Aug 10th, 12:00 AM

A TOGAF Based Chatbot Evaluation Metrics: Insights from Literature Review

Sagarika Suresh THIMMANAYAKANAPALYA
university at buffalo, sthimman@Buffalo.edu

Pavankumar Mulgund
University at Buffalo, pmulgund@buffalo.edu

Raj Sharman
University at Buffalo, SUNY, rsharman@buffalo.edu

A TOGAF Based Chatbot Evaluation Metrics: Insights from Literature Review

Completed Research

Sagarika Suresh Thimmanayakanapalya
University at Buffalo
sthimman@buffalo.edu

Pavankumar Mulgund
University at Buffalo
pmulgund@buffalo.edu

Raj Sharman
University at Buffalo
rsharman@buffalo.edu

Abstract

Chatbots have been used for basic conversational functionalities and task performance in today's world. With the surge in the use of chatbots, several design features have emerged to cater to its rising demands and increasing complexity. Researchers have grappled with the issues of modeling and evaluating these tools because of the vast number of metrics associated with their measure of successful. This paper conducted a literature survey to identify the various conversational metrics used to evaluate chatbots. The selected evaluation metrics were mapped to the various layers of *The Open Group Architecture Framework* (TOGAF) architecture. TOGAF architecture helped us divide the metrics based on the various facets critical to developing successful chatbot applications. Our results show that the metrics related to the business layer have been well studied. However, metrics associated with the data, information, and system layers warrant more research. As chatbots become more complex, success metrics across the intermediate layers may assume greater significance.

Keywords

Conversational agents, Chatbot evaluation, TOGAF, metrics classification

Introduction

Chatbots are dialogue systems that attempt to mimic human-like interactive conversation using Artificial Intelligence (AI). It is designed to be the ultimate virtual assistant, helping users complete tasks ranging from answering questions, getting driving directions, turning up the thermostat in a smart home, and playing one's favorite tunes. Chatbots are also increasingly becoming popular among businesses as they can reduce customer service costs and improve consumer experience (Folstad & Skjuve, 2019). Although the deployment and use of chatbots have surged significantly, the metrics for assessing and evaluating chatbots have not been well established.

Extant research papers have focused on specific issues such as the privacy of the data collected (Harkous et al., 2016) and the response time (Huang et al., 2018); others have adopted a more high-level business perspective, such as cost savings and increased business value. They have also accounted for several other factors that must be considered while assessing chatbots. For instance, chatbots that healthcare enterprises own should not only act in accordance with the needs of the enterprise but also comply with the government's rules and regulations for healthcare. Further, chatbots could be assessed for organizational alignment and ability to integrate with other information technology assets such as data and systems. Although prior research has developed specific metrics for assessing chatbots, there is a need for a study that synthesizes and classifies all the tested sets of metrics assessing chatbots at different hierarchical levels.

We need to adopt a comprehensive organization-wide outlook to assess and classify the chatbot applications. The enterprise architecture (EA) frameworks provide such a broad lens. It is a technology and management practice aimed at developing enterprise performance by helping them see themselves in the context of a holistic and integrated view of their technology resources, information flow, business practices, and strategic decisions. Enterprise architecture aims to optimize an enterprise's fragmented business processes into an integrated and responsive environment towards changes and support the delivery of business strategies (Bernard, 2012). There are currently many enterprise architecture frameworks that organizations or companies can use. However, based on Cameron & McMillan's research on the five most commonly used enterprise architecture frameworks, TOGAF frameworks are rated as far superior to other frameworks. The advantages of TOGAF include process completion, flexibility in the use of elements, integration/interconnection between layers, vendor neutrality, and alignment with industry standards (Mueller et al., 2013).

Therefore, in this paper, we collect chatbot metrics from prior studies using a literature review and categorize them based on the various layers of the open group architectural framework (TOGAF-NIST). Furthermore, we test for the validity of these metrics by conducting semi-structured interviews with varying levels of expertise from different organizations and revisit our formalized metrics framework. Subsequently, we elaborate on the uses of this metrics framework, the limitations of this paper, and discuss future scope.

Background

In this section, we speak about two main areas of literature: First, we discuss chatbot metrics that are used to measure the success of the chatbot at the different stages of development. Second, we will elaborate on the TOGAF enterprise architecture and classify the chatbot metrics based on this framework.

Chatbot Metrics

In the past, the Loebner Prize Competition has been used to evaluate chatbots' ability to fool people that they are speaking to humans. ALICE was one such chatbot that won the Loebner Prize Competition. ALICE tried to use explicit dialogue act linguistic expressions more than usual to reinforce the impression that users are speaking to humans (Shawar & Atwell, 2007). However, these days, many studies have shown that in some scenarios making the chatbot talk in a very humane manner backfires and leads to the uncanny valley wherein the users find it hard to trust a chatbot. Many studies have concluded that researchers must not adopt an evaluation methodology just because a standard has been established, such as the Loebner Prize evaluation methodology adopted by most chatbot developers (Shawar & Atwell, 2007). Instead, evaluation should be adapted to the application and the user's needs. If the chatbot is designed to provide a specific service for users, the best evaluation should assess whether it achieves that service or task. Therefore, our paper follows a method of broadly reviewing the various chatbot metrics described in previous literature and assessing their strengths at the various development stages of the chatbot. The findings from this paper would help system developers and executives make crucial decisions about chatbot's success at the user's level. Looking at metrics can further help system developers understand which metrics are needed when the chatbot is tested for its task achievement. For instance, if the end goal is to provide a list of pre-defined answers, the system developers must ensure that the chatbot has a good database in the backend and that some of the data-related metrics are robust. In other cases where the chatbot learns from the user's input, system developers might need to fine-tune machine learning model training and testing related metrics into their chatbots. Our paper argues how the chatbot metrics vary with different development stages. Moreover, developing an evaluation framework with various metrics at different stages could provide an evaluation framework for developers to seamlessly integrate their chatbot.

TOGAF enterprise architecture

TOGAF is a framework — a detailed method and a set of supporting tools — for developing an enterprise architecture (Josey, 2016). The Open Group is an established and maintained standard; an industry consortium focused on IT standards. A key aspect of TOGAF is the TOGAF Architecture Development Method (ADM), a tested and repeatable process for developing architectures. The ADM includes

establishing an architecture framework, developing architecture content, transitioning, and governing the realization of architecture.

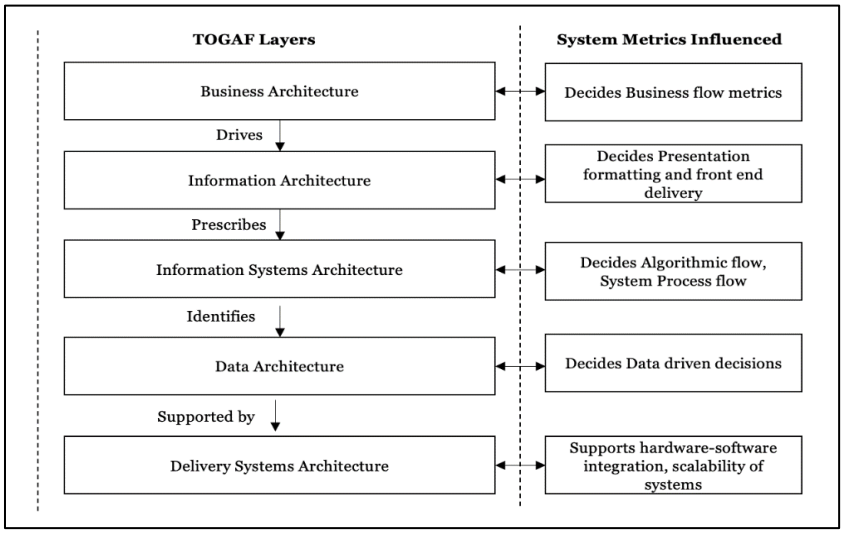

Figure 1. TOGAF NIST Enterprise Architecture

The TOGAF Architecture Content Framework (ACF) provides a structural model for architectural content, developed all along with the different steps of the ADM, which allows significant work products to be consistently defined, structured, and presented. The TOGAF ACF is structured according to its Content Metamodel. This metamodel is a single view that encompasses all four of the TOGAF architecture domains (Business, Data, Application; and Technology Architecture) and that defines a set of entities that allow architectural concepts to be captured, stored, filtered, queried, and represented in a way that supports consistency, completeness, and traceability. Figure 1 shows that the TOGAF model can be easily adapted to building any enterprise architecture. This architecture assists systems engineers in developing a series of models using predefined guidelines and guide the engineers through the systems development lifecycle; requirements engineering, design and analysis, and verification and validation.

Methodology

Literature review

As defined by Rowe (2014), a literature review 'synthesizes past knowledge on a topic or domain of interest, identifies important biases and knowledge gaps in the literature, and proposes corresponding future research directions. Prior research in IS has emphasized the importance and relevance of conducting literature reviews (Templier & Paré, 2018) and offered several methodological guidelines for authoring high-quality review articles (Paré et al., 2016). We systematically searched for literature published in the past 20 years (between January 1, 2000, and January 1, 2020) using the following databases: ACM Digital Library, EBSCO, Springer, IEEE Xplore, Web of Science, Google Scholar, PubMed, and JSTOR. The searches were performed using the following search terms: ('Chatbot and Evaluation" OR "Chatbot Evaluation Metrics" OR "Chatbot Success"). Initially, we used the title, abstract, and index terms to screen published journal articles, conference papers, proceedings, case studies, and book chapters. Two reviewers performed screening independently and met regularly to discuss the inclusion of studies. The third reviewer was consulted when there was disagreement between the reviewers. Furthermore, we performed hierarchical searches by identifying literature sources through references cited in the shortlisted papers selected from the keyword searches to find additional relevant articles. The inclusion criteria were limited to papers that specifically explored chatbot evaluation metrics to improve their quality. Exclusion criteria included white papers, metrics not related to chatbot quality improvement, and deemed as irrelevant to their evaluation.

We found 12 main papers which we referenced for the main chatbot metrics. We found 3 chatbot survey papers specific to metrics but not based on any specific layer of the TOGAF model and 17 papers that

allowed us to review various components needed to understand chatbot metrics and their implications. There were three survey papers on chatbot evaluation (Maroengsit et al., 2019; Denecke et al., 2021; Peras, 2018). However, these papers study the metrics across different developmental stages. Our paper uniquely positions a variety of chatbot metrics based on the TOGAF model allowing readers and developers to understand the type of metrics needed to be assessed at different enterprise architecture layers.

Results

In all, we looked at 32 papers to collect metrics that were specifically looked at and evaluated for the success of conversational agents. The literature review included chatbots which had not just textual results but also had multimodal methods of demonstrating output to the end-users. As per the architecture layers of the TOGAF enterprise model, we looked at business-oriented journals and papers on financial outcomes with market measures for understanding the metrics that industrial organizations used for the success of their chatbots.

Our findings show that most of the metrics at the business architecture layer of the TOGAF mode depended on external factors, competitor success, market standards, environmental problems, and political landscapes. Therefore, we classify that the chatbot's market and financial measures that fall under the business architecture layer are governed by external factors and evaluated summatively after the chatbot has been completely built. The remaining layers of the TOGAF model, which consisted of the information architecture, systems architecture, data architecture, and delivery systems architecture, all point to operational measures of the chatbot. These layers are assessed formatively during the developmental stages. To understand the information architecture layer, which focuses on the information displayed to the end-user, we looked at papers that catered to the UX development phases of the chatbot. The information systems architecture layer of the chatbot addresses how the algorithm is developed to ensure the proper working of the chatbot. We, therefore, reviewed papers on evaluating algorithm testing and evaluation. For the data architecture layer of the TOGAF model, we explored papers that addressed data manipulation techniques of chatbots. In the delivery systems architecture, we look at chatbots and their integration with hardware systems and network issues. Moreover, there were papers on how chatbots could be seamlessly integrated with other tools and external software in the delivery systems architecture.

Business Architecture

The business architecture component of the Enterprise Architecture describes the core business processes which support the organization's missions. The Business Processes component is a high-level analysis of the agency's work to support its mission, vision, and goals (Sofyana & Putera, 2019; Benbya et al., 2020). Analysis of the business processes determines the information needed and processed by the agency. Senior program managers usually develop this aspect of the enterprise architecture in conjunction with IT managers. Without a thorough understanding of its business processes and their relation to the agency's missions, the product will lack business alignment (Cabrera et al., 2016).

Four leading quality attributes that define chatbot success at the business architecture layer is presented in Table 1 (1) User-driven financial requirement depend on user engagement with it. Key metrics include the total number of users, engaged users, and the chat volume determined by the number of conversations that flow between the chatbot and the end-user. Further, goal completion explores whether the user goal was achieved. The number of bot sessions points to how many times a given user engages with a chatbot uniquely. Performance rate (Chakrabarti & Luger, 2013) studies how well the chatbot works according to the end-user, average chat time looks at how long the chatbot engages the user, most frequently asked questions help organizations decide what type of questions they must train their chatbot systems to answer frequently. Finally, ads clicked per session allow business decision-makers to understand how much revenue the chatbot has bought to their company (Waghmare, 2019). (2) The system and organization-driven financial requirements quality show how well the organization has resolved issues reported with the chatbot to increase the success rate of this system. For instance, total leads generated points for customer loyalty. Similarly, total issues resolved and cost per issue answer whether resolving some issues costs the company more than required (Kyale et al., 2019). (3). The conversation fulfillment quality attribute informs decision-makers if the chatbot is truly engaging users (Jwalapuram, 2017) based on the content it delivers to the end-user. The metrics that business decision-makers test are, for instance, looking at whether human agents are better for specific end user-related issues than the chat agent (human vs. chatbot interaction)

(Hung et al., 2009). Moreover, organizations emphasize conversation duration, interaction rate, fallback rate, and goal completion rate, as shown in the metric section of Table 1. Finally, organizations also look at statistics on how satisfied the customer has been by conducting surveys with the end-user after their communication with the chat agent has ended. (4) Customer satisfaction, therefore, looks at various metrics on how well the chat agent has satisfied the end-user (Eren, 2021; Belz & Reiter, 2006)) The metrics used here are generally rated by end-users using Likert scale values.

Quality Attribute	Metric
User driven financial requirement	Total number of users, Engaged users. No of new users, Chat volume, Goal completion, no of bot sessions initiated, Bounce rate, Performance rate, Average chat time, Most frequently asked questions, Ads clicked per session
System and organization driven financial requirement	Total leads generated, Total issues resolved, Cost per issue
Conversation Fulfillment	Human vs chatbot interaction, Conversation duration, Interaction rate, Fallback rate, Goal completion rate, Chatbot conversation length, Questions per conversation, Comprehension level
Customer satisfaction	Retention rate, Satisfaction score, Self-service rate, Performance rate, Usage rate per login, Net promoter score, Average number of interactions, non-response rate, Target audience session volume, Chatbot activity volume, User feedback

Table 1. Chatbot Metrics for Business Architecture Layer

Information Architecture

The information architecture layer focuses on the content type, presentation, and information format. Prior research has explored how conversation content must be formatted and presented to the end-user. Eleven quality attributes are critical to the meaningful evaluation of the information architecture layer. They are (1) Chatbot Behavior (Venkatesh et al., 2018) (2) Conversation conduct -trustworthiness was looked at in terms of how the wordings were formalized (Duijst, 2017). Next, (3) Conversational physicality (Liu & Dong, 2019; Benke et al., 2020) (4) Conversational coherence (Liu & Dong, 2019) (5) Conversational user control (6) Conversational consistency (Beriault-Poirier et al., 2018) and (7) Conversational repair were looked at in terms of how well, coherent and consistent the conversation was formatted and presented to the end-user. User control looked at metrics such as whether users must be given more control of the conversation, or the system must make recommendations to the user. Furthermore, the appearance of the conversational agent was extensively explored as a quality attribute for these systems. Keeping this in mind, quality attributes such as (8) Conversational appearance – Perspicuity which consisted of readability of text, length of text, audio length, was looked at (9) Conversation attractiveness, for instance, in cases where the chatbot had to be friendly it was designed to look friendly, had a name. Finally, (10) Conversational conduct and (Smestad & Volden, 2018) (11) Chatbot appearance in terms of integration with other systems is generally looked at as a quality attribute. In other words, is the chatbot compliant, and does it give out forms to fill out when it collects sensitive information, and also when the chatbot is integrated with other systems, how well does the integration look to the end-user (Schurink, 2019).

Quality Attribute	Metric
Chatbot Behavior	Conversational friendliness, Conversational proactivity, Conversational clarity, Conversational naturalness, Conversational robustness, Willingness to re-engage, Conversational relevance score
Conversation conduct, trustworthiness	Visibility of system status score, Value input score, Accuracy score, Relevance score

Conversation conduct, physicality	Appeal, Coherence score, Next turn management
Conversation conduct, coherence	Slot filling accuracy, Entity recognition, Knowledge building performance
Conversation conduct, user control	Flexibility score, System recommendation score
Conversation conduct, conversation consistency	Conversation maintenance score
Conversation Repair and Recovery	No of missed questions accounted by Conversation repair strategy, out of bounds topic, Rate of recovery, No of re-prompts
Chatbot appearance, perspicuity	Readability of text length, Acceptance of text length, Understandability of audio, Audio length
Chatbot appearance, attractiveness	Friendliness, Humane (Likert scale), Usability score, User friendly score
Conversation conduct, ethics and compliance	Compliance related queries, System data visibility, Terms and conditions contracts, Compliance justification
Chatbot appearance, integration of systems	Ease of integration, Connection performance

Table 2. Chatbot Metrics for Information Architecture Layer

Systems architecture

The systems architecture explores the procedure and flow in which the systems work. With chatbots, systems architecture focuses on the algorithm, flow of conversation, task completion processes, and system design as a whole. We noticed that nine important quality attributes are critical to chatbot system architecture (1) Depth of knowledge - For instance, rule-based versus generative chatbots. The rule-based chatbot works on predetermined rules, while generative chatbots respond based on data mining (Shawar & Atwell, 2007). (2) Topical diversity - refers to the breadth of domain knowledge. Domain-based chatbots contain information particular to a particular domain (such as healthcare or even more specific), and chitchat bots are more generic conversational systems (Liu et al., 2020). (3) Content management explores how well the process flows between conversations of the chatbot and end-user are maintained by the algorithms developed (Fang et al., 2018) (4) Context preservation refers to whether the chatbot response is contextually relevant and pertinent to the question at hand (Hung et al., 2009) (5) Turn management studies turn-taking between a bot and a human to balance the number of turns (Dippold et al., 2020). (6) Text generation is a sub-component of the chatbot's natural language processing unit, ensuring correct text generation. (7) Intent recognition looks at how well the algorithm can realize the intent of the end-user (8) Information processing involves how well the information provided by the end-user is processed by the natural language processing unit of the chatbot. Finally, (9) Information learning involves how well the system can learn from the end-user. It is a more advanced quality attribute that computer scientists are currently looking at (Novikova et al., 2017).

Quality Attribute	Metric
Dialogue Management –Conversational Depth	Knowledge coverage, Information extracted, User knowledge gain, Algorithmic competency scale
Dialogue Management –Topical diversity	Breadth of information
Content management	Content publication, Content leads generated
Dialogue Management -Context preservation	Slot filling, Dialogue efficiency, Topic Interleaving score

Dialogue Management- Turn Management	No of turns, System turns, turns per task, Topic shifting
Dialogue Generation – Text Generation	Word overlap, Word building accuracy, Document selection, no of texts per turn, number of correct texts
Dialogue Interpretation- Intent recognition	No of recognized intents, pair analysis, entity recognition, accuracy entity mapping
Dialogue Interpretation– Information processing	Lemmatization accuracy, stemming accuracy, tokenization accuracy
Dialogue Interpretation– Information Learning	Logic improvement performance, profanity filter rate

Table 3. Chatbot Metrics for System Architecture Layer

Data Architecture

The data architecture layer consists of how well the system can manage and extract data from external sources. Data architects play a crucial role in maintaining this architecture layer. We noticed three main quality attributes of a successful chatbot (1) Domain coverage refers to the volume of data covered by the chatbot (2) Information matching and retrieval as measured by how well the given user input matches with the correct output. It also refers to the chatbot's capability to fetch and retrieve data from external sources if the data output is not available in its database (Schumaker et al., 2007) (3). Data quality is crucial for the chatbot's success in processing correct information without bias promptly (Radziwill & Benton, 2017). Data quality also ensures the accuracy, timeliness (up-to information), uniqueness, validity, and security of the data itself (Lai et al., 2018).

Quality Attribute	Metric
Discovery of content- Domain coverage	Training data breadth
Information retrieval and matching	Interleaving of topics, Relevance across topics, Propensity score matching, Pairs analysis, Entity validation
Data quality	Data completeness, Data accuracy, Timeliness, Uniqueness count, Data Validity, Data bias score, Data Security.

Table 4. Chatbot Metrics for Data Architecture Layer

Delivery systems architecture

The technology infrastructure component describes and identifies the physical layer, including the functional characteristics, capabilities, and interconnections of the hardware, software, and communications, including networks, protocols, and nodes. We recognized the importance of five leading chatbot quality attributes (1) Time efficiency expressed in terms of the total time elapsed and time elapsed per task to understand if users were frustrated with wait times (2) Performance Efficiency expressed in terms of the number of turns taken to complete the goal. For instance, a higher number of turns to arrive at an answer for a simple query would leave users frustrated with the conversational agent (Kuligowska, 2015). (3) Reliability of technology expressed as the robustness of hardware and software components of the chatbot (Nguyen & Sidorova, 2017; Weber & Ludwig, 2020)(4) Accessibility measured in terms of its scalability to all geographical areas (Vanjani et al., 2019) and (5) Interoperability measured by the ease of integration of the chatbot with all external systems (Kuligowska, 2015). Further, network bandwidth and API performance are explored (Reshmi & Balakrishnan, 2018).

Quality Attribute	Metric

Efficiency of systems – time efficiency	Total time elapsed, Time elapsed per task
Efficiency of systems – performance efficiency	Task text length, Number of turns taken
Reliability of technology	Length of audio, audio processing speed, syntax former, syntactic tools, dictionary correctness, vocabulary and grammar builder tools, lexical analyzer, stemmers and lemmatization tools
Accessibility	Multi lingual scalability, no of languages trained, language performance, scalability of systems
Interoperability	Host integration, API performance, integrator performance, network performance

Table 5. Chatbot Metrics for Delivery Architecture Layer

Discussions and Conclusions

The results of this study indicate that the metrics at the business layer have been well explored. Our findings corroborate prior studies highlighting that businesses typically view chatbots as cost-saving alternatives without significant loss of consumer experience (Adam et al., 2021). Although recent research points to the potential of chatbots to offer companionship (Skjuve, 2021), have deep emotional conversations (Lee, 2017), organizations have utilized chatbots for mostly transactional tasks with a strong emphasis on cost-benefit analysis and underlying business drivers such as increased business value or reduced costs. Therefore, it is not surprising to find a plethora of business metrics for evaluating chatbots. Within the business layer metrics, there are two broad categories – 1. Cost-based metrics, and 2. Value-based metrics. Value-based metrics are particularly useful for a business model where chatbots engage in sales recommendations, promotion, or other value-added activities. On the other hand, chatbots used for troubleshooting and case management focus on task completion and reducing costs. Such metrics are highly valuable in call centers, and automated services with banks and financial institutions. The metrics at the lower layers, particularly around data, systems, and information architecture, warrant more research. These metrics are formative in nature and are used mainly by the technical team to ensure the better development of chatbot systems. While some prior work exists, as noted in previous sections, there is potential for considerable future work. For instance, several evaluation issues have been largely ignored at the data layers. For example, data variety and diversity, metrics for ethical data management, including timeliness and veracity, remain a significant challenge. Certain critical issues such as regulatory compliance and privacy management remain a major concern at the system level. Several critical issues warrant future research at the information presentation level, including metrics relating to message audience fit, framing, and audience fit.

There are several limitations to this study. Therefore, the readers are recommended to exercise caution when interpreting the conclusions of this study since they are derived from selected academic publications that fulfilled our selection criteria. Additionally, we excluded gray literature such as white papers and practitioner reports. The inclusion of such articles could have influenced our conclusions. Second, while we exercised significant caution in selecting our research, some publications could be missed owing to keyword mismatches, which might have influenced our findings.

REFERENCES

Adam, M., Wessel, M., & Benlian, A. (2021). AI-based chatbots in customer service and their effects on user compliance. Electronic Markets, 31(2), 427-445.

Belz, A., & Reiter, E. (2006, April). Comparing automatic and human evaluation of NLG systems. In *11th conference of the european chapter of the association for comp linguistics* (pp. 313-320).

Benbya, H., Nan, N., Tanriverdi, H., & Yoo, Y. (2020). Complexity and information systems research in the emerging digital world. *Mis Quarterly*, *44*(1), 1-17.

Benke, I., Knierim, M. T., & Maedche, A. (2020). Chatbot-based emotion management for distributed teams: A participatory design study. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1-30

Beriault-Poirier, A., Prom Tep, S., & Sénécal, S. (2018, October). Putting chatbots to the test: does the user experience score higher with chatbots than websites?. In *International Conference on Human Systems Engineering and Design: Future Trends and Applications* (pp. 204-212). Springer, Cham.

Bernard, S. A. (2012). *An introduction to enterprise architecture*. Authorhouse.

Cabrera, A., Abad, M., Jaramillo, D., Gómez, J., & Verdum, J. C. (2016). Definition and implementation of the Enterprise Business Layer through a Business Reference Model, using the architecture development method ADM-TOGAF. In *Trends and Applications in Software Engineering* (pp. 111-121). Springer, Cham.

Chakrabarti, C., & Luger, G. F. (2013, May). A framework for simulating and evaluating artificial chatter bot conversations. In *The Twenty-Sixth International FLAIRS Conference*.

Denecke, K., Abd-Alrazaq, A., Househ, M., & Warren, J. (2021). Evaluation Metrics for Health Chatbots: A Delphi Study. *Methods of Information in Medicine*, *60*(05/06), 171-179.

Dippold, D., Lynden, J., Shrubsall, R., & Ingram, R. (2020). A turn to language: How interactional sociolinguistics informs the redesign of prompt: response chatbot turns. *Discourse, Context & Media*, *37*, 100432.

Duijst, D. (2017). Can we improve the user experience of chatbots with personalisation. *Master's thesis. University of Amsterdam*.

Eren, B. A. (2021). Determinants of customer satisfaction in chatbot use: evidence from a banking application in Turkey. *International Journal of Bank Marketing*.

Fang, H., Cheng, H., Sap, M., Clark, E., Holtzman, A., Choi, Y., ... & Ostendorf, M. (2018). Sounding board: A user-centric and content-driven social chatbot. *Arxiv preprint arxiv:1804.10202*.

Feine, J., Morana, S., & Maedche, A. (2020, September). A chatbot response generation system. In *Proceedings of the Conference on Mensch und Computer* (pp. 333-341).

Følstad, A., & Skjuve, M. (2019, August). Chatbots for customer service: user experience and motivation. In *Proceedings of the 1st international conference on conversational user interfaces* (pp. 1-9).

Harkous, H., Fawaz, K., Shin, K. G., & Aberer, K. (2016). {pribots}: Conversational Privacy with Chatbots. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*.

Huang, C. Y., & Ku, L. W. (2018, December). Emotionpush: Emotion and response time prediction towards human-like chatbots. In *2018 IEEE Global Communications Conference (GLOBECOM)* (pp. 206-212). IEEE.

Hung, V., Elvir, M., Gonzalez, A., & demara, R. (2009). Towards a method for evaluating naturalness in conversational dialog systems. In *2009 IEEE International Conference on Systems, Man and Cybernetics* (pp. 1236–1241). San Antonio, TX, USA: IEEE. Https://doi.org/10.1109/ICSMC.2009.5345904

Hung, V., Gonzalez, A., & Demara, R. (2009, February). Towards a context-based dialog management layer for expert systems. In *2009 International Conference on Information, Process, and Knowledge Management* (pp. 60-65). IEEE.

Josey, A. (2016). *TOGAF® Version 9.1-A Pocket Guide*. Van Haren.

Jwalapuram, P. (2017, September). Evaluating dialogs based on Grice's maxims. In *Proceedings of the Student Research Workshop associated with RANLP* (pp. 17-24).

Kuligowska, K. (2015). Commercial chatbot: performance evaluation, usability metrics and quality standards of embodied conversational agents. *Professionals Center for Business Research*, *2*.

Kvale, K., Sell, O. A., Hodnebrog, S., & Følstad, A. (2019, November). Improving conversations: lessons learnt from manual analysis of chatbot dialogues. In *International workshop on chatbot research and design* (pp. 187-200). Springer, Cham.

Lai, S. T., Leu, F. Y., & Lin, J. W. 2018. "A Banking Chatbot Security Control Procedure for Protecting User Data Security and Privacy," In Proceedings of the International Conference on Broadband and Wireless Computing, Communication and Applications.

Liu, Q., Huang, J., Wu, L., Zhu, K., & Ba, S. (2020). CBET: design and evaluation of a domain-specific chatbot for mobile learning. *Universal Access in the Information Society*, *19*(3), 655-673.

Liu, R., & Dong, Z. (2019, February). A study of user experience in knowledge-based QA chatbot design. In *International Conference on Intelligent Human Systems Integration* (pp. 589-593). Springer, Cham.

Maroengsit, W., Piyakulpinyo, T., Phonyiam, K., Pongnumkul, S., Chaovalit, P., & Theeramunkong, T. (2019, March). A survey on evaluation methods for chatbots. In *Proceedings of the 2019 7th International Conference on Information and Education Technology* (pp. 111-119).
Meyer von Wolff, R., Hobert, S., Masuch, K., & Schumann, M. (2020). Chatbots at digital workplaces–a grounded-theory approach for surveying application areas and objectives. *Pacific Asia Journal of the Association for Information Systems*, *12*(2), 3.
Mueller, T., Schuldt, D., Sewald, B., Morisse, M., & Petrikina, J. (2013). Towards inter-organizational enterprise architecture management-applicability of TOGAF 9.1 for network organizations.
Nguyen, Q. N., & Sidorova, A. 2017. "AI Capabilities and User Experiences: A Comparative Study of User Reviews for Assistant and Non-assistant Mobile Apps," In Proceedings of the American Conference on Information Systems.
Novikova, J., Dušek, O., Curry, A. C., & Rieser, V. (2017). Why we need new evaluation metrics for NLG. *Arxiv preprint arxiv:1707.06875*.
Okuda, T., & Shoda, S. (2018). AI-based chatbot service for financial industry. *Fujitsu Scientific and Technical Journal*, *54*(2), 4-8.
Palanica, A., Flaschner, P., Thommandram, A., Li, M., & Fossat, Y. (2019). Physicians' perceptions of chatbots in health care: cross-sectional web-based survey. *Journal of medical Internet research*, *21*(4), e12887.
Peras, D. (2018). Chatbot evaluation metrics. *Economic and Social Development: Book of Proceedings*, 89-97.
Qurratuaini, H. (2018, September). Designing enterprise architecture based on TOGAF 9.1 framework. In *IOP Conference Series: Materials Science and Engineering* (Vol. 403, No. 1, p. 012065). IOP Publishing.
Radziwill, N., & Benton, M. (2017). Evaluating Quality of Chatbots and Intelligent Conversational Agents, 21.
Reshmi, S., & Balakrishnan, K. (2018). Empowering chatbots with business intelligence by big data integration. *International Journal of Advanced Research in Computer Science*, *9*(1).
Schuetzler, R. M., Grimes, G. M., Giboney, J. S., & Rosser, H. K. (2021). Deciding whether and how to deploy chatbots. *MIS Quarterly Executive*, *20*(1), 4.
Schumaker, R. P., Ginsburg, M., Chen, H., & Liu, Y. (2007). An evaluation of the chat and knowledge delivery components of a low-level dialog system:The AZ-ALICE experiment. *Decision Support Systems*, *42*(4), 2236–2246. Https://doi.org/10.1016/j.dss.2006.07.001
Schurink, E. (2019). *The role of perceived social presence in online shopping: The effects of chatbot appearance on perceived social presence, satisfaction and purchase intention* (Master's thesis, University of Twente).
Shawar, B. A., & Atwell, E. (2007, April). Different measurement metrics to evaluate a chatbot system. In *Proceedings of the workshop on bridging the gap: Academic and industrial research in dialog technologies* (pp. 89-96).
Singh, A., Ramasubramanian, K., & Shivam, S. (2019). *Building an enterprise chatbot: Work with protected enterprise data using open source frameworks*. New York: Apress.Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2021). My chatbot companion-a study of human-chatbot relationships. *International Journal of Human-Computer Studies*, *149*, 102601
Smestad, T. L., & Volden, F. (2018, October). Chatbot personalities matters. In *International Conference on Internet Science* (pp. 170-181). Springer, Cham.
Sofyana, L., & Putera, A. R. (2019, November). Business architecture planning with TOGAF framework. In *Journal of Physics: Conference Series* (Vol. 1375, No. 1, p. 012056). IOP Publishing.
Vanjani, M., Aiken, M., & Posey, J. (2019). An evaluation of a multilingual chatbot. *Issues in Information Systems*, *20*(1), 134-143.
Venkatesh, A., Khatri, C., Ram, A., Guo, F., Gabriel, R., Nagar, A., ... Raju, A. (2018). On Evaluating and Comparing Conversational Agents. *Arxiv:1801.03625 [Cs]*. Retrieved from http://arxiv.org/abs/1801.03625
Waghmare, C. (2019). Business benefits of using chatbots. In *Introducing Azure Bot service* (pp. 147-165). Apress, Berkeley, CA.
Weber, P., & Ludwig, T. 2020. "(Non-) Interacting with Conversational Agents: Perceptions and Motivations of Using Chatbots and Voice Assistants," In Proceedings of the Conference on Mensch und Computer.