The Impact of Generative AI on Test & Evaluation: Challenges and Opportunities

Laura Freeman
National Security Institute
Virginia Tech
Ballston, VA, USA
lamorgan@vt.edu

John Robert
Software Solutions Division
Software Engineering Institute
Pittsburgh, PA, USA
jer@sei.cmu.edu

Heather Wojton
Operational Evaluation Division
Institute for Defense Analyses
Alexandria, VA, USA
hwojton@ida.org

ABSTRACT

Generative Artificial Intelligence (AI) is transforming software development processes, including the critical domain of test and evaluation (T&E). From automating test case design to enabling continuous testing in DevOps pipelines, AI-driven tools can enhance the efficiency, accuracy, and speed of software testing. At the same time, however, the integration of AI components, especially generative models like large language models (LLMs), into software-reliant systems introduces new challenges for verification and validation. Traditional T&E methodologies must evolve to address issues such as AI bias, hallucinated outputs, and the complexity of validating non-deterministic behaviors. This paper examines how generative AI is changing T&E practices across the software development lifecycle (SDLC), explores the challenges of testing AI capabilities, and discusses key concerns like reliability and regulatory compliance. Drawing on insights from a recent webinar with leaders from the T&E domain and related research [1], this paper provides recommendations for refining T&E strategies to ensure that AI-augmented, software-reliant systems are trustworthy and effective in practice.

KEYWORDS

Generative AI, Large Language Models (LLMs), Test & Evaluation (T&E), AI-Driven Test Automation, Verification and Validation,

1. AI-Augmented Transformation of Test and Evaluation Practices

As AI continues to revolutionize software development, its impact on test and evaluation (T&E) will likely be equally profound. AI-augmented testing methodologies have the potential to reshape traditional verification and validation processes by introducing automation, accelerating test case generation, and enabling real-time monitoring. These advancements can address long-standing challenges in software assurance, including reducing manual effort, improving test coverage, and integrating user feedback earlier in the development cycle. Below, we summarize key areas where AI has the potential to drive critical shifts in T&E practices.

Automating and accelerating test design. AI-augmented tools can automatically generate test cases, test data, and even testing scripts by analyzing requirements and past defect patterns [20]. This automation can reduce the manual effort and human error in creating extensive test suites [11]. As software complexity increases, software verification and validation processes incur significant amounts of testing. Generative AI can help create test cases that cover the vast domain of software inputs. Generative AI can leverage natural language processing on specifications and code to produce a wide range of test scenarios, including edge cases, faster than traditional methods [9]. These capabilities contribute to continuous testing, where tests are run and updated iteratively throughout development, aligning with modern agile and DevOps practices. These advances yield a significant improvement in test coverage and frequency, enabling quicker feedback on software quality within CI/CD pipelines.

Shift-Left testing and early user involvement. AI is also enabling T&E to move earlier in the SDLC (so-called "shift-left" testing), blurring the line between development and testing. One potential example of the impact of generative AI is its ability to rapidly prototype user interfaces or system behaviors, allowing developers and end-users to evaluate and provide feedback on early mock-ups. In the national security context, getting operational users involved sooner has been a long-standing goal for better outcomes.

Generative models can also help develop what the actual interface would look like for operators to facilitate user testing much earlier in the acquisition process [15]. For example, instead of waiting for a fully mature system, teams can use AI-generated prototypes and simulated interfaces to conduct usability tests and gather user feedback in the early design stages. This early engagement through AI-augmented prototyping helps to shorten the feedback loop between users and developers, leading to interfaces and features that are more aligned with user needs.

Enhanced prototyping and simulation. Generative AI can serve as an accelerator for creating software prototypes and models that extend current systems in new directions. By quickly materializing new ideas in code or design, AI allows teams to conduct small-scale tests or experiments on innovative features [23]. Another transformative aspect is the use of AI in building synthetic test environments and scenarios. In operational testing—where software-reliant systems are evaluated under realistic conditions—constructing and executing a comprehensive set of test scenarios can be costly and labor-intensive.

AI can alleviate the time and effort associated with operational T&E by generating scenarios and conditions that reflect operational realism in earlier phases of testing from high-level test design goals. By automating parts of scenario creation, generative AI helps to ensure that test campaigns cover diverse operational conditions and edge cases that might otherwise be overlooked [5]. Similarly, AI-augmented simulation tools can generate synthetic data or simulate system inputs at scale, enabling stress testing and performance evaluation under varied conditions without the cost of physical exercises [12].

Continuous testing and monitoring. In addition to upfront test design, AI contributes to continuous monitoring and validation after

deployment. Machine learning techniques can monitor software behavior in real time, flag anomalies, and even perform root-cause analysis on test failures. For instance, AI tools can sift through logs and performance metrics from test runs to detect patterns that indicate bugs or security vulnerabilities, potentially suggesting likely causes and fixes [10]. These tools augment human testers by quickly pinpointing issues in complex, distributed systems.

In addition, the incorporation of AI in the SDLC supports a tighter integration of development, security, and operations (DevSecOps) in the T&E process. The DoD's recent AI adoption strategy emphasizes a continuous cycle of iteration, innovation, and improvement, where effective feedback loops among developers, users, and T&E experts ensure capabilities are more stable, secure, ethical, and trustworthy.[7]. In practice, these AI-augmented systems are tested not just in the pre-deployment phases but continuously throughout deployment, with insights feeding back into improvements, which is essential to evaluate key quality attributes of AI-augmented systems that learn and/or evolve.

2. Practices Challenges in Testing and Evaluating AI-Augmented Systems

Despite the gains from AI-augmented testing, independently testing and evaluating AI capabilities presents unique challenges. Traditional software has deterministic or at least well-specified behaviors against which testers can validate correctness. In contrast, AI-augmented systems—especially those based on machine learning or generative models—, on the other hand, behave probabilistically and can exhibit unexpected outputs or decision patterns that are hard to predict or exhaustively test. This section highlights key T&E challenges for AI-augmented systems.

Bias and fairness issues. AI models are only as good as the data and algorithms that shape them, and they can inadvertently learn biases that lead to unfair or ineffective outcomes. An AI model may perform well on average metrics yet still harbor unintended bias that may result in harm [3]. For example, if training data underrepresents certain conditions or populations, the AI's performance on those may be poor, raising concerns of equity and reliability [18].

In the national security context, representation bias can diminish the operational effectiveness and suitability of an AI-augmented system [6]. Testing for bias requires going beyond usual functional tests to include statistical analyses of model outputs across various subgroups and scenarios. There is no single metric for "fairness," however, as it often depends on context. Multiple definitions of bias (*e.g.*, fairness, representation, or statistical parity) must be considered, which makes the evaluation of AI fairness a complex, context-dependent process.

Independent T&E teams need methodologies to detect and measure bias in AI. Curated test sets that adequately cover the intended operational employment, including edge or corner cases, can characterize the performance of the model and suggest strategies to mitigate any biases found, such as model retraining or data augmentation. Ensuring an AI's decisions are fair and justified is as critical as testing its basic functionality.

Hallucinations and unpredictable behavior. AI-augmented systems (including—but not limited to—LLMs) tend to hallucinate by producing outputs that may be plausible-sounding but entirely fabricated or incorrect [19]. These hallucinations pose a challenge for evaluation and evaluators, *i.e.*, how to validate the output of an AI that can create content dynamically. Ideally, testers would check AI outputs against ground truth, but for generative tasks (*e.g.*, open-ended text generation), ground truth may be ambiguous or vast in scope. Moreover, AI-augmented systems may perform correctly most of the time and still occasionally produce errors or nonsensical results, especially when prompted in unfamiliar ways or when the actual data encountered during operations diverges substantially from the data used for training [14].

Testers must therefore design evaluation processes that detect these failure modes. These processes should include adversarial testing or intentionally prompting or inputting edge cases to see if the model responds robustly, and manual review by subject matter experts of AI outputs for accuracy. T&E professionals must maintain healthy skepticism and systematically verify AI outputs.

Moreover, an AI-augmented system should be tested on its ability to refuse inappropriate requests, handle incomplete or noisy data gracefully, and indicate uncertainty rather than guess. These are new facets of performance assessment unique to AI behavior. Ensuring rigorous evaluation of generative AI outputs should include human-in-the-loop assessment, where experts review and score the AI's responses as part of the test protocol [22].

Validation in complex, uncertain scenarios. AI components can introduce non-deterministic and context-dependent behaviors that complicate validation. A traditional piece of software will produce the same output every time when given the same input. In contrast, a machine learning model may either (1) not behave deterministically or (2) its performance might depend on subtleties of the input distribution [17]. Test results for an AI-augmented system may exhibit variability since running the same test twice might not yield identical outcomes if the model uses randomness or has state. Likewise, AI-augmented systems might encounter inputs in the real world that differ significantly from the training data, leading to unpredictable performance drops, a phenomenon related to distributional shift or model drift [4].

For these complex models, validating correctness is only a small part of evaluating AI-augmented systems. Factors like robustness, resilience, uncertainty, and bias must also be assessed. In particular, even if an AI passes basic functional tests, evaluators must probe deeper to address issues such as

- How does the model handle noisy or adversarial inputs?
- Does it explain its reasoning or provide confidence measures?
- Will its performance degrade over time or in new environments?

These subtleties require new test artifacts and metrics.

Testers are also increasingly challenged to design experiments that reveal how an AI handles the unknown unknowns [17]. Techniques like stress testing that varies inputs systematically to find breaking points, Monte Carlo simulations of model decisions, and statistical validation over many runs are critical in characterizing an AI's reliability [13]. In summary, independently evaluating an AI capability requires a multifaceted examination of its behavior beyond what is needed for static, rule-based software to ensure the system is robust under a range of conditions.

3. Ensuring AI Reliability and Compliance in the SDLC

As AI becomes increasingly integrated into software systems, ensuring its reliability, compliance, and trustworthiness has become a critical challenge. Traditional T&E methodologies are not sufficient to assess the unique risks associated with AI-augmented decision-making. In high-stakes domains, such as defense, healthcare, and finance, organizations must adopt rigorous approaches to verify AI performance, mitigate failure risks, and adhere to evolving regulatory and ethical standards. This section explores key strategies for AI reliability testing and regulatory compliance.

Reliability and trustworthiness. Ensuring the reliability of AI-augmented systems is a paramount concern, especially in mission-critical applications from high-stakes domains. For example, the US Department of Defense (DoD) has formally recognized reliable AI as one of its ethical principles for AI deployment, alongside responsible, traceable, equitable, and governable use [8]. Achieving reliability in AI goes hand-in-hand with rigorous T&E.

Unlike conventional software, where reliability might be measured by uptime or bug counts, AI reliability encompasses consistent performance and the absence of catastrophic failures or unsafe decisions. T&E practitioners must establish confidence bounds on AI performance. For example, verifying that an autonomous vehicle's perception model detects 99.9% of obstacles in varied conditions or that a decision aid AI has a quantifiably low error rate within its intended operating domain.

Building this trust requires iterative evaluations and a "trust but verify" mindset. In practice, it is essential to never rely blindly on the output of AI models without some form of verification or fallback. For software engineers, this means incorporating fail-safes or monitors that catch when the AI's output might be erroneous. For instance, a monitor should check whether AI-generated recommendations conflict with known constraints or invariants. For T&E professors, likewise, it is crucial to test not only normal operation but also how the system behaves under unseen contexts or when an AI-augmented component or subsystem returns an uncertain result.

Reliability testing for AI could include extensive *robustness testing,* which evaluates model performance under perturbations or in simulated adversarial conditions, and/or *resilience testing*, which measures the system's ability to recover or fail safely if the AI gives a bad output. Ultimately, ensuring AI reliability is about confidence through evidence, *i.e.*, gathering sufficient test evidence under diverse scenarios to demonstrate that the AI component will perform as intended with high probability and identifying the bounds within which that remains true [17].

Regulatory and ethical compliance. As organizations integrate AI into software-reliant systems, they must navigate an evolving landscape of regulations, standards, and ethical guidelines. In U.S. national security programs, for instance, policies now demand Responsible AI, which aligns with ethical principles and avoids undue bias or safety risks. The 2023 DoD AI Adoption Strategy emphasizes that sound assurance processes for testing, evaluation, validation, and verification are imperative for Responsible AI [8]. In this context, compliance is not just a documentation exercise but is directly linked to rigorous T&E practices.

T&E professionals may need to demonstrate that an AI-augmented system complies with specific standards. For example, a model used in personnel decisions should be tested for disparate impact on classes of individuals. Likewise, an autonomous drone's targeting AI should be tested to ensure it abides by the rules of engagement. In regulated industries like healthcare or finance, AI components might require certification or audit, which in turn requires comprehensive testing evidence (*e.g.*, proving a medical diagnostic AI meets a certain level of accuracy and safety).

One emerging practice is using AI itself to aid in compliance checking. Generative AI tools can ingest large policy and regulatory documents and compare system specifications or logs against these requirements to highlight potential inconsistencies [2]. For instance, an AI assistant could scan a new software release to flag if any change might violate a cybersecurity compliance rule, effectively acting as a compliance analyst. While this AI-augmented compliance checking can speed up verification against known standards, it remains the responsibility of the T&E team to ensure the AI-augmented system's behavior remains compliant when fielded. Checking for this compliance may involve scenario-based tests that exercise ethical edge cases (*e.g.*, does an AI respond to a command that would break a law or policy?) and verifying that the system gracefully declines or defers to human judgment.

In addition, testers and evaluators should be aware of emerging AI regulations (such as the EU's AI Act or U.S. federal guidance) that might impose new testing requirements, such as documentation of training data or provisions for explainability. Going forward, new test methodologies will likely be mandated to probe the ethical and legal compliance of AI behaviors beyond functional correctness.

Need for new T&E methodologies. The advent of AI-augmented, software-reliant systems is stretching the limits of traditional T&E methods and necessitating innovative approaches. One clear need is for operationally realistic scenario-based and simulation-based testing at scale. As discussed above, generative AI itself can assist by creating rich test scenarios and synthetic environments, but the test community must develop frameworks to integrate these AI-generated artifacts into test plans.

Methods like Monte Carlo testing (running thousands of randomized scenario simulations) can help assess the distribution of AI outcomes and identify rare failure cases. Likewise, high-fidelity simulations of the operational context paired with AI scenario generation can allow testers to examine system behavior in conditions that may be too costly or dangerous to reproduce live.

Another methodological shift involves incorporating adversarial testing (*i.e.,* red teaming) as a standard part of AI evaluation. Red teaming involves stress-testing AI-augmented systems by simulating real-world attacks to identify vulnerabilities before malicious actors can exploit them. For example, a red team might test an LLM–powered chatbot for susceptibility to malicious prompts that induce disallowed behavior or test a vision AI with specially crafted images designed to fool it.

AI models are susceptible to adversarial manipulation, data poisoning, and privacy attacks. Through adversarial testing, red teaming helps improve the resilience of AI models against threats like membership inference attacks, and model extraction, which pose significant risks to privacy and data integrity [21]. These techniques,

common in cybersecurity, are becoming essential in AI safety testing to ensure models cannot be easily tricked or subverted.

In addition, there is a growing need for explainability and transparency in testing. These test methods seek to inspect and explain the AI's decision process. Techniques from the AI research community, such as saliency maps for neural networks or logic extraction from models, could be adapted into testing procedures so that evaluators can verify why an AI made certain decisions, not just what decisions it made. These techniques are particularly important for debugging and building trust with stakeholders (including regulators) that the AI is making reasonable inferences.

Finally, given the continuous-learning nature of some AI, the T&E community should adopt a more continuous assessment model rather than one-time certification. This approach involves ongoing monitoring of AI performance in operations, automated re-testing when models are updated, and periodic re-validation throughout the system's life [16]. In summary, to adequately assess AI-augmented systems, T&E professionals must expand their toolkits to employ techniques from data science, statistics, security testing, and human factors to cover the new dimensions of quality introduced by AI.

4. Recommendations for Advancing AI Test & Evaluation

Addressing the challenges described in Sections 2 and 3 above to fully leverage AI's benefits for T&E requires a concerted effort in research, policy, and cross-disciplinary collaboration. This section presents our recommendations for refining T&E strategies to better support AI-augmented workflows.

Invest in specialized AI T&E research. Stakeholders should support research into new verification and validation techniques tailored for AI. This research includes developing metrics for qualities like AI fairness, explainability, and uncertainty quantification, as well as tools to automatically detect issues like hallucinations or concept drift in models. Research efforts could explore formal methods or model-checking for machine learning, improved methods for generating adversarial test cases, and techniques for validating AI-augmented systems that learn and adapt over time. By advancing the science of AI testing, we can enable more rigorous and scalable evaluation processes.

Enhance workforce training and education. The introduction of AI into the SDLC means that test engineers and developers need to master new skills. Organizations should provide training on AI fundamentals, data science, and how to effectively use AI tools in T&E. As individuals become more familiar with using AI, they will learn about its strengths and weaknesses, informing how to better test AI. Building this expertise will help teams avoid the misuse of AI and better understand its outputs. Training should also include key test considerations emphasizing the critical review of AI results. By upskilling the T&E workforce, organizations ensure that humans remain firmly in the loop, ready to interpret AI results and intervene when something seems off.

Foster multi-stakeholder collaboration. Effective T&E for AI-augmented systems requires tighter collaboration between software engineers, AI model developers, testers, and end-users, as well as oversight and regulatory organizations. Creating forums and working groups that bring together stakeholders can accelerate learning and consensus on best practices. For example, involving domain experts and end-users (such as doctors in healthcare systems) in the test design phase can ensure that operationally relevant scenarios and criteria are used. Conversely, having test and safety experts participate early in the development of AI models can guide developers to build with testability and transparency in mind.

On a larger scale, collaboration could take the form of joint research initiatives between government, academia, and industry to create open datasets and open-source tools for AI T&E, or cross-organization challenge problems to benchmark AI test techniques. Sharing lessons learned, including test cases where AI failed and how issues were fixed, will be vital so that the community can collectively progress.

Establish standards and guidelines for AI T&E. Given the novelty of AI in conventional software-reliant systems, there is a need to develop standardized frameworks and guidance for T&E. Government agencies and professional bodies should work on creating T&E protocols that address AI-specific aspects. This work should include standard definitions for levels of AI autonomy and the corresponding test requirements, guidelines on the minimum testing needed for deploying an AI-augmented system in safety-critical roles, or checklists for ethical risk assessment during T&E.

The development of an AI T&E practices drawing from frameworks like the NIST AI Risk Management Framework and DoD's AI principles can also help practitioners navigate the evaluation process. Standardized test suites and benchmarks for different classes of AI (*e.g.,* computer vision, language, and decision-making systems) could also provide baselines for comparison and improvement. By formalizing such standards, organizations will have clearer targets for what adequate testing entails.

Integrate continuous monitoring and feedback mechanisms., T&E should be developed as an ongoing activity across the lifecycle of AI-augmented systems, not a one-time "rite of passage" to full-rate production decisions and/or operational fielding. Moreover, after an AI-augmented system is fielded, mechanisms should be in place to monitor its performance and collect operational data, which can feed back into updates or improvements. If the system's environment or requirements change, the AI will likely need re-testing or re-training.

Organizations should also institute periodic audit cycles for AI to ensure it remains within acceptable bounds. For example, an AI model's accuracy and bias using new data should be conducted periodically. In mission- and safety-critical applications, real-time monitoring dashboards can track AI outputs for anomalies (*e.g.,* a sudden spike in error rates) and alert operators or trigger fallbacks.

By coupling deployment with a "test-as-you-go" philosophy, any degradation in AI performance can be caught early and corrected. This continuous evaluation aligns with the concept of "campaigns of learning" in deployment, wherein each use of the AI provides data to refine its future performance [7]. Such feedback loops will be essential as AI-augmented systems operate over long durations or interact with changing adversarial behaviors.

5. Concluding Remarks

Generative AI is poised to revolutionize not only how we develop software, but also how we conduct software test and evaluation. It

has the potential to enable faster prototyping, more exhaustive testing, and continuous quality assurance, which can significantly improve software reliability and time-to-deployment. However, the same technology introduces a new realm of complexity in ensuring that AI-augmented systems are trustworthy, unbiased, and compliant with requirements and policies. The performance evaluation of AI-augmented systems must expand beyond correctness to include issues like bias, robustness, and uncertainty.

This paper highlighted both the opportunities (e.g., automation, continuous testing, and smarter test design) and the challenges (e.g., independent AI evaluation, hallucinations, and validation complexity) that generative AI brings to T&E. Key concerns like reliability and ethical compliance demand attention, but they are surmountable with concerted effort in developing new test methods and collaboration across the AI and testing communities.

Moving forward, organizations should champion research, education, and partnerships that strengthen our collective ability to test AI-augmented systems rigorously. By doing so, we ensure that as software engineering enters this AI-augmented era, our verification and validation practices evolve in tandem, thereby ensuring that AI-augmented, software-reliant systems are not only innovative and efficient, but also safe, fair, and dependable for all users.

ACKNOWLEDGMENTS

This material is based upon work supported, in whole or in part, by the U.S. Department of Defense, Director, Operational Test and Evaluation (DOT&E) through the Office of the Assistant Secretary of Defense for Research and Engineering (ASD(R&E)) under Contract HQ0034-24-D-0023.

The initial draft of this manuscript was crafted with help from OpenAI's *Deep Research* based on a transcript of our webinar from [1]. We then thoroughly revised the initial draft and accept responsibility for the veracity and correctness of all material. We would like to thank Dr. Douglas C. Schmidt (Dean of Computing, Data Science & Physics at William & Mary) and Ms. Alexis Bonnell (Chief Information Officer at the Air Force Research Lab) for their contributions to the webinar and subsequent discussions that helped shape key aspects of this paper.

REFERENCES

[1] Acquisition Innovation Research Council Panel on Generative AI in the Acquisition Lifecycle. Retrieved from: https://www.youtube.com/watch?v=ZlCc94w-2bY

[2] Akshay Sekar Chandrasekaran. 2024. Harnessing the Power of Generative Artificial Intelligence (GenAI) in Governance, Risk Management, and Compliance (GRC). (2024).

[3] Jaganmohan Chandrasekaran, Erin Lanus, Tyler Cody, Laura J Freeman, Raghu N Kacker, MS Raunak, and D Richard Kuhn. 2024. Leveraging Combinatorial Coverage in the Machine Learning Product Lifecycle. Computer 57, 7 (2024), 16–26.

[4] Youngwon Choi, Wenxi Yu, Mahesh B Nagarajan, Pangyu Teng, Jonathan G Goldin, Steven S Raman, Dieter R Enzmann, Grace Hyun J Kim, and Matthew S Brown. 2023. Translating AI to clinical practice: overcoming data shift with explainability. Radiographics 43, 5 (2023), e220105.

[5] Tyler Cody, Erin Lanus, Daniel D Doyle, and Laura Freeman. 2022. Systematic training and testing for machine learning using combinatorial interaction testing. In 2022 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), IEEE, 102–109.

[6] Department of Defense, Chief Data and Artificial Intelligence Office, Test and Evaluation of Artificial Intelligence Models. Retrieved from: https://www.ai.mil/Portals/137/Documents/Resources%20Page/Test%20and%20Evaluation%20of%20Artificial%20Intelligence%20Models%20Framework.pdf

[7] Department of Defense, Data, Analytics, and Artificial Intelligence Adoption Strategy. Retrieved from: https://media.defense.gov/2023/Nov/02/2003333300/-1/-1/1/DOD_DATA_ANALYTICS_AI_ADOPTION_STRATEGY.PDF

[8] Department of Defense, Responsible Artificial Intelligence Strategy and Implementation Pathway. Retrieved from: https://media.defense.gov/2022/Jun/22/2003022604/-1/-1/0/Department-of-Defense-Responsible-Artificial-Intelligence-Strategy-and-Implementation-Pathway.PDF

[9] Yinlin Deng, Chunqiu Steven Xia, Chenyuan Yang, Shizhuo Dylan Zhang, Shujing Yang, and Lingming Zhang. 2024. Large language models are edge-case generators: Crafting unusual programs for fuzzing deep learning libraries. In Proceedings of the 46th IEEE/ACM international conference on software engineering, 1–13.

[10] Pradeep Dogga, Karthik Narasimhan, Anirudh Sivaraman, Shiv Saini, George Varghese, and Ravi Netravali. 2022. Revelio: Ml-generated debugging queries for finding root causes in distributed systems. Proceedings of Machine Learning and Systems 4, (2022), 601–622.

[11] Vahid Garousi, Nithin Joy, Alper Buğra Keleş, Sevde Değirmenci, Ece Özdemir, and Ryan Zarringhalami. 2024. AI-powered test automation tools: A systematic review and empirical evaluation. arXiv preprint arXiv:2409.00411 (2024).

[12] A Gupta and F Mohummed. 2023. Role of generative AI in augmented reality (AR) and virtual reality (VR) application testing. JAIMLD 1, (2023), 426–30.

[13] Yili Hong, Jiayi Lian, Li Xu, Jie Min, Yueyao Wang, Laura J Freeman, and Xinwei Deng. 2023. Statistical perspectives on reliability of artificial intelligence systems. Quality Engineering 35, 1 (2023), 56–78.

[14] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems 43, 2 (2025), 1–55.

[15] SA Mohaiminul Islam, MD Shadikul Bari, and Ankur Sarkar. 2024. Transforming Software Testing in the US: Generative AI Models for Realistic User Simulation. Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023 6, 1 (2024), 635–659.

[16] Erin Lanus, Ivan Hernandez, Adam Dachowicz, Laura J Freeman, Melanie Grande, Andrew Lang, Jitesh H Panchal, Anthony Patrick, and Scott Welch. 2021. Test and evaluation framework for multi-agent systems of autonomous intelligent agents. In 2021 16th International Conference of System of Systems Engineering (SoSE), IEEE, 203–209.

[17] Erin Lanus, Brian Lee, Luis Pol, Daniel Sobien, Justin Kauffman, and Laura J Freeman. 2024. Coverage for Identifying Critical Metadata in Machine Learning Operating Envelopes. In 2024 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), IEEE, 217–226.

[18] Jiayi Lian, Laura Freeman, Yili Hong, and Xinwei Deng. 2021. Robustness with respect to class imbalance in artificial intelligence classification algorithms. Journal of Quality Technology 53, 5 (2021), 505–525.

[19] Negar Maleki, Balaji Padmanabhan, and Kaushik Dutta. 2024. AI hallucinations: a misnomer worth clarifying. In 2024 IEEE conference on artificial intelligence (CAI), IEEE, 133–138.

[20] Filippo Ricca, Alessandro Marchetto, and Andrea Stocco. 2021. Ai-based test automation: A grey literature analysis. In 2021 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), IEEE, 263–270.

[21] Padmaksha Roy, Jaganmohan Chandrasekaran, Erin Lanus, Laura Freeman, and Jeremy Werner. 2023. A Survey of Data Security: Practices from Cybersecurity and Challenges of Machine Learning. arXiv preprint arXiv:2310.04513 (2023).

[22] Agus Sudjianto and Srinivas Neppalli. 2024. Human-Calibrated Automated Testing and Validation of Generative Language Models: An Overview. Available at SSRN (2024).

[23] Jules White, Sam Hays, Quchen Fu, Jesse Spencer-Smith, and Douglas C Schmidt. 2024. Chatgpt prompt patterns for improving code quality, refactoring, requirements elicitation, and software design. In Generative ai for effective software development. Springer, 71–108.