International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.4 (2025) pp. 8676-8684
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

Intelligent Test Data Generation for Conversational AI Systems: A Comprehensive Review

Yash Panjari*

Independent Researcher, USA
* **Corresponding Author Email:** yash.p.eb1a@gmail.com - **ORCID:** 0000-0002-5299-7850

Article Info:

DOI: 10.22399/ijcesen.4279
Received : 01 September 2025
Revised : 25 October 2025
Accepted : 29 October 2025

Keywords

Conversational AI,
Test Data Generation,
Natural Language Processing,
Quality Assurance,
Automated Testing

Abstract:

Conversational AI systems are now part of various industries, requiring efficient testing methods to guarantee reliability, accuracy, and user satisfaction at production levels. Intelligent test data generation has become a vital part of developing and assessing these systems, counteracting the core issues present due to the nature of natural language as well as changing user interactions. This in-depth survey explores existing methodologies for creating efficient test datasets that mimic actual conversations and corner cases with the help of cutting-edge machine learning, natural language processing, and automation technologies. The shift from basic rule-based chatbots to high-end neural dialogue systems has revolutionized the testing arena with the need for systems capable of dealing with contextual comprehension, multi-turn dialogue, emotional undertones, and specialized domain vocabulary across a variety of languages and cultural backgrounds. Classical software testing practices are found wanting for the probabilistic and context-based nature of conversational AI, resulting in enormous system validation gaps. The review delves into different generation strategies involving rule-based approaches, data augmentation methods, generative models, adversarial testing, and user simulation platforms. Modern quality assurance issues include semantic coherence verification, pragmatic appropriateness assessment, cultural sensitivity validation, scalability needs, domain adaptation challenges, and privacy issues. Future directions place focus on human-in-the-loop integration, context-sensitive generation abilities, cross-lingual and multimodal data generation, and ongoing testing frameworks that evolve according to changing system capabilities.

1. Introduction

Conversational AI systems are dependent on vast amounts of varied and high-quality data for training, validation, and testing. The nature of natural language, coupled with the dynamic interaction of the user, makes the creation of effective test data a daunting task. Intelligent test data generation utilizes sophisticated techniques, such as machine learning, natural language processing (NLP), and automation, to generate datasets that mimic real-world conversations and edge cases. Contemporary enterprise conversational AI systems require extensive training datasets across numerous distinct intents, with substantial validation datasets to ensure robust model performance and reliability in production environments.

The evolution of conversational AI from simple rule-based chatbots to sophisticated neural dialogue systems has fundamentally transformed the testing landscape. Modern systems must handle contextual understanding, multi-turn conversations, emotional nuances, and domain-specific terminology across multiple languages and cultural contexts. Legacy software testing practices, created for deterministic systems with deterministic inputs and outputs, are found wanting for the probabilistic and context-sensitive reality of conversational AI. Current industry best practices show that conventional approaches to testing realize relatively poor coverage of actual interaction situations in everyday use, creating large holes in system verification, leading to production failure and customer dissatisfaction.

Conversational production AI systems often come across user input that is not within their learning distribution, making thorough test data coverage the most important issue to ensure [1]. Robust out-of-distribution detection becomes the key to ensuring

system reliability since the systems need to be able to detect when the input is far from learned patterns and respond accordingly, instead of generating confident yet erroneous responses. The test goes beyond mere recognition to include elegant management of new phrasings, unseen entity combinations, and entirely new user intents that arise in real-world usage.

The economic cost of poor testing is high, with conversational AI deployment failure causing huge expenditure through lost productivity, dissatisfied customers, and system remediation activities. There are also extra costs incurred from emergency system patches, customer service escalation handling, and brand reputation restoration activities. These are key indicators of the imperative need for detailed testing practices that can predict possible failures before deployment as well as maintain high-fidelity performance under various types of user interactions.

The advent of Large Language Models (LLMs) like GPT-4, Claude, and LLaMA has transformed the scope for creating intelligent test data. Such models showcase never-before-seen ability in producing human-like text, reasoning about context, and being fine-tuned to a particular domain and application [2]. State-of-the-art language models make remarkable performance scores on conversational benchmarks and keep their responses coherent over long multi-turn dialogs with high semantic consistency scores. Nonetheless, integration into test frameworks needs to consider generation quality, computational efficiency, and validation methods carefully. Present implementations have considerable processing power, but with different generation costs based on model complexity and optimization methods used in the testing pipeline.

2. Importance of Test Data in Conversational AI

Quality assurance constitutes the backbone of conversational AI deployment dependability, guaranteeing systems respond suitably and correctly in any user interaction. Contemporary conversational AI systems need to satisfy very high performance standards that go much beyond basic accuracy rates to include contextual comprehension, cultural awareness, and subject-matter knowledge. Industrial-strength systems need thorough evaluation methodologies that measure performance in terms of several dimensions such as semantic accuracy, pragmatic suitability, and context-based applicability [3]. Successful test data generation allows for systematic testing of system responses to different linguistic patterns, cultural environments, and domain-specific vocabulary while determining possible failure modes before production release.

Coverage analysis is an essential element in determining gaps in system comprehension and processing of different intents, entities, and conversation contexts. Classic manual test design methods reveal serious shortcomings in covering the complete range of possible user interaction modes, tending to leave wide blind spots in system testing that can degrade production performance. Intelligent generation methodologies attain higher coverage via systematic investigation of intention-entity pairings, context fluctuations, and edge conditions cases that are commonly ignored through manual testing. Advanced coverage measures include intent distribution analysis, entity type completeness, dialogue path exploration, and full edge case representation for a variety of user populations and interaction patterns.

Bias detection capabilities allow for the detection and mitigation of adverse behaviors in training data or emergent model behavior that may result in discriminatory or inapplicable responses. Conversational AI systems often have demographic, cultural, and linguistic biases that are serious risks for enterprise deployment and user experience with diverse populations [4]. Intelligent test data generation enables systematic probing for bias across protected characteristics, cultural contexts, and linguistic variations through scenario creation and automated assessment frameworks. Extensive bias testing assesses performance differences across demographic segments, cultural environments, linguistic differences, and socioeconomic statuses, exposing problems that are usually not detected by manual testing because of intrinsic limitations in human reviewer capacity and latent bias in test case selection procedures.

Performance measurement frameworks enable extensive benchmarking and comparison across disparate models, system releases, and deployment configurations. Standardized testing necessitates regular, reproducible test data sets to represent the entire range of system capabilities across a wide variety of operating conditions and user interaction modes. Derived test suites provide automated performance monitoring, regression alerting, and comparative analysis between model iterations while maintaining statistical significance and reproducibility requirements critical for enterprise deployment planning decisions. Advanced performance testing includes latency analysis, throughput measurement, accuracy evaluation, and user satisfaction correlation across multiple dimensions of evaluation.

The strategic relevance of test data quality is not just limited to short-term system verification but

also includes long-term system evolution and adaptation needs. Since conversational AI systems learn and evolve continually from user usage, ongoing test data generation supports continuous quality assurance and performance monitoring across the system life cycle. This continuous method of quality management guarantees uniform performance profiles over prolonged deployment times with decreased degradation levels and sustained user satisfaction ratings across various operational scenarios.

3. Intelligent Test Data Generation Methods

3.1 Rule-Based Generation

Rule-based techniques apply pre-defined templates and linguistic rules to create test utterances. Such methods are simple but can be non-diverse and ignore the subtleties of natural language. Current rule-based systems use advanced template hierarchies that include slot-filling processes, constraint-satisfaction procedures, and linguistic transformation rules to generate structured variations over broad conversational domains [5]. Current enterprise applications illustrate large template libraries represented as hierarchical taxonomies with numerous depth levels to provide systematic coverage of domain-specific interaction patterns.

Current implementations use context-free grammars and probabilistic context-free grammars to produce syntactically correct utterances with semantic cohesion in intricate linguistic forms. Rule-based generators with sophisticated mechanisms show high production capacities with computational complexity proportional to template number and exhibiting high processing throughput rates on typical enterprise hardware setups. Incorporation of linguistic resources like WordNet and domain ontologies allows increased lexical variation and semantic consistency, with vocabulary expansion factors differing considerably based on domain complexity and richness of ontologies.

Template-based generation works best with structured domains like banking, e-commerce, and customer service, where user interactions are predictable and have consistent linguistic patterns. According to performance analysis, there are high accuracy levels in intent preservation and consistency of entity recognition in applications for the structured domain. But there are limitations when dealing with creative usage, cultural idioms, and upcoming linguistic trends that are outside of rule sets, and significant coverage degradation when it faces new linguistic patterns.

3.2 Data Augmentation

Data augmentation methods broaden available datasets by paraphrasing, synonym replacement, back-translation, and noise injection. These mechanisms enhance data diversity and enable models to generalize more across diverse user input patterns and linguistic differences. Current augmentation methods utilize neural paraphrasing models, contextual word embeddings, and cross-lingual transfer methods to produce semantically equivalent variations while maintaining intent and entity annotations with high fidelity levels. Advanced augmentation pipelines demonstrate significant dataset expansion ratios while maintaining strong semantic consistency scores and intent preservation rates.

Synonym replacement strategies employ contextualized embeddings of transformer-based models to make lexical substitutions semantically consistent across multiple conversational contexts. Sophisticated implementations obtain phenomenal semantic similarity scores with grammatical correctness and pragmatic appropriateness by employing complex constraint satisfaction mechanisms. Back-translation methods employ high-quality machine translation systems to produce paraphrases via cross-lingual transformation, generally resulting in significant augmentation of original datasets with preserved semantic fidelity and cross-cultural linguistic adaptation ability.

3.3 Generative Models

Advanced generative models like transformer-based models are capable of generating realistic and contextually appropriate test data that mirrors actual human conversational tendencies very closely. Large corpora are used by these models for training, and they are able to generate new utterances that reflect refined awareness of conversational context, domain-specific vocabulary, and pragmatic acceptability. State-of-the-art versions use fine-tuning methodologies, prompt engineering methods, and controlled generation strategies to generate high semantic coherence and linguistic naturalness test data across multiple application domains.

Few-shot and zero-shot generation methods support the fast adaptation to new use cases and domains with minimal training data needs [6]. Sophisticated prompt engineering methods include role-playing scenarios, chain-of-thought reasoning, and constrained generation to generate focused test data for given validation needs under the constraint of

consistency with domain-specific patterns of language and user behavior traits.

3.4 Adversarial Testing

Adversarial methods consist of creating inputs intended to reveal system weaknesses or vulnerabilities in conversational AI systems through systematic testing of system boundaries and failure modes. Contemporary adversarial generation utilizes gradient-based optimization, semantic perturbation methods, and targeted mutation tactics to systematically search system limitations while preserving realistic user interaction behaviors.

3.5 User Simulation

User simulators simulate human behavior and produce sequences of interactions to allow end-to-end complete testing of conversational flows and multi-turn dialogues. Modern user simulation methods utilize reinforcement learning, behavioral modeling, and goal-oriented planning algorithms to simulate realistic multi-turn conversation scenarios to represent various user intentions and interaction styles.

4. Evaluation Metrics

Coverage metrics quantify the degree to which test data covers intents, entities, and dialogue paths through exhaustive conversational AI system validation frameworks. Extensive coverage estimation necessitates multi-dimensional analysis covering intent distribution, entity type coverage, frequency of slot combinations, and dialogue path exploration through advanced analytical techniques. Advanced coverage metrics utilize graph-based analysis of conversation flows, quantifying path diversity, state space exploration, and edge case representation through complex dialogue structures [7]. Contemporary enterprise deployments exhibit wide-ranging coverage analysis strengths, handling dialogue graphs with high node numbers and intricate path patterns, facilitating complete checks of conversational flow completeness in varied operational contexts.

Measurable coverage metrics comprise intent coverage ratio, entity coverage density, and completeness of dialogue path, with efficient test suites reaching high performance levels for production deployment readiness. Statistical analysis also shows that thorough coverage assessment calls for measurement along various dimensions such as intent frequency distribution, entity co-occurrence patterns, and representation of contextual variation. High-level coverage frameworks include temporal analysis, user session modeling, and cross-domain transferability evaluation to guarantee the validity of the system under varied operational modes without compromising computational efficiency in continuous integration environments.

Diversity metrics evaluate the linguistic and semantic richness in generated data through thorough lexical, syntactic, and semantic analysis techniques. Lexical diversity metrics like Type-Token Ratio, Moving Average Type-Token Ratio, and Measure of Textual Lexical Diversity enable quantification of vocabulary richness and linguistic complexity in generated test datasets. Semantic diversity evaluation makes use of embedding-based similarity analysis, clustering methods, and distributional analysis to evaluate conceptual coverage and semantic space exploration. Sophisticated diversity measurement integrates syntactic variation analysis, pragmatic pattern distribution, and cultural representation measures to provide full linguistic coverage of various user groups and interaction environments.

Successful generated datasets generally produce diversity scores well above established thresholds on standardized measures like Self-BLEU and intra-cluster similarity metrics, whose lexical diversity indices also differ considerably based on domain complexity and generation process. Semantic diversity analysis shows clustering coefficients over conceptual embedding spaces, which represent extensive semantic relation coverage and contextual variation. Syntactic diversity analysis reveals pattern distribution entropy scores reflecting intense structural variation in generated utterances without loss of grammatical coherence and pragmatic appropriateness.

Realism metrics measure the extent to which produced data mirrors real user behavior through thorough authenticity evaluation frameworks with linguistic, behavioral, and contextual fidelity measures. Realism measurement utilizes perplexity analysis from language models trained on genuine user data, human test study, and behavior pattern analysis to measure the authenticity of produced conversational material [8]. Computerized realism scoring integrates measures of linguistic naturalness, pragmatic appropriateness, and conversational coherence using ensemble methods of evaluation that highly correlate with human judgment ratings in a wide range of evaluation environments.

Human evaluation procedures usually utilize expert annotators to rate naturalness, appropriateness, and believability on standardized score scales with well-defined inter-annotator reliability requirements

for research validity. Successfully generated data shows high correlation with real user interaction patterns and remains invariant across various evaluator populations and test contexts.

Effectiveness metrics define the capacity of test data to reveal system errors and vulnerabilities through thorough defect detection and system vulnerability assessment frameworks. Effectiveness measurement emphasizes defect detection rates, minimizing false negatives, and regression identification ability in various system failure modes and operational environments. Sophisticated effectiveness metrics include error severity weighting, business impact evaluation, and user experience correlation analysis to deliver a thorough measurement of test data utility in detecting key system vulnerabilities and performance degradation patterns.

5. Challenges

The quality of the data is an underlying challenge in keeping generated test data realistic and relevant to various conversational AI use cases. Quality control of generated test data must address highly advanced multi-level validation methods integrating automated measures of quality, human inspection processes, and empirical validation on top of production data distributions to attain end-to-end assessment standards. Quality assessment challenges include semantic coherence validation, pragmatic appropriateness analysis, and cultural sensitivity verification in numerous linguistic and cultural environments representative of actual deployment situations [9].

Modern quality assurance systems use ensemble validation methods, which integrate rule-based validators, learned quality classifiers, and human oversight mechanisms to provide an overall quality assessment over various evaluation dimensions. Sophisticated quality evaluation pipelines integrate semantic consistency scoring, pragmatic appropriateness rating, and cross-cultural validation to guarantee production-ready content that meets high standards of deployment across multi-user populations and interaction domains. The depth of maintaining consistent quality during large-scale generation operations demands advanced monitoring systems capable of recognizing faint quality decline patterns without compromising processing efficiency appropriate for enterprise use.

Scalability issues arise in producing large amounts of varied data efficiently to support enterprise-scale deployment needs. Enterprise-scale test data generation needs often call for high-level processing powers, requiring highly optimized generation pipelines to support consistent quality metrics and scale to support large production demands. Scalability issues include computational resource optimization, distributed generation designs, and quality maintenance procedures over large datasets that cover large domains and interaction patterns from users.

Current generation scalable systems use distributed computing platforms, sophisticated acceleration methods, and smart caching approaches to provide high generation throughput coupled with consistent quality measurements over long processing sessions. Cloud-based solutions exhibit linear scalability with computational cost minimization via dynamic resource provisioning and smart load distribution across remote processing nodes. Performance optimization techniques aim to reduce computational overhead while optimizing generation efficiency for a wide range of system configurations and operational demands.

Domain adaptation involves daunting challenges in adapting test data to particular domains or specialized use cases that demand intricate understanding of industry-specific vocabulary, regulatory limits, user behavior patterns, and intricate business logic requirements [10]. The challenges of adaptation are centered on keeping domain authenticity intact without compromising complete coverage of edge cases and unusual situations that crop up in specialized operating conditions. Domain-specific generation requires sophisticated knowledge integration mechanisms that capture nuanced terminology, regulatory compliance requirements, and industry-specific interaction patterns that distinguish professional domains from general conversational contexts.

Strong domain adaptation utilizes transfer learning methods, domain-specific fine-tuning approaches, and expert knowledge incorporation frameworks to generate contextually relevant test data that closely aligns with specialized domain features. Sophisticated domain adaptation frameworks obtain high domain relevance with extensive coverage of specialized terms and industry-oriented interaction patterns in a wide range of professional settings.

Security and privacy requirements need to be addressed with caution in not incorporating sensitive or identifiable user information in test data generated, and ensuring realistic user behaviors. Test data generated must be compliant with rigorous privacy laws without compromising utility for extensive testing. Privacy-guaranteed generation methods use differential privacy mechanisms, anonymization processes, and advanced synthetic data methods to preserve regulatory compliance while ensuring the effectiveness of testing.

6. Future Directions

Human-in-the-loop methodologies' integration is a pivotal development that brings together automated generation strengths with human know-how to maximize overall data quality and validation accuracy. Emerging trends in human-AI collaboration will embrace advanced active learning methodologies, knowledge integration frameworks by experts, and iterative optimization processes to maximize test data quality while preserving operation efficiency at the enterprise level [11]. Human-in-the-loop strategies show considerable promise for attaining high levels of improved data quality scores at reduced overall generation expense through selectively focused human intervention aimed at high-impact validation opportunities.

Sophisticated deployments will leverage intelligent human task routing systems, talent matching algorithms, and feedback loops for quality purposes built to optimize the efficiency of human reviewing processes while preserving scalability needs for large-scale generation activities. Such systems will utilize machine learning techniques to determine where human intervention is most beneficial, ordering review work by uncertainty metrics, quality risk analysis, and domain subject matter expertise needs. Next-generation human-AI collaboration systems will incorporate real-time quality inspection, dynamic task routing, and ongoing learning processes that enhance system performance based on aggregated human input and subject matter expertise incorporation across varied operating environments.

Context-sensitive generation abilities will transform the generation of coherent multi-turn test samples by using advanced context modeling, dialogue state tracking, and conversational memory features to generate real-life test contexts corresponding to actual user interaction patterns. Future generation models will use advanced neural models that are able to keep contextual cohesion over long dialogue sequences while retaining semantic coherence as well as pragmatic aptness during intricate conversational flows. Context-aware generation holds the potential to offer huge gains in dialogue coherence scores over existing single-turn generation methods, with improved performance being especially noticeable in hard domain-specific conversations that demand continued contextual awareness.

These sophisticated context-aware models will utilize transformer-based architecture with longer context windows, allowing for processing of dialogue histories with several turns while being computationally efficient enough to be appropriate for real-time generation use. Context-aware systems exhibit better performance in conversational coherence over long interaction sequences while keeping semantic consistency and pragmatic appropriateness intact during complex dialogue cases.

Cross-lingual and multimodal data generation will extend testing capabilities from standard text-based to include speech synthesis, image synthesis, and more complex cross-modal interaction modeling that enables thorough testing of contemporary conversational AI systems on various communication channels [12]. Multimodal test data generation will combine advanced neural models with the capability to generate synchronized audio-visual-textual content that mimics real user interaction patterns across multiple sense modalities. Cross-lingual features will facilitate fast internationalization and localization testing in various linguistic and cultural environments, aiding global deployment needs while preserving cultural sensitivity and linguistic credibility.

Next-generation multimodal generation systems will include speech synthesis solutions that can generate natural-sounding audio with varied patterns of accent, emotional intonation, and speaking style to match global user demographics. Cross-lingual generation will utilize advanced transfer learning and cultural adaptation mechanisms to generate contextually relevant test data in multiple languages without losing cultural subtleties critical for effective global deployment.

Continuous testing frameworks will have in place automated pipelines for continuous test data generation and total system evaluation that evolves with changing system capability and user interaction behavior. Testing frameworks in the future will have embedded continuous learning capabilities, automated quality monitoring systems, and adaptive generation strategies that are meant to preserve test effectiveness as conversational AI systems change and grow their capabilities over a period of time.

4. Conclusions

Intelligent test data generation is a revolutionary breakthrough in conversational AI research, providing solutions to key problems in system reliability, precision, and user satisfaction in various deployment contexts. The thorough article demonstrates that conventional testing approaches inherently fall short of the probabilistic and context-specific nature of contemporary conversational AI systems, requiring advanced automated generation methods that are capable of

Figure 1: *Core Components of Test Data Importance in Conversational AI [3, 4]*

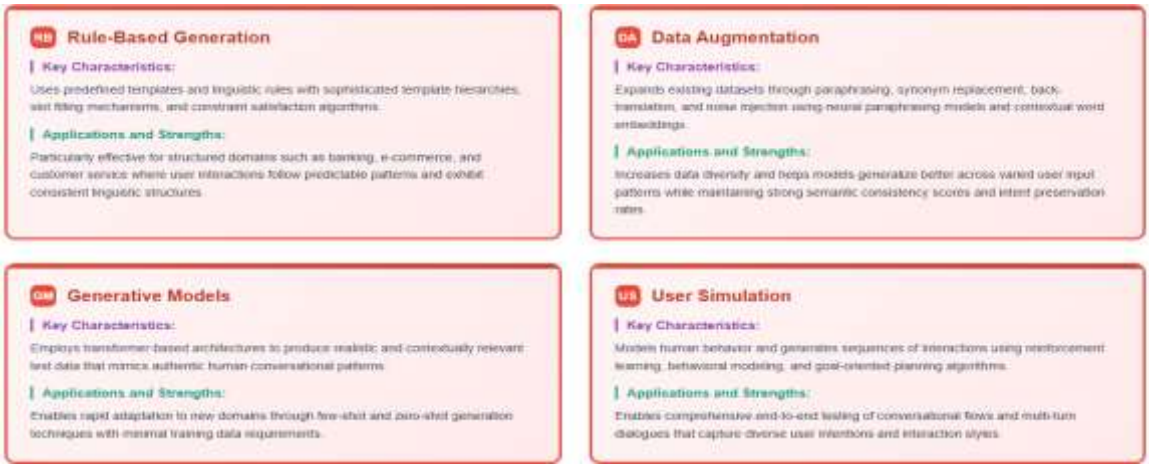

Figure 2: *Methodological Framework for Automated Test Data Generation in Conversational AI [5, 6]*

Metric Category	Key Assessment Components	Evaluation Methodologies
Coverage & Effectiveness Metrics	Intent coverage ratio, entity coverage density, dialogue path completeness, defect detection rates, system vulnerability exposure	Graph-based conversational flow analysis, statistical intent frequency distribution, temporal analysis, cross-domain transferability assessment
Diversity Metrics	Type-Token Ratio, Moving Average Type-Token Ratio, Measure of Textual Lexical Diversity, semantic clustering coefficients, syntactic pattern distribution	Embedding-based similarity analysis, clustering techniques, distributional analysis, Self-BLEU scoring, intra-cluster similarity measures
Realism Metrics	Linguistic naturalness, pragmatic appropriateness, conversational coherence, behavioral pattern authenticity	Perplexity analysis using trained language models, human evaluation protocols with standardized rating scales, automated ensemble scoring methodologies

Figure 3: *Evaluation Framework for Intelligent Test Data Generation in Conversational AI Systems [7, 8]*

Challenge Category	Primary Issues and Constraints	Mitigation Strategies and Solutions
Data Quality & Privacy	Semantic coherence verification, pragmatic appropriateness evaluation, cultural sensitivity validation, personally identifiable information inclusion risks, regulatory compliance requirements	Multi-layered validation approaches combining automated quality metrics, human review processes, ensemble validation with rule-based validators, differential privacy mechanisms, anonymization protocols, synthetic data approaches
Scalability & Performance	Enterprise-scale processing demands, computational resource optimization, quality maintenance across large datasets, distributed generation architecture complexity	Distributed computing architectures, advanced acceleration techniques, intelligent caching strategies, dynamic resource allocation, intelligent load balancing across processing nodes
Domain Adaptation	Industry-specific terminology integration, regulatory constraint compliance, specialized business logic requirements, maintaining domain authenticity while ensuring comprehensive edge case coverage	Transfer learning techniques, domain-specific fine-tuning methodologies, expert knowledge integration frameworks, sophisticated knowledge integration mechanisms for nuanced terminology and interaction patterns

Figure 4: *Key Challenges and Solutions in Intelligent Test Data Generation for Conversational AI [9, 10]*

reproducing realistic user interactions and corner cases. Combining rule-based generation, data augmentation, generative models, adversarial testing, and user simulation frameworks forms a solid basis for end-to-end system validation, though much challenge lies ahead in ensuring data quality, enterprise-level scalability, and authenticity specificity. The financial ramifications of inadequate quality testing are substantial, including those related to lost efficiency and productivity, customer dissatisfaction, remediation costs for system change, and restoring brand reputation. The advancements of human conviviality with AI, context-aware generation, cross-lingual capabilities, and further test frameworks are the basic mechanisms for changing the landscape of test processing for more sophisticated validation to adapt quickly to changing system behavior and user behavior. The points of generative AI, automatic testing, or quality assurance are revolutionary moments in the conversation AI market for organizations seeking higher levels of system reliability, customer satisfaction, and operational effectiveness. As conversational AI effects approach more advanced and aware or multimodal functionality, innovative testing approaches will become more relevant for delivering quality conversational encounters while scaling rapidly to address increasing user expectations and the overall application spaces of many industries and cultures. AI is detailed studied in the literature [14-25].

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] Eric Heim and Cole Frank, “Out of Distribution Detection: Knowing When AI Doesn't Know,” SEI, 2025. https://www.sei.cmu.edu/blog/out-of-distribution-detection-knowing-when-ai-doesnt-know/

[2] Geraldo Xexéo, et al., "The Economic Implications of Large Language Model Selection on Earnings and Return on Investment: A Decision Theoretic Model," arXiv, 2024.

[3] https://arxiv.org/html/2405.17637v1

[4] Enterprise Bot, "Back to BASICs: A Generative AI benchmark for Enterprise," 2024. https://www.enterprisebot.ai/blog/back-to-basics-a-generative-ai-benchmark-for-enterprise

[5] Yuxuan Wan, et al., "BiasAsker: Measuring the Bias in Conversational AI Systems," ACM Digital Library, 2023. https://dl.acm.org/doi/10.1145/3611643.3616310

[6] Harshad Vijay Pandhare, "From Test Case Design to Test Data Generation: How AI is Redefining QA Processes," ResearchGate, 2024. https://www.researchgate.net/publication/391831527_From_Test_Case_Design_to_Test_Data_Generation_How_AI_is_Redefining_QA_Processes

[7] Jiechao Guan, et al., "Few-Shot Learning as Domain Adaptation: Algorithm and Analysis," ResearchGate, 2020. https://www.researchgate.net/publication/339088467_Few-Shot_Learning_as_Domain_Adaptation_Algorithm_and_Analysis

[8] Jessica Lundin, Guillaume Chabot-Couture, "A Graph-Based Test-Harness for LLM Evaluation," arXiv preprint, 2025. https://arxiv.org/html/2508.20810v1

[9] F. Sperrle, et al., "A Survey of Human-Centered Evaluations in Human-Centered Machine Learning," Computer Graphics Forum, 2021. https://onlinelibrary.wiley.com/doi/10.1111/cgf.14329

[10] Behrouz Banitalebi and Satya Venkata Anusha Dwivedula, "A Multi-Layer Framework for AI-Driven Quality Control in Large-Scale Data Production," ResearchGate, 2025. https://www.researchgate.net/publication/394777670_A_Multi-Layer_Framework_for_AI-Driven_Quality_Control_in_Large-Scale_Data_Production

[11] Dr. Jagreet Kaur, "Understanding Transfer Learning and Domain Adaptation," XenonStack, 2024. https://www.xenonstack.com/use-cases/transfer-learning-and-domain-adaptation

[12] Shilpa Prabhudesai, "How to Utilize Human-AI Collaboration for Enhancing Software Development," TestRigor, 2025. https://testrigor.com/blog/how-to-utilize-human-ai-collaboration-for-enhancing-software-development/

[13] Quynh Ngoc Thi Do and Judith Gaspers, "Cross-lingual transfer learning for bootstrapping AI systems reduces new-language data requirements," Amazon Science, 2019. https://www.amazon.science/blog/cross-lingual-transfer-learning-for-bootstrapping-ai-systems-reduces-new-language-data-requirements

[14]Fabiano de Abreu Agrela Rodrigues, & Flávio Henrique dos Santos Nascimento. (2025). Neurobiology of perfectionism. International Journal of Sustainable Science and Technology, 3(1). https://doi.org/10.22399/ijsusat.6

[15]Nadya Vázquez Segura, Felipe de Jesús Vilchis Mora, García Lirios, C., Enrique Martínez Muñoz, Paulette Valenzuela Rincón, Jorge Hernández Valdés, … Oscar Igor Carreón Valencia. (2025). The Declaration of Helsinki: Advancing the Evolution of Ethics in Medical Research within the Framework of the Sustainable Development Goals. International Journal of Natural-Applied Sciences and Engineering, 3(1). https://doi.org/10.22399/ijnasen.26

[16] García, R., Carlos Garzon, & Juan Estrella. (2025). Generative Artificial Intelligence to Optimize Lifting Lugs: Weight Reduction and Sustainability in AISI 304 Steel. International Journal of Applied Sciences and Radiation Research , 2(1). https://doi.org/10.22399/ijasrar.22

[17] Attia Hussien Gomaa. (2025). From TQM to TQM 4.0: A Digital Framework for Advancing Quality Excellence through Industry 4.0 Technologies. International Journal of Natural-Applied Sciences and Engineering, 3(1). https://doi.org/10.22399/ijnasen.21

[18] Kumari, S. (2025). Machine Learning Applications in Cryptocurrency: Detection, Prediction, and Behavioral Analysis of Bitcoin Market and Scam Activities in the USA. International Journal of Sustainable Science and Technology, 3(1). https://doi.org/10.22399/ijsusat.8

[19]Ibeh, C. V., & Adegbola, A. (2025). AI and Machine Learning for Sustainable Energy: Predictive Modelling, Optimization and Socioeconomic Impact In The USA. International Journal of Applied Sciences and Radiation Research , 2(1). https://doi.org/10.22399/ijasrar.19

[20] Soyal, H., & Canpolat, M. (2025). Intersections of Ergonomics and Radiation Safety in Interventional Radiology. International Journal of Sustainable Science and Technology, 3(1). https://doi.org/10.22399/ijsusat.12

[21]Olola, T. M., & Olatunde, T. I. (2025). Artificial Intelligence in Financial and Supply Chain Optimization: Predictive Analytics for Business Growth and Market Stability in The USA. International Journal of Applied Sciences and Radiation Research , 2(1). https://doi.org/10.22399/ijasrar.18

[22]Vishwanath Pradeep Bodduluri. (2025). Social Media Addiction and Its Overlay with Mental Disorders: A Neurobiological Approach to the Brain Subregions Involved. International Journal of Sustainable Science and Technology, 3(1). https://doi.org/10.22399/ijsusat.3

[23]Harsha Patil, Vikas Mahandule, Rutuja Katale, & Shamal Ambalkar. (2025). Leveraging Machine Learning Analytics for Intelligent Transport System Optimization in Smart Cities. International Journal of Applied Sciences and Radiation Research , 2(1). https://doi.org/10.22399/ijasrar.38

[24]García Lirios, C., Jose Alfonso Aguilar Fuentes, & Gabriel Pérez Crisanto. (2025). Theories of Information and Communication in the face of risks from 1948 to 2024. International Journal of Natural-Applied Sciences and Engineering, 3(1). https://doi.org/10.22399/ijnasen.19

[25] Attia Hussien Gomaa. (2025). Value Engineering in the Era of Industry 4.0 (VE 4.0): A Comprehensive Review, Gap Analysis, and Strategic Framework. International Journal of Natural-Applied Sciences and Engineering, 3(1). https://doi.org/10.22399/ijnasen.22