Scientific Research Publishing

Journal of Software Engineering and Applications, 2025, 18(9), 358-372
https://www.scirp.org/journal/jsea
ISSN Online: 1945-3124
ISSN Print: 1945-3116

A Review of Agent Data Evaluation: Status, Challenges, and Future Prospects as of 2025

Shaohan Wang*

Research and Innovation Center, COSCO Shipping Technology Co., Ltd., Shanghai, China
Email: *wang.shaohan@coscoshipping.com

How to cite this paper: Wang, S.H. (2025) A Review of Agent Data Evaluation: Status, Challenges, and Future Prospects as of 2025. *Journal of Software Engineering and Applications*, **18**, 358-372.
https://doi.org/10.4236/jsea.2025.189021

Received: March 29, 2025
Accepted: July 1, 2025
Published: September 17, 2025

Abstract

With the rapid advancement of large language models (LLMs), agents capable of autonomous perception, decision-making, and action have emerged as a frontier paradigm in artificial intelligence. These entities are transitioning from academic research to complex real-world applications. However, the rapid iteration of agent capabilities poses severe challenges to evaluation methodologies—particularly in assessing their core competencies in data processing and evaluation. As of 2025, the field of agent data evaluation exhibits a dynamic yet fragmented landscape. Traditional static dataset-based evaluations are no longer sufficient to measure agent performance in open, dynamic environments. The research community is actively shifting toward more interactive and realistic benchmarking paradigms. Despite the emergence of innovative benchmarks such as ToolBench and MLAgentBench, there remains a widespread lack of unified evaluation standards, widely accepted metric systems, and mature methodologies. This paper systematically reviews the state of agent data evaluation in 2025, tracing the evolution from traditional metrics to emerging process-oriented ones. Building upon this, we delve into the methodology of dataset and benchmark design, with particular attention to key elements in experimental design, such as controlled experiments, sample size determination, and statistical analysis. Furthermore, we analyze the core challenges facing the field, including the "realism gap" between evaluation and real-world tasks, the scalability dilemma of automated evaluation, and the increasingly prominent issues of data privacy and security. Our findings indicate that although potential technologies such as differential privacy and federated learning exist, dedicated privacy-preserving frameworks for agent evaluation remain in their infancy. Finally, this report outlines future research directions, emphasizing the urgent need to establish unified evaluation frameworks, develop process-oriented evaluation metrics, and formulate standardized privacy and security auditing protocols—aiming to provide a scientific foundation for building more robust, trustworthy, and responsible agent systems.

Keywords

Agent Evaluation, Large Language Model, Benchmarks, Process-Oriented Evaluation

1. Introduction

As of 2025, artificial intelligence (AI) has entered a new era, marked by the rise of agents [1]. Defined as computational entities that can autonomously act within their environment to achieve goals [2] [3], agents are rapidly transforming from theoretical concepts into powerful tools capable of executing complex tasks. Their application scope—from automated scientific research [4] [5] and industrial asset operations to everyday computer task automation—is expanding at an unprecedented pace. This transformation is primarily driven by the powerful natural language understanding, reasoning, and planning capabilities that large language models (LLMs) bestow upon agents [6] [7].

The core value of agents lies in their ability to process and utilize data. Unlike traditional models that receive fixed inputs and produce outputs, agents must actively explore their environment, gather information, invoke external tools (e.g., APIs, databases), process unstructured data, and make a series of decisions and actions based on this data. Therefore, evaluating agents is fundamentally an assessment of their entire "perceive-decide-act" cycle, with data processing capability serving as the central thread throughout. How to scientifically and accurately evaluate agents' data processing capabilities—including the accuracy of data retrieval, depth of data analysis, effectiveness of data integration, and soundness of data-driven decisions—has become crucial for assessing their intelligence, reliability, and safety, directly impacting their applicability in critical domains.

However, as of 2025, the field of agent data evaluation faces a significant gap between "capability development" and "evaluation methodology". Existing evaluation methods and benchmarks largely originate from traditional machine learning or natural language processing tasks and are insufficient to fully accommodate the interactivity, autonomy, and emergent behaviors exhibited by agents [8] [9]. Studies indicate that there is currently no widely accepted unified evaluation framework; different research efforts often adopt self-built, task-specific environments and metrics, making results difficult to compare across studies and hindering the healthy development of the field. Meanwhile, as agents become increasingly embedded in real-world data flows, ensuring data privacy and security during evaluation has become an urgent ethical and technical challenge [10] [11].

This review aims to systematically survey and analyze the state of research on agent data evaluation as of September 2025. We begin by reviewing relevant literature, discussing the evolution of evaluation frameworks, benchmarks, and key metrics. Subsequently, we delve into evaluation methodologies, particularly the types of datasets used and principles of experimental design. Next, we synthesize

and analyze the results of current evaluation practices and discuss the core challenges in depth, especially realism, scalability, and privacy-security issues. Finally, based on this analysis, we offer forward-looking perspectives on future research directions to promote standardization and deeper development in the field of agent data evaluation.

2. Literature Review

This section aims to review core literature related to agent data evaluation, covering the conceptual evolution of agents, the development of evaluation frameworks and benchmarks, the evolution of key evaluation metrics, and the emerging trends and challenges in the field.

2.1. Evolution and Theoretical Foundations of Agents

The concept of agents is not new; its theoretical foundations can be traced back to early research in distributed artificial intelligence (DAI) and multi-agent systems (MAS) [2] [12]. These studies define agents as entities possessing characteristics such as autonomy, social ability, reactivity, and pro-activeness [3] [13]. Early agent architectures, such as the BDI (Belief-Desire-Intention) model, provided theoretical frameworks for goal-directed agent behavior [14]. However, the emergence of LLMs has greatly accelerated the development of agents. LLMs, serving as powerful "brains" or "central processors", have endowed agents with unprecedented language understanding, world knowledge, and complex reasoning capabilities, enabling them to handle more open and complex tasks [6] [7]. By 2025, agent research has moved beyond single-agent paradigms and begun to focus on collaboration and communication among multiple agents. For example, the Agent-to-Agent (A2A) protocol introduced in 2025 aims to provide standardized solutions for coordination, security, and scalability in multi-agent environments [15].

2.2. The Evolution of Evaluation Frameworks and Benchmarks

Evaluation of agents has undergone a transformation from static to dynamic, and from simple to complex. Traditional AI model evaluation heavily relies on static, labeled datasets, such as ImageNet for image recognition [16] and GLUE or SQuAD for natural language understanding [17]. In this paradigm, models make predictions on a fixed test set, which are then compared against a "gold standard" answer. However, this method cannot assess an agent's ability to change the state of the world through interaction with its environment. Consequently, research trends have clearly shifted toward dynamic and interactive evaluation environments [18] [19]. In these environments, each action taken by the agent alters the state of the environment, requiring multi-step reasoning and planning to complete tasks—better reflecting real-world performance.

To address this transformation, numerous agent-specific evaluation benchmarks have recently emerged. These benchmarks can be broadly categorized into several types.

Tool Usage and API Invocation Benchmarks: Examples include ToolBench [1] and MetaTool [20], which focus on evaluating agents' ability to learn and use various external tools (APIs) to complete complex tasks. Such evaluations directly relate to agents' ability to access and process external data.

Simulation and Embodied Intelligence Benchmarks: Examples include AgentSims [21] and CuisineWorld [22], which evaluate agents' planning, navigation, and object interaction capabilities in highly realistic virtual environments. These interactions themselves constitute a continuous form of data processing.

Domain-Specific Task Benchmarks: For instance, MLAgentBench [23] is specifically designed to evaluate AI agents as research assistants in performing computer science research tasks, including literature review, code generation, and experiment analysis—complex data processing activities. Similarly, DeepResearch Bench aims to evaluate agents in deep research scenarios [24].

Specialized Data Analytics Benchmarks: Recently, benchmarks specifically targeting data analysis capabilities have emerged, such as AgentAda [25] and InfiAgent-DABench [26]. These require agents to perform exploratory analysis, visualization, and insight discovery on given datasets, making them the most directly relevant benchmarks to the theme of "data evaluation".

Although these benchmarks have advanced the field, a prominent issue remains: the lack of standardized benchmarks. Many studies still rely on custom-built benchmarks, resulting in limited cross-study comparability. A 2025 survey paper also notes that while there is abundant research on evaluating LLM-based agents, systematic evaluation frameworks are still under exploration [27]. The summarization of benchmarks is shown in **Table 1**.

Table 1. Benchmarks' summarization.

Benchmark Type	Defining Features	Example Tasks	Representative Benchmarks
Tool Usage & API Invocation	Evaluates agents' ability to discover, select, and chain external tools (e.g., APIs)	Retrieve weather, book a flight, currency conversion	ToolBench [1], MetaTool [20]
Simulation & Embodied Intelligence	Dynamic environments with physical/spatial reasoning; state evolves with actions	Navigate a kitchen, assemble furniture, avoid obstacles	AgentSims [21], WorldCuisine [22]
Domain-Specific Task Benchmarks	Focus on professional/scientific workflows requiring multi-step planning and synthesis	Conduct literature review, generate code, analyze experiments	MLAgentBench [23], DeepResearch Bench [24]
Specialized Data Analytics Benchmarks	Emphasize exploratory analysis, visualization, and insight generation from datasets	Analyze sales trends, detect log anomalies, generate reports	AgentAda [25], InfiAgent-DABench [26]

2.3. Key Evaluation Metrics

Evaluation metrics for agent data processing exhibit diversity and multi-layered characteristics.

Task Completion Metrics: These are the most fundamental dimensions, includ-

ing Task Success Rate, Accuracy, Precision, Recall, and F1-Score. For multi-step tasks, metrics such as pass@k are used to measure the proportion of successful task completions within k attempts.

Efficiency and Cost Metrics: Beyond task completion, efficiency is equally important. Relevant metrics include response time, number of steps or time required to complete a task, API call costs, and resource utilization.

Data Processing Process Metrics: Specialized metrics have emerged for specific data processing stages. For example, in information retrieval tasks, nDCG@k is used to evaluate the quality of returned resource lists. In scenarios requiring agents to generate action sequences, partial action match scores are used as evaluation criteria [28].

Human-Centered Metrics: In human-agent interaction scenarios, user satisfaction, engagement, and feedback are crucial evaluation metrics [29]. For complex data analysis tasks, the quality of final outputs may even require manual assessment by domain experts.

Limitations of Metrics: Researchers are increasingly aware of the limitations of traditional metrics. For instance, word-overlap-based metrics like BLEU or METEOR struggle when evaluating diverse, plausible action sequences generated by agents with "one-to-many" characteristics [30]. Moreover, fast, automated proxy metrics may not correlate with costly but more realistic online human-agent interaction evaluations, highlighting the gap between offline and online performance [31].

2.4. Emerging Trends and Core Challenges

As of 2025, the field of agent data evaluation exhibits several notable trends and challenges.

Trend 1: Alignment with Real-World Complexity: An increasing number of studies emphasize that evaluation tasks should move beyond logic puzzles and video games to scenarios reflecting real-world complexity, ambiguity, and dynamics [32]. For example, evaluating agent robustness in handling incomplete or noisy real-world data.

Trend 2: From Single-Agent to Multi-Agent Evaluation: With the rise of multi-agent systems, evaluation focus is shifting from assessing individual agents to evaluating collaboration, communication, and negotiation efficiency among multiple agents [15] [33].

Challenge 1: Standardization and Reproducibility: This is currently the most severe challenge. The lack of unified evaluation protocols, environments, and metrics hinders the accumulation and comparison of research results, severely constraining scientific progress in the field [34] [35].

Challenge 2: Evaluating Generalization: An agent that performs well on a specific benchmark may experience a sharp performance drop when faced with new tools, data formats, or tasks outside its training data. Designing evaluation schemes to effectively assess generalization remains an open problem.

Challenge 3: Cost and Scalability of Evaluation: High-quality evaluation of complex tasks—especially those requiring human involvement—is extremely costly and slow [28]. Developing evaluation systems that ensure quality while enabling large-scale automated deployment is a key future direction.

Challenge 4: Security and Ethical Issues: Agents may introduce bias, leak privacy, or be maliciously exploited when processing data [10]. Systematically integrating assessments of these ethical and security risks into existing evaluation frameworks remains a major challenge.

3. Methodology: Datasets and Benchmarking Paradigms

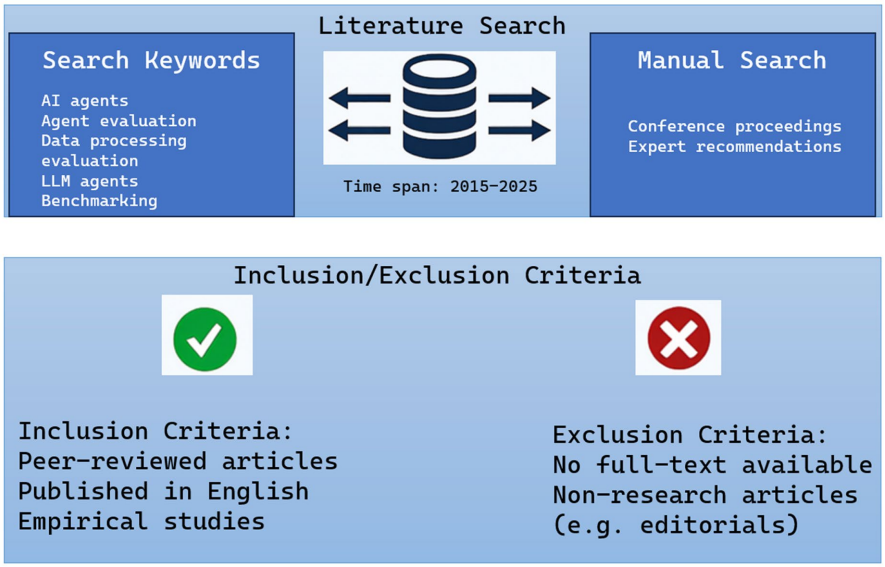

Figure 1. Methodology schematics.

This section details the methodologies used in agent data evaluation, focusing on the types of datasets relied upon, principles of controlled experimental design, and specific methods for the emerging dimension of data privacy and security, where the schematics of methodology is manifested in **Figure 1**.

3.1. Classification of Datasets and Environments for Evaluation

The data sources and environments used in agent evaluation are foundational to the methodology. Based on their characteristics, they can be classified as follows.

Static and domain-specific datasets: These datasets form the basis of evaluation, typically used to assess agents' foundational capabilities in specific subtasks. Examples include natural language inference datasets (e.g., GLUE) [17], visual question answering datasets (e.g., VQA), and code generation datasets. Although static, in multi-step tasks, agents may need to analyze such data to guide subsequent actions.

Interactive simulation environments: This is the mainstream paradigm in current agent evaluation. Environments such as AgentSims [21], CuisineWorld [22],

and game engines like Unity provide dynamic sandboxes. Agents act within them, receiving feedback and new states based on their actions. Data generation is dynamic and tightly coupled with agent behavior. These environments are particularly suitable for evaluating agents' planning, exploration, and long-term memory capabilities [18] [19].

Real-world tool-based benchmarks: Represented by ToolBench [1], these benchmarks provide agents with numerous real API interfaces (e.g., search engines, calculators, calendars). The core of evaluation lies in whether the agent can understand task requirements and correctly select, combine, and invoke these tools to acquire and process data. Such benchmarks significantly enhance evaluation realism and practicality.

Specialized data analytics benchmarks: As mentioned in Section 2.2, benchmarks like AgentAda [25] and InfiAgent-DABench [26] represent a significant advancement in evaluation methodology. They typically provide one or more datasets (e.g., CSV files) and pose open-ended data analysis questions (e.g., "Analyze this sales data and summarize key trends"). Evaluation focuses not on a single correct answer, but on the quality of the entire process—data cleaning, exploratory analysis, logical reasoning, and insight presentation. This methodology requires converting open-ended questions into closed-loop, automatically evaluable forms or involving human experts for assessment.

3.2. Principles of Controlled Experimental Design

To ensure scientific and reliable evaluation results, researchers typically follow strict principles in experimental design, although specific implementations vary across studies.

Experimental Setup: A typical controlled experiment clearly defines protocols, agents, environments, tasks, data collection, and analysis methods. Experiments usually include one or more experimental groups (using the agent under evaluation) and a control group (using a baseline agent or human) for performance comparison. All conditions unrelated to the variable being evaluated should remain consistent to minimize confounding effects.

Sample Size Determination: Determining an appropriate sample size (e.g., number of tasks, number of experimental runs) is a key and challenging aspect of experimental design. Studies show that in AI-related clinical trials, sample sizes range from 22 to 2352, indicating a lack of unified calculation standards. Nevertheless, some statistical methods are widely used. For example, power analysis using tools like G*Power, with predefined statistical power (e.g., 80%) and significance level (e.g., $\alpha = 0.05$), can calculate the required minimum sample size. Some studies also follow domain-specific heuristics—for instance, sample size should be 50 times the number of predicted variables, far exceeding the traditional 10x rule. Sample size choice directly affects statistical power, reliability, and generalizability of results.

Statistical Methods: After the experiment, statistical methods are used to ana-

lyze collected data and determine whether observed differences are statistically significant. Common methods include t-tests (for comparing two group means), ANOVA (for comparing multiple group means), and regression analysis (for exploring variable relationships). For binary outcomes like accuracy, confusion matrices and ROC curves are also used. Some studies explore finer-grained metrics, such as measuring "partial errors" to more precisely assess AI system prediction accuracy rather than simply classifying them as "correct" or "incorrect" [28].

3.3. Methodology for Evaluating Data Privacy and Security

This is a cutting-edge and critical area in evaluation methodology. As of 2025, systematic frameworks and benchmarks for agent privacy evaluation remain very limited. Current methodologies are still in the exploratory phase but can be summarized into several directions:

Attack-Based Evaluation: A "red team" approach, where specially designed attack strategies test whether agents leak privacy. For example, using carefully crafted prompts (prompt injection) to induce agents to leak sensitive information from their training data, or testing whether a malicious agent can steal private data from other agents in a multi-agent system.

Compliance-Based Evaluation: The core of this methodology is to check whether agent behavior complies with predefined privacy rules or regulations (e.g., GDPR). For instance, the RegulAI framework proposes using formal languages to express data privacy rules in multi-agent systems [36]. Evaluation experiments can be designed to verify whether agents consistently adhere to these rules across various scenarios.

Integration of Privacy-Enhancing Technologies (PETs): Methodologies are beginning to explore embedding PETs into agent architectures and evaluating their effectiveness.

Differential Privacy: A technique providing mathematically provable privacy guarantees [37] [38]. Evaluation can measure how well agents protect personal data when differential privacy is applied, and the impact of such protection on task performance (e.g., accuracy). Frameworks have been proposed to use differential privacy to protect sensitive data in multi-agent collaborative decision-making [39].

Federated Learning: Combining federated learning with multi-agent systems allows model training and collaboration without sharing raw data [40]. Evaluation can focus on communication overhead, model convergence speed, and privacy protection strength in this mode.

Information Flow Control: This technology ensures at the system level that data flows comply with security policies. Evaluation can verify whether information flow control mechanisms effectively prevent agents from leaking private information to public channels [11].

In summary, the methodology for evaluating agent data privacy and security is shifting from passive auditing to active defense and verification, but it remains far from forming standardized, operational evaluation processes.

4. Results and Discussion

This section synthesizes the literature review and methodology, distills key findings in agent data evaluation as of 2025, deeply analyzes core challenges, and proposes future research directions.

4.1. Synthesized Findings: A Dynamic yet Fragmented Evaluation Ecosystem

The most central conclusion from synthesizing all findings is: As of 2025, the field of agent data evaluation is in a vibrant yet highly fragmented and standards-deficient stage.

Rapid Evolution of Evaluation Paradigms: The community has clearly recognized that traditional, static dataset-based evaluation paradigms are insufficient for modern agents. The shift toward interactive, dynamic, real-world tool-based evaluation environments has become a consensus and mainstream trend. The emergence of innovative benchmarks such as AgentBench, ToolBench [1], and MLAgentBench [23] demonstrates the community's significant efforts in advancing evaluation methodology [1] [24] [34].

Severe Lack of Standardization: In stark contrast to the rapid evolution of paradigms, standardization lags severely. There is currently no industry-recognized "gold standard" benchmark for comprehensively evaluating agents' data processing capabilities. Different research teams tend to use different environments, tasks, and metrics, leading to "siloed" results that are difficult to compare and accumulate. This situation is especially severe in emerging areas like multi-agent system evaluation and privacy-security evaluation [34] [35].

Widening "Capability-Evaluation" Gap: Agent capabilities—especially reasoning and tool usage driven by LLMs—are evolving faster than our ability to scientifically and rigorously evaluate them. We can build seemingly powerful agents, but our understanding of their reliability, robustness, and boundary conditions lags behind. A typical example is that many evaluations still focus on final outcomes (success or failure), while methods for assessing agents' "chain of thought," decision rationale, and error attribution during data processing remain underdeveloped [41] [42].

Absence of Core Literature: A notable phenomenon is that, despite the highly specific topic, no clearly recognized "foundational" or "highly cited" review or paper on "agent data evaluation" could be identified during this research. This "negative result" indirectly confirms that this specific interdisciplinary field, as an independent and mature research direction, is still in a very early stage. Related work is scattered across conferences and journals in AI, HCI, and software engineering, without forming a cohesive core literature cluster.

4.2. In-Depth Analysis of Core Challenges

Based on the above findings, we can more deeply analyze several core challenges.

The "Realism" Gap: Although evaluation benchmarks strive to simulate the real

world, they are essentially simplified, controlled environments. Success in lab settings does not guarantee high performance when agents face infinite edge cases, subtle data format variations, API instability, or ambiguous task descriptions in the real world [32]. For example, can an agent proficient in using a weather API on ToolBench adapt when the real-world API documentation undergoes minor updates? A 2025 empirical study by Chen *et al.* [34] quantitatively evaluated agent performance on ToolBench versus live production APIs under controlled API version drift. Their results showed a mean performance drop of 38% in task success rate when agents trained on v1 of a financial data API were tested on v1.1 (with only field name and response format changes), highlighting a significant "realism gap" between benchmark and real-world deployment conditions. Systematically evaluating such adaptability to "unexpected" changes remains a major challenge.

The "Scalability" Dilemma: For tasks requiring deep reasoning and creative data analysis, human expert evaluation is considered the gold standard [28]. However, this method is extremely costly, time-consuming, and difficult to scale, unable to keep pace with agent iteration. Conversely, fully automated evaluation metrics (e.g., code similarity, keyword matching) are efficient but often too coarse to capture the semantic correctness, creativity, or depth of insight in agent solutions. Finding the optimal balance between evaluation "quality" and "efficiency" is a key bottleneck for large-scale, continuous agent evaluation.

The "Privacy" Paradox: Agent functionality relies on access to massive amounts of data, including personal sensitive data. This creates a paradox: to make agents more useful, we must give them more data—but this simultaneously increases privacy risks. Current evaluation methodologies mostly focus on performance and efficiency, with privacy protection not yet systematically integrated into mainstream evaluation frameworks [10] [11]. While technologies like differential privacy and federated learning offer theoretical solutions, how to deploy and verify them in complex agent systems and quantify their privacy protection levels remains an open, interdisciplinary research problem. As of 2025, we do not even have an accepted framework to measure the "privacy leakage risk" caused by agents [11].

4.3. Research Directions for 2025 and Beyond

Facing these challenges, future research should focus on breakthroughs in the following directions.

Developing Composable and Scalable Unified Evaluation Platforms: Future evaluations should not be single, rigid benchmarks, but modular platforms. Researchers should be able to "build" different environments (e.g., simulators, real APIs), tasks (e.g., data queries, report generation), data sources, and interference factors (e.g., noise, missing data) like building blocks. This will greatly enhance evaluation flexibility and coverage, promote community co-construction and share, and gradually establish de facto standards [34] [35].

Researching Process-Oriented Evaluation Metrics: Evaluation focus must shift from "outcome" to "process." New metrics and methods are needed to assess the entire data processing lifecycle. For example, evaluating the efficiency of information retrieval strategies, logical consistency of reasoning chains, handling of data uncertainty, and explainability of decisions. This may require combining log analysis, causal inference, and human-computer interaction techniques [41] [43].

However, process-oriented metrics—particularly those derived from agent logs or trace analysis—are susceptible to validity threats. A key concern is subjectivity in interpretation: different evaluators may assign varying scores to the same reasoning chain based on their domain expertise or cognitive biases. For instance, a log-based "reasoning coherence score" might be interpreted as high by one rater who values creativity, but low by another who prioritizes procedural correctness. This inter-rater variability can undermine metric reliability.

To mitigate this, inter-rater reliability checks should be institutionalized. One effective strategy is to conduct double or triple coding of agent decision logs by independent domain experts, followed by calculation of Cohen's kappa or Krippendorff's alpha to quantify agreement. Discrepancies can then be resolved through consensus meetings. Studies in human-AI interaction have shown that such protocols can increase metric reliability from $\kappa = 0.45$ (moderate) to $\kappa > 0.80$ (almost perfect) [42]. Incorporating such validation steps is essential for ensuring that process metrics are both meaningful and reproducible.

Establishing Standardized Privacy and Security Auditing Protocols: The community urgently needs to collaboratively develop standardized privacy and security auditing standards and procedures for agents. This should include: 1) a standard "attack test case library" to probe agents' privacy leakage risks under various inductions; 2) formal privacy policy description languages and automated verification tools to check agent compliance [36]; 3) quantitative evaluation standards for the trade-off between performance and privacy in agents using privacy-enhancing technologies.

Conducting Long-term, Longitudinal Evaluation Studies: Current evaluations are mostly "one-time" snapshot assessments. Future research needs to examine agent performance over longer time scales, including learning and adaptation capabilities, behavioral drift during continuous environmental interaction, and long-term stability and reliability. This is crucial for agents deployed in critical infrastructure or long-term missions.

5. Conclusions

In 2025, we stand at the dawn of the agent revolution. These autonomous entities powered by large language models hold the potential to unleash immense productivity in science, industry, and daily life. However, this review reveals that our scientific evaluation capabilities for these powerful tools—especially for their core data processing abilities—remain in a challenging, nascent stage.

We have systematically surveyed the state of agent data evaluation. We find that

evaluation paradigms are shifting from static to dynamic, and a series of innovative benchmarks have emerged. Yet the field exhibits a fragmented character, severely lacking unified evaluation standards, accepted metric systems, and mature methodologies. We have deeply analyzed the methodological aspects of datasets, experimental design, and emerging privacy-security considerations, identifying three core challenges: the "realism gap", the "scalability dilemma", and the "privacy paradox". Together, these challenges form a barrier that hinders our comprehensive, objective understanding and trust in the agents we create.

Looking ahead, the path to reliable and trustworthy agents must be built on more scientific, rigorous, and comprehensive evaluation methodologies. The research community must work together to construct standardized evaluation platforms, develop new metrics that illuminate agents' internal decision-making processes, and establish strict privacy and security auditing protocols. Only through continuous evaluation and iteration can we ensure that agent technology advances toward greater capability while also progressing toward greater safety and responsibility—ultimately fulfilling its grand vision of benefiting humanity.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Huang, Q., Vora, J., Liang, P. and Leskovec, J. (2023) Benchmarking Large Language Models as AI Research Agents. arXiv: 2310.03302v1.

[2] Wooldridge, M. and Jennings, N.R. (1995) Intelligent Agents: Theory and Practice. *The Knowledge Engineering Review*, **10**, 115-152. https://doi.org/10.1017/s0269888900008122

[3] Franklin, S. and Graesser, A. (1997) Is It an Agent, or Just a Program? A Taxonomy for Autonomous Agents. In: Müller, J.P., Wooldridge, M.J. and Jennings, N.R., Eds., *Intelligent Agents III Agent Theories, Architectures, and Languages*, Springer, 21-35. https://doi.org/10.1007/bfb0013570

[4] Kim, G.J., Wilf, A., Morency, L.P. and Fried, D. (2025) From Reproduction to Replication: Evaluating Research Agents with Progressive Code Masking. https://github.com/j1mk1m/AutoExperiment

[5] Qiu, R., Chen, S., Su, Y., Yen, P.Y. and Shen, H.W. (2025) Completing a Systematic Review in Hours Instead of Months with Interactive AI Agents. arXiv: 2504.14822. https://arxiv.org/pdf/2504.14822

[6] Cheng, Y., Zhang, C., Zhang, Z., Meng, X., Hong, S., Li, W., Wang, Z., Wang, Z., Yin, F., Zhao, J. and He, X. (2024) Exploring Large Language Model Based Intelligent Agents: Definitions, Methods, and Prospects. arXiv: 2401.03428. https://arxiv.org/pdf/2401.03428

[7] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E.H., Le, Q.V. and Zhou, D. (2022) Chain of Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, **35**, 24824-24836.

[8] Durante, Z., Huang, Q., Wake, N., Gong, R., Park, J.S., Sarkar, B., Taori, R., Noda, Y., Terzopoulos, D., Choi, Y., Ikeuchi, K., Vo, H., Fei-Fei, L. and Gao, J. (2024) Agent

AI: Surveying the Horizons of Multimodal Interaction. arXiv: 2401.03568. https://arxiv.org/pdf/2401.03568

[9] Ignise, A. and Vahi, Y. (2024) Tracking Intelligence and Effectiveness of Agents. *International Journal of Computer Science and Mobile Applications*, **12**, 41-48.

[10] Rani, A., Grover, N., Deepa, N. and Prajitha, C. (2024) A Smart Agent-Based Approach for Privacy Preservation and Threat Mitigation to Enhance Security in the Internet of Medical Things. *Journal of Autonomous Intelligence*, **7**, Article 1629. https://doi.org/10.32629/jai.v7i5.1629

[11] Costa, M., Köpf, B., Kolluri, A., Paverd, A., Russinovich, M., Salem, A., Zanel-la-Béguelin, S., *et al.* (2025) Securing AI Agents with Information-Flow Control. arXiv: 2505.23643.

[12] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., *et al.* (2015) Human-Level Control through Deep Reinforcement Learning. *Nature*, **518**, 529-533. https://doi.org/10.1038/nature14236

[13] Tweedale, J. and Ichalkaranje, N. (2005) Innovations in Intelligent Agents. *Proceedings of the 9th International Conference on Knowledge-Based Intelligent Information and Engineering Systems*, Melbourne, 14-16 September 2005, 821-824. https://doi.org/10.1007/11552451_112

[14] Rao, A.S. and Georgeff, M.P. (1995) BDI Agents: From Theory to Practice. *First International Conference on Multiagent Systems*, San Francisco, 12-14 June 1995, 312-319.

[15] Bansod, P.B. (2025) Distinguishing Autonomous AI Agents from Collaborative Agentic Systems: A Comprehensive Framework for Understanding Modern Intelligent Architectures. arXiv: 2506.01438.

[16] Deng, J., Dong, W., Socher, R., Li, L., Li, K. and Li, F.F. (2009) ImageNet: A Large-Scale Hierarchical Image Database. 2009 *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, 20-25 June 2009, 248-255. https://doi.org/10.1109/cvpr.2009.5206848

[17] Rajpurkar, P., Zhang, J., Lopyrev, K. and Liang, P. (2016) SQuAD: 100,000+ Questions for Machine Comprehension of Text. *Proceedings of the* 2016 *Conference on Empirical Methods in Natural Language Processing*, Austin, November 2016, 2383-2392. https://doi.org/10.18653/v1/d16-1264

[18] Bellemare, M.G., Naddaf, Y., Veness, J. and Bowling, M. (2013) The Arcade Learning Environment: An Evaluation Platform for General Agents. *Journal of Artificial Intelligence Research*, **47**, 253-279. https://doi.org/10.1613/jair.3912

[19] Tambe, M., Johnson, W.L., Jones, R.M., Koss, F., Laird, J.E., Rosenbloom, P.S. and Schwamb, K. (1995) Intelligent Agents for Interactive Simulation Environments. *AI Magazine*, **16**, 15.

[20] Wang, X., Li, D., Zhao, Y. and Wang, H. (2024) MetaTool: Facilitating Large Language Models to Master Tools with Meta-Task Augmentation. arXiv: 2407.12871.

[21] Lin, J., Zhao, H., Zhang, A., Wu, Y., Ping, H. and Chen, Q. (2023) AgentSims: An Open-Source Sandbox for Large Language Model Evaluation. arXiv: 2308.04026.

[22] Winata, G.I., Hudi, F., Irawan, P.A., Anugraha, D., Putri, R.A., Wang, Y., Ngo, C.W., *et al.* (2024) WorldCuisines: A Massive-Scale Benchmark for Multilingual and Multicultural Visual Question Answering on Global Cuisines. arXiv: 2410.12705.

[23] Du, M., Xu, B., Zhu, C., Wang, X. and Mao, Z. (2025) DeepResearch Bench: A Comprehensive Benchmark for Deep Research Agents. arXiv: 2506.11763.

[24] Patel, D., Lin, S., Rayfield, J., Zhou, N., Vaculin, R., Martinez, N., Kalagnanam, J., *et al.* (2025) AssetOpsBench: Benchmarking AI Agents for Task Automation in Industrial Asset Operations and Maintenance. arXiv: 2506.03828.

[25] Abaskohi, A., Ramesh, A.V., Nanisetty, S., Goel, C., Vazquez, D., Pal, C., Laradji, I.H., *et al.* (2025) AgentAda: Skill-Adaptive Data Analytics for Tailored Insight Discovery. arXiv: 2504.07421.

[26] Hu, X., Zhao, Z., Wei, S., Chai, Z., Ma, Q., Wang, G., Wu, F., *et al.* (2024) InfiAgent-DABench: Evaluating Agents on Data Analysis Tasks. arXiv: 2401.05507.

[27] Testini, I., Hernández-Orallo, J. and Pacchiardi, L. (2025) Measuring Data Science Automation: A Survey of Evaluation Tools for AI Assistants and Agents. arXiv: 2506.08800.

[28] Yadav, D., Jain, R., Agrawal, H., Chattopadhyay, P., Singh, T., Jain, A., Batra, D., *et al.* (2019) EvalAI: Towards Better Evaluation Systems for AI Agents. arXiv: 1902.03570.

[29] Elshan, E., Zierau, N., Engel, C., Janson, A. and Leimeister, J.M. (2022) Understanding the Design Elements Affecting User Acceptance of Intelligent Agents: Past, Present and Future. *Information Systems Frontiers*, **24**, 699-730. https://doi.org/10.1007/s10796-021-10230-9

[30] Papineni, K., Roukos, S., Ward, T. and Zhu, W. (2001) BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the* 40*th Annual Meeting on Association for Computational Linguistics—ACL*'02, Philadelphia, 7-12 July 2002, 311-318. https://doi.org/10.3115/1073083.1073135

[31] Abramson, J., Ahuja, A., Carnevale, F., Georgiev, P., Goldin, A., Hung, A., Yan, C., *et al.* (2022) Evaluating Multimodal Interactive Agents. arXiv: 2205.13274.

[32] Hartmann, M. and Koller, A. (2024) A Survey on Complex Tasks for Goal-Directed Interactive Agents. arXiv: 2409.18538.

[33] Moreno, R., Fernandez-Isabel, A., Diego, I.M.D., Moguerza, J.M., Lancho, C. and Teresa, M.C.S. (2022) Automatic Detection of Potential Customers by Opinion Mining and Intelligent Agents. *Proceedings of the* 17*th Conference on Computer Science and Intelligence Systems*, Sofia, 4-7 September 2022, 93-101. https://doi.org/10.15439/2022f131

[34] Chen, K., Ren, Y., Liu, Y., Hu, X., Tian, H., Xie, T., Mo, Z., *et al.* (2025) Xbench: Tracking Agents Productivity Scaling with Profession-Aligned Real-World Evaluations. arXiv: 2506.13651.

[35] Alves, P.H., Correia, F., Frajhof, I., De Souza, C.S. and Lopes, H. (2023) Designing Intelligent Agents in Normative Systems toward Data Regulation Representation. *IEEE Access*, **11**, 51590-51605. https://doi.org/10.1109/access.2023.3276294

[36] Dwork, C. and Roth, A. (2013) The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, **9**, 211-407. https://doi.org/10.1561/0400000042

[37] Dwork, C. (2006) Differential Privacy. In: Bugliesi, M., Preneel, B., Sassone, V. and Wegener, I. Eds., *International Colloquium on Automata, Languages, and Programming*, Springer, 1-12.

[38] Chen, B., Hawkins, C., Karabag, M.O., Neary, C., Hale, M. and Topcu, U. (2023) Differential Privacy in Cooperative Multiagent Planning. *Uncertainty in Artificial Intelligence*, Pittsburgh, 31 July-4 August 2023, 347-357.

[39] McMahan, B., Moore, E., Ramage, D., Hampson, S. and Arcas, B.A. (2017) Communication-Efficient Learning of Deep Networks from Decentralized Data. *Artificial Intelligence and Statistics*, Fort Lauderdale, 20-22 April 2017, 1273-1282.

[40] Pan, J., Zhang, Y., Tomlin, N., Zhou, Y., Levine, S. and Suhr, A. (2024) Autonomous Evaluation and Refinement of Digital Agents. arXiv: 2404.06474.

[41] Yang, Z., Bhatnagar, A., Qiu, Y., Miao, T., Tser Jern Kon, P., Xiao, Y., *et al.* (2025) Cloud Infrastructure Management in the Age of AI Agents. *ACM SIGOPS Operating Systems Review*, **59**, 1-8. https://doi.org/10.1145/3759441.3759443

[42] Chen, C., Zhang, Z., Khalilov, I., Guo, B., Gebreegziabher, S.A., Ye, Y., Li, T.J.J., *et al.* (2025) Toward a Human-Centered Evaluation Framework for Trustworthy LLM-Powered Gui Agents. arXiv: 2504.17934.

[43] Sangaraju, V.R. (2025) A Framework for Secure Data Processing Using AI Agents in Business Intelligence Applications. *Economic Sciences*, **21**, 914-924. https://doi.org/10.69889/srn8qw78