International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 6, Nov – Dec 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Enterprise AI Evaluation and Benchmarking Framework

A Comprehensive Enterprise Framework for Evaluating AI Quality, Safety, Reliability, Security, Fairness, Cost, Business Value, and Production Readiness

Sanjeeve Kumar Gajadi*
*(HCL Technologies, Bengaluru, Karnataka, India)
Email: sanjeevekumar.g@hcltech.com

****

Abstract:

Enterprise AI systems require evaluation methods that extend beyond accuracy because predictive models, Generative AI, Retrieval-Augmented Generation, and agentic systems exhibit different failure modes and business risks. Organizations often rely on isolated benchmark scores that do not reflect enterprise data, business processes, security constraints, user expectations, or operational cost. This paper proposes an Enterprise AI Evaluation and Benchmarking Framework that combines task quality, factuality, robustness, fairness, privacy, security, explainability, latency, cost, usability, business value, and policy compliance. The framework defines evaluation datasets, golden cases, red-team scenarios, human review, automated judges, statistical confidence, baseline comparison, regression testing, risk-tier-based acceptance criteria, and production monitoring. It introduces benchmark governance, score normalization, composite scorecards, release gates, and continuous evaluation. The proposed approach enables enterprises to compare models and AI solutions consistently, select fit-for-purpose technologies, detect regressions, and maintain defensible evidence of production readiness.

Keywords — AI Evaluation, AI Benchmarking, Generative AI Evaluation, Model Evaluation, RAG Evaluation, Agentic AI, Responsible AI, AI Quality, AI Safety, Enterprise AI

****

I. INTRODUCTION

Enterprise AI evaluation must answer whether an AI system is suitable for a specific business purpose, not merely whether it performs well on public benchmarks.

Evaluation should reflect enterprise data, workflows, risk appetite, user expectations, operational constraints, and failure consequences.

II. EVALUATION OBJECTIVES

The framework supports model selection, release approval, vendor comparison, regression detection, risk assessment, cost optimization, and continuous improvement.

Different AI patterns require tailored metrics while maintaining a common enterprise scorecard for governance.

III. EVALUATION DIMENSIONS

Core dimensions include task quality, factuality, relevance, robustness, fairness, privacy, security, explainability, usability, latency, availability, cost, and business value.

High-risk use cases may add safety, regulatory, resilience, and human-oversight criteria.

IV. EVALUATION DATASET DESIGN

Evaluation datasets should represent normal cases, edge cases, high-risk scenarios, sensitive data, rare events, and realistic enterprise language.

Golden datasets require versioning, provenance, ownership, confidentiality controls, and periodic refresh to avoid benchmark staleness.

V. PREDICTIVE MODEL EVALUATION

Predictive models may use accuracy, precision, recall, calibration, discrimination, error cost, fairness, robustness, and stability metrics.

 ©IJSRED: All Rights are Reserved

Metrics should be selected according to business consequences rather than convenience.

VI. GENERATIVE AI EVALUATION

Generative AI requires assessment of factuality, groundedness, completeness, instruction following, harmful output, style, consistency, and refusal behavior.

Human review and automated evaluators can be combined, but automated judges should themselves be validated and monitored.

VII. RAG EVALUATION

RAG evaluation should separate retrieval quality from generation quality. Measures can include recall, precision, relevance, context utilization, citation accuracy, groundedness, and answer completeness.

Enterprise tests should verify that retrieval preserves permissions and excludes unauthorized or stale content.

VIII. AGENTIC AI EVALUATION

Agentic systems require scenario-based evaluation of planning, tool selection, authorization compliance, task completion, recovery, delegation, reversibility, and human escalation.

Success should be measured across full workflows rather than individual model responses.

IX. SECURITY AND RED-TEAM TESTING

Evaluation should include prompt injection, data exfiltration, unsafe tool use, jailbreak attempts, malicious retrieval content, privilege escalation, and denial-of-service scenarios.

Security test results should influence release decisions and residual-risk scoring.

X. BENCHMARKING AND BASELINES

Models should be compared against stable baselines, previous production versions, alternative providers, and human performance where appropriate.

Benchmark results should include confidence intervals, dataset version, configuration, prompt version, model version, and evaluation date.

XI. COMPOSITE SCORECARDS

A composite scorecard can normalize metrics across dimensions while preserving visibility into critical failures that should never be hidden by averages.

Risk-based minimum thresholds should be defined for mandatory dimensions such as safety, privacy, security, or regulatory compliance.

XII. RELEASE GATES AND REGRESSION TESTING

Production release should require documented acceptance criteria and regression testing against previous approved versions.

Material model, prompt, data, retrieval, or tool changes should automatically trigger relevant evaluation suites.

XIII. CONTINUOUS EVALUATION

Production telemetry, user feedback, incidents, drift, business outcomes, and new failure cases should feed back into evaluation datasets.

Continuous evaluation transforms benchmarking from a one-time selection activity into an ongoing assurance process.

XIV. CONCLUSION

The Enterprise AI Evaluation and Benchmarking Framework provides a practical method for comparing and assuring AI systems across quality, risk, cost, and business value.

By combining representative datasets, risk-based metrics, release gates, and continuous evaluation, enterprises can make more defensible AI adoption decisions and detect regressions before they create material impact.

ACKNOWLEDGMENT

The author acknowledges the contributions of enterprise architecture practitioners, AI governance communities, security professionals, standards bodies, and technology researchers whose published work has informed the concepts synthesized in this paper.

REFERENCES

[1] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), 2023.

[2] National Institute of Standards and Technology, AI RMF Generative Artificial Intelligence Profile, 2024.

 ©IJSRED: All Rights are Reserved

[3] ISO/IEC 42001:2023, Information technology — Artificial intelligence — Management system.
[4] ISO 31000:2018, Risk management — Guidelines.
[5] ISO/IEC 27001:2022, Information security, cybersecurity and privacy protection — Information security management systems.
[6] National Institute of Standards and Technology, SP 800-207: Zero Trust Architecture, 2020.
[7] The Open Group, TOGAF Standard, 10th Edition, 2022.
[8] ISACA, COBIT 2019 Framework: Governance and Management Objectives, 2019.
[9] OWASP Foundation, OWASP Top 10 for Large Language Model Applications.
[10] MITRE, ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems.
[11] OECD, Recommendation of the Council on Artificial Intelligence.
[12] IEEE, Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems.

 ©IJSRED: All Rights are Reserved