Measure only what is measurable: towards conversation requirements for evaluating task-oriented dialogue systems

Emiel van Miltenburg[1], Anouck Braggaar[1], Emmelyn Croes[1], Florian Kunneman[2], Christine Liebrecht[1], Gabriëlla Martijn[2]

[1]Tilburg University, [2]Utrecht University

Correspondence: C.W.J.vanMiltenburg@tilburguniversity.edu

Abstract

Chatbots for customer service have been widely studied in many different fields, ranging from Natural Language Processing (NLP) to Communication Science. These fields have developed different evaluation practices to assess chatbot performance (e.g., *fluency, task success*) and to measure the impact of chatbot usage on the user's perception of the organisation controlling the chatbot (e.g., *brand attitude*) as well as their willingness to enter a business transaction or to continue to use the chatbot in the future (i.e., *purchase intention, reuse intention*). While NLP researchers have developed many automatic measures of success, other fields mainly use questionnaires to compare different chatbots. This paper explores the extent to which we can bridge the gap between NLP and Communication Science, and proposes a research agenda to further explore this question.

1 Introduction

We need to talk about measurement requirements. There is a vast body of literature on the evaluation of dialogue systems, with a wide range of different methods to assess different properties of the conversations that people have with chatbots and the impressions that are formed during those conversations. The goal of many Natural Language Processing (NLP) researchers seems to be to avoid asking people about their experiences because human evaluation is costly and time-consuming. However, most of the literature rests on the assumption that the properties that we are interested in are measurable from the conversational data. We question this assumption and ask: **to what extent do chatbot conversations contain useful cues to determine conversation quality (and beyond)?** Although this question is relevant for all kinds of chatbots, we will focus on task-oriented dialogue systems. As we will argue, NLP researchers mostly focus on intrinsic properties of these systems (§2) while organisations are often more interested in extrinsic evaluation (§3). An open challenge in dialogue research is to predict the users' opinion of the system based *only* on their conversation. To tackle this challenge, we believe it is essential to think about the requirements for us to be able to say more about what users think about chatbots. For this, we need to study conversational richness (§4,5). Based on an overview of the existing literature, we propose a roadmap for future research (§6).

2 Chatbot assessment in NLP

Let us first look at chatbots from a technological perspective. NLP researchers have a particular way of looking at the assessment of chatbots: they mostly care about the inner workings of the system and less on the effects that the system has on its users (see for example the work by Vijayaraghavan et al. (2020) which discusses different algorithms to evaluate separate components of dialogue systems). Common constructs of interest are, for example, *coherence* (of the conversation, e.g. Dziri et al. 2019), *robustness* (of the system, e.g. Cheng et al. 2019) *relevance* and *correctness* (of the generated utterances; discussed by Deriu et al. 2021). This leads us to:

Observation 1
NLP researchers tend to focus on constructs that are associated with intrinsic evaluation.

Where possible, NLP researchers tend to prefer automatic metrics to quantify system performance, since automatic approaches are generally cheaper and faster (Maroengsit et al., 2019). Depending on the purpose of their study, NLP researchers may even choose not to let their system interact with people at all, but rather to have the system engage in simulated (parts of) conversations (e.g., Vasconcelos et al. 2017; de Wit 2024). This allows them to compare how different systems respond to the same

Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM² 2025), pages 231–238
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

utterances. For example, researchers may investigate how *appropriate* the different responses are (e.g., Chen et al. 2023). When NLP researchers engage in human evaluation studies, they often do this through relatively superficial crowd-sourcing tasks, providing participants with responses in particular conversations, and asking questions about the quality of these individual responses (see e.g. Sedoc and Ungar (2020), who ask human annotators to select the better system answer to a specific prompt).[1] Additionally, NLP researchers have explored methods to automatically predict human judgements (Reddy, 2022; Wu and Chien, 2020; Deriu and Cieliebak, 2019). An interesting observation about this line of work is that nobody discusses the feasibility of the task; it is more or less assumed to be possible to predict the ratings, and the studies themselves show to what extent the authors' approach seems to work. Finally, NLP research typically does not specify the properties of conversations for which the proposed approach should work.[2]

3 The Communication Science perspective

Task-oriented chatbots are designed to be used, often by organisations aiming to alleviate the workload of their customer support agents. Communication scientists may study the use of chatbots in such an organisational context.[3]

The Communication perspective differs from that of NLP researchers. The review of Braggaar et al. (2023) shows that NLP typically does not assess either the attitude of the customer support agent or the user's attitude towards the brand. While NLP thus mostly seem to focus on the quality of the interaction, researchers in business communication are more interested in the impact of chatbot interactions on users' experience and their perceptions of the organization employing the chatbot. Subsequently, they also tend to focus more on evaluating the chatbot from the organisations' perspective. For example, research has explored chatbot implementation (e.g., Araujo et al. 2022) and chatbot collaboration (e.g., Martijn et al. 2024), as well as the distinction between the *drivers* of chatbot adoption and the *outcomes* of interacting with these systems (Mariani et al., 2023). Typical constructs of interest include *customer satisfaction* (e.g. Chung et al. 2020; Ruan and Mezei 2022), *user experience* (e.g. Chen et al. 2021; Trivedi 2019), *brand attitude* (e.g. Shahzad et al. 2024), *continuance usage intention* and *purchase intention* (e.g. Jiang et al. 2022; Li and Wang 2023; Akdemir and Bulut 2024). This leads us to:

Observation 2
Communication researchers focus on constructs associated with extrinsic evaluation.

Most research in this area relies on interviews or scenario-based experiments followed by questionnaires using validated scales. Yet, few studies have analysed the complete chatlogs from these interactions, which is a missed opportunity. A mixed-method approach that combines chatlog analysis with traditional surveys can yield a broader perspective on service quality by capturing both the internal dynamics of chatbot conversations and external customer perceptions (e.g., *customer experience, brand attitude, continuance usage intention*). Moreover, the current state of AI and NLP makes it possible to automate chatlog analysis and perhaps even predict evaluation scores.[4] However, for this to work, we need to consider media richness.

4 Media richness and chatbots

Media richness refers to the idea that our means of communication differ in the kind and number of cues that they can process (Daft and Lengel, 1986). For example, a text message does not carry any auditory information, whereas a telephone call does. Videoconferencing introduces visual information but may still lack other features of face-to-face conversations: *haptic cues* such as touch and smell, but also the *affordance* to interact with the real world and manipulate objects together. Researchers studying media richness may, for example, look at how the richness of different means of communi-

[1] Also note that there is typically only one question item per construct, leading to mono-operation bias.

[2] Another concern is that research on dialogue systems often uses controlled evaluations that often do not involve human participants. An example of this is work on conversational agents that guide a user through the different steps to prepare a meal. Although the context of a user standing in the kitchen is rather prominent for how a conversation may unfold and be appreciated by the user, performance on subtasks like intent detection, instruction ordering and response helpfulness is evaluated by comparing to an artificial dataset based on role-playing between crowd-workers (Le et al., 2023) or on a dataset augmented from user-system interactions that did not involve cooking (Glória-Silva et al., 2024).

[3] Business communication has three different application domains: business-to-consumer, business-to-business (also known by the acronym b2b), and internal communication. We focus on the business-to-consumer (b2c) domain.

[4] This goal is present in the literature at least since the introduction of the PARADISE framework (Walker et al., 1998), but it keeps re-appearing (e.g. recently in Ay et al. 2025).

Kind	Dimension	Potential values
Affordances	Form of interaction	Buttons, written, spoken, signed
	Available modalities	Text, audio, image, video
	Physical presence	Picture, moving avatar, embodied agent (i.e. face-to-face)
Implementation	Length of the interaction	Less than 5 turns, 5-10 turns, more than 10 turns
	Length of the responses	1-2 words, short phrases, full sentences, extended responses
	Conversational genre	Informational request, transaction, instruction, discussion
	Usage duration	Single interaction, repeated over a short period, extended use
	Scope	Narrow domain, broad domain, open domain

Table 1: Different dimensions that contribute to the richness of interactions with dialogue systems.

cation affects the kinds of interactions that people have, and the kinds of information that interlocutors are willing to disclose (Antheunis et al., 2012).

Traditionally, the media richness literature defines richness in terms of the ability a medium has to reproduce any given information. A criticism of this perspective is that the theory does not make any distinctions *within* a medium. This paper builds on the media richness literature and introduces different gradations in the richness of conversations that can arise within one single medium (in this case, human-chatbot interactions). We propose to consider the question of how rich an interactional setting needs to be before you can meaningfully analyse the interaction and draw conclusions about different kinds of constructs. These could be high-level constructs such as *customer satisfaction, brand attitude, reuse intention* and so on, but also lower-level constructs such as *fluency*; we just never seem to have any conversation about whether it makes sense to measure these constructs at all, based on the richness of the conversation.

A scale of interactional richness? Chatbots seem to exist on a scale of media richness. On the lower end, there are chatbots that are designed to answer queries as efficiently as possible, using mostly buttons or closed yes/no-questions. But conversations with such chatbots hardly contain any useful information about the user experience.[5] Thus we make the following observation:

Observation 3
Meaningful analyses require meaningful content; you cannot measure what is not measurable.

Fortunately, customer service chatbots have moved away from the rigid stereotype described above, and (informally) seem to be richer. Let us now operationalise what we mean by 'richness.'

What dimensions would be relevant to establish the richness of the interactions with a particular dialogue system? Table 1 provides a (preliminary) taxonomy, showing the axes along which we can measure the richness of any given chatbot. We make a general distinction between *affordances* (features that the system has, and that enable the user to carry out different actions) and *implementation* (how those features are used in practice), since the mere presence of particular affordances is not enough for a rich and satisfying conversation.

The dimensions in Table 1 are not equal; different dimensions may have different effects. For example, some dimensions are *facilitating* the conversation (e.g. form of interaction, available modalities) while others are *stimulating* the conversation and possibly *extending the range of topics* that may be discussed (e.g. length of the responses, scope). And some dimensions, such as *physical presence* may do both at the same time. More work needs to be done to establish a general framework to characterize the richness of chatbot interactions.

5 Conversation requirements

When installing a piece of software on a computer, the computer needs to meet a particular set of system requirements for the software to run properly. We do not yet have any equivalent to system requirements (perhaps we could call these 'conversation requirements') for evaluation metrics. That leads us to ask:

When is a conversation rich enough? Different constructs have different requirements that need to be fulfilled before they can be operationalized through behavioural data. As we said before: low-level constructs such as the fluency of the system

[5] There could be valid reasons for these design choices. Our point is that those choices have consequences for evaluation.

responses can already be measured in many cases. Through fully written conversations we may also be able to determine the smoothness of the conversation, and given the full conversation, we are also able to determine task success. But what about the user's mood or their level of satisfaction? What kind and amount of information do we need to establish these? And what else can we learn from conversations with chatbots?

Correlates of extrinsic constructs. We do not have to start from scratch. Researchers from different areas have found correlations between (non-)verbal behaviour and different mental states. For example, there is a long history in psychology of inferring writers' mental characteristics based on the words they use (Tausczik and Pennebaker, 2010). Researchers in the field of *affective computing* have worked to extract emotions from speech and facial expressions (for surveys, see: George and Muhamed Ilyas 2024; Ballesteros et al. 2024). Aside from emotions, other researchers have worked on audiovisual cues that signal uncertainty (e.g. Krahmer and Swerts 2005), which may be a good indicator for when users are confused about the actions of the chatbot. Similarly, previous work has compared textual cues of engagement to self-reported engagement, demonstrating that utterance level cues can predict engagement in chatbot conversations (He et al., 2024).[6] This leads us to:

Observation 4
There may be hope:
a. Proxies or antecedents of extrinsic constructs *can* be measured from interaction data.
b. Different researchers are working to identify relations between relevant variables (e.g. *language use* and *level of confusion*)

Of course, this kind of research is not without its drawbacks. Tausczik and Pennebaker (2010) are the first to admit that the relation between texts and authors' mental states is very complex, and existing text analysis methods are still relatively crude. Furthermore, Barrett et al. (2019) note that the relation between facial expressions and experienced[7] emotions may not be universal, and so it may be hard to draw reliable conclusions based on visual features alone.[8] Finally, we again emphasize that the question is not just about whether one can in principle make the connection between what someone says and how they feel. We should also look at how much text, video, or audio is needed in order to make any inference at all, and how much data is needed for that process to be reliable.

6 Discussion

In summary, we propose that research on evaluation metrics should pay more attention to the requirements for those metrics to work properly. These requirements should be operationalised using different richness dimensions, along the lines of Table 1 (presented on the previous page).

6.1 Research agenda

We propose the following research agenda. Evaluation researchers should: 1. Develop a standardized way to quantify the richness of chatbot interaction designs. 2. Investigate how and to what extent different properties of the conversations are related to constructs of interest, i.e. intrinsic evaluation targets and extrinsic business and communication objectives. 3. Establish basic requirements for different evaluation metrics. (If these requirements are not met, other approaches such as questionnaires should be used.) However, these goals are not without challenges. We discuss these below.

6.2 Is standardisation feasible?

The lack of standardization in NLP is a major challenge to the development of conversation requirements for evaluation metrics. There is a great deal of variation in the terminology used and the methods applied to evaluate natural language generation systems (Howcroft et al., 2020; Schmidtova et al., 2024) and task-oriented dialogue systems (Braggaar et al., 2023). How can we agree on the requirements for different evaluation metrics if we do not even agree on the relevant terminology and the way those metrics should be applied?

Reasons for optimism The recent observations about terminological and methodological confusion in our field make the standardisation of our evaluation practices seem like a daunting task. Still, the publication of these studies is a *good* sign: the field is changing and people are paying attention to the improvement and standardisation of our evaluation practices (the GEM workshop is another

[6] Again, the conversational data does need to be rich enough to be able to carry out such analyses.

[7] As opposed to *perceived* emotions that may only exist in the eyes of the observer.

[8] Given the lure of 'mind reading software' we need to be careful here, not least because of the ethical implications of such technology, but also from a purely scientific standpoint: extraordinary claims require extraordinary evidence.

case in point). Indeed the recent work of Belz et al. (2020, 2024) and Fitrianie et al. (2025) shows that we are making progress in the standardisation of terminology and approaches. Now we need to push through and determine when and how these approaches can be used. Another reason to be optimistic is that the conversational capacities of chatbots has been improving. This gives us another way forward.

6.3 Designing conversations for user insights

Our discussion so far has focused on conversation requirements for evaluation metrics, but we have not discussed the idea that we could also enrich conversations to make it easier to measure user engagement. The core question is this: *how can we ask users about their experiences, without asking them about their experiences?* Conversation designers may be able to implement conversational cues that invite the user to engage more with the chatbot, or at least to provide responses that indicate their stance towards the chatbot and the current conversation. We can then measure users' actual level of engagement more easily.

The use of **invitational rhetoric** may be useful to prompt users to respond in a particular way. Liebrecht et al. (2021) define six different ways in which organisations may elicit responses from chatbot users. These range from explicit questions (asking for feedback) to apologies (*sorry!*) and well-wishing (*have a good day!*) that may prompt users to respond in kind (*no problem; you too!*). Traditionally, this kind of rhetoric was introduced for users to perceive chatbots as more warm and human-like, which in turn might improve the users' brand attitude and purchase intention (e.g. Liebrecht and van der Weegen 2019). But a welcome side-effect of invitational rhetoric is that we may gain some insight into the users' thoughts through their responses.

6.4 Do current systems support rich dialogue?

It is currently unknown to what extent existing chatbots support or stimulate rich conversations. Future research should investigate the richness of chatbots that are currently deployed, so that we have a better sense of the kind of dialogue that is elicited by existing systems and the ways in which these systems actively stimulate conversation. This would serve two purposes. First, this would help to expand our taxonomy to better capture the ways in which chatbots may facilitate rich conversations. Second, we would have a better understanding of the context in which evaluation metrics may be deployed in the real world, and the limitations that are posed by the way the conversations are designed. The overview studies of Chaves and Gerosa (2021) and Janssen et al. (2020) are a good starting point, but then we still need to determine the extent to which different design characteristics support rich conversations.

6.5 Assessing richness

While it is relatively easy to gauge the richness of rule-based dialogue systems, it is harder to do the same for LLM-based chatbots. For example, with rule-based systems we can check how often the system asks closed versus open questions, and how often the opportunity arises for users to provide meaningful answers (where 'meaningful' could be defined as the extent to which the answer provides insight into the user's engagement and stance towards the system). For LLM-based systems it is not immediately clear how we could measure the extent to which the system offers users the opportunity to show their engagement in the conversation, particularly since it is notoriously hard to evaluate multi-turn interactions (Laban et al., 2025).

7 Conclusion

This paper has discussed the evaluation of chatbots from two perspectives, NLP and Communication Science. Communication Scientists tend to focus more on constructs that NLP researchers would consider extrinsic: a shift in BRAND ATTITUDE may be a *consequence* of an interaction with a chatbot, whereas intrinsic constructs such as FLUENCY are assumed to be measurable on the basis of interactions with the chatbot. We may be able to predict a user's BRAND ATTITUDE if the conversation is rich enough to contain clues about the user's stance towards the organisation that the chatbot represents. But when we take a step back, we have to acknowledge that this also holds for the measurement of FLUENCY and so many other intrinsic constructs that NLP researchers seem to take for granted. When can we meaningfully assess any property? Different constructs will have different conversation requirements, but we always have to take conversational richness into account.

Acknowledgments

This paper is part of the NWO-funded Smooth Operators project (KIVI.2019.009).

Ethical considerations

Although our paper aims to advance a theoretical discussion on the requirements for us to make valid measurements, the paper also touches on the idea that we may predict the mental state of chatbot users, specifically variables such as *customer satisfaction, purchase intention, brand attitude*. This idea should be handled with care. One way to reduce the risk of misuse is to work towards aggregate metrics that capture the *distribution of user experiences* and *common causes of those experiences* rather than predicting specific properties of individual users (Baldridge, 2017). We do not need to know any intimate details about the users; we just want to know how to improve the overall user experience for people interacting with chatbots.

References

Dilek Merve Akdemir and Zafer Arslan Bulut. 2024. Business and customer-based chatbot activities: The role of customer satisfaction in online purchase intention and intention to reuse chatbots. *Journal of Theoretical and Applied Electronic Commerce Research*, 19(4):2961–2979.

Marjolijn L. Antheunis, Alexander P. Schouten, Patti M. Valkenburg, and Jochen Peter. 2012. Interactive uncertainty reduction strategies and verbal affection in computer-mediated communication. *Communication Research*, 39(6):757–780.

Theo Araujo, Ward van Zoonen, and Claartje ter Hoeven. 2022. "a large playground": Examining the current state and implications of conversational agent adoption in organizations. *International Journal of Innovation and Technology Management*, 19(07):2250024.

Fehime Ceren Ay, Eleonora Freddi, Asbjørn Følstad, Stig Hodnebrog, Knut Kvale, Olav Alexander Sell, and Simen Ulsaker. 2025. Conversation logs as a source of insight: predicting user satisfaction for customer service chatbots. *Quality and User Experience*, 10(1):1.

Jason Baldridge. 2017. Practical and ethical considerations in demographic and psychographic analysis. Presented as a keynote at the First ACL Workshop on Ethics in Natural Language Processing. https://nlp.stanford.edu/seminar/details/jbaldridge.pdf.

Jesús A. Ballesteros, Gabriel M. Ramírez V., Fernando Moreira, Andrés Solano, and Carlos A. Pelaez. 2024. Facial emotion recognition through artificial intelligence. *Frontiers in Computer Science*, 6.

Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. 2019. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68. PMID: 31313636.

Anya Belz, Simon Mille, and David M. Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194, Dublin, Ireland. Association for Computational Linguistics.

Anya Belz, Simon Mille, Craig Thomson, and Rudali Huidrom. 2024. QCET: An interactive taxonomy of quality criteria for comparable and repeatable evaluation of NLP systems. In *Proceedings of the 17th International Natural Language Generation Conference: System Demonstrations*, pages 9–12, Tokyo, Japan. Association for Computational Linguistics.

Anouck Braggaar, Christine Liebrecht, Emiel van Miltenburg, and Emiel Krahmer. 2023. Evaluating task-oriented dialogue systems: A systematic review of measures, constructs and their operationalisations. *arXiv preprint arXiv:2312.13871*.

Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How should my chatbot interact? a survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction*, 37(8):729–758.

Bao Chen, Yuanjie Wang, Zeming Liu, and Yuhang Guo. 2023. Automatic evaluate dialogue appropriateness by using dialogue act. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7361–7372, Singapore. Association for Computational Linguistics.

Ja-Shen Chen, Tran-Thien-Y Le, and Devina Florence. 2021. Usability and responsiveness of artificial intelligence chatbot on online customer experience in e-retailing. *International Journal of Retail & Distribution Management*, 49:1512–1531.

Minhao Cheng, Wei Wei, and Cho-Jui Hsieh. 2019. Evaluating and enhancing the robustness of dialogue systems: A case study on a negotiation agent. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3325–3335, Minneapolis, Minnesota. Association for Computational Linguistics.

Minjeong Chung, Eunju Ko, Hye Jung Joung, and Seung Joo Kim. 2020. Chatbot e-service and customer satisfaction regarding luxury brands. *Journal of Business Research*, 117:587–595.

Richard L. Daft and Robert H. Lengel. 1986. Organizational information requirements, media richness and structural design. *Management Science*, 32(5):554–571.

Jan de Wit. 2024. Leveraging large language models as simulated users for initial, low-cost evaluations of designed conversations. In *Chatbot Research and Design*, pages 77–93, Cham. Springer Nature Switzerland.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54:755–810.

Jan Milan Deriu and Mark Cieliebak. 2019. Towards a metric for automated conversational dialogue system evaluation and improvement. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 432–437, Tokyo, Japan. Association for Computational Linguistics.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3806–3812, Minneapolis, Minnesota. Association for Computational Linguistics.

Siska Fitrianie, Merijn Bruijnes, Amal Abdulrahman, and Willem-Paul Brinkman. 2025. The artificial social agent questionnaire (asaq) — development and evaluation of a validated instrument for capturing human interaction experiences with artificial social agents. *International Journal of Human-Computer Studies*, 199:103482.

Swapna Mol George and P. Muhamed Ilyas. 2024. A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise. *Neurocomputing*, 568:127015.

Diogo Glória-Silva, Rafael Ferreira, Diogo Tavares, David Semedo, and João Magalhães. 2024. Plan-grounded large language models for dual goal conversational settings. *arXiv preprint arXiv:2402.01053*.

Linwei He, Anouck Braggaar, Erkan Basar, Emiel Krahmer, Marjolijn Antheunis, and Reinout Wiers. 2024. Exploring user engagement through an interaction lens: What textual cues can tell us about human-chatbot interactions. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, CUI '24, New York, NY, USA. Association for Computing Machinery.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Antje Janssen, Jens Passlick, Davinia Rodríguez Cardona, and Michael H. Breitner. 2020. Virtual assistance in any context: A taxonomy of design elements for domain-specific chatbots. *Business & Information Systems Engineering*, 62(3):211–225.

Kaiyuan Jiang, Min Qin, and Shuo Li. 2022. Chatbots in retail: How do they affect the continued use and purchase intentions of chinese consumers? *Journal of Consumer Behaviour*, 21(4):756–772.

Emiel Krahmer and Marc Swerts. 2005. How children and adults produce and perceive uncertainty in audiovisual speech. *Language and Speech*, 48(1):29–53. PMID: 16161471.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. Llms get lost in multi-turn conversation. *Preprint*, arXiv:2505.06120.

Duong Minh Le, Ruohao Guo, Wei Xu, and Alan Ritter. 2023. Improved instruction ordering in recipe-grounded conversation. *arXiv preprint arXiv:2305.17280*.

Ming Li and Rui Wang. 2023. Chatbots in e-commerce: The effect of chatbot language style on customers' continuance usage intention and attitude toward brand. *Journal of Retailing and Consumer Services*, 71:103209.

Christine Liebrecht, Christina Tsaousi, and Charlotte van Hooijdonk. 2021. Linguistic elements of conversational human voice in online brand communication: Manipulations and perceptions. *Journal of Business Research*, 132:124–135.

Christine Liebrecht and Evi van der Weegen. 2019. Menselijke chatbots: een zegen voor online klantcontact? *Tijdschrift voor Communicatiewetenschap*, 47(3).

Marcello M. Mariani, Novin Hashemi, and Jochen Wirtz. 2023. Artificial intelligence empowered conversational agents: A systematic literature review and research agenda. *Journal of Business Research*, 161:113838.

Wari Maroengsit, Thanarath Piyakulpinyo, Korawat Phonyiam, Suporn Pongnumkul, Pimwadee Chaovalit, and Thanaruk Theeramunkong. 2019. A survey on evaluation methods for chatbots. In *Proceedings of the 2019 7th International conference on information and education technology*, pages 111–119.

Gabriella Martijn, Charlotte Van Hooijdonk, Florian Kunneman, and Hans Hoeken. 2024. Reconfiguring the customer service domain: Perspectives of managers, conversational designers, and human agents on human–chatbot collaboration. *International Journal of Innovation and Technology Management*, 21(04):2450028.

Sujan Reddy. 2022. Automating human evaluation of dialogue systems. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 229–234.

Yuwei Ruan and József Mezei. 2022. When do ai chatbots lead to higher customer satisfaction than human frontline employees in online shopping assistance? considering product attribute type. *Journal of Retailing and Consumer Services*, 68:103059.

Patricia Schmidtova, Saad Mahamood, Simone Balloccu, Ondrej Dusek, Albert Gatt, Dimitra Gkatzia, David M. Howcroft, Ondrej Platek, and Adarsa Sivaprasad. 2024. Automatic metrics in natural language generation: A survey of current evaluation practices. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 557–583, Tokyo, Japan. Association for Computational Linguistics.

João Sedoc and Lyle Ungar. 2020. Item response theory for efficient human evaluation of chatbots. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 21–33, Online. Association for Computational Linguistics.

Muhammad Farooq Shahzad, Shasha Xu, Xiaolong An, and Imran Javed. 2024. Assessing the impact of ai-chatbot service quality on user e-brand loyalty through chatbot user trust, experience and electronic word of mouth. *Journal of Retailing and Consumer Services*, 79:103867.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Jitender Trivedi. 2019. Examining the customer experience of using banking chatbots and its impact on brand love: The moderating role of perceived risk. *Journal of Internet Commerce*, 18(1):91–111.

Marisa Vasconcelos, Heloisa Candello, Claudio Pinhanez, and Thiago dos Santos. 2017. Bottester: testing conversational systems with simulated users. pages 1–4.

Varadharajan Vijayaraghavan, Jack Brian Cooper, and Rian Leevinson J. 2020. Algorithm inspection for chatbot performance evaluation. *Procedia Computer Science*, 171:2267–2274. Third International Conference on Computing and Network Communications (CoCoNet'19).

M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella. 1998. Evaluating spoken dialogue agents with PARADISE: Two case studies. *Computer Speech & Language*, 12(4):317–347.

Shih-Hung Wu and Sheng-Lun Chien. 2020. Learning the human judgment for the automatic evaluation of chatbot. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1598–1602.