Transparency for Agentic AI: A Study of Benchmarks, Metrics, and Oversight

Anonymous ACL submission

Abstract

Agentic AI is an emerging class of systems that perceive, plan, and act to achieve complex real-world goals (e.g., web navigation, gaming, and software engineering). However, the the transparency of these systems is fragile because behavior unfolds across multi-step trajectories and interacting sub-agents; a failure in a single component propagate and degrade the whole execution. This study reviews benchmarking, evaluation, and governance approaches for agentic AI through a transparency lens. Existing surveys largely focus on agent architectures and capabilities; in contrast, we organize the literature around transparency and show that outcome-only evaluation fails to support auditability. To address this gap, we outline trace-based signals and metrics that better capture agent behavior across the full execution trajectory. Overall, we aim to provide a practical resource for researchers and practitioners and to support the development of more transparent and auditable agentic AI systems.

1 Introduction

Recent advances in large language models (LLMs) have accelerated a shift from single-turn, static prediction systems toward autonomous, goal-directed agents that can reason, plan, and act in interactive environments (Plaat et al., 2025; Xi et al., 2025). Agentic AI systems refer to LLM-driven systems capable of (i) maintaining state across steps (e.g., memory or persistent context), (ii) decomposing objectives into subgoals, (iii) invoking external tools and APIs, and (iv) adapting their behavior based on observations and feedback (Bandi et al., 2025). Early systems such as AutoGPT (Yang et al., 2023) and BabyAGI (Nakajima, 2023) illustrated this paradigm, while more recent frameworks (Zhu et al., 2025a; Yoran et al., 2024) incorporate richer memory, multi-agent coordination, and human-in-the-loop (HITL) controls. However, as autonomy increases, it becomes harder to inspect how decisions are formed and to govern agent behavior, especially in high-stakes settings.

Transparency in AI has traditionally focused on explaining individual predictions using techniques such as attention visualization, saliency maps, or feature attribution (Hulsen, 2023). While valuable for static models, these approaches are insufficient for agentic AI systems, whose behavior involves planning, tool invocation, memory updates, and coordination across components. The years 2024 and 2025 have witnessed rapid progress in agentic AI systems across both research (Yang et al., 2025) and industry (Grand View Research, Inc., 2025). However, despite this growth, systematic approaches to transparency in agentic AI are overlooked, largely because the field is still emerging. Existing surveys have mainly focused on agentic architectures, planning mechanisms, and design patterns (as shown in **Table 1**), but transparency is often treated as a secondary concern rather than a first-class design and evaluation objective.

Current work on agentic AI evaluation has largely emphasized outcome-based metrics such as accuracy, success rate, and efficiency (Liang et al., 2023). While useful, trustworthy deployment requires evaluation beyond final outcomes, including trajectory traceability, verifiable tool use and provenance, and governance controls; explainability and interpretability can further support human oversight when available. This gap is increasingly consequential under emerging governance frameworks such as the EU AI Act (European Union, 2024), the NIST AI Risk Management Framework (NIST, 2023), and ISO/IEC 42001 (ISO, 2023), which emphasize transparency, accountability, and human oversight for higher-risk AI systems. Together, these developments motivate transparency mechanisms for agentic AI that are **lifecycle-aware** and spans design-time instrumentation, run-time monitoring, and post-hoc auditing. A working instance

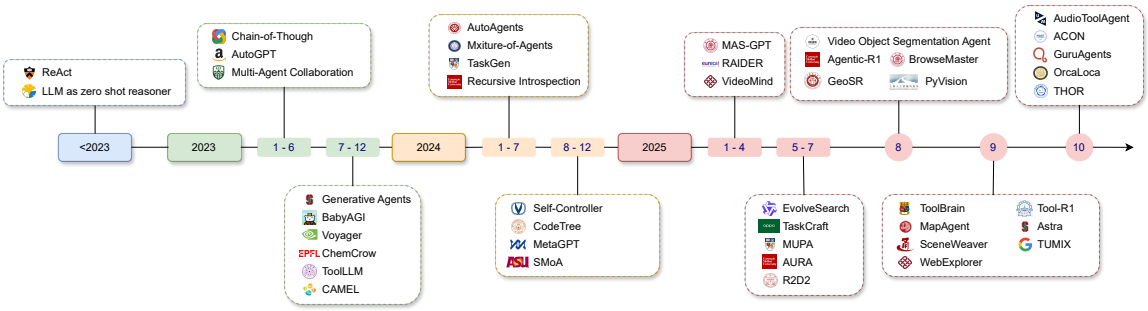

Figure 1: Timeline of Agentic AI papers which are released from 2023 to 2025. The frequency of agentic papers have increased in the year 2025, and every month the number of papers introducing new agents are increasing.

Legend: ✓ = Comprehensive coverage, ◐ = Partial/limited discussion, ✗ = Not addressed.

Survey	Primary Focus	Interpretability	Explainability	Auditability	Governance
(Wang et al., 2024a)	LLM-based agent architectures	◐	✗	✗	✗
(Yehudai et al., 2025)	Agent evaluation methods	◐	◐	✗	◐
(Plaat et al., 2025)	Agentic LLM capabilities	◐	◐	✗	◐
(Mohammadi et al., 2025)	Benchmarks and metrics	✗	◐	✗	✗
(Xi et al., 2025)	Agent taxonomy and applications	✗	✗	✗	◐
(Guo et al., 2024)	Multi-agent systems	◐	✗	✗	✗
This Survey	**Transparency in agentic AI across the lifecycle**	✓	✓	✓	✓

Table 1: Comparison of prior surveys on agentic AI with respect to transparency-related dimensions.

example is given in **Figure 2**.

We structure this review around three pillars: benchmarks, metrics, and governance, because these are the concrete artifacts used to evaluate and control agentic systems in practice. We review transparency mechanisms for agentic AI with an emphasis on traceability, auditability and provenance, and governance, and we provide a small, reproducible reference audit using a fixed evaluation harness.[1]

Our experiments are included to support the review article rather than provide a comprehensive leaderboard study. The goal is to apply the transparency lens and highlight common failure patterns in agent executions.

Our findings indicate that current evaluation mainly emphasizes outcome-level metrics (e.g., task success rate, final accuracy). In contrast, existing process-level metrics (e.g., number of tool calls, timeout rate), which aim to capture the end-to-end pipeline, are often not intuitive enough for diagnosing failure modes. This is especially true when errors emerge from tool-use interactions or resource exhaustion, which accuracy-only evaluation can miss. To address this gap, we propose three trace-based behavioral metrics, $\mathcal{R}imp$, $\mathcal{B}inf o$, and $\mathcal{C}_{ops}$ (Section 5.2), which audit agent behavior from execution traces alone, without requiring explicit reasoning logs. Together, these observations and metrics motivate transparency-aware evaluation protocols that can audit and assess the full agentic lifecycle.

[1]Code available at: Code

2 Background & Scope

Agentic AI, types, and autonomy We define *agentic AI* as LLM-based systems that act over multiple steps: they keep state (e.g., memory), plan and break down goals, call tools or APIs, and interact with an environment. Agentic systems can also be grouped by autonomy. *Retriever–generator agents* are mostly reactive and use retrieval to support generation (Lewis et al., 2020). *Tool-augmented agents* extend capabilities through external functions and APIs (Xu et al., 2023). *Planning agents* perform goal decomposition and multi-step decision making (Zhang et al., 2025). *Autonomous agents* aim for long-horizon behavior with persistent state and sometimes self-reflection (Li et al., 2024a). Multi-agent systems distribute tasks across specialized agents (Qu et al., 2025), but make transparency harder due to coordination and harder-to-audit communication. These systems are often harder to inspect as errors can compound over time, tools can fail or drift (Arike et al., 2025), and memory can introduce inconsistencies that blur the decision trail.

Scope and coverage This review article focuses on interactive, tool-using agentic systems where decisions unfold over multiple steps and where

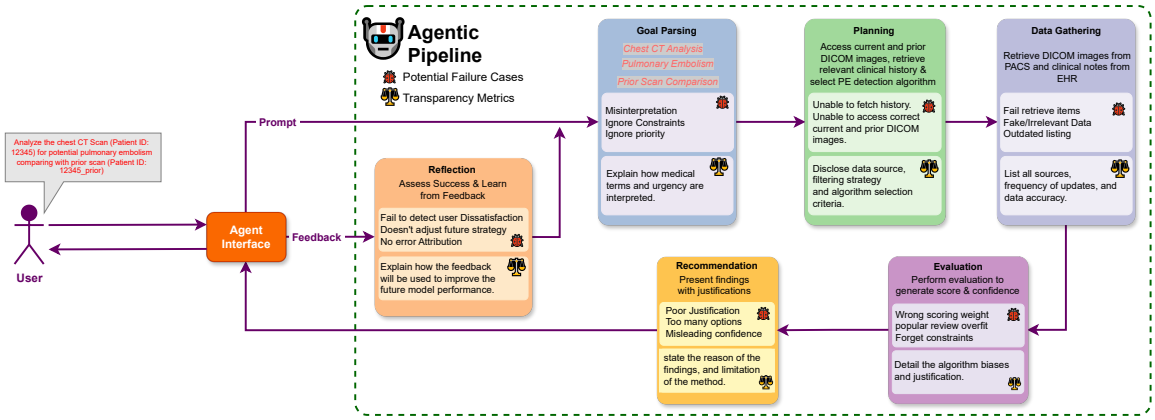

Figure 2: Working instance of healthcare assistant agent for radiologist to evaluate the evolution of pulmonary embolism in chest CT Scan. The red bugs show potential failure cases that can happen at each pipeline step, and the transparency metrics show the explainable objective of specialist sub-agent at each step.

Family	Representative Benchmarks	Verifiable Transparency Signals
Web / GUI	(Yao et al., 2022; Gou et al., 2025; Deng et al., 2023; Zhou et al., 2023a; Koh et al., 2024)	State checks; action logs; tool receipts; parseable execution traces
OS / App	(Xie et al., 2024; Trivedi et al., 2024; Kapoor et al., 2024)	Replayable state diffs; environment resets; recovery traces
Tool / API	(Qin et al., 2023; Li et al., 2023; Patil et al., 2024; Huang et al., 2023b)	Schema-valid calls; argument validation; tool-sequence correctness
Planning	(Stein et al., 2023; Shen et al., 2024; Zheng et al., 2024; Mialon et al., 2023; Wu et al., 2024a)	Plan–trace alignment; long-horizon error propagation
Multi-agent	(Zhu et al., 2025a; Wang et al., 2024b; Zhou et al., 2023b; Hyun et al., 2025)	Message logs; attribution; shared-memory consistency

Table 2: Benchmark families and the transparency signals they support (examples; see Appendix Table A.1 for the full list).

Metric	Operational Definition	Evidence
Task success (Zhou et al., 2023a)	Final task completion (binary or graded) per benchmark definition.	Benchmark state
Pass@k (Chan et al., 2024)	At least one successful outcome within k independent trials.	Trial outcomes
Efficiency (Huang et al., 2023a)	Steps, tokens, or time per successful run (optionally cost).	Runtime trace
Trace completeness (Deshpande et al., 2025)	Fraction of required fields present per step, averaged per run.	Schema checks
Trace consistency (Deshpande et al., 2025)	Valid ordering and linkage; no orphaned or missing steps.	Trace validator
Attribution coverage (Yao et al., 2022)	Proportion of steps with identifiable agent/module origin.	Trace metadata
Tool-call validity (Xu et al., 2023)	Tool calls conforming to declared schema and permissions.	Tool logs
Tool-call correctness (Xu et al., 2023)	Appropriateness of tool choice for the subgoal/context.	Verifier / rubric
Provenance grounding (NIST, 2023)	Actions or claims linked to retrievable evidence sources.	Trace + evidence
Policy adherence (Yao et al., 2024)	Actions compliant with declared policies; violation rate.	Policy logs
Intervention rate	Human or gate interventions per run (approve/deny).	Oversight trace
Auditability	Sufficiency of traces and versions to enable replay.	Trace + manifests
Recovery rate	Runs that recover from an error and still succeed.	Trace + outcome
Time-to-recovery	Steps or time from failure to resumed correct execution.	State diffs

Table 3: Trace-derived transparency metrics consolidated into a unified reporting schema (full list in Appendix Table A.2).

execution traces can be logged (e.g., plans, tool calls, observations/state changes, memory events, and human-oversight interventions). We primarily focus on the post-LLM agent era (2023–2025), while including a small number of earlier foundational works only when necessary. We conducted a keyword-based literature search (e.g., "Agentic AI", "autonomous agents", "LLM-based agents", "AI evaluation") and provide the full search details in Appendix A. This process yielded approximately 180 candidate articles. Each article was independently reviewed by the authors for relevance and quality. After further screening based on predefined inclusion criteria, 120 studies were retained for detailed analysis. **Figure 1** provides a timeline of agentic AI evolution.

In this study, we do not aim to cover single-turn chat evaluation, static text-only tasks without tool interaction, or transparency proposals that are not grounded in agent execution traces. Our goal is to connect transparency mechanisms to measurable signals across the agent lifecycle.We summarize current state-of-the-art agentic AI leaderboards and benchmarks in **Figure 3**.

How to read this review article We organize the review around benchmarks (**Table 2**) metrics (**Table 3**), and governance (**Table 4**), and we use a shared transparency lens (traceability, explainability when available, auditability/provenance, and governance/oversight) to evaluate what each artifact can verify across the agent lifecycle. We

Control Family	Trace-Generating Mechanisms	Auditable Artifacts & Signals
Identity & Access *(AuthZ, Secrets)*	Policy-as-code (OPA); Scoped capability tokens; Workload identity (OIDC); Zero-trust guards	Decision logs (allow/deny); Policy override rates; Token revocation lists; Secret rotation proofs
Runtime Safety *(Sandboxing, HITL)*	Tool sandboxing; Circuit breakers; Human-in-the-Loop (HITL) gates; Shadow policies	Egress/syscall logs; Sandbox escape alerts; HITL transcripts; Intervention rates; Kill-switch activation logs
Supply Chain *(Provenance, Data)*	Cryptographic signing (in-toto/SLSA); SBOM/ABoM; Data lineage tracking; Immutable logging	Provenance graphs; Tamper-evident hash chains; Dataset flow maps; Verified build manifests
Governance *(Risk, Eval)*	Risk error budgets (SLO); Red-teaming suites; Programmatic assertions; Post-deployment monitoring	SLO breach reports; Jailbreak success rates; Incident post-mortems; Compliance matrices

Legend: **AuthZ**: Authorization; **OPA**: Open Policy Agent; **OIDC**: OpenID Connect; **HITL**: Human-in-the-Loop; **SLSA**: Supply-chain Levels for Software Artifacts; **SBOM/ABoM**: Software/Agent Bill of Materials; **SLO**: Service Level Objective.

Table 4: Taxonomy of control families mapping active mechanisms to verifiable audit artifacts. This summary aggregates the regulatory controls (see Appendix Table A.3) into trace-based categories.

further organize evaluation signals into outcome, process-trace, and governance to cover the agent lifecycle and present a small evaluation study.

3 Transparency Mechanisms in Agentic AI: A Review

We review transparency mechanisms across five categories that inform our taxonomy (Section 4). Our analysis reveals three key findings: **(1)** existing Explanable AI (XAI) methods focus on single-turn predictions and are insufficient for multi-step agent behavior, **(2)** classical planning systems provided explicit traceability that modern neural agents have lost, and **(3)** current evaluation emphasizes outcomes over process, creating an auditability gap.

3.1 Traceability in Planning and Multi-Agent Systems

Traceability has evolved from static feature attribution to the dynamic analysis of decision sequences. Early XAI methods, such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017), successfully approximated local decision boundaries but failed to capture the causal chains required for planning. While recent mechanistic interpretability approaches, such as Sparse Autoencoders (SAEs) (Shu et al., 2025), reveal concepts within the latent space, they do not inherently explain the temporal logic of an agent's trajectory. To address this, Multi-Agent Reinforcement Learning (MARL) frameworks have begun utilizing LLMs to externalize thought processes. For instance, DEXTER-LLM (Zhu et al., 2025b) demonstrates that planning in natural language renders decision traces human-readable, serving as a "window" into agent internal states (Zhu et al., 2024). However, these traceability mechanisms remain fragmented; current literature lacks a unified audit trail that connects low-level mechanistic states to these high-level natural language plans.

3.2 Benchmarking & Evaluation of Agentic AI

The transition to agentic AI necessitates a shift from outcome-based metrics to process-oriented observability. This approach prioritizes "cognitive" and "reasoning" benchmarks over traditional accuracy scores (Du et al., 2024). The core challenge is assessing whether an agent's Chain-of-Thought (CoT) follows a valid logical path rather than simply achieving a final goal. Consequently, new metric frameworks like Mean Time To Recover (MTTR) (Or, 2025) have emerged to debug reasoning errors within execution chains. Despite these specific debugging metrics, the field currently lacks a comprehensive set of "transparency metrics" designed to systematically quantify the auditability of an agent's reasoning process across diverse tasks.

3.3 AI Governance and Oversight

Governance mechanisms are evolving from static safety alignment to active, runtime oversight of autonomous goal pursuit. As agents operate with increasing independence, risks such as deceptive alignment and misevolution require continuous monitoring. The MI9 Agent Intelligence Protocol (Wang et al., 2025a) exemplifies this by establishing a distinct governance layer for real-time monitoring. To enforce this, recent proposals suggest architectural interventions, such as private governance layers (Ball, 2025) or specialized "enforcement agents" (Tamang and Bora, 2025) tasked solely with monitoring the planning traces of other agents. While these frameworks propose *who* should monitor agents, they fail to define the standardized *protocols* and *auditability pillars* required to make such governance universally effective.

3.4 Synthesis and Positioning

Despite advancements in mechanistic interpretability and governance protocols, a critical gap remains. Current XAI techniques are predominantly static, designed for single-turn classification, and fail to capture the temporal, multi-step decision-making characteristic of agentic workflows. There is a lack of unified frameworks that bridge the granularity of circuit discovery with the high-level auditability, required for multi-agent systems. To address this, the field must shift toward transparency approach that prioritizes internal mechanisms over pure performance. This survey addresses the gaps identified above by providing a comprehensive analysis of the necessary frameworks and metrics. Ultimately, we position transparency and auditability as central pillars for the next generation of agentic AI.

4 Taxonomy of Transparency for Agents

Building on the scope in Section 2, we formalize *transparency in agentic AI* as the availability of process artifacts, including plans, tool invocations, observations, memory updates, and oversight decisions, that are recorded throughout execution (Becerra Sandoval and Jing, 2025). We refer to this record as an *execution trajectory*[2]. We organize transparency into four complementary dimensions used throughout the survey: (1) *Traceability* (what happened?), (2) *Explainability* (why did it do that, when such signals are available), (3) *Auditability and provenance* (can we verify it?), and (4) *Governance and oversight* (was it allowed and controlled?).

Lifecycle view These dimensions can be instantiated at each stage of an agent pipeline, from goal parsing to planning, tool use and data gathering, acting, evaluation, and feedback or reflection (Manuvinakurike et al., 2025). **Figure 4** summarizes where measurable signals arise across this lifecycle, and we use the four dimensions above to structure the remainder of the survey.

Traceability Mechanisms Traceability mechanisms enable an agent's behavior to be reconstructed end to end across multi step execution trajectories, connecting plans to tool calls and state updates. Because failures often arise at intermediate stages, effective instrumentation is required to identify where and why a run deviates or fails (Barrak, 2025), rather than merely whether it succeeds overall. Common mechanisms include structured execution traces that log observations, actions, tool calls, and intermediate outputs, module level telemetry for planners, reasoners, tools, and memory components, and stage wise reporting of error propagation and goal drift (Yehudai et al., 2025). Traceability also benefits from standardized testbeds that support trace comparison across runs and systems.

Explainability Mechanisms Explainability mechanisms make key decisions understandable to humans. In agentic settings, explanations should be tied to trace steps and focus on actionable questions (Turpin et al., 2023), such as why a tool was selected, why a subgoal was prioritized, and why evidence was accepted or rejected. Common mechanisms include modular explanations aligned with agent components, reflection or self critique summaries that justify plan revisions (Manuvinakurike et al., 2025), causal abstractions that isolate the internal states driving specific actions (Meng et al., 2022), and process oriented evaluation that scores reasoning quality using rubrics or judge models (Lightman et al., 2023) as used in prior work. These mechanisms aim to produce concise rationales that support debugging and oversight rather than long narratives.

Auditability and Provenance Mechanisms Auditability mechanisms provide verifiable evidence for post hoc review (Marks et al., 2025). Because agents can change external state, audit logs and provenance become central. Typical mechanisms include provenance logs that link actions and claims to retrieved sources, tool receipts that capture inputs, outputs, and side effects, and run manifests that record model, version, and configuration so executions can be replayed. Some work further proposes packaging these artifacts into standardized bundles (NTIA, 2021), such as an Agent Bill of Materials, to support organizational handoffs and compliance reporting.

Governance and Oversight Mechanisms Governance and oversight mechanisms control what an agent may do, under which conditions, and with what approval (Agent, 2024). Common patterns include action gating via a control plane that approves or denies tool calls, least privilege credentials and scoped permissions, policy as code checks, and human in the loop review for higher risk ac-

[2]An execution trajectory captures the sequence of an agent's actions, observations, and intermediate artifacts over time.

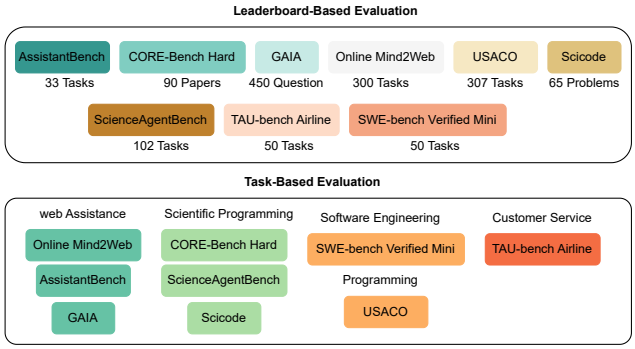

Figure 3: Major agentic benchmarks and their leaderboard-based and task-based summaries.

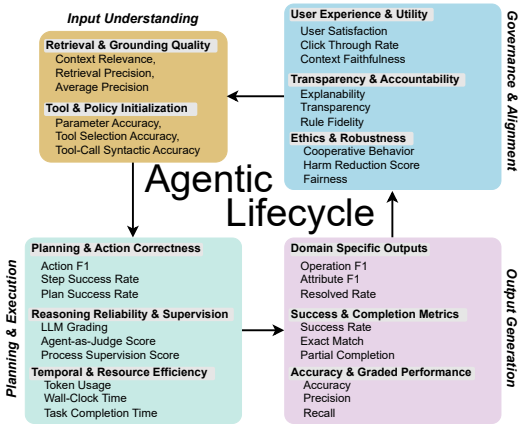

Figure 4: Evaluation metrics across stages of the agentic lifecycle.

tions. Oversight events, including approvals, denials, overrides, and violations, should be logged as part of the trace to enable auditing. Recent proposals (SLSA Community, 2023; VCWG, 2025) also operationalize governance via measurable targets, such as policy violation rate and unauthorized action rate, linking transparency artifacts to enforceable reliability and risk SLOs (Service Level Objectives).

5 Agentic AI Evaluation through a Transparency Lens

We map the transparency mechanisms (Section 4) to evaluation settings that provide *verifiable* signals beyond final answers. We group representative agentic AI benchmarks by the evidence they expose and propose a unified reporting schema spanning outcome, trace, and governance metrics. We then instantiate this schema via a lightweight, trace-based reference evaluation. Our experimental settings are given in Appendix B.

5.1 Evaluation Benchmarks

To illustrate trace-based transparency evaluation in a controlled setting, we use the **HAL Harness** framework (Kapoor et al., 2025) (details in Appendix B.1). We evaluate two benchmarks that expose complementary failure modes. For **GAIA** (Mialon et al., 2023), we run 100 validation tasks that require multi-step information gathering and file-based manipulation. For **AssistantBench** (Yoran et al., 2024), we run 33 tasks targeting web-navigation robustness, including persistency failures where agents enter repetitive, resource-consuming loops.

Metrics We report results under two evaluation settings: (1) *default outcome evaluation*, and (2) *trace-based transparency evaluation*. For (1), we use benchmark outcome metrics (success/failure) and efficiency (token usage). In this setting, we use **GPT-4o mini** for AssistantBench and **GPT-4.1 mini** for GAIA. For (2), we compute trace-derived transparency metrics from HAL execution logs using **GPT-4.1 mini** as the backbone model to control for model effects during trace analysis. We additionally use an LLM judge to evaluate the sub-task failure recovery rate, thoroughness, ability to pivot at dead-ends, and appropriate context switching rate. We use **Gemini-3-pro-preview** as the judge model. We also reference publicly reported HAL leaderboard baselines (e.g., Claude Sonnet 4.5 and GPT-5 Medium) for context, without re-running all baselines. Appendix **Table B.2** consolidates all evaluation metrics used in this study. Metrics marked with † represent our novel contributions for auditing black-box agents, defined below.

5.2 Trace-Based Behavioral Metrics

Since many high-performance agents (like GPT-4.1 mini) often optimize for speed by skipping explicit reasoning steps (producing empty "Thought" logs), we propose three behavioral metrics to audit their performance based solely on their execution traces.

Implicit Error Resilience ($\mathcal{R}_{imp}$) While agents may not explicitly verbalize errors in a "Thought" stream, an auditable system must demonstrate resilience. We define *Implicit Error Resilience* as the probability that an agent, upon encountering an environment error (e.g., `Command not found`, `403 Forbidden`), immediately attempts a *corrective action* (a different tool or modified argument) in the

subsequent step, rather than repeating the failed action or halting. Equation 1 is the mathematical formulation of this metric.

$$\mathcal{R}_{imp} = \frac{\sum_{t \in E} \mathbb{I}(\text{a}_{t+1} \neq \text{a}_t \vee \text{args}_{t+1} \neq \text{args}_t)}{|E|} \tag{1}$$

where E is the set of steps returning an error code or failure message. A score of 1.0 indicates perfect autonomous recovery behavior.

Information Gathering Breadth ($\mathcal{B}_{info}$) To audit the thoroughness of an agent's research without relying on citations, we measure *Information Gathering Breadth*. This metric quantifies the density of unique external information nodes (e.g., unique URLs visited, unique files read) accessed per successful trajectory. High $\mathcal{B}_{info}$ values indicate an agent that cross-references multiple sources rather than hallucinating or relying on a single data point.

$$\mathcal{B}_{info} = \frac{1}{N} \sum_{i=1}^{N} |\{u \mid u \in \text{UniqueSources}(\tau_i)\}| \tag{2}$$

where $\texttt{UniqueSources}(\tau_i)$ represents the set of distinct domains or file paths accessed during trajectory τ_i.

Operational Complexity Score ($\mathcal{C}_{ops}$) A sophisticated agentic system should seamlessly weave together different modes of operation (e.g., web browsing, code execution, file manipulation). The *Operational Complexity Score* measures the rate of "Context Switching" between different tool categories within a single trace. A higher score implies the agent is capable of multi-modal problem solving.

$$\mathcal{C}_{ops} = \frac{1}{T-1} \sum_{t=1}^{T-1} \mathbb{I}(\text{ToolType}_t \neq \text{ToolType}_{t+1}) \tag{3}$$

where T is the total number of steps in the trajectory and `ToolType` represents the category of the tool used (e.g., `bash` vs. `browser`).

5.3 Results and Discussion

We evaluate our trace-based transparency framework on two benchmarks to demonstrate that outcome metrics alone miss critical failure patterns.

Success Rates Obscure Divergent Failure Behaviors **Table 5** contrasts GPT-4o mini on AssistantBench with GPT-4.1 mini on GAIA. Despite similar framing as agentic tasks, failures exhibit opposite behavioral signatures. AssistantBench reveals a "Persistency Trap" where agents enter recursive loops on intractable states such as CAPTCHAs, consuming 3× more tokens on failures (∼112k) than successes (∼38k). In contrast, GAIA exhibits "Fast Refusal" with immediate failures from multimodal context errors and negligible token waste. This dichotomy is invisible to accuracy-only evaluation and is unearthed by the trace-based transparency evaluation.

Metric	AssistantBench	GAIA
Model	GPT-4o mini	GPT-4.1 mini
Success Rate	39.4%	9.0%
Avg. Tokens (Fail)	∼112k	Negligible
Primary Failure	Navigation Loops	Context Blindness

Table 5: Outcome metrics obscure distinct failure phenotypes across benchmarks.

Benchmark	Model	Acc. (%)
GAIA	Claude Sonnet 4.5	74.55
	GPT-5 Medium	62.80
AssistantBench	o3-medium	38.81
	GPT-5 Medium	35.23
CORE-Bench Hard	Claude Opus 4.1	51.11
	Gemini 3 Pro Preview	40.00

Table 6: HAL Harness leaderboard results (accuracy).

Accuracy-Only Leaderboards Miss Process-Level Signals **Table 6** reports HAL Harness leaderboard scores across four benchmarks. These accuracy-only rankings cannot distinguish between resource-efficient failures and costly loops, nor do they capture process-level transparency. This limitation motivates our trace-based metrics.

Trace-Based Metrics Reveal Latent Agent Capabilities **Table 7** reports our proposed metrics (Section 5.2) for GPT-4.1 mini on GAIA. The agent achieved high error resilience ($\mathcal{R}_{imp} = 0.92$), demonstrating recovery from failures without explicit reasoning traces. For instance, when encountering a `command not found` error, the agent autonomously executed an install command in the subsequent step. The information gathering breadth ($\mathcal{B}_{info} = 5.33$) confirms that success is grounded in multiple sources rather than parametric memorization. Moderate tool diversity ($\mathcal{C}_{ops} = 0.42$) indicates agents coordinate across tool types for complex tasks.

Summary Overall, we observe that trace-based metrics reveal behavioral properties, such as re-

Metric	Score	Interpretation
$\mathcal{R}_{imp}$	0.92	High error recovery
$\mathcal{B}_{info}$	5.33	Broad source coverage
$\mathcal{C}_{ops}$	0.42	Multi-tool coordination

Table 7: Trace-based behavioral metrics for GPT-4.1 mini on GAIA.

silience, thoroughness, and operational complexity, that outcome metrics miss. Because our audit uses different backbones across some benchmarks, we avoid causal claims about benchmark difficulty; our goal is to demonstrate how trace-level transparency signals surface distinct failure patterns under common agent deployments. We recommend future benchmarks adopt multi-dimensional reporting that couples accuracy with process-level transparency.

6 Discussion: Open Challenges and Future Directions

Summary of findings Our survey and empirical analysis reveal three key observations. First, outcome-level metrics (accuracy, success rate) consistently fail to capture process-level failures such as resource exhaustion loops and context blindness. Second, existing benchmarks expose varying transparency signals, but few provide unified reporting schemas that span outcome, trace, and governance dimensions. Third, regulatory frameworks increasingly demand transparency properties that current evaluation practices do not measure.

Gaps in current evaluation A critical limitation in current agentic AI evaluation lies in the weak linkage between performance metrics and governance requirements (Mohammadi et al., 2025). While recent benchmarks treat agents as decision-making systems, legacy metrics fail to capture end-to-end behavior across perception, planning, and control (Kokel et al., 2025). As governance frameworks increasingly demand transparency and risk-aware operation (Tian et al., 2025), future evaluations must shift from single outcome scores to multidimensional scorecards that couple task success with policy adherence and auditability.

Future research directions We identify two promising directions:

Trajectory-first evaluation. Evaluation must move beyond terminal outputs to inspect the entire reasoning process, thoughts, actions, and observations, enabling "process-trace audits." This includes assessing agents under distributional shift, reporting out-of-distribution gaps, and mitigating test contamination (Raza et al., 2025a,b).

State-grounded scoring Tasks should be containerized with versioned manifests and step-level correctness reporting (Van Wynsberghe, 2021). Reporting unit economics (tokens, time, cost) is essential to reveal trade-offs on the efficiency frontier (Khan et al., 2025).

Recommendations for practitioners Based on our analysis, we offer three recommendations:

1. **Report efficiency alongside accuracy**: Token usage, failure-mode analysis, and resource consumption should be standard in evaluation reports.

2. **Adopt trace-based auditing**: Metrics like $\mathcal{R}_{imp}$, $\mathcal{B}_{info}$, and $\mathcal{C}_{ops}$ complement outcome metrics for reliability assessment.

3. **Align with governance frameworks**: Link technical metrics to regulatory requirements (EU AI Act, NIST AI RMF, ISO/IEC 42001) through evaluation cards that uniformly report outcome, process, efficiency, and governance metrics.

7 Conclusion

This survey highlight the need for trace-based metrics for transparency and auditability for agentic AI. Our empirical analysis reveals that accuracy alone masks distinct failure patterns, such as persistency traps and context blindness, necessitating a "trajectory-first" approach that inspects the entire execution lifecycle. We addressed this gap by proposing a unified transparency taxonomy and three novel trace-based metrics, (1) Implicit Error Resilience ($\mathcal{R}_{imp}$), (2) Information Gathering Breadth ($\mathcal{B}_{info}$), and (3) Operational Complexity Score ($\mathcal{C}_{ops}$). These metrics enable the auditing of agent behaviors directly from execution logs. As governance frameworks like the EU AI Act and NIST AI RMF evolve, the field must adopt these multidimensional evaluation strategies to ensure agents are not only performant but also verifiable and compliant.

Limitations

This survey has some limitations. First, we focus on the post-LLM agent era (2023–2025) and may miss relevant earlier work on multi-agent systems and cognitive architectures. We do not aim to cover single-turn chat evaluation, static text-only tasks, or transparency proposals not grounded in execution traces. Second, our experiments cover only two benchmarks (GAIA, AssistantBench) with single-agent configurations; multi-agent coordination and communication failures remain untested. Due to computational cost constraints, we use mid-tier models (GPT-4o mini, GPT-4.1 mini) rather than frontier models, which may exhibit different failure patterns and transparency properties.

Third, we use Gemini-3-pro-preview as the LLM judge for trace-based evaluation, which may introduce model-specific biases in scoring. LLM-as-judge approaches are known to exhibit self-preference and positional biases that could affect our behavioral metric assessments (Raza et al., 2025c). We also do not evaluate safety, fairness, or ethical alignment dimensions, which are increasingly important for responsible deployment. Our proposed metrics ($\mathcal{R}_{imp}$, $\mathcal{B}_{info}$, $\mathcal{C}_{ops}$) focus on operational transparency rather than value alignment. Finally, the rapidly evolving nature of agentic AI means some recent developments may not be captured, and benchmark leaderboards may shift significantly after publication.

Ethics Statement

This work primarily involves a comprehensive survey and a limited computational evaluation of existing publicly available benchmarks. As such, this study did not involve human subjects, and no personally identifiable information (PII) was collected or processed, exempting it from Institutional Review Board (IRB) oversight. We acknowledge that the execution of agentic workflows and the utilization of LLM-as-a-judge frameworks can bring more computational costs and energy consumption (Powers et al., 2025). To address this, our proposed evaluation framework explicitly incorporates efficiency metrics (e.g., token usage and resource consumption) alongside accuracy to encourage the development of more sustainable agentic systems. Furthermore, while we aim to enhance auditability, we recognize that our metrics focus on operational transparency and do not fully substitute for safety alignment regarding fairness or bias; we caution practitioners to use these transparency signals as part of a broader, holistic governance strategy.

References

2023. Information technology – artificial intelligence – management system. 51 pp.

Saaket Agashe, Yue Fan, Anthony Reyna, and Xin Eric Wang. 2023. Llm-coordination: evaluating and analyzing multi-agent coordination abilities in large language models. *arXiv preprint arXiv:2310.03903*.

Open Policy Agent. 2024. Open policy agent: Policy as code. https://openpolicyagent.org/docs/. Accessed 2025-10-18.

Rauno Arike, Elizabeth Donoway, Henning Bartsch, and Marius Hobbhahn. 2025. Technical report: Evaluating goal drift in language model agents. *arXiv preprint arXiv:2505.02709*.

Dean W Ball. 2025. A framework for the private governance of frontier artificial intelligence. *arXiv preprint arXiv:2504.11501*.

Ajay Bandi, Bhavani Kongari, Roshini Naguru, Sahitya Pasnoor, and Sri Vidya Vilipala. 2025. The rise of agentic ai: A review of definitions, frameworks, architectures, applications, evaluation metrics, and challenges. *Future Internet*, 17(9):404.

Amine Barrak. 2025. Traceability and accountability in role-specialized multi-agent llm pipelines. *arXiv preprint arXiv:2510.07614*.

Juana Catalina Becerra Sandoval and Felicia S Jing. 2025. Historical methods for ai evaluations, assessments, and audits. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 1371–1386.

Tara Bogavelli, Roshnee Sharma, and Hari Subramani. 2025. Agentarch: A comprehensive benchmark to evaluate agent architectures in enterprise. *arXiv preprint arXiv:2509.10769*.

Ben Bogin, Kejuan Yang, Shashank Gupta, Kyle Richardson, Erin Bransom, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Super: Evaluating agents on setting up and executing tasks from research repositories. *arXiv preprint arXiv:2409.07440*.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, and 1 others. 2024. Mle-bench: Evaluating machine learning agents on machine learning engineering. *arXiv preprint arXiv:2410.07095*.

Ma Chang, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. 2024. Agentboard: An analytical evaluation board of multi-turn llm agents.

Advances in neural information processing systems, 37:74325–74362.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Xing Gao, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, Fei Huang, and 1 others. 2024a. Socialbench: Sociality evaluation of role-playing conversational agents. *arXiv preprint arXiv:2403.13679*.

Junzhe Chen, Xuming Hu, Shuodi Liu, Shiyu Huang, Wei-Wei Tu, Zhaofeng He, and Lijie Wen. 2024b. Llmarena: Assessing capabilities of large language models in dynamic multi-agent environments. *arXiv preprint arXiv:2402.16499*.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, and 1 others. 2024c. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. *arXiv preprint arXiv:2410.05080*.

De Chezelles, Thibault Le Sellier, Sahar Omidi Shayegan, Lawrence Keunho Jang, Xing Han Lù, Ori Yoran, Dehan Kong, Frank F Xu, Siva Reddy, Quentin Cappart, and 1 others. 2024. The browsergym ecosystem for web agent research. *arXiv preprint arXiv:2412.05467*.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36:28091–28114.

Darshan Deshpande, Varun Gangal, Hersh Mehta, Jitin Krishnan, Anand Kannappan, and Rebecca Qian. 2025. Trail: Trace reasoning and agentic issue localization. *arXiv preprint arXiv:2505.08638*.

Yubo Dong, Xukun Zhu, Zhengzhe Pan, Linchao Zhu, and Yi Yang. 2024. Villageragent: A graph-based multi-agent framework for coordinating complex task dependencies in minecraft. *arXiv preprint arXiv:2406.05720*.

Hung Du, Srikanth Thudumu, Rajesh Vasa, and Kon Mouzakis. 2024. A survey on context-aware multi-agent systems: techniques, challenges and future directions. *arXiv preprint arXiv:2402.01968*.

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.

Longling Geng and Edward Y Chang. 2025. Realm-bench: A real-world planning benchmark for llms and multi-agent systems. *arXiv preprint arXiv:2502.18836*.

Ran Gong, Qiuyuan Huang, Xiaojian Ma, Hoi Vo, Zane Durante, Yusuke Noda, Zilong Zheng, Song-Chun Zhu, Demetri Terzopoulos, Li Fei-Fei, and 1 others. 2023. Mindagent: Emergent gaming interaction. *arXiv preprint arXiv:2309.09971*.

Boyu Gou, Zanming Huang, Yuting Ning, Yu Gu, Michael Lin, Weijian Qi, Andrei Kopanev, Botao Yu, Bernal Jiménez Gutiérrez, Yiheng Shu, and 1 others. 2025. Mind2web 2: Evaluating agentic search with agent-as-a-judge. *arXiv preprint arXiv:2506.21506*.

Grand View Research, Inc. 2025. Ai agents market size and share, industry report, 2033. https://www.grandviewresearch.com/industry-analysis/ai-agents-market-report. Report ID: GVR-4-68040-471-3; Market forecast and trends for AI agents from 2026 to 2033; accessed Jan 2026.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Benhao Huang, Yingzhuo Yu, Jin Huang, Xingjian Zhang, and Jiaqi W Ma. 2025. Dca-bench: A benchmark for dataset curation agents. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5482–5492.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. 2023a. Mlagentbench: Evaluating language agents on machine learning experimentation. *arXiv preprint arXiv:2310.03302*.

Yue Huang, Jiawen Shi, Yuan Li, Chenrui Fan, Siyuan Wu, Qihui Zhang, Yixin Liu, Pan Zhou, Yao Wan, Neil Zhenqiang Gong, and 1 others. 2023b. Metatool benchmark for large language models: Deciding whether to use tools and which to use. *arXiv preprint arXiv:2310.03128*.

Tim Hulsen. 2023. Explainable artificial intelligence (xai): concepts and challenges in healthcare. *Ai*, 4(3):652–666.

Jonathan Hyun, Nicholas R Waytowich, and Boyuan Chen. 2025. Crew-wildfire: Benchmarking agentic multi-agent collaborations at scale. *arXiv preprint arXiv:2507.05178*.

IBM. Agentic AI evaluation. https://www.ibm.com/docs/en/watsonx/saas?topic=sdk-agentic-ai-evaluation. Accessed: 5 October 2025.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2023. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*.

Raghav Kapoor, Yash Parag Butala, Melisa Russak, Jing Yu Koh, Kiran Kamble, Waseem AlShikh, and Ruslan Salakhutdinov. 2024. Omniact: A dataset and benchmark for enabling multimodal generalist autonomous agents for desktop and web. In *European Conference on Computer Vision*, pages 161–178. Springer.

Sayash Kapoor, Benedikt Stroebl, Peter Kirgis, Nitya Nadgir, Zachary S Siegel, Boyi Wei, Tianci Xue, Ziru Chen, Felix Chen, Saiteja Utpala, Franck Ndzomga, Dheeraj Oruganty, Sophie Luskin, Kangheng Liu, Botao Yu, Amit Arora, Dongyoon Hahm, Harsh Trivedi, Huan Sun, and 12 others. 2025. Holistic agent leaderboard: The missing infrastructure for ai agent evaluation. https://github.com/princeton-pli/hal-harness.

Tahniat Khan, Soroor Motie, Sedef Akinli Kocak, and Shaina Raza. 2025. Optimizing large language models: Metrics, energy efficiency, and case study insights. *arXiv preprint arXiv:2504.06307*.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. 2024. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. *arXiv preprint arXiv:2401.13649*.

Harsha Kokel, Michael Katz, Kavitha Srinivas, and Shirin Sohrabi. 2025. Acpbench: Reasoning about action, change, and planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26559–26568.

Ido Levy, Ben Wiesel, Sami Marreed, Alon Oved, Avi Yaeli, and Segev Shlomov. 2024. St-webagentbench: A benchmark for evaluating safety and trustworthiness in web agents. *arXiv preprint arXiv:2410.06703*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Keyu Li, Mohan Jiang, Dayuan Fu, Yunze Wu, Xiangkun Hu, Dequan Wang, and Pengfei Liu. 2025. Datasetresearch: Benchmarking agent systems for demand-driven dataset discovery. *arXiv preprint arXiv:2508.06960*.

Lingyu Li, Yixu Wang, Haiquan Zhao, Shuqi Kong, Yan Teng, Chunbo Li, and Yingchun Wang. 2024a. Reflection-bench: Evaluating epistemic agency in large language models. *arXiv preprint arXiv:2410.16270*.

Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Erran Li Li, Ruohan Zhang, and 1 others. 2024b. Embodied agent interface: Benchmarking llms for embodied decision making. *Advances in Neural Information Processing Systems*, 37:100428–100534.

Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. 2023. Api-bank: A comprehensive benchmark for tool-augmented llms. *arXiv preprint arXiv:2304.08244*.

Percy Liang, Rishi Bommasani, and et al. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2023. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv: 2308.03688*.

Yanming Liu, Xinyue Peng, Jiannan Cao, Yuwei Zhang, Xuhong Zhang, Sheng Cheng, Xun Wang, Jianwei Yin, and Tianyu Du. 2024. Tool-planner: Task planning with clusters across multiple tools. *arXiv preprint arXiv:2406.03807*.

Xiaoya Lu, Zeren Chen, Xuhao Hu, Yijin Zhou, Weichen Zhang, Dongrui Liu, Lu Sheng, and Jing Shao. 2025. Is-bench: Evaluating interactive safety of vlm-driven embodied agents in daily household tasks. *arXiv preprint arXiv:2506.16402*.

Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*.

Zixian Ma, Weikai Huang, Jieyu Zhang, Tanmay Gupta, and Ranjay Krishna. 2024. m & m's: A benchmark to evaluate tool-use for m ulti-step m ulti-modal tasks. In *European Conference on Computer Vision*, pages 18–34. Springer.

Zhao Mandi, Shreeya Jain, and Shuran Song. 2024. Roco: Dialectic multi-robot collaboration with large language models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 286–299. IEEE.

Ramesh Manuvinakurike, Emanuel Moss, Elizabeth Anne Watkins, Saurav Sahay, Giuseppe Raffa, and Lama Nachman. 2025. Thoughts without thinking: Reconsidering the explanatory value of chain-of-thought reasoning in llms through agentic pipelines. *arXiv preprint arXiv:2505.00875*.

Samuel Marks, Johannes Treutlein, Trenton Bricken, Jack Lindsey, Jonathan Marcus, Siddharth Mishra-Sharma, Daniel Ziegler, Emmanuel Ameisen, Joshua Batson, Tim Belonax, Brian Chen, Hoagy Cunningham, Carson Denison, Satvik Golechha, Akbir Khan, Jan Kirchner, Jan Leike, Austin Meek, Kei Nishimura-Gasparian, and 13 others. 2025. Auditing language models for hidden objectives. *Preprint*, arXiv:2503.10965.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. Gaia: a benchmark for general ai assistants. In *The Twelfth International Conference on Learning Representations*.

Samuel Miserendino, Michele Wang, Tejal Patwardhan, and Johannes Heidecke. 2025. Swe-lancer: Can frontier llms earn $1 million from real-world freelance software engineering? *arXiv preprint arXiv:2502.12115*.

MITRE Corporation. 2025. Atlas: Adversarial threat landscape for artificial-intelligence systems. https://atlas.mitre.org/. Accessed 2025-10-18.

Mahmoud Mohammadi, Yipeng Li, Jane Lo, and Wendy Yip. 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*, pages 6129–6139.

Atsunori Moteki, Shoichi Masui, Fan Yang, Yueqi Song, Yonatan Bisk, Graham Neubig, Ikuo Kusajima, Yasuto Watanabe, Hiroyuki Ishida, Jun Takahashi, and 1 others. 2025. Fieldworkarena: Agentic ai benchmark for real field work tasks. *arXiv preprint arXiv:2505.19662*.

Yohei Nakajima. 2023. Babyagi: An autonomous task management system. https://github.com/yoheinakajima/babyagi. Original release date: March 28, 2023. Accessed: [Insert Current Date].

Vaskar Nath, Pranav Raja, Claire Yoon, and Sean Hendryx. 2025. Toolcomp: A multi-tool reasoning & process supervision benchmark. *arXiv preprint arXiv:2501.01290*.

NIST. 2023. Artificial intelligence risk management framework (ai rmf 1.0). Technical Report NIST AI 100-1, NIST.

NTIA. 2021. The minimum elements for a software bill of materials (sbom). Technical report, U.S. Department of Commerce.

Barak Or. 2025. Mttr-a: Measuring cognitive recovery latency in multi-agent systems. *arXiv preprint arXiv:2511.20663*.

OWASP Foundation. 2025. Owasp top 10 for large language model applications (2025). https://genai.owasp.org/llm-top-10/. Accessed 2025-10-18.

Yichen Pan, Dehan Kong, Sida Zhou, Cheng Cui, Yifei Leng, Bing Jiang, Hangyu Liu, Yanyi Shang, Shuyan Zhou, Tongshuang Wu, and 1 others. 2024. Webcanvas: Benchmarking web agents in online environments. *arXiv preprint arXiv:2406.12373*.

Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. 2024. Gorilla: Large language model connected with massive apis. *Advances in Neural Information Processing Systems*, 37:126544–126565.

Ralph Peeters, Aaron Steiner, Luca Schwarz, Julian Yuya Caspary, and Christian Bizer. 2025. Webmall–a multi-shop benchmark for evaluating web agents. *arXiv preprint arXiv:2508.13024*.

Aske Plaat, Max van Duijn, Niki van Stein, Mike Preuss, Peter van der Putten, and Kees Joost Batenburg. 2025. Agentic large language models, a survey. *arXiv preprint arXiv:2503.23037*.

Maximus Powers, Shaina Raza, Alex Chang, Rehana Riaz, Umang Mavani, Harshitha Reddy Jonala, Ansh Tiwari, and Hua Wei. 2025. Responsible ai in nlp: Gus-net span-level bias detection dataset and benchmark for generalizations, unfairness, and stereotypes.

Siyuan Qi, Shuo Chen, Yexin Li, Xiangyu Kong, Junqi Wang, Bangcheng Yang, Pring Wong, Yifan Zhong, Xiaoyuan Zhang, Zhaowei Zhang, and 1 others. 2024. Civrealm: A learning and reasoning odyssey in civilization for decision-making agents. *arXiv preprint arXiv:2401.10568*.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, and 1 others. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

Xiaodong Qu, Andrews Damoah, Joshua Sherwood, Peiyan Liu, Christian Shun Jin, Lulu Chen, Minjie Shen, Nawwaf Aleisa, Zeyuan Hou, Chenyu Zhang, and 1 others. 2025. A comprehensive review of ai agents: Transforming possibilities in technology and beyond. *arXiv preprint arXiv:2508.11957*.

Muhammad Shihab Rashid, Christian Bock, Yuan Zhuang, Alexander Buchholz, Tim Esler, Simon Valentin, Luca Franceschi, Martin Wistuba, Prabhu Teja Sivaprasad, Woo Jung Kim, and 1 others. 2025. Swe-polybench: A multi-language benchmark for repository level evaluation of coding agents. *arXiv preprint arXiv:2504.08703*.

Shaina Raza, Rizwan Qureshi, Marcelo Lotif, Aman Chadha, Deval Pandya, and Christos Emmanouilidis. 2025a. Just as humans need vaccines, so do models: Model immunization to combat falsehoods. *arXiv preprint arXiv:2505.17870*.

Shaina Raza, Caesar Saleh, Azib Farooq, Emrul Hasan, Franklin Ogidi, Maximus Powers, Veronica Chatrath, Marcelo Lotif, Karanpal Sekhon, Roya Javadi, Haad Zahid, Anam Zahid, Vahid Reza Khazaie, and Zhenyu Yu. 2025b. Vilbias: Detecting and reasoning about bias in multimodal content. *Preprint*, arXiv:2412.17052.

Shaina Raza, Ashmal Vayani, Aditya Jain, Aravind Narayanan, Vahid Reza Khazaie, Syed Raza Bashir,

Elham Dolatabadi, Gias Uddin, Christos Emmanouilidis, Rizwan Qureshi, and 1 others. 2025c. Vldbench evaluating multimodal disinformation with regulatory alignment. *arXiv preprint arXiv:2502.11361*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. *Preprint*, arXiv:1602.04938. Published in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016).

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J Maddison, and Tatsunori Hashimoto. 2023. Identifying the risks of lm agents with an lm-emulated sandbox. *arXiv preprint arXiv:2309.15817*.

Yongliang Shen, Kaitao Song, Xu Tan, Wenqi Zhang, Kan Ren, Siyu Yuan, Weiming Lu, Dongsheng Li, and Yueting Zhuang. 2024. Taskbench: Benchmarking large language models for task automation. *Advances in Neural Information Processing Systems*, 37:4540–4574.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2020. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*.

Dong Shu, Xuansheng Wu, Haiyan Zhao, Daking Rai, Ziyu Yao, Ninghao Liu, and Mengnan Du. 2025. A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models. *arXiv preprint arXiv:2503.05613*.

Manish Shukla. 2025. Adaptive monitoring and real-world evaluation of agentic ai systems. *arXiv preprint arXiv:2509.00115*.

Zachary S Siegel, Sayash Kapoor, Nitya Nagdir, Benedikt Stroebl, and Arvind Narayanan. 2024. Core-bench: Fostering the credibility of published research through a computational reproducibility agent benchmark. *arXiv preprint arXiv:2409.11363*.

SLSA Community. 2023. Slsa specification v1.0. https://slsa.dev/spec/v1.0/. Accessed 2025-10-18.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, and 1 others. 2025. Paperbench: Evaluating ai's ability to replicate ai research. *arXiv preprint arXiv:2504.01848*.

Katharina Stein, Daniel Fišer, Jörg Hoffmann, and Alexander Koller. 2023. Autoplanbench: Automatically generating benchmarks for llm planners from pddl. *arXiv preprint arXiv:2311.09830*.

Haochen Sun, Shuwen Zhang, Lujie Niu, Lei Ren, Hao Xu, Hao Fu, Fangkun Zhao, Caixia Yuan, and Xiaojie Wang. 2025. Collab-overcooked: Benchmarking and evaluating large language models as collaborative agents. *arXiv preprint arXiv:2502.20073*.

Sagar Tamang and Dibya Jyoti Bora. 2025. Enforcement agents: Enhancing accountability and resilience in multi-agent ai frameworks. *arXiv preprint arXiv:2504.04070*.

Steven Thurgood, Betsy Beyer, and David Ferguson. 2018. Example error budget policy. https://sre.google/workbook/error-budget-policy/. In *The Site Reliability Workbook*, Google SRE.

Chunlin Tian, Xinpeng Qin, Kahou Tam, Li Li, Zijian Wang, Yuanzhe Zhao, Minglei Zhang, and Chengzhong Xu. 2025. Clone: Customizing llms for efficient latency-aware inference at the edge. *Preprint*, arXiv:2506.02847.

Shulin Tian, Ziniu Zhang, Liangyu Chen, and Ziwei Liu. 2024. Mmina: Benchmarking multihop multimodal internet agents. *arXiv preprint arXiv:2404.09992*.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. 2024. Appworld: A controllable world of apps and people for benchmarking interactive coding agents. *arXiv preprint arXiv:2407.18901*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Preprint*, arXiv:2305.04388.

Ikechukwu Uchendu, Jason Jabbour, Korneel Van den Berghe, Joel Runevic, Matthew Stewart, Jeffrey Ma, Srivatsan Krishnan, Izzeddin Gur, Austin Huang, Colton Bishop, and 1 others. 2025. A2perf: Real-world autonomous agents benchmark. *arXiv preprint arXiv:2503.03056*.

European Union. 2018. General data protection regulation. [Accessed 01-10-2024].

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 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.

Aimee Van Wynsberghe. 2021. Sustainable ai: Ai for sustainability and the sustainability of ai. *AI and Ethics*, 1(3):213–218.

VCWG. 2025. Verifiable credentials data model 2.0. W3c recommendation, World Wide Web Consortium (W3C).

Charles L Wang, Trisha Singhal, Ameya Kelkar, and Jason Tuo. 2025a. Mi9: An integrated runtime governance framework for agentic ai. *arXiv preprint arXiv:2508.03858*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, and 1 others. 2024a. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Ning Wang, Zihan Yan, Weiyang Li, Chuan Ma, He Chen, and Tao Xiang. 2025b. Advancing embodied agent security: From safety benchmarks to input moderation. *arXiv preprint arXiv:2504.15699*.

Wei Wang, Dan Zhang, Tao Feng, Boyan Wang, and Jie Tang. 2024b. Battleagentbench: A benchmark for evaluating cooperation and competition capabilities of language models in multi-agent systems. *arXiv preprint arXiv:2408.15971*.

Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. 2023. Mint: Evaluating llms in multi-turn interaction with tools and language feedback. *arXiv preprint arXiv:2309.10691*.

Cheng-Kuang Wu, Zhi Rui Tam, Chieh-Yen Lin, Yun-Nung Vivian Chen, and Hung-yi Lee. 2024a. Streambench: Towards benchmarking continuous improvement of language agents. *Advances in Neural Information Processing Systems*, 37:107039–107063.

Xixi Wu, Yifei Shen, Caihua Shan, Kaitao Song, Siwei Wang, Bohang Zhang, Jiarui Feng, Hong Cheng, Wei Chen, Yun Xiong, and 1 others. 2024b. Can graph learning improve planning in llm-based agents? *Advances in Neural Information Processing Systems*, 37:5338–5383.

Zhiheng Xi, Yiwen Ding, Wenxiang Chen, Boyang Hong, Honglin Guo, Junzhe Wang, Xin Guo, Dingwen Yang, Chenyang Liao, Wei He, and 1 others. 2025. Agentgym: Evaluating and training large language model-based agents across diverse environments. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27914–27961.

Ruixuan Xiao, Wentao Ma, Ke Wang, Yuchuan Wu, Junbo Zhao, Haobo Wang, Fei Huang, and Yongbin Li. 2024. Flowbench: Revisiting and benchmarking workflow-guided planning for llm-based agents. *arXiv preprint arXiv:2406.14884*.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, and 1 others. 2024. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094.

Mingzhe Xing, Rongkai Zhang, Hui Xue, Qi Chen, Fan Yang, and Zhen Xiao. 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*, pages 6061–6072.

Qiantong Xu, Fenglu Hong, Bo Li, Changran Hu, Zhengyu Chen, and Jian Zhang. 2023. On the tool manipulation capability of open-source large language models. *arXiv preprint arXiv:2305.16504*.

Hui Yang, Sifu Yue, and Yunzhong He. 2023. Auto-gpt for online decision making: Benchmarks and additional opinions. *arXiv preprint arXiv:2306.02224*.

Yingxuan Yang, Huacan Chai, Yuanyi Song, Siyuan Qi, Muning Wen, Ning Li, Junwei Liao, Haoyi Hu, Jianghao Lin, Gaowei Chang, and 1 others. 2025. A survey of ai agent protocols. *arXiv preprint arXiv:2504.16736*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5369–5380. Dataset available at: https://hotpotqa.github.io/.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. Webshop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems (NeurIPS)*. Dataset available at: https://github.com/princeton-nlp/WebShop.

Asaf Yehudai, Lilach Eden, Alan Li, Guy Uziel, Yilun Zhao, Roy Bar-Haim, Arman Cohan, and Michal Shmueli-Scheuer. 2025. Survey on evaluation of llm-based agents. *arXiv preprint arXiv:2503.16416*.

Sheng Yin, Xianghe Pang, Yuanzhuo Ding, Menglan Chen, Yutong Bi, Yichen Xiong, Wenhao Huang, Zhen Xiang, Jing Shao, and Siheng Chen. 2024. Safeagentbench: A benchmark for safe task planning of embodied llm agents. *arXiv preprint arXiv:2412.13178*.

Ori Yoran, Samuel Joseph Amouyal, Chaitanya Malaviya, Ben Bogin, Ofir Press, and Jonathan Berant. 2024. Assistantbench: Can web agents solve realistic and time-consuming tasks? *arXiv preprint arXiv:2407.15711*.

Xueqiao Zhang, Chao Zhang, Jianwen Sun, Jun Xiao, Yi Yang, and Yawei Luo. 2025. Eduplanner: Llm-based multi-agent systems for customized and intelligent instructional design. *IEEE Transactions on Learning Technologies*.

Yikai Zhang, Siyu Yuan, Caiyu Hu, Kyle Richardson, Yanghua Xiao, and Jiangjie Chen. 2024. Timearena: Shaping efficient multitasking language agents in a time-aware simulation. *arXiv preprint arXiv:2402.05733*.

Huaixiu Steven Zheng, Swaroop Mishra, Hugh Zhang, Xinyun Chen, Minmin Chen, Azade Nova, Le Hou, Heng-Tze Cheng, Quoc V Le, Ed H Chi, and 1 others. 2024. Natural plan: Benchmarking llms on natural language planning. *arXiv preprint arXiv:2406.04520*.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, and 1 others. 2023a. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and 1 others. 2023b. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

Yifei Zhou, Song Jiang, Yuandong Tian, Jason Weston, Sergey Levine, Sainbayar Sukhbaatar, and Xian Li. 2025. Sweet-rl: Training multi-turn llm agents on collaborative reasoning tasks. *arXiv preprint arXiv:2503.15478*.

Changxi Zhu, Mehdi Dastani, and Shihan Wang. 2024. A survey of multi-agent deep reinforcement learning with communication. *Autonomous Agents and Multi-Agent Systems*, 38(1):4.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, and 1 others. 2025a. Multiagentbench: Evaluating the collaboration and competition of llm agents. *arXiv preprint arXiv:2503.01935*.

Yuxiao Zhu, Junfeng Chen, Xintong Zhang, Meng Guo, and Zhongkui Li. 2025b. Dexter-llm: Dynamic and explainable coordination of multi-robot systems in unknown environments via large language models. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10182–10189. IEEE.

A 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 180 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 included.

Benchmark Name*	Description
WebArena (Zhou et al., 2023a)	Realistic web environment benchmark simulating tasks in e-commerce, social forums, collaborative coding, and content management; evaluates functional correctness on 812 web-based tasks.
ToolLLM (Qin et al., 2023)	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.
AgentBoard (Chang et al., 2024)	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.
CuisineWorld (Gong et al., 2023)	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.
MINT (Wang et al., 2023)	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
AppWorld (Trivedi et al., 2024)	Evaluates whether agents can navigate real-world computer systems, execute complex tasks, and coordinate actions across multiple applications.
AgentBench (Liu et al., 2023)	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).
VisualWebArena (Koh et al., 2024)	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.
GAIA (Mialon et al., 2023)	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
Roco (Multi-robot) (Mandi et al., 2024)	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.
ColBench (Zhou et al., 2025)	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
PaperBench (Starace et al., 2025)	Benchmarks AI systems on their ability to replicate the methodology and results of existing AI research papers.
Webshop (Yao et al., 2022)	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
EAsafetyBench (Wang et al., 2025b)	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.
ToolEmu (Ruan et al., 2023)	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
A2Perf (Uchendu et al., 2025)	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.
BFCL (Function-Calling) (Patil et al., 2024)	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.
DCA-Bench (Huang et al., 2025)	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.
MetaTool (Huang et al., 2023b)	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
Reflection-Bench (Li et al., 2024a)	Assesses intrinsic epistemic agency by decomposing reflection into seven cognitive components, including belief updating, prediction, and meta-reflection.
AgentArch (Bogavelli et al., 2025)	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
Natural Plan (Zheng et al., 2024)	Designed to evaluate how LLMs handle real-world planning tasks when presented solely in natural language.
CREW-Wildfire (Hyun et al., 2025)	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.
API-Bank (Li et al., 2023)	Provides a large-scale benchmark for tool-augmented LLMs, evaluating their capacity to invoke and coordinate multiple real-world APIs effectively.
TaskBench (Shen et al., 2024)	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.
SWELancer (Miserendino et al., 2025)	Targets freelance coding tasks, representing the latest trend in benchmark development by linking agent performance to monetary value and long-term reasoning.
m&m's (Ma et al., 2024)	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.

Continued on next page

Benchmark Name*	Description
OSWorld (Xie et al., 2024)	A scalable, real computer environment with 369 tasks designed to evaluate multimodal agents' navigation and execution across various operating systems and applications.
Collab-Overcooked (Sun et al., 2025)	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.
ScienceAgentBench (Chen et al., 2024c)	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.
LLM-Coordination (Agashe et al., 2023)	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.
Tool-Planner (Liu et al., 2024)	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.
CivRealm (Qi et al., 2024)	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.
Android Environment Benchmark (Xing et al., 2024)	Analyzes LLM agents' weaknesses in complex Android operating environments, focusing on robustness, adaptability, and task success rate.
LLMARENA (Chen et al., 2024b)	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.
AssistantBench (Yoran et al., 2024)	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.
BattleAgentBench (Wang et al., 2024b)	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.
SafePlan-Bench (Yin et al., 2024)	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.
VillagerBench (Dong et al., 2024)	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.
DatasetResearch (Li et al., 2025)	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.
MultiAgentBench (Zhu et al., 2025a)	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.
MMINA (Tian et al., 2024)	Benchmarks multimodal internet agents on multi-hop reasoning tasks involving text, images, and web-based information retrieval.
Embodied Agent Interface (EAI) (Li et al., 2024b)	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.
SocialBench (Chen et al., 2024a)	Evaluates the social intelligence and role-playing abilities of conversational agents through structured multi-role social interaction scenarios.
AutoPlanBench (Stein et al., 2023)	Evaluates LLM agent planning abilities specifically in everyday scenarios, demonstrating agents lag behind classical symbolic planners.
IS-Bench (Lu et al., 2025)	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.
SWE-bench (Jimenez et al., 2023)	Benchmarks agent performance on software development tasks, requiring agents to resolve real-world software issues found in public repositories.
BrowserGym (Chezelles et al., 2024)	Offers a unified, open ecosystem for training and benchmarking web agents under controlled browser simulation environments.
ACPBench (Kokel et al., 2025)	A benchmark focused on evaluating LLMs on core reasoning skills across 7 reasoning tasks and 13 planning domains using formally synthesized problems.
FlowBench (Xiao et al., 2024)	Evaluates workflow planning abilities, specifically targeting expertise-intensive tasks that require sophisticated sequencing and orchestration.
OmniACT (Kapoor et al., 2024)	Tests agents' capacity to coordinate actions and execute complex tasks across multiple applications within full-scale computer operating environments.
WebCanvas (Pan et al., 2024)	Tests web agents' ability to perform tasks in realistic browser-based environments with interactive, dynamic web pages.
Sotopia (Zhou et al., 2023b)	Provides an interactive benchmark for evaluating the social intelligence and cooperation skills of language agents through multi-agent social simulations.
MLAgentBench (Huang et al., 2023a)	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.
StreamBench (Wu et al., 2024a)	A challenging benchmark assessing agents' ability to continuously improve performance by leveraging external memory components and previous interactions over time.

Continued on next page

Benchmark Name*	Description
Can-Graph (Wu et al., 2024b)	Introduces a benchmark to test whether graph-structured reasoning can enhance LLM-based agents' planning and decision-making in multi-step tasks.
REALM-Bench (Geng and Chang, 2025)	Evaluates LLMs and multi-agent systems on dynamic, complex real-world planning and scheduling tasks, focusing on coordination and adaptation to disruption.
Core-Bench (Siegel et al., 2024)	Assesses the computational reproducibility of AI research by benchmarking agents that attempt to reproduce published experiments.
TimeArena (Zhang et al., 2024)	Evaluates multitasking LLM agents in a time-constrained simulated environment to measure efficiency, prioritization, and scheduling capabilities.

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

Table A.1: List of all the benchmarks associated with agentic AI from year 2023–2025

Paper Name	Evaluation metric	Description
FieldWorkArena (Moteki et al., 2025)	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
	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
ScaleAI ToolComp (Nath et al., 2025)	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
	Exact match	Programmatically checks sorted and unsorted lists, numbers (within tolerance) and strings in the final answer to ensure exact equality with ground truth
	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
MLE-Bench (Chan et al., 2024)	Pass@k	Measures success when the agent is given multiple attempts; pass@k improvements quantify reliability of repeated attempts.
	Medal rate	Agents earn Bronze/Silver/Gold medals based on Kaggle leaderboards; the headline metric is the fraction of attempts that receive any medal
	Resource-scaling metrics	Evaluates how longer time budgets or multiple seeds affect performance (e.g., 24-h runs vs. short runs).
	Raw competition scores	Reports the agent's score on each competition's own metric (e.g., AUROC, F1, RMSE), enabling fair comparison across tasks
ToolBench (Xu et al., 2023)	Tool Selection Accuracy	Accuracy of choosing appropriate tools for given tasks.
	Tool Execution Success	Rate of successful tool executions without errors.
	Parameter Accuracy	Correctness of API call parameters when using tools.
WebShop (Yao et al., 2022)	Task Success Rate	Success rate in completing realistic online shopping tasks.
	Attribute F1	F1 score for correctly identifying required product attributes.
ST-WebAgentBench (Levy et al., 2024)	Risk Ratio	Quantifies policy violations across safety and trustworthiness dimensions.
	Completion Under Policy (CuP)	Success rate when both completing task and respecting all policy constraints.
MLAgentBench (Huang et al., 2023a)	Token Usage	Number of input and output tokens used by the agent, another efficiency metric
	Wall-clock time	Time taken by the agent to complete its run; used to measure efficiency
	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
MINT (Wang et al., 2023)	Improvement rate	Slope obtained from regressing success rate against the number of allowed interaction turns; quantifies how additional turns improve performance
	Success rate	Percentage of tasks successfully completed under a fixed interaction limit
SWE-PolyBench (Rashid et al., 2025)	CST Node-level Retrieval	Precision in locating specific code structures needing changes.
	File-level Localization	Accuracy in identifying correct files requiring modification.
SUPER (Bogin et al., 2024)	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
	Accuracy	Fraction of tasks whose final outputs (numerical or textual) exactly match the gold solution
WebArena (Zhou et al., 2023a), VisualArena (Koh et al., 2024)	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
	Must exclude	Rewards the agent if specific undesirable strings are absent from the output
	eval fuzzy image match	Compares images using structural similarity (SSIM) to assess whether the agent produced the correct image
	Exact match	Text evaluation function that returns 1 if the agent's answer string exactly equals the ground-truth answer
	Task success rate	Overall percentage of tasks that are completed successfully across all categories; derived from per-task rewards in WebArena/VisualWebArena.
	Must include	Checks that the generated response contains required keywords or phrases; partial match yields positive reward
	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
Mind2Web (Deng et al., 2023)	Operation F1	F1 score for correct operation prediction on web elements.
	Element Accuracy	Accuracy of selecting correct webpage elements for interaction.
HotPotQA (Yang et al., 2018)	Supporting Fact F1	F1 score for identifying relevant supporting facts across reasoning steps.
	Multi-hop F1	F1 score for multi-step reasoning chains in question answering.

Continued on next page

Paper Name	Evaluation metric	Description
Mind2Web2 (Gou et al., 2025)	Agent-as-a-Judge Score	LLM-based evaluation score using tree-structured rubrics.
	Partial Completion	Average root score measuring partial task completion.
IBM WatsonX (IBM)	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).
	Task success / completion rate	fraction of tasks the agent completes successfully (standard task-level performance).
	Risk / severity score	aggregated indicator of potential harms/violations produced by agent behavior (operational risk).
	Average steps / action count	mean number of actions taken to finish a task (efficiency).
ToolLLM (Qin et al., 2023)	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
	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
WebMall (Peeters et al., 2025)	Cross-Shop Success Rate	Success rate for tasks requiring navigation across multiple e-commerce sites.
	Price Comparison Accuracy	Accuracy in identifying best prices across different online shops.
TRAIL (Deshpande et al., 2025)	Joint Accuracy	Accuracy of identifying both error category and location in agent traces.
	Error Detection Rate	Rate of correctly identifying errors in multi-step agent workflows.
AgentBench (Liu et al., 2023)	Action F1	In web-browsing tasks, token-level F1 between predicted operations and ground truth; used as micro-level metric
	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
	Success Rate	Primary metric across eight environments; percentage of tasks in which the agent achieves the goal
	Answer F1	For knowledge-graph queries, computes F1 between the set of answers predicted by the agent and the ground truth
	Step success rate	Fraction of steps where the agent selected the correct element and action
	Win rate / Reward score	In digital card games, counts winning rounds, total rounds, win rate and damage rate, then computes a final reward score
	Progress (lateral puzzles)	Measures the portion of guessed-out bullets when the agent fails to solve a puzzle within the allowed rounds
GAIA (Mialon et al., 2023)	Quasi Exact Match	Automatically evaluates answers by normalizing the predicted answer and checking exact equality with the ground truth; ensures fast, factual scoring.
Rise of Agentic AI (Bandi et al., 2025)	Graph Edit Distance (GED)	Measures structural similarity between the agent-generated task graph and the ground truth; lower GED indicates closer alignment.
	Cooperative behavior	Evaluates how well multiple agents collaborate and coordinate to achieve common goals.
	Rule fidelity	Evaluates how accurately the symbolic rules learned by the agent mirror the actual decision-making process.
	User satisfaction	Measures how well the system meets user preferences; assessed via user ratings or Net Promoter Score
	Click-through rate (CTR) / Gross Merchandise Value (GMV)	Application-specific metrics that measure user engagement (CTR) and monetary impact of agent recommendations (GMV)
	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
	Fairness / Bias mitigation	Metrics and techniques used to detect and reduce demographic bias in agent outputs
	Transparency	Degree to which the agent's internal decision processes are open for inspection; evaluated through user-facing clarity or external audits.
	Adaptability	Measures how quickly an agent can adjust to changing tasks or environments; tested via dynamic policy simulations.
	Task completion time (TCT)	Measures the time taken for the agent to plan and execute a task, providing an operational efficiency metric.
	Explainability	Qualitative metric evaluating whether an agent's reasoning steps are understandable; assessed via self-reflection or cross-agent reflection for foundation models.

Continued on next page

Paper Name	Evaluation metric	Description
Adaptive Monitoring (Shukla, 2025)	Harm Reduction Score	Quantifies agent's ability to minimize potential negative outcomes.
	Adaptive Multi-Dimensional Score	Composite score using exponentially weighted moving averages across dimensions.
	Goal Drift	Measure of how much an agent deviates from intended objectives over time.
ALFWorld (Shridhar et al., 2020)	Success Rate	Task completion rate in text-based embodied household environments.
	Progress Rate	Partial completion rate when full task success is not achieved.
PlanBench(Valmeekam et al., 2023)	Partial correctness	The authors suggest future metrics that give credit for partially correct plans when not all preconditions/effects are satisfied.
	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)
τ-Bench (Yao et al., 2024)	Pass^k	Reliability metric that measures how often the same task is completed successfully over k independent trials; highlights consistency across runs
	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)
SWEET-RL (Zhou et al., 2025)	Win Rate	Rate of preferred responses in human-AI collaborative task completion.
	Turn-wise Advantage	Advantage score measuring quality of each decision in multi-turn interactions.
SWE-Bench	Resolved Rate	Percentage of GitHub issues successfully resolved by passing tests.

Table A.2: Papers and their evaluation metric along with a description.

Control Domain	Concrete Mechanisms (e.g.)	SLOs (e.g.)	Audit Artifacts	Standards Mapping	Lifecycle Phase
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 (NIST, 2023); ISO 42001: 8.10 (ISO, 2023); EU AI Act: Serious incident reporting (European Union, 2024).	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 (NIST, 2023); ISO 42001: 8.7 (ISO, 2023); EU AI Act: Security & robustness (European Union, 2024).	Build, Operate
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 (NIST, 2023); ISO 42001: 9.1 (ISO, 2023); EU AI Act: Technical documentation (European Union, 2024).	Test, 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 (NIST, 2023); ISO 42001: 6.1, 8.2 (ISO, 2023); EU AI Act: Technical documentation & controls (European Union, 2024).	Design, 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 (MITRE Corporation, 2025); OWASP LLM (OWASP Foundation, 2025); NIST: Measure/Manage (NIST, 2023); ISO 42001: 8.9. (ISO, 2023)	Pre-release, Continuous	—
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 (NIST, 2023): Govern/Manage; ISO/IEC 42001: 8.3, 8.5 (ISO, 2023); EU AI Act: Risk mgmt & post-market monitoring (European Union, 2024).	Design, Deploy, Operate	—
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 (NIST, 2023); ISO 42001: 8.4 (ISO, 2023); NTIA SBOM (NTIA, 2021); SLSA (SLSA Community, 2023).	Plan, 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 (Union, 2018); NIST: Map / Measure (NIST, 2023); ISO 42001: 6.2, 8.8 (ISO, 2023); EU AI Act: Risk mgmt/data (European Union, 2024).	Design, 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 (NIST, 2023); ISO 42001: 8.1 (ISO, 2023); EU AI Act (European Union, 2024): QMS practices.	Build, Release
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 (OWASP Foundation, 2025); NIST: Manage (NIST, 2023); ISO 42001: 8.5 (ISO, 2023); EU AI Act: Security controls (European Union, 2024).	Build, Operate

Continued on next page

Control Domain	Concrete Mechanisms (e.g.)	SLOs (e.g.)	Audit Artifacts	Standards Mapping	Lifecycle Phase
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), PJR (prompt-jailbreak rate), MTTC; SLO docs; budget dashboards; review minutes; corrective actions; trend analyses.	NIST: Govern/Manage (NIST, 2023); ISO 42001: 6.1, 9.1 (ISO, 2023); SRE error-budgets (Thurgood et al., 2018).	Plan, 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 (NIST, 2023); ISO 42001: 8.6 (ISO, 2023); EU AI Act: Technical documentation (European Union, 2024); NTIA SBOM/SLSA (NTIA, 2021).	Build, Operate
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 (NIST, 2023); ISO 42001: 7.2 (ISO, 2023); EU AI Act: Human oversight (European Union, 2024).	Design, 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 (NIST, 2023); EU AI Act: Post-market monitoring (European Union, 2024); ISO 42001: 9.1. (ISO, 2023)	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 (European Union, 2024); NIST: Govern (NIST, 2023); ISO 42001: 8.6. (ISO, 2023)	Release, Operate

Legend: 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.

Table A.3: Regulatory alignment controls mapped to mechanisms, measurable SLOs, audit artifacts, standards, and lifecycle phase.

B Experimental setup

B.1 Hal Harness Environment Configuration

To ensure reproducible evaluation of agent capabilities, we utilize the HAL (Holistic Agent Leaderboard) harness. The local execution environment requires a rigorous configuration to support requirements of interactive benchmarks such as `AssistantBench` and `GAIA`.

The primary environment relies on a standard Anaconda distribution with Python 3.10+ isolation. However, these benchmarks impose strict containerization requirements to sandbox the agent's tool usage (e.g., file system manipulation, web browsing) and prevent host contamination. Consequently, the local machine must host a Docker daemon configured with at least 4GB of memory and 2 CPU cores per container. Unlike purely code-based benchmarks, effective evaluation of `AssistantBench` and `GAIA` often necessitates a container environment with configured network policies to support realistic information retrieval and external API interactions without compromising security.

For agent logging and cost tracking, the environment is integrated with Weave and Wandb. API keys for proprietary models (e.g., OpenAI, Anthropic) must be exported as environment variables prior to the initialization of the `hal-eval` CLI.

For each benchmark, we use a representative model configuration aligned with the task profile (a stronger reasoning model for GAIA and a lightweight model for web navigation). Where relevant, we contextualize our results using publicly reported baselines from the HAL leaderboard (e.g., Claude Sonnet 4.5 and GPT-5 Medium), rather than re-running all baselines. **Table B.1** summarizes those requirements.

Category	**Requirement / Specification**
	Hardware Constraints
Memory (RAM)	$\geq$ 4GB per Docker container
Processor (CPU)	$\geq$ 2 Cores per Docker container
Network	Open policies for external API/Web access
	Software Environment
OS	Ubuntu 24.04.2 LTS
Runtime	Python 3.10+ (Anaconda Distribution)
Containerization	Docker Daemon
Logging	Weave, Wandb
CLI Tool	`hal-eval`
Configuration	Environment Variables (OpenAI/Anthropic Keys)

Table B.1: Hardware and Software Specifications for Local HAL Harness Evaluation

Metric	Definition	Evidence	Transp. Dim.
	Outcome Metrics		
Success Rate	Percentage of tasks completed successfully	Benchmark state	–
Accuracy	Correctness of final answer (exact/fuzzy match)	Benchmark verifier	–
Token Usage	Total input/output tokens consumed per run	Runtime logs	Efficiency
Task Duration	Wall-clock time to task completion	Runtime logs	Efficiency
	Process Metrics		
Trace Completeness	Fraction of required fields present per step	Schema checks	Traceability
Trace Consistency	Valid ordering and linkage; no orphaned steps	Trace validator	Traceability
Tool-Call Validity	Tool calls conforming to declared schema	Tool logs	Auditability
Tool-Call Correctness	Appropriateness of tool choice for subgoal	Verifier / rubric	Explainability
Attribution Coverage	Steps with identifiable agent/module origin	Trace metadata	Auditability
	LLM-as-Judge Metrics		
Recovery Rate	Ability to recover from intermediate failures	Judge eval.	Traceability
Pivot Ability	Appropriate redirection at dead-ends	Judge eval.	Explainability
Context Switching	Appropriate transitions between task contexts	Judge eval.	Traceability
	Governance Metrics		
Policy Adherence	Actions compliant with declared policies	Policy logs	Governance
Intervention Rate	Human/gate interventions per run	Oversight trace	Governance
	Trace-Based Metrics (Ours)		
$R^{\dagger}_{\text{imp}}$	Implicit Error Resilience: probability of corrective action following environment error (Eq. 1)	Execution trace	Traceability
$B^{\dagger}_{\text{info}}$	Information Gathering Breadth: unique external sources accessed per trajectory (Eq. 2)	Execution trace	Auditability
$C^{\dagger}_{\text{ops}}$	Operational Complexity Score: rate of context-switching between tool categories (Eq. 3)	Execution trace	Traceability

Table B.2: Evaluation metrics used in this study, organized by category and mapped to transparency dimensions. Metrics marked with † are proposed in this work.