and [47] emphasize adaptive performance tracking and dataset curation ([70]).
Multi-Agent Systems and Collaboration	Studies addressing coordination, competition, and collective intelligence. [195], [158], and [53] assess large-scale cooperation/competition. [4], [79], and [35] explore coordination dynamics across agents. [18] and [133] focus on collaborative reasoning, while [42] extends to gaming contexts. Social-science-driven multi-agent paradigms are discussed in [46].
Tool Use and Embodied Environments	Embodied decision-making, tool integration, and interactive multimodal contexts. Tool-use exploration benchmarks include [103], [51], [98], and [76]. Embodied agent evaluations are led by [72], [175], and [178]. Realistic interactive testbeds such as [192], [144], and [59] simulate complex decision environments. Foundational task automation and planning are benchmarked in [132], [64], and [78]. Early open-source systems like [86] illustrate autonomous task execution foundations.

Table 3
Preliminary terminology for Agentic AI (LLM-based agents).

Term	Brief Definition
Agentic AI [32]	LLMs exhibiting autonomy through reasoning, planning, and acting loops.
LLM-based Agent [156]	LLM integrated with memory, tool use, and control modules for autonomous tasks.
Perception–Reason–Act Loop [180]	Core operational cycle where the agent observes, reasons, and acts in the environment.
Planner / Controller [156]	Translates high-level goals into executable sub-tasks or plans.
Executor / Tool Caller [122]	Invokes APIs, code, or tools to perform concrete actions.
Reflexion [125]	Mechanism for self-evaluation and improvement via feedback on past actions.
Retrieval-Augmented Generation (RAG) [69]	Enhances factual grounding via external document retrieval.
Memory (Short/Long-term) [156]	Stores contextual information across multiple reasoning steps.
Toolformer [122]	LLM trained to decide autonomously when and how to use tools.
Task Decomposition [156]	Divides complex tasks into smaller, manageable subgoals.
Observation / State [101]	Representation of the agent's perceived environment or belief.
Judge / Critic / Verifier [108]	LLM-based evaluation of correctness, coherence, or safety.
Guardrails / Policy Engine [34]	Enforces ethical or safety constraints on agent behavior.
Reward Model / Preference Signal [92]	Provides feedback to optimize or fine-tune agent policies.
Chain-of-Thought (CoT) [160]	Explicit reasoning traces that improve logical consistency.
Program-of-Thought (PoT) [22]	Executes reasoning as code for precision and verifiability.
Multi-Agent System [96]	Collaboration or competition among multiple autonomous agents.
Evaluation Trajectory [156]	Complete record of thoughts, actions, and observations for evaluation.

checks, and red-team loops that can interrupt or reshape plans; we reference these as governance hooks (§7). Some of the key terms used in this review article are given in Table 3. Taxonomy of benchmarks and evaluations terms for agentic AI lifecycle are shown in Figure 2.

3.1. LLM-based Agents and Types

LLMs have textual interface, and are transformer-based architectures trained on massive text corpora to perform reasoning, planning, and generative tasks through contextual learning [45]. They enable flexible, few-shot generalization across diverse domains but remain prone to hallucination and limited grounding. VLMs extend this paradigm by integrating

visual modality, enabling multimodal reasoning in tasks such as image captioning, visual question answering, and perception-grounded dialogue [114]. RAG based systems further enhance factual reliability by coupling parametric memory (LLM knowledge) with non-parametric memory (retrieval index) [69]. This allows agents to access up-to-date information dynamically, mitigating hallucinations but introducing dependency on retrieval quality and latency.

In AI, an agent is a system capable of perceiving its environment, reasoning over objectives, and taking actions to achieve those objectives [163]. Modern LLM-based agents combine language understanding with external tool use, executing API calls, browsing, coding, or querying databases

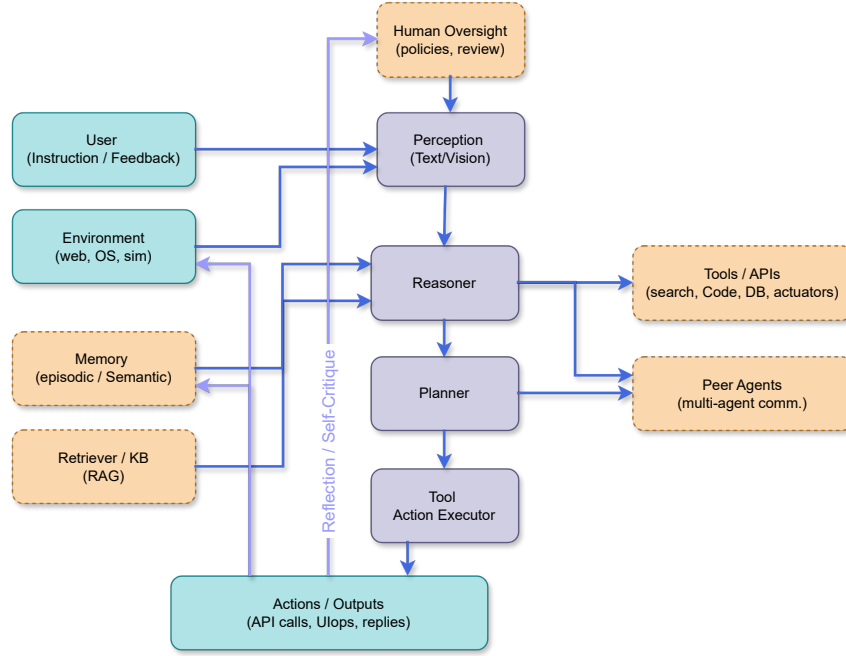


Figure 6: Agentic AI overview: perception → reasoning → planning → tool/execution with memory/RAG, human oversight, and multi-agent communication.

through natural language reasoning. These agents can plan, decompose, and act within dynamic environments, but their autonomy remains limited by due to insufficient prompts, temporal memory loss, and lack of consistent governance frameworks. An agentic AI overview is given in Figure 6. Table 4 shows comparison on different aspects for LLMs, VLMs, RAG, and agents.

Multi-agent systems generalize single-agent setups by introducing multiple interacting entities, each specialized in perception, planning, or execution to collectively solve complex problems [105]. This paradigm supports specialization, robustness, and emergent cooperation [185]; however, it also raises questions about explainability and interpretability. Multi-agent frameworks are now being explored for collaborative reasoning, human–agent teaming, and self-improving ecosystems, marking a shift from isolated reasoning models to interactive cognitive systems [10].

AI agents can be understood across four progressive stages of intelligence and autonomy [152]. At the foundation lie **Retriever–Generator Agents**, which combine retrieval and generation to provide contextually relevant responses from existing data sources. They are purely reactive [69], responding only to user inputs without memory or self-improvement. Building upon this, **Tool-Enhanced Agents** integrate external capabilities, such as APIs and databases [173, 122, 104], allowing them to perform concrete tasks like searching, calculating, or booking. While more capable, they still depend entirely on user prompts. The next evolution, **Strategic Agents**, introduces planning and reasoning: these agents can decompose complex goals into structured steps, adapt their actions based on results, and refine their approach

dynamically [188, 23] – yet they remain confined to single – session workflows without true persistence. At the frontier are **Autonomous Agents**, which can sustain context across sessions, initiate actions independently, and adapt through feedback [71]. They represent a shift from reactive to proactive intelligence – agents that can, in theory, self-direct improvement and generate novel solutions – though fully realized autonomy of this kind remains a vision of the near future.

3.2. Challenges in Agentic AI Works

Based on our review of the literature, some of the challenges associated with agentic AI frameworks are given below:

- **Reasoning reliability and grounding.** Agents still hallucinate, struggle with temporal grounding (what changed since time t), and mis-handle uncertainty calibration under tool use.
- **Tool-use brittleness.** Chain-of-thought (CoT) plans often degrade when external APIs are slow; error handling and recovery policies are under-specified, which causes cascading failures.
- **Memory and provenance.** Short/long-term memory is inconsistent; there is lack of provenance which makes it hard to trace why an action was taken or to reproduce outcomes.
- **Multi-agent coordination.** Division of labor helps performance but introduces coordination costs (conflicting beliefs, shared-memory races) and unclear accountability across agents.

Table 4

Comparison between different AI technologies across multiple aspects. **T** refers to Text modality, **I** is for image modality, **D** is for document modality, **MM** is for multimodal, **✓** is for the full support, **●** is for the partial support and **✗** is for the no support.

Aspect	LLM	VLM	RAG	LLM-Based Agent	Multiagent Framework
Input Modality	T	T+I	T+D	MM	MM
Contextual Reasoning	✓	✓	✓	✓	✓
Grounded Knowledge	✗	●	✓	✓	✓
Tool Use / External APIs	✗	✗	●	✓	✓
Autonomous Planning	✗	✗	✗	✓	✓
Collaboration / Coordination	✗	✗	✗	●	✓
Long-Term Memory	✗	✗	●	●	✓
Robustness to Tool/Env Drift	✗	●	●	●	●
Provenance / Audit Logs	✗	✗	●	●	●

- **Evaluation gaps.** Most benchmarks score final answers, not the *process*: planning quality, tool selection, collaboration, and safety under distribution shift remain under-measured.
- **Governance and auditability.** Existing checks are static and single-shot; agentic settings need continuous oversight, action logs, decision rationales, and verifiable data lineage.
- **RAG dependency and data quality.** Retrieval improves fidelity but creates new failure modes: stale indices, biased sources, latency/throughput trade-offs, and query-construction errors.
- **Efficiency and cost.** Long-horizon tasks inflate latency, carbon, and spend; few works report cost/energy alongside quality, complicating responsible deployment decisions.

These issues motivate lifecycle-aware benchmarks and governance-aligned evaluation (§5, §6).

3.3. Benchmarks in Agentic AI

In current LLM-based agentic AI research, a variety of benchmarks are used to assess the capabilities of autonomous agents. Existing agentic-AI benchmarks assess interactive capabilities beyond static QA. For example, AgentBench [75], which measures multi-turn reasoning across domains; τ -Bench [179], which evaluates long-horizon and human-in-the-loop interactions; and TRAIL [33], which focuses on trace-based reasoning and error localization. Other notable frameworks such as WebArena [192], OSWorld [170], and FieldWorkArena [85] test agents in realistic or simulated environments that require dynamic adaptation and sustained goal pursuit. Collectively, these benchmarks reflect the growing shift from static LLM evaluation toward interactive, task-oriented, and behavior-level assessment of agentic intelligence. Despite, seminal progress in the field, current benchmarks do not provide a unified, governance-aligned, lifecycle view of agent behavior. We therefore introduce a taxonomy (Fig. 2) and a consolidated metric suite (Table 6, and §6) that jointly evaluate *process*, *governance*, *robustness*,

and *efficiency*, enabling apples-to-apples comparisons across LLM, VLM, RAG, and multi-agent systems.

3.4. Evaluations in Agentic AI

Current evaluation methods for agents borrow metrics from adjacent domains like NLP (accuracy, BLEU, perplexity) [116], reinforcement learning (RL) (reward, success rate), and multimodal reasoning (VQA accuracy, grounding IoU) [110]. While these provide partial insights, they remain fragmented and fail to capture the holistic performance of agentic systems that operate across multiple components and modalities.

Evaluating agentic AI systems requires precise and multi-criteria metrics that capture not only task success but also efficiency, reliability, and alignment with human values. Core measures such as success rate and accuracy quantify an agent’s ability to achieve defined goals, forming the baseline for competence. Pass@k [15] and policy-adherence [179] metrics extend this by assessing the agent’s consistency and rule following behavior across repeated trials or stateful environments. Efficiency-oriented metrics like token usage [50] and latency [54] evaluate how effectively agents utilize computational resources and respond within real-time constraints. Finally, emerging human-centric dimensions, such as explainability and fairness [110], reflect a growing emphasis on transparency, safety, and social alignment in autonomous systems. More details on evaluations of agentic AI systems are given in §5 and §6 and in Tables 5 and 6.

4. Framework for Evaluating Agentic AI Systems

We are proposing an agentic AI system evaluation framework 7. The framework is logically divided into four sections, each designed to evaluate a particular aspect of the agentic AI system.

Agent Type Identification The first stage of the evaluation is Agent Type Identification. The most popular categories for agent types as we discussed in Section 3.1 include retrieval generator, tool-calling, planning, and autonomous agents.

These broad categories can often represent any general-purpose agent. Although, a particular agent is not necessarily confined to only one type. An agent may combine different types, for instance, an airline recommendation agent could comprise retrieval generator, planning, and tool-calling sub-agents.

Tasks and Environment Following agent type identification is the evaluation of the task and environment of the agent. A range of publicly available benchmarks fit into this category. Section 5 provides a comprehensive discussion of benchmarks for agentic AI system evaluation. In regard to the agentic environment, similar types of agents can report different performance if provided with variable environmental testbeds. The Holistic Agentic Leaderboard (HAL) [60] reports the performance of main agentic systems in relation to a specific environment, referred to as *scaffold*. Different agentic AI scaffolds can impact the performance assessment of the agentic system. Thus, environmental consistency is of immense importance for a valid agentic system evaluation.

Instrumentation & Tracking In the next stage, the metrics corresponding to the agentic pipeline are used to evaluate each sub-stage of the pipeline. This is accomplished in the Instrumentation and Tracing stage. At this point, the agent AI overall capability to accomplish the task can be measured. The overall task success of the agentic AI system relies on a range of intermediate decisions. Poor performance on these intermediate steps would deteriorate the overall agentic performance, as shown in Figure 3. Comprehensive metrics for this evaluation are provided in Section 6. The overall success of the agentic AI system will be a culmination of the micro-successes of the agent in those small, atomic tasks.

Service Level Objectives (SLOs) and Governance The final stage of the evaluation is related to governance. This stage checks if the agent complies with global standards for safety and risk. Service Level Objectives (SLOs) determine the criteria the agentic AI system must comply with to be regarded as successful. Regulatory organizations like *ISO* and *NIST* have introduced comprehensive standards for AI regulations [88, 1, 148] to ensure the safe propagation of the technology. There are also qualitative metrics, which provide more fine-grained regulation on AI systems. Extensive detail on agentic system alignment and regulation can be found in Section 7.

Throughout the entire evaluation phase, the datasets, metrics, and case studies of the agentic AI system are continuously monitored and evaluated. All the components and flow the evaluation framework is provided in Figure 7

Following the agentic evaluation framework, we examine state-of-the-art best performing agentic AI systems and their results on various publicly available benchmarks. Quantitative data for comparing these systems across multiple benchmarks is sourced from the HAL [60] and consolidated in Figure 8. It can be seen Figure 8 there is still significant room for improvement for top models in domain-specific

areas like scientific programming. Although models are showing better performance on composite benchmarks, such as GAIA [81], substantial progress is still needed in specialized benchmarks like scientific programming. Overall, Claude models dominate most reasoning and coding benchmarks, while GPT-5 and o3 models perform well on general and web-based tasks at lower costs. Specifically, o3 Medium leads in AssistantBench (38.8%) and ScienceAgentBench (33.3%). Claude Opus 4.1 performs best in CORE-Bench Hard (51.1%) and ties on SWE-bench (54%). Claude Sonnet 4.5 achieves the highest score in GAIA (74.6%). GPT-5 Medium ranks first in Online Mind2Web (42.3%) and USACO (69.7%), showing strong semantic and procedural reasoning. Performance in Scicode remains low overall, with o3 Medium reaching only 9.2%. On τ -Bench Airline, Claude-3.7 Sonnet and o4-mini High tie at 56%. In summary, most closed-source models excel in complex reasoning tasks, while GPT-5 and o3 models offer strong performance-efficiency tradeoffs.

5. Benchmarks and Datasets for Autonomous LLM Agents

We distinguish three artifact types used to evaluate autonomous LLM agents: (i) *interactive environments/benchmarks* (simulated or real systems that agents act within), (ii) *static datasets* (logs, trajectories, tool-call traces) for training and offline evaluation, and (iii) *task suites/wrappers* that orchestrate multi-step tasks across tools or environments. Benchmarks provide controlled but diverse settings to measure how agents perceive, reason, plan, and act. With consistent metrics: task success, long-horizon efficiency, tool-use reliability, memory retention, and safety; the benchmarks translate agency into comparable, reproducible indicators across prompts, architectures, and tool-use paradigms.

The importance of benchmarks lies in their role as unifying frameworks that guide reproducible evaluation and accelerate the evolution of general-purpose agentic systems. For example, AgentBench [75] introduces a multi-environment testbed for assessing reasoning and decision-making in open-ended digital tasks, whereas WebArena [192] focuses on realistic, goal-driven web interaction. Similarly, AgentBoard [16] provides structured evaluation pipelines capturing cognitive, reasoning, and social dimensions of agency. As summarized in Table 5, recent benchmarks (2023–2025) are being developed to evaluate different perspectives on how agency can be measured, compared, and improved across evolving real-world and simulated environments. A broad level taxonomy of these benchmarks is also depicted in figure 9 and below we discuss them.

Multi-Domain Agent Benchmarks test versatility, adaptability, and reasoning consistency of agents across diverse environments. It evaluates an agent ability to perceive, reason, and act across diverse interactive environments and task types within a unified framework. AgentBench is a comprehensive benchmark with 8 interactive environments for evaluating LLMs-as-agents [75]. It spans new domains like operating system control, database querying, knowledge graphs, digital

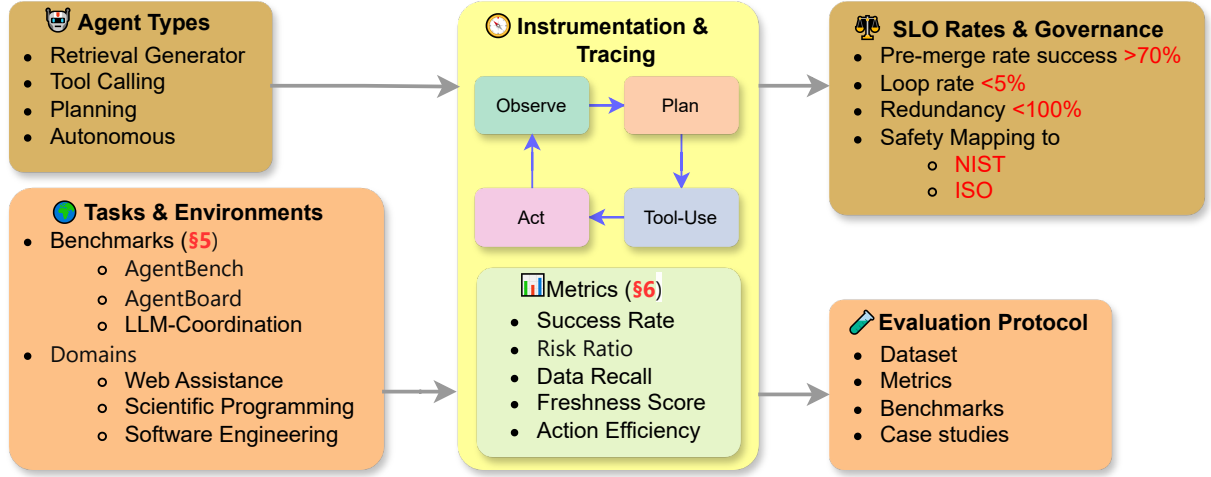


Figure 7: Evaluation Framework for Agentic AI system

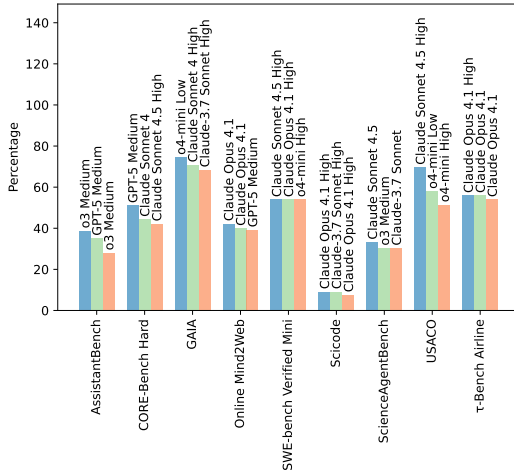


Figure 8: Percentage performance of top performing model on range of agentic AI benchmarks, quantitative data obtained from HAL [60]

card games, lateral thinking puzzles, as well as tasks adapted from prior datasets (e.g. ALFWorld for household tasks [126], WebShop for web shopping [178], Mind2Web for web browsing [31]). This provides a standardized testbed for multi-turn reasoning, tool use, and decision-making across diverse scenarios.

Web Interaction Environments evaluate how well agents navigate and act within web-based or GUI-driven ecosystems, replicating real-world human-computer interactions [68]. These benchmarks measure interface understanding, sequential reasoning, and the ability to complete complex web tasks autonomously. These benchmarks belong to a simulated or real-world digital platforms that evaluate an agent's ability to perceive, navigate, and act within web-based or GUI interfaces to accomplish specific tasks. Specialized benchmarks assess an agent's ability to navigate websites and GUI applications. For example, BrowserGym and WebArena

test general web navigation [26, 192], while WebCanvas adds GUI interactions [95]. VisualWebArena [63] and MMInA [141] introduce multimodal web tasks (combining text and images). The AssistantBench suite targets realistic, long-horizon web tasks (like complex, time-consuming online activities) to evaluate goal completion under real-world conditions [184].

Code and Research Tasks measure an agent's competence in coding, debugging, and conducting scientific reasoning or automation workflows. These benchmarks test logical planning, problem-solving, and the ability to interface with software tools or research pipelines. Several datasets target planning/execution in coding and scientific domains. SWE-Bench [57] consists of software engineering tasks like resolving GitHub issues, while ScienceAgentBench [23] focuses on automating scientific data analysis and programming. For research assistance, CORE-Bench [128] and PaperBench [131] challenge agents to reproduce or summarize academic results (e.g. reading papers, running experiments). AppWorld [144] provides interactive coding tasks within app interfaces, testing how well an agent can integrate with software tools.

Coding benchmarks can be further extended tool use as well. Tool use benchmarks evaluate function calling and API integration. FlowBench [169], ToolBench [76], and API-Bank [73] is a compilation of task which requires external function calls and API invocation. For instance, ToolBench includes an instruction-tuning dataset with 16,000 real-world APIs and an automated ToolEval to measure success rates and solution quality. These benchmarks often provide ground-truth tool call sequences and expected parameters, enabling fine-grained evaluation of whether an agent chooses the correct tools and inputs.

Memory and Long-Horizon Tasks focus on testing an agent's capacity to maintain continuity, recall, and strategic context over prolonged or sequential interactions. They highlight temporal reasoning, context retention, and state management across evolving scenarios. To assess an agent's

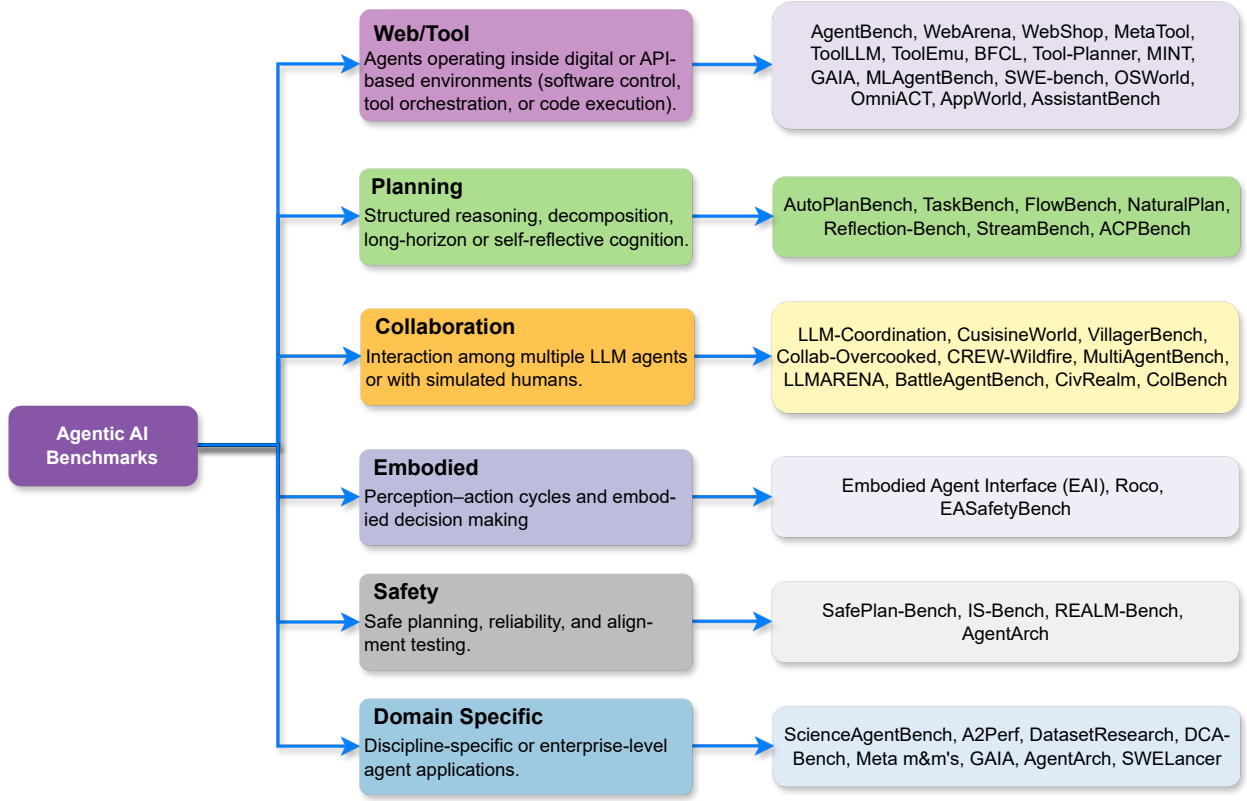


Figure 9: Agentic AI benchmark landscape

memory and context retention, benchmarks like SocialBench [17] use extended dialogues (40+ turns) where the agent later must answer questions requiring recall of early conversation details. Another example is TIME-Arena[189], a text-based simulation with time-sensitive multi-tasking (e.g. managing cooking and household tasks) that stresses temporal reasoning. Similarly, AndroidArena[171] provides a simulated mobile OS environment to test long-horizon task execution across multiple apps with user constraints. Success on these benchmarks indicates an agent's ability to maintain state over long interactions.

Multi-Agent and Social Simulations Benchmark assess how effectively multiple agents coordinate, communicate, and exhibit social intelligence in shared or cooperative environments. These benchmarks evaluate collaboration dynamics, emergent behavior, and collective decision-making. The SOTOPIA[193] platform (2024) creates "societies" of LLM-driven agents that learn and interact socially; it comes with an evaluation suite (SOTOPIA-Eval) that uses human and LLM judges to rate outcomes like collective goal completion and agent believability in social roles, [17]. Other works test emergent cooperation and communication by having multiple agents tackle tasks from domains like MMLU (language understanding), MATH, or even play games like chess and prisoner's dilemma [96, 162]. Metrics include whether agent teams outperform solos and if they exhibit human-like behaviors (e.g. forming consensus or trust).

Safety and Robustness Benchmarks test how resilient and ethically aligned agents remain when exposed to adversarial, unsafe, or manipulative scenarios. These benchmarks focus on identifying vulnerabilities, failure modes, and alignment risks in autonomous systems. A few datasets specifically probe risky or adversarial scenarios. AgentHarm [6] evaluates how often an agent produces harmful or toxic outputs when given unsafe prompts [84]. AgentDojo [30] tests an agent's resilience to prompt injection and jailbreaking attacks by simulating malicious inputs and seeing if the agent can be manipulated. Such benchmarks are used to audit safety. They complement general benchmarks by focusing on failure modes and alignment under stress.

6. Evaluation Metrics for Agentic AI Systems

Researchers employ a mix of quantitative and qualitative metrics to assess autonomous LLM agent's performance, safety, and reliability. Quantitative metrics provide objective numerical measure of an agent's efficiency, accuracy, or consistency, such as Success Rate, F1 Score, Wall-Clock Time, or Token Usage in benchmarks like AgentBench [75] and MAgentBench [50]. These capture measurable aspects of task completion and resource utilization.

In contrast, qualitative metrics capture subjective or interpretive dimensions of agent behavior, such as explainability, transparency, user satisfaction, or fairness [9, 115], evaluated through human judgment, reflection, or audits. Together,

Table 5: List of all the benchmarks associated with agentic AI from year 2023–2025

Benchmark Name*	Tasks	Description
AgentBoard [16]	Jericho, Tool-Query, Tool-Operation, Alfworld, ScienceWorld, BabyAI	Provides a fine-grained progress rate metric to capture incremental improvements during multi-turn interactions, along with a modular evaluation framework to analyze LLM agents across multiple behavioral dimensions.
AgentBench [75]	Operating System, Database, Knowledge Graph, Digital Card Game, Lateral Thinking Puzzles, House-holding, Web Shopping, Web Browsing	Assesses the ability of an LLM “agent” to reason and make decisions in multi-turn, open-ended settings across eight environments (e.g. OS, database, web tasks).
WebArena [192]	E-commerce, Social Forum Discussions, Collaborative Software Development, Content Management	Realistic web environment benchmark simulating tasks in e-commerce, social forums, collaborative coding, and content management; evaluates functional correctness on 812 web-based tasks.
GAIA [81]	Question Answering (requiring reasoning, multi-modality, web browsing, and tool-use)	466 real-world assistant tasks requiring multi-step reasoning, handling multimodal inputs (text + images/files), and proficient tool use; tasks range from simple queries to complex multi-tool problems
MINT [159]	Coding, Reasoning, Decision Making, Question Answering, Math Reasoning, Code Generation	Evaluates LLMs’ ability to solve tasks via multi-turn interactions using tools and dynamic feedback; repurposes tasks (reasoning QA, code generation, decision-making) to require iterative tool use and user feedback integration
ColBench [194]	Backend Programming, Frontend Design	Multi-turn benchmark where an LLM collaborates with a simulated human partner on coding/design tasks, proposing drafts, receiving feedback, and refining iteratively – simulating a realistic step-by-step development workflow
ToolEmu [120]	Risk Identification	Sandbox benchmark with 36 high-stakes tools and 144 test cases to probe risky tool-use behavior; simulates tool execution and uses an LM-based evaluator to examine agent failures and quantify associated risks
Webshop [178]	E-commerce Shopping (find, customize, and purchase an item)	Simulated e-commerce shopping environment (1.18M products, 12k+ user instructions) evaluating LLM agents on realistic web navigation—searching, filtering, and purchasing items to fulfill user requests
MetaTool [51]	Deciding Whether to Use Tools and Which to Use	Evaluates whether LLMs “know” when to use tools and can select the correct tool; includes 21k prompts with ground-truth tool usage (single-tool & multi-tool), covering tool-use awareness and nuanced tool-selection scenarios
BFCL (Function-Calling) [98]	API Call Generation	Tests LLMs on accurate function/API calls across 2000 QA pairs in multiple languages (Python, Java, JS, REST); evaluates correct function selection, argument formatting, and appropriate abstention from calling.
ToolLLM [103]	Single-tool API use, Multi-tool API use	Framework and benchmark for advanced API/tool use: introduces ToolBench, a dataset of 16k+ REST APIs (49 categories) with auto-generated instructions, to test multi-step planning, correct API invocation (including multi-tool sequences), and the ability to abstain when needed.
CREW-Wildfire [53]	Wildfire Response Scenarios	Open-source multi-agent benchmark using wildfire response scenarios with large maps, heterogeneous agents, partial observability, and long-horizon objectives; evaluates scalable coordination, communication, spatial reasoning, and planning under real-world complexity.
AgentArch [12]	Customer Request Routing, Requesting Time-Off	Enterprise-focused benchmark evaluating 18 distinct LLM-agent architectures (varying orchestration: single vs. multi-agent, prompting style, memory, tool integration) on complex workflow tasks, to reveal how design choices impact performance
m&m’s [78]	Multi-step Multi-modal Tool-Use, Tool-augmented Visual Question Answering	4,000+ multi-step multi-modal tasks (text+image+audio) using 33 tools (vision models, APIs, etc.), with 1,565 human-verified executable plans; enables evaluation of LLM planners under different strategies (one-shot vs. stepwise planning), plan formats (JSON vs. code), and feedback types.
TaskBench [123]	Task Decomposition, Tool Invocation, Parameter Prediction	Comprehensive framework for evaluating LLMs in task automation by breaking user instructions into sub-tasks; assesses performance in three stages (task decomposition, tool selection, parameter prediction) using a Tool Graph representation and multi-faceted (automated + human) evaluation.
LLM-Coordination [4]	Agentic Coordination, Coordination Question Answering, Environment Comprehension, Theory of Mind Reasoning, Joint Planning	Benchmark for pure multi-agent coordination games; includes an Agentic Coordination suite (LLMs act as agents in four cooperative games) and a Coordination QA set (198 questions) to test environment understanding, theory-of-mind reasoning, and joint planning capabilities.
Collab-Overcooked [133]	Collaborative Cooking	LLM-based multi-agent benchmark built on the Overcooked game environment with 30 collaborative tasks; supports natural language communication between agents and introduces process-oriented metrics for fine-grained evaluation of collaboration (coordination, adaptation) beyond task success.
Roco (Multi-robot) [79]	Sort Cubes, Multi-robot Collaboration	Benchmark for multi-robot collaboration using LLMs as controllers; evaluates how language-model agents coordinate multiple robots through dialogue-based planning and action to achieve shared goals in physical or simulated tasks.
VillagerBench [35]	Multi-agent Collaboration (in Minecraft)	Minecraft-based multi-agent benchmark where LLM agents (as “villagers”) must coordinate on complex, interdependent tasks; uses a graph-based task structure to evaluate planning and coordination in an open-ended sandbox environment.

Continued on next page

Evaluating and Regulating Agentic AI

Benchmark Name*	Tasks	Description
LLMARENA [18]	TicTacToe, ConnectFour, Texas Hold'em, Undercover, Bargain, Bid (First-price Sealed-Bid Auction), Hanabi	Benchmark of dynamic multi-agent scenarios to test LLMs' collaboration in changing environments; provides virtual interactive tasks that require agents to handle evolving state and other agents' behaviors in real time.
CivRealm [102]	Full Game (Civilization), Mini-games (Development, Battle, Diplomacy)	Uses a Civilization-game environment as a benchmark for long-horizon planning and reasoning; evaluates LLM-based agents on strategic decision-making and learning in a complex, multi-step world with open-ended goals.
BattleAgentBench [158]	Single-agent Scenario Navigation, Paired-agent Task Execution, Multi-agent Collaboration and Competition	Benchmark of multi-agent game scenarios designed to test both cooperation and competition among LLM agents; evaluates how well models can form alliances or adversarial strategies and adapt to other agents' actions.
CuisineWorld [42]	Multi-agent Collaboration in Gaming	A multi-agent gaming benchmark (within MindAgent) focusing on collaborative cooking tasks; requires several LLM-driven agents to coordinate efficiently to complete recipes, and introduces a Collaboration Score (CoS) to quantify team performance.
MultiAgentBench [195]	Collaboration, Competition	Comprehensive benchmark for LLM-based multi-agent systems across diverse interactive scenarios; measures task success as well as quality of collaboration and competition via milestone-based metrics, and evaluates various coordination protocols (star, chain, graph, etc.) for their effect on performance.
Embodied Agent Interface (EAI) [72]	Goal Interpretation, Subgoal Decomposition, Action Sequencing, Transition Modeling	Generalized interface and benchmark for evaluating LLMs on embodied decision-making tasks; unifies a wide range of embodied tasks with a common formalism (e.g. LTL goals, modular sub-tasks) and provides fine-grained error metrics (missing steps, wrong order, etc.) to diagnose LLMs' planning and reasoning in interactive physical environments.
AutoPlanBench [132]	Natural Language Planning Tasks (from PDDL domains like Blocksworld, Ferry)	Evaluates LLM agent planning abilities specifically in everyday scenarios, demonstrating agents lag behind classical symbolic planners.
SWE-bench [57]	Resolving Real-world Github Issues	Benchmarks agent performance on software development tasks, requiring agents to resolve real-world software issues found in public repositories.
ACPBench [64]	Action Applicability, Progression, Reachability, Action Reachability, Validation, Justification, Landmarks	A benchmark focused on evaluating LLMs on core reasoning skills across 7 reasoning tasks and 13 planning domains using formally synthesized problems.
OSWorld [170]	Open-ended Computer Tasks (web/desktop apps, file I/O, multi-app workflows)	A scalable, real computer environment with 369 tasks designed to evaluate multimodal agents' navigation and execution across various operating systems and applications.
OmniACT [59]	Generating Executable Programs (for Desktop and Web Tasks)	Tests agents' capacity to coordinate actions and execute complex tasks across multiple applications within full-scale computer operating environments.
AppWorld [144]	Interactive Coding Tasks (using day-to-day apps)	Evaluates whether agents can navigate real-world computer systems, execute complex tasks, and coordinate actions across multiple applications.
FlowBench [169]	Workflow-Guided Planning (e.g., customer service, logistics, healthcare)	Evaluates workflow planning abilities, specifically targeting expertise-intensive tasks that require sophisticated sequencing and orchestration.
Natural Plan [191]	Trip Planning, Meeting Planning, Calendar Scheduling	Designed to evaluate how LLMs handle real-world planning tasks when presented solely in natural language.
StreamBench [164]	Continuous Stream of Tasks	A challenging benchmark assessing agents' ability to continuously improve performance by leveraging external memory components and previous interactions over time.
Reflection-Bench [71]	Prediction, Decision-making, Perception, Memory, Counterfactual thinking, Belief updating, Meta-reflection	Assesses intrinsic epistemic agency by decomposing reflection into seven cognitive components, including belief updating, prediction, and meta-reflection.
REALM-Bench [40]	Real-world Planning Scenarios	Evaluates LLMs and multi-agent systems on dynamic, complex real-world planning and scheduling tasks, focusing on coordination and adaptation to disruption.
SWELancer [82]	Independent Engineering Tasks (bug fixes, feature implementations), Managerial Tasks (selecting technical proposals)	Targets freelance coding tasks, representing the latest trend in benchmark development by linking agent performance to monetary value and long-term reasoning.
EASafetyBench [157]	Input Moderation (for embodied agents)	A safety benchmark for embodied agents that provides a curated risk taxonomy, datasets, and evaluation suite for training/evaluating input-moderation systems for embodied LLMs.
SafePlan-Bench [183]	Safe Task Planning (Hazardous and Safe Tasks)	A safety-aware task-planning benchmark of 750 executable tasks (covering 10 hazard types) plus a universal SafeAgentEnv and evaluation metrics to measure embodied LLM agents' hazard recognition and safe rejection behavior.
IS-Bench [77]	Interactive Safety (Pre-caution and Post-caution safety risks)	A multimodal, process-oriented interactive-safety benchmark (161 scenarios, 388 risks) that tests whether VLM-driven embodied agents perceive emergent hazards and execute correct mitigation steps in order.
A2Perf [145]	Computer Chip Floorplanning (Circuit Training), Web Navigation, Quadruped Locomotion	A real-world autonomous-agents suite with three realistic environments (chip floorplanning, web navigation, quadruped locomotion) and metrics for task performance, generalization, efficiency, and reliability.

Benchmark Name*	Tasks	Description
DatasetResearch [70]	Dataset Discovery, Dataset Synthesis	2A demand-driven dataset-discovery benchmark (208 real-world dataset requirements) that evaluates agents' ability to find or synthesize datasets matching complex, knowledge-intensive user needs.
DCA-Bench [47]	Identifying Hidden Dataset Issues, Autonomous Dataset Curation	A dataset-curation benchmark (221 real-world test cases) that measures LLM agents' capability to detect and diagnose data quality issues in the wild with an automated evaluation pipeline.
Tool-Planner [76]	Task Planning with Clusters Across Multiple Tools	A tool-planning framework/benchmarking approach that clusters tools into functionally similar toolkits so LLMs can plan across toolsets and robustly recover from tool errors.
ScienceAgentBench [23]	Data-driven Scientific Discovery Tasks	A scientifically-grounded benchmark of 102 validated, domain-specific tasks (converted to self-contained Python programs) for rigorously assessing language agents on data-driven scientific discovery workflows.
AssistantBench [184]	Realistic and Time-Consuming Web Tasks (information-seeking)	A 214-task benchmark of realistic, time-consuming web tasks (multi-page/web-navigation problems) designed to evaluate web agents' ability to autonomously solve real user scenarios.
VisualWebArena [63]	Realistic Visually Grounded Web Tasks (e.g., Classifieds)	A large suite of visually-grounded web tasks (hundreds of realistic, self-hosted scenarios) for evaluating multimodal web agents on image+text perception plus web interaction skills.
MLAgentBench [50]	Machine Learning Experimentation Tasks	A suite of 13 machine-learning experimentation tasks that evaluates agents' ability to run experiments, modify code, and iterate to improve ML models end-to-end.
Can-Graph [166]	Task Planning	Introduces a benchmark to test whether graph-structured reasoning can enhance LLM-based agents' planning and decision-making in multi-step tasks.
SocialBench [17]	Sociality Evaluation (individual and group), Multi-choice Questions, Open-domain Generation Questions	Evaluates the social intelligence and role-playing abilities of conversational agents through structured multi-role social interaction scenarios.
Core-Bench [128]	Computational, Code and Result Reproducibility	Assesses the computational reproducibility of AI research by benchmarking agents that attempt to reproduce published experiments.
PaperBench [131]	Replicating AI Research Papers, Understanding Paper Contributions, Developing a Codebase, Executing Experiments	Benchmarks AI systems on their ability to replicate the methodology and results of existing AI research papers.
API-Bank [73]	Planning API calls, Retrieving APIs, Calling APIs	Provides a large-scale benchmark for tool-augmented LLMs, evaluating their capacity to invoke and coordinate multiple real-world APIs effectively.
MMINA [141]	Multihop Multimodal Internet Tasks (e.g., shopping, travel, event planning)	Benchmarks multimodal internet agents on multi-hop reasoning tasks involving text, images, and web-based information retrieval.
WebCanvas [95]	Dynamic Web Tasks	Tests web agents' ability to perform tasks in realistic browser-based environments with interactive, dynamic web pages.
BrowserGym [26]	Efficient Multitasking, Cooking, Household Activities, Laboratory Work	Offers a unified, open ecosystem for training and benchmarking web agents under controlled browser simulation environments.
TimeArena [189]	Daily Tasks on Android OS, Cross-APP Collaboration	Evaluates multitasking LLM agents in a time-constrained simulated environment to measure efficiency, prioritization, and scheduling capabilities.
Android Environment Benchmark [171]	Daily Tasks on Android Operating System, Cross-Application Collaboration	Analyzes LLM agents' weaknesses in complex Android operating environments, focusing on robustness, adaptability, and task success rate.
Sotopia [193]	Open-ended Social Interactions, Social Goal-driven Behavior (cooperative, competitive, and mixed)	Provides an interactive benchmark for evaluating the social intelligence and cooperation skills of language agents through multi-agent social simulations.

*For most benchmarks, the dataset name is the same as the benchmark name, so an extra column is not created.

these complementary approaches enable holistic assessment of both how well agents perform and how responsibly they act within complex, real-world environments [112]. A detailed taxonomy and comprehensive list of both qualitative and quantitative evaluation metrics, along with their corresponding benchmark sources, are presented in Table 6 and the accompanying evaluation metrics taxonomy diagram in this Figure 10.

Task Success and Goal Completion: This category captures whether an agent effectively fulfills the assigned objectives, representing the most fundamental indicator of

performance. It measures how consistently the agent achieves intended outcomes across different runs or environments. This is often reported as a success rate or task goal completion score per task. Many evaluations use binary success indicators (1 if the agent's actions satisfy the goal, 0 otherwise) or reward functions that flag goal achievement. For probabilistic agents, variants like Pass@N measure the probability of success within N attempts. For example, an agent might get multiple tries to solve a problem, and Pass@5 would indicate the percentage of cases it succeeds at least once in five tries. Metrics such as the *Task Success Rate* in



Figure 10: Agentic Lifecycle and taxonomy of metric on each step of the agentic lifecycle.

AgentBench [75], the *Success Rate* in MINT [159], and *Success Rate* in PlanBench [149] are representative examples used to quantify task completion performance.

Output Quality (Accuracy and Coherence): This dimension evaluates how correct, coherent, and contextually relevant an agent’s outputs are, beyond mere task success. It focuses on the content’s factual accuracy [13], fluency [192, 63], and logical soundness to reflect user-perceived quality. This metric includes traditional NLP metrics: accuracy and relevance of generated content, clarity/fluency of language, and logical coherence of the reasoning provided. An agent could complete a task but still produce a confusing or suboptimal solution [85], so these metrics capture the user experience. For instance, in conversational agents, one may rate the fluency of responses and the logical consistency of the agent’s explanations or steps. If the agent uses external tools or knowledge bases, standard retrieval-augmentation metrics apply: e.g. factual correctness (does the answer align with verified information) and contextual relevance of the response to the query. Often, human evaluators or LLM-based judges are used to qualitatively rate outputs when automatic metrics (like BLEU or ROUGE) are insufficient, for example, ranking the better of two agent’s solutions in terms

of correctness or preferences in user studies. Metrics such as *Accuracy* in SUPER [13], *Answer F1* in AgentBench [75], and *Correctness Score* in FieldWorkArena [85] exemplify this dimension.

Efficiency (Latency and Cost): Efficiency measures the agent’s ability to achieve goals quickly and with minimal computational [16] or financial cost. It reflects system responsiveness and resource optimization, both essential for real-world deployment. For interactive agents, latency is critical. Researchers measure Time to First Token (TTFT⁴) metric that measures how long a user needs to wait before seeing the model output. This is the time it takes from submitting a question to receiving the first token (if the response is not empty). In agentic AI setting, it measures, how quickly the agent begins responding, especially for streaming interactions. End-to-end latency [140] (total time to complete a task), is another important measure if the agent executes a long tool-use sequence or a multi-step plan. Cost is another practical metric: many LLM agents run on API calls, so one can estimate monetary cost by counting tokens or tool invocations [58]. Some works report the average number of model queries

⁴<https://docs.nvidia.com/nim/benchmarking/llm/latest/metrics.html>

or tokens consumed per task, since agents that solve problems with fewer calls are more efficient for real-world deployment. In sum, throughput and resource usage metrics help determine if an agent is not only effective but also practical to deploy at scale. Examples include *Wall-clock Time* and *Token Usage* in MLAGentBench [50] and *Latency/Response Time* in IBM WatsonX [54].

Tool Use Accuracy: This category assesses the precision and appropriateness of the agent's external tool or API utilization. Since tool use is central to any agentic AI system, specialized metrics evaluate each aspect of that process. Evaluation metric in context of tool use generally answer question like "*Did the agent decide to use a tool when it should? Did it pick the correct tool and use it properly?*". Metrics include Invocation Accuracy, whether the agent correctly determines if and when a tool/API call is needed [24], and Tool Selection Accuracy [104], whether it chooses the appropriate tool from an available set. In contexts where the agent must pick from a large toolkit or plugin library.

Retrieval Accuracy is another measure is a ranking metric [186] (e.g. Mean Reciprocal Rank or NDCG) for the target tool given a query. After selecting a tool, the agent must generate the right parameters for the API call⁵; here evaluators use metrics like parameter name F1-score (does the agent supply all required parameters correctly?) and argument accuracy. For example, if an agent calls a function `book_flight(destination, date)`, it should provide both destination and date in correct format, missing or wrong fields would lower the F1. Next step of evaluation is execution based by actually running the agent's tool calls in a sandbox and checking if the outcomes are correct. This catches semantic errors (the call runs but produces a wrong result) that mere syntax checks would miss. Representative metrics include *Parameter Accuracy*, *Tool Execution Success*, and *Tool Selection Accuracy* from ToolBench [173] and the *Pass Rate* from ToolLLM [103].

Planning and Reasoning Measures: These metrics measure the logical structure and multi-step reasoning quality behind an agent's decision-making process. This shows the quality of the agent's plan, not just the final result (whether it won or lost). To evaluate planning quality, researchers compare the agent's sequence of actions to an expert or reference plan [196]. One approach represents the ideal plan as a graph (with nodes = actions/tools and edges = order/dependencies). They then compute Node F1 (did the agent include the correct actions?) and Edge F1 (did it execute them in the correct order?) [166]. A high Node F1 but low Edge F1 means the agent chose mostly relevant steps but in the wrong sequence.

Similarly, a Normalized Edit Distance [55] between the agent's action sequence and the optimal sequence can indicate planning efficiency. Another metric is Step Success Rate [50, 75], the percentage of the agent's actions that are valid or lead closer to the goal. For instance, in a cooking task, every correctly executed recipe step would count toward this step

success rate [42]. These metrics, often used in benchmarks like ScienceWorld [23] or ALFRED [126], focusing on the agent's reasoning fidelity, not only the final answer but how it got there. Mistakes in planning (even if corrected later) may be penalized to encourage robust reasoning. Planning metrics like *Step Success Rate* from AgentBench [75] and *Partial Correctness* from PlanBench [149] exemplify this dimension.

Memory and Long-Term Consistency: This kind of evaluation evaluates whether the agent can retain and appropriately use historical context over long interactions. Sustained context understanding reflects cognitive continuity crucial for realistic deployments. For long-running agents, an important evaluation is whether the system remembers and utilizes prior context. In long-dialogue tests like SocialBench [17], after a 40-turn conversation, the agent might be quizzed on something from turn 5, a high score means it retained that detail and responded correctly.

Metrics here include information retention accuracy (*Did the agent recall facts or decisions made earlier?*) and context utilization (*Does the agent's behavior remain consistent with earlier states*). For example, an agent with a user profile should not contradict facts the user provided earlier. Some works also track memory footprint (how much context window or external memory the agent needed to store information) and whether it can retrieve relevant memories when needed. Essentially, these tests ensure the agent can handle long-horizon dependencies without forgetting or repeating mistakes. The *Stateful Evaluation* metric in τ -Bench [179] and the *Goal Drift* score in Adaptive Monitoring [127] both exemplify this category by quantifying consistency over time.

Multi-Agent Collaboration Metrics: This dimension measures the effectiveness of coordination, communication, and cooperation between agents or between humans and agents. When multiple agents work together or an agent interacts with humans, evaluation must capture coordination and social dynamics. Common metrics include team success rate [75, 16] (*Did the group achieve the goal, e.g. winning a game or completing a task, and how does that compare to a single agent alone?*). Time to completion or efficiency [159, 54] is measured for collaborative scenarios (e.g. two agents solve a puzzle faster than one).

Qualitatively, researchers look at emergent behaviors like, do agents develop communication protocols or show role specialization? Some studies use human judges or LLM critics to rate dialogues between agents for traits like coherence, persuasiveness, or adherence to expected human norms [108, 109]. There are also metrics for specific social phenomena like measuring conformity [167] (*will an agent change its correct answer if peers disagree?*), trust (in economic games, do agents exhibit trust and reciprocity similar to humans?), or consensus-building (how often a group of agents reaches agreement). These evaluations, though sometimes domain-specific, are crucial as agents become collaborators with humans or with each other. Collaborative evaluation is reflected as *Win Rate* in SWEET-RL [194].

⁵<https://blog.quotientai.co/evaluating-tool-calling-capabilities-in-large-language-models-a-literature-review/>

Robustness and Reliability: These metrics assess the stability and consistency of an agent’s behavior under perturbations or repeated trials. Autonomy requires reliability under varied conditions. Consistency [179] is a measure to check if the agent is run multiple times on the same input or task, does it produce the same outcome? Given the probabilistic nature of LLMs [114], identical queries might yield different answers, so researchers quantify this variance. A consistent agent is more predictable and trustworthy.

Robustness evaluation involve introducing perturbations, e.g. rephrasing the prompt, adding irrelevant data, or changing environmental variables, and seeing if the agent’s performance holds up [167, 68]. If a small change causes a large drop in success, the agent is brittle. Some adversarial evaluations send purposely tricky or misleading inputs to see if the agent can avoid traps. The degree of performance degradation under such perturbations is a key metric for robustness. Additionally, for long-run agents, researchers watch for error accumulation [33], meaning does the agent recover from mistakes or do they compound over time? In summary, these metrics aim to ensure the agent behavior is stable and predictable in non-ideal circumstances. A Typical example include *Pass^k* from τ -Bench [179] for repeatability for resilience to adversarial input.

Safety and Alignment Metrics: This category evaluates whether the agent’s behavior aligns with ethical norms, safety constraints, and policy rules. To assess whether an agent behaves in alignment with ethical and policy expectations, multiple measures are used. Toxicity and harm checks count how often the agent’s outputs contain hate speech, harassment, or other harmful content. Evaluations like AgentHarm [6] directly measure the frequency of unsafe responses or compliance with dangerous instructions.

Bias and fairness metrics examine if the agent’s decisions or content are biased against certain groups [113], for instance, does it yield different outcomes for different demographic profiles given the same task? Researchers may use bias benchmarks or constructed tests (e.g. changing the gender/race in a prompt and seeing if the answer quality changes) to quantify unfair behavior.

Compliance and policy adherence is another angle [68, 179], agents are often given guidelines, such as “do not reveal confidential info” or “refuse if asked for illegal advice”), and metrics can track the rate of policy violations. For example, how often does an agent complied to a disallowed request or leak private data?. AgentDojo [30] is a benchmark that deliberately attempt prompt injections and then scored the agent based on whether it resisted those attacks. Finally, some alignment tests involve red teaming the agent with creative adversarial prompts and then having humans rate the severity of any misbehavior [54]. A holistic evaluation of a goal-driven agent thus considers not just task proficiency but also trustworthiness, is it doing what it should and not doing what it shouldn’t? Safety and policy adherence metrics like the *Policy-Compliance/Safety Score* in IBM WatsonX [54] and the *Risk Ratio* in ST-WebAgentBench [68] exemplify this dimension.

7. Alignment with Regulation

We treat *governance* as who/what an agent may do under which conditions, *policy* as the enforceable rules (policy-as-code), and *audit* as verifiable evidence of what occurred. Unlike chat-style models where risk is concentrated in textual outputs, agentic AI systems change external state: they log into services, modify files, trigger workflows, and coordinate with other agents. Governance therefore shifts from content acceptability to action authorization, provenance, and accountability: what an agent is allowed to do, under which conditions, with what oversight, and how those actions are recorded and audited for post-hoc review. We show the overall pipeline of regulatory alignment using agentic AI in Figure 11 with details in Table 7 that shows mapping of regulatory alignment controls to mechanisms, measurable SLOs, audit artifacts, standards, and lifecycle phase.

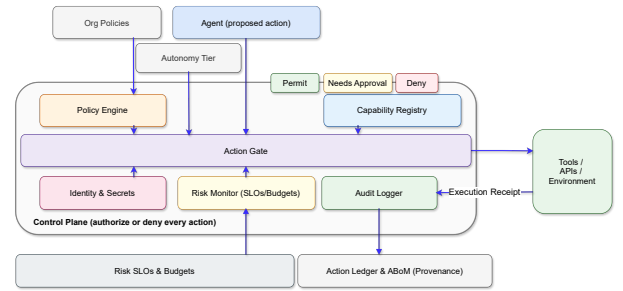


Figure 11: Regulatory alignment with Agentic AI

Control-plane view (core idea). We frame governance as a control plane that approves or denies every external action an agent proposes. The control plane enforces least-privilege credentials, scoped capabilities, policy-as-code checks, tool whitelists, and mandatory human approval at elevated risk. A policy engine evaluates each tool call against organizational rules and context (data sensitivity, user role, environment), aligning with zero-trust principles for continuous authorization rather than one-time approvals [119, 91].

Provenance and ABoM (making audits objective). To make audits objective, every step carries signed, query-able provenance: the agent signs intent (prompt, rationale, policy state) and the tool or API signs execution receipts (what was done, inputs/outputs, side effects). These dual attestations populate an Agent Bill of Materials (ABoM) analogous to software SBOMs, with SLSA/in-toto style provenance and verifiable credentials for capabilities. The result is a tamper-evident action trail suitable for accountability and cross-organisational handoff [148, 129, 142, 151]. In practice, Figure 11 provides the path; Table 7 lists what to measure and retain.

Risk SLOs and budgets (operational safety). Oversight is operationalization with risk service-level objectives (for example, unauthorized-action rate, policy-override incidence,

Table 6: Papers and their evaluation metric along with a description.

Paper Name	Evaluation metric	Description
MLAgentBench [50]	Success Rate	Percentage of runs where the agent improves the task-specific performance metric (e.g., test accuracy) by $\geq 10\%$ over a baseline
	Average Improvement	Mean relative improvement in the downstream metric (e.g., accuracy) over baseline across runs
	Wall-clock time	Time taken by the agent to complete its run; used to measure efficiency
	Token Usage	Number of input and output tokens used by the agent, another efficiency metric
MLE-Bench [15]	Medal rate	Agents earn Bronze/Silver/Gold medals based on Kaggle leaderboards; the headline metric is the fraction of attempts that receive any medal
	Raw competition scores	Reports the agent's score on each competition's own metric (e.g., AUROC, F1, RMSE), enabling fair comparison across tasks
	Pass@k	Measures success when the agent is given multiple attempts; pass@k improvements quantify reliability of repeated attempts.
	Resource-scaling metrics	Evaluates how longer time budgets or multiple seeds affect performance (e.g., 24-h runs vs. short runs).
SUPER [13]	Accuracy	Fraction of tasks whose final outputs (numerical or textual) exactly match the gold solution
	Landmark metric	Awards partial credit when the agent reaches intermediate "landmark" states (e.g., finishing training) even if final answers are wrong
	Script-Executed metric	Proxy measure for tasks without gold solutions; a task counts as successful if the provided script runs without exceptions for a minimum duration
PlanBench [149]	Success rate	Measures the proportion of planning tasks for which the generated plan is entirely correct (e.g., GPT-4 solved 26/600 deceptive Blocksworld instances)
	Partial correctness	The authors suggest future metrics that give credit for partially correct plans when not all preconditions/effects are satisfied.
WebArena [192], VisualArena [63]	Exact match	Text evaluation function that returns 1 if the agent's answer string exactly equals the ground-truth answer
	Must include	Checks that the generated response contains required keywords or phrases; partial match yields positive reward
	Must exclude	Rewards the agent if specific undesirable strings are absent from the output
	Fuzzy match	Uses an LLM to judge semantic equivalence between the agent's output and the ground truth; yields a binary reward when the answers are close
	eval_vqa	For visual tasks, queries a vision-language model with a question and rewards 1 if the returned answer contains the ground-truth answer
	eval fuzzy image match	Compares images using structural similarity (SSIM) to assess whether the agent produced the correct image
	Task success rate	Overall percentage of tasks that are completed successfully across all categories; derived from per-task rewards in WebArena/VisualWebArena.
FieldWorkArena [85]	Correctness score	Each response is labelled Correct, Incorrect or Partially Correct; partial answers receive a score between 0 and 1 representing the degree of agreement
	Weighted accuracy	Final accuracy calculated as a weighted average of the correctness score and a continuous numerical score; improves fuzzy matching for real-world tasks
τ -Bench [179]	Stateful evaluation/Policy Adherence	Compares the final database state after a task with the expected outcome to determine if the agent accomplished the goal (no human or LLM judge required)
	Pass ^k	Reliability metric that measures how often the same task is completed successfully over k independent trials; highlights consistency across runs
AgentBench [75]	Success Rate	Primary metric across eight environments; percentage of tasks in which the agent achieves the goal
	Win rate / Reward score	In digital card games, counts winning rounds, total rounds, win rate and damage rate, then computes a final reward score
	Answer F1	For knowledge-graph queries, computes F1 between the set of answers predicted by the agent and the ground truth
	Action F1	In web-browsing tasks, token-level F1 between predicted operations and ground truth; used as micro-level metric
	Step success rate	Fraction of steps where the agent selected the correct element and action
	Task success rate	Task is counted as successful only if all steps in the chain are correct; due to difficulty, AgentBench often reports step success rate instead
	Progress (lateral puzzles)	Measures the portion of guessed-out bullets when the agent fails to solve a puzzle within the allowed rounds
MINT [159]	Success rate	Percentage of tasks successfully completed under a fixed interaction limit
	Improvement rate	Slope obtained from regressing success rate against the number of allowed interaction turns; quantifies how additional turns improve performance
GAIA [81]	Quasi Exact Match	Automatically evaluates answers by normalizing the predicted answer and checking exact equality with the ground truth; ensures fast, factual scoring.
ToolLLM [103]	Pass rate	Automatic evaluator labels each tool-use solution path as Pass, Fail or Unsure based on whether the instruction was solvable and whether the agent retrieved correct information; pass rate is the proportion of Pass solutions
	Win rate	Pairwise comparison metric where two solution paths are judged on information richness, factuality, reasoning quality, milestones, exploration and cost; a majority vote across ChatGPT evaluations determines which path wins
ScaleAI ToolComp [87]	LLM grading	Uses a GPT-4 judge to compare the agent's final answer with the ground truth, classifying it as Correct, Correct with bad formatting or Incorrect; both Correct categories count as a win

Continued on next page

Evaluating and Regulating Agentic AI

Paper Name	Evaluation metric	Description
	Exact match	Programmatically checks sorted and unsorted lists, numbers (within tolerance) and strings in the final answer to ensure exact equality with ground truth
	Process supervision score	Evaluates how well an agent ranks human-corrected steps over model-generated steps; each evaluation yields 0 for a loss, 0.5 for a tie and 1 for a win
IBM WatsonX [54]	Task success / completion rate	fraction of tasks the agent completes successfully (standard task-level performance).
	Average steps / action count	mean number of actions taken to finish a task (efficiency).
	Latency / response time	time between request and agent's final action or decision (responsiveness).
	Policy-compliance / safety score	a scored measure of how often agent actions obey configured safety/policy checks (risk control).
	Risk / severity score	aggregated indicator of potential harms/violations produced by agent behavior (operational risk).
Rise of Agentic AI [9]	Explainability	Qualitative metric evaluating whether an agent's reasoning steps are understandable; assessed via self-reflection or cross-agent reflection for foundation models.
	Transparency	Degree to which the agent's internal decision processes are open for inspection; evaluated through user-facing clarity or external audits.
	User satisfaction	Measures how well the system meets user preferences; assessed via user ratings or Net Promoter Score
	Fairness / Bias mitigation	Metrics and techniques used to detect and reduce demographic bias in agent outputs
	Cooperative behavior	Evaluates how well multiple agents collaborate and coordinate to achieve common goals.
	Adaptability	Measures how quickly an agent can adjust to changing tasks or environments; tested via dynamic policy simulations.
	Robustness	Assesses the agent's ability to maintain performance despite internal failures or adversarial input; e.g., sandboxed execution and rollback for coding agents.
	Accuracy / Precision / Recall / F1 score	Standard classification metrics used to quantify correctness of agent outputs
	Graph Edit Distance (GED)	Measures structural similarity between the agent-generated task graph and the ground truth; lower GED indicates closer alignment.
	Rule fidelity	Evaluates how accurately the symbolic rules learned by the agent mirror the actual decision-making process.
	Task completion time (TCT)	Measures the time taken for the agent to plan and execute a task, providing an operational efficiency metric.
	Click-through rate (CTR) / Gross Merchandise Value (GMV)	Application-specific metrics that measure user engagement (CTR) and monetary impact of agent recommendations (GMV)
ALFWorld [126]	Progress Rate	Partial completion rate when full task success is not achieved.
	Success Rate	Task completion rate in text-based embodied household environments.
Adaptive Monitoring [127]	Adaptive Multi-Dimensional Score	Composite score using exponentially weighted moving averages across dimensions.
	Harm Reduction Score	Quantifies agent's ability to minimize potential negative outcomes.
	Goal Drift	Measure of how much an agent deviates from intended objectives over time.
HotPotQA [177]	Multi-hop F1	F1 score for multi-step reasoning chains in question answering.
	Supporting Fact F1	F1 score for identifying relevant supporting facts across reasoning steps.
Mind2Web [31]	Element Accuracy	Accuracy of selecting correct webpage elements for interaction.
	Operation F1	F1 score for correct operation prediction on web elements.
Mind2Web2 [44]	Partial Completion	Average root score measuring partial task completion.
	Agent-as-a-Judge Score	LLM-based evaluation score using tree-structured rubrics.
ST-WebAgentBench [68]	Completion Under Policy (CuP)	Success rate when both completing task and respecting all policy constraints.
	Risk Ratio	Quantifies policy violations across safety and trustworthiness dimensions.
SWE-PolyBench [107]	File-level Localization	Accuracy in identifying correct files requiring modification.
	CST Node-level Retrieval	Precision in locating specific code structures needing changes.
SWE-Bench	Resolved Rate	Percentage of GitHub issues successfully resolved by passing tests.
SWEET-RL [194]	Turn-wise Advantage	Advantage score measuring quality of each decision in multi-turn interactions.
	Win Rate	Rate of preferred responses in human-AI collaborative task completion.
TRAIL [33]	Error Detection Rate	Rate of correctly identifying errors in multi-step agent workflows.
	Joint Accuracy	Accuracy of identifying both error category and location in agent traces.
ToolBench [173]	Parameter Accuracy	Correctness of API call parameters when using tools.
	Tool Execution Success	Rate of successful tool executions without errors.
	Tool Selection Accuracy	Accuracy of choosing appropriate tools for given tasks.

Paper Name	Evaluation metric	Description
WebMall [99]	Price Comparison Accuracy	Accuracy in identifying best prices across different online shops.
	Cross-Shop Success Rate	Success rate for tasks requiring navigation across multiple e-commerce sites.
WebShop [178]	Attribute F1	F1 score for correctly identifying required product attributes.
	Task Success Rate	Success rate in completing realistic online shopping tasks.

prompt-injection success). Each agent or service receives a risk budget; when it is consumed, the control plane automatically tightens guardrails through rate limits, capability revocation, or human-in-the-loop escalation. This adapts SRE error-budget mechanics to agentic autonomy and makes safety thresholds a measurable, precommitted contract [139].

Continuous assurance (test update re-scope). Governance embeds adversarial testing against LLM-specific risks (prompt injection, insecure output handling, data poisoning) and AI threat tactics (model theft or evasion) using standard catalogs and red-team playbooks. Checks run pre-deployment and continuously at runtime (canary tasks, shadow policies), with findings feeding policy updates and capability scopes [93, 83].

Regulatory mapping and responsible scaling. Regulatory mapping and responsible scaling. The control-plane model maps to current guidance: lifecycle risk management and post-market monitoring (EU AI Act), management systems for AI (ISO/IEC 42001), and the NIST AI RMF functions of Govern, Map, Measure, and Manage. For higher-capability systems, responsible-scaling proposals tie increased autonomy to stronger safeguards and disclosure obligations [38, 1, 88, 7].

Practical checklist. (i) Define autonomy tiers and initial capability scopes. (ii) Route every tool call through a policy-as-code gate with continuous authorisation. (iii) Issue scoped, revocable capability credentials to agents and record dual-signed receipts. (iv) Publish an ABoM with provenance for each release. (v) Set risk SLOs and error budgets; wire circuit breakers to autonomy levels. (vi) Run continuous AI red teaming and feed results back into policy and scopes.

Autonomy level frameworks. Instead of treating “more autonomy” as an unregulated byproduct of capability, recent work proposes explicit classification and management. A five-level autonomy framework for agents, inspired by safety-critical domains, ranges from Level 1 (human is the operator; the agent acts only on explicit instructions) to Level 5 (the human is an observer of a fully self-directed agent). The taxonomy helps developers and policymakers specify permitted decision-making scope, and introduces autonomy certificates—labels issued by a trusted third party to make an agent’s autonomy level and key behaviours legible to integrators and other agents [39].

Industry governance analyses. Organisations are mapping the risk shift from assistants to agents and recommending controls. A 2025 industry whitepaper argues that tool

execution and goal-directed behaviour introduce new security and accountability concerns, and recommends stronger oversight checkpoints, restricted privileges, continuous monitoring, and stage-appropriate audits as deployments move from sandboxed pilots to production [124].

Regulatory and policy perspectives. Policy and safety proposals emphasise that autonomy increases accountability requirements. Guidance includes pre-deployment risk assessments, continuous monitoring, auditability of decisions, and assured human intervention or shutdown paths for higher-risk agents [38, 88, 7]. These directions align with the control-plane approach in which authorisation, provenance, and fallback mechanisms are first-class design elements.

Automated auditing and red-teaming. Recent research explores using specialised auditor agents to probe other systems for hidden goals, unsafe behaviours, or prompt-injection susceptibility. This combines breadth (large-scale scenario generation) with depth (trace inspection and targeted stress tests), offering a scalable complement to human reviews and enabling regression-style safety testing over time [80].

Practical oversight mechanisms. Practitioners report embedded compliance (rule checks and allowed/disallowed action models inside the agent loop), comprehensive action logging with rationales, manual-intervention triggers above defined risk thresholds, and adaptive governance structures that update policies as capabilities or regulations evolve [5, 39].

8. Discussion

In this section we synthesize progress and limitations in agentic AI evaluation and outline practical directions for the field.

8.1. Progress and Limitations in Agent Evaluation

Recent benchmarks and metric taxonomies increasingly treat agents as *decision-making systems* rather than text generators, however, the landscape remains fragmented. While, legacy NLP/RL metrics are useful baselines but they rarely capture end-to-end behavior across perception, planning, tool use, and control, motivating richer, process-aware schemes [64]. So, there is need for studying evaluation metrics for agentic AI, which motivated this study.

One solution to have more works on agentic AI evaluations is to study this shift from single outcome scores to *multidimensional scorecards* that combine task success with efficiency, reliability, and policy adherence. Beyond success rate or accuracy, stateful and governance-linked indicators (cf. Sections 5 and 6)—e.g., policy compliance, risk/severity

Table 7: Regulatory alignment controls mapped to mechanisms, measurable SLOs, audit artifacts, standards, and lifecycle phase.

Control Domain	Concrete Mechanisms (e.g.)	SLOs (e.g.)	Audit Artifacts	Standards Mapping	Lifecycle Phase
Action authorization	Policy-as-code (OPA/Rego); zero-trust continuous authZ; per-tool allow/deny lists; environmental guards (prod/stage); contextual checks (data sensitivity, user role); mandatory HIL for high-risk actions.	Unauthorized-action rate (UAR) per 1,000 actions; policy-override incidence/week; mean time to decision (MTTD); Policy bundle snapshots; evaluation logs (decision, inputs, effect); signed approval records; HIL transcripts; change tickets linking policy versions.	NIST AI RMF [88]: Govern/Manage; ISO/IEC 42001: 8.3, 8.5 [1]; EU AI Act: Risk mgmt & post-market monitoring [38].	Design, Deploy, Operate	—
Capability scoping & least privilege	Scoped, revocable capabilities (capability tokens / VC); per-tool scopes (read/list/write/execute); time-bounded privileges; kill-switch/circuit breakers bound to scopes.	Scope creep rate; Capability registry; VC/attestation lists; revocation logs; scope-to-policy matrix; breaker activation logs.	NIST: Map/Manage [88]; ISO 42001: 6.1, 8.2 [1]; EU AI Act: Technical documentation & controls [38].	Design, Operate	—
Identity & secrets management	Workload identity (OIDC/SPIFFE); secret vaulting/rotation; short-lived creds; device posture checks; mTLS between agent/control plane/tools.	Secret rotation interval compliance; secrets-in-logs incidence; unauthorized credential use; mTLS coverage.	KMS/VAULT rotation records; access logs; cert lifetimes; failed authZ attempts; secret scanning reports.	NIST: Govern/Manage [88]; ISO 42001: 8.7 [1]; EU AI Act: Security & robustness [38].	Build, Operate
Data governance & privacy	DPIA/TRA; data minimization; DLP on tool boundaries; PII redaction; purpose limitation tags; retention policies; dataset lineage.	PII leakage rate; DLP block rate; retention compliance; purpose-tag adherence; cross-border transfers tracked.	DPIA reports; data inventory; lineage graphs; redaction configs; retention/audit logs; access reviews.	GDPR / PIPEDA [147]; NIST: Map / Measure [88]; ISO 42001: 6.2, 8.8 [1]; EU AI Act: Risk mgmt/data [38].	Design, Operate
Provenance & ABoM (attested)	Dual-signed receipts (agent intent + tool execution); in-toto/SLSA attestations; JSON-LD/VC ABoM (version, tools, datasets, policies); hash-chained logs.	Provenance coverage % (actions with receipts); attestation verification rate; tamper-detect MTTC.	ABoM per release; in-toto metadata; SLSA provenance; hash chain proofs; verification reports.	NIST: Measure/Manage [88]; ISO 42001: 8.6 [1]; EU AI Act: Technical documentation [38]; NTIA SBOM/SLSA [148].	Build, Operate
Tool governance & sandboxing	Tool registry with schemas; static/dynamic policy checks; syscall/file/Network sandboxing; outbound egress allowlists; safe output handling (no-raw-exec).	Tool-call precision/recall; unsafe-output execution incidence; blocked egress attempts; sandbox escape rate.	Tool schemas; sandbox configs; egress logs; block/allow decisions; unsafe-output test results.	OWASP LLM [93]; NIST: Manage [88]; ISO 42001: 8.5 [1]; EU AI Act: Security controls [38].	Build, Operate
Runtime assurance & kill-switches	Canary tasks; shadow policies; runtime monitors; rate limiting; automatic de-scoping; kill-switch with human confirm.	MTTD/MTTC for unsafe patterns; breaker activation frequency; success degradation under shadow policy.	Monitor dashboards; anomaly alerts; breaker audits; shadow-policy replay traces.	NIST: Manage [88]; EU AI Act: Post-market monitoring [38]; ISO 42001: 9.1. [1]	Operate
Red teaming & adversarial testing	Prompt-injection suites; jailbreak corpora; insecure-output handling tests; data poisoning checks; SSRF/file-write probes; scheduled chaos tests.	Prompt-injection success rate; jailbreak rate; IOH incident rate; Red-team reports; scenario catalogs; reproduction seeds; mitigation diffs; regression dashboards.	MITRE ATLAS [83]; OWASP LLM [93]; NIST: Measure/Manage [88]; ISO 42001: 8.9. [1]	Pre-release, Continuous	—
Evaluation & judging (quality)	Programmatic assertions; trace-based scoring; diversified LLM judges + blinded human review; rubric calibration; leakage checks.	Inter-annotator κ ; LLM-human agreement; judge leakage rate; CI width for success@k; evaluation cost/task.	Rubrics; judge prompts; blinded samples; agreement stats; leakage tests; eval manifests.	NIST: Measure [88]; ISO 42001: 9.1 [1]; EU AI Act: Technical documentation [38].	Test, Operate
Risk SLOs & budgets	Pre-committed risk SLOs; budgets tied to autonomy tier; automatic policy tightening when budget consumed; weekly error review.	UAR, PIR (policy-override rate), PIR (prompt-jailbreak rate), MTTC; SLO docs; budget dashboards; review minutes; corrective actions; trend analyses.	NIST: Govern/Manage [88]; ISO 42001: 6.1, 9.1 [1]; SRE error-budgets [139].	Plan, Operate	—
Incident response & forensics	Runbooks; immutable logs; snapshotting state; comms templates; regulator notification workflow; post-mortems with CAPA.	MTTD/MTTR; IR runbooks; ticketing trails; snapshots; post-mortems; CAPA logs; regulator filings.	NIST: Manage [88]; ISO 42001: 8.10 [1]; EU AI Act: Serious incident reporting [38].	Operate	—
Change mgmt & versioning	Model/prompt/tool versioning; gated rollouts; evaluation gates; rollback plans; policy diffs; approvals.	Change failure rate; rollback success time; gate coverage %; un-gated change incidence.	Changelogs; approvals; eval reports; rollout logs; rollback evidence; policy diff history.	NIST: Govern/Manage [88]; ISO 42001: 8.1 [1]; EU AI Act [38]; QMS practices.	Build, Release
Third-party & supply chain	Vendor risk reviews; SBOM/ABoM ingestion; license compliance; API quotas/SLAs; cryptographic provenance verification.	% deps with verified provenance; vendor SLA breaches; license non-compliance rate.	Vendor assessments; SBOM/ABoM archives; license scans; SLA reports; verification logs.	NIST: Map/Manage [88]; ISO 42001: 8.4 [1]; NTIA SBOM [148]; SLSA [129].	Plan, Operate

Control Domain	Concrete Mechanisms (e.g.)	SLOs (e.g.)	Audit Artifacts	Standards Mapping	Lifecycle Phase
Human-in-the-loop & training	Risk tiering for HIL; escalation SLAs; operator training/cert; dual-control for high-impact actions; UX to reveal rationale.	HIL latency; override correctness rate; trained-operator coverage; dual-control adherence.	Training records; certification logs; HIL transcripts; UX screenshots; audit samples.	NIST: Govern/Manage [88]; ISO 42001: 7.2 [1]; EU AI Act: Human oversight [38].	Design, Operate
Disclosure & transparency	User-facing capability cards; autonomy level certificate; changelog of safeguards; data use notices; opt-out paths.	Timeliness of disclosures; completeness checklist score; user complaint rate.	Capability cards; L4–L5 autonomy labels; release notes; DPIA summaries; DSR logs.	EU AI Act: Transparency [38]; NIST: Govern [88]; ISO 42001: 8.6. [1]	Release, Operate

Abbreviations: HIL—Human-in-the-Loop; UAR—Unauthorized-Action Rate; PJR—Prompt-Jailbreak Rate; PIR—Policy-Override Rate; MTTD/MTTR/MTTC—Mean Time to Detect/Recover/Contain; DLP—Data Loss Prevention; VC—Verifiable Credential; SBOM/ABoM—(Software/Agent) Bill of Materials; SLSA—in-toto/Supply-chain Levels for Software Artifacts.

rates, and *completion under policy*—explicitly test whether an agent achieves goals while remaining within constraints. These families of metrics make failure modes legible for deployment stakeholders by tying behavior to organizational rules rather than only to ground truth.

It is also observed that evaluation settings are also evolving. Simulation and *HITL studies* probe temporal properties, coherence, goal drift, trust, and sustainability. This shows that agentic AI quality depends on strategies over trajectories, not just terminal outputs [84]. Emerging frameworks, therefore, track interaction histories and longitudinal effects to surface where plans degrade, safety constraints are bypassed, or coordination fails. Community testbeds now span realistic web tasks, OS/tool environments, and multi-environment suites [112]. This breadth improves ecological validity but complicates comparability. For example, task designs, reward functions, and supervision styles vary widely, which makes cross-benchmark generalization claims weaker unless evaluations report both *outcome* and *process* metrics with shared definitions.

A key limitation in prior seminal work lies in the weak linkage between performance metrics and governance requirements [84]. On one hand, agentic AI systems are rapidly gaining traction in the market for their autonomy and ease of use; on the other hand, governance frameworks increasingly demand transparency, accountability, and risk-aware operation [140]. Prior surveys (Table 1) have largely focused on component taxonomies and application domains, offering limited guidance on how to evaluate robustness, reliability, efficiency, safety, alignment, and governance in an integrated manner. This gap underscores the need to align technical performance signals with oversight mechanisms, such as coupling task success or pass@k with policy adherence, risk ratios, and auditability indicators. In this work, we highlight the gap and propose recommendations that how current evaluation practises can be applied on it.

However, as with any study, our review also has certain limitations. First, it focuses on the 2023–2025 period and primarily includes publicly documented benchmarks. This scope was chosen to capture the most recent developments in the field; however, some foundational studies and rapidly evolving industrial practices may be underrepresented. Second, while we attempted to include a diverse range of model families; both proprietary and open-source, proprietary

evaluations and confidential red-teaming exercises remained beyond our scope. Consequently, our synthesis may not fully reflect the depth of safety, reliability, and governance testing occurring within private ecosystems.

To make results deployable, outcome and process metrics should bind to governance levers (as discussed in Section 7). Some recommendations are to : (i) promote *completion-under-policy* to a first-class metric; (ii) report *risk SLOs* (e.g., unauthorized-action rate, policy-override, prompt-injection success) alongside success@k; and (iii) surface *trace auditability* (presence of signed actions, rationale visibility, reproducible state checks). This turns benchmark scores into inputs for risk budgets and release gates rather than standalone leaderboards.

8.2. Open Challenges and Future Directions

Two design directions appear especially promising. (i) Trajectory-first evaluation: instead of only judging the final answers, evaluation should look at the entire process: including the model’s thoughts, actions, observations, and intermediate steps. This helps create a clearer picture of how decisions are made and allows better auditing and supervision of the full reasoning process (for example, through “process-trace audits” when possible). (ii) State-grounded scoring: evaluations should rely on verifiable facts, such as database entries or file-system states, so results are less subjective. This approach allows us to check both whether the goal was achieved and whether the model followed the correct rules and policies. Several open challenges merit further attention, together with brief remedies, which we discuss below:

- Agents are seldom assessed under distributional shift, in recovery from errors, or during post-failure adaptation. We recommend including leave-one-domain-out and counterfactual splits, adding recovery-after-fault tasks, and reporting the out-of-distribution (OOD) gap, recovery rate, and regret alongside success.
- Reporting on robustness and reliability is inconsistent in AI solutions including Agentic AI. We recommend disclosing seed sweeps and prompt/parameter ablations; running input-perturbation tests; reporting confidence intervals and variance; and releasing configurations for reproduction [114].

- Human-centered outcomes (trust, satisfaction, cognitive load) require clearer operational definitions and instrumentation tied to real interaction logs. We recommend predefining constructs and rubrics, using validated instruments where possible, and pairing surveys with log-derived proxies (e.g., intervention counts, help requests) [106].
- Risks of test contamination (training exposure) and judge leakage (LLM or human access to gold labels) may inflate results. We advise de-duplicating against training corpora (hash/MinHash), tracking dataset provenance, blinding judges, and maintaining concealed holdout sets with leakage checks [111].
- Environment drift (changes to web, OS, or UI) and tool-affordance confounds (permissions, quotas) complicate comparability [150]. We recommend containerizing tasks with versioned manifests, recording environment fingerprints and tool permissions, and providing replayable traces.
- Multi-step settings introduce error propagation, whereby early mistakes disproportionately affect outcomes. We recommend reporting step-level correctness, recovery rate, and time-to-recovery, and using plan/trace verification and targeted restart tests to localize failures.
- Stochastic factors (seeds, decoding choices, network latency) and user-interface instrumentation can modulate behavior. We recommend standardizing decoding and seeds, running multi-seed evaluations with confidence intervals, controlling latency where feasible, and documenting instrumentation effects.
- Efficiency and sustainability are also often under-reported. We recommend reporting unit economics for each setting: tokens, wall-clock time, steps, memory, and \$ per successful task [61]; and, where relevant, adding energy/CO₂e estimates to reveal trade-offs on the efficiency frontier (e.g., success versus cost/energy).

Looking ahead, we advocate *evaluation cards* that uniformly report: (i) outcome metrics (success, graded accuracy, pass@k), (ii) process metrics (plan/step correctness, error localization, recovery rate), (iii) resource and temporal efficiency (steps, tokens, wall-clock, \$ per success), and (iv) governance metrics (policy adherence, risk ratios, trace auditability). Applied consistently across web, OS/tool, and multi-agent settings, these additions align evaluation with the control-plane view in Section 7, turning benchmark results into actionable risk SLOs, release gates, and auditable traces.

8.3. Cognitive Architectures and LLM-driven Agentic AI

While recent research in Agentic AI and AI Agents is rapidly progressing in relation to LLM-based agents, further promising research threads are offered through lessons

learned from literature on Intelligent, Autonomous, or Cognitive Agents [62, 146]. Research in this area benefits from philosophical [43] and psychological [165] perspectives, which build into cognitive architectures motivational elements regarding what drives agentic actions towards goal generation and consequent planning to attain goals [11, 36, 66]. Motivational concepts can be relevant to architectures following both a cognitive science perspective [28, 89, 118, 135, 165], benefitting from a meta-cognition layer [4], as well as emergent architectural development through learning [130]. Most recent developments in cognitive systems architectures combine numerical, symbolic and sub-symbolic computation, and hybridize aspects of cognitive systems and emergent structures, shaped by learning [67]. LLM-based architectures are clearly of emergent type.

In practice, for the adoption of Agentic AI, an operational perspective is needed for auditing and evaluation. Under the assumption that a key aim in Agentic AI is to offer operational autonomy, benchmarking would need to consider both domain-agnostic and domain-specific aspects of such autonomy, with the autonomy essentially resulting from computational operations [36]. Domain-specific evaluation would need to look deeper into what is operationally meaningful in the content of the applicable application domain. Overall, further work on benchmarking Agentic AI could look at the autonomy from the operational, computational, and alignment perspectives. Therefore, it would require looking into (a) operational efficiency and effectiveness of Agentic AI; (b) quality and performance in terms of cognitive functions (e.g. memory, attention mechanisms, goal generation, decision-making, planning etc); (c) computational resources; and (d) values, ethical, policies, controllability and regulatory alignment [27].

9. Conclusion

This study highlights the growing need for structured evaluation and regulation to guide the development of agentic AI systems. While prior surveys have primarily focused on definitions, architectures, and applications, our work bridges a critical gap by systematically organizing benchmarks, metrics, and governance approaches for assessing autonomous, LLM-base agents. We show that evaluating agentic AI requires multi-dimensional assessment that goes beyond accuracy or task success to include reasoning depth, adaptability, cooperation, ethical alignment, and safety. Furthermore, linking these evaluations with emerging regulatory frameworks such as the EU AI Act, NIST AI RMF, and ISO/IEC 42001 offers a pathway toward responsible deployment. Future research should aim to develop unified, process-aware evaluation suites and policy-linked metrics that capture the full lifecycle of agentic systems. We encourage researchers, policymakers, and developers to advance trustworthy, auditable, and human-aligned agentic AI through rigorous technical evaluation and proactive governance integration.

Declarations

Acknowledgments and Funding Funding for the research was partly provided through Horizon Europe project AIXpert: An agentic, multi-layer, GenAI-powered backbone to make an AI system explainable, accountable, and transparent (ID:101214389).

Declaration of Competing Interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contributions **A.F., S.R.:** Conceptualization, Methodology, Literature review, Writing. **A.F.** Tables, Figures, Bibliographic analysis, Writing, Re-writing, Editing. **S.R.** Supervision, Detailed review, Re-writing, Organizing. **N.K.:** Conceptualization, Review. **H.I.:** Review, Figures. **A.V., C.E.:** Supervision, Detailed Review, Edits. All authors have read and approved the final manuscript.

References

- [1] , 2023. Information technology – artificial intelligence – management system. 51 pp.
- [2] Acharya, D.B., Kuppan, K., Divya, B., 2025. Agentic ai: Autonomous intelligence for complex goals—a comprehensive survey. IEEe Access .
- [3] Achim, T., Best, A., Der, K., Fédérico, M., Gukov, S., Halpern-Leister, D., Henningsgard, K., Kudryashov, Y., Meiburg, A., Michelsen, M., et al., 2025. Aristotle: Imo-level automated theorem proving. arXiv preprint arXiv:2510.01346 .
- [4] Agashe, S., Fan, Y., Reyna, A., Wang, X.E., 2023. Llm-coordination: evaluating and analyzing multi-agent coordination abilities in large language models. arXiv preprint arXiv:2310.03903 .
- [5] Ahmed, S.Q., 2025. Agentic ai: A governance wake-up call. Directorship Magazine, Online Exclusive URL: <https://www.nacdonline.org/all-governance/governance-resources/directorship-magazine/online-exclusives/2025/q3-2025/autonomous-artificial-intelligence-oversight/>. accessed: September 30, 2025.
- [6] Andriushchenko, M., Souly, A., Dziemian, M., Duenas, D., Lin, M., Wang, J., Hendrycks, D., Zou, A., Kolter, Z., Fredrikson, M., et al., 2024. Agentharm: A benchmark for measuring harmfulness of llm agents. arXiv preprint arXiv:2410.09024 .
- [7] Anthropic, 2025. Responsible Scaling Policy, Version 2.1. Technical Report. Anthropic. URL: <https://www-cdn.anthropic.com/17310f6d70ae5627f55313ed067afc1a762a4068.pdf>.
- [8] Artacho, B., Savakis, A., 2021. Omnipose: A multi-scale framework for multi-person pose estimation. arXiv preprint arXiv:2103.10180 .
- [9] Bandi, A., Kongari, B., Naguru, R., Pasnoor, S., Vilipala, S.V., 2025. The rise of agentic ai: A review of definitions, frameworks, architectures, applications, evaluation metrics, and challenges. Future Internet 17, 404.
- [10] Belcak, P., Heinrich, G., Diao, S., Fu, Y., Dong, X., Muralidharan, S., Lin, Y.C., Molchanov, P., 2025. Small language models are the future of agentic ai. arXiv preprint arXiv:2506.02153 .
- [11] Berto, L., Tanevska, A., Cirne, A., Costa, P., Simoes, A., Gudwin, R., Rea, F., Colombini, E., Sciutti, A., 2025. Curiosity and affect-driven cognitive architecture for hri. IEEE Transactions on Affective Computing doi:10.1109/TAFFC.2025.3551512.
- [12] Bogavelli, T., Sharma, R., Subramani, H., 2025. Agentarch: A comprehensive benchmark to evaluate agent architectures in enterprise. arXiv preprint arXiv:2509.10769 .
- [13] Bogin, B., Yang, K., Gupta, S., Richardson, K., Bransom, E., Clark, P., Sabharwal, A., Khot, T., 2024. Super: Evaluating agents on setting up and executing tasks from research repositories. arXiv preprint arXiv:2409.07440 .
- [14] Cervantes, S., López, S., Cervantes, J.A., 2020. Toward ethical cognitive architectures for the development of artificial moral agents. Cognitive systems research 64, 117–125.
- [15] Chan, J.S., Chowdhury, N., Jaffe, O., Aung, J., Sherburn, D., Mays, E., Starace, G., Liu, K., Maksin, L., Patwardhan, T., et al., 2024. Mle-bench: Evaluating machine learning agents on machine learning engineering. arXiv preprint arXiv:2410.07095 .
- [16] Chang, M., Zhang, J., Zhu, Z., Yang, C., Yang, Y., Jin, Y., Lan, Z., Kong, L., He, J., 2024. Agentboard: An analytical evaluation board of multi-turn llm agents. Advances in neural information processing systems 37, 74325–74362.
- [17] Chen, H., Chen, H., Yan, M., Xu, W., Gao, X., Shen, W., Quan, X., Li, C., Zhang, J., Huang, F., et al., 2024a. Socialbench: Sociality evaluation of role-playing conversational agents. arXiv preprint arXiv:2403.13679 .
- [18] Chen, J., Hu, X., Liu, S., Huang, S., Tu, W.W., He, Z., Wen, L., 2024b. Llmarena: Assessing capabilities of large language models in dynamic multi-agent environments. arXiv preprint arXiv:2402.16499 .
- [19] Chen, J., Yuen, D., Xie, B., Yang, Y., Chen, G., Wu, Z., Yixing, L., Zhou, X., Liu, W., Wang, S., et al., 2024c. Spa-bench: A comprehensive benchmark for smartphone agent evaluation, in: NeurIPS 2024 Workshop on Open-World Agents.
- [20] Chen, L., Gu, J., Huang, L., Huang, W., Jiang, Z., Jie, A., Jin, X., Jin, X., Li, C., Ma, K., et al., 2025. Seed-prover: Deep and broad reasoning for automated theorem proving. arXiv preprint arXiv:2507.23726 .
- [21] Chen, L., Yan, F., Zhong, Y., Chen, S., Jie, Z., Ma, L., 2024d. Mindbench: A comprehensive benchmark for mind map structure recognition and analysis. arXiv preprint arXiv:2407.02842 .
- [22] Chen, W., Ma, X., Wang, X., Cohen, W.W., 2023. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. URL: <https://arxiv.org/abs/2211.12588>, arXiv:2211.12588.
- [23] Chen, Z., Chen, S., Ning, Y., Zhang, Q., Wang, B., Yu, B., Li, Y., Liao, Z., Wei, C., Lu, Z., et al., 2024e. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. arXiv preprint arXiv:2410.05080 .
- [24] Chen, Z., Du, W., Zhang, W., Liu, K., Liu, J., Zheng, M., Zhuo, J., Zhang, S., Lin, D., Chen, K., Zhao, F., 2024f. T-eval: Evaluating the tool utilization capability of large language models step by step. URL: <https://arxiv.org/abs/2312.14033>, arXiv:2312.14033.
- [25] Cherepanov, E., Kachaev, N., Kovalev, A.K., Panov, A.I., 2025. Memory, benchmark & robots: A benchmark for solving complex tasks with reinforcement learning. arXiv preprint arXiv:2502.10550 .
- [26] Chezelles, D., Le Sellier, T., Shayegan, S.O., Jang, L.K., Lù, X.H., Yoran, O., Kong, D., Xu, F.F., Reddy, S., Cappart, Q., et al., 2024. The browsergym ecosystem for web agent research. arXiv preprint arXiv:2412.05467 .
- [27] Choi, D., Langley, P., 2018. Evolution of the icarus cognitive architecture. Cognitive Systems Research 48, 25–38. doi:10.1016/j.cogsys.2017.05.005.
- [28] Cox, M.T., Dannenhauer, D., Kondrakunta, S., 2017. Goal operations for cognitive systems URL: www.aaai.org.
- [29] Dannenhauer, D., Gogineni, V.R., Kondrakunta, S., Mitchell, A., Cox, M.T., 2020. Midca version 1.5: User manual and tutorial for the metacognitive integrated dual-cycle architecture. Tech. Rep. No. COLAB2-TR-5 .
- [30] Debenedetti, E., Zhang, J., Balunovic, M., Beurer-Kellner, L., Fischer, M., Tramèr, F., 2024. Agentdojo: A dynamic environment to evaluate prompt injection attacks and defenses for llm agents. Advances in Neural Information Processing Systems 37, 82895–82920.
- [31] Deng, X., Gu, Y., Zheng, B., Chen, S., Stevens, S., Wang, B., Sun, H., Su, Y., 2023. Mind2web: Towards a generalist agent for the web. Advances in Neural Information Processing Systems 36, 28091–28114.

- [32] Derouiche, H., Brahmi, Z., Mazeni, H., 2025. Agentic ai frameworks: Architectures, protocols, and design challenges. *arXiv preprint arXiv:2508.10146*.
- [33] Deshpande, D., Gangal, V., Mehta, H., Krishnan, J., Kannappan, A., Qian, R., 2025. Trail: Trace reasoning and agentic issue localization. *arXiv preprint arXiv:2505.08638*.
- [34] Dong, Y., Mu, R., Jin, G., Qi, Y., Hu, J., Zhao, X., Meng, J., Ruan, W., Huang, X., 2024a. Building guardrails for large language models. *arXiv preprint arXiv:2402.01822*.
- [35] Dong, Y., Zhu, X., Pan, Z., Zhu, L., Yang, Y., 2024b. Villageragent: A graph-based multi-agent framework for coordinating complex task dependencies in minecraft. *arXiv preprint arXiv:2406.05720*.
- [36] Estany, A., Martínez, S., 2014. "scaffolding" and "affordance" as integrative concepts in the cognitive sciences. *Philosophical Psychology* 27, 98–111. doi:10.1080/09515089.2013.828569.
- [37] et.al., G.T., 2024. Gemma: Open models based on gemini research and technology. URL: <https://arxiv.org/abs/2403.08295>, *arXiv:2403.08295*.
- [38] European Union, 2024. Regulation (eu) 2024/1689 of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act). Official Journal of the European Union, OJ L 2024/1689, 12 July 2024. URL: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>.
- [39] Feng, K., McDonald, D.W., Zhang, A.X., 2025. Levels of autonomy for ai agents. *Essays and Scholarship*, 25-15 Knight First Amend. Inst. URL: <https://knightcolumbia.org/content/levels-of-autonomy-for-ai-agents-1>, accessed: September 30, 2025.
- [40] Geng, L., Chang, E.Y., 2025. Realm-bench: A real-world planning benchmark for llms and multi-agent systems. *arXiv preprint arXiv:2502.18836*.
- [41] Gioacchini, L., Siracusano, G., Sanvito, D., Gashteovski, K., Friede, D., Bifulco, R., Lawrence, C., 2024. Agentquest: A modular benchmark framework to measure progress and improve llm agents. *arXiv preprint arXiv:2404.06411*.
- [42] Gong, R., Huang, Q., Ma, X., Vo, H., Durante, Z., Noda, Y., Zheng, Z., Zhu, S.C., Terzopoulos, D., Fei-Fei, L., et al., 2023. Mindagent: Emergent gaming interaction. *arXiv preprint arXiv:2309.09971*.
- [43] González-Santamarta, M.A., Rodríguez-Lera, F.J., Ángel Manuel Guerrero-Higueras, Matellán-Olivera, V., 2023. Integration of large language models within cognitive architectures for autonomous robots URL: <http://arxiv.org/abs/2309.14945>.
- [44] Gou, B., Huang, Z., Ning, Y., Gu, Y., Lin, M., Qi, W., Kopanav, A., Yu, B., Gutiérrez, B.J., Shu, Y., et al., 2025. Mind2web 2: Evaluating agentic search with agent-as-a-judge. *arXiv preprint arXiv:2506.21506*.
- [45] Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N.V., Wiest, O., Zhang, X., 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.
- [46] Haase, J., Pokutta, S., 2025. Beyond static responses: Multi-agent llm systems as a new paradigm for social science research. *arXiv preprint arXiv:2506.01839*.
- [47] Huang, B., Yu, Y., Huang, J., Zhang, X., Ma, J.W., 2025a. Dca-bench: A benchmark for dataset curation agents, in: *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 5482–5492.
- [48] Huang, C., Yu, W., Wang, X., Zhang, H., Li, Z., Li, R., Huang, J., Mi, H., Yu, D., 2025b. R-zero: Self-evolving reasoning llm from zero data. *arXiv preprint arXiv:2508.05004*.
- [49] Huang, J., Chen, L., Guo, T., Zeng, F., Zhao, Y., Wu, B., Yuan, Y., Zhao, H., Guo, Z., Zhang, Y., et al., 2024. Mmevalpro: Calibrating multimodal benchmarks towards trustworthy and efficient evaluation. *arXiv preprint arXiv:2407.00468*.
- [50] Huang, Q., Vora, J., Liang, P., Leskovec, J., 2023a. Mlagentbench: Evaluating language agents on machine learning experimentation. *arXiv preprint arXiv:2310.03302*.
- [51] Huang, Y., Shi, J., Li, Y., Fan, C., Wu, S., Zhang, Q., Liu, Y., Zhou, P., Wan, Y., Gong, N.Z., et al., 2023b. Metatool benchmark for large language models: Deciding whether to use tools and which to use. *arXiv preprint arXiv:2310.03128*.
- [52] Hughes, L., Dwivedi, Y.K., Malik, T., Shawosh, M., Albashrawi, M.A., Jeon, I., Dutot, V., Appenderanda, M., Crick, T., De', R., et al., 2025. Ai agents and agentic systems: A multi-expert analysis. *Journal of Computer Information Systems*, 1–29.
- [53] Hyun, J., Waytowich, N.R., Chen, B., 2025. Crew-wildfire: Benchmarking agentic multi-agent collaborations at scale. *arXiv preprint arXiv:2507.05178*.
- [54] IBM, . Agentic AI evaluation. <https://www.ibm.com/docs/en/watsonx/saas?topic=sdk-agentic-ai-evaluation>. Accessed: 5 October 2025.
- [55] Jain, E., Roy, I., Meher, S., Chakrabarti, S., De, A., 2024. Graph edit distance with general costs using neural set divergence. URL: <https://arxiv.org/abs/2409.17687>, *arXiv:2409.17687*.
- [56] Jha, S., Arora, R., Watanabe, Y., Yanagawa, T., Chen, Y., Clark, J., Bhavya, B., Verma, M., Kumar, H., Kitahara, H., et al., 2025. Itbench: Evaluating ai agents across diverse real-world it automation tasks. *arXiv preprint arXiv:2502.05352*.
- [57] Jimenez, C.E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., Narasimhan, K., 2023. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*.
- [58] Kang, H., Zhang, Q., Cai, H., Xu, W., Krishna, T., Du, Y., Weissman, T., 2025. Win fast or lose slow: Balancing speed and accuracy in latency-sensitive decisions of llms. URL: <https://arxiv.org/abs/2505.19481>, *arXiv:2505.19481*.
- [59] Kapoor, R., Butala, Y.P., Russak, M., Koh, J.Y., Kamble, K., AlShikh, W., Salakhutdinov, R., 2024. Omniact: A dataset and benchmark for enabling multimodal generalist autonomous agents for desktop and web, in: *European Conference on Computer Vision*, Springer. pp. 161–178.
- [60] Kapoor, S., Stroebl, B., Kirgis, P., Nadgir, N., Siegel, Z.S., Wei, B., Xue, T., Chen, Z., Chen, F., Utpala, S., Ndzomga, F., Oruganty, D., Luskin, S., Liu, K., Yu, B., Arora, A., Hahm, D., Trivedi, H., Sun, H., Lee, J., Jin, T., Mai, Y., Zhou, Y., Zhu, Y., Bommasani, R., Kang, D., Song, D., Henderson, P., Su, Y., Liang, P., Narayanan, A., 2025. Holistic agent leaderboard: The missing infrastructure for ai agent evaluation. <https://github.com/princeton-plti/hal-harness>.
- [61] Khan, T., Motie, S., Kocak, S.A., Raza, S., 2025. Optimizing large language models: Metrics, energy efficiency, and case study insights. *arXiv preprint arXiv:2504.06307*.
- [62] Khodaygani, M., Ali, A.T., Dohnke, T., Groth, T., Baake, E., Leucker, M., Russwinkel, N., 2025. Cognitive modeling of agents: Integrating emotions, goals, needs, and decision-making.
- [63] Koh, J.Y., Lo, R., Jang, L., Duvvur, V., Lim, M.C., Huang, P.Y., Neubig, G., Zhou, S., Salakhutdinov, R., Fried, D., 2024. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. *arXiv preprint arXiv:2401.13649*.
- [64] Kokel, H., Katz, M., Srinivas, K., Sohrabi, S., 2025. Acpbench: Reasoning about action, change, and planning, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 26559–26568.
- [65] Kumar, S., Jeon, H.J., Lewandowski, A., Van Roy, B., 2024. The need for a big world simulator: A scientific challenge for continual learning. *arXiv preprint arXiv:2408.02930*.
- [66] Langley, P., 2017. Interactive cognitive systems and social intelligence. *IEEE Intelligent Systems* 32, 22–30. doi:10.1109/MIS.2017.3121556.
- [67] Langley, P., Laird, J.E., Rogers, S., 2009. Cognitive architectures: Research issues and challenges. *Cognitive Systems Research* 10, 141–160. URL: <http://dx.doi.org/10.1016/j.cogsys.2006.07.004>, doi:10.1016/j.cogsys.2006.07.004.
- [68] Levy, I., Wiesel, B., Marreed, S., Oved, A., Yaeli, A., Shlomov, S., 2024. St-webagentbench: A benchmark for evaluating safety and trustworthiness in web agents. *arXiv preprint arXiv:2410.06703*.
- [69] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al., 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33, 9459–9474.

- [70] Li, K., Jiang, M., Fu, D., Wu, Y., Hu, X., Wang, D., Liu, P., 2025. Datasetresearch: Benchmarking agent systems for demand-driven dataset discovery. arXiv preprint arXiv:2508.06960 .
- [71] Li, L., Wang, Y., Zhao, H., Kong, S., Teng, Y., Li, C., Wang, Y., 2024a. Reflection-bench: Evaluating epistemic agency in large language models. arXiv preprint arXiv:2410.16270 .
- [72] Li, M., Zhao, S., Wang, Q., Wang, K., Zhou, Y., Srivastava, S., Gokmen, C., Lee, T., Li, E.L., Zhang, R., et al., 2024b. Embodied agent interface: Benchmarking llms for embodied decision making. Advances in Neural Information Processing Systems 37, 100428–100534.
- [73] Li, M., Zhao, Y., Yu, B., Song, F., Li, H., Yu, H., Li, Z., Huang, F., Li, Y., 2023. Api-bank: A comprehensive benchmark for tool-augmented llms. arXiv preprint arXiv:2304.08244 .
- [74] Lin, J., Xia, Y., Zhang, J., Yan, K., Lu, L., Luo, J., Zhang, L., 2024. Ct-glip: 3d grounded language-image pretraining with ct scans and radiology reports for full-body scenarios. arXiv preprint arXiv:2404.15272 .
- [75] Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., Zhang, S., Deng, X., Zeng, A., Du, Z., Zhang, C., Shen, S., Zhang, T., Su, Y., Sun, H., Huang, M., Dong, Y., Tang, J., 2023. Agentbench: Evaluating llms as agents. arXiv preprint arXiv:2308.03688 .
- [76] Liu, Y., Peng, X., Cao, J., Zhang, Y., Zhang, X., Cheng, S., Wang, X., Yin, J., Du, T., 2024. Tool-planner: Task planning with clusters across multiple tools. arXiv preprint arXiv:2406.03807 .
- [77] Lu, X., Chen, Z., Hu, X., Zhou, Y., Zhang, W., Liu, D., Sheng, L., Shao, J., 2025. Is-bench: Evaluating interactive safety of vlm-driven embodied agents in daily household tasks. arXiv preprint arXiv:2506.16402 .
- [78] Ma, Z., Huang, W., Zhang, J., Gupta, T., Krishna, R., 2024. m & m's: A benchmark to evaluate tool-use for m multi-step m multi-modal tasks, in: European Conference on Computer Vision, Springer. pp. 18–34.
- [79] Mandi, Z., Jain, S., Song, S., 2024. Roco: Dialectic multi-robot collaboration with large language models, in: 2024 IEEE International Conference on Robotics and Automation (ICRA), IEEE. pp. 286–299.
- [80] Marks, S., Treutlein, J., Bricken, T., Lindsey, J., Marcus, J., Mishra-Sharma, S., Ziegler, D., Ameisen, E., Batson, J., Belonax, T., Chen, B., Cunningham, H., Denison, C., Golechha, S., Khan, A., Kirchner, J., Leike, J., Meek, A., Nishimura-Gasparian, K., Ong, E., Olah, C., Pearce, A., Roger, F., Salle, J., Shih, A., Tong, M., Thomas, D., Rivoire, K., Jermyn, A., MacDiarmid, M., Henighan, T., Hubinger, E., 2025. Auditing language models for hidden objectives. URL: <https://arxiv.org/abs/2503.10965>, arXiv:2503.10965.
- [81] Mialon, G., Fourrier, C., Wolf, T., LeCun, Y., Scialom, T., 2023. Gaia: a benchmark for general ai assistants, in: The Twelfth International Conference on Learning Representations.
- [82] Miserendino, S., Wang, M., Patwardhan, T., Heidecke, J., 2025. Swe-lancer: Can frontier llms earn \$1 million from real-world freelance software engineering? arXiv preprint arXiv:2502.12115 .
- [83] MITRE Corporation, 2025. Atlas: Adversarial threat landscape for artificial-intelligence systems. <https://atlas.mitre.org/>. Accessed 2025-10-18.
- [84] Mohammadi, M., Li, Y., Lo, J., Yip, W., 2025. Evaluation and benchmarking of llm agents: A survey, in: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, pp. 6129–6139.
- [85] Moteki, A., Masui, S., Yang, F., Song, Y., Bisk, Y., Neubig, G., Kusajima, I., Watanabe, Y., Ishida, H., Takahashi, J., et al., 2025. Fieldworkarena: Agentic ai benchmark for real field work tasks. arXiv preprint arXiv:2505.19662 .
- [86] Nakajima, Y., 2023. Babyagi: An autonomous task management system. <https://github.com/yoheinakajima/babyagi>. URL: <https://github.com/yoheinakajima/babyagi>. original release date: March 28, 2023. Accessed: [Insert Current Date].
- [87] Nath, V., Raja, P., Yoon, C., Hendryx, S., 2025. Toolcomp: A multi-tool reasoning & process supervision benchmark. arXiv preprint arXiv:2501.01290 .
- [88] National Institute of Standards and Technology (NIST), 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Technical Report NIST AI 100-1. NIST. URL: <https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf>.
- [89] Naya-Varela, M., Faina, A., Duro, R.J., 2021. Morphological development in robotic learning: A survey. IEEE Transactions on Cognitive and Developmental Systems 13, 750–768. doi:10.1109/TCDS.2021.3052548.
- [90] Nisa, U., Shirazi, M., Saip, M.A., Pozi, M.S.M., 2025. Agentic ai: The age of reasoning—a review. Journal of Automation and Intelligence .
- [91] Open Policy Agent Project, 2024. Open policy agent: Policy as code. <https://openpolicyagent.org/docs/>. Accessed 2025-10-18.
- [92] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al., 2022. Training language models to follow instructions with human feedback. Advances in neural information processing systems 35, 27730–27744.
- [93] OWASP Foundation, 2025. Owasp top 10 for large language model applications (2025). <https://genai.owasp.org/llm-top-10/>. Accessed 2025-10-18.
- [94] Padigela, H., Shah, C., Juyal, D., 2025. MI-dev-bench: Comparative analysis of ai agents on ml development workflows. arXiv preprint arXiv:2502.00964 .
- [95] Pan, Y., Kong, D., Zhou, S., Cui, C., Leng, Y., Jiang, B., Liu, H., Shang, Y., Zhou, S., Wu, T., et al., 2024. Webcanvas: Benchmarking web agents in online environments. arXiv preprint arXiv:2406.12373 .
- [96] Park, J.S., O'Brien, J., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S., 2023. Generative agents: Interactive simulacra of human behavior, in: Proceedings of the 36th annual acm symposium on user interface software and technology, pp. 1–22.
- [97] Pasukonis, J., Lillicrap, T., Hafner, D., 2022. Evaluating long-term memory in 3d mazes. arXiv preprint arXiv:2210.13383 .
- [98] Patil, S.G., Zhang, T., Wang, X., Gonzalez, J.E., 2024. Gorilla: Large language model connected with massive apis. Advances in Neural Information Processing Systems 37, 126544–126565.
- [99] Peeters, R., Steiner, A., Schwarz, L., Caspar, J.Y., Bizer, C., 2025. Webmall—a multi-shop benchmark for evaluating web agents. arXiv preprint arXiv:2508.13024 .
- [100] Piccialli, F., Chiaro, D., Sarwar, S., Cerciello, D., Qi, P., Mele, V., 2025. Agentai: A comprehensive survey on autonomous agents in distributed ai for industry 4.0. Expert Systems with Applications , 128404.
- [101] Plaat, A., van Duijn, M., van Stein, N., Preuss, M., van der Putten, P., Batenburg, K.J., 2025. Agentic large language models, a survey. arXiv preprint arXiv:2503.23037 .
- [102] Qi, S., Chen, S., Li, Y., Kong, X., Wang, J., Yang, B., Wong, P., Zhong, Y., Zhang, X., Zhang, Z., et al., 2024. Civrealms: A learning and reasoning odyssey in civilization for decision-making agents. arXiv preprint arXiv:2401.10568 .
- [103] Qin, Y., Liang, S., Ye, Y., Zhu, K., Yan, L., Lu, Y., Lin, Y., Cong, X., Tang, X., Qian, B., et al., 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. arXiv preprint arXiv:2307.16789 .
- [104] Qu, C., Dai, S., Wei, X., Cai, H., Wang, S., Yin, D., Xu, J., Wen, J.R., 2024. Tool learning with large language models: A survey. URL: <https://arxiv.org/abs/2405.17935>, doi:https://doi.org/10.1007/s11704-024-40678-2, arXiv:2405.17935.
- [105] Qu, X., Damoah, A., Sherwood, J., Liu, P., Jin, C.S., Chen, L., Shen, M., Aleisa, N., Hou, Z., Zhang, C., et al., 2025. A comprehensive review of ai agents: Transforming possibilities in technology and beyond. arXiv preprint arXiv:2508.11957 .
- [106] Qureshi, R., Sapkota, R., Shah, A., Muneer, A., Zafar, A., Vayani, A., Shoman, M., Eldaly, A., Zhang, K., Sadak, F., et al., 2025. Thinking beyond tokens: From brain-inspired intelligence to cognitive foundations for artificial general intelligence and its societal impact. arXiv preprint arXiv:2507.00951 .
- [107] Rashid, M.S., Bock, C., Zhuang, Y., Buchholz, A., Esler, T., Valentin, S., Franceschi, L., Wistuba, M., Sivaprasad, P.T., Kim, W.J., et al.,

2025. Swe-polybench: A multi-language benchmark for repository level evaluation of coding agents. arXiv preprint arXiv:2504.08703 .
- [108] Raza, S., Bamgbose, O., Ghuge, S., Tavakoli, F., Reji, D.J., Bashir, S.R., 2025a. Developing safe and responsible large language model: can we balance bias reduction and language understanding? *Machine Learning* 114, 140.
- [109] Raza, S., Khan, T., Chatrath, V., Paulen-Patterson, D., Rahman, M., Bamgbose, O., 2024a. Fakewatch: a framework for detecting fake news to ensure credible elections. *Social Network Analysis and Mining* 14, 142.
- [110] Raza, S., Narayanan, A., Khazaie, V.R., Vayani, A., Chettiar, M.S., Singh, A., Shah, M., Pandya, D., 2025b. Humanibench: A human-centric framework for large multimodal models evaluation. arXiv preprint arXiv:2505.11454 .
- [111] Raza, S., Qureshi, R., Lotif, M., Chadha, A., Pandya, D., Emmanouilidis, C., 2025c. Just as humans need vaccines, so do models: Model immunization to combat falsehoods. arXiv preprint arXiv:2505.17870 .
- [112] Raza, S., Qureshi, R., Zahid, A., Kamawal, S., Sadak, F., Fioresi, J., Saeed, M., Sapkota, R., Jain, A., Zafar, A., et al., 2025d. Who is responsible? the data, models, users or regulations? a comprehensive survey on responsible generative ai for a sustainable future. arXiv preprint arXiv:2502.08650 .
- [113] Raza, S., Reji, D.J., Ding, C., 2024b. Dbias: detecting biases and ensuring fairness in news articles. *International Journal of Data Science and Analytics* 17, 39–59.
- [114] Raza, S., Saleh, C., Farooq, A., Hasan, E., Ogidi, F., Powers, M., Chatrath, V., Lotif, M., Sekhon, K., Javadi, R., Zahid, H., Zahid, A., Khazaie, V.R., Yu, Z., 2025e. Vilbias: Detecting and reasoning about bias in multimodal content. URL: <https://arxiv.org/abs/2412.17052>, arXiv:2412.17052.
- [115] Raza, S., Sapkota, R., Karkee, M., Emmanouilidis, C., 2025f. Trism for agentic ai: A review of trust, risk, and security management in llm-based agentic multi-agent systems. arXiv preprint arXiv:2506.04133 .
- [116] Raza, S., Shaban-Nejad, A., Dolatabadi, E., Mamiya, H., 2024c. Exploring bias and prediction metrics to characterise the fairness of machine learning for equity-centered public health decision-making: A narrative review. *IEEE Access* .
- [117] Rein, D., Becker, J., Deng, A., Nix, S., Canal, C., O'Connell, D., Arnott, P., Bloom, R., Broadley, T., Garcia, K., et al., 2025. Hcast: Human-calibrated autonomy software tasks. arXiv preprint arXiv:2503.17354 .
- [118] Ritter, F.E., Tehranchi, F., Oury, J.D., 2019. Act-r: A cognitive architecture for modeling cognition. doi:10.1002/wcs.1488.
- [119] Rose, S., Borchert, O., Mitchell, S., Connelly, S., 2020. Zero Trust Architecture. Technical Report Special Publication 800-207. National Institute of Standards and Technology (NIST). URL: <https://nvlpubs.nist.gov/nistpubs/specialpublications/NIST.SP.800-207.pdf>.
- [120] Ruan, Y., Dong, H., Wang, A., Pitis, S., Zhou, Y., Ba, J., Dubois, Y., Maddison, C.J., Hashimoto, T., 2023. Identifying the risks of lm agents with an lm-emulated sandbox. arXiv preprint arXiv:2309.15817 .
- [121] Russell, S.J., Norvig, P., 2021. *Artificial Intelligence: A Modern Approach*. 4th ed., Pearson.
- [122] Schick, T., Dwivedi-Yu, J., Dessi, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., Scialom, T., 2023. Toolformer: Language models can teach themselves to use tools. URL: <https://arxiv.org/abs/2302.04761>, arXiv:2302.04761.
- [123] Shen, Y., Song, K., Tan, X., Zhang, W., Ren, K., Yuan, S., Lu, W., Li, D., Zhuang, Y., 2024. Taskbench: Benchmarking large language models for task automation. *Advances in Neural Information Processing Systems* 37, 4540–4574.
- [124] Sherman, E., Shattuck, S., Singh, N., Eisenberg, I., 2025. From assistant to agent: Navigating the governance challenges of increasingly autonomous ai. *Credo AI Whitepaper*. URL: <https://www.credo.ai/recourseslongform/from-assistant-to-agent-navigating-the-governance-challenges-of-increasingly-autonomous-ai>.
- [125] Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., Yao, S., 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems* 36, 8634–8652.
- [126] Shridhar, M., Yuan, X., Côté, M.A., Bisk, Y., Trischler, A., Hausknecht, M., 2020. Alfworld: Aligning text and embodied environments for interactive learning. arXiv preprint arXiv:2010.03768 .
- [127] Shukla, M., 2025. Adaptive monitoring and real-world evaluation of agentic ai systems. arXiv preprint arXiv:2509.00115 .
- [128] Siegel, Z.S., Kapoor, S., Nagdir, N., Stroebel, B., Narayanan, A., 2024. Core-bench: Fostering the credibility of published research through a computational reproducibility agent benchmark. arXiv preprint arXiv:2409.11363 .
- [129] SLSA Community, 2023. Slsa specification v1.0. <https://slsa.dev/spec/v1.0/>. Accessed 2025-10-18.
- [130] Sreenath, R.M., Singh, M.P., 2004. Agent-based service selection. *Journal of Web Semantics* 1, 261–279.
- [131] Starace, G., Jaffe, O., Sherburn, D., Aung, J., Chan, J.S., Maksin, L., Dias, R., Mays, E., Kinsella, B., Thompson, W., et al., 2025. Paperbench: Evaluating ai's ability to replicate ai research. arXiv preprint arXiv:2504.01848 .
- [132] Stein, K., Fišer, D., Hoffmann, J., Koller, A., 2023. Autoplanbench: Automatically generating benchmarks for llm planners from pddl. arXiv preprint arXiv:2311.09830 .
- [133] Sun, H., Zhang, S., Niu, L., Ren, L., Xu, H., Fu, H., Zhao, F., Yuan, C., Wang, X., 2025. Collab-overcooked: Benchmarking and evaluating large language models as collaborative agents. arXiv preprint arXiv:2502.20073 .
- [134] Sun, R., Helie, S., 2013. Psychologically realistic cognitive agents: taking human cognition seriously. *Journal of Experimental & Theoretical Artificial Intelligence* 25, 65–92.
- [135] Sun, R., Hélie, S., 2013. Psychologically realistic cognitive agents: Taking human cognition seriously. *Journal of Experimental and Theoretical Artificial Intelligence* 25, 65–92. doi:10.1080/0952813X.2012.661236.
- [136] Team, D.A., 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. URL: <https://arxiv.org/abs/2501.12948>, arXiv:2501.12948.
- [137] Team, M., 2023. Mistral 7b. URL: <https://arxiv.org/abs/2310.06825>, arXiv:2310.06825.
- [138] Team, P., 2024. Phi-3 technical report: A highly capable language model locally on your phone. URL: <https://arxiv.org/abs/2404.14219>, arXiv:2404.14219.
- [139] Thurgood, S., Beyer, B., Ferguson, D., 2018. Example error budget policy. <https://sre.google/workbook/error-budget-policy/>. In *The Site Reliability Workbook*, Google SRE.
- [140] Tian, C., Qin, X., Tam, K., Li, L., Wang, Z., Zhao, Y., Zhang, M., Xu, C., 2025. Clone: Customizing llms for efficient latency-aware inference at the edge. URL: <https://arxiv.org/abs/2506.02847>, arXiv:2506.02847.
- [141] Tian, S., Zhang, Z., Chen, L., Liu, Z., 2024. Mmina: Benchmarking multihop multimodal internet agents. arXiv preprint arXiv:2404.09992 .
- [142] Torres-Arias, S., Afzali, H., Kuppusamy, T.K., Curtmola, R., 2019. in-toto: Providing farm-to-table guarantees for bits and bytes, in: 28th USENIX Security Symposium (USENIX Security 2019). URL: <https://www.usenix.org/system/files/sec19-torres-arias.pdf>.
- [143] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G., 2023. Llama: Open and efficient foundation language models. URL: <https://arxiv.org/abs/2302.13971>, arXiv:2302.13971.
- [144] Trivedi, H., Khot, T., Hartmann, M., Manku, R., Dong, V., Li, E., Gupta, S., Sabharwal, A., Balasubramanian, N., 2024. Appworld: A controllable world of apps and people for benchmarking interactive coding agents. arXiv preprint arXiv:2407.18901 .
- [145] Uchendu, I., Jabbour, J., Berghe, K.V.d., Runevic, J., Stewart, M., Ma, J., Krishnan, S., Gur, I., Huang, A., Bishop, C., et al., 2025.

- A2perf: Real-world autonomous agents benchmark. arXiv preprint arXiv:2503.03056 .
- [146] Uddin, A., Salam, H., 2025. Beyond rule-based context awareness: Large language models as adaptive cognitive layers in cyber-physical systems. URL: www.aaai.org.
- [147] Union, E., 2018. General data protection regulation. URL: <https://gdpr-info.eu/>. [Accessed 01-10-2024].
- [148] U.S. National Telecommunications and Information Administration (NTIA), 2021. The Minimum Elements for a Software Bill of Materials (SBOM). Technical Report. U.S. Department of Commerce. URL: <https://www.ntia.gov/report/2021/minimum-elements-software-bill-materials-sbom>.
- [149] Valmeekam, K., Marquez, M., Olmo, A., Sreedharan, S., Kambhampati, S., 2023. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Advances in Neural Information Processing Systems* 36, 38975–38987.
- [150] Van Wynsberghe, A., 2021. Sustainable ai: Ai for sustainability and the sustainability of ai. *AI and Ethics* 1, 213–218.
- [151] Verifiable Credentials Working Group, 2025. Verifiable Credentials Data Model 2.0. W3C Recommendation. World Wide Web Consortium (W3C). URL: <https://www.w3.org/TR/vc-data-model-2.0/>.
- [152] Vongthongsri, K., 2025. Llm agent evaluation: Assessing tool use, task completion, agentic reasoning, and more. URL: <https://www.confident-ai.com/blog/llm-agent-evaluation-complete-guide#evaluating-tool-use>.
- [153] Wang, F., Chen, B., Xu, K., Tang, B., Xiong, F., Li, Z., 2025a. Text2mem: A unified memory operation language for memory operating system. arXiv preprint arXiv:2509.11145 .
- [154] Wang, H., Zou, H., Song, H., Feng, J., Fang, J., Lu, J., Liu, L., Luo, Q., Liang, S., Huang, S., et al., 2025b. Ui-tars-2 technical report: Advancing gui agent with multi-turn reinforcement learning. arXiv preprint arXiv:2509.02544 .
- [155] Wang, J., Zhou, J., Wen, M., Mo, X., Zhang, H., Lin, Q., Jin, C., Wang, X., Zhang, W., Peng, Q., 2024a. Hammerbench: Fine-grained function-calling evaluation in real mobile device scenarios. arXiv preprint arXiv:2412.16516 .
- [156] Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., et al., 2024b. A survey on large language model based autonomous agents. *Frontiers of Computer Science* 18, 186345.
- [157] Wang, N., Yan, Z., Li, W., Ma, C., Chen, H., Xiang, T., 2025c. Advancing embodied agent security: From safety benchmarks to input moderation. arXiv preprint arXiv:2504.15699 .
- [158] Wang, W., Zhang, D., Feng, T., Wang, B., Tang, J., 2024c. Battleagentbench: A benchmark for evaluating cooperation and competition capabilities of language models in multi-agent systems. arXiv preprint arXiv:2408.15971 .
- [159] Wang, X., Wang, Z., Liu, J., Chen, Y., Yuan, L., Peng, H., Ji, H., 2023. Mint: Evaluating llms in multi-turn interaction with tools and language feedback. arXiv preprint arXiv:2309.10691 .
- [160] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D., 2022. Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171 .
- [161] Wang, Y., Zhao, S., Wang, Z., Fan, M., Zhang, Y., Zhang, X., Wang, Z., Huang, H., Liu, T., 2025d. Rag+: Enhancing retrieval-augmented generation with application-aware reasoning. arXiv preprint arXiv:2506.11555 .
- [162] Wijk, H., Lin, T., Becker, J., Jawhar, S., Parikh, N., Broadley, T., Chan, L., Chen, M., Clymer, J., Dhyani, J., et al., 2024. Re-bench: Evaluating frontier ai r&d capabilities of language model agents against human experts. arXiv preprint arXiv:2411.15114 .
- [163] Wooldridge, M., 2009. *An Introduction to MultiAgent Systems*. 2nd ed., John Wiley & Sons.
- [164] Wu, C.K., Tam, Z.R., Lin, C.Y., Chen, Y.N.V., Lee, H.y., 2024a. Streambench: Towards benchmarking continuous improvement of language agents. *Advances in Neural Information Processing Systems* 37, 107039–107063.
- [165] Wu, S., Oltramari, A., Francis, J., Giles, C.L., Ritter, F.E., 2025. Cognitive llms: Toward human-like artificial intelligence by integrating cognitive architectures and large language models for manufacturing decision-making. *Neurosymbolic Artificial Intelligence 1*. URL: <https://journals.sagepub.com/doi/10.1177/29498732251377341>, doi:10.1177/29498732251377341.
- [166] Wu, X., Shen, Y., Shan, C., Song, K., Wang, S., Zhang, B., Feng, J., Cheng, H., Chen, W., Xiong, Y., et al., 2024b. Can graph learning improve planning in llm-based agents? *Advances in Neural Information Processing Systems* 37, 5338–5383.
- [167] Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., et al., 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences* 68, 121101.
- [168] Xia, M., Ruehle, V., Rajmohan, S., Shokri, R., 2025. Minerva: A programmable memory test benchmark for language models. arXiv preprint arXiv:2502.03358 .
- [169] Xiao, R., Ma, W., Wang, K., Wu, Y., Zhao, J., Wang, H., Huang, F., Li, Y., 2024. Flowbench: Revisiting and benchmarking workflow-guided planning for llm-based agents. arXiv preprint arXiv:2406.14884 .
- [170] Xie, T., Zhang, D., Chen, J., Li, X., Zhao, S., Cao, R., Hua, T.J., Cheng, Z., Shin, D., Lei, F., et al., 2024. Oworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems* 37, 52040–52094.
- [171] Xing, M., Zhang, R., Xue, H., Chen, Q., Yang, F., Xiao, Z., 2024. Understanding the weakness of large language model agents within a complex android environment, in: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6061–6072.
- [172] Xu, F.F., Song, Y., Li, B., Tang, Y., Jain, K., Bao, M., Wang, Z.Z., Zhou, X., Guo, Z., Cao, M., et al., 2024. Theagentcompany: benchmarking llm agents on consequential real world tasks. arXiv preprint arXiv:2412.14161 .
- [173] Xu, Q., Hong, F., Li, B., Hu, C., Chen, Z., Zhang, J., 2023. On the tool manipulation capability of open-source large language models. arXiv preprint arXiv:2305.16504 .
- [174] Xu, W., Mei, K., Gao, H., Tan, J., Liang, Z., Zhang, Y., 2025. A-mem: Agentic memory for llm agents. arXiv preprint arXiv:2502.12110 .
- [175] Yang, H., Yue, S., He, Y., 2023. Auto-gpt for online decision making: Benchmarks and additional opinions. arXiv preprint arXiv:2306.02224 .
- [176] Yang, R., Chen, H., Zhang, J., Zhao, M., Qian, C., Wang, K., Wang, Q., Koripella, T.V., Movahedi, M., Li, M., et al., 2025. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. arXiv preprint arXiv:2502.09560 .
- [177] Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W.W., Salakhutdinov, R., Manning, C.D., 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. arXiv preprint arXiv:1809.09600 .
- [178] Yao, S., Chen, H., Yang, J., Narasimhan, K., 2022. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems* 35, 20744–20757.
- [179] Yao, S., Shinn, N., Razavi, P., Narasimhan, K., 2024. tau-bench: A benchmark for tool-agent-user interaction in real-world domains. arXiv preprint arXiv:2406.12045 .
- [180] Yao, S., Zhao, J., Yu, D., Du, N., Shafraan, I., Narasimhan, K., Cao, Y., 2023. React: Synergizing reasoning and acting in language models, in: *International Conference on Learning Representations (ICLR)*.
- [181] Ye, D., Zhou, F., Lv, J., Ma, J., Zhang, J., Lv, J., Li, J., Deng, M., Yang, M., Fu, Q., et al., 2025. Yan: Foundational interactive video generation. arXiv preprint arXiv:2508.08601 .
- [182] Yehudai, A., Eden, L., Li, A., Uziel, G., Zhao, Y., Bar-Haim, R., Cohan, A., Shmueli-Scheuer, M., 2025. Survey on evaluation of llm-based agents. arXiv preprint arXiv:2503.16416 .
- [183] Yin, S., Pang, X., Ding, Y., Chen, M., Bi, Y., Xiong, Y., Huang, W., Xiang, Z., Shao, J., Chen, S., 2024. Safeagentbench: A benchmark

- for safe task planning of embodied llm agents. arXiv preprint arXiv:2412.13178 .
- [184] Yoran, O., Amouyal, S.J., Malaviya, C., Bogin, B., Press, O., Berant, J., 2024. Assistantbench: Can web agents solve realistic and time-consuming tasks? arXiv preprint arXiv:2407.15711 .
 - [185] Yu, M., Meng, F., Zhou, X., Wang, S., Mao, J., Pan, L., Chen, T., Wang, K., Li, X., Zhang, Y., et al., 2025. A survey on trustworthy llm agents: Threats and countermeasures, in: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, pp. 6216–6226.
 - [186] Yuan, S., Song, K., Chen, J., Tan, X., Shen, Y., Kan, R., Li, D., Yang, D., 2024. Easytool: Enhancing llm-based agents with concise tool instruction. URL: <https://arxiv.org/abs/2401.06201>, arXiv:2401.06201.
 - [187] Zhang, T., Eysenbach, B., Salakhutdinov, R., Levine, S., Gonzalez, J.E., 2021. C-planning: An automatic curriculum for learning goal-reaching tasks. arXiv preprint arXiv:2110.12080 .
 - [188] Zhang, X., Zhang, C., Sun, J., Xiao, J., Yang, Y., Luo, Y., 2025a. Eduplanner: Llm-based multi-agent systems for customized and intelligent instructional design. IEEE Transactions on Learning Technologies .
 - [189] Zhang, Y., Yuan, S., Hu, C., Richardson, K., Xiao, Y., Chen, J., 2024. Timearena: Shaping efficient multitasking language agents in a time-aware simulation. arXiv preprint arXiv:2402.05733 .
 - [190] Zhang, Z., He, Y., Sun, Y., Shi, J., Liu, L., Nie, Q., 2025b. Roboact-clip: Video-driven pre-training of atomic action understanding for robotics. arXiv preprint arXiv:2504.02069 .
 - [191] Zheng, H.S., Mishra, S., Zhang, H., Chen, X., Chen, M., Nova, A., Hou, L., Cheng, H.T., Le, Q.V., Chi, E.H., et al., 2024. Natural plan: Benchmarking llms on natural language planning. arXiv preprint arXiv:2406.04520 .
 - [192] Zhou, S., Xu, F.F., Zhu, H., Zhou, X., Lo, R., Sridhar, A., Cheng, X., Ou, T., Bisk, Y., Fried, D., et al., 2023a. Webarena: A realistic web environment for building autonomous agents. arXiv preprint arXiv:2307.13854 .
 - [193] Zhou, X., Zhu, H., Mathur, L., Zhang, R., Yu, H., Qi, Z., Morency, L.P., Bisk, Y., Fried, D., Neubig, G., et al., 2023b. Sotopia: Interactive evaluation for social intelligence in language agents. arXiv preprint arXiv:2310.11667 .
 - [194] Zhou, Y., Jiang, S., Tian, Y., Weston, J., Levine, S., Sukhbaatar, S., Li, X., 2025. Sweet-rl: Training multi-turn llm agents on collaborative reasoning tasks. arXiv preprint arXiv:2503.15478 .
 - [195] Zhu, K., Du, H., Hong, Z., Yang, X., Guo, S., Wang, Z., Wang, Z., Qian, C., Tang, X., Ji, H., et al., 2025. Multiagentbench: Evaluating the collaboration and competition of llm agents. arXiv preprint arXiv:2503.01935 .
 - [196] Zhuge, M., Wang, W., Kirsch, L., Faccio, F., Khizbullin, D., Schmidhuber, J., 2024. Language agents as optimizable graphs. URL: <https://arxiv.org/abs/2402.16823>, arXiv:2402.16823.