AOSIS

Chatbot evaluation for effectiveness in customer query resolution

Check for updates

Authors:
Tryphosa B. Mashigo[1] ●
Wafeequa Dinath[1] ●
Sithembiso Khumalo[1] ●

Affiliations:
[1]Department of Information and Knowledge Management, College of Business and Economics, University of Johannesburg, Johannesburg, South Africa

Corresponding author:
Sithembiso Khumalo,
skhumalo@uj.ac.za

Background: The adoption of artificial intelligence technologies, specifically chatbots, has grown tremendously in various industries and is expected to transform how businesses communicate with and resolve customer queries. Yet, fewer empirical studies have been conducted on how chatbots can be assessed for their effectiveness in resolving customer queries.

Objectives: This study aims to uncover and identify research gaps through bibliometric analysis. Bibliometric analysis allows for the exploration of conversational chatbot evaluation in resolving customer queries.

Method: A comprehensive analysis of 27 literature articles published between 2015 and 2024 was conducted using data retrieved from the Web of Science database. The study encompasses various analytical approaches, such as performance analysis and science mapping techniques, and includes keyword co-occurrence analysis and citation network visualisation, to elucidate the distribution of publications, influential authors and institutions, and critical research disciplines.

Results: Findings reveal a growing body of research on chatbot evaluation predominantly in well-developed countries, spanning diverse disciplines such as Business and Economics, Computer Science, and Engineering. Examining citation networks suggests interconnectedness in the literature, with specific articles emerging as central nodes of influence.

Conclusion: The study's implications for future research include the importance of interdisciplinary collaboration, a deeper examination of aspects of chatbot design, user experience, and interaction dynamics, and prioritising context-sensitive approaches to effective chatbot deployment and evaluation in emerging countries.

Contribution: Overall, this bibliometric analysis offers valuable insights into the current state of research on chatbots and provides a foundation for future endeavours in this rapidly evolving research domain.

Keywords: chatbots; customer service; query resolution; bibliometric analysis; artificial intelligence.

Introduction

In recent years, the proliferation of conversational chatbots has revolutionised customer service practices across various industries (Følstad & Taylor 2021:1; Rossmann, Zimmermann & Hertweer 2020:237). These automated systems, powered by artificial intelligence (AI) and natural language processing (NLP) technologies (Følstad & Skjuve 2019:2), aim to enhance customer satisfaction and streamline query resolution processes (Le & Rajah 2022:164; Miklosik, Evans & Qureshi 2021:106230). However, the effectiveness of chatbots in addressing customer queries remains a topic of considerable debate and investigation. Established industry standards and expectations guide the implementation of chatbots for customer service enhancement (Janssen et al. 2022). Yet, there exists a noticeable gap between industry expectations and the actual performance of chatbots. This highlights the critical need for organisations to assess and monitor chatbots effectively (Adam, Wessel & Benlian 2020:427; Følstad & Taylor 2021:1). The identified gap emphasises the significance of rigorous evaluation methods to measure the efficiency of chatbots in customer query resolution accurately (Følstad & Taylor 2021:1). As chatbot technologies evolve, researchers emphasise that current service evaluation tools cannot measure conversational depth, coherence and emotional responsiveness in AI-driven customer interactions (Møller et al. 2024).

The evaluation of chatbot effectiveness in customer query resolution requires a multifaceted approach that goes beyond quantitative metrics to encompass qualitative aspects such as customer interactions and experiences (Jadeja & Varia 2017:2; Haugeland et al. 2022:2).

Understanding the intricacies of how conversational chatbots perform in addressing customer queries is essential for organisations striving to provide superior customer service experiences (Følstad & Skjuve 2019:3). Recent literature also emphasises the importance of user-centric constructs such as perceived helpfulness, trustworthiness, and system reliability in shaping satisfaction with AI chatbot services (Møller et al. 2024; Przegalinska et al. 2019). Performance models should not only evaluate technical accuracy but also account for user expectations and communication style preferences in customer service contexts (Møller et al. 2024). This bibliometric analysis seeks to address this gap in the literature by systematically examining the body of research related to the evaluation of chatbot effectiveness in resolving customer queries. By employing bibliometric methods, this study aims to uncover trends, patterns and gaps in the existing literature, providing valuable insights into the state of research in this domain.

Through a comprehensive analysis of article publications, this study aims to address the following research questions:

- *What is the current state of research on chatbot evaluation for its effectiveness in resolving customer queries?*
- *Which research disciplines predominantly contribute to the literature on chatbot evaluation in resolving customer queries?*
- *Which journal publishers, authors, countries and institutions are prominent contributors in chatbot evaluation for resolving customer queries?*
- *What are the research gaps and potential avenues for future research and development in chatbot evaluation and customer service practices?*

Subsequent sections detail the methodology used in this study, present and discuss results, and conclude by discussing contributions, limitations and conclusions.

Research methods and design

Bibliometric analysis systematically examines publication patterns, citation, collaboration, and research impact within a specific research area or discipline (Aria & Cuccurullo 2017:959).

Furthermore, a bibliometric analysis enables the researcher to understand the state of research in the domain of study and identify research gaps (Aria & Cuccurullo 2017:959). In this study, bibliometric analysis provides a robust framework for investigating the body of literature related to chatbot evaluation and its effectiveness in resolving customer queries. By analysing bibliographic data extracted from publications, this methodology facilitates the identification of key trends, influential works and research gaps in the domain, contributing to a deeper understanding of chatbot adoption in customer service.

The steps followed in conducting the bibliometric analysis are summarised in Table 1 to enhance transparency and ensure reproducibility of the study.

Data collection

The data for this bibliometric analysis was collected from Web of Science (WoS) using predefined search strings. The search strings included combinations of keywords related to chatbots, customer service, query resolution, and evaluation within the specified publication date range of 01 January 2000 to 20 February 2024. Publications meeting the inclusion criteria were exported in plain text format, including full record content and citation references. This systematic approach ensures comprehensive coverage of relevant literature, enabling a thorough analysis of research trends and developments in the domain.

The search strategy employed Boolean operators to combine keywords and refine search queries, aiming to retrieve publications that specifically evaluated chatbots' effectiveness in customer query resolution. The search string ('chatbots' OR 'conversational agents') AND ('customer service' OR 'customer query' OR 'query resolution') AND ('evaluation' OR 'evaluate' OR 'assess' OR 'assessment') was designed to capture relevant literature while minimising the risk of overlooking pertinent studies. By restricting the publication year range to 2000–2024, the analysis encompasses the evolution of research trends and recent advancements in chatbot technology and customer service practices, providing insights into current research trends.

Although the search range was set to include publications from 2000 to 2024 to provide extensive initial coverage,

TABLE 1: The bibliometric analysis process.

Step	Description
1	Define the research focus
2	Develop and formulate the search query using keywords relevant to the research topic
3	Conduct the search in Clarivate Web of Science (WoS) Core Collection
4	Apply inclusion and exclusion criteria to filter results
5	Set language to English
6	Screen articles manually based on title, abstract and full text
7	Select relevant articles and export data in a plaintext file – set record content to 'Full Record and Cited References'
8	Download and install VOSviewer
9	Launch VOSviewer – create a map based on bibliographic data and select the downloaded WoS text file
10	Choose the preferred type of analysis (such as co-authorship, keyword co-occurrence) and set the counting method to full counting
11	Set thresholds and verify results
12	Click 'finish' to produce visualisations
13	Conduct analysis on the network visualisations – select items to have clusters and specific items generated
14	Save visualisations by using the 'Screenshot' function
15	Repeat steps 10 to 14 as required

TABLE 2: Criteria for selecting relevant literature.

Inclusion criteria	Exclusion criteria
- Peer-reviewed journal articles - Published in English - Related to chatbot use in customer service or query resolution - Published between 2015 and 2024 - Indexed in the WoS Core Collection	- Conference proceedings, book chapters, dissertations, and grey literature - Non-English publications - Studies unrelated to chatbot effectiveness or customer service contexts - Duplicate or irrelevant articles based on title and abstract review

WoS, Web of Science.

the search yielded publications from 2015 onwards. These search results validate that research in chatbot technology in the domain of query resolution started emerging from 2015 onwards. Thus, 2015 to 2024 provides an effective analysis window to capture research developments and ensure data relevance (Table 2). To maintain the integrity and relevance of the data that are analysed, the study only considers English-language peer-reviewed journal articles. These publications need to explicitly address the evaluation of chatbots within the context of customer service or query resolution. Including English-language articles as part of the selection criterion provides consistency in the research and enables direct interaction with the content without translation.

Some publication types are deliberately excluded to maintain a focused and methodologically sound dataset (Table 2). Conference proceedings are excluded as they often represent preliminary findings or extended abstracts that are not peer-reviewed. Book chapters are also excluded because of their broad scope and limited relevance to the research focus. Grey literature, including reports, theses, working papers, and non-peer-reviewed documents, is excluded to minimise the risk of bias and ensure data validity. In addition, non-English publications are excluded, as the aim is to include literature that is available for a broad academic audience and considers the language restrictions of the research instruments and analytic processes used.

The extracted data encompassed various bibliographic elements, including publication year, author names, affiliations, journal or conference title, abstract, keywords, and citation counts. Utilising export formats in plain text with full record content and citation references facilitated further analysis and processing of the collected data. Each publication was carefully reviewed to ensure alignment with the research focus and adherence to the predefined inclusion criteria, ensuring the integrity and validity of the dataset for subsequent analysis. A manual screening was conducted on each article to confirm relevance. The title and abstract were examined for all entries, and in instances where relevance was unclear, the full text was consulted. This manual screening ensured that only literature that aligned with study objectives was retained.

In view of the limited number of available literature retrieved using the search string, the data cleaning process involved browsing through the list of search results to identify possible duplicate publications. Through this manual process, no discrepancies or duplicate entries were identified, and the original dataset was retained intact to provide confidence in its completeness and accuracy for subsequent bibliometric procedures. Additionally, all extracted keywords from the dataset were reviewed for variations and synonymous expressions to ensure a coherent analysis. Both singular and plural forms of terms, such as 'chatbot' and 'chatbots', were retained to preserve contextual meaning but grouped as a single term to ensure consistency and prevent artificial fragmentation of term frequency.

Synonyms such as 'conversational agent' and 'virtual assistant' were standardised under the primary term 'chatbot'. Similarly, variations in terminology related to 'evaluation', including terms such as 'performance assessment' and 'effectiveness measurement', were grouped as the common term 'chatbot evaluation', and related concepts such as 'customer satisfaction' and 'user satisfaction' were standardised to 'satisfaction' as a single keyword.

Data analysis

This study utilised the results from the WoS search and employed VOSviewer version 1.6.20 to analyse extracted data. Performance analysis and science mapping techniques are the two bibliometric methods used for data analysis. Performance analysis quantifies productivity and impact in terms of publication counts and citations (Carpenter, Cone & Sarli 2024), and as such, descriptive statistics were used to analyse the distribution of publications over time, literature contribution by publishers, primary research areas, keyword analysis, and citation counts. These statistics provide insights into the volume and impact of research output in the domain. Science mapping techniques facilitate the exploration, description, or interpretation of the current state and evolution of research and visualise intellectual, conceptual, and social structures through co-citation, co-authorship, and keyword co-occurrence networks (Pradhan 2016:19). Citation analysis enabled the assessment of the influence and significance of individual publications (Naveen et al. 2021:228). These analysis techniques contribute to a comprehensive understanding of the research landscape surrounding evaluating chatbot effectiveness in customer query resolution.

Ethical considerations

Ethical clearance to conduct this study was obtained from the College of Business and Economics Research Ethics Committee (CBEREC), University of Johannesburg (Ref. No. 2023SCiiS011) to ensure that ethical standards were applied when conducting the study.

Results

This bibliometric analysis aims to provide a comprehensive overview of the current state of research on chatbot evaluation methodologies. By identifying trends, gaps, and emerging themes, the analysis offers valuable insights that can inform the development of more effective chatbots for customer query resolution. The literature search was conducted on WoS using the specified search string and date range and yielded 27 literature studies, including articles, reviews, and early-access articles. These literature studies were published between 2015 and 2024 (Figure 2) by a diverse range of publishers. Notably, there was an observable increase in the number of publications over the years, indicating a growing interest in and research activity in the domain of chatbot evaluation for effective customer query resolution.

The distribution of literature across various publishers highlights the diverse research landscape on chatbot evaluation for customer query resolution. With six articles each, Elsevier and Springer Link demonstrate their prominence in academic research publishing by offering a wealth of scholarly resources. The contribution from Multidisciplinary Digital Publishing Institute (MDPI) reflects the growing influence of open-access platforms in disseminating research on emerging topics such as chatbot evaluation and effectiveness. Additionally, articles from Emerald Group Publishing and Taylor and Francis emphasise the interdisciplinary nature of this research domain. The literature from the other publishers indicate the wide range of perspectives and approaches within the domain and showcase inclusivity and dynamism. Overall, this distribution provides a view of the accessibility and range of knowledge available in the chatbot research domain, specifically in the evaluation context for effectively resolving queries.

The analysis of published literature by authors revealed 95 contributors across the 27 literature studies. A co-authorship analysis using VOSviewer was conducted with a minimum threshold of one document and five citations per author and resulted in only 42 authors meeting this criterion. Subsequently, a network visualisation of co-authorship identified five authors with the most extensive set of connection links (Figure 3), while the remaining 37 authors had no link connections. These findings illustrate the collaborative patterns among authors within the analysed literature and highlight the presence of a cohesive cluster of authors with established connections alongside a significant proportion of authors with no collaborative links. As illustrated, the authors with connections are Przegalinska et al. (2019).

The analysis of literature contributions by author affiliations indicates participation from several institutions. The California State University System, Central South University, and Soochow University emerge as prominent contributors, each with two "participation" works (Table 3). Additionally, 62 other institutions have made singular contributions. The presence of multiple institutions suggests a degree of

robustness and diversity in the academic research institutions and indicates a broad engagement with the research domain across different academic settings and geographical regions. However, the predominance of single contributions from many institutions may suggest a lack of concentrated research efforts and collaboration. This indicates potential opportunities for increased collaboration and synergy among
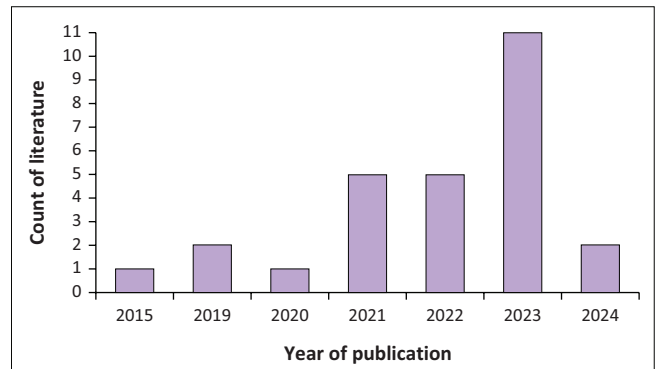

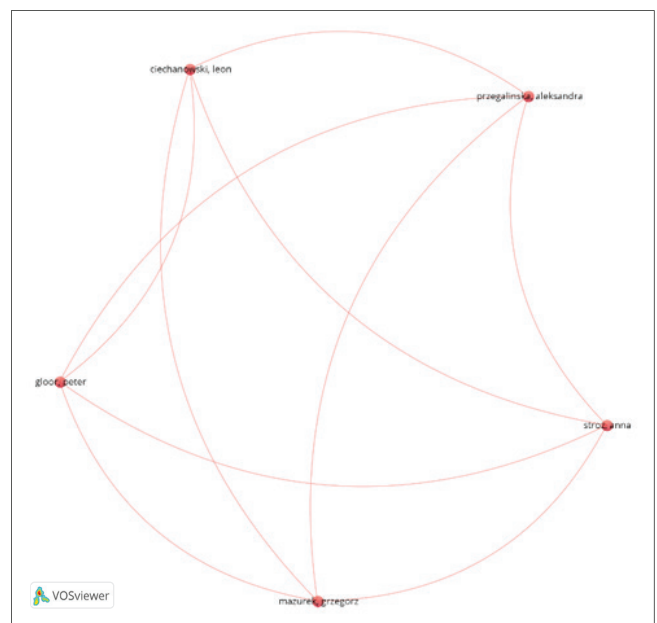**FIGURE 2:** The evolution of research evaluating chatbot effectiveness in customer query resolution.

FIGURE 3: Co-authorship network visualisation.

FIGURE 1: Screenshot of the search strategy.

institutions to foster deeper exploration and advancement of research in the domain.

The analysis of literature contributions by country indicates broad international participation in the research domain, with 20 countries contributing to the literature. The United States and Germany emerge as the top contributors, reflecting these countries' significant presence in research output in the domain. Overall, these contributions from diverse geographical regions enrich the body of knowledge in chatbot evaluation and highlight the global scope of research interest.
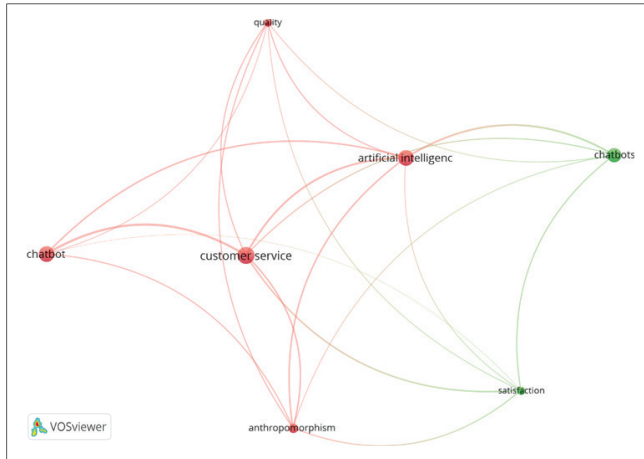

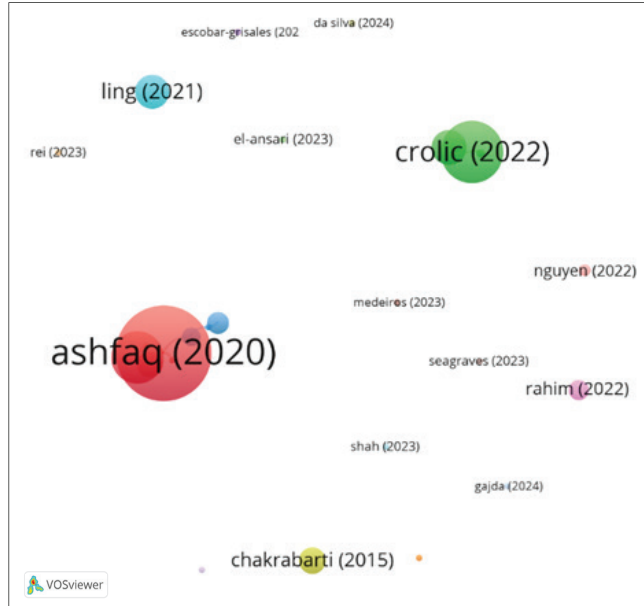**FIGURE 4:** Keyword co-occurrence network visualisation.

FIGURE 5: Citation network visualisation with the minimum threshold of zero.
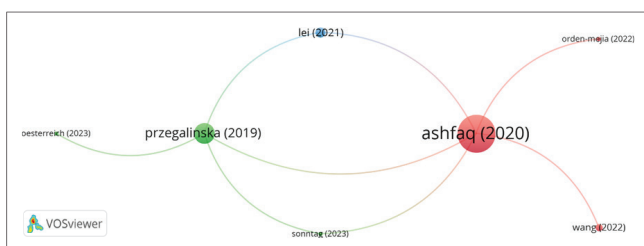

FIGURE 6: Citation network visualisation with the minimum threshold of one.

Analysis of research disciplines

The analysis of extracted literature by research disciplines is significant in identifying the concentration and understanding of the broader landscape of chatbot evaluation in resolving customer queries. This analysis aims to identify the top five priority research disciplines that shape the landscape of the research domain. The conducted analysis reveals a diverse research landscape and indicates that research spans multiple research disciplines.

Table 4 visually represents the distribution of the extracted literature across research disciplines, although the chart areas may not strictly correspond to the values of each category. These findings provide valuable insights into the multidisciplinary nature of chatbot evaluation research and signify the extensive impact of chatbot technology in various fields.

Business and Economics emerged as the predominant research domain with the most literature. This result demonstrates the practical application of chatbots and suggests a strong interest in optimising business operations and improving overall business efficiency through chatbot implementation. Computer Science, Engineering, Information Science and Library Science, and Environmental Sciences and Ecology are closely behind. The presence of other disciplines highlights the interdisciplinary nature of chatbot evaluation and recognises the significance of these disciplines in the comprehensive and effective development of chatbot solutions that address both technological and human-centric aspects of customer query resolution.

Research trends through keyword analysis

From a dataset comprising 212 keywords extracted from relevant literature, the keyword co-occurrence analysis conducted with a minimum threshold of five occurrences yielded eight unique keywords. Five was the minimum occurrence threshold, as keyword occurrences below five were deemed not value-adding based on the aim of the analysis. Seven of the eight unique keywords that emerged through the analysis were selected for visualisation (Figure 4).

TABLE 3: Top 3 contributing institutions in the research domain.

Institutions	Country	Literature count	% of 27 literature studies
California State University System	United States	2	7.401
Central South University	China	2	7.401
Soochow University	China	2	7.401

TABLE 4: Leading research disciplines in the publication of chatbot evaluation for customer query resolution.

Research discipline	Literature count	% of 27 literature studies
Business and Economics	10	38.04
Computer Science	7	25.93
Engineering	6	22.22
Information Science and Library Science	3	11.11
Environmental Sciences and Ecology	2	7.71

TABLE 5: Literature metadata and citation counts, order prioritised by connection links count.

Literature title	Author(s)	Journal/Source	Year	Publisher	Citations	Links
I, Chatbot: Modeling the determinants of users' satisfaction and continuance intention of AI-powered service agents.	Ashfaq, M., Yun, J., Yu, S.B., and Loureiro, S.M.C.	Telematics and Informatics, 54.	2020	Elsevier	238	5
In bot we trust: A new methodology of chatbot performance measures	Przegalinska, A., Ciechanowski, L., Stroz, A., Gloor, P., and Mazurek, G.	Business Horizons, 62(6): 787–797.	2019	Elsevier	101	4
Deriving trust-supporting design knowledge for AI-based chatbots in customer service: a use case from the automotive industry.	Sonntag, M., Mehmann, J., and Teuteberg, F.	Journal of Organisational Computing and Electronic Commerce, 33(3–4):178–210.	2023	Taylor & Francis Inc	1	2
A comparison between chatbot and human service: customer perception and reuse intention.	Lei, S.I., Shen, H.L., and Ye, S.	International Journal of Contemporary Hospitality Management, 33(11):3977–3995.	2021	Emerald Group Publishing Ltd	36	2
Understanding and Improving Consumer Reactions to Service Bots.	Castelo, N., Boegershausen, J., Hildebrand, C., and Henkel, A.P.	Journal of Consumer Research, 50(4):848–863.	2023	Oxford Univ Press Inc	8	1
Blame the Bot: Anthropomorphism and Anger in Customer-Chatbot Interactions.	Crolic, C., Thomaz, F., Hadi, R., and Stephen, A.T.	Journal of Marketing, 86(1):132–148.	2022	Sage Publications	130	1
How does artificial intelligence create business agility? Evidence from chatbots.	Wang, X.Q., Lin, X.L., and Shao, B.	International Journal of Information Management, 66.	2022	Elsevier	22	1
How can I help you? Design principles for task-oriented speech dialogue systems in customer service.	Oesterreich, T.D., Anton, E., Schuir, J., Brehm, A., and Teuteberg, F.	Information Systems and E-business Management, 21(1):37–79.	2023	Springer	2	1
Evaluation of the attributes of the chatbots that most effectively interact with the tourist: a case study of the chatbot 'Victoria la Malagueña'.	Orden-Mejía, M., and Huertas, A.	Cuadernos De Turismo, 50:119–141.	2022	Univ Murcia	1	1
Understanding the user satisfaction and loyalty of customer service chatbots.	Hsu, C.L., and Lin, J.C.C.	Journal of Retailing and Consumer Services, 71.	2023	Elsevier	28	0
Estimating the Impact of 'Humanizing' Customer Service Chatbots.	Schanke, S., Burtch, G., and Ray, G.	Information Systems Research, 32(3):736–751.	2021	INFORMS	55	1
Artificial conversations for customer service chatter bots: Architecture, algorithms, and evaluation metric.	Chakrabarti, C., and Luger, G.F.	Expert Systems with Applications, 42(20): 6878–6897.	2015	Pergamon-Elsevier Science Ltd	38	0
Factors influencing users' adoption and use of conversational agents: A systematic review.	Ling, E.C., Tussyadiah, I., Tuomi, A., Stienmetz, J., and Ioannou, A.	Psychology and Marketing, 38(7): 1031–1051.	2021	Wiley	54	0
Quantifying chatbots' ability to learn business processes.	Kecht, C., Egger, A., Kratsch, W., and Röglinger, M.	Information Systems, 113.	2023	Pergamon-Elsevier Science Ltd	1	0
AI-Based Chatbots Adoption Model for Higher-Education Institutions: A Hybrid PLS-SEM-Neural Network Modelling Approach.	Rahim, N.I.M., Iahad, N.A., Yusof, A.F., and Al-Sharafi, M.A.	Sustainability, 14(19).	2022	MDPI	25	0
'Don't Neglect the User!' – Identifying Types of Human-Chatbot Interactions and their Associated Characteristics.	Nguyen, T.H., Waizenegger, L., and Techatassanasoontorn, A.A.	Information System Frontiers, 24(3):797–838.	2022	Springer	12	0
Sentiment Analysis for Personalized Chatbots in E-Commerce Applications.	El-Ansari, A., and Beni-Hssane, A.	Wireless Personal Communications, 129(3): 1623–1644.	2023	Springer	3	0
Dynamic Working Memory for Context-Aware Response Generation.	Xu, Z., Sun, C.J., Long, Y.N., Liu, B.Q., Wang, B.X., Wang, M.J., Zhang, M., and Wang, X.L.	IEEE-ACM Transactions on Audio Speech and Language Processing, 27(9): 1419–1431.	2019	IEEE-Inst Electrical Electronics Engineers Inc	3	0
A review of natural language processing in contact centre automation.	Shah, S., Ghomeshi, H., Vakaj, E., Cooper, E., and Fouad, S.	Pattern Analysis and Applications.	2023	Springer	1	0
Detecting Fine-Grained Emotions in Literature.	Rei, L., and Mladenic, D.	Applied Sciences-Basel, 13(13).	2023	MDPI	2	0
Is the AI revolution a real estate boon or bane?	Seagraves, P.	Journal of Property Investment and Finance.	2023	Emerald Group Publishing Ltd	1	0

Note: Please see the full reference list of this article Mashigo, T.B., Dinath, W. & Khumalo, S., 2025, 'Chatbot evaluation for effectiveness in customer query resolution', *South African Journal of Information Management* 27(1), a1963. https://doi.org/10.4102/sajim.v27i1.1963 for more information.

AI, artificial intelligence; PLS-SEM, Partial Least Squares Structural Equation Modeling.

These keywords include 'chatbot', 'chatbots', 'customer service', 'satisfaction', 'quality', 'anthropomorphism', and 'artificial intelligence'. Notably, 'technology' was the eighth keyword deliberately excluded from the analysis because of its broad and overarching nature, which lacks specificity to the research focus. The resulting network visualisation revealed two distinct clusters comprising 20 links with a total link strength of 64.

The keyword 'customer service' emerges as a central theme, reflecting the primary objective of chatbots in enhancing service delivery and their crucial role in the context of chatbot interactions and customer experiences. The presence of 'satisfaction' and 'quality' emphasises the importance of user satisfaction and service quality in evaluating chatbot performance and effectiveness. 'Satisfaction' and 'quality' highlight the importance of customer experience and the role of chatbots in meeting user expectations. 'Anthropomorphism' and 'artificial intelligence' highlight the technological foundation of chatbot interactions and explore aspects of human-like behaviour and AI-driven capabilities.

The decision to include both 'chatbot' and 'chatbots' in the analysis stems from recognising variations in terminology usage within the literature. While 'chatbot' represents the singular form commonly used to refer to automated conversational agents, 'chatbots' encompasses the plural form, accounting for instances where multiple chatbots are discussed collectively. By including both variations, the analysis ensures a comprehensive exploration of chatbot-related literature and captures diverse research findings and perspectives. This approach acknowledges the subtle differences in language usage within the research domain and enables a more inclusive examination of chatbot effectiveness in resolving customer queries.

Literature influence on research through citation analysis

A citation analysis was conducted to gauge the influence of existing literature and authors on the broader research landscape of chatbot evaluation and its effectiveness in resolving customer queries. Cited documents were employed as units of analysis for network visualisation. Initially, a minimum citation threshold of zero was set to explore the interconnections among the 27 literature studies. All 27 literature studies met the threshold, yielding a network visualisation comprising 17 clusters and 13 links (Figure 5).

Subsequently, the minimum threshold was adjusted to 1 citation, which resulted in 21 literature studies meeting the criterion. The resulting 21 literature studies are listed in Table 5.

Only seven of the literatures were selected from this subset because they formed the most extensive set of connected items, resulting in a network visualisation characterised by four clusters and eight links (Figure 6). Out of the eight links, three are unique, while five are common and shared across all four clusters. These shared links indicate a high level of interconnectivity between the clusters.

The high interconnectivity indicated by shared literature between the clusters suggests that the literature is influential across multiple thematic areas. Table 6 lists literature that comprises the four clusters, and links.

The significance of this analysis lies in its ability to uncover the interplay between cited documents to reveal patterns of influence and connections within the literature (Naveen et al. 2021:288). This analysis indicates that, while a substantial amount of literature was cited, only a few were extensively interlinked. These results suggest a fragmented landscape of research.

The scarcity of links in the literature may indicate a lack of cohesion or consensus within this field of research, potentially reflecting on the different research approaches or findings. With just 27 literature studies published between 2015 and 2024 and 762 citations accumulated over the years, this analysis signifies the relatively limited scope of research in the domain. However, the concentration of citations among a few highly interconnected articles may signify a focused and impactful body of literature and suggest that research in this domain is advancing. Figure 7 indicates that research steadily increased in both the number of

TABLE 6: Linked articles per cluster, prioritised by link strength.

Clusters by literature	Links per cluster	Linked literature
Cluster 1: *I, Chatbot: Modelling the determinants of users' satisfaction and continuance intention of AI-powered service agents.*	5	• *In bot we trust: A new methodology of chatbot performance measures.* • *A comparison between chatbot and human service: customer perception and reuse intention.* • *Deriving trust-supporting design knowledge for AI-based chatbots in customer service: a use case from the automotive industry.* • *Evaluation of the attributes of the chatbots that most effectively interact with the tourist: a case study of the chatbot 'Victoria la Malagueña'.* • *How does artificial intelligence create business agility? Evidence from chatbots.*
Cluster 2: *In bot we trust: A new methodology of chatbot performance measures.*	4	• *I, Chatbot: Modelling the determinants of users' satisfaction and continuance intention of AI-powered service agents.* • *A comparison between chatbot and human service: customer perception and reuse intention.* • *Deriving trust-supporting design knowledge for AI-based chatbots in customer service: a use case from the automotive industry.* • *How can I help you? Design principles for task-oriented speech dialogue systems in customer service.*
Cluster 3: *A comparison between chatbot and human service: customer perception and reuse intention.*	2	• *I, Chatbot: Modelling the determinants of users' satisfaction and continuance intention of AI-powered service agents.* • *In bot we trust: A new methodology of chatbot performance measures.*
Cluster 4: *Deriving trust-supporting design knowledge for AI-based chatbots in customer service: a use case from the automotive industry.*	2	• *I, Chatbot: Modelling the determinants of users' satisfaction and continuance intention of AI-powered service agents.* • *In bot we trust: A new methodology of chatbot performance measures.*

AI, artificial intelligence.

publications and citations over time. These results highlight the relevance and novelty of the research domain analysed.

Discussion

This bibliometric analysis aimed to uncover trends, patterns, and gaps in existing literature, providing valuable insights into the state of research on chatbot evaluation for resolving customer queries. The analysis revealed a multifaceted landscape of research spanning diverse disciplines. The extracted bibliographic data resulted in 27 literature studies sourced from 11 publishers between 2015 and 2024, encompassing 762 citations over the same period.

Growing interest and relevance of research

The analysis highlighted a notable trend of growing interest and relevance in chatbot research, particularly with a focus on addressing customer queries, their approach to resolving these queries, and the evaluation of their performance. The number of citations accumulated over the years emphasises the increasing recognition of chatbots as a viable solution for enhancing customer service and satisfaction across various industries. The observed increase in publications and citations (Figure 7) reflects a broader acknowledgement of the potential benefits offered by chatbots in improving customer service efficiency and effectiveness. The analysis of increased publication and citation rates addresses RQ1: 'What is the current state of research on chatbot evaluation for its effectiveness in resolving customer queries?' by demonstrating the expanding academic attention given to chatbot effectiveness and query resolution.

The growing trend signifies a shift in focus towards exploring the capabilities and applications of chatbots in addressing customer queries and indicates a growing emphasis on understanding chatbot effectiveness and impact in real-world scenarios. As such, researchers and practitioners alike are directing their efforts towards investigating the various aspects of chatbots and their implications for resolving customer queries. The observed increase in interest highlights the need for continued research efforts to address gaps in knowledge and explore emerging trends and challenges in chatbot implementation
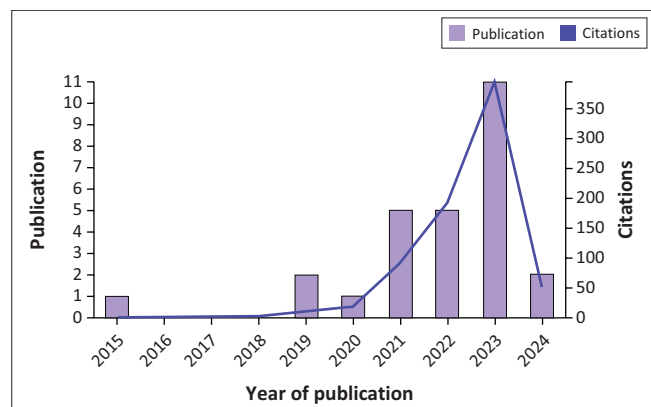

FIGURE 7: The evolution and impact of chatbot evaluation for effectiveness in customer query resolution research.

and performance evaluation. A review by Shah et al. (2023) provides insight into the role of machine learning and NLP in enhancing customer service through self-service voice portals and chatbots, indicating the potential of utilising transformer models in sentiment analysis for chatbot evaluation and offering valuable insights for future research. Researchers are encouraged to delve deeper into aspects of chatbot design, user experience, and interaction dynamics for an enhanced understanding of chatbots' role in customer service, such as in the exemplary case of Sonntag, Mehmann and Teuteberg (2023) and Oesterreich et al. (2023).

The dominance of the Business and Economics discipline

Business and Economics emerged as the dominant research discipline for chatbot evaluation in resolving customer queries. This prominence reflects the growing significance of chatbots as integral tools for enhancing customer service within commercial contexts. The prevalence of research in Business and Economics emphasises the strategic importance of chatbots in driving business innovation and improving customer experiences. Customer service organisations are increasingly turning to chatbots as a means to streamline customer query resolution processes and improve overall customer experiences, resulting in greater customer loyalty and ultimately a stronger competitive position in within their respective industries. (Miklosik et al. 2021:106535; Misischia, Poecze & Strauss 2022:423). These findings address RQ2: 'Which research disciplines predominantly contribute to the literature on chatbot evaluation in resolving customer queries?' by identifying Business and Economics as the most active discipline in chatbot evaluation literature.

The pivotal role of chatbots in enhancing customer experience and satisfaction is attributed to their integration into business operations within various industries. Kecht et al. (2023) argue that the extent to which organisations embrace chatbots depends on how effectively the underlying chatbot model can adapt and adhere to business processes. Chatbots offer businesses the opportunity to provide timely and personalised support to customers for increased satisfaction and loyalty (Chen et al. 2022:2; Hsu & Lin 2022:4). By rigorously evaluating chatbot performance in resolving customer queries within diverse industries, organisations can ensure their investment in chatbot technology yields desired outcomes. Therefore, understanding and assessing chatbot effectiveness is crucial in leveraging the chatbot's potential to enhance customer experience and satisfaction. To this point, Orden-Mejía and Huertas (2022) focus on evaluating tourists' experience with the chatbot 'Victoria la Malagueña' across attributes such as informativeness, empathy, accessibility, and interactivity, and utilise statistical techniques such as exploratory and confirmatory factor analysis for processing the data. Escobar-Grisales, Vásquez-Correa and Orozco-Arroyave (2023) also highlight the significance of using advanced neural network architectures for evaluating human-chatbot conversations.

While Business and Economics emerge as dominant research areas, collaboration with researchers from other disciplines, such as Computer Science, Linguistics, and Psychology, may be necessary to develop comprehensive and effective chatbot solutions. This interdisciplinary engagement further addresses RQ2 by highlighting the relevance of multiple research domains in contributing to chatbot evaluation. Cross-disciplinary collaboration can offer valuable insights into the technological and human-centred aspects of chatbot evaluation and effectiveness. Exemplary cases of significant cross-disciplinary collaborations include Chakrabarti and Luger (2015) and Ling et al. (2021). The former is from the Computer Science discipline and significantly contributes to the field of customer service chatbots by providing a comprehensive exploration of artificial conversations, architecture, algorithms, and evaluation metrics. The latter is from the Psychology discipline and provides an understanding of customer engagement with technology, which can be instrumental in guiding businesses towards developing more effective conversational agent strategies to enhance user experiences and adoption rates. Moreover, cross-disciplinary and collaborative efforts should be made to address issues of standardisation and benchmarking in chatbot evaluation, enabling comparability and reproducibility of research findings across studies.

Geographical influence on research

Chatbot evaluation for resolving customer queries is a dynamic and evolving field that continues to attract scholarly interest and attention, and the predominance of literature contributions from authors and institutions in well-developed countries, such as the United States, Germany, and England, indicates a geographical bias in the current scholarly discourse. These insights address RQ3: 'Which journal publishers, authors, countries and institutions are prominent contributors in chatbot evaluation for resolving customer queries?' by identifying leading contributor countries and institutions in chatbot research.

However, the emergence of contributions from countries like China and Brazil suggests an opportunity for broader global participation in the domain. This shift holds significant implications for future research, particularly in understanding the impact of contextual factors such as socio-economic and technological infrastructure on chatbot effectiveness. The technology divide between well-developed and emerging countries, including African countries, significantly influences chatbot research. Well-developed countries boast advanced digital ecosystems while emerging countries face challenges such as limited Internet access and lower smartphone penetration (Mogaji et al. 2021:2). Despite these challenges, this divide presents opportunities for innovation and collaboration. This observation directly informs RQ4: 'What are the research gaps and potential avenues for future research and development in chatbot evaluation and customer service practices?' by highlighting research gaps in geographical representation and digital accessibility. Future research should prioritise context-sensitive approaches to effective chatbot deployment and evaluation, to foster inclusive solutions and collaboration across diverse technological landscapes.

Future research suggestions

Future research should address the gaps identified in the current chatbot evaluation literature, with a focus on scope, geography, standardisation, and interdisciplinarity. The identified gaps limit the generalisability of findings and present opportunities for substantially advancing the research domain. Firstly, future research should explore underrepresented geographic regions, particularly developing countries and the African context, to provide a more holistic and inclusive understanding of chatbot effectiveness (linked to RQ3 and RQ4). The geographical bias uncovered in the analysis limits the applicability of findings in less technologically advanced regions, where the infrastructure and user behaviour differ significantly. Conducting chatbot evaluations within diverse socio-economic and technological contexts will offer richer insights into how environmental factors influence chatbot design, user adoption and satisfaction.

Secondly, future research should foster interdisciplinary collaborations to enhance the robustness and comprehensiveness of chatbot evaluation frameworks (linked to RQ2 and RQ4). Interdisciplinary collaborations can provide deeper insights into user experience, system interaction, emotional engagement, and NLP, and would help bridge technical and human-centric aspects of chatbot performance and foster the development of more effective, empathetic, and context-aware conversational agents. Thirdly, there is a need to develop and adopt standardised evaluation frameworks and benchmarking tools across studies (addressing gaps in RQ1 and RQ4). The literature shows a lack of consistency in evaluating chatbot effectiveness, with variations in performance metrics, methodologies, and user engagement parameters. Standardised tools would allow for comparability across studies, improved reliability of results, and better identification of best practices. These tools should be adaptable across sectors and sensitive to context-specific needs, thereby enhancing generalisability.

Fourthly, future research could investigate chatbot interaction quality from both a user-centric and technical performance perspective, combining qualitative insights, such as user satisfaction and perceived empathy, with quantitative measures, such as resolution rate and response accuracy (supporting RQ1 and RQ4). This mixed-methods approach can uncover the complex dynamics of human–chatbot interactions to address the current limitations of studies that heavily rely on either technical or perceptual evaluation only. Lastly, researchers should explore longitudinal studies to assess how chatbot effectiveness evolves in response to updates, retraining and changing user expectations. This recommendation is grounded in the observed trend of increasing academic interest and

technological development in chatbot systems over the last decade (derived from RQ1 findings). Longitudinal studies will allow for a deeper understanding of sustainability and continuous improvement in chatbot service delivery.

Conclusion

In conclusion, this bibliometric analysis offers a comprehensive review of the research landscape surrounding conversational chatbots in customer query resolution within different research disciplines. Exploring published literature and sources, keyword co-occurrence analysis, and citation networks provided valuable insights into the interdisciplinary nature of chatbot evaluation research, key thematic clusters, and influential literature. The study was, however, subject to limitations. The analysis relied exclusively on the WoS database and peer-reviewed articles which may have excluded other relevant publications and types indexed on other databases. These constraints may limit the generalisability of the findings.

Despite these limitations, the findings emphasise the importance of interdisciplinary collaboration, knowledge exchange, and innovation in advancing research in this domain. Researchers should prioritise addressing gaps in the literature, exploring emerging research disciplines, and disseminating findings to accelerate progress and drive innovation in chatbot technology and customer service.

Acknowledgements

Competing interests

The authors declare that they have no financial or personal relationships that may have inappropriately influenced them in writing this article. The author, S.K., serves as an associate editor of this journal. The peer review process for this submission was handled independently and the author had no involvement in the editorial decision-making process for this manuscript. The authors have no other competing interests to declare.

Authors' contributions

T.B.M., W.D. and S.K. contributed equally to this article. T.B.M. was an MPhil student, who was supervised by S.K. and co-supervised by W.D.

Funding information

Data availability

The data that support the findings of this study are available from the corresponding author, S.K., upon reasonable request.

Disclaimer

The views and opinions expressed in this article are those of the authors and are the product of professional research. They do not necessarily reflect the official policy or position of any affiliated institution, funder, agency, or that of the publisher. The authors are responsible for this article's results, findings, and content.

References

Adam, M., Wessel, M. & Benlian, A., 2020, 'AI-based chatbots in customer service and their effects on user compliance', *Electronic Markets* 31, 427–445. https://doi.org/10.1007/s12525-020-00414-7

Aria, M. & Cuccurullo, C., 2017, 'bibliometrix: An R-tool for comprehensive science mapping analysis', *Journal of Informetrics* 11(2017), 959–975. https://doi.org/10.1016/j.joi.2017.08.007

Carpenter, C.R., Cone, D.C. & Sarli, C.C., 2014, 'Using publication metrics to highlight academic productivity and research impact', *Academic Emergency Medicine* 21(10), 1160–1172. https://doi.org/10.1111/acem.12482

Chakrabarti, C. & Luger, G.F., 2015, 'Artificial conversations for customer service chatter bots: Architecture, algorithms, and evaluation metric', *Expert Systems with Applications* 42(20), 6878–6897. https://doi.org/10.1016/j.eswa.2015.04.067

Chen, Q., Lu, Y., Gong, Y. & Xiong, J., 2022, 'Can AI chatbots help retain customers? Impact of AI service quality on customer loyalty', *Internet Research* 33(6), 2205–2243. https://doi.org/10.1108/INTR-09-2021-0686

Escobar-Grisales, D., Vásquez-Correa, J.C. & Orozco-Arroyave, J.R., 2023, 'Evaluation of effectiveness in conversations between humans and chatbots using parallel convolutional neural networks with multiple temporal resolutions', *Multimedia Tools and Applications* 83(2), 5473–5492. https://doi.org/10.1007/s11042-023-14896-y

Følstad, A. & Skjuve, M.B., 2019, 'Chatbots for customer service: User experience and motivation', *CUI '19: Proceedings of the 1st International Conference on Conversational User Interfaces* 1, 1–9. https://doi.org/10.1145/3342775.3342784

Følstad, A. & Taylor, C., 2021, 'Investigating the user experience of customer service chatbot interaction: A framework for qualitative analysis of chatbot dialogues', *Quality and User Experience* 6, 6. https://doi.org/10.1007/s41233-021-00046-5

Haugelawd, I.K.F., Følstad, A., Taylor, C & Bjørkli, C.A., 2022, 'Understanding the user experience of customer service chatbots: An experimental study of chatbot interaction design', *International Journal of Human-Computer Studies* 161, 102788. https://doi.org/10.1016/j.ijhcs.2022.102788

Hsu, C. & Lin, J., 2023, 'Understanding the user satisfaction and loyalty of customer service chatbots', *Journal of Retailing and Consumer Services* 71(4), 1–10. https://doi.org/10.1016/j.jretconser.2022.103211

Jadeja, M. & Varia, N., 2017, *Perspectives for evaluating conversational AI*, viewed 09 February 2024, from https://www.semanticscholar.org/paper/Perspectives-for-Evaluating-Conversational-AI-Jadeja-Varia/5ebde6580941d9e0d16ad8cc6ec78aae16912510.

Janssen, A., Cardona, D.R., Passlick, J. & Breitner, M.H. 2022, 'How to make chatbots productive – A user-oriented implementation framework', *International Journal of Human–Computer Studies* 168, 102921. https://doi.org/10.1016/j.ijhcs.2022.102921

Kecht, C., Egger, A., Kratsch, W. & Röglinger, M., 2023, 'Quantifying chatbots' ability to learn business processes', *Information Systems* 113, 102176. https://doi.org/10.1016/j.is.2023.102176

Le, T.P.A. & Rajah, E. 2022, 'Using chatbots in customer service: A case study of Air New Zealand', in E. Papoutsaki & M. Shannon (eds.), *Proceedings: Rangahau Horonuku Hou – New Research Landscapes, Unitec/MIT Research Symposium 2021*, 06 and 07 December, pp. 161–176. ePress, Unitec, Te Pūkenga, Auckland. https://doi.org/10.34074/proc.2206011

Ling, E.C., Tussyadiah, I., Tuomi, A., Stienmetz, J. & Ioannou, A., 2021, 'Factors influencing users' adoption and use of conversational agents: A systematic review', *Psychology and Marketing* 38(7), 1031–1051. https://doi.org/10.1002/mar.21491

Miklosik, A., Evans, N. & Qureshi, A.M.A., 2021, 'The use of chatbots in digital business transformation: A systematic literature review', *IEEE Access* 9(2021), 106530–106539. https://doi.org/10.1109/ACCESS.2021.3100885

Misischia, M.V., Poecze, F. & Strauss, C., 2022, 'Chatbots in customer service: Their relevance and impact on service quality', *Procedia Computer Science* 201(2022), 421–428. https://doi.org/10.1016/j.procs.2022.03.055

Mogaji, E., Balakrishnan, J., Nwoba, A. & Nguyen, N., 2021, 'Emerging-market consumers' interactions with banking chatbots', *Telematics and Informatics* 65(2021), 101711. https://doi.org/10.1016/j.tele.2021.101711

Møller, C.G., Ang, K.E., Bongiovanni, M.L., Khalid, M.S. & Wu, J., 2024, 'Metrics of success: Evaluating user satisfaction in AI chatbots (ACM ICAAI 2024)', in *International Conference on Advances in Artificial Intelligence (ICAAI 2024),* October 17–19, 2024, Association for Computing Machinery, New York, NY.

Naveen, D., Satish, K., Debmalya, M., Nitesh, P. & Weng, M.L., 2021, 'How to conduct a bibliometric analysis: An overview and guidelines', *Journal of Business Research* 113(2021), 285–296. https://doi.org/10.1016/j.jbusres.2021.04.070

Oesterreich, T.D., Anton, E., Schuir, J., Brehm, A. & Teuteberg, F., 2023, 'How can I help you? Design principles for task-oriented speech dialog systems in customer service', *Information Systems and E-business Management* 21(1), 37–79. https://doi.org/10.1007/s10257-022-00570-7

Orden-Mejía, M. & Huertas, A., 2022, 'Evaluation of the attributes of the chatbots that most effectively interact with the tourist: A case study of the chatbot "Victoria la Malaguena"', *Cuadernos De Turismo* 50, 119–141. https://doi.org/10.6018/turismo.541891

Pradhan, P., 2016, 'Science mapping and visualization tools used in bibliometric & scientometric studies: An overview', *INFLIBNET Newsletter* 23, 19–33.

Przegalinska, A., Ciechanowski, L., Stroz, A., Gloor, P. & Mazurek, G., 2019, '2024 in bot we trust: A new methodology of chatbot performance measures', *Business Horizons* 62(6), 785–797. https://doi.org/10.1016/j.bushor.2019.08.005

Rossmann, A., Zimmermann, A. & Hertweer, D., 2020, 'The impact of chatbots on customer service performance', in J. Spohrer & C. Leitner (eds.), *Advances in the human side of service engineering*, Advances in Intelligent Systems and Computing, Springer, Florida, pp. 237–243.

Shah, S., Ghomeshi, H., Vakaj, E., Cooper, E. & Fouad, S., 2023, 'A review of natural language processing in contact centre automation', *Pattern Analysis and Applications* 26, 823–846. https://doi.org/10.1007/s10044-023-01182-8

Sonntag, M., Mehmann, J. & Teuteberg, F., 2023, 'Deriving trust-supporting design knowledge for AI-based chatbots in customer service: A use case from the automotive industry', *Journal of Organisational Computing and Electronic Commerce* 33(3–4), 178–210. https://doi.org/10.1080/10919392.2023.2276631