Evaluation and Benchmarking of Generative and Agentic AI Systems: A Comprehensive Survey

Manish Shukla
Independent Researcher, TX, USA
manishshukla.ms18@gmail.com

December 15, 2025

Abstract

The rapid emergence of generative and agentic artificial intelligence (AI) has outpaced traditional evaluation practices. While large language models excel on static language benchmarks, real-world deployment demands more than accuracy on curated tasks. Agentic systems use planning, tool invocation, memory and multi-agent collaboration to perform complex workflows. Enterprise adoption therefore hinges on holistic assessments that include cost, latency, reliability, safety and multi-agent coordination. This survey provides a comprehensive taxonomy of evaluation dimensions, reviews existing benchmarks for generative and agentic systems, identifies gaps between laboratory tests and production requirements, and proposes future directions for more realistic, multi-dimensional benchmarking.

Keywords: Agentic AI, Generative AI, AI Evaluation and Benchmarking, Autonomous Intelligent Systems, Tool-Augmented Language Models, Multi-Agent Systems, Long-Horizon Planning, Memory and Reasoning Evaluation, Adaptive Monitoring, AI Safety and Robustness, Multimodal AI Evaluation

1 Introduction

Large language models (LLMs) such as GPT-4/5, Claude, Gemini and Llama have achieved remarkable success in natural language understanding, reasoning and code generation. These capabilities have enabled a new class of *agentic AI systems* that plan tasks, invoke tools, maintain memory and collaborate with other agents to achieve goals autonomously. The shift from static text generation to autonomous action raises new evaluation challenges. Traditional metrics such as BLEU, ROUGE or exact match focus on output quality but ignore the agent's internal reasoning process, tool usage or adherence to user goals. Safety, cost and latency become critical when agents interact with real environments. The 2025 Stanford AI Index reports that AI performance on demanding benchmarks such as MMMU, GPQA and SWE-bench improved sharply between 2023 and 2024, with scores rising by 18.8, 48.9 and 67.3 percentage points respectively. At the same time, researchers introduced new safety benchmarks such as HELM Safety, AIR-Bench and FACTS to assess factuality and robustness. These developments signal a maturing landscape of evaluation but also highlight gaps: complex reasoning remains challenging and there is little agreement on how to assess agentic behaviour.

This survey aims to unify the fragmented evaluation landscape. We define a taxonomy of dimensions relevant to generative and agentic AI, review existing benchmarks in each category, discuss evaluation methodologies and limitations, and outline future research directions.

2 Taxonomy of Evaluation Dimensions

We classify evaluation into five broad dimensions: output-level, behavioural, interaction-level, multi-agent and safety/performance. Each dimension captures different aspects of generative and agentic systems (Figure 1).

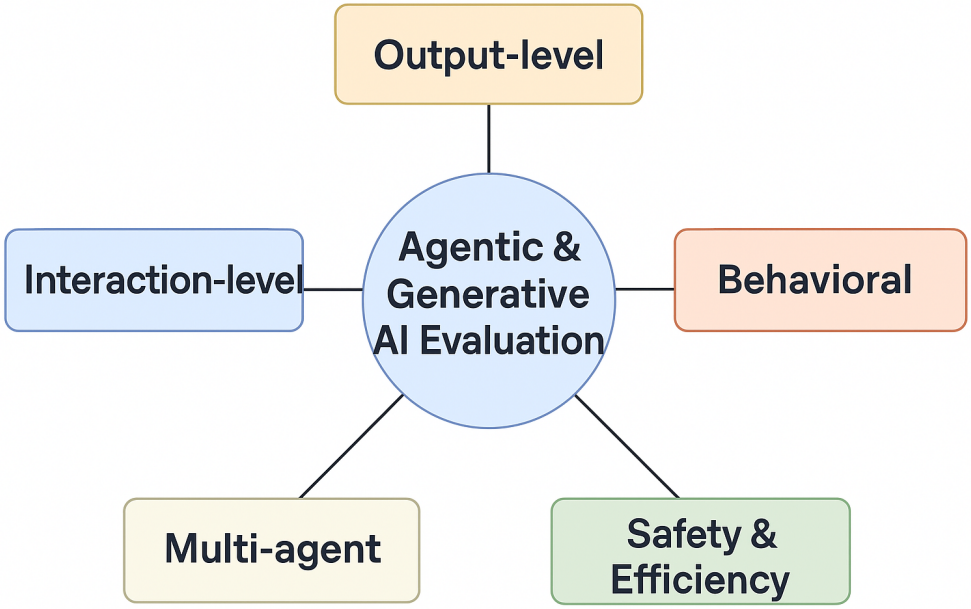

Figure 1: High-level evaluation dimensions for generative and agentic AI. The diagram illustrates five axes—output-level, behavioural, interaction-level, multi-agent and safety/performance—which together capture the multifaceted nature of modern AI evaluation.

2.1 Output-Level Evaluation

Generative models are commonly assessed on the quality of their final outputs. Typical criteria include fluency, coherence, factuality, relevance, completeness and creativity. Metrics such as BLEU, ROUGE, METEOR and BERTScore quantify surface similarity to reference texts. More recently, model-based evaluators like G-Eval or GPT-as-a-judge have been adopted to judge free-form responses. Output-level evaluation is suitable for tasks where a single best answer exists (e.g., translation, summarization) but fails to capture the process or decision-making behind an agent's behaviour.

2.2 Behavioural Evaluation

Agentic systems require evaluation of *how* tasks are accomplished. Key aspects include planning depth, number of steps to completion, tool selection accuracy, error handling, memory retrieval correctness and compliance with constraints. Benchmark metrics such as task success rate, step efficiency and tool-call correctness measure these properties. The CLEAR framework proposed by Mehta *et al.* introduces five dimensions—Cost, Latency, Efficacy, Assurance and Reliability—to assess enterprise agents holistically. Their analysis shows that optimizing solely for accuracy can yield agents that are 4.4–10.8× more expensive than cost-aware alternatives.
Note. In practice, enterprise deployments should adopt the CLEAR framework to balance capability metrics against operational costs [4].

2.3 Interaction-Level Evaluation

Many tasks require multi-turn interactions where the agent must maintain context across turns. Evaluations therefore measure dialogue state tracking, knowledge retention, response consistency and user satisfaction. Long-horizon consistency benchmarks test whether models maintain coherent world states over extended conversations.

2.4 Multi-Agent Evaluation

In collaborative settings, multiple agents coordinate to complete tasks. Metrics assess coordination success, communication efficiency, role fidelity and conflict resolution. Multi-agent frameworks like ChatDev demonstrate the potential of specialized agents (architect, programmer, tester) communicating via language to design, code and test software.

2.5 Safety and Robustness

Deployment requires evaluation of safety, fairness, robustness and compliance. Adversarial prompts, jailbreak attempts, biased data and harmful tool misuse must be tested. The AI Index notes that benchmarks like HELM Safety and AIR-Bench are emerging to assess factuality and safety. Cost and latency also affect usability; web-interactive agents spend 53.7

3 Benchmarks for Generative AI

Table 1 summarizes prominent benchmarks for generative models.

Category	Example Benchmarks	Key Metrics
Reasoning	MMLU, GSM8K, BigBench/BBH, ARC Challenge	Subject accuracy, reasoning depth
Language Understanding	GLUE, SuperGLUE, SQuAD	Accuracy, F1 score
Summarization	XSum, CNN/Daily Mail	ROUGE, BERTScore
Multimodal	MMMU, SEED, MathVista, ChartQA	Accuracy across modalities
Coding	HumanEval, MBPP, RepoBench	Pass@k, functional correctness
Safety	HELM Safety, AIR-Bench, FACTS	Toxicity rates, factuality

Table 1: Representative benchmarks for generative AI. Safety benchmarks measure harmful content and factuality.

Reasoning tasks test mathematical and commonsense reasoning. Language understanding benchmarks measure comprehension and inference in natural language. Multimodal datasets assess a model's ability to integrate text, images and charts. Coding benchmarks evaluate program synthesis and problem solving. Safety benchmarks quantify biases and harmful output. These tasks primarily evaluate final outputs; they provide limited insight into tool usage, planning or process fidelity.

4 Benchmarks for Agentic AI

Agentic evaluation is still nascent. We review categories for tool use, planning and long-horizon tasks, memory, multi-agent collaboration and self-evaluation.

4.1 Tool-Use Benchmarks

Tools extend LLMs with APIs such as calculators, web browsers and code execution. Benchmarks like ToolBench, API-Bank and BrowserBench test whether agents select appropriate tools and provide correct arguments. Metrics include tool selection accuracy and dependency chain quality. StableToolBench builds on these efforts by providing a virtual API server and caching mechanism that enable reproducible tool-use evaluation, along with pass and win rate metrics measured via LLM-based judges [8].

4.2 Planning and Long-Horizon Tasks

Planning benchmarks evaluate the agent's ability to decompose tasks into steps and execute them. ALFWorld and MiniWoB++ simulate household and GUI tasks. WebArena offers 812 tasks on real websites with functional correctness evaluation. PlanBench, introduced in 2023, uses classical automated-planning domains

to test reasoning about actions and change. The 2025 AI Index notes that LLMs still struggle with complex reasoning tasks like PlanBench. Early embodied environments such as ALFWorld and MiniWoB++ provide interactive settings for manipulation and interface control [11]. The WebArena benchmark further increases task diversity with hundreds of real-website scenarios [12]. PlanBench extends these efforts with formal planning domains drawn from the International Planning Competition and demonstrates that LLMs still fall short in generating valid plans [7]. Long-horizon benchmarks include embodied environments such as Habitat and BEHAVIOR. Success metrics consider completion rate, number of actions and plan optimality.

4.3 Memory Benchmarks

Memory enables agents to accumulate knowledge across interactions. Tan *et al.* propose MemBench, a dataset with factual and reflective memory tasks across participation and observation scenarios. The benchmark evaluates effectiveness, efficiency and capacity, offering metrics such as accuracy, recall and temporal efficiency. Their analysis underscores the need for multi-scenario and multi-level evaluation to capture both explicit and implicit memory. MemBench is publicly available and has quickly become a standard for evaluating memory modules in LLM agents [5].

4.4 Multi-Agent Benchmarks

Multi-agent benchmarks test collaboration among specialized agents. AgentBench evaluates LLM-as-agent across eight distinct environments (operating system, database, knowledge graph, games and web tasks). Tests over 27 API-based and open-source LLMs reveal large performance disparities; top commercial models outperform open-source competitors, and failures often stem from poor long-term reasoning and decision-making. ChatDev demonstrates a chat-powered software development framework where agents communicate via language to design, code and test software, showing the potential of language as a unifying bridge.
These findings highlight the importance of multi-environment evaluation suites such as AgentBench and collaborative platforms like ChatDev [6, 9].

4.5 Self-Evaluation and Meta-Cognition

Self-evaluation involves agents assessing their own outputs, detecting errors and refining plans. Benchmarks such as SelfCheckGPT and MetaQA test confidence estimation and error correction. Reflexion tasks evaluate the ability of agents to introspect and improve their performance over iterations. Recent work emphasises that self-reflection is more than error detection: AutoMeco proposes intrinsic meta-cognition metrics and a Markovian Intrinsic Reward Adjustment (MIRA) algorithm to encourage models to evaluate and adjust their reasoning processes [13]. StoryBench introduces dynamic narrative environments where agents must retain and reason over long-term context across interactive fiction tasks, exposing limitations in memory retention [14]. Reflection-Bench examines epistemic agency across seven cognitive dimensions—including prediction, decision-making, perception, memory, counterfactual thinking, belief updating and meta-reflection—and reveals that state-of-the-art models still struggle with meta-reflection and global pattern recognition [15]. Hallucination detection techniques like SelfCheckGPT rely on sampling multiple candidate responses to identify non-factual content [16].

5 Evaluation Methodologies

5.1 Human Evaluation

Human judgments remain the gold standard. Evaluators score outputs or behaviour based on accuracy, usefulness, trustworthiness, safety and efficiency. Methods include pairwise comparisons, Likert scales and rubric-based assessment. Human evaluation is critical for open-ended tasks, safety assessments and user satisfaction studies but is costly and subject to annotator bias.

5.2 Automated Evaluation

Automated evaluators provide scalable assessment. These include rule-based programs, unit tests (for code) and large models acting as judges. The Azure AI Evaluation library introduces metrics for Task Adherence, Tool Call Accuracy and Intent Resolution, capturing whether an agent answers the right question, uses tools correctly and understands user goals.

5.3 Simulation-Based Evaluation

Simulators provide safe environments to test agentic behaviour. Web simulators, game environments and API sandboxes allow controlled evaluation of tool use, planning and collaboration. AgentBench provides eight environments for such evaluation.

5.4 Real-World Evaluation

Ultimately, agents must be tested in real deployments. Real-world evaluation measures action correctness, error recovery, cost, latency and user satisfaction. Enterprises often find a mismatch between benchmark success and production reliability. Mehta *et al.* show that leading agents exhibit up to 50× cost variation for similar accuracy and that reliability drops dramatically when measured across multiple runs.

6 Challenges and Gaps

Despite rapid progress, several challenges remain:

1. **Limited Realism.** Many benchmarks use synthetic tasks that do not reflect complex enterprise workflows. Realistic multi-step tasks with noisy data, ambiguous instructions and external dependencies are rare.

2. **Environment Diversity.** Few benchmarks cover multiple domains or multi-agent settings. AgentBench is a notable exception but still limited in scope.

3. **Process-Level Evaluation.** Existing metrics focus on final outcomes. There is no consensus on how to evaluate planning traces, decision chains or intermediate tool calls. Pass@k metrics reveal reliability issues but do not explain failure modes.

4. **Safety and Alignment.** Safety benchmarks often measure text toxicity but neglect tool misuse, prompt injection attacks and emergent behaviours. Cost, latency and policy compliance are seldom reported.

5. **Standardization.** Benchmark definitions and scoring methods vary widely. Lack of standard metrics hinders comparison across studies.

7 Future Directions

To address the identified gaps, we suggest the following directions:

1. **Multi-Domain Agentic Benchmark Suites.** Develop unified benchmark suites spanning web tasks, embodied environments, enterprise workflows, knowledge-intensive tasks and cooperative games. Such suites should provide diverse tasks with clear metrics.

2. **Continuous and Long-Term Evaluation.** Assess agents over days or weeks to capture robustness to environment drift, memory retention and evolving instructions.

3. **Safety Stress Testing.** Incorporate adversarial prompts, jailbreak attempts, harmful tool scenarios and policy compliance checks. Evaluate both textual and action-level safety.

4. **Multi-Agent Ecosystem Simulations.** Create large-scale simulations of economies, supply chains or social systems to study emergent collaboration and competition. Metrics should evaluate coordination efficiency, fairness and emergent behaviours.

5. **Self-Evaluation Benchmarks.** Design tasks that require agents to estimate confidence, detect errors and correct themselves. Investigate how self-reflection improves reliability and safety.

6. **Enterprise-Oriented Metrics.** Adopt frameworks like CLEAR that integrate cost, latency, reliability, assurance and compliance. Provide tools to measure these metrics on real workloads.

8 Balanced Framework and Adaptive Monitoring

Recent research emphasises that evaluating agentic AI requires not only technical accuracy but also holistic assessment across multiple axes. Shukla proposes a balanced framework that spans five dimensions—capability and efficiency, robustness and adaptability, safety and ethics, human-centred interaction, and economic and sustainability—as a means to capture the full impact of agentic systems in high-stakes domains [1]. The framework introduces indicators such as goal-drift (how far an agent's behaviour deviates from its intended objectives over time) and harm-reduction (the agent's ability to avoid harmful outcomes), and underscores that current evaluations often neglect these sociotechnical aspects. Case studies suggest that agentic systems can achieve 20–60 % productivity gains but rarely assess fairness, trust or sustainability, motivating a shift to multidimensional evaluation.

Building on the balanced framework, Shukla develops an *Adaptive Multi-Dimensional Monitoring* (AMDM) algorithm to dynamically aggregate heterogeneous metrics and detect anomalies [2]. AMDM normalises disparate metrics using exponentially weighted moving averages, applies Mahalanobis distance to measure deviations from expected behaviour and triggers alerts when patterns exceed learned thresholds. Experiments demonstrate that AMDM reduces anomaly detection latency from 12.3 s to 5.6 s and lowers false-positive rates from 4.5 % to 0.9 %, enabling more responsive monitoring. Shukla also highlights a pervasive imbalance in evaluation reporting: roughly 83 % of studies emphasise capability and efficiency metrics, while only 30 % consider human-centred or economic dimensions. Addressing these gaps requires adaptive monitoring frameworks that combine automated metrics, human-in-the-loop scoring and economic analysis to ensure responsible adoption of agentic AI.

Beyond these frameworks, multi-metric safety initiatives such as HELM emphasise comprehensive evaluation across accuracy, calibration, robustness, fairness, bias, toxicity and efficiency [10]. Incorporating such multidimensional safety metrics into enterprise deployments will help bridge the gap between laboratory benchmarks and real-world expectations. In a related vein, the emerging field of agentic sign language translation integrates multimodal data acquisition, spatio-temporal sign recognition, LLM-based translation, generative sign synthesis and adaptive monitoring to deliver end-to-end communication for deaf and hard-of-hearing users [3]. This inclusive application underscores the need for evaluation frameworks that support diverse modalities, continuous monitoring and human-centred accessibility.

9 Conclusion

Generative and agentic AI systems hold immense promise but pose unique evaluation challenges. Benchmarks have evolved from static language tasks to multi-environment agentic tests, yet they still overlook critical aspects such as cost, latency, reliability and safety. Holistic evaluation frameworks like CLEAR and multi-environment suites like AgentBench highlight the path forward. To ensure trustworthy deployment, researchers and practitioners must develop benchmarks that reflect real-world complexity, integrate process-level metrics, assess multi-agent collaboration and prioritize safety. Only through rigorous, multidimensional evaluation can agentic AI systems transition from experimental demonstrations to reliable partners in complex workflows.

References

[1] Shukla, M. A. (2025). *Evaluating Agentic AI Systems: A Balanced Framework for Performance, Robustness, Safety and Beyond*. TechRxiv. Available at `https://doi.org/10.36227/techrxiv.175693283.32347108/v1`.

[2] Shukla, M. (2025). *Adaptive Monitoring and Real-World Evaluation of Agentic AI Systems*. arXiv preprint arXiv:2509.00115. Available at `https://arxiv.org/abs/2509.00115`.

[3] Shukla, M., & Yemi Reddy, J. (2025). *Agentic Sign Language: Balanced Evaluation and Adaptive Monitoring for Inclusive Multimodal Communication*. Preprints.org. DOI:10.20944/preprints202512.0690.v1.

[4] Mehta, S., et al. (2025). *CLEAR: A Holistic Evaluation Framework for Enterprise Agentic AI*. arXiv preprint arXiv:2511.14136.

[5] Tan, X., et al. (2025). *MemBench: Evaluating Memory in Large Language Model Agents*. arXiv preprint arXiv:2506.21605.

[6] Liu, B., et al. (2023). *AgentBench: Benchmarking LLMs as Agents*. arXiv preprint arXiv:2308.03688.

[7] Valmeekam, K., et al. (2022). *PlanBench: A Benchmark for Reasoning about Change Using LLMs*. arXiv preprint arXiv:2206.10498.

[8] Guo, D., et al. (2025). *StableToolBench: Enabling Stable Tool Learning for Large Language Models*. arXiv preprint arXiv:2403.07714.

[9] Qian, J., et al. (2023). *ChatDev: Collaborating with Large Language Model Agents for Software Development*. arXiv preprint arXiv:2307.07924.

[10] Liang, P., et al. (2022). *Holistic Evaluation of Language Models*. arXiv preprint arXiv:2211.09110.

[11] Shridhar, M., et al. (2020). *ALFWorld: Aligning Text and Embodied Environments for Interactive Instruction Following*. arXiv preprint arXiv:2010.03768.

[12] Yao, A., et al. (2023). *WebArena: A Realistic Web Environment for Building Autonomous Agents*. arXiv preprint arXiv:2307.09288.

[13] Ma, W., et al. (2025). *AutoMeco: Evaluating and Improving Intrinsic Meta-Cognition in Large Language Models*. arXiv preprint arXiv:2506.08410.

[14] Wan, Y., & Ma, W. (2025). *StoryBench: A Dynamic Benchmark for Long-Term Memory Reasoning*. arXiv preprint arXiv:2506.13356.

[15] Anonymous. (2025). *Reflection-Bench: Evaluating Epistemic Agency in Large Language Models*. Submitted manuscript.

[16] Manakul, P., et al. (2023). *SelfCheckGPT: Zero-Resource Hallucination Detection for Generative Language Models*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2023).

[17] Hendrycks, D., et al. (2020). *Measuring Massive Multitask Language Understanding*. arXiv preprint arXiv:2009.03300.

[18] Cobbe, K., et al. (2021). *Training Verifiers to Solve Math Word Problems*. arXiv preprint arXiv:2110.14168.

[19] Srivastava, A., et al. (2023). *Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models*. Proceedings of the AAAI Conference on Artificial Intelligence, 37(13), 12474–12485.