Towards Standardized Evaluation of Large Language Model-Based Agents

Dr.Farheen Mohammed
Assistant professor, Bapatla Engineering College, Bapatla, A.P, India
Email: farheen0122@gmail.com

Abstract

The rise of Large Language Model (LLM)-based agents marks a major shift in artificial intelligence, enabling autonomous systems to plan, reason, use external tools, and retain memory while interacting with dynamic environments. This paper presents a comprehensive survey of evaluation methodologies for such agents. We systematically examine benchmarks and frameworks across four key dimensions: (1) fundamental agent capabilities, including planning, tool use, memory, and self-reflection; (2) application-specific benchmarks for domains such as web navigation, software engineering, scientific discovery, and conversational agents; (3) benchmarks for generalist agents; and (4) frameworks for evaluating agent performance. Our study highlights emerging trends, such as the move toward more realistic and continuously updated evaluation settings, while also identifying critical gaps in assessing safety, robustness, cost-efficiency, and scalability. By mapping the rapidly evolving landscape of LLM-agent evaluation, this survey outlines current limitations and proposes promising directions for future research.

1. Introduction

Recent advances in Large Language Models (LLMs) have dramatically expanded their ability to handle complex tasks across diverse domains. However, traditional LLMs operate as static, single-turn, text-to-text models with limited context handling. In contrast, LLM-based agents extend these capabilities by maintaining shared states across multiple interactions, enabling coherent multi-step reasoning. They can also integrate external tools to perform computations, retrieve knowledge, and interact with real-world environments. This agentic ability empowers them to autonomously design, execute, and adapt complex plans, opening new opportunities across domains such as science, software engineering, and conversational AI.

Reliable evaluation of LLM-based agents is crucial to ensure their effectiveness in real-world applications and to guide their future development. While some evaluation overlaps with traditional LLM benchmarks, agents introduce unique challenges due to their sequential operation, dynamic environments, and broad applicability. As a result, new evaluation methodologies, benchmarks, and metrics are required to capture their true capabilities. This paper addresses these challenges by surveying the existing evaluation landscape, identifying gaps, and highlighting opportunities for advancing the field

Scope

This survey specifically focuses on evaluation methodologies for LLM-based agents. As such, widely adopted single-call LLM benchmarks—such as MMLU, AlpacaEval, GSM8K, and similar standardized datasets—are not discussed in depth. Likewise, detailed introductions to LLM agents, their architectures, and modeling choices fall outside the scope of this work, as they have been thoroughly covered in prior surveys (e.g., Wang et al., 2024a). While related areas such as multi-agent interactions, game-playing agents, and embodied agents are briefly mentioned, they are not the central emphasis of this survey. Instead, our objective is to provide a structured and comprehensive overview of evaluation methods tailored to LLM-based agents.

2. Evaluation of Agent Capabilities

LLM-based agents are built upon design patterns that leverage a core set of fundamental LLM abilities. Evaluating these foundational capabilities is essential for understanding both their strengths and limitations. In this section, we focus on four primary dimensions of agent capabilities: **planning**, **tool use**, **self-reflection**,

and **memory**. Together, these serve as the building blocks for assessing the effectiveness and reliability of LLM-based agents in dynamic environments.

Self-Reflection

An emerging research direction investigates the ability of agents to **self-reflect** and enhance their reasoning through iterative feedback. The goal is to reduce errors in multi-step interactions by enabling agents to critically assess their own outputs. Effective self-reflection requires the model to not only interpret external or internal feedback but also to dynamically revise its intermediate beliefs, plans, or reasoning steps. By adapting its behavior over extended task trajectories, a self-reflective agent can achieve more reliable performance in complex and evolving environments.

Memory

Memory mechanisms play a crucial role in enhancing the performance of LLM-based agents by enabling them to retain context, retrieve information, and reason effectively in dynamic scenarios (Park et al., 2023). Unlike tool use, which connects agents to external resources, memory supports **context preservation** across extended interactions, such as document processing or long-term conversations.

LLM agents typically employ two complementary forms of memory: **short-term memory**, which facilitates real-time responses by maintaining recent context, and **long-term memory**, which allows the agent to accumulate knowledge, build deeper understanding, and reuse past experiences over time. Together, these mechanisms empower agents to adapt, learn, and make well-informed decisions in tasks that require persistent access to prior informatio

3. Application-Specific Agents Evaluation

The landscape of application-specific LLM-based agents is expanding rapidly, with specialized systems emerging across domains such as tools, web, software, gaming, embodied environments, and scientific discovery (Wang et al., 2024a). In this section, we highlight four representative categories that demonstrate the diversity and potential of these agents, while examining the evaluation frameworks and performance metrics tailored to their respective applications.

Benchmarks for application-specific agents generally incorporate three core components. **First**, they define a dataset of tasks—ranging from website navigation to complex scientific problem-solving—that specifies the intended goals of the agent. **Second**, they establish an operating environment, which may be simulated (static or dynamic) or real-world, often integrating user simulations, external tools, or domain-specific constraints. **Third**, they employ evaluation metrics, such as **success rate, efficiency, and accuracy**, applied at varying levels of granularity. These may include step-by-step action tracking, milestone completion, or holistic end-to-end task evaluation.

Web Agents

Web agents are designed to interact with online platforms to perform tasks such as booking flights, making purchases, or retrieving structured information. Their evaluation focuses on the agent's ability to **successfully complete tasks, navigate complex web environments, and comply with safety, privacy, and ethical guidelines**.

As these systems have advanced, so too have the benchmarks developed to assess them. Early evaluations focused on simple, rule-based tasks, whereas recent benchmarks capture a broader range of **realistic, dynamic, and multi-step interactions**, better reflecting the challenges of real-world web navigation and automation.

4. Generalist Agents Evaluation

Building on the evaluation of fundamental capabilities and application-specific agents, we now consider **general-purpose agents**. As LLMs have evolved from task-specific systems to versatile, general-purpose models, LLM-based agents are likewise transitioning beyond narrowly defined applications toward broader, more adaptable functionalities. These agents combine core LLM abilities with advanced skills such as web navigation, information retrieval, and code execution to solve complex, multi-domain challenges. This shift necessitates comprehensive benchmarks capable of assessing a wide spectrum of agentic capabilities.

A major category of generalist benchmarks focuses on **multi-step reasoning, interactive problem-solving, and proficient tool use**. For instance, **GAIA** (Mialon et al., 2023) provides 466 human-authored, real-world questions that test reasoning, multimodal understanding, web navigation, and general tool integration. Similarly, **Galileo's Agent Leaderboard** (Galileo, 2025) evaluates function calls and API usage in practical applications such as database queries, online calculators, and web services. **AgentBench** (Liu et al., 2023a) introduces a diverse suite of interactive environments spanning operating system commands, SQL databases, digital games, and household tasks. Collectively, these benchmarks emphasize the flexibility, reasoning depth, and adaptive tool use required for effective general agents.

Beyond reasoning and tool integration, another critical evaluation dimension concerns an agent's ability to function within **full-scale computer operating environments**. Benchmarks such as **OSWorld** (Xie et al., 2024), **OmniACT** (Kapoor et al., 2024a), and **AppWorld** (Trivedi et al., 2024) test whether agents can navigate computer systems, execute multi-application tasks, and manage workflows across diverse software environments. These benchmarks often require agents to generate and modify code, handle complex control flows, and ensure reliable execution while minimizing unintended system disruptions.

5. Conclusion

The evaluation of LLM-based agents is a rapidly evolving research area, motivated by the growing complexity and autonomy of these systems. Considerable progress has been achieved in developing benchmarks that are more realistic, dynamic, and representative of real-world challenges. However, several critical gaps persist. In particular, current methodologies often fall short in addressing **safety, fine-grained performance analysis, cost-efficiency, and scalability**. Bridging these gaps will be essential for ensuring that LLM-based agents are developed and deployed responsibly, with robustness and reliability as core design principles.

Looking ahead, advancing evaluation practices will require interdisciplinary approaches, standardized frameworks, and continuous updates to benchmarks that reflect real-world dynamics. By addressing these challenges, the research community can better support the effective integration of LLM-based agents into practical applications, unlocking their full potential while safeguarding against risks.

References

- Toufique Ahmed, Martin Hirzel, Rangeet Pan, Avraham Shinnar, and Saurabh Sinha. 2024. **TDD-bench verified: Can LLMs generate tests for issues before they get resolved**
- Reem Aleithan, Haoran Xue, Mohammad Mahdi Mohajer, Elijah Nnorom, Gias Uddin, and Song Wang. 2024. **SWE-bench+: Enhanced coding benchmark for LLMs.** *arXiv preprint*
- Jacob Andreas, John Bufe, David Burkett, Charles Chen, Josh Clausman, Jean Crawford, Kate Crim, Jordan DeLoach, Leah Dorner, Jason Eisner, Hao Fang, Alan Guo, David Hall, Kristin Hayes, Kellie Hill, Diana Ho, Wendy Iwaszuk, Smriti Jha, Dan Klein, Jayant Krishnamurthy, Theo Lanman, Percy Liang, Christopher H. Lin, Ilya Lintsbakh, Andy McGovern, Aleksandr Nisnevich, Adam Pauls, Dmitrij Petters, Brent Read, Dan Roth, Subhro Roy, Jesse Rusak, Beth Short, Div Slomin, Ben Snyder, Stephon Striplin, Yu Su, Zachary Tellman, Sam Thomson, Andrei Vorobev, Izabela Witoszko, Jason Wolfe, Abby Wray, Yuchen

Zhang, and Alexander Zotov. 2020. **Task-oriented dialogue as dataflow synthesis.** *Transactions of the Association for Computational Linguistics*

- Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, Eric Winsor, Jerome Wynne, Yarin Gal, and Xander Davies. 2024. **AgentHarm: A benchmark for measuring harmfulness of LLM agents.**
- Samuel Arcadinho, David Oliveira Aparicio, and Mariana S. C. Almeida. 2024. **Automated test generation to evaluate tool-augmented LLMs as conversational AI agents.** In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 54–68. Miami, Florida, USA. Association for Computational Linguistics.
- Arize AI, Inc. 2025. **Agent evaluation.**
- Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie J. Cai, Michael Terry, Quoc V. Le, and Charles Sutton. 2021. **Program synthesis with large language models.**
- Kinjal Basu, Ibrahim Abdelaziz, Subhajit Chaudhury, Soham Dan, Maxwell Crouse, Asim Munawar, Sadhana Kumaravel, Vinod Muthusamy, Pavan Kapanipathi, and Luis A. Lastras. 2024a. **API-blend: A comprehensive corpora for training and benchmarking API LLMs.**
- Kinjal Basu, Ibrahim Abdelaziz, Kiran Kate, Mayank Agarwal, Maxwell Crouse, Yara Rizk, Kelsey Bradford, Asim Munawar, Sadhana Kumaravel, Saurabh Goyal, et al. 2024b. **Nestful: A benchmark for evaluating LLMs on nested sequences of API calls.**
- Pratik Bhavsar. 2025. **Agent Leaderboard.** Available at: Hugging Face Spaces.
- Daniel Gomez Blanco. 2023. **Practical Open Telemetry.** Springer.