



# Improving Conversational User Interfaces for Citizen Complaint Management through enhanced Contextual Feedback

Kai Karren  
s8kakarr@stud.uni-saarland.de  
Saarland University  
Saarbrücken, Germany

Dr. Michael Schmitz  
m.schitz@hbksaar.de  
Experimental Media Lab, University  
of Fine Arts Saar  
Saarbrücken, Germany

Dr. Stefan Schaffer  
German Research Center for Artificial  
Intelligence (DFKI)  
Berlin, Germany

## ABSTRACT

As cities transform, disrupting citizens' lives, their participation in urban development is often undervalued despite its importance. Citizen complaint systems exist but are often limited in fostering meaningful dialogue with municipalities. Meanwhile, smart cities aim to improve living standards, efficiency, and sustainability by integrating digital twins with physical infrastructures, potentially enhancing transparency and enriching communication between cities and their inhabitants with real-time data. Complementing these developments, technologies realizing Conversational User Interfaces (CUIs) are becoming more capable in providing a conversational and feedback-oriented approach such as complaint management processes.

The improvement of CUIs for citizen complaint management through enhanced contextual feedback is explored in this work. The term contextual feedback has been developed and defined as all information (for example, background, conditions, explanations, timelines, and the existence of similar complaints) related to a complaint and or the underlying problem that could potentially be relevant for the user. The solution proposed in this paper gathers data from users about their issues via a CUI, which subsequently queries various data sources to obtain relevant contextual information. Following this, a Large Language Model processes the collected data to produce the corresponding feedback. In the study, a static CUI without contextual data as the baseline has been compared to a CUI that includes contextual data, analyzing their impact on pragmatic and hedonic quality, reuse intention, and potential influence on the citizens' trust in their municipality. The study has been conducted in cooperation with the German municipality of Wadgassen. The good performance of the baseline system shows the general potential of LLMs in the citizen complaint domain even without data sources. The results show that contextual feedback performed better overall, with significant improvements in the pragmatic and hedonic quality, attractiveness, reuse intention, feeling that the complaint is taken seriously, and the citizens' trust in their municipality.

## CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces**; **Empirical studies in HCI**; • **Information systems** → **Personalization**.

## KEYWORDS

Conversational UI, Smart City, Context Information, Citizen Engagement

### ACM Reference Format:

Kai Karren, Dr. Michael Schmitz, and Dr. Stefan Schaffer. 2024. Improving Conversational User Interfaces for Citizen Complaint Management through enhanced Contextual Feedback. In *ACM Conversational User Interfaces 2024 (CUI '24)*, July 8–10, 2024, Luxembourg, Luxembourg. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3640794.3665562>

## 1 INTRODUCTION

Most cities are constantly growing, which causes various problems like traffic congestion, pollution, and the need for new and improved infrastructure. The adaption of a city to solve these problems and needs is an interdisciplinary task and the topic of various initiatives around the world. The authors of [4] state that to "make the city smart" can be seen as commonplace for these efforts. The smart city concept has gained momentum in public opinion, increased research interest, and moved onto the agenda of city authorities worldwide in the last years [4]. However, there is no consensus on when a city could be called a smart city and what it means. The authors of [2] defined it as "ideas about how information and communication technologies might improve the functioning of cities". It is important to involve the citizens and not ignore their requirements when building smart city services in favor of technological and strategic developments [28]. Because of this and because a smart city is an extremely complex and highly interconnected system of systems [28], there are lots of things that could potentially go wrong, or that could be improved. Therefore, there should be a way for citizens to voice their complaints and ideas to the authorities. Increasing the participation of citizens in their cities and municipalities using digitalization has been a goal and discussion topic in different forms for over a decade [15]. One such option where citizens can provide feedback and ideas to their local government is citizen complaint management [15]. The importance of complaint reporting in the citizen context has been shown in various studies [1, 3]. However, to report a complaint often requires lots of time and effort, having to install a special app, doing a one-time registration, searching for the website, lacking feedback, and so on. All these points are barriers that discourage citizens from reporting complaints [1, 3, 23]. The number of unvoiced complaints

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

*CUI '24*, July 8–10, 2024, Luxembourg, Luxembourg

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0511-3/24/07

<https://doi.org/10.1145/3640794.3665562>

has been reported to go from 50% to 90% in the business context [23] and around 61% in a study exploring it in the civic domain [3]. Especially in the citizen complaint context, where the direct benefits for the individual are minimal when reporting a complaint. These barriers should be minimized [3]. The complaint management process is conversational in nature and is a back-and-forth between the user and the agent. Therefore, using conversational user interfaces (CUIs) at least to assist human agents sounds beneficial, as discussed in [23]. CUIs have seen rapid growth in number from chatbots to smart assistants like Amazon's Alexa as well as rapid growth in research interest over the last years [5, 18, 27]. The interest in CUIs also includes governments and is not limited to researchers and private companies. The German government, for example, runs various Chatbots to provide certain information to the citizens and for internal use<sup>1</sup>. Previous works like [1] explored text-based conversational user interfaces (CUI) in the form of chatbots that allowed citizens to report complaints and [7] in the customer support domain. The authors of [15, 25] stated that it is important for the acceptance of e-participant systems to provide a short response time and that meaningful feedback is needed so that users think that their complaints are taken seriously. The results of the study of [25] about automated feedback for suggestions for tree planting locations indicate that users also have high expectations of the precision of automated feedback and not only the response time. In addition, smart cities provide many data sources, including real-time data, providing a vast potential to utilize this information in different ways [2]. This would also include conversations with CUIs maintained by the smart city authorities. Providing such meaningful feedback to citizen complaints in an automated way has not been explored yet. So there are many open questions like which information could be helpful, how they should be presented, potential technical challenges and solutions, effects on the user experience, perceived usefulness of the system, or changes in the citizens' opinion about their local government. This motivated this work to explore the improvement of CUIs for citizen complaint management through enhanced contextual feedback.

## 2 RELATED WORK

In the related work section of our research, we examine existing work at the intersection of CUIs and citizen complaint management within the broader context of smart cities. This related work section offers an overview of existing approaches and technologies that have laid the groundwork for enhancing citizen engagement and streamlining complaint management processes.

**Citicafe** introduces an omnichannel CUI named Citicafe for managing citizen complaints [1], highlighting the necessity of effective feedback mechanisms to engage citizens in civic improvement efforts. The study underscores the discouragement faced by citizens due to cumbersome complaint registration processes and the absence of responsive feedback, offering a solution through a multi-platform approach that enhances user involvement through improved feedback and communication strategies.

**Querimonia** addresses the challenges in complaint management processes overwhelmed by volume, proposing a semi-automated

system to assist human agents using Natural Language Processing (NLP) to generate response suggestions [13]. This prototype exemplifies the potential of AI in streamlining complaint resolution, despite its limitations in real-world evaluation.

**COTA** demonstrates advancements in customer support ticket resolution at *Uber* through machine learning and deep learning models [19]. This work evidences the significant improvements in support efficiency and customer satisfaction achievable through intelligent systems, thereby validating the application of similar technologies in public sector complaint management.

A study on customer service via social media introduces a sequence-to-sequence LSTM model for automating responses to customer inquiries on Twitter [29]. This research highlights the model's comparative efficacy in addressing emotional versus informational requests, underlining the complexities involved in automating customer service interactions. **Towards Automated Customer Support** evaluates models for automating Twitter-based customer support, particularly comparing retrieval, sequence-to-sequence with attention, and transformer models using a dataset of 3 million conversations [8]. The sequence-to-sequence model emerged as the most effective, demonstrating the potential of generative neural models in customer support automation despite challenges with limited data. This research aligns closely with [29], as both studies focus on automating social media customer interactions using advanced machine learning, specifically highlighting the effectiveness of sequence-to-sequence models in processing and responding to customer inquiries on platforms like Twitter.

**Automated Self-Learning Chatbot Database IRS** presents a knowledge-based chatbot for FAQ retrieval, which improves through the incorporation of human agent responses, underscoring the potential of self-learning systems in customer support contexts [16].

Lastly, the **Smarticipate** project explores automated feedback in citizen participation systems, illustrating the impact of responsive, precise, and immediate feedback on citizen engagement and trust in local governance [24, 25].

Together, these works illustrate a landscape where effective communication and feedback are substantial for engaging citizens and efficiently managing their concerns. Despite the diverse methodologies and domains of application, a common thread is the pursuit of systems that not only automate but also enhance the quality of interactions between citizens and their governments. Our research builds on this foundation, exploring the untapped potential of Large Language Models (LLMs) for providing contextual feedback in citizen complaint management. Building on the capabilities of LLMs for generating nuanced and informative responses, we aim to bridge the gap between technological advancements and the practical needs of citizens, thereby contributing to the development of more responsive and intelligent smart city infrastructures.

## 3 CONCEPT

The following section will elaborate on providing contextual feedback to citizen complaints. First, different definitions of context will be presented to provide a basis for the definition of contextual feedback.

<sup>1</sup><https://background.tagesspiegel.de/digitalisierung/ki-des-bundes-nicht-nur-freundliche-chatbots>

### 3.1 Context Definition

Lots of different definitions for "context" exist. Some of the definitions of context in the Cambridge Dictionary<sup>2</sup> are "the situation within which something exists or happens, and that can help explain it" and "the influences and events related to a particular event or situation". In [6], the author defines context as follows:

"Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves."

In contrast to the definitions from the Cambridge Dictionary and other possible definitions of context, e.g., as referred to in the corresponding section in [6], this definition has been developed from the perspective of human-computer interaction (HCI) and so-called "context-aware application development".

As stated in [31] with the definition of context provided in [6], any information that is relevant to the interaction would be context information. Because of this, [31] suggests five categories for context information: "individuality, activity, location, time, and relations". For this work, the relations category was the most interesting one because it describes the relations to surrounding entities as context information and that these relations may change over time. All these definitions, their reasoning, and perspectives have been used to create the "contextual feedback" definition for this work.

### 3.2 Contextual Feedback Definition

Contextual Feedback is defined in this work as all information (for example, background, conditions, and explanations) related to a complaint and or the underlying problem that could potentially be relevant to know for the user and can be provided to the user. It does **not** include general information that would always be true and non-concrete information like "we are working on solving the problem as fast as possible". Such generalizable phrases could add value to a response but are not considered contextual feedback in this work. Contextual feedback in the citizen complaint domain can also only be a form of automated initial or partial feedback because usually, problems need the involvement of humans to be resolved because an automated system can not fix a pothole or clean up garbage.

The following information could be fetched from external data sources and provided as contextual feedback:

- Existence of complaints about the same problem
- Assigned severity
- Estimated resolution time
- Reason for the problem
- Responsible department
- Links to more information

The items have been inspired by the social awareness discussed in [25], the usage of severity in [1], software project management and or (bug) ticket prioritization predictions [30], as well as by discussions with the municipality of Wadgassen.

Providing the issuer of external data sources would also be an option. As discussed in [24], this could provide background evidence

on the accuracy and trustworthiness of the information. However, it may not be understandable for everyone, and the results of [24] showed no significant differences when providing the data source in the response. In [15], capabilities of an automated feedback system have been envisioned to explain the reason why a suggestion from the citizen has to be rejected, e.g., because of legal reasons or because of underground pipes that would prevent a tree from being planted directly above. Such information could also be classified as contextual feedback. However, this would require the availability of a wide range of information in a machine-readable way and the ability to reason about it. It sounds like a potential task for fine-tuned LLMs. For this work, this was considered out-of-scope. Mainly because the needed data simply does not exist. It is also to note that the trade-off between the amount of information and the complexity of the response, as stated in [26], has to be considered.

## 4 IMPLEMENTATION

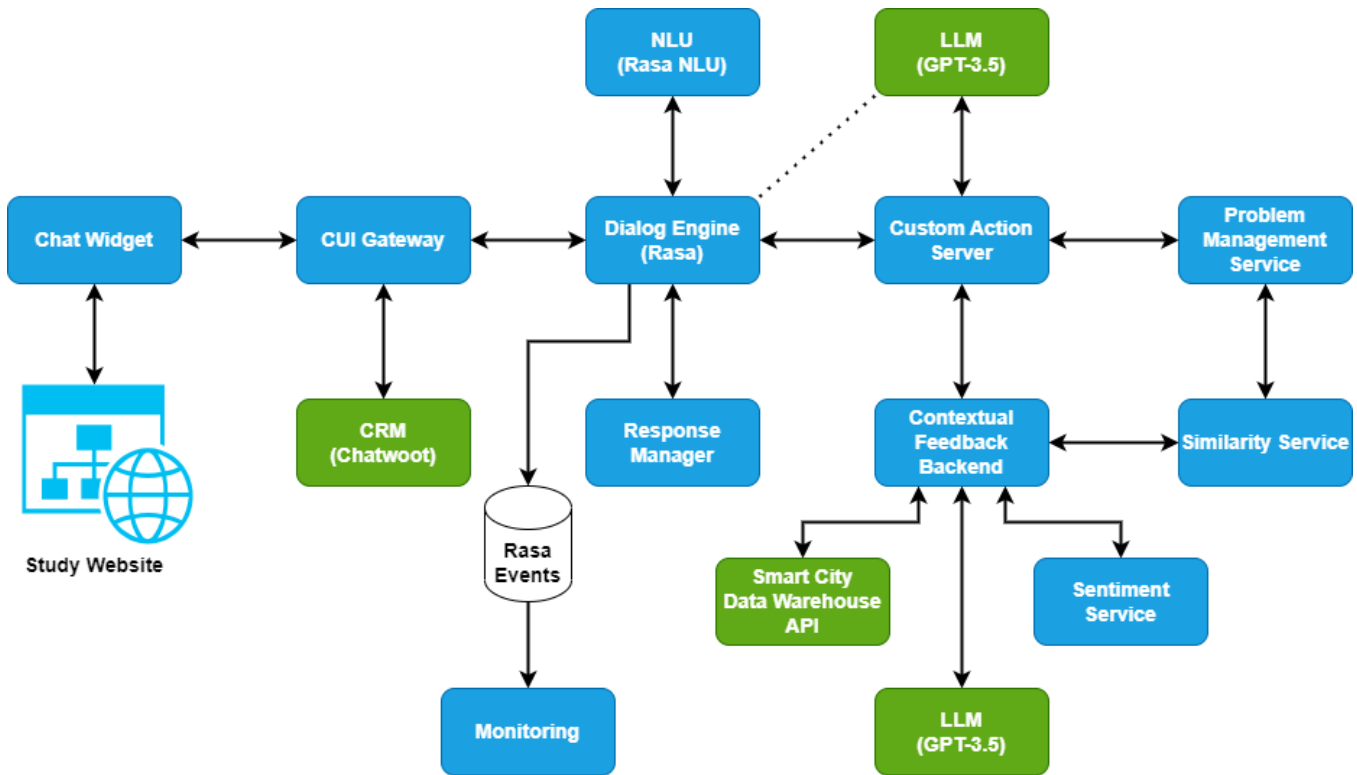
The architecture as shown in Figure 1 has been designed based on the requirements of the municipality of Wadgassen to act as a guideline for the realization of the researched functionality in the municipality. One focus point was to allow the easy integration of multiple interaction channels following the advice from [3] to reduce this aspect of the "complaint barrier". Another important aspect of the architecture was to keep it as generalizable and modular as possible to allow the easy integration of different components and services. Integrating the open-source Customer Relationship Management (CRM) tool Chatwoot<sup>3</sup> was done on the one hand for easier monitoring of the study but also to show the integration with a CRM tool in the form of the same solution that Wadgassen is migrating to. All services have been containerized as Docker containers and deployed within a multi-cloud environment, utilizing three servers in Oracle Cloud and Digital Ocean. To address the challenge of limited data availability, contextual feedback items were simulated (mocked). Consequently, the corresponding services depicted in the architecture diagram were not implemented in this work. The architecture and used technologies are not compute and cost intensive and can be easily scaled up to fit the needs of larger cities. To minimize the response times, the usage of an LLM like GPT3.5 should be preferred over GPT4. If the GPU compute power and enough data is available fine-tuning a base-LLM may also be an option for a city. However, the main problem remains the availability of relevant data in a digital and structured form.

In the following, we describe the individual components in more detail:

- The **study website** has been realized with Bootstrap34 together with plain HTML, CSS, and Javascript, including a main page with chat widget, descriptions, and instructions, as well as links to the survey and the data protection declaration. After the survey has been completed, the results are stored in the results database.
- The entry point for the interaction with the CUI is a custom **chat widget** that has been implemented based on the React

<sup>2</sup><https://dictionary.cambridge.org/dictionary/english/context>

<sup>3</sup><https://www.chatwoot.com>



**Figure 1: Architecture Overview: Green Boxes Represent External Components, Blue Boxes are Internal of the Developed System.**

components from the MIT-licensed Chat Widget from Voiceflow<sup>4</sup>. User inputs are sent to the CUI Gateway. Responses are displayed in the chat widget.

- The so-called **CUI Gateway** is a service that connects the Chat Widget with the Dialog Engine and with the customer relationship management (CRM) solution. It has been implemented as a Golang<sup>5</sup> application. For the study, the CRM integration of Chatwoot had been enabled to allow easier monitoring of the conversations.
- **Chatwoot**<sup>6</sup> is an open-source customer relationship management (CRM) solution. It is used to provide a web-based user interface to monitor the system’s conversations with the users.
- For Natural Language Understanding (NLU) and as the **Dialog Engine**, Rasa<sup>7</sup> is used in its open-source version. Rasa is a powerful and flexible platform for building complex CUIs. Training data for the intents and entities has been created by hand combined with data generated by ChatGPT and Github Copilot<sup>8</sup>. A total of 16 intents with 191 samples have been defined. The intents cover the reporting of three different

complaint categories in the form of pollution, noise, and infrastructure complaints, as well as intents for extracting the required information step by step and intents for handling questions about the capabilities of the system, affirmations, denying, greetings and goodbyes. The street name and the house number have been defined as entities to resemble a location.

- **Dialog Engine:** In the system, reporting a complaint with the chatbot could be completed in a single turn. If neither a complaint category nor a location can be extracted, the system asks for the missing information step by step. The usage of implicit or explicit confirmation strategies for the user input has been discarded because of the high accuracy of the NLU component in the pretests. In addition, intents and the corresponding RASA stories to handle also questions about the system’s capabilities have been defined. In case of a so-called fallback or non-understanding in which the system’s confidence for the intent classification falls below a threshold of 0.5, a custom multi-stage fallback handling has been implemented. In the fallback responses, the system first excuses itself and informs the users that it cannot understand the input. After that, depending on how many fallbacks have already been triggered in the current dialogue state, the system tries to help the user recover from the non-understanding. The first attempt is more general but still more specific compared to the initial utterance for the

<sup>4</sup><https://www.surveycircle.com>

<sup>5</sup><https://go.dev>

<sup>6</sup><https://www.chatwoot.com>

<sup>7</sup><https://rasa.com>

<sup>8</sup><https://github.com/features/copilot>

state. The second fallback response for the same state either switched to a step-by-step approach in which the user is asked if they could provide a category for this complaint, and or an example of a location is provided. If three fallbacks are triggered after another, this has been defined as a non-recoverable state, and the system excuses itself and informs the user that a human agent will take over the conversation.

- Rasa's **custom actions** respond with lists of response and event objects. The response objects contain the text that should be sent to the user. The events allow to perform additional functionality like setting slots, stopping the conversation, restarting the conversation, or scheduling a reminder. The action server implementation provided as part of the Rasa SDK (Python) has been used.
- **Contextual Feedback Generation:** Following VUI and GUI design guidelines (e.g., [22, 24, 26]), including a highly iterative experimentation phase, the prompt in Fig. 2. was developed to generate contextual feedback for a citizen complaint.
- The **Contextual Feedback Backend** has been implemented as a Java application using the Spring Boot<sup>9</sup> framework. It exposes a REST API to generate contextual feedback for a complaint. The corresponding custom action in Rasa makes a POST request to the backend with the classified complaint category and the extracted location.

## 5 STUDY

The study has been designed as a between-subjects online study using the feedback as the independent variable. Condition A is the control condition and Condition B is the experimental condition that provides the contextual feedback. Two tasks were defined together with the municipality of Wadgassen to ensure that the problems that the citizens should report were realistic.

A Pretest has been conducted with a total of six participants that led to improvements in some wordings and the replacement of the previously used chatwoot chat widget with a custom one based on the components from the MIT-licensed chat widget from Voiceflow. For the main study, the participants were recruited in cooperation with the municipality of Wadgassen over the internal employee mailing list and publicly over the social media channels of the municipality (Instagram and Facebook) as well as a post in the local newspaper "Wadgasser Rundschau" by the municipality's PR team. Because of a lower participation rate from the citizens in Wadgassen than expected, as well as a high rate of participants that did not follow the task instructions, the study has been shared on SurveyCircle<sup>10</sup>, mailing lists, and groups. The following hypotheses have been investigated.

### 5.1 H1.1 - Pragmatic Quality

Following the definition from [10], a product has pragmatic quality if it supports task completion in an effective and efficient way. The findings of [25] suggest that users were more satisfied with a system that provides additional (helpful) information even if they would have to wait longer for a reply. Because contextual feedback

provides this potentially relevant information about the context of the complaint, it may be seen as a partial resolution and, therefore, as a more effective and efficient way to complete the task. Based on this, the system that provides contextual feedback to the users is expected to be rated with the highest pragmatic quality.

**H1.1:** The system that provides contextual feedback to the users is rated with the highest pragmatic quality.

### 5.2 H1.2 - Hedonic Quality

A product has hedonic quality if it goes beyond pure usefulness so that users enjoy it and have fun [10]. Because the idea of contextual feedback is to improve the perceived usefulness and the feeling of the citizen that the complaint is taken seriously, an effect beyond that is unclear. However, it may exist because reducing the negative feelings related to why citizens do not report complaints, as previously discussed, could improve not only the perceived usefulness but also the perceived enjoyment and fun.

**H1.2:** The system that provides contextual feedback to the users is rated with the highest hedonic quality.

### 5.3 H2 - Reuse Intention

As stated in [9, 15, 17] the responsiveness of the authorities in e-participation systems is a key factor for the acceptance and satisfaction of the users with such systems. Based on their results, the authors of [25] state that "feedback can be considered a highly contributing factor to promoting public participation". In [15] it is stated that resignation is one of the main reasons citizens decide not to raise complaints. The citizens do not think their complaints will change something and are not taken seriously. The availability of contextual feedback could help to reduce this resignation, e.g., by providing information if others have already reported the complaint if the problem is already being worked on, and when it is expected to be solved. The paper [20] from the e-participation domain suggests that the quality of the feedback and the responsiveness of the authorities motivates e-participants to continue to participate. The hypotheses H2.1 and H2.2 have been formulated based on these points. H2.1 is about the expectation that the availability of contextual feedback will increase the willingness of the users to use such a system again. H2.2 is about the expectation that the availability of contextual feedback will increase the feeling that the complaint is taken seriously.

**H2.1:** Contextual Feedback increases the user's willingness to use the system again.

**H2.2:** Contextual Feedback increases the feeling that the complaint is taken seriously.

### 5.4 H3 - Meaningful Feedback

Following the requirements of timely and meaningful feedback in the e-participation domain formulated in [25] and the idea of contextual feedback to provide additional, relevant information, the expectation would be that the contextual feedback is perceived as more meaningful by the users even though it will not solve the underlying problem.

**H3:** The contextual feedback from the chatbot is perceived by the users as more meaningful feedback.

<sup>9</sup><https://spring.io/projects/spring-boot>

<sup>10</sup><https://www.surveycircle.com>

You are the automated complaint feedback module of a conversational user interface for citizen complaint management. The system is deployed by the municipality "Wadgassen" in Germany. Please respond in the style of a chatbot and from the perspective of the authorities responsible for solving the problem. Keep your response on point and try not to generate too long responses which are hard to read. Follow the following steps for the response generation:

Step 1: Determine which information is available in the JSON object

Step 2: Generate a polite and on-point response using the following guidelines based on the information in the JSON object. Start a new paragraph for the points below

- a: Acknowledge the complaint and express understanding of the concern.
- b: If available, mention the assigned severity level to indicate the municipality's attention to the issue.
- c: If similar complaints exist, inform the citizen about the existence of similar complaints.
- d: If available, provide an estimated resolution time to set expectations.
- e: Try not to exceed 10 sentences.

Please respond in German.

Don't address the user with their name or include formulations like "[Name der Stadt/municipality]".

Don't provide contact information or mention this point.

You are the initial contact point and human agents will take over if needed.

Don't include internal information like ids in the response.

Start the response with "Vielen Dank" and don't add a signature at the end of the response

because it is no letter but the response of a chatbot within a running conversation.

The following information in the user message has been fetched for the complaint as a JSON object :

```
{
  "category": "pollution",
  "summary": "Beschwerde wegen einer großen Menge von Sperrmüll bei den Recyclingcontainern in der Lindenstraße",
  "location": {
    "street": "Lindenstraße"
  },
  "number of similar complaints": 4,
  "severity": "high",
  "estimated resolution duration": "one week"
}
```

Figure 2: Prompt with Example Input

## 5.5 H4 - Trust

Based on the chances that complaint management could help the legitimacy of as well as the trust into the local governments as described in the papers [9, 12, 15], the availability of a complaint management system in general should improve the trust into the local government. The only work that could be found that looked into in the e-participation context is [25]. Here, the participants were asked about the impact of automated feedback on their trust, e.g., to the city administration. The results showed that the participants were skeptical about the impact of feedback on their trust. One of the findings of [15] was that citizens who feel connected to their municipality are more likely to raise complaints. These points motivated the evaluation of the impact on the trust into the local government in this work. On the one hand, in general, and on the other hand, in the context of the availability of contextual feedback. Resulting in the following hypothesis.

**H4:** The availability of contextual feedback in a conversational complaint management system will improve the citizens' trust in their municipality.

## 6 TASKS

The study participants were asked to visit the study website, where they were given the task for their assigned condition. The task

description and instructions were provided on this page. For condition A with "normal" llm-based feedback, the task was to report a complaint about a large amount of garbage in a forest near a street. The feedback was generated with the same prompt as condition B, except that the steps about the contextual information were removed to prevent hallucinations (see Figure 3). The task for condition B with the contextual feedback was reporting a complaint about garbage around recycling containers at one of the recycling centers in Wadgassen (see Figure 4). The interaction with the CUI happened over the custom chat widget that has been developed. After the conversation was completed, the button to the questionnaire was enabled, and the participants were asked to fill out the questionnaire. In addition to the items related to the hypotheses, additional questions were asked about the participants previous experience with chatbots, LLMs, and some demographic questions.

## 7 RESULTS

### 7.1 Participants

In total, 62 interactions (one per participant) with the system have been recorded. In 8 conversations, the participants did not follow the instructions and reported a complaint unrelated to the task. These conversations have been excluded from the analysis. As previously mentioned, this high rate of participants not following the

### Baseline Feedback

Thank you for your report. We understand your concern about the large amount of illegal waste in the forest at Butterborn. Disposing of waste in nature is not only illegal but also harmful to the environment. We will take immediate action to remove the waste and ensure that this does not happen again.

Your report is very important to us, and we will do our best to resolve the problem as quickly as possible. Thank you for your support in maintaining the cleanliness and beauty of our city.

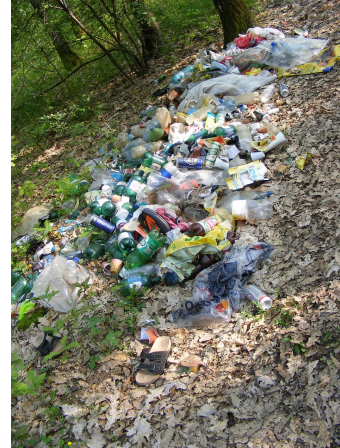


Figure 3: Condition A - Example Image and Feedback

### Contextual Feedback

Thank you for your complaint about a large amount of bulky waste at the recycling containers on Lindenstraße. We understand your concern and take it seriously.

We have classified the urgency of your issue as high to ensure that it is addressed as quickly as possible.

There are already four similar complaints about this issue. We are aware that this is a recurring problem and are working to find a solution.

The estimated processing time is one week. We will do our best to resolve the problem as quickly as possible and improve the situation. Please be patient while we address your concern. Thank you for your assistance in maintaining a clean environment in our city.



Figure 4: Condition B - Example Image and Feedback

task instructions was one of the reasons why the participant recruitment had to be extended. 7 of the 8 conversations happened while the study was only shared in Wadgassen. Out of the 32 conversations from participants that were recruited over the other channels, it happened once. Because the complaints were related to completely different topics like rats, the parking situation, bad street conditions, or parked cars, almost all of these conversations failed, and all except one did not fill out the questionnaire. One of the eight conversations has been classified as a troll conversation. Seven conversations have been successfully completed, but the participants did not fill out the survey. In total, 46 conversations following the task instructions have been completed, and one failed with a human hand-off because of 3 fallbacks in a row. The one participant who faced the simulated human hand-off did not fill out the survey. After removing four further results in which the

attention check has not been passed, 42 usable results remained (A = 22, B = 20). The participants were between 23 and 62 years old ( $M = 30.1$ ,  $SD = 18.05$ ). 22 males and 18 females, and 2 did not want to disclose their gender. The average previous experience with chatbots was 3.45 ( $SD = 1.24$ ) measured on a 5-point Likert scale, indicating moderate previous experience with chatbots. 23 participants (54.76%) answered that they had experience with LLMs. 17 participants (40.48%) answered yes to the optional question if they live in Wadgassen.

## 7.2 Descriptive Statistics

The descriptive statistics for the items measured for the two conditions can be found in the table below. The table has been created

by jamovi<sup>11</sup>. The usability of the chatbot has been measured with the Chatbot Usability Questionnaire (CUQ) [11]. The pragmatic and hedonic quality has been measured with the small version of the AttrakDiff questionnaire [10]. Whether the participants would be willing to use such a system again was measured with a single question on a 5-point Likert scale. The same goes for the feeling that the complaint is taken seriously. If the participants rated the feedback to the complaint as meaningful, it was measured by combining the question from the CUQ that the responses were "useful, appropriate and informative" with the question if the responses were irrelevant. The trust in the local government was measured by adapting the items from [14].

### 7.3 Hypotheses

Hypotheses H1.1, H1.2, H2, and H4 showed statistical significance with  $p$  values  $< 0.05$ , supporting their acceptance. Medium effect sizes were observed for the first three hypotheses, while H4 exhibited a large effect size. Conversely, H3, that contextual feedback is perceived as more meaningful, could not be accepted based on the results.

### 7.4 Qualitative Feedback

The participants were also asked three (optional) open questions about which information they would have liked to get from the chatbot, after reporting the complaint, if they got unwanted information, and if they had any other comments. In total, 17 comments for A and 20 for B have been collected. Because the included information was the variable that was changed between the two conditions, the answers to the first question about the desired information were especially interesting. The difference between the two conditions was directly visible. For condition A, seven participants (31.81% of A) responded, while for condition B, 5 participants (25% of B) responded. For condition A, 3 (42.86%) of the participants mentioned that they would have liked to get information about a rough time frame for the resolution of the complaint. Representing one of the contextual feedback elements that have been implemented. Two participants stated that they would have liked to get a contact person. One participant wanted more variation in the complaint categories and better knowledge of the locations, which is interpreted as improvement suggestions for the system and does not match the intention of the question. For condition B, one participant stated that the provided existence of similar complaints and the resolution time frame was sufficient for him/her, and one participant answered just with "no," indicating that he/she had no desire for additional information. Another participant wanted to get a process number for later inquiries. This was also commented on by one participant that it would have been helpful to know how the progress of the complaint could be tracked. The last participant for B stated that he/she was wondering how the chatbot detected that the complaint was about recycling containers even though she just mentioned containers and there are also other types of containers. The participant would have expected a selection process and a confirmation. In the study, this was caused by the fact that the feedback was pre-generated. However, in a real system, implementing the functionality to search for existing similar complaints, this information

could be fetched, for example, from the problem summary and the summaries of similar complaints. Alternatively, from a knowledge base (for example, simply the municipality website) that contains the container types and their location in the municipality. A similar comment was also made by one participant from condition A. This participant also stated that he/she guesses that this was potentially caused by previous complaints and suggested including this information in the response, followed by a verification if the information matches the complaint that should be reported.

For the second question about unwanted information, there were more responses for condition B, with 9 compared to 7 for condition A. For B, four participants (44% of the B comments, 20% of B) stated that the response of the system to the complaint was too long. One even described it as "overwhelming". This can be explained by the trade-off between providing additional information and simplicity mentioned in the concept chapter based on [26]. Two participants criticized the presence of "general" information in the response and would have preferred more specific information. A participant suggested that even though the information that four similar complaints already exist is helpful, he/she would have liked to know how these other complaints are handled. Three responses (33%) explicitly rejected the presence of unwanted information. For condition A, two participants (28.57% of A comments, 10% of A) described the response as too long and or too complex. One participant mentioned that part of the response raised more questions, such as how the removal of the garbage would be ensured. No participant mentioned any unwanted information present in the response for condition A. However, one participant of A mentioned that the response expressed a strong opinion that the pollution at this place was unacceptable and did not match his input. He further stated that this only increased the response size and created a feeling that his complaint was interpreted without the option to verify or deny it.

In response to the question about additional comments, the responses for A were the following points. Improvement suggestions like the support to send images and send responses similar to ChatGPT as a stream of words instead of single or multiple messages were made. One participant mentioned that he/she was surprised by the information that other similar complaints about the problem already exist but that he/she thinks that it is helpful. The same person also suggested that the system should also be able to handle shared locations so that, e.g., defective street lights could be reported with the GPS coordinates shared in WhatsApp. Another response mentioned that the information about the existence of similar complaints was unexpected but helpful, especially because you don't feel like the local police officer, according to the participant. The same response also suggested that a feature to share the location would be useful. Two responses stated that they liked the idea of the citizen complaint management chatbot. One of them also mentioned that he/she would like to see the system extended into an active participation platform and or to get information. The other mentioned that the responses created the feeling that the complaint is taken seriously but that he/she would have liked to get more detailed information about the further handling and resolution of the complaint.

<sup>11</sup><https://www.jamovi.org/>



**Table 1: Descriptives**

## Descriptives

	Condition	N	Mean	Median	SD	Minimum	Maximum
CUQ	A	22	73.011	73.438	13.101	31.25	90.63
	B	20	77.969	81.250	13.535	50.00	93.75
Attractiveness	A	22	0.341	0.000	0.836	-1.00	2.00
	B	20	0.975	1.000	1.106	-2.00	3.00
Pragmatic Quality	A	22	0.659	0.750	1.209	-1.00	2.25
	B	20	1.325	1.250	1.067	-1.00	3.00
Hedonic Quality	A	22	0.170	0.125	0.781	-1.75	1.50
	B	20	0.775	0.625	0.765	0.00	3.00
Reuse Intention	A	22	3.909	4.000	0.971	2	5
	B	20	4.550	5.000	0.605	3	5
Complaint is taken Seriously	A	22	3.227	3.000	1.066	1	5
	B	20	3.800	4.000	0.951	2	5
Meaningful Feedback	A	22	3.864	4.000	1.082	1.00	5.00
	B	20	4.300	4.500	0.894	2.00	5.00
Trust	A	22	3.080	3.000	0.418	2.25	3.75
	B	20	3.462	3.375	0.391	3.00	4.00

## 8 DISCUSSION

For H1.1, it was anticipated that the pragmatic quality in the contextual feedback condition would receive higher ratings. This expectation was substantiated with a medium effect size. Similarly, H1.2, positing higher hedonic quality in the contextual feedback condition, was confirmed with a large effect size. Moving on to H2.1, the expectation that users exposed to contextual feedback would demonstrate a greater willingness to use the system again was affirmed with a medium effect size. Condition B achieved an excellent average score, while Condition A was well-rated. Both conditions exhibited good CUQ scores, indicating favorable prospects for accepting such a system. Addressing the reluctance of citizens to report complaints due to concerns about their seriousness (as cited in [15]), H2.2 results indicated that contextual feedback significantly enhances the perception that complaints are taken seriously, confirming this hypothesis with a medium effect size. Contrary to expectations, H3, predicting higher ratings for the meaningfulness of contextual feedback, could not be substantiated. Notably, the contextual feedback condition had a slightly higher average rating of

4.3 compared to 3.86, with a small effect size (Cohen's  $d$ ). This inconclusiveness might be attributed to Condition A's already perceived good feedback and potential negative effects from some participants in Condition B finding the contextual feedback response excessively long. Future investigations may explore reducing response size to enhance perceived meaningfulness, but this could introduce trade-offs, potentially diminishing the seriousness of complaints. Remarkably, both conditions' responses were not perceived as irrelevant, a positive outcome considering the significant impact stated in [3]. Although initially contemplated, a comparison involving obviously false information was omitted in this study. Hypothesis H4, proposing that contextual feedback improves citizens' trust in their municipality, was validated with a large effect size. However, it's noteworthy that trust scores for both conditions were neutral for A and neutral to slightly positive for B, indicating the need for additional initiatives to augment or regain trust, aligning with the suggestions in [15]. As previously highlighted, (conversational) citizen complaint management stands as one essential component in this broader context. A notable number of participants deviated from task instructions, reporting unexpected problems. The cause

**Table 2: t-Test Results**

## Independent Samples T-Test

		Statistic	df	p	Mean difference	SE difference	95% Confidence Interval		Effect Size	
							Lower	Upper		
CUQ	Student's t	-1.21	40.0	0.118	-4.957	4.112	-Inf	1.9665	Cohen's d	-0.372
Attractiveness	Student's t	-2.11	40.0	0.021	-0.634	0.301	-Inf	-0.1275	Cohen's d	-0.651
Pragmatic Quality	Student's t	-1.88	40.0	0.033	-0.666	0.353	-Inf	-0.0709	Cohen's d	-0.582
Hedonic Quality	Student's t	-2.53	40.0	0.008	-0.605	0.239	-Inf	-0.2023	Cohen's d	-0.782
Reuse Intention	Student's t	-2.54	40.0	0.008	-0.641	0.253	-Inf	-0.2153	Cohen's d	-0.783
Complaint is taken Seriously	Student's t	-1.83	40.0	0.037	-0.573	0.313	-Inf	-0.0456	Cohen's d	-0.565
Meaningful Feedback	Student's t	-1.42	40.0	0.082	-0.436	0.308	-Inf	0.0825	Cohen's d	-0.438
Trust	Student's t	-3.06	40.0	0.002	-0.383	0.125	-Inf	-0.1719	Cohen's d	-0.944

Note.  $H_a: \mu_A < \mu_B$

remains to be clarified, possibly due to inadequate task description comprehension or participants disregarding it, assuming a fully operational complaint management system. This misconception may stem from the study's promotion in local media, yielding insights into potential complaint topics despite unusable interactions. High interest in progress updates and resolution notifications suggests implementing these features in a real system. Real-time monitoring may not be essential, given users' neutral-to-positive ratings. Reviewing failed conversations and formulating manual responses for progress updates is recommended. WhatsApp and a website chat widget are preferred interaction channels, with additional channels added based on demand. Notifications via WhatsApp and email for the website could enhance citizens' perception of the value of complaint reporting, aligning with recommendations from [3]. Some in Condition B found responses too long (20%), indicating a trade-off between simplicity and information. Evaluating the response size carefully is advised to avoid negative impacts, considering participants' varying opinions on length. Surprisingly, even with significantly longer responses in Condition B, some in Condition A (10%) also found them too lengthy, emphasizing the need for cautious adjustments to prevent adverse effects.

## 9 CONCLUSION AND FUTURE WORK

In conclusion, our study significantly advances the understanding of how contextual feedback in CUIs impacts user perception and engagement in the domain of citizen complaint management. The hypotheses testing revealed that contextual feedback positively influences both the pragmatic and hedonic quality of the user experience, with medium to large effect sizes confirming the anticipated benefits. Furthermore, our findings indicate a marked increase in users' willingness to re-engage with the system and perceive their complaints as being taken seriously when contextual feedback is employed, underscoring its value in enhancing user satisfaction and trust towards municipal systems.

Despite these positive outcomes, the study also uncovered challenges to address in future research, particularly in optimizing feedback length without compromising the depth of information. The mixed reactions to the feedback's length highlight the delicate balance required to maintain user engagement without overwhelming them. The work of this paper also revealed practical considerations for implementing such systems, including the importance of clear task descriptions and the potential for misinterpretations among participants. The expressed preference for specific communication

channels and features, like progress updates and resolution notifications, provides valuable direction for developing more responsive and user-friendly complaint management systems.

The positive results achieved by the baseline system show the potential of LLMs for complaint management systems and related tasks, even in the absence of additional data sources, relying solely on the LLM's inherent knowledge of the world. An additional advantage lies in the system's deployability without the need for supplementary data collection and integrations. Regarding the preliminary release of GPT-4V, the multi-modal variant of OpenAI's GPT-4, it is currently accessible for preview and demonstrates promising outcomes, particularly in its proficiency in comprehending accompanying images. However, it is essential to acknowledge that GPT-4V is still in its initial stages. As emphasized in [21], the integration of multimodality introduces new risks and challenges that require careful consideration to mitigate potential model exploits. Exploring the synergy between contextual feedback and multimodality presents an intriguing avenue for future research. In alignment with ideas from [3], further consideration could be given to evaluating whether the introduction of such a system could enhance citizen engagement across various domains, potentially contributing to overall citizen satisfaction. The prospect of LLMs providing automated feedback on complex subjects, as discussed in [15], becomes more feasible if the necessary data for reasoning is made readily accessible.

## ACKNOWLEDGMENTS

This research is funded by the German Federal Ministry of Education and Research (BMBF) project PYSA (16SV8850).

## REFERENCES

- [1] Shubham Atreja, Pooja Aggarwal, Prateeti Mohapatra, Amol Dumrewal, Anwesh Basu, and Gargi B Dasgupta. 2018. Citicafe: An Interactive Interface for Citizen Engagement. *International Conference on Intelligent User Interfaces, Proceedings IUI* (2018), 617–628. <https://doi.org/10.1145/3172944.3172955>
- [2] M. Batty, K. W. Axhausen, F. Giannotti, A. Pozdnoukhov, A. Bazzani, M. Wachowicz, G. Ouzounis, and Y. Portugali. 2012. Smart cities of the future. *European Physical Journal: Special Topics* 214 (12 2012), 481–518. Issue 1. <https://doi.org/10.1140/epjst/e2012-01703-3>
- [3] Christian Brock, Julia Meik, Janine Kaiser, Sadi Al-Dari, and Bernd Handke. 2012. *Der Bürger als Kunde Anliegenmanagement im öffentlichen Sektor*. Springer Fachmedien Wiesbaden. 109–124 pages. [https://doi.org/10.1007/978-3-531-19347-2\\_6](https://doi.org/10.1007/978-3-531-19347-2_6)
- [4] Andrés Camero and Enrique Alba. 2019. Smart City and information technology: A review. *Cities* 93 (10 2019), 84–94. <https://doi.org/10.1016/j.cities.2019.04.014>
- [5] Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How Should My Chatbot Interact? A Survey on Social Characteristics in Human–Chatbot Interaction Design. *International Journal of Human–Computer Interaction* 37 (2021), 729–758. Issue 8. <https://doi.org/10.1080/10447318.2020.1841438>
- [6] Anind Dey. 2001. Understanding and Using Context. *Personal and Ubiquitous Computing* 5 (02 2001), 4–7. <https://doi.org/10.1007/s007790170019>
- [7] Anne Fota, Katja Wagner, Tobias Roeding, and Hanna Schramm-Klein. 2022. “Help! I Have a Problem” – Differences between a Humanlike and Robot-like Chatbot Avatar in Complaint Management. *Proceedings of the 55th Hawaii International Conference on System Sciences* 7 (2022), 4273–4282. <https://doi.org/10.24251/hicss.2022.522>
- [8] Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2018. Towards Automated Customer Support. (2018), 48–59. [https://doi.org/10.1007/978-3-319-99344-7\\_5](https://doi.org/10.1007/978-3-319-99344-7_5)
- [9] Mike Harding, Bran Knowles, Nigel Davies, and Mark Rouncefield. 2015. HCI, Civic Engagement and Trust. *Conference on Human Factors in Computing Systems - Proceedings* 2015-April (4 2015), 2833–2842. <https://doi.org/10.1145/2702123.2702255>
- [10] Marc Hassenzahl, Michael Burmester, and Franz Koller. 2003. *AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität*. Vieweg+Teubner Verlag, 187–196. [https://doi.org/10.1007/978-3-322-80058-9\\_19](https://doi.org/10.1007/978-3-322-80058-9_19)
- [11] Samuel Holmes, Anne Moorhead, Raymond Bond, Huiyu Zheng, Vivien Coates, and Michael Mctear. 2019. Usability Testing of a Healthcare Chatbot: Can We Use Conventional Methods to Assess Conversational User Interfaces? (*ECCE '19*). Association for Computing Machinery, New York, NY, USA, 207–214. <https://doi.org/10.1145/3335082.3335094>
- [12] Soonhee Kim and Jooho Lee. 2012. E-Participation, Transparency, and Trust in Local Government. *Public Administration Review* 72, 6 (2012), 819–828. <https://doi.org/10.1111/j.1540-6210.2012.02593.x>
- [13] Maximilien Kintz, Claudia Dukino, Matthias Blohm, and Marc Hanussek. 2020. Make Your Customers Happy Again: AI and NLP for a Customer Complaint Management Platform. (2020). <https://publica-rest.fraunhofer.de/server/api/core/bitstreams/222a46f8-cbcb-4514-8da5-c08feb984c1/content>
- [14] Geok Theng Lau and Sook Han Lee. 1999. Consumers' Trust in a Brand and the Link to Brand Loyalty. , 341–370 pages. <https://doi.org/10.1023/A:1009886520142>
- [15] Maria Leitner. 2018. *Digitale Bürgerbeteiligung Forschung und Praxis-Chancen und Herausforderungen der elektronischen Partizipation*. Springer Vieweg Wiesbaden. <https://doi.org/10.1007/978-3-658-21621-4>
- [16] Alessandro Massaro, Vincenzo Maritati, and Angelo Galiano. 2018. Automated Self-Learning Chatbot Initially Built as a FAQs Database Information Retrieval System: Multi-level and Intelligent Universal Virtual Front-Office Implementing Neural Network. *Informatica (Slovenia)* 42 (2018), 515–525. Issue 4. <https://doi.org/10.31449/inf.v42i4.2173>
- [17] Anna S. Mattila and Daniel J. Mount. 2003. The impact of selected customer characteristics and response time on E-complaint satisfaction and return intent. *International Journal of Hospitality Management* 22, 2 (2003), 135–145. [https://doi.org/10.1016/S0278-4319\(03\)00014-8](https://doi.org/10.1016/S0278-4319(03)00014-8)
- [18] Michael Mctear. 2020. *Conversational AI*. Springer Cham. <https://doi.org/10.2200/S01060ED1V01Y202010HLT048>
- [19] Piero Molino, Huaixiu Zheng, and Yi-Chia Wang. 2018. COTA: Improving the Speed and Accuracy of Customer Support through Ranking and Deep Networks. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 586–595. <https://doi.org/10.1145/3219819.3219851>
- [20] Jae Yun Moon and Lee S. Sproull. 2008. The Role of Feedback in Managing the Internet-Based Volunteer Work Force. *Information Systems Research* 19, 4 (2008), 494–515. <https://doi.org/10.1287/isre.1080.0208>
- [21] OpenAI. 2023. GPT-4V(ision) System Card. [https://cdn.openai.com/papers/GPTV\\_System\\_Card.pdf](https://cdn.openai.com/papers/GPTV_System_Card.pdf)
- [22] Cathy Pearl. 2016. *Designing Voice User Interfaces: Principles of Conversational Experiences*. " O'Reilly Media, Inc."
- [23] Holger Roschk, Masoumeh Hosseinpour, and Golnaz Rezaei. 2021. Supporting Complaint Management with Artificial Intelligence. (2021), 213–232. [https://doi.org/10.1007/978-3-658-34326-2\\_8](https://doi.org/10.1007/978-3-658-34326-2_8)
- [24] Andreas Sackl, Sarah Thiel, and Peter Fröhlich. 2019. Smarticipate Citizen Dialog Requirements. [https://www.smarticipate.eu/wp-content/uploads/smarticipate\\_D2.2-Citizen-dialog-design-requirements.pdf](https://www.smarticipate.eu/wp-content/uploads/smarticipate_D2.2-Citizen-dialog-design-requirements.pdf)
- [25] Andreas Sackl, Sarah Kristin Thiel, Peter Fröhlich, and Manfred Tscheligi. 2018. “Thanks for Your Input. We Will Get Back to You Shortly.” How to Design Automated Feedback in Location-Based Citizen Participation Systems. *Lecture Notes in Geoinformation and Cartography* 0, 257–268. Issue 208669. [https://doi.org/10.1007/978-3-319-71470-7\\_13](https://doi.org/10.1007/978-3-319-71470-7_13)
- [26] Dirk Schelle and Fernando Lyardet. 2006. Voice User Interface Design Patterns. <https://pdfs.semanticscholar.org/6e08/7ddc8a262659fa211a2c9dc26a41e758b989.pdf>
- [27] Ryan M. Schuetzler, G. Mark Grimes, Justin Scott Giboney, and Holly K. Rosser. 2021. Deciding Whether and How to Deploy Chatbots. *MIS Quarterly Executive* (Jan. 2021), 1–15. <https://doi.org/10.17705/2msqe.00039>
- [28] Priyanka Singh, Fiona Lynch, and Markus Helfert. 2021. Role of Citizens in the Development of Smart Cities: Benefit of Citizen's Feedback for Improving Quality of Service. (2021), 35–44. Issue Smartgreens. <https://doi.org/10.5220/0010442000350044>
- [29] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A New Chatbot for Customer Service on Social Media. *Conference on Human Factors in Computing Systems - Proceedings* 2017-May, 3506–3510. <https://doi.org/10.1145/3025453.3025496>
- [30] Hongyu Zhang, Liang Gong, and Steve Versteeg. 2013. Predicting Bug-Fixing Time: An Empirical Study of Commercial Software Projects. In *2013 35th International Conference on Software Engineering (ICSE)*. IEEE, 1042–1051. <https://doi.org/10.1109/icse.2013.6606654>
- [31] Andreas Zimmermann, Andreas Lorenz, and Reinhard Oppermann. 2007. An Operational Definition of Context. In *Modeling and Using Context*, Boicho Kokinov, Daniel C. Richardson, Thomas R. Roth-Berghofer, and Laure Vieu (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 558–571. [https://doi.org/10.1007/978-3-540-74255-5\\_42](https://doi.org/10.1007/978-3-540-74255-5_42)