International Journal of Computer Science and Engineering Innovations
ISSN: 3107-9458 | Volume 2 Issue 4, 13-16, October-December 2026
DOI: https://doi.org/10.64137/31079458/IJCSEI-V2I4P103
Received: 12/09/2026 Revised: 24/09/2026 Accepted: 30/09/2026 Published: 03/10/2026

Scholastic Research Publication

Original Article

From UAT to AI Evaluation: A Requirements-Based Framework for Human Validation of Enterprise AI Systems

[1]AISHWARYA KANNOTH PUTLUMBATH, [2]VISHWAJEET KHAROTE, [3]UTSAV KALE, [4]GAURAV BHARDWAJ, [5]RUGVED GRAMOPADHYE, [6]SAJEEL SHAMSUDEEN
[1]Department of Computer Science, University of Wisconsin Milwaukee, USA.
[2]Computer Science, San Jose University, California, USA.
[3]Government Polytechnic Amravati, Amravati, Maharashtra, India.
[4]APJ Abdul Kalam Technological University, Thiruvananthapuram, Kerala, India.
[5]University of Dallas at Texas, Dallas, USA.
[6]Independent Researcher, USA.

ABSTRACT: *Enterprise organizations are increasingly embedding generative and agentic AI into business applications historically validated through User Acceptance Testing (UAT). Conventional UAT is typically structured around predefined requirements and sufficiently predictable expected outcomes evaluated using pass/fail criteria; generative AI complicates this model because outputs are probabilistic and context-sensitive, and a response may be fluent and confident while incorrect, non-compliant with a business rule, or inconsistent with prior behavior. This conceptual, design-science paper synthesizes requirements engineering, software testing, human-AI interaction, AI evaluation, and AI governance literature into a requirements-based framework for human validation of enterprise AI systems, comprising requirements elicitation, AI-specific acceptance criteria, risk-tiered human validation, and feedback/remediation loops under a governance layer supporting continuous revalidation. The paper contributes an explicit mapping from conventional UAT concepts to AI-specific extensions and an AI defect taxonomy that routes findings to remediation owners. Human validation is positioned as complementary to, not a replacement for, automated evaluation. A generic, composite scenario illustrates the framework's application. Explicit limitations, including the absence of empirical validation, are discussed.*

KEYWORDS: *Acceptance Criteria, AI Defect Taxonomy, AI Evaluation, AI Governance, Business Rule Compliance, Continuous Evaluation, Design Science Research, Enterprise AI, Hallucination Detection, Human-AI Collaboration, Human-in-the-Loop, Requirements Engineering, Risk-Tiered Validation, User Acceptance Testing.*

1. INTRODUCTION

Enterprise software delivery has long relied on UAT as the final validation gate before go-live. Conventional UAT assumes a correct outcome can be specified in advance and checked deterministically for a given case. Embedding generative and agentic AI into enterprise applications complicates this model: outputs vary across otherwise identical invocations, and an output can be fluent and confident while factually wrong, while violating an undocumented business rule, or while drifting after deployment as models, prompts, or data change.

A substantial AI-evaluation literature addresses benchmarking, preference alignment, and agent evaluation; a separate, established literature addresses requirements engineering and UAT. The connective layer between them- how enterprises should extend, not discard, RE/UAT discipline to govern AI features remains underdeveloped, a gap corroborated by agent-evaluation work that independently flags missing enterprise concerns such as compliance and role-based access [10].

This paper contributes: (1) an explicit mapping from conventional UAT concepts to AI-specific extensions (Sec. III); (2) an operational framework, requirements elicitation, AI-specific acceptance criteria, risk-tiered validation, and feedback loops under governance with defined stakeholders and artifacts (Sec. IV); (3) an AI defect taxonomy routing findings to remediation owners (Sec. IV); and (4) a positioning of human validation as complementary to automated evaluation (Sec. VI). This is a conceptual contribution; Section VII discusses what empirical validation would require.

2. BACKGROUND AND RELATED WORK

2.1. UAT AND REQUIREMENTS ENGINEERING

UAT is the final validation stage in most enterprise delivery lifecycles, structured around requirements traceability and predefined pass/fail criteria [1], [2]. This model suits deterministic business logic well but assumes a correct result is knowable in advance for a given input.

This work is an open access and licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License
(https://creativecommons.org/licenses/by-nc-nd/4.0/)

2.2. HUMAN-AI INTERACTION AND AUTOMATED EVALUATION

Amershi et al. and Bansal et al. show that effective human-AI collaboration requires calibrated trust and designed review checkpoints rather than informal oversight [3], [4]. Automated methods RLHF-based alignment [5] and LLM-as-judge scoring [6] are validated against human preference data but exhibit documented biases distinct from human reviewer biases [7]; hallucination detection remains an active research problem [8], [9]. Agent-evaluation surveys explicitly note that current benchmarks omit enterprise concerns such as compliance and role-based access [10].

2.3. AI GOVERNANCE

The NIST AI RMF and its Generative AI Profile call for continuous governance across the AI lifecycle rather than a single check [11], [12]; ISO/IEC 42001 specifies an organizational AI management system [13]; the EU AI Act imposes risk-tiered human-oversight obligations [14]; ISO/IEC 25010 offers a general, pre-generative-AI quality model [15].

2.4. IDENTIFIED GAP

Existing literature is model-centric (evaluation) or organization-centric (governance); enterprises lack an operational bridge extending RE/UAT discipline into AI feature validation. Sections III–IV propose that bridge.

3. RESEARCH METHOD OR FRAMEWORK DEVELOPMENT

This paper follows a design-science approach: the contribution is a designed artifact, justified through literature synthesis rather than new empirical observation [16], and is not a systematic review. Three bodies of literature were synthesized: RE/UAT standards [1], [2] supplied the baseline lifecycle; human-AI interaction research [3], [4] supplied principles for positioning human checkpoints; AI evaluation and governance literature [5]–[15] supplied AI-specific failure modes (non-determinism, hallucination, judge bias, business-rule violation, drift) absent from the conventional lifecycle. Framework components were derived by mapping each failure mode onto the UAT lifecycle stage it most disrupts (Table 1).

TABLE 1 Mapping from Conventional UAT Concepts to AI-Specific Extensions

Traditional UAT Concept	AI-Specific Challenge	Framework Mechanism
Predefined expected result	Multiple outputs may be acceptable	AI acceptance criteria (Sec. IV)
Functional requirement	Behavior on ambiguous input is unspecified	Requirements elicitation (Sec. IV)
Scripted test execution	Exhaustive scripting is infeasible	Risk-tiered validation (Sec. IV)
Pre-release sign-off	Behavior drifts after deployment	Governance & revalidation (Sec. IV)

4. PROPOSED REQUIREMENTS-BASED AI VALIDATION FRAMEWORK

The framework comprises four components: requirements elicitation, AI-specific acceptance criteria, risk-tiered human validation, and feedback/remediation loops under a governance layer managing sign-off and continuous revalidation. Table 2 summarizes each component's objective, primary stakeholders, key artifact, and escalation mechanism.

TABLE 2 Operational Summary of the Framework's Components

Component	Objective	Stakeholder(s)	Key Artifact	Escalation
Requirements Elicitation	Capture functional, behavioral, failure-mode requirements	BA/product owner; SME; compliance	Requirements register, severity-tagged	Undocumented rules escalate to governance
AI Acceptance Criteria	Translate requirements into scoreable, risk-appropriate thresholds	Business owner; compliance; QA lead	Scored rubric per risk tier	Inconsistently scored dimensions return to rubric design
Risk-Tiered Validation	Apply review capacity where it reduces most risk	SME/compliance reviewers; QA lead	Reviewed sample; logged findings	Findings passed to feedback loop, tagged by severity
Feedback & Remediation	Route findings to correct owner via defect taxonomy	QA lead; owning team; governance	Classified defect log; remediation record	Recurring findings trigger component review
Governance & Revalidation	Certify go-live; sustain confidence post-launch	Business/compliance owners; governance body	Sign-off record; audit trail	Failed revalidation returns to risk-tiered validation

Candidate acceptance-criteria dimensions include factual accuracy, groundedness, business-rule compliance, hallucination, appropriate abstention, completeness, relevance, tone, policy/compliance adherence, and permission/security boundaries; appropriate dimensions and thresholds are use-case- and risk-dependent, and no universal numerical thresholds are proposed. Risk-tier assignment is informed by business impact, customer impact, regulatory impact, reversibility of error, financial consequence, and autonomy level, rather than a formally validated sampling ratio.

Findings are classified by an AI defect taxonomy before routing (Table 3), since a wrong output may originate in the requirement, prompt, retrieved data, source data, model behavior, tool integration, a business rule, or an undocumented policy.

TABLE 3 AI Defect Taxonomy and Typical Remediation Owner

Defect Category	Typical Remediation Owner
Requirement defect	Business analyst/product owner
Prompt / instruction defect	Prompt / product team
Retrieval / knowledge-source defect	Data/knowledge engineering team
Source-data defect	Data owner / upstream system owner
Model-behavior defect	AI engineering team
Integration / tool-use defect	Engineering/integration team
Business-rule defect	Business owner; rubric revision
Governance / policy-ambiguity defect	Governance body

5. ILLUSTRATIVE ENTERPRISE APPLICATION

The following composite, illustrative scenario of an AI assistant drafting billing-inquiry responses for human-agent review is not drawn from any real deployment and is not empirical evidence.

TABLE 4 Illustrative Walkthrough of the Framework (Illustrative Only)

Stage	Illustrative Content
Requirement	Must cite source policy field for refund amounts; flag out-of-scope requests.
AI acceptance criterion	Zero-tolerance for unauthorized discounts; required abstention on low-confidence questions.
Automated evaluation	Rubric-based check flags outputs lacking citation or containing discount claims.
Human review	SME checks flagged and sampled outputs against source account data.
Defect classification	Unsupported discount claims classified as a business-rule or model-behavior defect.
Remediation & revalidation	Rubric/prompt update or engineering fix; re-tested against original sample.
Sign-off / monitoring	Business and compliance sign-off; reduced-sampling post-launch monitoring.

6. DISCUSSION

6.1. HUMAN AND AUTOMATED EVALUATION AS COMPLEMENTARY LAYERS

TABLE 5 Complementary Strengths of Automated and Human Evaluation

Dimension	Automated Evaluation	Human Evaluation
Scale/repeatability	Continuous, consistent given a fixed rubric	Limited by capacity; variable across reviewers
Strength	Monitoring; regression/drift detection	Contextual judgment; ambiguous-rule interpretation
Known weakness	Documented judge biases [6], [7]	Cost; inter-rater disagreement; reviewer bias
Best suited to	High-volume, well-specified checks	Novel failure modes; accountability/sign-off

Automated evaluation should carry continuous, high-volume monitoring, while human validation is reserved for risk-tiered sampling, uncertain cases, and governance checkpoints (Sec. IV).

6.2. ORGANIZATIONAL IMPLICATIONS AND ADOPTION

Adopting the framework changes existing roles: UAT coordinators need training in risk-tiered sampling and rubric scoring; QA functions need the defect taxonomy in Table 3 to route findings correctly; compliance needs earlier visibility, since business-rule criteria are set during elicitation. Adoption is likely incremental, applied first to one high-risk feature to calibrate rubric design before wider rollout. Where business rules are undocumented, the elicitation stage surfaces that gap before it reaches acceptance testing.

7. LIMITATIONS AND FUTURE RESEARCH

This is a conceptual framework, not an empirically validated one. It has not been tested against real deployments for defect-detection effectiveness, cost, or adoption barriers. Risk-tier factors (Sec. IV) are qualitative guidance, not a validated sampling formula, and reviewer inter-rater reliability is acknowledged but not measured. Acceptance-criteria thresholds are domain-dependent, and no universal values are proposed; establishing ground truth for accuracy and groundedness can itself be difficult, and undocumented business rules limit near-term applicability of the rule-compliance dimension. Both human and automated evaluators carry their own biases [7]. Future work should empirically test the framework against real or simulated enterprise features to measure reliability, defect-detection effectiveness, and adoption cost, and examine how risk-tier thresholds should evolve as AI governance regulation matures [12], [14].

8. CONCLUSION

Enterprise adoption of generative and agentic AI does not make UAT discipline obsolete, but requires extending it to accommodate probabilistic, context-sensitive, and occasionally fluent-but-wrong outputs. The framework proposed here requirements elicitation, AI-specific acceptance criteria, risk-tiered validation, and feedback or remediation loops under governance offers a structured, auditable bridge between established RE/UAT practice and emerging AI evaluation techniques, positioning human validation as complementary to automated evaluation. Empirical validation remains necessary future work.

CONFLICTS OF INTEREST

The author(s) declare(s) that there is no conflict of interest concerning the publication of this paper.

ACKNOWLEDGEMENTS

The author(s) received no specific funding for this work.

REFERENCES

[1] "IEEE Standard for Software and System Test Documentation," IEEE Std 829-2008, pp. 1–150, July 2008, doi: 10.1109/IEEESTD.2008.4578383.

[2] "ISO/IEC/IEEE 29119-1:2022 Software and systems engineering Software testing," 2022. [Online]. Available: https://www.iso.org/standard/81291.html

[3] S. Amershi et al., "Guidelines for Human-AI Interaction," Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19, 2019, doi: 10.1145/3290605.3300233.

[4] G. Bansal et al., "Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance," Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–16, May 2021, doi: 10.1145/3411764.3445717.

[5] "Ouyang, L., Wu, J., Jiang, X., et al., Training Language Models to Follow Instructions With Human Feedback. arXiv 2203.02155. - References - Scientific Research Publishing," 2022. https://www.scirp.org/reference/referencespapers?referenceid=4079808

[6] L. Zheng et al., "Judging LLM-as-a-Judge With MT-Bench and Chatbot Arena," arXiv.org, 2023. https://doi.org/10.48550/arXiv.2306.05685

[7] G. H. Chen, S. Chen, Z. Liu, F. Jiang, and B. Wang, "Humans or LLMs as the Judge? A Study on Judgement Bias," Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 8301–8327, Jan. 2024, doi: 10.18653/v1/2024.emnlp-main.474.

[8] Z. Ji et al., "Survey of Hallucination in Natural Language Generation," ACM Computing Surveys, vol. 55, no. 12, Nov. 2022, doi: 10.1145/3571730.

[9] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, "Detecting Hallucinations in Large Language Models Using Semantic Entropy," Nature, vol. 630, no. 8017, pp. 625–630, June 2024, doi: 10.1038/s41586-024-07421-0

[10] M. Mohammadi, Y. Li, J. Lo, and W. Yip, "Evaluation and Benchmarking of LLM Agents: A Survey," Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, pp. 6129–6139, Aug. 2025, doi: 10.1145/3711896.3736570

[11] NIST, "AI Risk Management Framework," Artificial Intelligence Risk Management Framework (AI RMF 1.0), vol. 1, no. 1, Jan. 2023, doi: 10.6028/nist.ai.100-1

[12] NIST, "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile," Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, July 2024, doi: 10.6028/nist.ai.600-1

[13] International Organization for Standardization, "ISO/IEC 42001:2023," ISO, 2023. https://www.iso.org/standard/42001

[14] European Union, "Regulation - EU - 2024/1689 - EN - EUR-Lex," eur-lex.europa.eu, June 13, 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[15] "ISO ISO/IEC 250102011, 'Systems and Software Engineering Systems and Software Quality Requirements and Evaluation (SQuaRE) System and Software Quality Models,' 2011. - References - Scientific Research Publishing," Scirp.org, 2026. https://www.scirp.org/reference/referencespapers?referenceid=891857

[16] A. Hevner, S. March, J. Park, and S. Ram, "Design Science in Information Systems Research," MIS Quarterly, vol. 28, no. 1, pp. 75–105, 2004, doi: 10.2307/25148625.