Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Online Evaluation of Conversational Agents Using Machine-Learned Metrics

Guneet Singh Kohli

Independent Researcher, USA

ARTICLE INFO

Received:03 Sept 2025

Revised:07 Oct 2025

Accepted:17 Oct 2025

ABSTRACT

The article discusses the new paradigm of metrics learned by machines to be used in the assessment of conversational agents in a real-time setting. With voice assistants and chatbots playing an ever-larger role in human interactions with computers in many domains, the classical evaluation techniques are fatally limited in their range, precision, and dynamism. Manual feedback mechanisms are primarily retrospective, introducing significant delays between defect identification and corrective action. Machine-learned evaluation methods, in contrast, utilize computational models that have been trained over historical interactions to automatically evaluate the satisfaction of users solely based on the content of a dialogue, conversation metadata, and behavioral cues. Hierarchical systems process information at multiple temporal resolutions. This enables both detailed and summary-level evaluations of dialogue quality. Applied in real time, these measures allow conversational systems to adapt dynamically during interactions. Remediation strategies may include response adjustment or escalation to human operators. Empirical studies explain that such approaches have a stronger association with user satisfaction, spot deficiencies in quality earlier, extrapolate to other areas of application, and continue to enhance their performance through online education. However, these benefits are accompanied by major challenges such as ensuring explainability, adapting to cultural contexts, evaluating multimodal interactions, modeling long-term engagement, preserving user privacy, and establishing standardized evaluation frameworks.

Keywords: Conversational Agents, Machine Learning Evaluation, Real-Time Quality Assessment, Adaptive Dialogue Systems, User Satisfaction Metrics

1. Introduction: The Challenge of Evaluating Conversational AI

The spread of conversational agents into the digital environment has fundamentally changed the way users interact with technology and thereby introduced the need to develop complex assessment approaches that can work efficiently at scale. The modern-day world already uses such systems: Alexa, Google Assistant and enterprise chatbots, to name a few, mediate millions of interactions between people and their computers each day, provide the main interface to access information, finish tasks, and even entertain them. This large-scale use has also cast doubt on the fundamental shortcomings of the evaluation methodologies that have been used to test the previous, simpler dialogue systems. According to in-depth reviews on the evolution of conversational AI, the increasing complexity and nuance of these interactions has surpassed the traditional models of evaluation [1]. Although the manual annotation processes remain invaluable due to deeper insights that are gleaned as a result, the process is simply too resource-demanding to be used as a production process in scenarios where millions of conversations are taking place in tandem. Binary task completion or latency metrics reduce the multidimensional nature of conversational quality to oversimplified signals, failing to capture factors such as coherence, engagement, and relevance.

The social and industrial implications of this challenge of evaluation are acute. The user dissatisfaction with conversational experiences results in a cycle of increasingly negative outcomes, the immediate failure to fulfill user expectations, and becoming disengaged with an experience, coupled with longer-term brand trust loss and platform abandonment. The challenge is whether such

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

indicators of dissatisfaction could be identified reliably, in situations where they may be more informative, e.g., in prolonged multi-dialogues when contextual information plays an essential role. Recent dialogue evaluation research has shown that heuristic metrics struggle to capture the complexity of modern multi-turn conversations [2]. The user would hardly express dissatisfaction directly and express it by some behavioral cues (interrupting an answer given by the system, rephrasing the request with other words, pausing before getting an answer, giving up a conversation). The implicit signals form a rich yet difficult dataset for quality assessment. Machine-learned metrics offer a promising solution to this assessment challenge, having computational models trained on previous interaction information with the goal that they could be used to automatically identify these patterns. Using the latest breakthroughs in natural language processing, sequence modeling, and behavior analysis, these systems seek to perform real-time evaluation of conversation quality without explicit interruption of a feedback system that breaks the natural flow of interaction. Such functionality allows reactive response options with the potential of reversing problematic conversations before the loss of users, a radical improvement over system operators who only have a global evaluation option that can only detect anomalies after the fact.

2. Limitations of Traditional Evaluation Methods

Conventional methods of assessing conversation systems bear large limitations when extended to new dialogue systems, especially at the stage when a system has to work with millions of interactions per day. Such approaches cannot give the same-time scalable assessment required to keep tabs on continuous improvement.

Although offering abundant qualitative input on conversation quality, the annotation of real human experts is basic because of its restrictive scope. Even under thoroughly designed rubrics, inter-rater reliability--the level of agreement among annotators--is in practice not sufficient to permit high-quality measurements. More importantly, manual assessment is done on a retroactive basis and thus, no chances of achieving live assessment of quality are possible. As seen in the analysis of the dialogue evaluation metrics extensively studied by Liu et al. [3], human evaluation can be seen as the best indicator of quality, but is, at the same time, unrealistic to use in production systems because of such limitations.

Interactive feedback systems, such as post-interaction surveys, are the real views of users, though they have some flaws associated with them, such as response biases, as the reviews are conditioned to scale to either end of end. The problem of low completion rates (often only 10–20% without active solicitation) introduces sampling biases that undermine the statistical validity of survey-based evaluation [4]. These methods also flatten and reify multidimensional interactions to a single rating value and neglect to model underlying moment-to-moment quality fluctuations that describe real conversations.

Task success measures capture transactional interactions but fail to account for exploratory or social exchanges, where user intentions may be ambiguous or shift dynamically. A study conducted by Venkatesh et al. underlined the gap between the traditional measures of task completion and reception under reality, as ~ 43% of conversations that the users have declared as satisfying would have been labeled improperly if measured using the traditional measures [4].

Built-in heuristics that rely on easily observable conversational signals are computationally efficient but lack versatility across different topics and often fail to capture subtle indicators of user engagement or frustration. In the meantime, automated metrics (BLEU, ROUGE, perplexity) exhibit no correlation with user satisfaction to distracted levels, as they only consider superficial linguistic

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

properties rather than conversational dynamics. Consistent with this observation is the fact that Liu et al. in their detailed meta-analysis of measures of automatic evaluation found that such measures normally demonstrate correlation coefficients with human assessments of dialogue quality of less than 0.4 [13].

The inherent weakness that ties these strategies together is that they are retrospective. Traditional methods also have a major lag in improving quality by acting post-conversation in nature as compared to a real-time evaluation mechanism. Such a time lag is a crippling restriction on the possibility of adaptive response to user requirements, a restriction that machine-learned evaluation measures especially seek to remove. Figure 1 summarizes these traditional limitations, showing how manual annotation, survey bias, and heuristic metrics fail to capture the dynamic, multi-turn nature of modern dialogues.

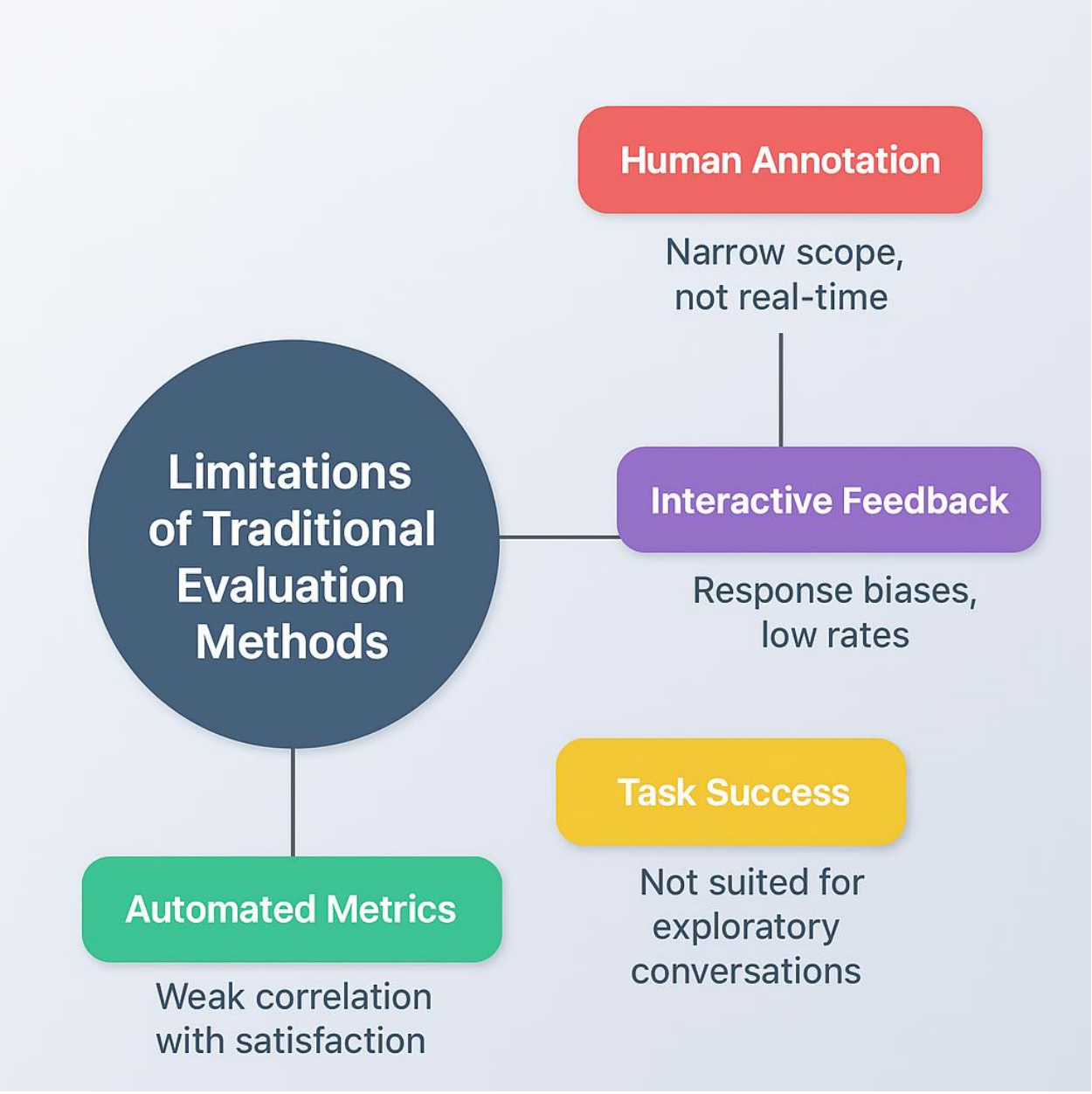

Figure 1: Limitations of Traditional Evaluation Methods for Conversational AI [3, 4]

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

3. Machine-Learned Metrics: Core Components and Architecture

Machine-learned metrics represent a significant advancement in the evaluation of AI conversational agents, where the heuristic metrics and signatures may be replaced with data models that learn with experience and exposure to new interaction patterns. Such systems can combine several streams of information to create overall judgments of quality that correlate with the real user satisfaction.

These models have the multidimensional nature of conversational quality represented by variables comprising the foundation of their models. The middle ground is the content of dialogue: the raw text of both user utterances and the system responses used to provide dialogue turn-taking and situational context in the dialogue. This is typically backend processed into embedding models, which compress linguistic content into dense vector representations carrying semantic meaning and not just the occurrence of keywords. In complement to this textual document, conversation metadata offers structural cues such as turn counts, measurements of response latency, ASR confidences, and entity recognitions existing to contextualize the conversation itself. As also exhibited in the article of Hancock et al. and their assessment of multi-channel quality figures, behavior patterns related to users, such as interruptions, barge-ins, reformulations, and avoidance sequences, are the most robust predictors of dissatisfaction, especially when considering the past user interaction, establishing a baseline of related interaction patterns [5]. In voice-based systems, acoustic characteristics (such as prosodic aspects, such as alteration in pitch, speech rate as well as volume changes) offer extra user-sentiment degrees that cannot be measured by text.

Modern architectures are hierarchical in design and thus process at different levels of time. This is generally initiated through turn-level processing, where single utterances and responses are processed, followed by processing by the sequence modeling component that operates on the temporal dynamics of many conversation turns and is typically performed by using recurrent neural networks or transformer-based networks. Recent work on unsupervised multi-turn evaluation has demonstrated that hierarchical models better capture coherence and dialogue flow [6]. The architecture usually ends up with user-level aggregation modules, which combine turn-level judgments into overall measures of good conversation quality whilst being able to distinguish any problematic exchanges.

Supervised learning algorithms are commonly used to train these models, on data sets where quality labels are acquired in one of several ways: through explicit user ratings or survey data, through implicit feedback in the form of conversational completion or repeated usage, through expert annotations of reasonably-sized subsets of conversations, or through proxy variables which are correlated with satisfaction, such as subscription retention or feature usage rates. The resultant models add a level of granular turn-by-turn measurement of quality as well as a macro conversation level measure of quality to allow this information to be used both to intervene in specific ways as well as make system-wide improvement plans. Figure 2 illustrates the overall architecture of these machine-learned metrics, highlighting how textual content, metadata, and behavioral cues are integrated through hierarchical modeling to produce real-time quality assessments.

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

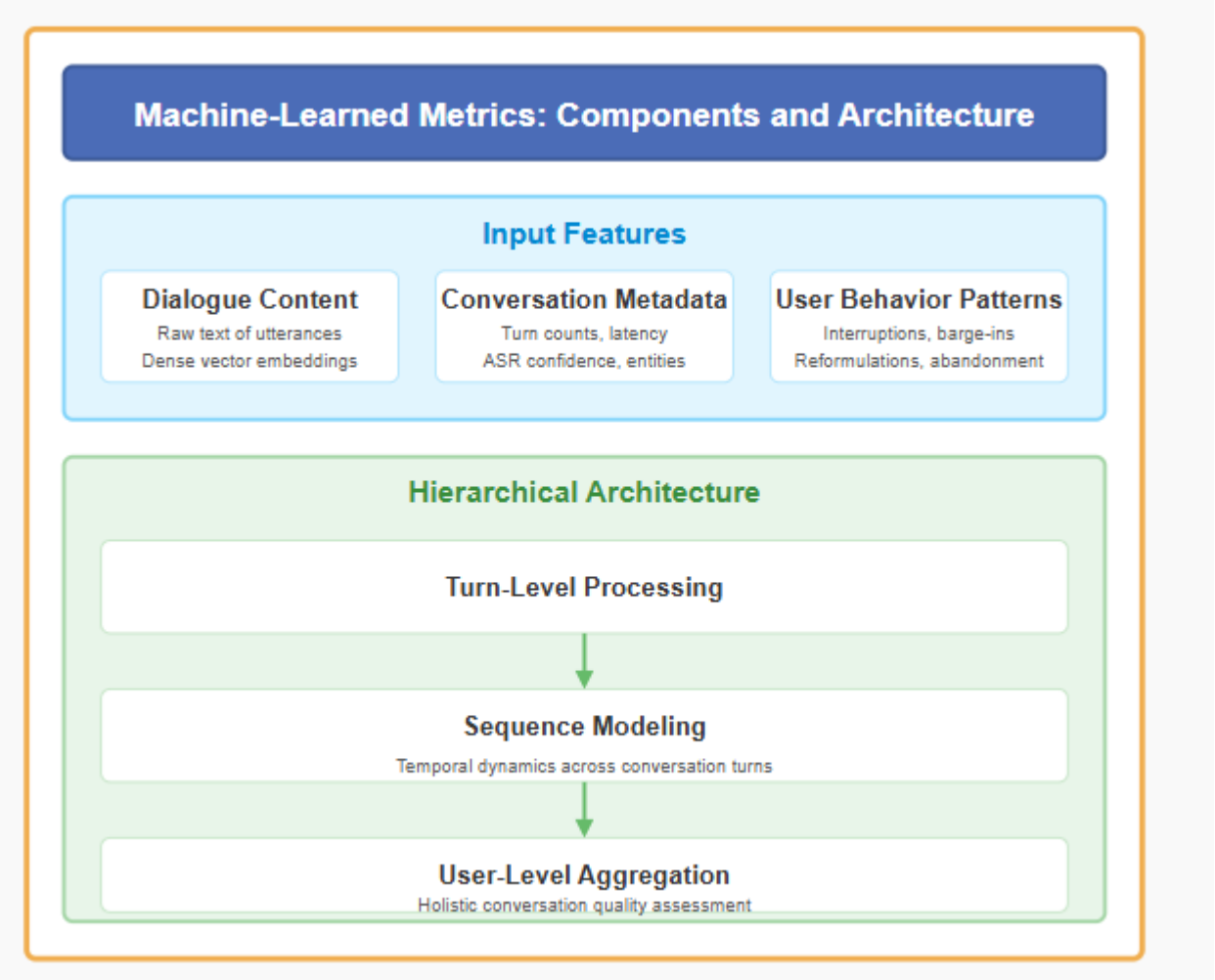

Figure 2: Machine-Learned Metrics: Core Components and Architecture [5, 6]

4. Real-Time Implementation: From Analysis to Action

The transformative potential of machine-learned metrics is fully realized when operating in a real-time setting, where conversational systems can respond dynamically to live interactions rather than retrospectively analyzing quality. This real-time capability represents the key advancement over traditional methods, though it requires sophisticated engineering to balance computational efficiency, evaluation accuracy, and response latency.

Effective implementations require structured processing pipelines beginning with efficient feature extraction that captures signals at strategic points during conversations. Input standardization ensures consistent scaling of diverse data types before passing them to the evaluation system, which updates quality predictions after each turn rather than waiting for conversation completion. Recent work highlights the importance of contextual uncertainty and confidence estimation in avoiding premature interventions during live dialogue [7]. These uncertainty metrics inform calibrated decision thresholds that trigger various intervention strategies based on prediction severity and confidence levels.

Once quality degradation is identified with sufficient confidence, systems can implement various corrective measures tailored to the specific issues detected. Response adaptation strategies

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

dynamically adjust generation parameters to make vague responses more specific, simplify responses when cognitive overload is detected, or modify tone when emotional mismatches occur. For deeper quality issues, conversational repair processes enable systems to identify potential misunderstandings and offer alternative interpretations. The CREPE framework (Conversational Repair Evaluation for Person-Centered Dialogue Systems) demonstrated significant improvements in subsequent user satisfaction when explicit repair strategies were applied selectively using high-confidence quality predictions, compared to non-adaptive baseline systems [8]. In cases where automated remediation proves insufficient, escalation protocols transfer challenging interactions to human operators, while personalization adjustments modify user models to prevent similar issues in future interactions.

Practical implementation considerations include latency management, where each evaluation process must complete quickly enough to influence the next generation process without introducing noticeable delay. Computational resources should be dynamically allocated based on conversation complexity and criticality, while robust A/B testing platforms enable systematic comparison of different evaluation models and intervention approaches. Most importantly, closed-loop feedback systems must track intervention outcomes to continuously refine quality predictions and remediation strategies. Organizations typically introduce these capabilities in phases, starting with offline analysis, progressing to shadow mode operation, and finally implementing full closed-loop systems. Figure 3 presents the real-time evaluation and intervention workflow, showing how live conversational signals feed into prediction, decision, and remediation stages to enhance user experience.

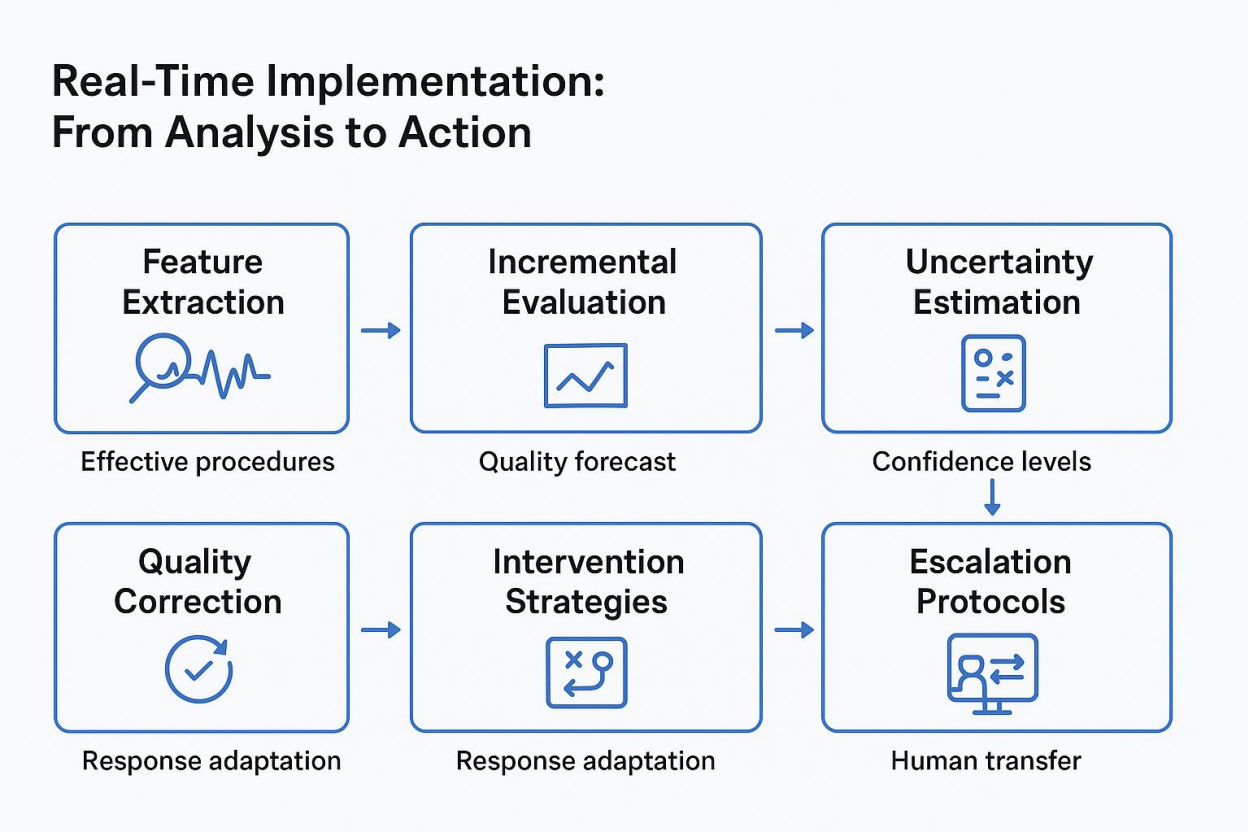

Figure 3: Real-Time Conversational Quality Evaluation and Intervention Workflow [7, 8]

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

5. Empirical Results: Evidence of Effectiveness

The effectiveness of machine-learned measurement of conversational assessment lies in the fact that it has been and continues to rely heavily on empirical studies and has proven to provide serious end-user benefits and alternatives to past assessment systems. The improvements in performance are evident in various aspects of quality and evaluation and business impact.

A core benefit is seen in the stronger correlation between machine-learned predictions and actual user satisfaction. Surveyed benchmarking studies consistently show that learned metrics achieve substantially higher correlations with human judgments (often exceeding 0.6, and in some cases approaching 0.8) compared to rule-based heuristics (around 0.3–0.5) and traditional n-gram metrics such as BLEU or ROUGE (typically below 0.3) [9]. Such correspondence as reported by Lowe in his comprehensive assessment scheme of dialogue systems denotes the success of learned metrics to appreciate the multidimensionality of conversation quality as determined by real users and not its solitary technical components [9]. Such enhanced correlation is directly applicable to enhanced quality evaluation in large-scale domains.

Probably the most significant operational aspect, the advanced models also have highly impressive abilities when it comes to early identification of developing quality challenges. Early corpus-based studies demonstrate that conversational breakdown patterns can be detected within the first few turns, enabling proactive intervention [10]. Their research shows that implementing such early detection systems significantly reduces conversation abandonment rates in commercial applications compared to systems without these capabilities.

These approaches demonstrate effectiveness in practical applications because they can be generalized across disciplines. While domain-specific tuning typically enhances performance, several studies suggest that models trained on diverse conversation types may retain significant predictive capabilities when applied to new domains. This potential for cross-domain generalization indicates that certain conversational quality signals might transcend specific use cases, though further research is needed to establish precise transfer learning effectiveness across different verticals.

A further opportunity is continuous improvement trajectories, in which systems with online learning mechanisms will gradually improve, as they accrue interaction data. During initial deployment periods, research suggests that quality assessment accuracy typically improves incrementally as models adapt to user behaviors and expectations, though the precise rate of improvement varies across implementations [10]. A study by Sardi et al. analyzing the performance evolution of conversational systems found that quality metrics typically show rapid improvement during initial deployment phases, with gains gradually plateauing as the models mature [10].

The potential business impact of these technical improvements may be reflected in various key performance indicators, including reduced escalation rates, higher task completion, improved user retention, and enhanced satisfaction scores. While promising, the precise magnitude of these benefits requires rigorous evaluation across different implementation contexts and should be validated through controlled studies before making definitive claims about business performance improvements. Figure 4 summarizes these empirical findings, comparing the correlation of machine-learned metrics with user satisfaction and highlighting improvements in early issue detection and operational efficiency.

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

Figure 4: Empirical Results: Evidence of Effectiveness [9, 10]

6. Future Directions and Open Challenges

Although machine-learned metrics represent a significant advancement in conversational AI evaluation, numerous research questions remain to be addressed to guide the field's future direction.

The most imperative methodological issues are explainability and transparency. Current evaluation models typically function as black boxes, producing quality scores without clear explanations for their judgments. According to IBM's comprehensive survey on AI interpretability, this opacity creates significant obstacles for system designers attempting to diagnose and fix problems [11]. Developing explainable assessment systems that can clearly articulate why a conversation was rated poorly would substantially enhance system improvement efforts and increase stakeholder confidence in automated evaluation.

To address these challenges, interdisciplinary collaboration across machine learning, linguistics, human-computer interaction, and ethics will be necessary to establish more holistic and reliable evaluation frameworks.

The possibility of cultural and contextual adaptation is also a major challenge since conversation norms and satisfaction indicators tend to differ significantly across cultures and demographics.

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

Studies unanimously show that assessment frameworks that are developed based mostly on Western and English-speaking interactions usually fail to work effectively when implemented for non-Western languages and cultural contexts. It is still difficult to develop sufficient evaluation systems to take into consideration such differences in an appropriate manner without further aggravating the same existing biases when trying to roll out conversational platforms on a worldwide scale.

There is more complexity yet to deal with as multimodal interfaces are gaining popularity, and conversational systems are increasingly multimodal, supporting visual output, and gestures as the means of interaction, in addition to text or voice. The existing evaluation methods tend to evaluate these modalities individually instead of measuring the interactions that occur across modalities. The establishment of a framework that would be able to measure quality in a holistic way by using numerous synchronized modalities is an important line of research.

Another interesting direction is long-term engagement modeling. Current models are effective at assessing discrete dialogues. However, they lack the ability to evaluate how these individual interactions contribute to lasting user relationships. A study conducted by Zhang et. al. involving trustworthy human-AI conversation enforces the fact that meaning and trust help in maintaining a fruitful long term interaction between the user and the conversational systems [12]. Such a strategy will lead to a more detailed explanation of the relationship between conversations and user retention as well as platform loyalty in the long-term positioning.

As users become keen on data privacy, privacy-preserving evaluation techniques are going to take a center stage. Gathering the data on conversation in order to perform a robust evaluation, respecting the privacy of the users would entail careful considerations in realms of data minimization, anonymization, even the application of federated learning methods that store sensitive conversation data on user devices but still allow the improvement of the models.

Lastly, standardization would go a long way in spurring developments on the entire front. In contrast to other research fields like computer vision and machine translation, conversational AI metrics do not have universally agreed upon standards and methodologies that can allow a useful comparison between methods and systems. The formulation of such standards would make innovation take place much faster and the quality goals in commercial applications much more specific. Ongoing shared tasks such as the Dialogue State Tracking Challenge (DSTC) [14] and the ConvAI competitions [15] provide promising venues for benchmarking and standardizing evaluation methods, and their continued expansion will be critical for advancing the field.

Conclusion

The development of machine-learned measures for evaluating conversational agents represents a significant shift in dialogue system assessment. These methods allow adaptive management of conversation by going beyond retrospective approaches to methods of evaluation that are dynamic and real time so that they can respond promptly to user needs as they arise. Combining a broader variety of input features of mutual linguistic quality content, behavioral signals, and contextual metadata into a single, unified assessment model, however, produces more comprehensive measurements of quality that are reflective of the measures that a user actually experiences.

Empirical outcomes demonstrate better correlation to satisfaction, early issue detection, and cross-domain applicability. However, significant challenges remain. Explainability and transparency present major methodological hurdles. Cultural adaptation across different regions and languages requires further investigation. Multimodal interaction evaluation needs substantial development. Longitudinal engagement modeling demands deeper understanding of user relationships over time. Additionally, privacy preservation remains a critical area requiring extensive research. With

conversational agents increasingly being adopted across healthcare, education, and enterprise sectors, robust evaluation frameworks must be developed to ensure user trust and engagement. Future research in this area will likely involve interdisciplinary efforts that integrate aspects of linguistics, psychology, and human-computer interaction.

These collaborative approaches will help create responsive and reliable conversational experiences that can continually adapt to user needs and expectations. Future work should prioritize standardized benchmarks, cross-lingual evaluation methods, and explainable models to facilitate reproducibility and fair comparison across conversational AI systems.

References

[1] Y. Deng, et al., "Proactive Conversational AI: A Comprehensive Survey of Advancements and Opportunities," ACM Trans. Inf. Syst., vol. 43, no. 3, 2025. doi: 10.1145/3715097

[2] S. Mehri and E. Eskénazi, "USR: An Unsupervised and Reference-Free Evaluation Metric for Dialog Generation," in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL), 2020, pp. 681–707. doi: 10.18653/v1/2020.acl-main.563

[3] A. C. Curry, H. Hastie, and V. Rieser, "A Review of Evaluation Techniques for Social Dialogue Systems," arXiv:1709.04409, 2017. [Online]. Available: https://arxiv.org/abs/1709.04409

[4] A. Venkatesh, et al., "On Evaluating and Comparing Conversational Agents," in Proc. NeurIPS Workshop on Conversational AI, 2017. [Online]. Available: https://assets.amazon.science/56/82/4385d9b74a73b9d23dd49bcdefcb/on-evaluating-and-comparing-conversational-agents.pdf

[5] B. Hancock, et al., "Learning from Dialogue after Deployment: Feed Yourself, Chatbot!," arXiv:1901.05415, 2019. [Online]. Available: https://arxiv.org/abs/1901.05415

[6] S. Mehri and E. Eskénazi, "Unsupervised Evaluation of Interactive Dialog with DialoGPT," in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol. (NAACL-HLT), 2021, pp. 3464–3479. doi: 10.18653/v1/2021.naacl-main.235.

[7] S. Ghazarian, S. Poria, E. Hovy, and A. Galstyan, "What Makes Good In-Context Examples for GPT-3?," arXiv:2201.12886, 2022. [Online]. Available: https://arxiv.org/abs/2201.12886

[8] E. Alghamdi, et al., "System and User Strategies to Repair Conversational Breakdowns of Spoken Dialogue Systems: A Scoping Review," ACM Trans. Interact. Intell. Syst., 2024. doi: 10.1145/3640794.3665558.

[9] J. Deriu, et al., "Survey on evaluation methods for dialogue systems," Artif. Intell. Rev., vol. 54, pp. 755–810, 2021. doi: 10.1007/s10462-020-09866-x.

[10] R. Lowe, N. Pow, I. V. Serban, and J. Pineau, "The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems," in Proc. SIGDIAL, 2015, pp. 285–294. doi: 10.18653/v1/W15-4640.

[11] A. McGrath and A. Jonker, "What is AI interpretability?," IBM Think, 2024. [Online]. Available: https://www.ibm.com/think/topics/interpretability

[12] G. Zhang, X. Yao, and X. Xiao, "On Modeling Long-Term User Engagement from Stochastic Feedback," arXiv:2302.06101, 2023. [Online]. Available: https://arxiv.org/abs/2302.06101

[13] C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue

Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

Response Generation," in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2016, pp. 2122–2132. doi: 10.18653/v1/D16-1230.

[14] J. Williams, A. Raux, D. Ramachandran, and A. Black, "The Dialogue State Tracking Challenge," *Proc. SIGDIAL*, 2013, pp. 404–413. doi: 10.18653/v1/W13-4065.

[15] V. Serban, et al., "A Deep Reinforcement Learning Chatbot (Short Version)," *NeurIPS Conversational Intelligence Challenge (ConvAI)*, 2017. arXiv:1709.02349. [Online]. Available: https://arxiv.org/abs/1709.02349