IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Customer Perceived Reliability For Conversational AI Assistants: Measurement, Challenges And Opportunities With Generative AI

Abhai Pratap Singh[1], Adit Jamdar[2], Prerna Kaul[3]
[1 3]Product Leader, [2]Software Engineer
[1]Independent Researcher, Milpitas, CA, USA

Abstract: This paper discusses the challenges associated with measuring customer-perceived reliability for conversational AI assistants and proposes a framework to measure the reliability of such assistants. We first provide an overview of conversational AI assistants and propose a dimensional approach to defining reliability that covers the broad range of use-cases served by these assistants. We analyze existing mechanisms used to measure reliability of assistants and discuss challenges presented by the new wave of assistants powered by Large Language Models (LLMs). The proposed solution framework leverages end-to-end telemetry, feedback collection through real-time sentiment analysis, and offline analysis of interaction logs by generative models to provide a comprehensive view of reliability. This research's framework includes further analysis of data to identify the source of defects and feed information back to the assistant, enabling the assistant to adjust its responses as needed. By building on recent progress in LLMs, this framework tackles issues such as context-awareness, external tool integration and scalability. Finally, this work offers tangible ideas for implementation while considering real world limitations around cost, privacy, and compliance with regulations.

Index Terms - artificial intelligence, chatbots, conversational AI, customer experience, reliability measurement, voice assistants

I. INTRODUCTION

Stand-alone conversational AI assistants have become ubiquitous in modern society, with platforms like Amazon Alexa, Apple Siri, and Google Assistant serving over 150 million active US users alone [9]. They can communicate with the user through everyday language and perform a multitude of functions, from setting alarms, playing music and controlling smart home devices. The technology landscape is quickly changing with emergent technologies based on Large Language Models (LLMs), leading to new assistants, including OpenAI's ChatGPT, which gained 180 million active users around the world just 2 years post-launch [13]. Initially released in the form of chatbots, many of these modern conversational AI assistants can process vast amounts of contextual information to execute progressively sophisticated functions such as answering complex questions, recalling past conversations [7]. and creating imaginative content [17], including summarizing email messages, writing computer programming.

As these assistants become more sophisticated and integrated into daily life, the need for robust measurement methodologies for their reliability has become imperative. Traditional software reliability metrics are often rendered inadequate by the unique characteristics of conversational AI systems (e.g. open-endedness, subjectivity and heavy reliance on context [2]). This study fills the gap by analyzing existing measurement approaches and challenges posed by advances in AI assistant technology to create a comprehensive framework that provides insights into measuring and improving reliability as perceived by the customer.

II. CONVERSATIONAL ASSISTANTS OVERVIEW

First, this research presents a technical overview of ChatGPT and similar AI assistants (see Fig 1). User inputs are provided to the Assistant through voice or text; these inputs are delivered from a Mobile application, web interface or smart device. If interfaced through voice, a Speech Recognition component extracts text tokens from audio input. For conventional assistants, these text tokens will go through a Language Processing component to semantically analyze the text, then various layers of business logic that take care of the user intent and produce a text, speech or visual response to the user. In newer assistants, some or all this processing could be done by an LLM. The assistant interacts with multiple complementing pieces of software that provide context augmenting the prompt with information of interest for the user's request (past conversations, location metadata, user settings etc.) and factual knowledge access using Retrieval Augmented Generation (RAG). As the last step, the assistant may interact with third-party applications (ChatGPT Functions, Alexa Skills, Gemini Extensions etc.) to perform user-requested actions.

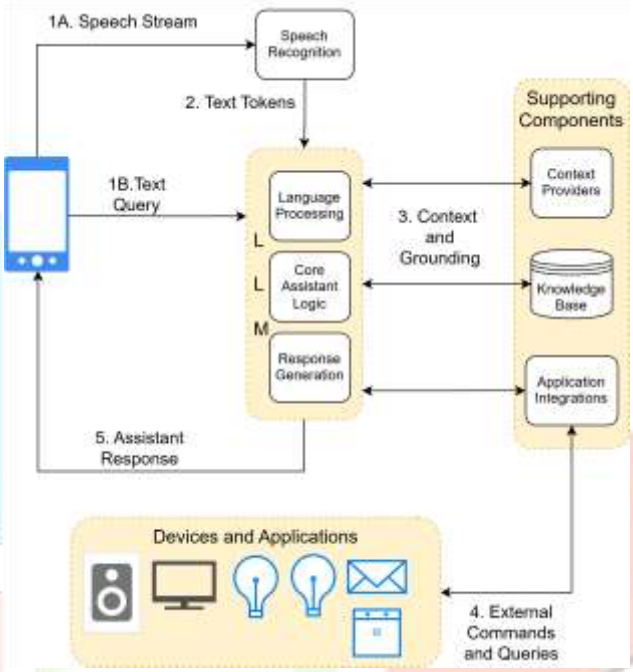

Figure 1. Technical overview of Conversational AI assistants

III. RELIABILITY DIMENSIONS AND MEASUREMENT CHALLENGES

3.1 Definition of Reliability

This research defines customer-perceived reliability, in the context of conversational AI assistants, as the user's subjective assessment of the assistant's ability to consistently and accurately understand their intent, provide relevant and helpful responses, and successfully complete requested tasks. To consider reliability across the broad range of functionalities provided by modern AI assistants, this paper proposes separating its measurements across different dimensions. The first dimension is on conversational quality; effective assistants should conduct fluid, natural conversations while accurately interpreting the users' intentions and keeping the context in mind. The second dimension is information retrieval; assistants must maintain accuracy in sourcing information, precision in presentation, and must also avoid misinformation. The third dimension involves content creation; assistants must produce results that align with subjective user expectations for style, tone and creativity. The fourth-dimension deals with task execution; the assistant must select the appropriate external tools and orchestrate actions to successfully complete the requested task end-to-end.

3.2 Current Measurement Approaches

Traditional measurement approaches range from simple methods like human evaluation or explicit customer feedback, to complex approaches with machine learning models. Several AI assistants rely on having human evaluators review a sample of interactions with users to assess the quality of the assistants [15]. While this method provides a nuanced understanding of reliability, it is labor-intensive, time-consuming, and can be prone to subjectivity. Assistants may also use in-conversation or in-app prompts to gather explicit customer feedback. Companies may rely on more formal survey mechanisms such as the User Experience Quality framework put forth by Klein, Andreas M., et al. [3]. These methods offer direct insight into an individual user's subjective perception of quality but can suffer from low response rates (~5-30%) [11] and potential sampling and response biases [12]. More sophisticated approaches involve using machine learning

models trained to analyze interaction logs and identify patterns indicative of reliability issues. Unlike the more labor-intensive methods described above, models can process large volumes of data quickly but may struggle to capture the nuances of human language and context. As explored by Khaziev, Rinat, et al, [6] such models may use less sophisticated approaches such as word-overlap metric models or more complex approaches that analyze interactions across several dialogs to assess quality.

3.3 Emerging Challenges

AI assistant usage is growing exponentially, introducing unparalleled challenges for measuring reliability. Recent industry analysis shows that ChatGPT alone gives rise to over 100 billion words each day [13], rendering traditional evaluation with human annotation and surveys increasingly impractical. AI assistants are increasingly consuming more contextual information about their users to provide a richer and more personalized experience. Context may also include large volumes of data such as the user's documents, emails and calendar appointments or memory of past conversations with the assistant, further increasing complexity for both human annotators and other machine learning models. Conversational AI assistants often rely on external tools and services to perform tasks in the real world, such as controlling smart home devices or playing music. Failures in these external components are hidden from the user and may be perceived as failures of the assistant itself. Pinpointing the exact cause of errors can be challenging with the unpredictability of generative AI models, unconstrained user inputs, integration with external tools and the context unique to every interaction. This presents challenges with assigning responsibility and implementing targeted improvements. Any framework that seeks to improve measurement of reliability should also aid in identifying problem areas to drive resolution of unreliable experiences.

IV. Proposed Framework

Due to the potential for open-ended interactions with the assistant and diverse challenges around these interactions, the research proposes a well-rounded reliability measurement framework. As Fig. 2 shows, this architecture is built upon 6 components working closely with contemporary LLMs. Every part resolves challenges described previously while also allowing for scalable analysis of reliability. The components of the framework are defined below-

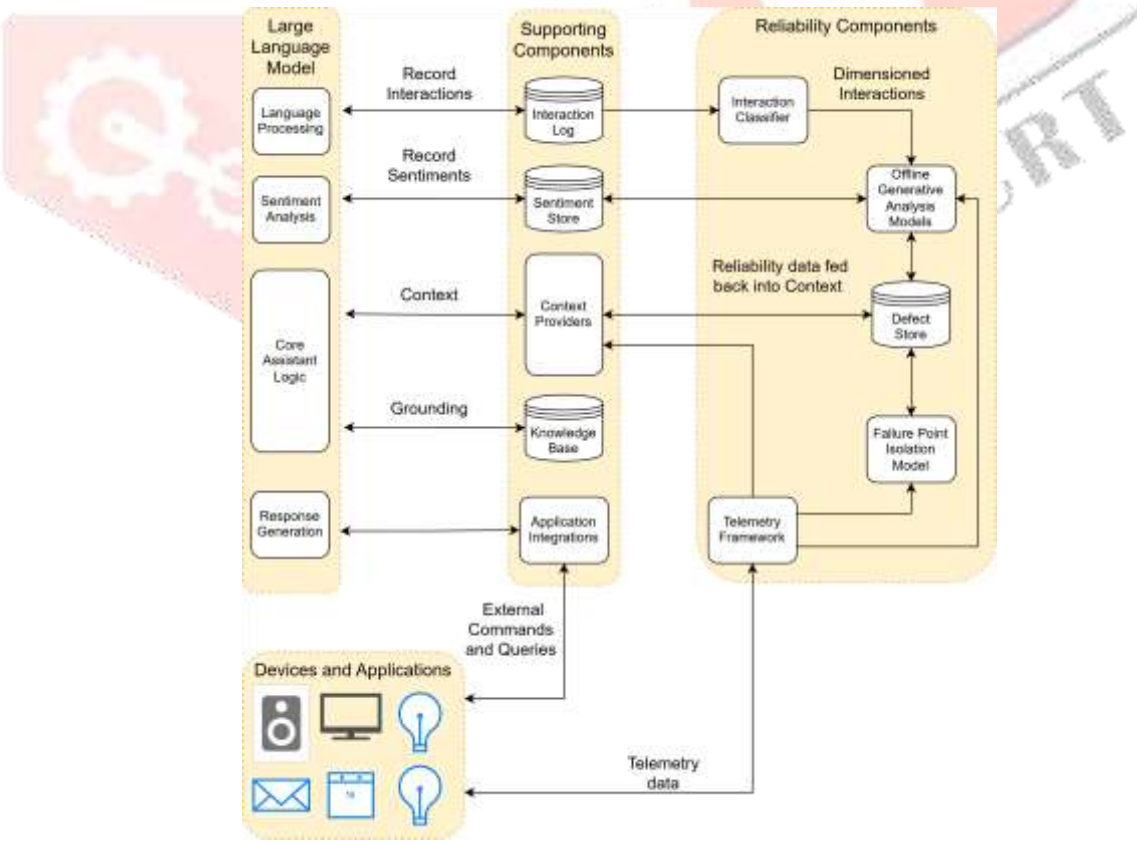

Figure 2. Proposed Architecture for Conversational AI reliability

4.1 Dimensional Reliability Analysis

The framework initially defines independent reliability metrics across the main assistant dimensions. A lightweight classifier is developed to classify interactions across the four dimensions discussed in Section III, allowing measurement strategies that are targeted for each dimension. The framework uses fact-checking algorithms to verify the accuracy and credibility of information retrieval tasks while real-world outcome measurements are used to judge task-oriented interactions. The overall measure of the assistant's reliability is the composition of the individual reliability measurements across each dimension.

4.2 End-to-End Telemetry

Particularly for task-oriented use-cases, the reliability of the user experience is defined by its end-to-end success, not just by the output of the LLM. The framework requires end-to-end telemetry tracking, requiring that all integrated applications and devices report the outcomes of their actions and proactively report any changes in state to the assistant. This provides full visibility into the user experience and allows offline analysis of potential defects.

4.3 Enhanced Feedback Collection

The framework incorporates both explicit and implicit feedback mechanisms. Explicit prompts continue to be important to collect user feedback [8,16]. In addition, modern generative models demonstrate remarkable capability in sentiment analysis [1], allowing for continuous collection of implicit feedback signals during conversations. This creates a rich stream of real-time user sentiment data that improves both immediate responses and long-term performance. Finally, generative models are used to create targeted, individualized surveys based on an individual user's interaction history to collect detailed and personalized feedback from customers.

4.4 Offline Analysis System

Leveraging advanced generative models, the offline analysis system processes interaction logs incorporating user context, historical data, telemetry, implicit feedback and explicit survey responses, thus scaling traditional human annotations to handle the volume of data generated by modern assistants. The system creates comprehensive user profiles from past interaction logs enabling personalized reliability assessment. Human annotators verify a small sample of interactions to ensure system accuracy.

4.5 Defect Attribution

When reliability issues are detected, a specialized model analyzes the likely cause by comparing all the parameters described above across past interactions for the user and similar interactions from matching user cohorts. This defect attribution system can leverage either generative models or enhanced versions of existing Failure Point Isolation frameworks [6].

4.6 Feedback Integration

The final component establishes a complete feedback loop, feeding reliability insights back into context providers for future interactions. This allows the assistant to maintain awareness of prior issues and adjust responses accordingly, while providing appropriate transparency to users regarding known limitations or previous difficulties.

V. FUTURE CONSIDERATIONS

The field of conversational AI assistants and generative AI is rapidly evolving. While the key components of this research's reliability framework address current challenges, several important considerations merit attention for future implementations:

5.1 Cost-Benefit Analysis

For real-world use cases in the present context, costs that would be incurred from implementing every tenet of the reliability framework often outweigh the short-term benefits of doing so. However, as a16z report [14] show, computation costs keep falling, with LLM inference costs dropping about 10x year-on-year at present. This reallocation continues to make the case for investing in holistic reliability measurement infrastructure more compelling.

5.2 Privacy and Regulatory Compliance

The proposed framework collects, stores and analyses implicit user feedback and telemetry data across the ecosystem. As AI assistants become more integrated into daily life, increasing governmental scrutiny and new regulations will affect implementation. More research is necessary to ensure that such systems maintain privacy requirements whilst providing sufficient data access to improve reliability.

5.3 Pattern of Integration with the Assistant

AI Assistants and generative AI models are continuously evolving. This research expects to see specialized assistants (e.g. Gemini Gems, custom GPTs) that are trained to better handle specific categories of tasks. OpenAI's integration with Siri [10] highlights how modern AI assistants may leverage each other's capabilities to extend assistant functionality (e.g. ChatGPT for complex information retrieval, Siri for task execution). This evolution in how AI assistants operate and integrate with various systems brings new dimensions to measuring reliability, particularly in terms of how these systems process and analyze user interactions.

These future considerations help clarify that this research's framework development is tempered by practical limits surrounding cost, privacy, and integration complexity. Given the fast pace of evolution in conversational AI technology, a more frequent re-evaluation of the framework will likely be required to ensure its ongoing efficacy.

VI. Conclusion

This work contributes key components of a new framework for both measuring and optimizing customer-perceived reliability for conversational AI assistants. This research's analysis shows the inadequacy of traditional metrics in assessing modern AI assistant reliability given the scale of data generated, integration with external tools and contextual data informing the user experience. While the proposal is a step forward in the increasingly complex world of conversational AI assistants, it also raises questions about cost and privacy. This has implications for future work on conversational AI assistants, particularly sentiment analysis techniques, telemetry tools, defect attribution and effective privacy-centric feedback aggregation strategies. The proposed framework lays the groundwork for further progress in reliability measurement as AI assistants become more and more embedded in everyday life.

References

[1] Krugmann, Jan Ole, and Jochen Hartmann. "Sentiment Analysis in the Age of Generative AI." Customer Needs and Solutions, vol. 11, no. 1, 2024, doi:10.1007/s40547-024-00143-4.

[2] Faruk, Lawal Ibrahim Dutsinma, et al. "A Review of Subjective Scales Measuring the User Experience of Voice Assistants." IEEE Access: Practical Innovations, Open Solutions, vol. 12, 2024, pp. 14893–14917, doi:10.1109/access.2024.3358423.

[3] Klein, Andreas M., et al. "Measuring User Experience Quality of Voice Assistants." 2020 15th Iberian Conference on Information Systems and Technologies (CISTI), IEEE, 2020, pp. 1–4.

[4] Park, Kunwoo, et al. "Positivity Bias in Customer Satisfaction Ratings." Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18, ACM Press, 2018, pp. 631–638.

[5] Johnson, Timothy, et al. "The Relation between Culture and Response Styles: Evidence from 19 Countries." Journal of Cross-Cultural Psychology, vol. 36, no. 2, 2005, pp. 264–277, doi:10.1177/0022022104272905.

[6] Khaziev, Rinat, et al. "FPI: Failure Point Isolation in Large-Scale Conversational Assistants." Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track, edited by Anastassia Loukina et al., Association for Computational Linguistics, 2022, pp. 141–148.

[7] Openai.com, "Memory and new controls for ChatGPT", 13 Feb. 2024, https://openai.com/index/memory-and-new-controls-for-chatgpt/.

[8] Xiao, Ziang, et al. "Let Me Ask You This: How Can a Voice Assistant Elicit Explicit User Feedback?" Proceedings of the ACM on Human-Computer Interaction, vol. 5, no. CSCW2, 2021, pp. 1–24, doi:10.1145/3479532.

[9] Kumar, Naveen. "67 Voice Search Statistics 2024: Users & Trends Data." DemandSage, 12 Nov. 2024, https://www.demandsage.com/voice-search-statistics/.

[10] Openai.com, "OpenAI and Apple announce partnership to integrate ChatGPT into Apple experiences", 12 Jun. 2024, https://openai.com/index/openai-and-apple-announce-partnership/.

[11] Chung, Lucia. "What Is a Good Survey Response Rate for Online Customer Surveys?" Delighted, 17 Feb. 2022, https://delighted.com/blog/average-survey-response-rate.

[12] Chung, Lucia. "The 7 Types of Sampling and Response Bias to Avoid in Customer Surveys." Delighted, 14 June 2019, https://delighted.com/blog/avoid-7-types-sampling-response-survey-bias.

[13] Duarte, Fabio. "Number of ChatGPT Users (Nov 2024)." Exploding Topics, 30 Mar. 2023, https://explodingtopics.com/blog/chatgpt-users.

[14] Appenzeller, Guido. "Welcome to LLMflation - LLM Inference Cost Is Going down Fast □." Andreessen Horowitz, 12 Nov. 2024, https://a16z.com/llmflation-llm-inference-cost

[15] Statt, Nick. "Amazon's Alexa Isn't Just AI — Thousands of Humans Are Listening." The Verge, 11 Apr. 2019, https://www.theverge.com/2019/4/10/18305378/amazon-alexa-ai-voice-assistant-annotation-listen-private-recordings

[16] Deng, Yuqi, and Sudeeksha Murari. "When a Voice Assistant Asks for Feedback: An Empirical Study on Customer Experience with A/B Testing and Causal Inference Methods." Companion Publication of the 2021 International Conference on Multimodal Interaction, ACM, 2021, pp. 183–191.

[17] "The Potential of AI Generative Models: How They Are Changing the Game." Prymatica, 4 Mar. 2023, https://www.prymatica.com/articles/the-potential-of-ai-generative-models-how-they-are-changing-the-game/