Evaluating Agentic AI Systems: A Balanced Framework for Performance, Robustness, Safety and Beyond

Manish A. Shukla
Independent Researcher, Plano, Texas, USA
manishshukla.ms18@gmail.com

August 18, 2025

Abstract

Agentic artificial intelligence (AI)—multi-agent systems that combine large language models with external tools and autonomous planning—are rapidly transitioning from research labs into high-stakes domains. Existing evaluations emphasise narrow technical metrics such as task success or latency, leaving important sociotechnical dimensions like human trust, ethical compliance and economic sustainability under-measured. We propose a balanced evaluation framework spanning five axes (capability&efficiency, robustness&adaptability, safety&ethics, human-centred interaction and economic&sustainability) and introduce novel indicators including goal-drift scores and harm-reduction indices. Beyond synthesising prior work, we identify gaps in current benchmarks, develop a conceptual diagram to visualise interdependencies and outline experimental protocols for empirically validating the framework. Case studies from recent industry deployments illustrate that agentic AI can yield 20–60 % productivity gains yet often omit assessments of fairness, trust and long-term sustainability. We argue that multidimensional evaluation—combining automated metrics with human-in-the-loop scoring and economic analysis—is essential for responsible adoption of agentic AI.

Keywords : Agentic AI, Multi-Agent Systems, Evaluation Framework, Goal-Drift, Ethical AI, AI Governance

1 Introduction

Large language models (LLMs) have enabled a new class of AI agents that can parse natural-language instructions, call external tools and carry out multistep tasks. Over the past few years these agents have evolved into *agentic* AI systems that coordinate multiple specialised agents to achieve high-level goals through planning, memory and delegation[1]. A recent taxonomy distinguishes between generative AI, single-agent systems and agentic AI: generative models produce prompt–response interactions; single-agent systems incorporate tool calls; agentic AI adds autonomous goal-setting, multi-agent collaboration and persistent memory. While generative systems are evaluated by accuracy and latency, agentic AI requires assessing long-horizon reasoning, inter-agent coordination and human impacts.

However, existing benchmarks such as MMLU and HELM focus on static question-answering and simple tool invocation. They fail to capture long-horizon tasks, interaction loops or tool orchestration. Industry reports proclaim that agentic AI will deliver double-digit productivity gains and multi-trillion-dollar economic potential, yet the validity of these claims is contentious[2]. A systematic review of 84 studies from 2023–2025 found that 83 % of evaluations focused on technical performance, while only 30 % considered human-centred factors and 30 %

considered economic impacts[2]. This imbalance produces a disconnect between benchmark success and real-world value; deployments that excel on technical metrics may fail due to poor trust, workflow integration or ethical risks.

The goal of this paper is twofold: (1) to synthesise existing evaluation practices for agentic AI and identify their strengths and limitations, and (2) to propose a balanced evaluation framework that integrates performance, robustness, safety, human factors and economic sustainability. We also outline an empirical evaluation protocol and apply the framework to real-world case studies drawn from the banking and market-research domains. Our contributions include: a five-axis evaluation framework; novel metrics such as goal-drift and harm-reduction scores; a conceptual diagram illustrating interdependencies; and a discussion of experimental design and governance implications.

2 Related Work

2.1 Existing Evaluation Metrics

Accuracy and effectiveness. Accuracy measures how often an agent produces the correct output or makes the right decision[3]. Effectiveness assesses whether the agent achieves its goals in a specific context, often measured by task-completion rates, adaptive task evaluations or sequence matching. These metrics treat tasks as isolated functions rather than sequences embedded in dynamic environments and are therefore necessary but insufficient.

Efficiency and scalability. Efficiency gauges how well an agent uses resources such as time and computing power, while scalability measures its ability to handle increasing workload or complexity[3, 4]. Metrics include latency, throughput, cost per interaction (e.g. token usage) and success rate over time. These are relevant for organisations seeking to scale agentic systems but must be balanced against robustness and sustainability.

Output quality and hallucination. Beyond correctness, evaluators examine the relevance, coherence and fluency of generated content. For LLM-powered agents, hallucination rate and groundedness are critical: hallucination counts invented facts, while groundedness assesses whether responses are based on verifiable sources[4]. Retrieval-augmented generation and factual-consistency evaluation help mitigate hallucinations.

Robustness and reliability. Robustness captures an agent's ability to operate under varying conditions such as noisy inputs or adversarial attacks; reliability refers to consistent performance over time and across repetitions. Metrics include consistency scores, error rate, resilience to adversarial examples and recovery from failures[3]. Evaluating robustness is especially important for autonomous systems interacting with unpredictable environments.

Safety, ethics and fairness. Safety metrics detect harmful outputs, toxic language and security vulnerabilities. Bias detection measures identify unfair treatment across demographic groups, while fairness metrics assess equitable outcomes[4]. Ethical evaluation also considers transparency, accountability and compliance with legal standards. The AIMultiple survey emphasises that automated scores must be complemented with structured human evaluations and custom tests for bias, fairness and toxicity to assess both quantitative performance and qualitative risks[8].

User experience and human factors. User satisfaction and trust are critical for adoption. Metrics such as customer satisfaction (CSAT) or net promoter score (NPS) evaluate how end-users perceive agentic systems. Human-in-the-loop evaluations, where human reviewers judge tone, coherence or creativity, complement automated scoring[3]. Instruments like the TrAAIT model measure clinicians' trust in AI based on information credibility, perceived application value and reliability; the model provides a dashboard to identify barriers to adoption and can help organisations implementing AI intercept trust issues[6].

2.2 Existing Frameworks and Benchmarks

Multiple benchmarks have been proposed for agentic AI. MLAgentBench, ML-Bench and SUPER evaluate task success, efficiency and end-to-end execution using predefined scripts or repository-grounded tasks. PlanBench introduces symbolic validation for plan structure, and VisualWebArena tests multimodal agents in web environments[1]. Commercial platforms such as Galileo, QAwerk and Orq.ai provide dashboards and tests for performance, hallucination and bias. However, these frameworks primarily measure technical competence; they rarely assess integration into human workflows or long-term sustainability.

In response to the limitations of technical benchmarks, HCI and social-computing researchers have proposed instruments for evaluating trust, usability and alignment[2]. Examples include the TrAAIT survey for clinician trust and guidelines emphasising transparency and observability. AIMultiple's review of LLM evaluation stresses the need for multidimensional strategies that integrate automated scores, structured human evaluations and custom tests for fairness, toxicity and bias[8]. Nevertheless, these approaches remain fragmented and are not yet widely incorporated into industry practice. Our work aims to bridge this gap by integrating technical and sociotechnical dimensions into a unified framework and providing guidance for empirical evaluation.

3 A Balanced Evaluation Framework for Agentic AI

3.1 Framework Overview

To address the measurement imbalance we propose a five-axis evaluation framework (Figure 1). Each axis represents a set of metrics that capture different aspects of agent performance, environment resilience and societal impact. We emphasise that these axes are interdependent; improvements in one dimension (e.g. efficiency) may come at the expense of another (e.g. safety). Following AIMultiple's recommendation for multidimensional evaluation[8], a comprehensive assessment should measure all axes and analyse their interactions.

- **Capability&Efficiency** — assesses whether the agent accomplishes its tasks effectively and efficiently. Core metrics include task-completion rate, latency, throughput, resource utilisation and cost per interaction[3, 4].

- **Robustness&Adaptability** — measures the agent's resilience to changing conditions, adversarial inputs and unexpected events. Metrics include success rate under noisy inputs, recovery time from failures, ability to adapt to new goals and resilience to adversarial examples[3].

- **Safety&Ethics** — evaluates whether the agent avoids harmful actions, mitigates biases and adheres to ethical norms. Metrics encompass hallucination rate, harmful-content generation, fairness scores and compliance with regulatory requirements[4]. A harm-reduction index integrates hallucination, toxicity and fairness measures.

- **Human-Centred Interaction** — captures how users perceive and interact with the agent. Metrics include user satisfaction (CSAT/NPS), trust scores, transparency, explainability and cognitive load. Human-in-the-loop assessments use instruments like TrAAIT to measure trust[6].

- **Economic&Sustainability Impact** — examines cost–benefit trade-offs and long-term sustainability of deployment. Metrics include productivity gain, return on investment, carbon footprint of compute resources and alignment with organisational goals. This axis addresses the economic dimension often overlooked in technical benchmarks[2].

3.2 Measuring Complex Behaviours

Goal drift and alignment adherence. Agentic systems must maintain alignment with user intent over long horizons. We use a *goal-drift* score that penalises deviations from the initial goal across intermediate steps by comparing plan states and actions to the original specification. Evaluating goal drift is crucial because agents may gradually adopt new objectives in response to competing pressures; recent research demonstrates that even state-of-the-art language-model agents exhibit measurable goal drift under adversarial pressures and long context windows[7]. A high goal-drift score indicates poor alignment; evaluators should aim for low scores.

Resilience to environment shifts. To capture adaptability we evaluate agents in noisy and adversarial environments, including perturbations to inputs, changes in available tools and dynamic goal alterations. Recovery time and success rate under these perturbations quantify resilience. Backtracking efficiency measures how quickly an agent abandons an unproductive plan and adopts a new strategy.

Safety and ethical auditing. Rather than simply counting hallucinations, our framework computes a *harm-reduction index* that integrates hallucination rate, toxicity score and fairness metrics. Evaluators should also measure compliance with domain-specific regulations and track how often the agent escalates uncertain decisions to human oversight. Custom tests for bias and toxicity are essential to assess both quantitative performance and qualitative risks[8].

Human-agent co-evaluation. Human judges provide qualitative assessments of explanations, empathy and trustworthiness. We recommend using validated instruments such as TrAAIT to measure trust[6]. Combining human scores with technical metrics yields a more comprehensive understanding of system performance.

Economic sustainability. Productivity gains should be contextualised with deployment costs and environmental impact. For instance, cost per completed task can be normalised by energy consumption or carbon emissions. Return on investment must consider long-term maintenance, training costs and the opportunity cost of human labour.

3.3 Visualising the Framework

Figure 1 illustrates the interdependencies among the five axes as an abstract network. Each node represents a dimension and is connected to the others, reflecting that improvements on one axis may influence outcomes on another (e.g. robustness improvements may increase compute cost, reducing efficiency). The illustration was created specifically for this paper and can be replaced with a higher-fidelity diagram if desired.

4 Proposed Experiments and Validation

To strengthen the novelty of our framework we outline a set of empirical evaluations that practitioners can perform on agentic systems. These experiments are designed to test goal adherence, resilience, safety and user trust in realistic scenarios.

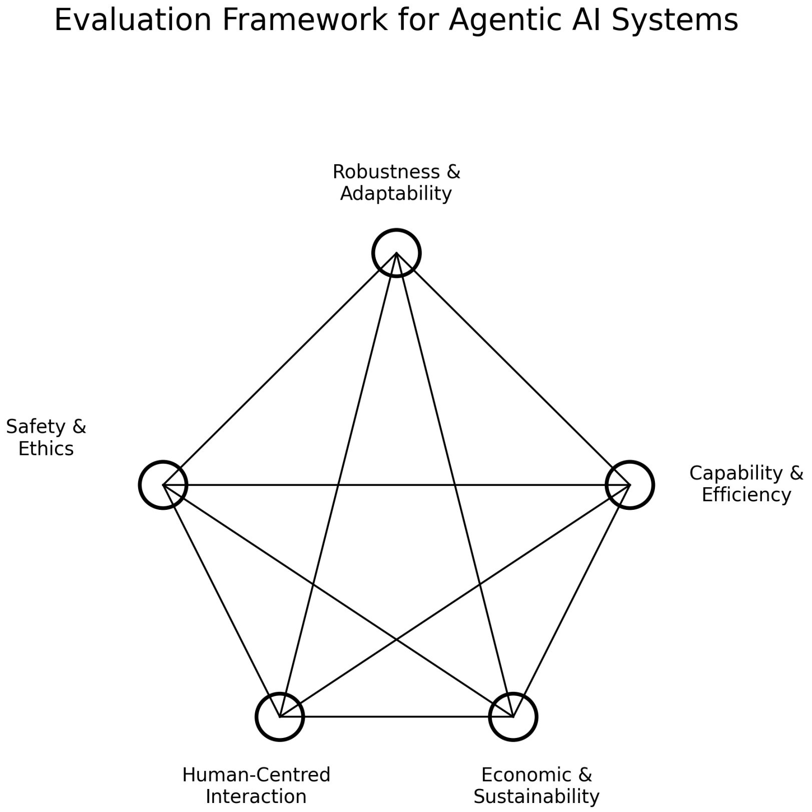
Evaluation Framework for Agentic AI Systems
Robustness & Adaptability
Safety & Ethics
Capability & Efficiency
Human-Centred Interaction
Economic & Sustainability

Figure 1: Interconnected evaluation dimensions for agentic AI systems. The abstract network shows how the five axes—capability&efficiency, robustness&adaptability, safety&ethics, human-centred interaction and economic&sustainability—are interdependent. Improvements in one dimension may affect others.

4.1 Goal-Drift Evaluation

Inspired by recent studies on goal drift in language-model agents[7], we propose an experimental environment where an agent is given a primary goal (e.g. "write code to add logging to a module") and later exposed to competing objectives (e.g. "improve performance" or "refactor unrelated files"). The agent's actions are logged and compared against the original specification. A *goal-drift score* is computed as the divergence between the intended plan and the executed actions over time. Evaluators can vary the difficulty of competing goals and measure how quickly different agent architectures drift from their assigned objectives. This test quantifies alignment adherence and informs the design of safety mechanisms such as explicit goal reminders or constrained planning.

4.2 Robustness Under Perturbations

To assess robustness and adaptability, agents should be evaluated under noisy and adversarial inputs. For example, web-browsing agents can be tested with corrupted HTML pages or misleading links; code-generation agents can be fed faulty dependency graphs. Recovery time, success rate and backtracking efficiency serve as metrics. Experiments should include tool outages, dynamic goal changes and adversarial prompts. The evaluation should report both aggregate metrics and worst-case behaviours to reveal brittleness.

4.3 Safety and Ethical Auditing

Evaluators should use red-teaming exercises and toxicity probes to measure harmful behaviour. Prompting the agent with borderline queries (e.g. ethically sensitive scenarios) allows measurement of a harm-reduction index. Fairness tests can simulate varied demographic profiles and quantify disparities in outcomes. Automated scoring should be complemented by human review to ensure nuanced interpretation[8].

4.4 Human-Centred Trust Evaluation

We recommend conducting user studies where participants interact with the agent in realistic tasks (e.g. drafting reports or answering questions). After completion, participants complete trust surveys such as TrAAIT[6], rating the agent's information credibility, perceived application value and reliability. Qualitative interviews can reveal reasons for high or low trust and highlight design improvements. Combining trust scores with objective metrics illuminates the trade-offs between efficiency and user acceptance.

4.5 Economic and Sustainability Analysis

Finally, deployment cost and environmental impact should be measured by tracking compute utilisation, energy consumption and carbon footprint. Cost per completed task, return on investment and productivity gains are analysed in conjunction with the other metrics. This analysis encourages responsible adoption and counters purely short-term efficiency gains.

5 Case Studies: Agentic AI in the Wild

To demonstrate the applicability of our framework we examine three case studies from a 2025 McKinsey report on agentic AI deployments[5]. These cases span software modernisation, market research and banking, and provide quantitative impact estimates.

5.1 Legacy Application Modernisation

Context. A large bank sought to modernise its legacy core system comprising hundreds of pieces of software. Manual coding and documentation made coordination across silos difficult and slowed progress. Although early generative-AI tools accelerated individual tasks, overall velocity remained low[5]. In the agentic approach, human workers became supervisors overseeing squads of AI agents. Each squad documented legacy applications, wrote new code, reviewed others' code and integrated features. By elevating humans to strategic oversight and delegating repetitive tasks to agents, the bank achieved more than a 50 % reduction in time and effort for early-adopter teams[5].

Framework perspective. This case exhibits high capability and efficiency (dramatic time reduction) and improved robustness via multiple agents cross-validating outputs. Safety risk was low in this controlled environment. However, human-centred metrics (e.g. developer trust in agent code) were not reported, indicating an evaluation gap.

Table 1: Summary of case studies from McKinsey[5]. Reported impacts focus on productivity gains but rarely include human-centred or safety metrics.

Case	**Agentic approach**	**Reported impact**
Legacy modernisation	Humans supervise squads of agents to document, code, review and integrate features	> 50 % reduction in time/effort
Data quality & insights	Agents detect anomalies, analyse internal/external signals and synthesise drivers	> 60 % productivity gain; > $3 M annual savings
Credit-risk memos	Agents extract data, draft sections, generate confidence scores; humans supervise	20–60 % productivity; 30 % faster decisions

5.2 Market-Research Data Quality and Insight Generation

Context. A market-research firm employed more than 500 people to gather, structure and codify data; 80 % of errors were detected by clients[5]. A multi-agent solution autonomously identified anomalies, analysed internal signals (e.g. product taxonomy changes) and external events (e.g. recalls, severe weather) and synthesised key drivers for decision-makers. The system promised more than 60 % potential productivity gain and over US3 M*inannualsavings*[5].

Framework perspective. Capability and economic impact are strong; robustness improved through anomaly detection. Safety (bias) and human-factors metrics (analyst trust) were not reported, demonstrating the need for multidimensional assessments.

5.3 Credit-Risk Memo Generation

Context. Relationship managers at a retail bank spent weeks writing credit-risk memos, manually extracting information from multiple data sources and reasoning across interdependent sections[5]. An agentic proof of concept extracted data, drafted memo sections, generated confidence scores and suggested follow-up questions, shifting human analysts toward strategic oversight and exception handling. Reported gains were 20–60 % productivity and 30 % faster credit decisions[5].

Framework perspective. Capability and efficiency improved; however, safety and ethics are critical (e.g. fairness, compliance). Transparent rationales and bias monitoring are necessary for deployment.

5.4 Cross-Case Analysis

Table 1 summarises the three cases and their reported impacts. All show substantial efficiency gains but omit human-centred and safety metrics, underscoring the measurement imbalance.

6 Implications and Future Directions

6.1 Towards Balanced Benchmarks

Our framework underscores the necessity of benchmarks that go beyond task success and latency. Evaluations should include long-horizon planning, tool usage and inter-agent communication; robustness under noisy/adversarial inputs; human-in-the-loop trust scoring; and economic sustainability metrics such as energy consumption and cost per outcome. The AI research community should develop public leaderboards that report all five axes and provide full evaluation scripts and data. This echoes calls for multidimensional evaluation in the LLM community[8].

6.2 Reproducibility and Open Evaluation

Because agentic systems involve stochastic LLMs and external tools, evaluators should fix random seeds, log tool calls and specify environment configurations to enable repeatable experiments. All experiments should be open-sourced, and evaluation platforms should publish datasets, prompts and scoring scripts. Standards bodies could recommend minimal reporting requirements for agentic evaluations.

6.3 Human-Agent Collaboration and Trust

Integrating trust instruments like TrAAIT[6] into evaluation pipelines will help capture satisfaction, transparency and acceptance. Observing when humans accept, override or request explanations informs design choices and accountability. Empowering end users to calibrate autonomy levels can mitigate goal drift and reduce over-reliance.

6.4 Policy and Governance

Regulators and organisations need evaluation standards for high-impact domains (finance, healthcare, education). Our framework highlights metrics that should be mandatory: fairness, harmful-action avoidance, transparency logs, alignment adherence and environmental impact. Governance must balance innovation with accountability and incorporate human oversight at critical decision points.

7 Conclusion

Agentic AI promises to transform work by coordinating multiple agents with memory, planning and tool use. Yet evaluation practices have lagged behind. Current benchmarks privilege technical performance and often omit human-centred, ethical and economic dimensions[2]. We proposed a balanced framework across capability&efficiency, robustness&adaptability, safety&ethics, human-centred interaction and economic&sustainability. We introduced novel metrics such as goal-drift and harm-reduction scores, provided an abstract visualisation of the framework and outlined experimental protocols for empirical validation. Case studies suggest substantial productivity gains (20–60 %)[5] but reveal missing evaluations of robustness, fairness and trust. Measuring agentic AI holistically is essential to ensure it delivers value safely and equitably. We hope this framework will guide researchers, practitioners and policymakers toward more responsible adoption of agentic AI systems.

Acknowledgements

We thank the open-source communities and researchers whose insights informed this work. This paper is an independent contribution and does not represent the views of any organisation.

References

[1] R. Sapkota, G. Tambwekar, A. Crespi, A. Ramachandran and H. Long. "AI Agents vs Agentic AI: A Conceptual Taxonomy, Applications and Challenges." *arXiv preprint arXiv:2505.10468*, 2025.

[2] K. J. Meimandi, N. Arsenlis, S. Kalamkar and A. Talwalkar. "The Measurement Imbalance in Agentic AI Evaluation Undermines Industry Productivity Claims." *arXiv preprint arXiv:2506.02064*, 2025.

[3] C. Bronsdon. “AI Agent Evaluation: Methods, Challenges, and Best Practices.” Galileo, 2025. https://www.galileo.ai/blog/ai-agent-evaluation.

[4] QAwerk. “AI Agent Evaluation: Metrics That Actually Matter.” Blog, 2025. https://qawerk.com/blog/ai-agent-evaluation-metrics/.

[5] McKinsey&Company. “Seizing the Agentic AI Advantage: A CEO Playbook.” Report, 2025. https://www.mckinsey.com/featured-insights/mckinsey-technology-and-innovation/seizing-the-agentic-ai-advantage.

[6] A. F. Stevens, P. Stetson and colleagues. “Theory of trust and acceptance of artificial intelligence technology (TrAAIT): An instrument to assess clinician trust and acceptance of artificial intelligence.” *Journal of Biomedical Informatics*, 148:104550, 2023. (Cited for the TrAAIT model and trust instrument.)

[7] R. Arike, E. Donoway, H. Bartsch and M. Hobbhahn. “Technical Report: Evaluating Goal Drift in Language Model Agents.” *arXiv preprint arXiv:2505.02709*, 2025. (Cited for the goal-drift evaluation methodology.)

[8] C. Dilmegani. “Large Language Model Evaluation in 2025: 10+ Metrics & Methods.” AIMultiple, 2025. https://research.aimultiple.com/large-language-model-evaluation/. (Cited for the need to combine automated metrics with human and fairness evaluations.)