TPM Vol. 32, No. S7, 2025
ISSN: 1972-6325
https://www.tpmap.org/

Open Access

SYNTHETIC AUDIO DATASETS FOR EVALUATING CONVERSATIONAL AI SYSTEMS

GUNEET SINGH KOHLI
INDEPENDENT RESEARCHER

Abstract
This article examines the evolving landscape of conversational AI evaluation through synthetic audio datasets. Traditional evaluation methods relying on human-graded interactions face significant limitations in scalability, coverage, and resource efficiency, creating a bottleneck in the development pipeline for voice-based systems. The article explores how synthetic datasets generated through text-to-speech systems and scripted dialogue generation offer promising alternatives by enabling systematic coverage of diverse interaction patterns, including rare edge cases that often reveal critical system limitations. The article encompasses the approaches to synthetic data generation, highlighting how modern neural TTS technologies and sophisticated dialogue simulation frameworks can create realistic conversational corpora with controllable parameters. The benefits of synthetic datasets are analyzed, including enhanced coverage, scalability, and automatic quality labeling capabilities. Implementation considerations focus on balancing realism with systematic exploration, while acknowledging the remaining challenges in bridging the authenticity gap between synthetic and real conversations. We conclude by examining the future trajectory of hybrid evaluation methodologies that strategically combine synthetic and real-world data throughout the development lifecycle.
Keywords: Conversational AI Evaluation, Synthetic Audio Datasets, Text-To-Speech Systems, Dialogue Simulation, Hybrid Evaluation Methodologies

1. INTRODUCTION

Voice assistants have become woven into the fabric of contemporary digital interactions, changing the way people connect with technology in numerous fields. Traditionally human-graded audio interactions have been used to train quality estimation models on an effective evaluation method of these systems. The speed at which voice-based interfaces are evolving has posed unprecedented challenges to quality assessment methodologies, especially as these systems start to get more and more sophisticated and begin to deal with more and more complex conversations. Denise Sogemeier et al. indicate that the conventional methods of evaluation are struggling to support the rapid pace of development among the modern voice assistants and, in turn, pose a big disconnect between development capacity and solid initiative of quality assurance [1, 11]. This discrepancy is a result of inherent inadequacies in the evaluation of conversational systems today, using methodologies developed to assess systems that followed the more limited command-response paradigm, and now being tested in the far more in-depth conversational setting.

Manual collection and annotation of real human-agent dialogues is an investment of large amounts of time and money. The complete rigorous studies on evaluation methodologies emphasize that proper evaluation requires the recruitment of a diverse participant panel, control of realistic interaction scenarios, recording of high-quality audio samples, and employment of trained annotators who apply consistent frameworks when examining different conversations. This data-rich process creates inevitable subjectivity in determining quality, and not all human annotators will agree on the same meaning of concepts of conversational quality like coherence, pertinence, and naturalness. A comparative study in [2] has shown that these humanistic-based methods of evaluation have the tendency to favor some interaction patterns and deliberately avoid others, while traditional metrics have known limitations in correlating with human judgments.

The shortcomings of the conventional techniques do not end at the practical limits of resources, but also relate to essential coverage issues. Collected datasets often overrepresent common interaction patterns while lacking coverage of the entire range of conversational conditions. Such sampling bias leads to assessment frameworks that work in ordinary circumstances but fail when they are asked to deal with out-of-the-box user behavior or uncommon requests, or culturally specific conversational styles. The study by Sogemeier et al. [1] describes how mainstream voice assistant functionalities are unable to cover adequately the broad array of interaction types in a methodical fashion, meaning that the coverage of multi-turn conversations, error-recovery sequences, and dialogues using domain-specific lexicon is most deficient. The issue of these blind spots in evaluation is even more critical with conversational AI systems entering more specialised areas like healthcare, finance, and education, where applying evaluations may have serious actual implications on the quality of verbal communication.

The problem with the traditional evaluation methods mentioned above, namely the resource requirements, puts a significant breaking point on the pipeline of developing conversational AI systems. Steve Bickley et al. [2] underscores

that different development teams have to very closely balance the pressures to evaluate in a thorough way on the one hand against the practical pressures of cost and timing, which leads to the inability to keep up with the model development bottleneck through evaluation cycles. This testing slow point prevents the attempt of new conversational tactics and hinders the capability to quickly iterate on quality upwards. With the conversational AI systems evolving in terms of their capabilities, this gap in evaluation risks becoming a key weakness of the domain itself, potentially limiting the innovation and restricting the deployment of the systems in sectors presenting the highest stakes in terms of quality assurance, ensuring their reliability.

2. The Limitations of Traditional Evaluation Methods

Human-marked datasets, while valuable, present numerous challenges in evaluating conversational AI applications. Data collection is labor-intensive, requiring participant recruitment, session monitoring, interaction recording, and manual labeling by trained evaluators. Research comparing evaluation approaches across leading voice assistant systems has shown that extensive participant selection procedures are necessary to ensure demographic diversity and expose a statistically significant variety of interaction patterns [3]. This resource-intensive process creates substantial obstacles to rapid iteration, as industry timelines dictate that thorough evaluation adds considerable development time—often incompatible with agile development frameworks.

Traditional approaches typically produce limited datasets that inadequately cover edge cases and rare interactions. Systematic quantitative examination of standard evaluation benchmarks reveals consistent coverage deficiencies, as datasets frequently oversample common interaction classes while lacking sufficient examples of multi-turn dialogues, domain-specific queries, and culturally diverse communication patterns [4]. These constraints are particularly problematic when assessing modern conversational systems expected to handle increasingly complex and open-ended communications. Experimental results presented in [4] highlight that even heavily annotated datasets capture only a fraction of interaction types likely to occur in deployment, leaving significant blind spots in quality assessment models. Furthermore, human annotators often apply inconsistent thresholds when rating conversations, introducing variability into training data. Close analysis of how evaluators rate subjective quality attributes such as naturalness, coherence, and appropriateness reveals substantial variation between raters, with inter-rater consistencies declining sharply in complex or ambiguous cases [3]. This subjectivity complicates the development of consistent evaluation frameworks, as the ground truth labels used to train automated quality estimation models inevitably introduce noise during model training. Comparative examination of evaluation techniques suggests this annotation variability disproportionately impacts the assessment of advanced conversational capabilities due to potential systematic bias against innovative dialogue strategies that deviate from established patterns.

The cost and resources required to develop these datasets ultimately constrain the iterative refinement of dialogue systems, as each evaluation cycle demands significant investment. Analysis of development practices across various conversational AI platforms indicates that evaluation protocols often become bottlenecks in release cycles, forcing teams to make difficult tradeoffs between thorough quality testing and development speed [4]. This conflict has intensified as conversational AI systems adopt increasingly complex architectures requiring more comprehensive assessment than their simpler predecessors. The inherent limitations of traditional evaluation systems have driven demand for alternative approaches that reduce constraints while maintaining high quality standards. Figure 1 summarizes the fundamental drawbacks of conventional evaluation methodologies, emphasizing their limited scalability, inconsistent human judgments, and restricted representational breadth.

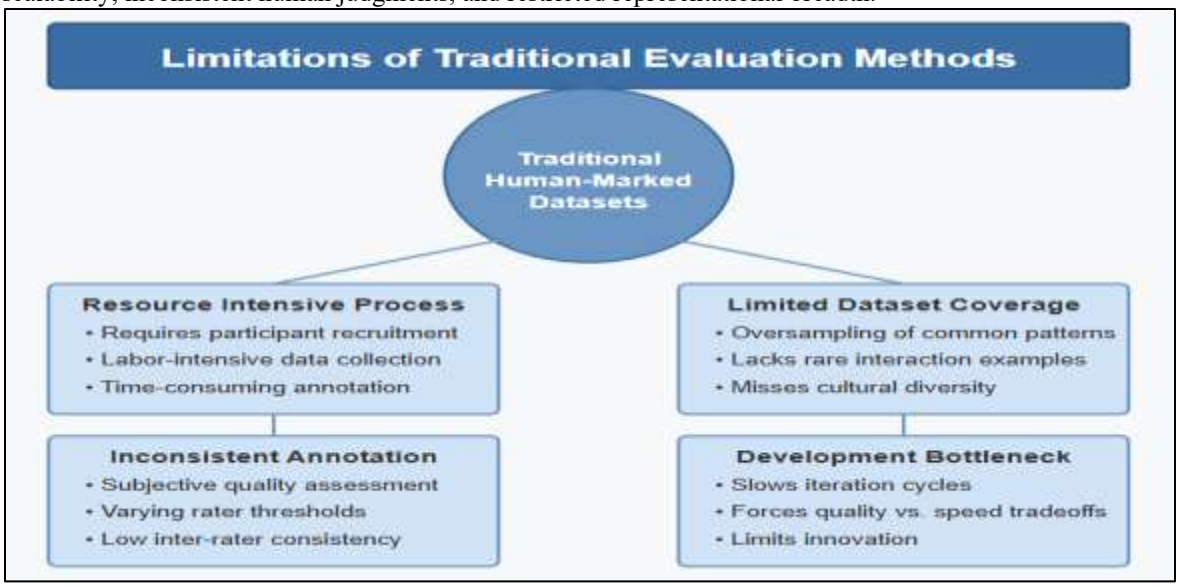

Figure 1: Limitations of Traditional Evaluation Methods [3, 4]

3. Synthetic Data Generation Approaches

One promising alternative to real-world data evaluation is the use of text-to-speech (TTS) and scripted AI-generated speech to create synthetic data. Modern TTS technologies have experienced tremendous advancement in recent years, with neural-based methods demonstrating impressive progress in naturalness and expressivity. Contemporary TTS systems can render speech with controlled prosody characteristics, simulating various speaker types through adjustments in pitch, speaking pace, accent, and emotional tone [5, 9]. These developments enable the generation of artificial voices that closely approximate human speech patterns across diverse demographic samples and communication styles—crucial since manually collected datasets are limited by practical constraints on participant recruitment.

These synthetic voices can be paired with automatically generated conversation scripts that systematically explore various user intents, system responses, and conversational patterns. Recent dialogue generation systems employ advanced natural language generation methods to produce diverse yet realistic conversations [6]. These approaches typically combine template-based generation for context-appropriate utterances with statistical models to handle both regular interaction patterns and edge cases. Current script generation systems can produce large numbers of coherent, natural dialogues that human evaluators often judge as plausible in controlled settings [6]. This capability allows the creation of evaluation datasets that comprehensively explore the interaction space, covering both common scenarios represented in real-world data and important edge cases often omitted but crucial for testing system robustness.

The generation process can be parameterized to control dialogue complexity, errors, interruptions, and other conversational phenomena of interest. This parametric approach enables systematic exploration of the conversational space, allowing researchers to create specialized evaluation datasets targeting specific aspects of system performance [6]. For example, generation frameworks can be configured to produce conversations with specific levels of user interruptions, speech disfluencies, background noise, or domain-specific terminology—factors that significantly impact real-world system performance but are difficult to assess through standard methods. Advanced neural TTS technologies can introduce speech nuances that emulate human communication behaviors such as pauses, emphasis, and emotional modulation, which substantially influence how dialogue systems interpret and respond to user input [5]. Manipulating these parameters creates targeted evaluation datasets that scrutinize specific aspects of system performance while controlling for confounding variables.

This parametric generation technique represents a paradigm shift in evaluation methodology, moving from opportunistic sampling of real-world interactions to systematic exploration of the conversational space. The framework described by Heydar Soudani et al. [6] enables the creation of evaluation sets designed to explicitly probe system limitations rather than merely reflecting likely usage patterns, yielding more informative measures of robustness and generalization performance. This approach addresses a fundamental drawback of conventional evaluation methods, which tend to overemphasize generic interaction patterns while providing little insight into outliers and edge cases. Synthetic data generation techniques offer a powerful complement to existing assessment methods by enabling controlled exploration of the conversation space, potentially accelerating the development of more robust and pragmatic conversational AI systems. Figure 2 illustrates the main components of the synthetic data generation process, including text-to-speech synthesis, scripted dialogue generation, and parametric control mechanisms for exploring conversational complexity.

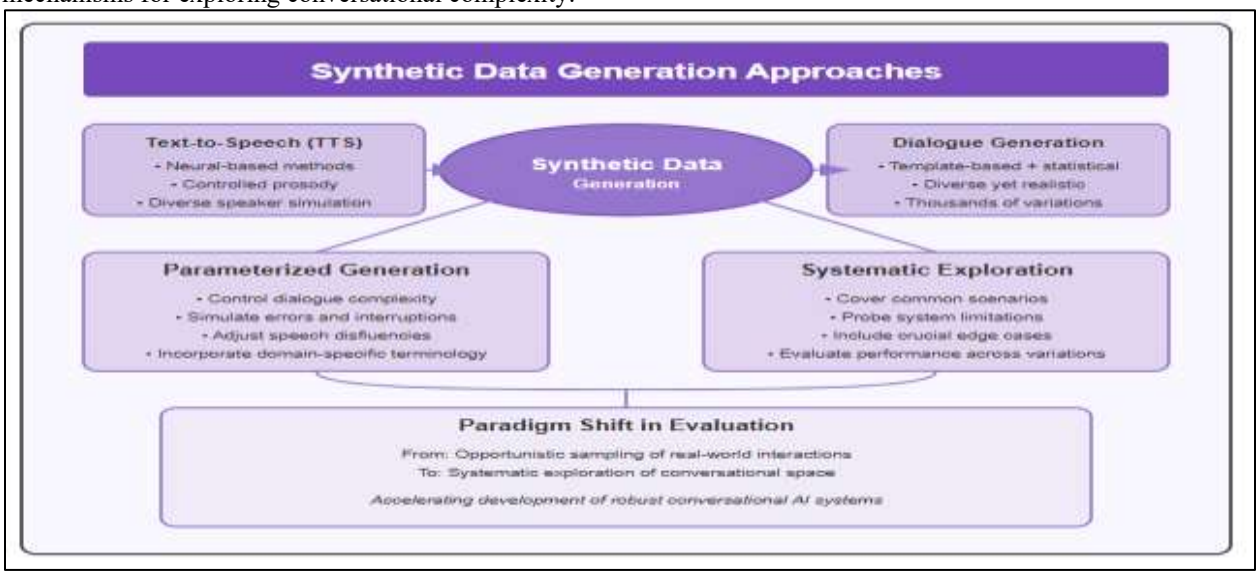

Figure 2: Key Components of Synthetic Data Generation for Conversational AI [5, 6]

4. Benefits of Synthetic Conversational Corpora

Synthetic data offers several benefits in relation to AI conversational simulation. These properties first allow systematic coverage of user behavior that may be poorly represented in naturally collected data and is either rare or complex, e.g., interruptions, hesitations, misrecognitions, and repair strategies. Comparative studies of synthetic and natural datasets used by evaluators have shown that synthetic methods make possible far-reaching coverage of corners and odd interaction scenarios [7]. Expanded coverage is crucial, as rare interactions often expose weaknesses that traditional testing overlooks. Existing frameworks and surveys indicate that synthetic datasets can enhance the ability to detect system weaknesses—especially for complex error recovery and atypical multi-turn interactions—relative to conventional evaluation alone [7]. The systematic observation of the interaction space overcomes a core weakness of opportunistic data collection methods, which often cannot be used to represent all varieties of usage patterns found in the real world.

Further, synthetic data can be generated in large quantities—thousands or millions of diverse interactions—without the inherent biases and resource constraints associated with human data collection methods. The scalability of state-of-the-art dialogue generation and text-to-speech systems enables the creation of large-scale evaluation datasets with substantially reduced reliance on manual collection compared to traditional pipelines [7, 8]. Such scalability transforms the practical feasibility of conversational AI assessment, allowing for more comprehensive and frequent evaluation cycles throughout the development process. Recent studies suggest that synthetic data generation can reduce evaluation costs while broadening the range of interaction types, particularly in specialized domains where recruiting qualified participants is challenging [7]. This cost-effectiveness enables sustained, detailed evaluation regimes throughout the entire development cycle rather than limiting rigorous testing to major milestones.

Third, dialogue systems may produce automatically labeled satisfaction scores or other quality measures by pre-determined heuristics or based upon an existing model, resulting in a massive amount of pseudo-labeled data with which to train evaluators [7]. This labeling method can alleviate another severe bottleneck within the form of the conventional evaluation pipelines, which is manual quality measurement annotation [7]. This can be achieved by the incorporation of evaluation criteria themselves into the generation process, allowing researchers to create datasets with homogeneous and transparent evidence of their quality concerning the immediate evaluation goals. Recent surveys indicate that with sufficiently large synthetic datasets, the performance gap with human-labeled data can narrow for certain objective measures such as task completion and information accuracy [11]. Comprehensive analysis indicates that the volume and diversity advantages of synthetic data can outweigh potential naturalism limitations in many evaluation scenarios.

The method achieves remarkable performance, especially in offline assessment and pre-deployment phases, in which the generalizability of possible interaction patterns is more important than high naturalism. The framework outlined in [7] shows that synthetic datasets have the potential to be effectively utilized as part of a multi-stage pipeline that can essentially be used as the initial screening mechanism that would help define any problems before engaging in the much more resource-intensive human-based evaluation activities. Such a hybrid strategy will provide the advantages of both synthetic and natural evaluation procedures, with tractable synthetic procedures providing wide-ranging coverage and human evaluation resources being reserved to tackle more qualitative forms of assessment that may gratify subjective judgment. Integrating these methods selectively, researchers can create more comprehensive and effective evaluation frameworks that will fulfill all the inherent weaknesses of the conventional methodologies, at the same time, encompassing a high level of quality control. The incorporation is a breakthrough in assessing conversational AI, and it may provide faster creation of robust and capable dialogue systems. Figure 3 depicts the major benefits of synthetic conversational corpora, emphasizing their scalability, extended coverage, and capacity for automatic labeling that complement human-based evaluation.

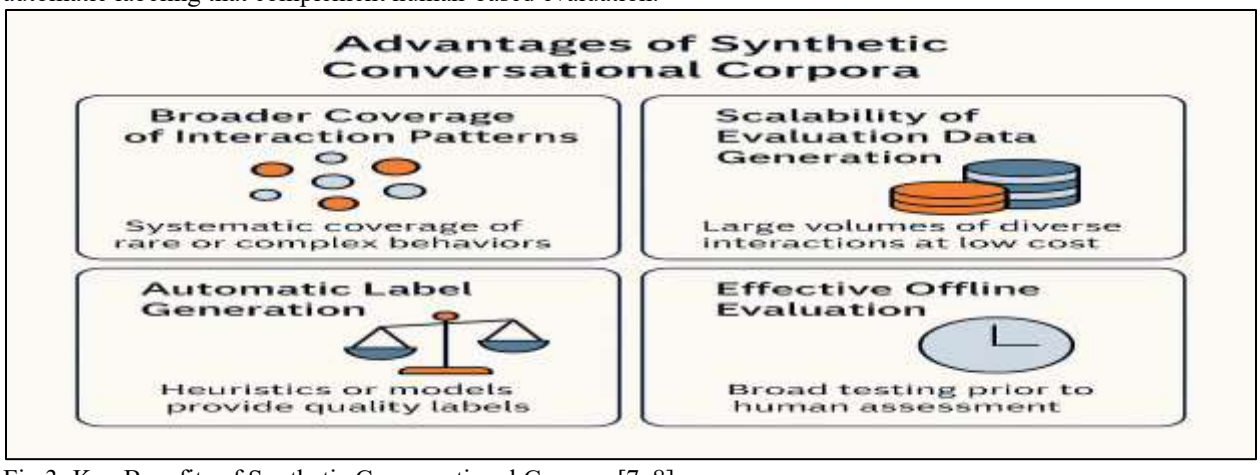

Fig 3: Key Benefits of Synthetic Conversational Corpora [7, 8]

5. Implementation Considerations

Constructing good synthetic datasets demands a well-chosen design so as to achieve a plausible realism, variety, and control. TTS voices strongly influence the quality of the content produced, and they determine the extent to which the synthetic data reflects real situations involving users. The reality of the voices generated depends on the level of naturalness, prosody, and emotional nuances. Recent advances in modern TTS engines, such as VITS [9], highlight quality dimensions that strongly affect the usefulness of synthetic data for evaluating conversational AI systems. The analyses reveal that these are indeed important variables in addition to mere intelligibility in the construction of synthetic conversations that will be effective in mimicking real-world conversation. The reported experiments [9] indicate that although perfect naturalism remains challenging, modern neural TTS systems can achieve high perceptual quality when trained and tuned appropriately. Thereby, this study has identified particular measures of quality in certain evaluation conditions, making it possible to gain practical advice on applying synthetic data methods to various application areas.

In the same way, the generation of dialogue scripts should be between a controlled variation and the reality of a conversation. Powerful methods tend to merge the use of rule-based templates with statistical models in order to arrive at varying but realistic dialogues. The generalized structure, described in [6], shows how hierarchical generation methods can preserve conversational coherence while structurally covering diverse interaction forms [6]. Comparisons of template-based, statistical and hybrid approaches suggest that hybrids often better capture realistic error sequences and recovery actions [6]. Such hybrid approaches allow researchers to be in control of key parameters of critical evaluation with the statistical models used to produce natural language variations finding the middle ground between systematic coverage and realistic conversational aspects. In the corresponding process of creating pseudo-labels of such conversations, metrics should include various factors that have been shown to have a positive correlation with the user satisfaction, including task success, efficiency of a conversation, or correct system responses to user requirements. The labeling scheme described in [11] illustrates how multiple evaluation aspects can be embedded into generation to yield multi-dimensional quality labels. Experimental studies surveyed in [11] indicate that models trained with such labels can approach the performance of systems trained on human annotations for objective tasks, especially on metrics of task accomplishment and correct information. Yet, recent work also notes limitations for more subjective scales like naturalness and engagingness, indicating that hybrid evaluation (synthetic + human) remains essential for complete assessment.

It is also necessary to view the introduction of synthetic evaluation frameworks in terms of practical integration with current development processes. Recent frameworks and surveys [7, 11] point out some key success factors associated with the implementation of synthetic approaches to evaluation, such as: consistency with the particular evaluation goals, calibration with human judgments on key metrics, transparent reporting of generation parameters and limitations, and tactical integration with traditional approaches. It is these practical concerns of implementation that make a key difference to the extent to which synthetic datasets are useful as add-ons to existing evaluation schemes, or whether they create novel sources of bias and other restraints. Figure 4 illustrates implementation considerations for synthetic data evaluation, outlining the relationships among TTS realism, dialogue script generation, pseudo-labeling strategies, and integration with existing assessment workflows.

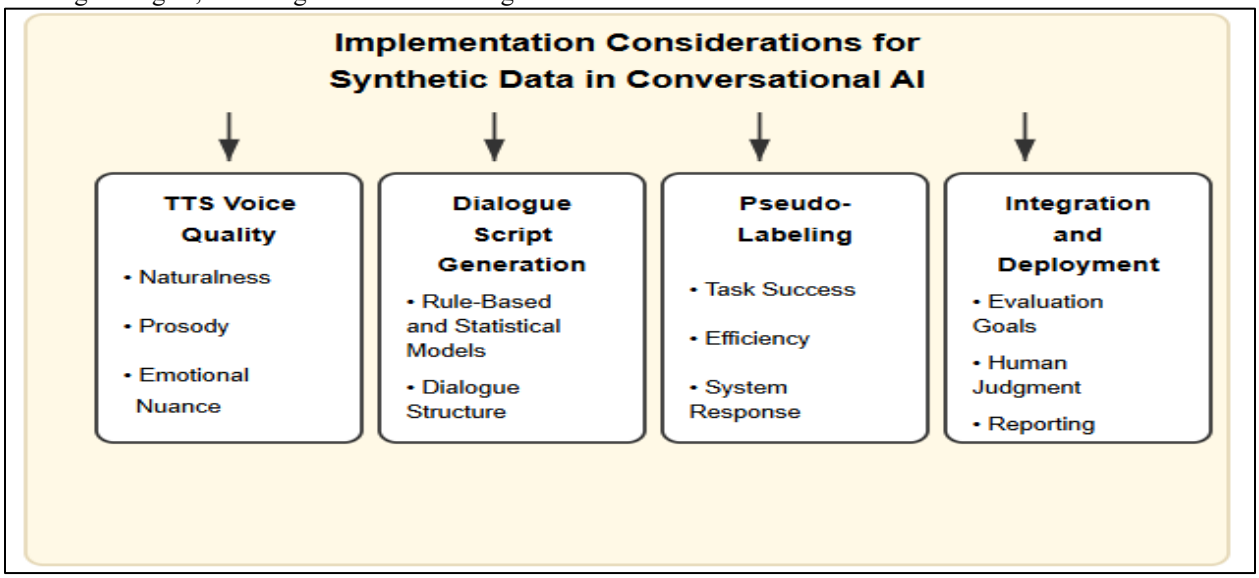

Figure 4: Implementation Considerations for Synthetic Data in Conversational AI [7, 11]

6. FUTURE DIRECTIONS AND LIMITATIONS

Although synthetic datasets can bring tremendous advantages, there are still various issues when it comes to implementing the technology and using it in practice. The gap between synthesized and real conversation authenticity is not crossed yet but keeps getting smaller. In the comparison of the human understanding of synthetic and real conversation, it has been observed that there are some dimensions that synthetic conversations have not achieved, especially in terms of natural variation and sensitivity of conversation as observed in real human communication [6]. Such weaknesses tend to be the most significant in emotionally complicated situations, culturally specific communication styles, and ultra-contextual communication activities with reliance upon background knowledge. Such perception gaps, highlighted in holistic benchmarks like HELM [4], demonstrate that synthetic conversations still fall short on natural variation, cultural sensitivity, and nuanced context. Although such authenticity gaps are substantial limitations, they have to be balanced against the major advantages in coverage, scale and control that synthetic methods offer, especially with regard to systematic testing of conversational system robustness and generalizability levels.

Synthetic data is very useful during the pretraining and offline evaluation tasks, but usually needs the addition of real-world data in order to tune and finalize the model. Hybrid methods combining synthetic and real data, as described in comprehensive surveys of dialogue evaluation [11], have the potential to capitalize on the respective strengths of each approach. The combined strategy normally comprises deploying synthetic data to cover large areas of possible interaction patterns and edge conditions, and then doing a specific, real-world examination precisely on those aspects of quality that synthetic information has failed to capture properly. Recent benchmark studies highlight that hybrid methods can offer more balanced and scalable evaluation than either synthetic or real-world approaches alone [4]. The implications of these results are that conversational AI evaluation in the future does not seem to be a matter of picking one or the other approach, but the methodological sophistication of how these two data sources can be effectively integrated strategically throughout the development cycle. Most recent studies indicate that synthetic conversations are useful in training dialogue quality estimators and response selection models, where performance is comparable to that on human-labeled datasets. Recent surveys [6] describe advances that have narrowed gaps between models trained on synthetic versus human-labeled data in some settings. In parallel, benchmark work emphasizes conditions under which evaluation comparability improves across datasets and tasks [4]. These findings hint that eventually, synthetic data methodology might already be relevant enough to apply practically to a wide range of assessment tasks and progress to make them even more relevant will come with the advances of the constituent technologies as well.

With ongoing research of TTS technology and dialogue simulation tools, synthetic data will probably play a more significant role in scaling evaluation methodology used in conversational AI systems that may revolutionize the development and improvement of voice-based agents. Recent surveys [6] point to several promising research directions that could make synthetic data methods even more useful, such as enhancing emotion modeling in TTS systems, simulating more natural conversational dynamics (e.g., interruptions and backchanneling), and incorporating cultural context in dialogue generation. In addition, emerging work on socially aware synthetic dialogues [10] highlights the importance of grounding automatic labeling schemes in subjective quality dimensions. These developments, along with increased access to computational resources capable of running large-scale simulations, indicate that the role of synthetic evaluation methodologies in conversational AI development will only grow, perhaps with the benefit of accelerated innovation cycles at large scale. The evolution marks one of the paradigm changes in evaluation, shifting to highly blended evaluation techniques and methods that combine synthetic and real-world evaluation in complementary ways and across the development lifecycle.

CONCLUSION

Composing synthetic audio data sets is a radical change in testing conversational AI systems, both expanding essential constraints of conventional, human-focused methods, and holding to high quality control standards. Despite the persistent difficulties to achieve the perfect simulation of human communication (and its intricate particularities), the current rate of development of TTS-technologies and dialogue generation systems reduces this level of authenticity even further. Thus, hybrid methods of evaluation that make the best use of the unique advantages of both synthetic and real-world data, with synthetic data used to cover all possible forms of interaction patterns and edge cases comprehensively, and human evaluation used to provide subjective dimensions of quality where synthetic methods continue to suffer, is the most promising avenue ahead. This evolution toward hybrid evaluation methodologies allows systematic assessment of conversational capabilities, which may both shorten innovation cycles and enable more robust quality evaluation. With the maturation of these technologies, it is probable that synthetic datasets will increasingly become a core component in the development stack of voice-based systems, changing the approaches of how conversational agents are developed, tested, refined and deployed as they are applied to an ever-expanding, multi-faceted and complex variety of applicative areas.

REFERENCES

[1] Denise Sogemeier, et al. 2024. Short Assessment of Voice Assistants Scale (SAVAS): A Screening Instrument for Human-Technology Interaction Assessment. In AutomotiveUI '24 Adjunct: Adjunct Proceedings of the 16th International Conference on Automotive User Interfaces and Interactive Vehicular Applications. https://doi.org/10.1145/3641308.3685031

[2] Steven J. Bickley, Ho Fai Chan, Bang Dao, Benno Torgler, Son Tran, and Alexandra Zimbatu. 2024. Comparing human and synthetic data in service research: using augmented language models to study service failures and recoveries. Journal of Services Marketing, 39(1), 36–52. https://doi.org/10.1108/JSM-11-2023-0441

[3] Deepika Chauhan, Chaitanya Singh, Romil Rawat, and Manoj Dhawan. 2024. Evaluating the Performance of Conversational AI Tools: A Comparative Analysis. Wiley. https://doi.org/10.1002/9781394200801.ch24

[4] Percy Liang, Rishi Bommasani, Tony Lee, et al. 2022. Holistic Evaluation of Language Models (HELM). arXiv:2211.09110. https://arxiv.org/abs/2211.09110

[5] Jonathan Shen, Ruoming Pang, Ron J. Weiss, et al. 2018. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions (Tacotron 2). In ICASSP 2018 – IEEE International Conference on Acoustics, Speech and Signal Processing, 4779–4783. https://doi.org/10.1109/ICASSP.2018.8461368

[6] Heydar Soudani and Roxana Petcu. 2024. A Survey on Recent Advances in Conversational Data Generation. arXiv:2405.13003. https://arxiv.org/abs/2405.13003

[7] Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. 2025. Comprehensive Framework for Evaluating Conversational AI Chatbots. arXiv:2502.06105. https://arxiv.org/abs/2502.06105

[8] Geonyeong Son and Misuk Kim. 2024. A Simple and Efficient Dialogue Generation Model Incorporating Commonsense Knowledge. Expert Systems with Applications, 249(B), 122198. https://doi.org/10.1016/j.eswa.2023.122198

[9] Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. VITS: Conditional Variational Generation for End-to-End Text-to-Speech. In Advances in Neural Information Processing Systems (NeurIPS 2021). https://proceedings.neurips.cc/paper/2021/hash/6a2016a2ef5c5e0b494bb0a72e0c1f6d-Abstract.html

[10] Chengfei Wu and Dan Goldwasser. 2024. Hiding in Plain Sight: Designing Synthetic Dialog Generation for Uncovering Socially Situated Norms. arXiv:2410.00998. https://arxiv.org/abs/2410.00998

[11] Jan Deriu, Alvaro Rodrigo, Arash Eshghi, et al. 2021. Survey on Evaluation Methods for Dialogue Systems. Artificial Intelligence Review, 54, 7451–7492. https://doi.org/10.1007/s10462-021-10030-3