International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V72I11P126

Volume 72 Issue 11, 241-247, November 2024

Original Article

Measuring and Improving AI Performance in Conversational Shopping Assistants

Abhai Pratap Singh[1], Prerna Kaul[2]

[1]*Product Leader, Independent Researcher, Sunnyvale, California, USA.*
[2]*Product Leader, Independent Researcher, Seattle, Washington, USA*

[1]*Corresponding Author : prernakaul1@gmail.com*

Received: 12 October 2024 Revised: 13 November 2024 Accepted: 26 November 2024 Published: 30 November 2024

Abstract *- Generative AI has transformed the domain of conversational AI, opening doors for a new breed of e-commerce shopping assistants. Developments in Generative AI have the potential to improve customer experience through more natural, dynamic and context-aware dialog exchange. However, the current implementations have critical flaws — limited access to real-time data and not enough contextual knowledge, resulting in difficulties in attaining trust and satisfaction from the user. Addressing these gaps is key to unlocking the potential of generative AI in building seamless shopping experiences. Using the Natural Conversation Framework (NCF), this paper evaluates the performance gaps and recommends how generative AI can create world-class shopping assistants. By combining conversational UX principles with a holistic measurement framework, this study provides a structured way to improve reliability, personalization, and conversational depth. This study incorporates a blend of technical and user-centric metrics, such as product accuracy, latency, user engagement, and satisfaction; this provides us with a holistic view of a conversation AI system's performance. Beyond mapping challenges, this study elaborates on the path towards scalable human-centered conversational agents. This study also illustrates how to design shopping assistants with interaction patterns that are scaffolding for a centralized technical architecture to drive long-term engagement and trust. This work has practical implications for researchers and builders aspiring to harness generative AI for e-commerce applications.*

Keywords *- Conversational AI, E-commerce, Engagement metrics, Generative AI, Shopping assistants.*

1. Introduction

Conversational AI assistants have facilitated a tremendous change within the e-commerce landscape. Current market research estimates the market will reach $19.21 billion by 2025 — up from the hedge of $15.5 billion in 2024 [15] and shopping assistants are becoming more core to online shopping in the years ahead.

We have come a long way since basic chatbots; these systems have evolved into fully-fledged advisory platforms capable of delivering product recommendations with contextual suggestions and product comparisons throughout the shopping journey.

The research is intended to address three critical limitations faced by Conversational AI Shopping Assistants currently:

1. Data freshness and synchronization, where existing systems have trouble keeping product information up to date in real-time, leads to loss of user frustrations and loss of trust [16, 17]. Wang et al. [10] emphasized that users experience trust concerns due to inaccurate product details.
2. Context that comprehends the context of the conversation in which the assistants cannot keep the context in long conversations and gives vague and impersonal recommendations [18, 19]. Chen and Liu [18] show that current systems retain minimal context after 3-4 conversation turns.
3. The absence of standard metrics to evaluate the success of shopping assistants results in diminishing return on investment, especially when the user experience is ridden with latency and lacks observability of the solution [20, 21]. In their recent work, Kumar and Zhang [20] emphasize the dire need for benchmark evaluation frameworks.

Considering that research gap, we introduce a comprehensive framework to measure and improve the performance of AI Shopping Assistants [22]. The metrics proposed and trade-offs discussed are novel for the eCommerce domain and represent a holistic avenue for commercialization for the AI assistants to balance user trust [23, 24]. AI Shopping Assistants are increasingly becoming more advanced with Generative AI, memory, and long

context windows. However, there is a tension between direct access to real-time data sources (e.g., latest pricing, inventory and product definition) and improvements in natural language understanding. This results in irrelevant recommendations, wrong information and a decline in customer trust even as these Gen AI based solutions proliferate.

According to industry reports [10], at least 65% of drop-off in order completion can be attributed to finding erroneous or weak information offered by AI assistants. This study aims to fulfill three key objectives: First, this study performs a broad evaluation of shopping assistance through generative AI across both technological capabilities [5, 6] and real-world limitations. Second, it creates an organized framework to quantify and assess performance through technical precision benchmarks and user-focused engagement metrics.

Third, it presents creative approaches to improving shopping assistant performance when the use of real-time data is limited. This research builds upon the Natural Conversation Framework (NCF) [3], which synthesizes principles from conversation analysis [11, 12] and interaction studies [2, 7] to outline well-defined principles for context-aware and user-adaptive conversational systems. Through these use cases, this study formulates approaches for maintaining context in conversation, keeping response relevant and building trust with a user as it pertains to shopping assistance.

2. Current State of Shopping Assistants

2.1. How Shopping Assistance Has Changed

Shopping assistants are among the most substantial shifts in e-commerce technology — and have evolved through phases of development. Systems initially used simple rule-based logic, with narrow decision trees that matched user keywords in predetermined questions to provide short pre-populated answers [8]. Then came the ubiquitous AI-based systems, bringing a paradigm shift and providing Natural Language Processing (NLP) features and in-depth context comprehension. Today, AI shopping assistants employ complex methods and advanced machine learning models that can parse more complicated language structures and comprehend nuances in context and continuity across many exchanges. They now include sentiment analysis, intent detection and on the fly response creation. They can respond to complex issues such as customer product queries [9]. See Figure 1 for details. Natural Conversation Framework (NCF) patterns have also been integrated into these capabilities so that assistants can carry context across several turns, de-reference implicit references, and produce contextually relevant responses [3]. The most important challenge, however, is data synchronization. Even the most sophisticated AI systems can only give outdated and incorrect information without real-time access to product information, leaving a big trust gap between system capabilities and user expectations. This limitation implies the need to be cautious when doing caching and confidence scoring for the service to remain reliable [4].

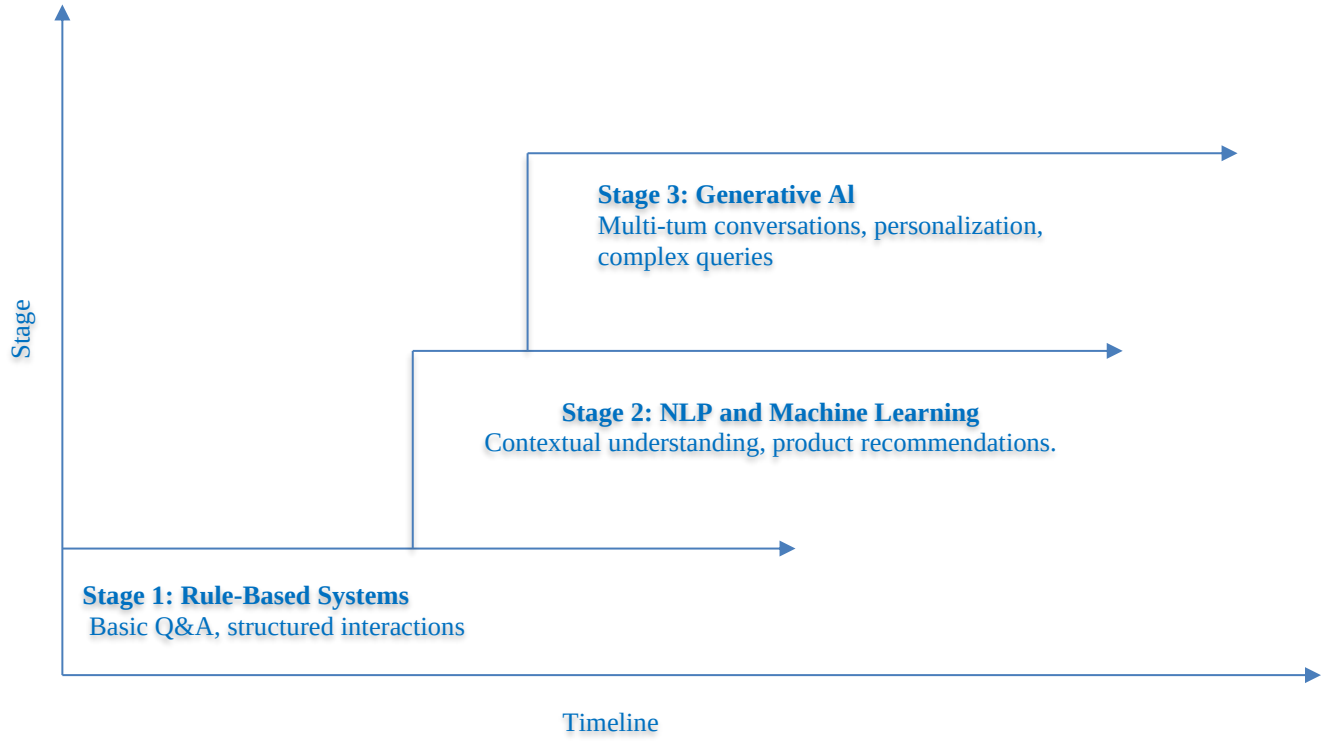

Fig. 1 Evolution of shopping assistants

2.2. Important Features and User Expectations

Modern shopping assistants must satisfy an intricate matrix of user needs, including product discovery, advanced comparison, inventory checking, and attribute analysis. Generative AI drove up the bar on user expectations that information should be delivered and interpreted according to each user's needs and personalized recommendation strategies [8]. By retaining conversation history and keeping track of users (with constant learning of what users prefer/like), these systems are bound to prove that they understand the context well while using smart filter algorithms for product discovery matching. First-generation assistants provide only static price comparisons and simple technical specifications.

However, modern assistants are expected to deliver dynamic price analysis, use conversational language in commercial terms, and provide instant inventory intelligence along with alternative recommendations if the given items are unavailable [8]. In line with NCF principles, these assistants personalize their communication mode, information richness and recommendation approach according to the patterns of user behavior and stated preferences. Such a solution is enabled through high-level user modeling abilities and dynamic answer generation systems, resulting in an interactive experience that instils confidence in users as they sense that their needs are endorsed during the decision-making process [3]. The underlying success of these systems is more about implementing the right conversation-design principles, listening to users, and continually polishing response generation algorithms – balancing both technical capability and natural language interaction patterns [1].

3. Design Principles for Conversation AI Shopping Assistants

The design of AI shopping assistants must incorporate well-defined principles to help builders make trade-offs between technical and user capabilities and create conversational systems that are reliable, engaging, and trustworthy.

3.1. Core Architectural Principles

Content retention is the first architectural element, allowing shopping assistants to keep conversations coherent over several exchanges. Applications with content retention typically utilize state management, where the user preference, chat log and product context are remembered [4]. If a user asks subsequent questions about product variations or pricing variations, the system retains a reference to previously discussed items without the need for explicit repetition. Expandable sequence is the second principle, with which conversations can be dynamic, and users can progressively specify their queries as they look for products or services. This framework caters both to general questions ("Show me laptops") and fine-grained refinement ("How about gaming laptops below $1000?"). The research relies on a structured high-level abstraction with sufficient context to permit varying detail [7]. Deploying necessitates effective query understanding processes and adaptable response generation frameworks.

3.2. Providing Elements to the User

Personalized interactions should be built around content recipient behavior patterns and stated preferences. Using this approach, historical interactions are considered, and users' level of expertise is analyzed to generate more complex or less complex responses [8]. It must be able to keep track of user profiles and change how it communicates-technical specifications when addressing knowledgeable users and simplified descriptions for newcomers. The DCAF design framework consists of transparency and trust-building mechanisms. With a potentially stale information landscape or lack of access to data, systems need to state their confidence and freshness explicitly. This approach fosters trust among users, even though technical limitations exist [3]. Possessing use cases in place can mean displaying timestamps, recommendations, confidence values, and how often data updates occur.

3.3. Implementation Considerations at the Technical Level

Keeping the user busy and their heads in the game as much as possible while the system hits a limitation or failure mode must be designed into error recovery mechanisms. This involves implementing graceful degradation protocols, clear error communication, and alternative suggestion pathways. If exact matches are unavailable, the system should also allow users to reformulate their queries or present alternative products [10]. Performance optimization lets us utilize advanced functionality, but it comes at a tradeoff of response time.

It is imperative to consider approaches such as efficient caching, lightweight processing algorithms, and prioritized data access patterns. While Zhang et al. (2023) are mainly focused on systems, the system should be able to answer common queries of sub-second latency quickly and, at the same time, deal with complex product comparison and personalization tasks efficiently.

3.4. Integration and Monitoring

Ongoing system improvement is made possible by continuously integrating performance monitoring and feedback [8]. This involves Frequent analysis of user behavior patterns. This helped with Completion rates and other user satisfaction metrics. Systematic analysis and collection of Fails Automated performance tuning routines Operating in harmony, these design principles enable shopping assistants to be technically functional and create compelling, trustworthy, and fulfilling shopping experiences. The better they are implemented, the more they affect user satisfaction indicators and system adoption rates.

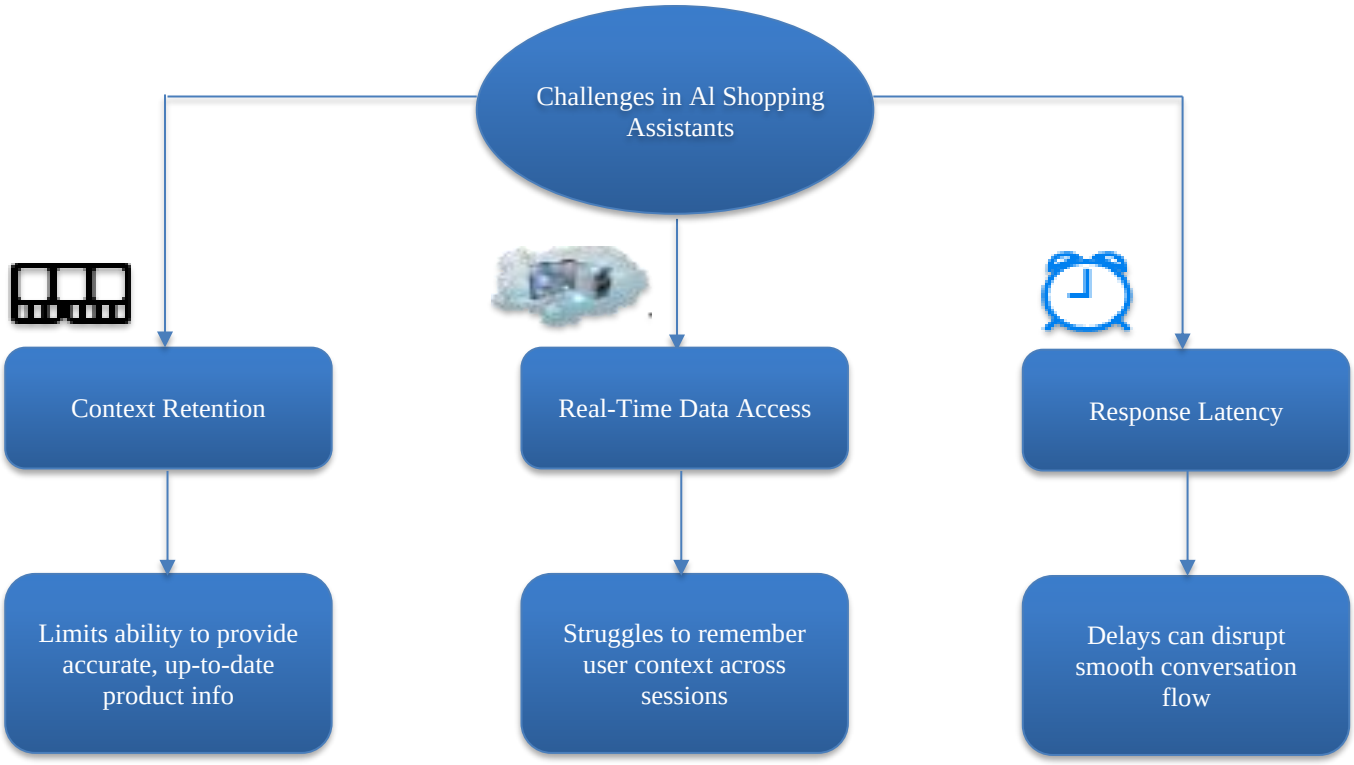

Fig. 2 Key challenges in using generative AI for shopping

4. Current Gap and Challenges

4.1. Accuracy of Knowledge and User Trust

In shopping conversations, users usually narrow down their choices, compare items to each other, or request more details about a particular item. NCF also facilitates sequence-level management patterns, so the AI assistant tracks the user's questions and preferences throughout the conversation to answer follow-up questions, like, "Does it come in blue?" that provide negative experiences. Even if the user directly uses a different term for that exact product. Such a technique encourages the assistant to become more conversational and fun to interact with [13].

4.2. Understanding Context and Improving Engagement

In shopping conversations, users usually narrow down their choices, compare items to each other, or request more details about a particular item. NCF also facilitates sequence-level management patterns, so the AI assistant tracks the user's questions and preferences throughout the conversation to answer follow-up questions, like, "Does it come in blue?" that provide negative experiences. even if the user directly uses a different term for that exact product. Such a technique encourages the assistant to become more conversational and fun to interact with [13].

4.3. Technical Constraints

Shopping assistants without access to current data find it hard to provide information. To deal with this, the NCF's pattern language recommends splitting responses into smaller pieces for the AI to answer quickly, thus giving a natural flow to the conversation. A technique such as model distillation to downsize the AI would also result in shorter response times, maintaining requirements for an assistant to be useful and responsive during fast-paced shopping [13]. (See Figure 2 for the key challenges in using Generative AI for shopping)

5. Measurement Framework

This paper introduces a framework for evaluating shopping assistants by measuring technical accuracy and user-centered engagement. This combination of metrics aligns with the NCF principles and shows how well the assistant performs overall.

5.1. Accuracy Metrics

5.1.1. Product Information Accuracy (PIA)

This measures how well the assistant provides accurate, current product details, focusing on price, stock, and features. It prioritizes the information that matters most to the user.

5.1.2. Price Range Accuracy (PRA)

This measures how accurately the assistant can quote prices, including acceptable ranges for variations.

5.1.3. Feature Description Accuracy (FDA)

This evaluates how well the assistant describes product features, ensuring the descriptions match official product information.

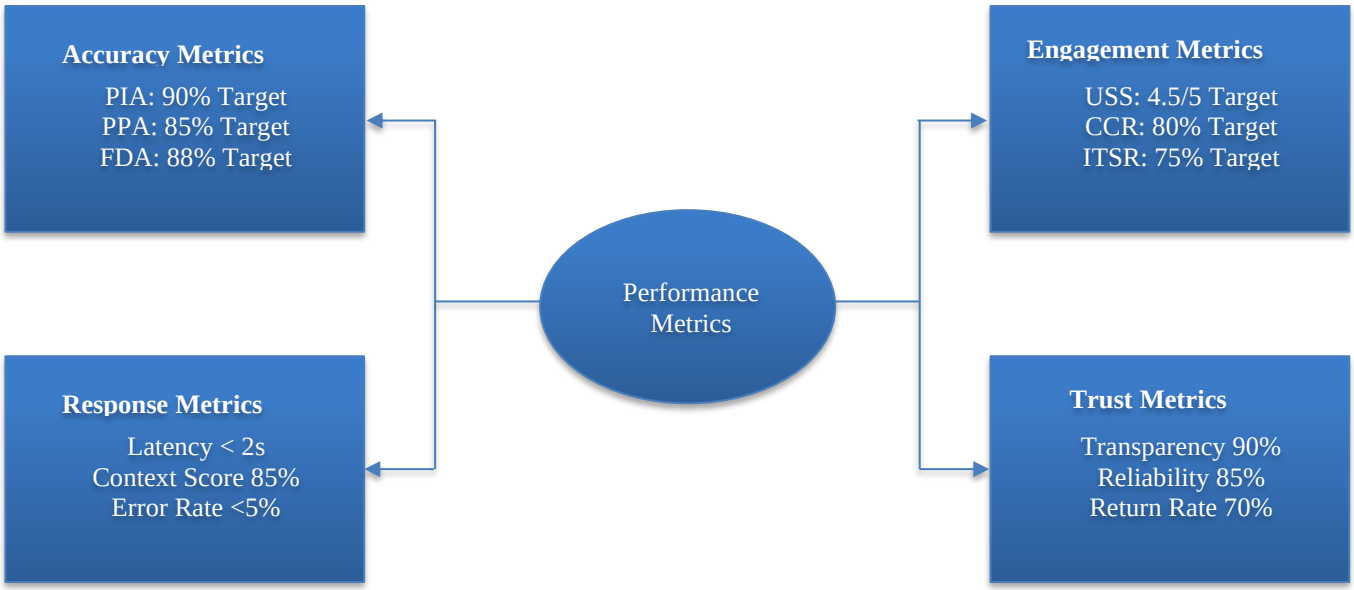

Fig. 3 Conversational AI shopping assistant performance framework

5.2. User-Centered Engagement Metrics

5.2.1. User Satisfaction Score (USS)

This metric captures how happy users feel after interacting with the assistant, providing insight into the overall experience.

5.2.2. Conversation Completion Rate (CCR)

This tracks how often user questions are fully answered, clearly measuring the assistant's effectiveness.

5.2.3. Interaction Time and Session Retention (ITSR)

This measures the average time users spend in a session and how likely they are to return, reflecting trust and satisfaction. High retention usually means users find the assistant helpful.

6. Comparative Analysis of Measurement Approaches

Table 1 shows how the framework proposed is novel since it extends beyond existing approaches by introducing:

1. Caching mechanisms to ensure real-time data availability
2. Context awareness and state management
3. Adaptive behavior of the model based on user responses

Table 1. Comparison of shopping assistant frameworks

Framework	Real-time Data	Context Retention	User Adaption
Rule-Based	Limited	None	Static
ML-Based	Moderate	Partial	Partially Dynamic
NCF-Based	Advanced	Full	Dynamic
Proposed	Advanced	Full	Adaptive

7. Common Method Bias and Implications

Common Method Bias (CMB) is a concern in empirical research. This bias arises from the measurement method rather than the constructs the measures are intended to represent, potentially threatening the validity of research findings. In the proposal context, CMB is critical due to this study's multifaceted data sources and evaluation approaches [25].

1. *Single Source of Data Collection:* Many evaluations of conversational AI rely heavily on user feedback collected through surveys or interviews, which can introduce biases such as social desirability bias or respondent fatigue [26].

2. *Homogeneous Measurement Contexts:* When performance metrics for conversational AI are collected exclusively within controlled testing environments or simulated scenarios, results may not reflect the variability and unpredictability of real-world contexts [27].

3. *Temporal and Procedural Influences:* Assessments conducted over short time frames or using similar procedural formats (e.g., identical question phrasing, response formats, or scales) can amplify common method variance, conflating actual performance differences with measurement artifacts [28, 29].

To address these biases, several measures were implemented:

1. *Triangulation of Data Sources:* We incorporated multiple data streams, including user-reported feedback, system-generated logs, and expert evaluations, to ensure a holistic and unbiased assessment of shopping assistant performance [28].

2. *Temporal and Spatial Variation:* Performance data were collected across time frames, including peak and non-peak shopping periods, to account for context-dependent variations. This approach helps to minimize procedural uniformity and enhance the generalizability of findings [29].

3. *Post-Hoc Statistical Adjustments:* Statistical techniques such as Harman's single-factor test and partial correlation analysis were applied to detect and account for common method variance in the collected data. This allowed us to isolate and examine the true relationships between constructs of interest [26].

4. *Designing Robust Survey Instruments:* Surveys and questionnaires were designed to include reverse-coded items and varying question formats to mitigate response biases and encourage thoughtful participant engagement [27].

8. Opportunities of Improvement

Here are some strategies this study recommends for making generative AI fit shopping-specific needs:

Periodic Data Broadcast and Caching

Periodic data updates and caching (copying of the data being kept) can avoid out-of-date information problems, which enables assistants to become more reliable [3].

Reduction of Model Complexity for Quicker Response

Simplifying the model with techniques such as model distillation makes responses quicker, rendering the interaction more conversational and efficient [2].

Context retention for personalization

Due to NCF's recipient design approach, the assistant can remember the user across sessions [4], enabling more tailored-feeling conversations in the future.

Feedback Loops for Continuous Learning

Users provide feedback on the experience, enabling the assistant to learn and grow. This process reflects the Testing Phase of NCF, whereby the assistant continuously trains based on real users' events [7].

Conflicts of Interest

There is no conflict of interest regarding the publication of this paper.

References

[1] Ayse Pinar Saygin, and Ilyas Cicekli, "Pragmatics in Human-Computer Conversations," *Journal of Pragmatics*, vol. 34, no. 3, pp. 227-258, 2002. [CrossRef] [Google Scholar] [Publisher Link]

[2] Harold Garfinkel, *Studies in Ethnomethodology*, 3rd ed., Social Theory Re-Wired, pp. 1-9, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[3] Robert J. Moore, Raphael Arar, *Conversational UX Design: A Practitioner's Guide to the Natural Conversation Framework*, ACM Books, Association for Computing Machinery, New York, NY, United States, pp. 1-320, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[4] Emanuel A. Schegloff, Gail Jefferson, and Harvey Sacks, *The Preference for Self-Correction in the Organization of Repair in Conversation*," Language, Linguistic Society of America, vol. 53, no. 2, pp. 361-382, 1977. [CrossRef] [Google Scholar] [Publisher Link]

[5] A. M. Turing, "Computing Machinery and Intelligence," *Mind*, vol. 59, no. 236, pp. 433-460, 1950. [CrossRef] [Google Scholar] [Publisher Link]

[6] Joseph Weizenbaum, "ELIZA-A Computer Program for The Study of Natural Language Communication Between Man and Machine" *Communications of the ACM*, vol. 9, no. 1, pp. 36-45, 1966. [CrossRef] [Google Scholar] [Publisher Link]

[7] Ludwig Wittgenstein, *Philosophical Investigations: The German Text, with a Revised English Translation 50th Anniversary Commemorative Edition*, 3thed., Wiley, pp. 1-464, 1991[Google Scholar] [Publisher Link]

[8] Xinmeng Song, and Ting Xiong, "A Survey of Published Literature on Conversational Artificial Intelligence," *2021 7th International Conference on Information Management (ICIM)*, London, United Kingdom, pp. 113-117, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[9] Xusen Cheng et al., "Exploring Consumers Response to Text-Based Chatbots in E-commerce: The Moderating Role of Task Complexity and Chatbot Disclosure," *arXiv*, pp. 1-22, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[10] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan, "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems," *ACM Transactions on Interactive Intelligent Systems*, vol. 11, no. 3-4, pp. 1-45, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Harvey Sacks, Emanuel A. Schegloff and Gail Jefferson, *A Simplest Systematics for the Organization of Turn-Taking for Conversation*, Language, Linguistic Society of America, vol. 50, no. 4, pp. 696-735, 1974. [CrossRef] [Google Scholar] [Publisher Link]

[12] Harold Garfinkel, *Studies in Ethnomethodology*, Prentice-Hall, Englewood Cliffs, New Jersey, pp. 1-288, 1967. [Google Scholar] [Publisher Link]

[13] Emanuel A. Schegloff, *The Relevance of Repair to Syntax-for-Conversation*, Syntax and Semantics, Discourse and Syntax, New York: Academic Press vol. 12, pp. 261-286, 1979. [Google Scholar] [Publisher Link]

[14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the Knowledge in a Neural Network," *Arxiv*, pp. 1-9, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[15] Emilia Lesiak et al., "Digital Assistant in a Point of Sales," *arXiv*, pp. 1-15, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[16] Four Ways Conversational AI Transforms Digital Experiences, Grid Dynamic, [Online]. Available: https://www.griddynamics.com/blog/conversational-ai-ecommerce

[17] FAN Min et al., "Research on Users' Trust of Chatbots Driven by AI: An Empirical Analysis Based on System Factors and User Characteristics," *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, Guangzhou, China, vol. 66, no. 3, pp. 215-228, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[18] Salla Syv€anen, and Chiara Valentini, "Conversational Agents in Online Organization–Stakeholder Interactions: A State-of-the-Art analysis and Implications for Further Research," *Journal of Communication Management*, vol. 24, no. 4, pp. 339-362, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[19] Arshit Gupta et al., "CASA-NLU: Context-Aware Self-Attentive Natural Language Understanding for Task-Oriented Chatbots," *arXiv*, pp. 1-7, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[20] Rodrigo Bavaresco et al., "Conversational Agents in Business: A Systematic Literature Review and Future Research Directions," *Computer Science Review*, vol. 36, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[21] Justina Sidlauskiene, Yannick Joye, and Vilte Auruskeviciene, "AI-based Chatbots in Conversational Commerce and Their Effects on Product and Price Perceptions," *Electronic Markets*, vol. 33, pp. 1-21, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[22] Vai Rawool, Pantea Foroudi, and Maria Palazzo, "AI-Powered Voice Assistants: Developing A Framework for Building Consumer Trust and Fostering Brand Loyalty," *Electronic Commerce Research*, pp. 1-33, 2024.. [CrossRef] [Google Scholar] [Publisher Link]

[23] Mateusz Dubiel, Sylvain Daronnat, and Luis A. Leiva, "Conversational Agents Trust Calibration: A User-Centred Perspective to Design," *Proceedings of the 4th Conference on Conversational User Interfaces (CUI 2022)*, Glasgow United Kingdom, pp. 1-6, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[24] S. Jayakrishnan, "Artificial Intelligence (AI) in Retailing-A Systematic Review and Research Agenda," *Contemporary Research in Management*, vol. 65, pp. 1-14, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[25] Podsakoff, et al., "Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies," *Journal of Applied Psychology*, vol. 88, no. 5, pp. 879-903, 2003. [CrossRef] [Google Scholar] [Publisher Link]

[26] Sea-Jin Chang, Arjen van Witteloostuijn, and Lorraine Eden, "From the Editors: Common Method Variance in International Business Research," *Journal of International Business Studies*, vol. 41, no. 2, pp. 178-184, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[27] Hettie A. Richardson, Marcia J. Simmering, and Michael C. Sturman, "A Tale of Three Perspectives: Examining Post Hoc Statistical Techniques for Detection and Correction of Common Method Variance," *Organizational Research Methods*, vol. 12, no. 4, pp. 762-800, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[28] J. Williams Larry, K. Brown Barbara, "Method Variance in Organizational Behavior and Human Resources Research: Effects on Correlations, Path Coefficients, and Hypothesis Testing," *Organizational Behavior and Human Decision Processes*, vol. 57, no. 2, pp. 185-209, 1994. [CrossRef] [Google Scholar] [Publisher Link]

[29] Ned Kock, "Common Method Bias in PLS-SEM: A Full Collinearity Assessment Approach," *International Journal of e-Collaboration (IJEC)*, vol. 11, no. 4, pp. 1-10, 2015. [CrossRef] [Google Scholar] [Publisher Link]