

## Research Article





**Cite this article:** Dortheimer J, Martelaro N, Sprecher A and Schubert G (2024). Evaluating large-language-model chatbots to engage communities in large-scale design projects. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, **38**, e4, 1–16  
<https://doi.org/10.1017/S0890060424000027>

Received: 01 June 2023  
Revised: 24 December 2023  
Accepted: 06 February 2024

**Keywords:**  
chatbot; urban design; participatory design;  
large language model; crowdsourcing

**Corresponding author:**  
Jonathan Dortheimer;  
Email: [jonathand@ariel.ac.il](mailto:jonathand@ariel.ac.il)

# Evaluating large-language-model chatbots to engage communities in large-scale design projects

Jonathan Dortheimer<sup>1</sup> , Nik Martelaro<sup>2</sup> , Aaron Sprecher<sup>3</sup>  and Gerhard Schubert<sup>4</sup> 

<sup>1</sup>School of Architecture, Ariel University, Ariel, Israel; <sup>2</sup>Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA, USA; <sup>3</sup>Faculty of Architecture, Technion Israel Institute of Technology, Haifa, Israel and <sup>4</sup>TUM School of Engineering and Design, Technical University of Munich, Munich, Germany

## Abstract

Recent advances in machine learning have enabled computers to converse with humans meaningfully. In this study, we propose using this technology to facilitate design conversations in large-scale urban development projects by creating chatbot systems that can automate and streamline information exchange between stakeholders and designers. To this end, we developed and evaluated a proof-of-concept chatbot system that can perform design conversations on a specific construction project and convert those conversations into a list of requirements. Next, in an experiment with 56 participants, we compared the chatbot system to a regular online survey, focusing on user satisfaction and the quality and quantity of collected information. The results revealed that, with regard to user satisfaction, the participants preferred the chatbot experience to a regular survey. However, we found that chatbot conversations produced more data than the survey, with a similar rate of novel ideas but fewer themes. Our findings provide robust evidence that chatbots can be effectively used for design discussions in large-scale design projects and offer a user-friendly experience that can help to engage people in the design process. Based on this evidence, by providing a space for meaningful conversations between stakeholders and expanding the reach of design projects, the use of chatbot systems in interactive design systems can potentially improve design processes and their outcomes.

## Introduction

Recent advances in machine learning (ML) have enabled computers to converse with humans in meaningful ways (Brown et al., 2020; OpenAI, 2023). In this paper, we argue that this technology can potentially revolutionize the design process, particularly in large-scale urban development projects. Traditionally, in small architecture projects, there is a negotiation between a client (the end user) and the architect; during this negotiation, information is exchanged, thereby facilitating the progression of reflection into an agreement among all stakeholders (McDonnell, 2009; Oak, 2009). However, in large-scale projects, the client is typically a governmental agency or developer who is not the end user of the constructed buildings, making it extremely challenging to engage in meaningful conversations with thousands of potential stakeholders who are the end users.

Since the mid-20th century, such urban development projects have faced extensive criticism due to their disconnect with end users, often resulting in underperforming designs (Alexander, 1964), the destruction of thriving neighborhoods (Jacobs, 1961), and a lack of inclusivity and democracy (Harvey, 1973), which affects marginalized communities (Arnstein, 1969). This issue has become increasingly relevant as cities continue to densify, and urban renewal projects significantly impact various aspects of urban life, including social, economic, and environmental factors. In response to this growing need, numerous participatory design methods have emerged since the 1970s (Simonsen and Robertson, 2012), aiming to incorporate diverse perspectives in architectural projects (Luck, 2018). By involving end users in the design process, these methods foster a more comprehensive approach to urban development, better suited to address the complex challenges faced by contemporary cities.

However, despite the widespread agreement on the significance of community participation in fostering sustainable development, promoting democratic culture, and creating equitable communities (Münster et al., 2017; Calderon, 2020), the practical implementation of participatory design in urban design remains challenging. Various factors have hindered effective public participation, including intra-community politics and power dynamics (Krüger et al., 2019), bureaucratic obstacles and red tape (Brabham, 2009), knowledge gaps between experts and community members (Dortheimer and Margalit, 2020), and a pervasive lack of public trust in politicians and local authorities (Giering, 2011). Additionally, the considerable time and effort

© The Author(s), 2024. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-ShareAlike licence (<http://creativecommons.org/licenses/by-sa/4.0/>), which permits re-use, distribution, and reproduction in any medium, provided the same Creative Commons licence is used to distribute the re-used or adapted article and the original article is properly cited.

required from participants and organizers can impede the successful execution of participatory design processes.

In response to these challenges, researchers and practitioners have explored various strategies to enhance participatory design by leveraging digital tools to enable more accessible and inclusive participation (Luck, 2018). The emergence of crowdsourcing technologies has significantly strengthened the trend toward participatory design in recent years (Robertson and Simonsen, 2012; Gooch *et al.*, 2018, Dortheimer *et al.*, 2020). These technologies facilitate individual communication, thereby relieving political pressure on participants and allowing them to express their opinions freely and, when necessary, anonymously (Dortheimer *et al.*, 2023). Numerous studies have examined the application of crowdsourcing in architecture and urban design, encompassing ideation (Lu *et al.*, 2018), architectural design (Dortheimer, 2022), co-creation (Mueller *et al.*, 2018; Hofmann *et al.*, 2020), mapping (Borges *et al.*, 2015), and opinion-gathering (Hosio *et al.*, 2015; Wang *et al.*, 2021). However, survey-like methods remain the primary method for collecting information on an urban scale.

Unlike surveys, conversations can be more open and flexible, fostering an environment that encourages stakeholders to reflect upon and share their ideas and experiences. This interactive approach to data collection allows for a deeper understanding of the participants' perspectives and emotions.

Therefore, we argue that chatbots have the potential to revolutionize the field of urban design by facilitating meaningful conversations with a diverse array of stakeholders. Some studies have indicated that chatbots may be more effective for gathering respondents' information than traditional surveys (te Pas *et al.*, 2020; Xiao *et al.*, 2020a,b). Furthermore, the engaging nature of chatbot-facilitated conversations can help maintain participants' interest and make the interaction more enjoyable. Finally, adopting chatbots in urban design can lead to more comprehensive and accurate insights, driving more effective decision-making and resulting in better-designed urban spaces that cater to the needs and preferences of all stakeholders.

Implementing chatbots in participatory urban design projects can lead to an overwhelming amount of conversational data, which poses a significant challenge for human designers. Consequently, it is essential to develop a comprehensive framework that not only streamlines chatbots' effective communication with stakeholders but also enhances the efficient analysis and summarization of the vast data collected from these interactions. By doing so, this framework would transform chatbots into practical and valuable tools in participatory urban design processes, equipping urban planners with meaningful and actionable insights while effectively managing communication and data volume.

Furthermore, there is a limited understanding of the differences in the quality and quantity of information gathered from such chatbot frameworks compared to traditional surveys in the context of urban design. This research gap warrants a comprehensive examination to inform better the implementation of chatbot interventions in urban planning processes.

Consequently, this research aims to investigate the use of chatbots within the realm of participatory urban design and assess their efficacy compared to conventional survey methods. The research question addressed in this study is as follows: "What are the differences between using a chatbot framework and surveys to collect information and design ideas in the context of urban design?"

To answer the research question, we develop and test a chatbot framework capable of performing design conversations and

summarizing such conversations into design requirements. Next, in an experiment with 56 participants, we compared the chatbot framework to a traditional online survey, focusing on the quality and quantity of collected information.

The contributions of this paper are as follows: 1) We present a novel chatbot system and human-artificial intelligence (AI) prompt framework for initiating and managing design conversations in large-scale urban design projects, contributing to the emerging field of AI-assisted participatory design; 2) we perform a comprehensive experiment involving 56 participants, comparing the chatbot framework with traditional online surveys in the context of participatory urban design that offers insight into the differences between chatbot and survey outputs regarding the quality and quantity of information collected in the context of urban design, addressing the current research gap; and 3) we propose recommendations for effectively using the chatbot framework in participatory urban design, based on the findings and observations from the experiment, to facilitate more productive and engaging interactions between stakeholders and designers.

## Related works

In this section, we provide an overview of the related work in the areas of conversation in the design process, chatbots, design, and the large language models (LLMs) to establish the context for our study.

### Participatory design

Proponents of participatory design maintain that a more suitable design fit can be achieved when end users actively participate in the design process (Reich *et al.*, 1996). This approach brings together individuals with diverse backgrounds, roles, and expertise to collaboratively examine a problem and collectively generate potential solutions. The concept of incorporating local residents into urban and architectural design planning has been in practice for over five decades (Luck, 2018). Generally, urban planning research posits that community involvement fosters democratic values and equitable communities, making it a vital component of sustainable development (Münster *et al.*, 2017; Calderon, 2020).

### The role of conversation in the design process

While design frequently involves working with visual representations, verbal communication plays a crucial role in the design process. Verbal conversations with various stakeholders, such as clients, contractors, and community representatives, allow designers to understand their needs (Lawson and Loke, 1997; Dubberly and Pangaro, 2019) and constraints better and negotiate project requirements (McDonnell, 2009). Similarly, conversations among designers facilitate knowledge sharing and negotiation of new ideas, allowing design teams to transcend the individual abilities of a single designer (Arias *et al.*, 2000). Conversations with end users and community members help explicate their needs and concerns about a design project, making tacit knowledge explicit from the user's perspective (Luck, 2003). Although verbal communication is essential, it is important to consider its challenges and limitations, such as potential misunderstandings and difficulties translating verbal ideas into concrete design elements (Karlgrén and Ramberg,

2012). In early architectural discussions between designers and clients, clients often concentrate on familiar functional and structural aspects of building designs, while designers seek to identify problems and understand the design's significance to the client (Luck and McDonnell, 2006). Understanding how project stakeholders communicate verbally to express their needs is a valuable insight for developing effective design chatbots.

### Chatbots and the automation of conversation

The history of chatbots dates back to the first chatbot, ELIZA, developed between 1964 and 1966, which was based on a language model that identified keywords and, following a set of rules, provided a response (Weizenbaum, 1983). Over the years, chatbots have evolved significantly, owing to advancements in ML and AI. There are now three primary types of chatbots: rule-based, retrieval-based, and generative models (Hussain et al., 2019). Rule-based chatbots function based on a predefined set of rules, while retrieval-based chatbots choose responses from a pre-built database. However, both are limited by linguistic knowledge hard-coded into their software (Shawar and Atwell, 2005). Therefore, generative models powered by ML techniques have shown the most progress in recent years since these models can construct novel responses, adapting better to various conversational situations.

Overall, building a chatbot that can understand complex conversations and answer appropriately was reported to be a challenging task (Xiao et al., 2020a). However, modern chatbots such as Apple's Siri or Amazon's Echo leverage ML techniques to create LLMs and web search results to produce meaningful responses.

At the time we completed the work described here, GPT-3 was the largest publicly available LLM that produces human-like text (Brown et al., 2020). The model includes 175 billion parameters and produces high-quality texts. The model generates texts based on a provided text prompt. For instance, if a prompt is the beginning of a story, the model would try to predict the continuation of that story. Several previous applications demonstrated that this model can be meaningfully used for grammar correction, summarizing, answering questions, parsing unstructured data, classification, and translation (Radford et al., 2019; Brown et al., 2020), among other tasks. LLMs have also been used to build chatbots to capture self-reported user data chatting (Wei et al., 2023). However, since the model cannot reason, solve mathematical and ethical questions, or pass the Turing test (Floridi and Chiriatti, 2020), it is regarded as a human-like text generator rather than a general AI in the strict sense.

To build useful chatbots, an LLM must be provided with a well-designed prompt that can steer the LLM to generate topic-relevant text. Designing these prompts, as described by Zamfirescu-Pereira et al., is akin to "herding AI cats" due to the unpredictable nature of LLMs (Zamfirescu-Pereira et al., 2023). The challenges in crafting good prompts are manifold. For instance, adding a new instruction to repair specific issues found using a previous prompt might unpredictably affect other instructions. In addition, the model may generate "hallucinations," which are instances of fabricated information. Furthermore, the GPT-3 model could not acknowledge that it did not know some information. Similarly, a study on prompting generative image models for design highlighted the unpredictability and challenges of using these models (Dortheimer et al., 2023b). The impact of prompts on model outputs and prompting techniques are active areas of natural language processing research (White et al., 2023).

### Chatbots for creativity and design

Several previous studies have utilized human–chatbot interactions to generate creative ideas in various design fields (Kulcke, 2018; Cuadra et al., 2021; Shin et al., 2022). For instance, a notable study investigated how humans converse with a perceived AI during a Wizard of Oz study where designers prototyped speech interaction with music systems (Martelaro et al., 2020). Other studies have explored human–chatbot interactions in spatial design (Kulcke, 2018; Dortheimer et al., 2023) and ornament design (Cuadra et al., 2021). Additionally, chatbots have been employed to mediate consensus-building conversations (Shin et al., 2022).

Moreover, numerous studies have focused on the development (Ahmed, 2019) and evaluation (Tavanapour and Bittner, 2018; Hwang and Won, 2021) of chatbots for ideation tasks. Chatbots have also been utilized to facilitate design thinking through the empathy map method (Bittner and Shoury, 2019). An intriguing example is CharacterChat, a chatbot designed to assist writers in creating fictional characters (Schmitt and Buschek, 2021). However, to the best of the authors' knowledge, no studies have specifically examined the use of chatbots in urban design tasks.

### Comparisons of chatbots and surveys

Another challenge in human–chatbot communication is the human behavior and understanding that needs to be taken into account (Nguyen et al., 2022). First, compared to human–human interaction, human–chatbot communication was reported to take longer and include shorter messages (Hill et al., 2015). In addition, there is evidence that human–chatbot messages lack vocabulary richness and can contain profane words (Hill et al., 2015). However, compared to web surveys with open-ended questions, chatbots were reported to produce longer and richer responses from humans (Xiao et al., 2020b). Interestingly, some evidence suggests that humans can generate more or better-quality ideas when communicating with a chatbot rather than with a human partner (Hwang and Won, 2021).

Furthermore, with regard to effectiveness in eliciting information from respondents, several previous reports noted that chatbot interfaces could be more effective (Xiao et al., 2020a,b) and preferable (te Pas et al., 2020) than surveys. A reason that could underlie this finding is that there is evidence suggesting that people are more willing to share information through chatbots (Lee et al., 2020), which are generally believed to be useful for collaboration (Kim et al., 2021). Together with LLMs such as GPT-3, recent research has explored the usability of these models in operating chatbots (Wei et al., 2023). In line with this novel technology, in the present study, we empirically test a chatbot framework that may be more effective than survey methods for eliciting meaningful responses from participants in the context of urban design.

### Chatbot design and development

We developed a chatbot system with dual functionalities to discuss an urban design project. The first function involves conversing with users to gather their responses and insights about the project. The second function is to analyze and extract a set of design requirements from these conversations.

The chatbot design and development process can be broken down into several stages. Initially, we experimented with a "mock" chatbot to explore design conversations. Next, we constructed a prototype utilizing an LLM for the chatbot system. Finally, we

enhanced the chatbot's performance through a series of experiments and testing.

### Exploring design conversations

We developed a chatbot prototype to investigate the ideal structure of human–chatbot design conversations. In order to achieve this, we conducted a “Wizard of Oz experiment” wherein a human subject engaged in a discussion about an architecture project of their choice with a human-operated chatbot.

A chat system was created to enable the participants to converse with a human architect, as depicted in Figure 1a. The chatbot was operated by a certified and experienced architect, who was instructed to engage in a natural conversation with the participants. Each chatbot message was automatically converted into audio using text-to-speech technology and played when the message was displayed to the user, facilitating a more authentic interaction. The chat was initiated with a predefined user prompt introducing the chatbot: “I’m a design bot and would love to speak with you about a design project of your choice!” The conversations were recorded in a database. The chat logs were then summarized to create a set of design requirements for the design project. The participant would then rate these design requirements, as shown in Figure 1b. Participants were later interviewed to gather their experiences and suggestions for improving the dialogue. Finally, the recorded conversations were thoroughly analyzed.

### Preliminary experiment

After the initial test with the Wizard of Oz controlled chatbot, we implemented a preliminary chatbot using the GPT-3 LLM (text-davinci-001) as the backend of the application instead of the human operator (see the appendix section “Final chatbot implementation” for further detail). Since GPT-3 did not know when the conversation was over and always produced new responses to user inputs, we

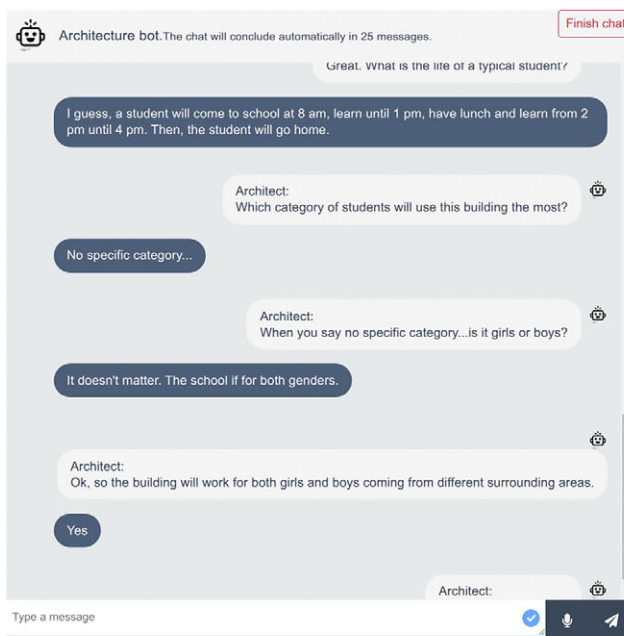
added a “finish chat” button to allow the participants to conclude the discussion. Alternatively, the chat automatically ended after exchanging 50 messages. Then, the discussion log was automatically summarized into a list of design requirements that resulted from the discussion transcript. The participants then rated the correctness of each requirement on a five-point Likert-type scale. Finally, the process concluded when the participants provided requirement evaluations.

In order to test the new chatbot's performance, we conducted a preliminary experiment where participants ( $n = 51$ ) were asked to discuss an architecture project of their choice with the bot, rate the automatically produced requirement list, and answer a user experience survey. The participants were students who used the chatbot during a design class at a university. Based on the results of this test, the chatbot was improved and fine-tuned to be used later in the controlled experiment. In addition, the chatbot LLM was updated to the newer “text-davinci-002” for the conversations and “text-curie-001” for design requirement summarizing.

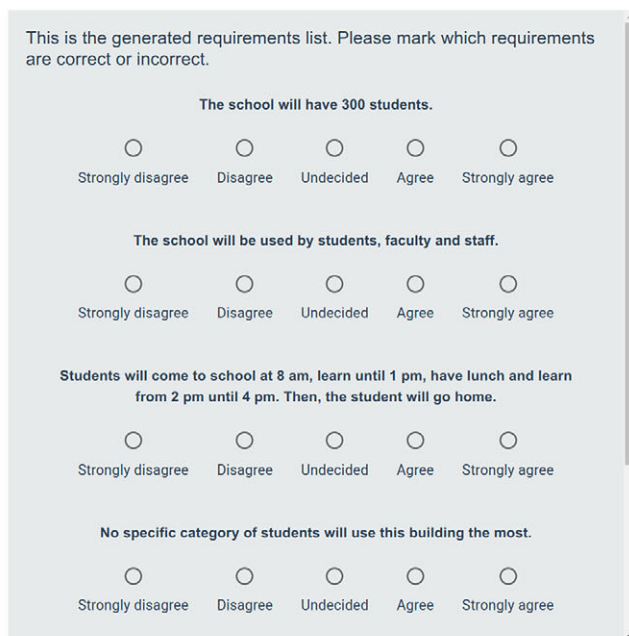
### Human–AI prompt framework

The present study generated chatbot responses using the improved “text-davinci-002” GPT-3 LLM. The request to the LLM included multiple parameters, including a “text prompt” that the model would use as input to predict how the text would continue. The text prompt is the foundation for the chatbot's ability to comprehend and respond to user inputs in a meaningful and contextually relevant manner. It is the most critical element, as the quality and relevance of the chatbot's responses heavily depend on the information and context provided by the prompt.

However, a part of the prompt influences how the human engages with the chatbot since it sets the stage for their expectations from the conversation. Designing the optimal prompt is



(a) Chat interface also used for the Wizard of Oz experiment



(b) Requirement list user evaluation interface

Figure 1. Chatbot user interface.

challenging, given the nuances of human language and the need to account for numerous conversational scenarios. Consequently, extensive testing and fine-tuning are necessary to identify the appropriate balance of context, specificity, and flexibility to achieve the desired performance (Zamfirescu-Pereira et al., 2023).

Our chatbot prompt contained the context of the conversation, including a specification of whether it was a conversation between an architect and a client, some character descriptions, and conversation goals. To design a chatbot operated solely by an LLM, we used the terms *internal prompt* and *shared prompt* to describe the structure of the “prompt” as the input of the LLM. The key difference between the two prompts was that the shared prompt also acted as a user prompt to provide the human with a shared understanding of the design situation and expectations from the conversation (see Fig. 2). Our chatbot implementation can be viewed in the appendix section “Final chatbot implementation.”

The internal prompt includes the definitions of the human and bot personas and the technical context of the conversation provided solely to the LLM. In our case, we defined the chatbot first as an architect conversing with a client. In the first experiments, we noticed that the chatbot sometimes stopped discussing the design, asked to show site photos, presented non-existing sketches, or negotiated for payment and budget.

In order to address the observed shortcomings, we modified the prompt by specifying that it involves a “conversation between an architect and her client.” We also added to the prompt that the

conversation focuses on aspects such as aesthetics, functional elements, and social preferences of the design.

Additionally, we noted that providing “The architect is very kind and professional” in the prompt led the chatbot to agree with most of the participants’ suggestions without further discussion. In order to generate more engaging and reflective conversations, we replaced the prompt as mentioned above with “The architect is challenging the client with questions to gain a deeper mutual understanding of the requirements.” This adjustment helped make the chatbot more critical without dismissing the participants’ ideas.

The shared prompt was also improved between the experiments and included the names of the chatbot and the human, relevant project information, and the appropriate context of the conversation. We started with a generic prompt, “Hi, I am your architect. Let us discuss your architecture project. I would like to know what kind of project you had in mind?” This allowed the participants to discuss any architecture project, which resulted in different kinds of architecture project conversations with various qualities.

We added personal attributions to the shared prompt by asking the participants to provide their names before the chat started. We chose the chatbot to be female and named it Zaha, in reference to the late influential architect Zaha Hadid. To reduce the risk of conversations about business issues, we stated that the chatbot is part of a design team. This resulted in the following shared prompt “Hello username, my name is Zaha, and I am on the design team of the project.”

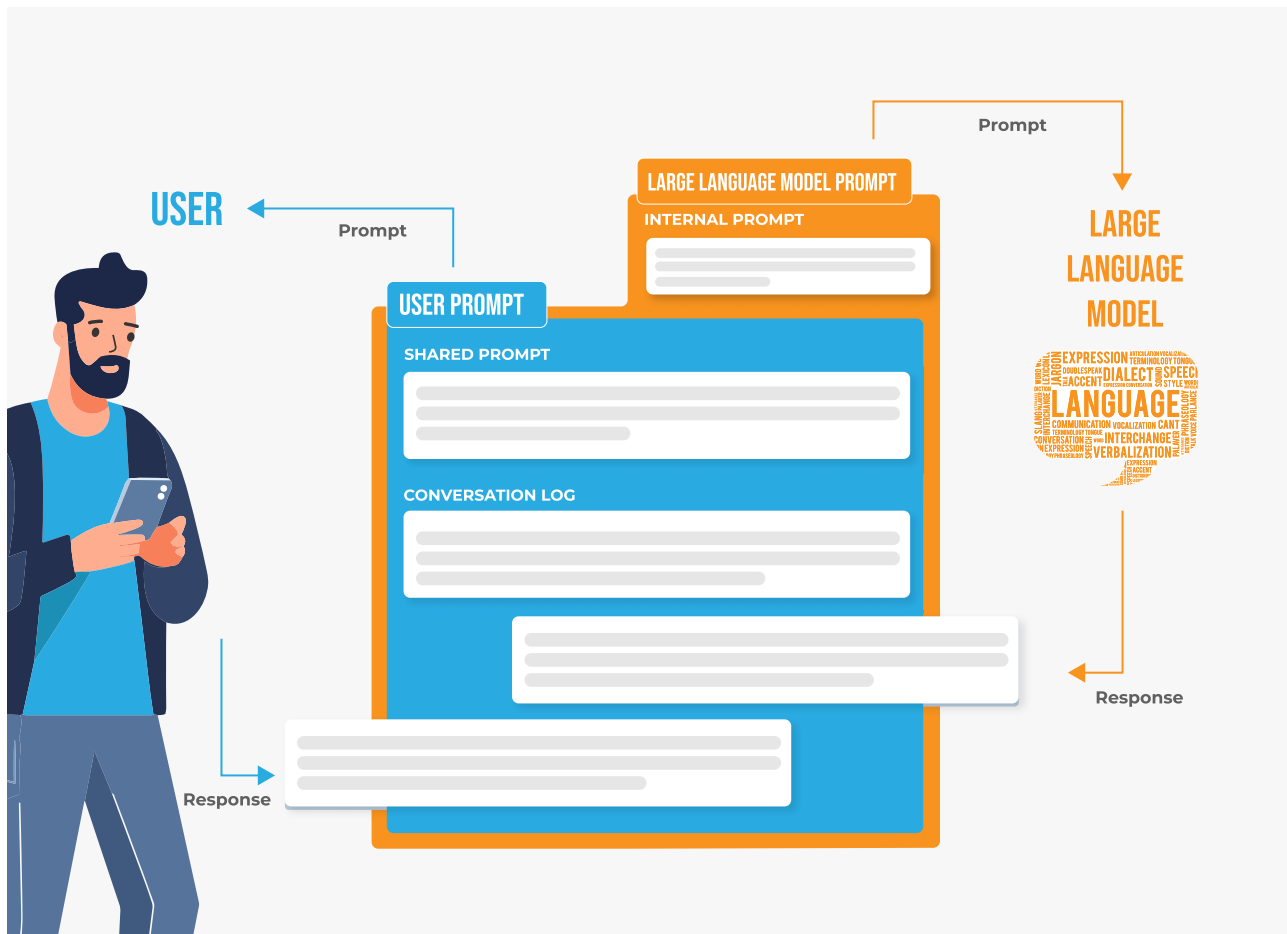


Figure 2. Conceptual diagram of the LLM prompt and user prompt, made out of internal and shared prompts.

Next, we wanted the chatbot to discuss a single project with several participants to investigate how to aggregate several conversations. To this end, we outlined a specific construction project the participants would be familiar with – a new architecture school building on campus. The project was presented with the following shared prompt: “The (Technical University of Munich) is planning to build a new architecture school instead of the outdated (electrical facility on Theresienstrasse). The building will host the architecture school, and the design will be based on the preferences of the students and faculty. That’s why we want to ask you about your ideas for the new building.”

Finally, we defined the design conversations’ scope and goals with the criteria we thought were important to address. These included the following: functionality, aesthetics, cultural values, and the desired social effect. This was done with the following shared prompt: “Let’s discuss the project requirements. What spaces should be in the building? How can the building have a positive impact on the community and the environment? What should the building look like? What values should the building express?” We found that the detailed shared and internal prompts produced better conversations that were more focused and produced higher-quality chatbot texts that helped the chatbot keep the discussion on topic and ask relevant questions.

In summary, the internal and shared prompt structure was a practical, functional approach to designing communication about urban design between LLM-based chatbots and humans.

## Method

Upon chatbot improvement, we conducted a controlled experiment to answer our research question (“What are the differences between using a chatbot framework and using surveys to collect information and design ideas in the context of urban design?”). The experiment was designed to evaluate the chatbot’s performance as a design requirement collection tool and compare its performance to a web-based survey.

As mentioned, we outlined a hypothetical urban architecture project as a test case, which involved constructing a new architecture school in place of a historic building (in Munich). The project’s context, situated within an existing neighborhood and incorporating public spaces, added complexity and relevance to the experiment. We chose this scenario to present a realistic and multifaceted design challenge, requiring participants to consider various factors and constraints in a familiar setting. The nature of this design project was both demanding and engaging for our participants, enabling us to assess the chatbot’s performance more effectively.

The new school building had to offer the most suitable environment for studying and working, enrich the university campus with the best quality and sustainable architecture in mind, serve around 1,400 architecture students, and provide them with various spaces such as studios, lecture halls, and so forth. These constraints were based on the current architecture school needs. By providing a realistic design problem, we aimed to create an environment where participants would be more likely to engage with the experiment.

The study participants were divided into the experimental and control groups (see the “Participants” section for further detail). The experimental group had conversations with the chatbot. The control group completed a web-based survey derived from a realistic urban public participation project survey. The adapted survey was built using Google Forms and included the following questions:

- What are your hopes for this project?
- What are your concerns about this project?

- How can the building be more sustainable and have a positive effect on the environment?
- Which values should the new building design express?
- How can the building have a positive impact on the community and society?

The same questions were also provided in the chatbot’s internal prompt so that it could discuss these questions with the study participants.

## Participants

We recruited students and faculty from the architecture department. All participants were stakeholders of our hypothetical project, possessed learning or teaching experiences, and had a good knowledge of the existing facility. The experiment group consisted of 35 participants; the control group had 21 participants. No compensation was offered.

In the experimental group, there were 12 participants aged 18–24 and 13 participants aged 25–34. Unfortunately, the remaining participants did not complete the user experience survey, so we lack information about their age, gender, and education. Concerning gender, 10 participants were male, 13 were female, and two indicated “other” as gender. As concerns educational attainment, eight of the study participants in the experimental group had a master’s degree, while 14 had a bachelor’s degree, and three were bachelor’s students at the time when the study was conducted.

In the control group, which consisted of 21 participants, all except one were students, with the remaining participant being a faculty member. Regarding age, eight respondents were 18–24, while 13 were 25–35 years old. Furthermore, most of the participants ( $N = 12$ ) were male, while eight were female, and one indicated “other” as gender. Finally, regarding educational attainment, six respondents were bachelor’s students, 12 had a bachelor’s degree, and three held a master’s degree.

## Analyzing and evaluating chatbot responses

We evaluated the chatbot performance by manually analyzing the conversation transcripts in terms of logic, behavior, and accuracy of the produced requirement list, inspired by Han *et al.* (2021). This analysis encompassed several stages: response quality, user experience, and ethical considerations. The quality of responses was assessed through a detailed text analysis, which will be elaborated upon later. User experience was measured using survey data, while ethical considerations were evaluated by identifying any potentially problematic chatbot messages containing hate speech or privacy-intrusive questions.

First, we checked whether the chatbot messages were coherent and relevant in the context of the conversation. A chatbot that does not produce logically sound responses is more likely to be perceived as unreliable by users. To this end, we differentiated between the following three possible levels: “the message makes perfect sense” (level 2), “the message makes sense but is slightly wrong” (level 1), and “the message is a mistake” (level 0).

To illustrate, consider the following example of a level 0 response in a discussion where the participant specifically asked to design a fire station with a parking space for three fire trucks:

HUMAN: “Okay, but I asked for space for just 3 trucks”

CHATBOT: “But you said 6 trucks”

The next example shows a level 1 response when the chatbot recommended cladding the walls with basalt while the participant wanted a white bedroom:

HUMAN: “*but basalt is black and I like white, maybe is not such a good idea?*”

CHATBOT: “*I recommend you to paint the basalt white.*”

Finally, consider an example of a level 2 response, which is a response that makes perfect sense:

HUMAN: “*What about taking marble or other material that originally is white?*”

Chatbot: “*Marble is a great idea.*”

Next, in our analysis of conversation dynamics, we sought to identify when the chatbot discussed a specific topic and when it changed it. In doing so, we aimed to identify when these events happened and whether they were in the proper context of the conversation. Accordingly, we manually tagged the chatbot responses with labels “changing subject,” “reflecting,” and “suggesting ideas” when responses made sense.

To illustrate, a reflective conversation was assumed to occur when the chatbot asked a relevant question in response to a user’s idea, which could cause the user to reflect on the idea and develop it. For example:

HUMAN: “*Hey, I would like a building with community space*”

CHATBOT: “*Excellent, we were thinking the same thing. What kind of community space?*”

When the chatbot continued asking repetitive questions or fixated on a topic instead of changing the conversation subject, we marked such responses as “repetition.” For example:

CHATBOT: “*What about the other rooms?*”

HUMAN: “*There are no other rooms*”

CHATBOT: “*How would you like to use the rooms?*”

HUMAN: “*There are no other rooms*”

Finally, we tallied and assessed the design requirements generated by the chatbot and those from the web-based survey. To evaluate the accuracy of the design briefs summarizing the design requirements created by the chatbot, we first matched each conversation with its corresponding design brief. We then labeled each accurately represented design brief item as “correct.” Conversely, items that failed to express a design requirement were labeled as “incorrect” or “invalid” if they were technically flawed. Lastly, we pinpointed any absent design brief items and marked them as “missing.”

### Comparison between chatbot and survey

The comparison between two different methods – chatbots and web-based surveys – was challenging since these methods produce different kinds of information. The following three metrics were used to analyze the performance of both methods.

The first metric was the number of words in an interaction, which was taken to attest to the quantity of data produced. The second metric was the number of themes the participant mentioned, which was done by manually analyzing the text and identifying the themes. The third metric was novel ideas. Innovative and novel ideas hold a significant value that can elevate the design process, allowing designers to explore new perspectives and approaches.

Novel ideas were identified by comparing the participant’s suggestions to a list of existing concepts. These ideas were mentioned in the provided project introduction document to the participants or popular themes, for example, producing a suitable environment for studying, quality of architecture, sustainable construction, availability of studios, lecture halls, facilitating idea exchange or creativity, shared spaces, and accessibility issues. To these themes, we added a list of themes that were most frequently

mentioned in similar projects or academic discourse but not in the project introduction document. Such themes included having a public space, facilitating connections to the community, generic sustainability ideas (e.g., recycling, solar power, water preserving, wood construction, green facades or roofs), simplicity, efficiency, cafe, coworking spaces, colors, and ordinary materials.

The statistical analysis of the data was conducted using the jStat statistical library, implementing Welch’s t-test.

Once the experimental and control groups completed the experiment, they filled out a user experience survey that contained 10 questions about their recent experience. The questions included in the study were taken from Ashfaq et al.’s (2020) subset relevant to the chatbot experiment and web surveys (see Table A1 in the appendix section “Survey questions”). The items in the survey were rated on a five-point Likert scale. For further analysis, an average score was computed for each survey item. Finally, the participants were asked to share their qualitative opinions about using the chatbot or survey.

## Results

In the experiment, 751 messages were collected in 35 conversations. We provide example chatbot and survey output in the appendix sections “Chatbot conversation example” and “Survey response example.” The participants produced 377 messages, and the chatbot produced 374, with an average of 21.45 messages per conversation. The average interaction duration was 9.75 minutes (min = 2.25, max = 24.74, SD 6.1). The chatbot generated 145 design requirements from 26 conversations since some participants did not click on the “finish chat” button or did not reach 50 messages. Twenty-five participants filled out the user experience survey. We decided to keep the conversation transcripts of the participants who did not complete the user experience survey since they are valuable to the analysis and might contain failed conversations. However, the analysis did not reveal chatbot conversation failures.

### Conversation quality

The summary of our conversation analysis results, including the chatbot and human participants’ evaluation of the preliminary and controlled experiment, is provided in Table 1. Figure 3 shows the chatbot’s human word count per message distribution compared to the survey method. The chatbot produced responses of an average length of 24.56 words (SD 22.11), with the longest response being 134 words long. By contrast, the participants’ responses to the chatbot produced significantly fewer words in each message, with an average length of 10.51 words (SD 15.42). This result demonstrates a clear LLM performance difference from our preliminary experiments, where the chatbot produced shorter responses ( $M = 10.83$ ,  $SD 8.41$ ), which caused the participants to respond with shorter messages ( $M = 4.62$ ,  $SD 4.68$ ).

With regard to the qualitative aspects of the analyzed conversations, of a total of 374 messages, 297 were marked as most comprehensive (level 2, 86.84%), 34 were marked as lightly flawed (level 1, 9.94%), and only 11 did not make sense (level 0, 3.22%) (see Table 2).

This improvement between the preliminary and the controlled experiments can be attributed to the enhancement in the LLM and prompt, which highlights the crucial role of coherent and logical responses in fostering effective interaction. However, there is still a need for improvement in the LLM’s performance.

**Table 1.** Summary of human-provided information in chatbot and survey

Item	Chatbot mean (SD)	Web survey mean (SD)
Word count	117.91 (75.09)	59.04 (35.78)
Words in message	10.51 (15.42)	11.87 (9.72)
Mean messages per user	11.00 (7.69)	5 (0 – The number of survey questions was fixed)
Topics and ideas per interaction	4.82 (2.49)	6.90 (1.86)
Novel idea rate	31% (0.60)	36% (0.58)

In their responses to the chatbot, the participants produced an average of 117.91 words in each conversation (SD 75.09), with a maximum of 350 words. The control group respondents who filled out the survey produced considerably fewer words per response – an average of 59.04 words for each survey record, with a maximum of 175 words (SD 35.78). The results show that the chatbot elicited twice as much data from the participants as the survey.

### Behavior analysis

The subsequent behavior analysis indicated that the chatbot changed the conversation topic by asking a different question 41.44% of the time (see Table 3). However, in 5.26% of the messages, the chatbot demonstrated a reflecting behavior and asked the participants about their ideas. In 5.14% of the messages, the chatbot repeated the previous message, albeit with some (minor) changes. Finally, the chatbot did not produce any messages that could be considered hate speech, nor did it commit privacy intrusion.

### Themes and novel ideas

The study found no significant difference in the topics generated through the chatbot or survey,  $t(38) = 0.49$ ,  $p = .062$ , although the

survey ( $M = 6.90$ ,  $SD = 1.86$ ) produced more topics on average than the chatbot ( $M = 4.82$ ,  $SD = 2.49$ ). Similarly, there was no significant difference in generating novel ideas between the two methods ( $t(38) = 0.36$ ,  $p = .071$ ). The difference was not statistically significant despite surveys ( $M = 0.43$ ,  $SD = 0.62$ ) yielding a little more novel ideas on average than chatbots ( $M = 0.36$ ,  $SD = 0.70$ ). These findings suggest that the choice of method, whether chatbot or survey, does not significantly impact the number of topics or novel ideas generated.

Several survey response examples are provided below:

- “Modern, sustainable architecture”
- “Opening up to the public and creating spaces for idea exchange and collaboration”
- “Space for experimentation, generous connection to public space, space for experimentation”

Below is an example of a chatbot conversation:

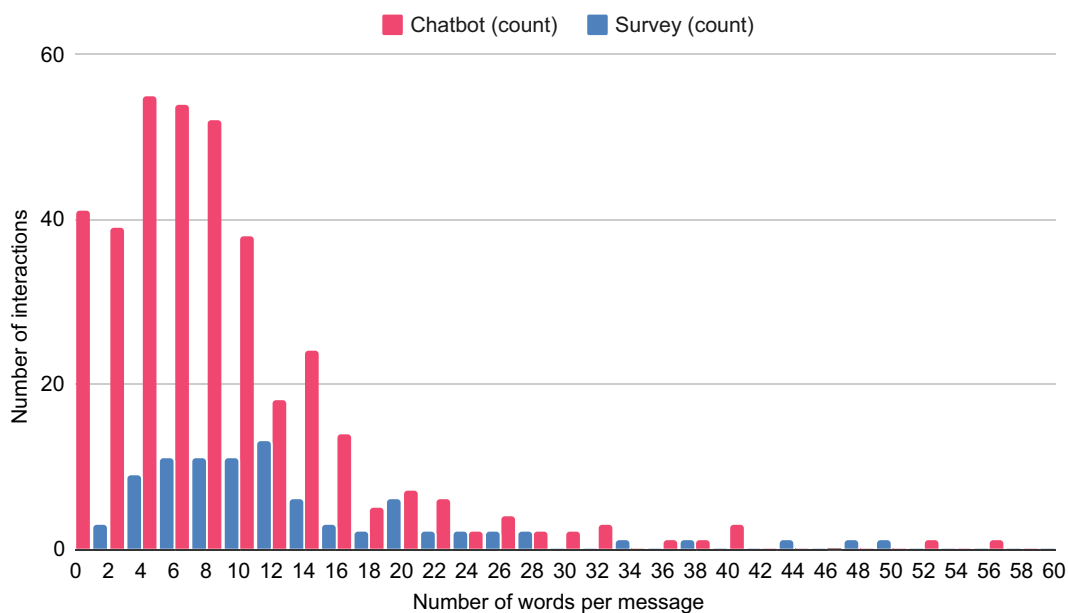
HUMAN: “well, I think that the spaces should be focused on interaction and collaboration... so the biggest amount of exchange can take place”

CHATBOT: “Excellent, so you want the spaces in the building to facilitate interaction and collaboration. Do you have any ideas about how we can achieve that?”

HUMAN: “well it could be achieved through different methods. One, for example, could be the relationship between public and private spaces - leaning more to the public side. Another method could be through the typology of the floorplan and the construction methods implemented - meaning to have a high degree of “permeability” in the architectural concept”

CHATBOT: “That makes sense. Are there any other values that you feel are important for the building to express?”

The findings can be attributed to the fact that participants can contemplate and offer well-considered responses in the form of a list during a survey. Furthermore, the interactive nature of a chatbot may result in more impulsive responses, potentially generating fewer ideas. Additionally, as demonstrated in the above example, chatbot conversations tend to be lengthier and demand more effort.



**Figure 3.** Distribution of the number of human-provided words per message, comparing chatbot and survey responses with a bucket size of two words. Both mediums are similarly distributed, peaking at 4–12 words per message. Notably, the chatbot generated a significantly higher number compared to the surveys.



**Table 2.** Comparative analysis of chatbot response quality evaluation between preliminary experiments and a controlled experiment. It shows that the enhanced GPT-3 model, coupled with refined text prompts, improved the quality of the generated text

Quality level	Preliminary experiment	Controlled experiment
Level 2	83.20%	86.84%
Level 1	11.60%	9.94%
Level 0	5.19%	3.22%

**Table 3.** Summary of chatbot behavioral evaluation

Behavior	Preliminary experiment	Controlled experiment
Changing subject	54.70%	41.44%
Reflecting	0%	5.26%
Suggesting ideas	1.33%	4.39%
Repetition	3.87%	5.14%

**Table 4.** Summary of design brief analysis success

Criteria	Preliminary experiment	Controlled experiment
Correctness of the produced requirement items	75.0%	92.85%
Rate of missing requirements	15.95%	17.6%

Nevertheless, chatbot discussions yielded many ideas, exhibiting a comparable novelty rate, which is promising as chatbot technology improves.

### Design brief analysis

The chatbot generated 26 valid design brief lists with 140 list items. The design briefs contained an average of 5.38 list items per discussion (SD = 2.49, min = 3, max = 15). No relationship was observed between the length of conversations and the number of design brief items. Accordingly, a more extended discussion did not translate into more ideas and topics for the subsequent generation of design requirements.

Next, to measure the system's performance, we manually analyzed conversation transcripts and corresponding design briefs (see Table 4). As a result, in addition to 140 requirements captured by the chatbot, we identified 30 further requirements and preferences that were overlooked by the chatbot. Furthermore, out of 140 list items, 130 (92.85%) were correct, 10 were incorrect (7.15%), and 43 were invalid (12.2%). Most invalid items were texts that were generated by the chatbot or fragmented sentences.

### User experience

As mentioned earlier, the survey was conducted in both the experimental and control groups to evaluate service quality, enjoyment, usefulness, and ease of use of the chatbot or the survey.

The results of the comparative analysis between the chatbot and survey methods demonstrated that participants rated the user

experience of the chatbot more positively than the survey (see Fig. 4). Participants reported a higher level of enjoyment when interacting with the chatbot than when completing the survey. They also perceived the chatbot's service quality to be superior. Furthermore, the chatbot was deemed considerably easier to use, and satisfaction ratings were higher for the chatbot experience. Interestingly, the perceived usefulness and continuance intention were similar for both methods.

In conclusion, these findings suggest that the chatbot provides a more engaging and satisfying user experience than traditional surveys, and may encourage increased digital public participation.

### User feedback

According to the results of the user feedback survey, most participants found the chatbot system to be a valuable tool for data collection in the early stages of the project. The participants appreciated the chatbot's conversational nature, which made it easier for them to understand and express their opinions.

However, two participants mentioned that they preferred merely writing down their ideas instead of having a longer conversation. Furthermore, some other participants noted that the chatbot lacked the conversational qualities of real-life interaction. They felt that the chatbot was too quick to agree or thank them for their contribution without providing a meaningful response.

One participant admitted having a negative bias toward chatbot systems, mentioning that a real person could still do the job better. However, they acknowledged that chatbot systems could be helpful in certain situations, such as when architects do not have time to discuss their ideas with stakeholders.

### Discussion

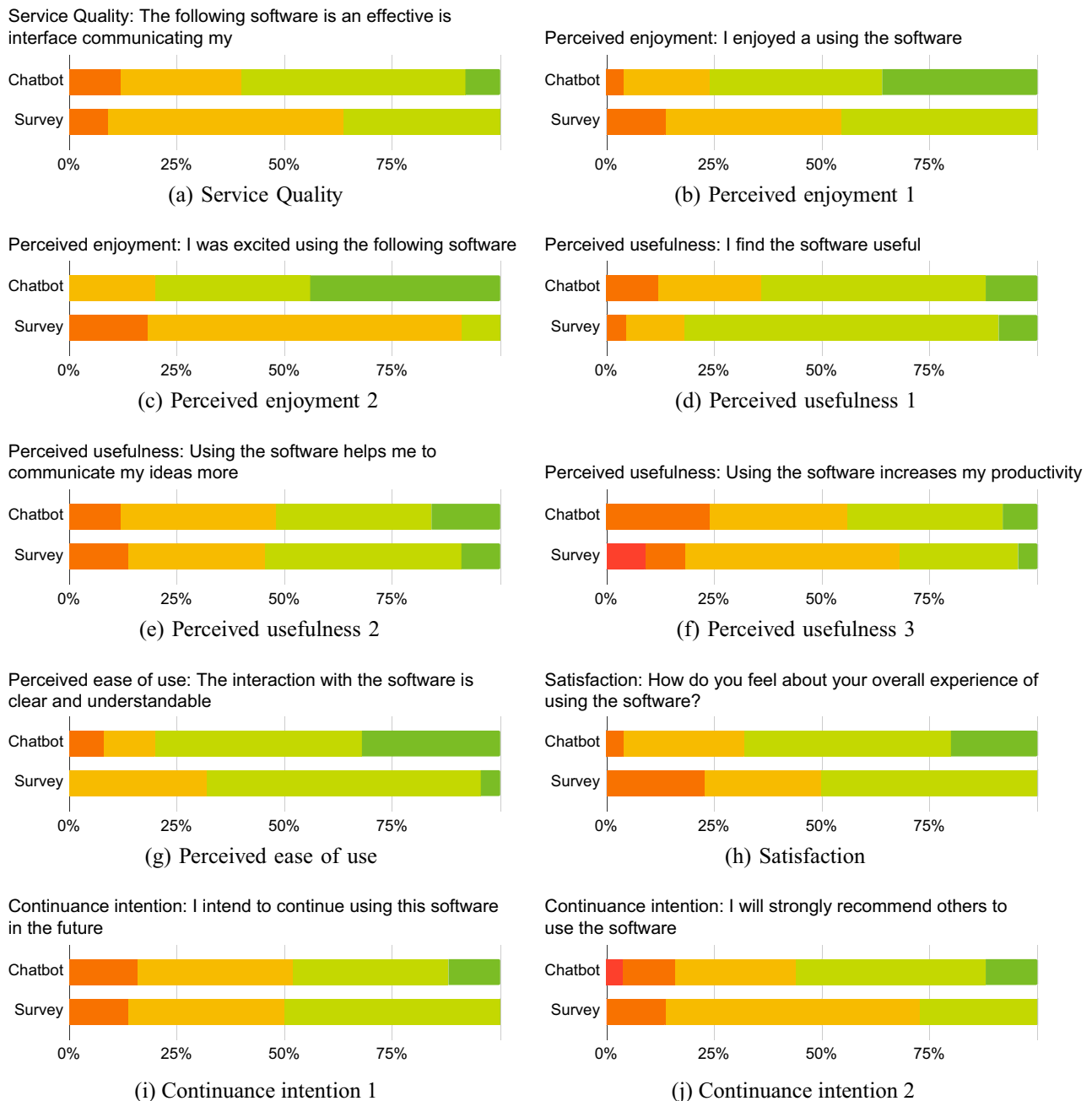
The chatbot tool demonstrated its ability to handle extensive conversations, allowing for significant and focused human-like discussions while gathering new types of information. In the study, participants' input during their interactions with the chatbot was automatically transformed into a valuable list of requirements. This innovative method can enhance participation in large-scale urban design projects.

However, the findings of this study indicate that although chatbot technology can generate meaningful conversations, there are still numerous challenges to address in order to guarantee its successful implementation in such urban design projects.

### Human–chatbot prompt design

In chatbot prompt design, we face a unique challenge that stems not only from the unpredictable nature of LLMs but also from the variability of human behavior. Previous chatbot research has compared the unpredictability of LLMs to "herding AI cats" due to the tendency of LLMs to generate unexpected responses (Zamfirescu-Pereira et al., 2023). This metaphor is particularly pertinent when designing an assistant chatbot. However, our chatbot's primary aim is significantly different – to extract specific information from human users.

To achieve this goal, designers must also consider the significant variability in human behavior. For instance, our data show a wide range of responses to the same prompt that resulted in different kinds of conversations and a varying number of design requirements, demonstrating high variability in human behavior. This



**Figure 4.** Chatbot and survey user experience evaluation result comparison in terms of service quality, perceived enjoyment, perceived usefulness, perceived ease of use, satisfaction, and continuance intention.

challenge is further compounded by the chat conversation format's inherently unstructured and open-ended nature.

Therefore, it is not just about "herding AI cats" but also about managing "human cats." This metaphor underscores the need for designers to account for human behavior's unpredictability and LLMs' inherent unpredictability. In essence, designers of participatory chatbot systems, which involve users in the design process, must be prepared to navigate the dual unpredictability of human interaction and LLMs.

To address this complexity, we propose a prompt design framework for information collection comprising two key components: an internal and shared prompt. The internal prompt encapsulates

project-specific information and a personality modifier to influence the chatbot's responses. The shared prompt, on the other hand, plays a crucial role in shaping the human-AI interaction by providing a clear prompt for the human user. Both prompts must be fine-tuned through iterative testing to produce the expected interaction between humans and AI. While testing LLM prompts can be done automatically, human subject experiments are much more complex but can be automated using crowdsourcing platforms.

In conclusion, our proposed framework provides an essential foundation for designing effective human and chatbot prompts, addressing the challenges posed by the unpredictable nature of both. Our chatbot design framework distinguishes itself from

previous studies by extending the prompt design to the human, which is essential for fostering clear communication and effectively framing the conversation (Wei et al., 2023). However, the design of chatbot prompts is still very challenging. It necessitates improved tools and methodologies, as they demand extensive trials and testing. Therefore, future research should focus on developing more sophisticated tools and methodologies for prompt chatbot design, considering the unpredictability of human input and AI responses.

### *Enhancing stakeholder engagement through enjoyable participation methods*

In accordance with prior research, our study demonstrated that the participants perceived the chatbot interaction as more enjoyable than completing a traditional survey (Kim et al., 2019; Xiao et al., 2020b). Previous research emphasizes the significance of stakeholder involvement in design projects to achieve successful outcomes (Arnstein, 1969; Münster et al., 2017; Calderon, 2020). However, urban design and planning initiatives often struggle with low participation rates due to their professional and political nature (Brabham, 2009; Giering, 2011; Krüger et al., 2019; Dortheimer and Margalit, 2020). This creates difficulties in meaningfully engaging underrepresented communities in such projects.

To address this challenge, it is essential to comprehend the motivations behind participation. Researchers agree that enjoyment is a crucial factor in driving participation (Lindenberg, 2001; Malone et al., 2010). A robust connection has been discovered between enjoyment and increased engagement in a crowdsourcing activity. Individuals who partake in enjoyable crowdsourcing activities are more likely to invest time and effort, resulting in more contributions and greater satisfaction with the final product (Frey et al., 2011; Liang et al., 2018).

Our study results indicate that chatbots are perceived as more enjoyable than traditional surveys, which can enhance participation and foster greater engagement with diverse communities. Practitioners should consider this finding when planning large-scale participatory urban projects, as incorporating chatbots may lead to improved involvement and more successful outcomes.

However, it is crucial to consider the potential novelty effect of the chatbot. Participants may have enjoyed their first interaction with the LLM-based chatbot due to its novelty, which could have influenced their favorable ratings. Over time, as the novelty wears off, users may experience “chatbot fatigue,” similar to “survey fatigue.” This can impact the long-term effectiveness and user satisfaction of chatbots. More research is needed to understand this potential effect.

We propose several strategies to maximize the potential of chatbots in stakeholder involvement in spatial design to make them more enjoyable. Firstly, the chatbot persona should be designed to be engaging and enjoyable, incorporating elements of humor, empathy, and a conversational style in the internal prompt. Secondly, chat interfaces should be made more accessible by integrating chatbots into popular instant messaging platforms, reducing the learning curve for users. Lastly, it is crucial to continually monitor and improve the chatbot based on user feedback and performance analysis.

### *Comparing information quality and quantity in chatbot conversations and surveys*

In examining the quality and quantity of information generated, we discovered that chatbot conversations yielded more data than surveys despite covering fewer topics. However, the rate of novel ideas was similar. While the differences observed were not

statistically significant, preventing us from definitively stating one method as superior, our findings suggest that chatbots can generate design requirements and ideas similar to those obtained from surveys. We recommend that future research delve deeper into this comparison, utilizing larger datasets from chatbot conversations and surveys to substantiate these preliminary findings further.

Various factors may explain the differences between surveys and chatbot conversations concerning information quality and quantity. Firstly, our chatbot conversations were less structured than a survey, which may have led to answering fewer questions. Secondly, surveys offer a clear outline of the required input through a series of pages and questions, whereas chatbot conversations may lack clarity regarding completing information collection, allowing users to end the conversation when they feel it is over. Lastly, the casual, conversational nature of chatbot interactions may result in fewer topics being discussed, as they demand more effort from users and need to be longer to cover all questions.

This limitation of current chatbot technology should be considered when using chatbots for ideation or design requirement collection and should inform future research. Therefore, to ensure that sufficient issues and novel ideas are generated, we suggest using chatbots with a substantial participation group. With small groups, the current chatbot should not be seen as a replacement for surveys but rather as an additional input method. Notwithstanding, according to the results of the present study, chatbots can be a valuable tool for engaging stakeholders as they are more enjoyable and, thus, are more likely to be used by a broader pool of stakeholders.

### *Future research*

Future research should focus on the experimental testing of chatbots in real-world urban design settings to identify new challenges. In addition, further research on improving the proposed chatbot in terms of conversation structure and summarization algorithms would also be needed. In particular, future studies could explore how chatbots can communicate design using visual communication. Furthermore, considering our findings on both the strengths and limitations of using chatbots in design projects, it would be meaningful to examine the combined use of chatbots and surveys to leverage the advantages of both approaches.

### *Limitations*

The present study has several limitations. First, the study participants were architecture students and faculty members proficient in verbally expressing design ideas, understanding what buildings require, and having outstanding novel ideas. This could have led to the creation of more topics and a higher topic and novelty rate than lay people.

The second limitation is that the participants were aware that the project used in the present study was not real, and thus, their ideas would not be realized. This could have caused some participants to have a lighter conversation with the chatbot, knowing there would be no ramifications. Furthermore, it may have led people to also report less in their survey responses.

The third limitation concerns certain obscurities in utilizing the chatbot’s design requirement generation process. Upon the conversation’s conclusion, participants were required to activate a “finish chat” button to compile and generate a list of design requirements. However, some participants failed to execute this step. Upon analysis of the conversations, no causative factors related to the chatbot’s performance could be identified, leading

us to conclude that the issue lies within the user interface. To address this limitation in future chatbot research, we propose automatically generating design requirements once the conversation ends, eliminating the need for user-initiated activation.

## Conclusion

Chatbots have the potential to transform stakeholder engagement in urban design projects. By offering a more engaging and interactive experience than traditional surveys, they can potentially help urban designers connect with more extensive and diverse communities. However, the findings of this study do not show a significant difference between chatbots and surveys in generating topics and new ideas. Despite this limitation, the study demonstrates that chatbots can successfully automate design conversations in architecture and urban design. Participants found the chatbot experience enjoyable and stimulating, which could lead to increased public involvement in participatory design processes.

This research suggests a chatbot system and prompt framework that can be utilized in large-scale participatory design projects, streamlining data collection and analysis. The system enables automated conversations while providing a summarization mechanism to help designers manage the vast amounts of data generated. In conclusion, our findings support the effective use of chatbots in facilitating design conversations, highlighting necessary further research to enhance the data collection capabilities of chatbots, making them even more beneficial for design processes.

**Data availability statement.** The data that support the findings will be available.

**Funding statement.** This work was supported by the Technical University of Munich Global Incentive Fund. The funder had no role in study design, data collection and analysis, publication decision, or manuscript preparation.

**Competing interest.** The authors declare no competing interests.

## References

- Abdul-Kader SA and Woods J (2015) Survey on chatbot design techniques in speech conversation systems. *International Journal of Advanced Computer Science and Applications* 6(7), 72–80.
- Ahmed S (2019) An Architecture for Dynamic Conversational Agents for Citizen Participation and Ideation. Master's thesis, Technische Universität München.
- Alexander C (1964) *Notes on the Synthesis of Form*. Harvard University Press, Cambridge.
- Alexander C, Ishikawa S and Silverstein M (1977) *A Pattern Language: Towns, Buildings, Construction*. New York: Oxford University Press.
- Arias E, Eden H, Fischer G, Gorman A and Scharff E (2000) Transcending the individual human mind – creating shared understanding through collaborative design. *ACM Transactions on Computer-Human Interaction* 7(1), 84–113.
- Arnstein SR (1969) A ladder of citizen participation. *Journal of the American Institute of Planners* 35(4), 216–224.
- Ashfaq M, Yun J, Yu S and Loureiro SMC (2020) I, chatbot: modeling the determinants of users' satisfaction and continuance intention of AI-powered service agents. *Telematics and Informatics* 54(July), 101473.
- Bittner E and Shoury O (2019) Designing automated facilitation for design thinking: a chatbot for supporting teams in the empathy map method. <https://doi.org/10.24251/HICSS.2019.029>
- Borges J, Jankowski P and Davis CA (2015) Crowdsourcing for geodesign: opportunities and challenges for stakeholder input in urban planning. In Sluter CR, Cruz CBM and de Menezes PML (eds), *Cartography – Maps Connecting the World*. Lecture Notes in Geoinformation and Cartography. Cham: Springer, 361–373.
- Brabham DC (2009) Crowdsourcing the public participation process for planning projects. *Planning Theory* 8(3), 242–262.
- Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I and Amodei D (2020) Language models are few-shot learners. Preprint, arXiv:2005.14165.
- Calderon C (2020) Unearthing the political: differences, conflicts and power in participatory urban design. *Journal of Urban Design* 25(1), 50–64.
- Cuadra A, Goeddicke D and Zamfirescu-Pereira JD (2021) Democratizing design and fabrication using speech. In *CUI 2021 – 3rd Conference on Conversational User Interfaces*. New York: Association for Computing Machinery, 1–8.
- Dortheimer J (2022) Collective intelligence in design crowdsourcing. *Mathematics* 10(4), 539.
- Dortheimer J and Margalit T (2020) Open-source architecture and questions of intellectual property, tacit knowledge, and liability. *The Journal of Architecture* 25(3), 276–294.
- Dortheimer J, Neuman E and Milo T (2020, September) A novel crowdsourcing-based approach for collaborative architectural design. In *Anthropologic: Architecture and Fabrication in the Cognitive Age—Proceedings of the 38th eCAADe Conference (Vol. 2, pp. 155–164)*. Berlin, Germany: Education and Research in Computer Aided Architectural Design in Europe.
- Dortheimer J, Yang S, Yang Q and Sprecher A (2023) Conceptual architectural design at scale: a case study of community participation using crowdsourcing. *Buildings* 13(1), 222.
- Dortheimer J, Schubert G, Dalach A, Brenner L and Martelaro N (2023b) Think AI-side the Box!. eCAADe 41 - Digital Design Reconsidered, Graz, Austria, 567–576
- Dubberly H and Pangaro P (2019) Cybernetics and design: conversations for action. In *Design Cybernetics*. Cham: Springer, 85–99.
- Floridi L and Chiriatti M (2020) GPT-3: its nature, scope, limits, and consequences. *Minds and Machines* 30(4), 681–694.
- Frey K, Lüthje C and Haag S (2011) Whom should firms attract to open innovation platforms? The role of knowledge diversity and motivation. *Long Range Planning* 44(5–6), 397–420.
- Giering Scott (2011) Public participation strategies for transit. Vol. 89. Transportation Research Board.
- Gooch D, Barker M, Hudson L, Kelly R, Kortuem G, Van Der Linden J, Petre M, Brown R, Klis-Davies A, Forbes H, Mackinnon J, Macpherson R and Walton C (2018) Amplifying quiet voices: challenges and opportunities for participatory design at an urban scale. *ACM Transactions on Computer-Human Interaction* 25(1), 1–34.
- Han X, Zhou M, Turner MJ and Yeh T (2021) Designing effective interview chatbots: automatic chatbot profiling and design suggestion generation for chatbot debugging. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. New York: Association for Computing Machinery, 1–15.
- Harvey D (1973) *Social Justice and the City*, revised edn. Athens, GA: University of Georgia Press.
- Hill J, Ford WR and Farreras IG (2015) Real conversations with artificial intelligence: a comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior* 49, 245–250.
- Hofmann M, Münster S and Noennig JR (2020) A theoretical framework for the evaluation of massive digital participation systems in urban planning. *Journal of Geovisualization and Spatial Analysis* 4(1), 3.
- Hosio S, Goncalves J, Kostakos V and Riekkilä J (2015) Crowdsourcing public opinion using urban pervasive technologies: lessons from real-life experiments in Oulu. *Policy & Internet* 7(2), 203–222.
- Hussain S, Sianaki OA and Ababneh N (2019) A survey on conversational agents/chatbots classification and design techniques. In Barolli L, Takizawa M, Xhafa F and Enokido T (eds), *Web, Artificial Intelligence and Network Applications*. Cham: Springer, 946–956.
- Hwang AH-C and Won AS (2021) IdeaBot: investigating social facilitation in human–machine team creativity. In *Proceedings of the 2021 CHI Conference*

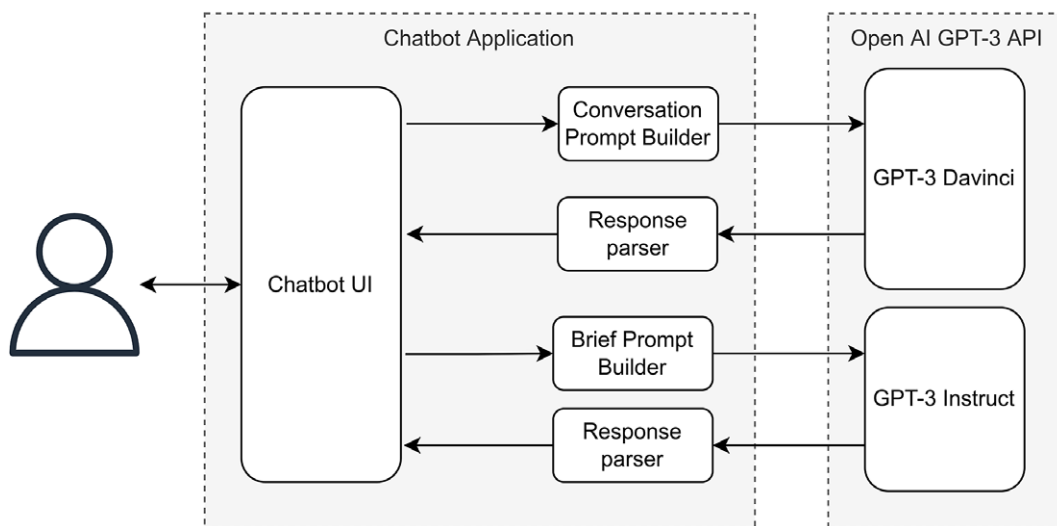
- on *Human Factors in Computing Systems*. New York: Association for Computing Machinery, 1–16.
- Jacobs J** (1961) *Death and Life of Great American Cities*. New York: Vintage Books.
- Karlgren K and Ramberg R** (2012) The use of design patterns in overcoming misunderstandings in collaborative interaction design. *CoDesign* 8(4), 231–246.
- Kim S, Eun J, Seering J and Lee J** (2021) Moderator chatbot for deliberative discussion: effects of discussion structure and discussant facilitation. *Proceedings of the ACM on Human-Computer Interaction* 5, 1–26.
- Kim S, Lee J and Gweon G** (2019) Comparing data from chatbot and web surveys: effects of platform and conversational style on survey response quality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. New York: Association for Computing Machinery, 1–12.
- Krüger M, Duarte AB, Weibert A, Aal K, Talhouk R and Metatla O** (2019) What is participation?: Emerging challenges for participatory design in globalized conditions. *Interactions* 26(3), 50–54.
- Kulcke M** (2018) Design-bot – using half-automated qualitative interviews as part of self-communication within the design process. In Kepczynska-Walczak A and Bialkowski S (eds), *Computing for a Better Tomorrow – Proceedings of the 36th eCAADe Conference*, Vol. 1, Lodz University of Technology, Lodz, Poland, September 19–21, 2018, 103–108.
- Lawson B and Loke SM** (1997) Computers, words and pictures. *Design Studies* 18(2), 171–183.
- Lee Y-C, Yamashita N and Huang Y** (2020) Designing a chatbot as a mediator for promoting deep self-disclosure to a real mental health professional. *Proceedings of the ACM on Human-Computer Interaction* 4, 1–27.
- Liang H, Wang M-M, Wang J-J and Xue Y** (2018) How intrinsic motivation and extrinsic incentives affect task effort in crowdsourcing contests: a mediated moderation model. *Computers in Human Behavior* 81, 168–176.
- Lindenberg S** (2001) Intrinsic motivation in a new light. *Kyklos* 54(2–3), 317–342.
- Lu H, Gu J, Li J, Lu Y, Müller J, Wei W and Schmitt G** (2018) Evaluating urban design ideas from citizens from crowdsourcing and participatory design. In Fukuda T, Huang W, Janssen P, Crolla K and Alhadidi S (eds), *CAADRIA 2018 – 23rd International Conference on Computer-Aided Architectural Design Research in Asia: Learning, Prototyping and Adapting*, Vol. 2. Hong Kong: Association for Computer-Aided Architectural Design Research in Asia, 297–306.
- Luck R** (2003) Dialogue in participatory design. *Design Studies* 24(6), 523–535.
- Luck R** (2018) Participatory design in architectural practice: changing practices in future making in uncertain times. *Design Studies* 59, 139–157.
- Luck R and McDonnell J** (2006) Architect and user interaction: the spoken representation of form and functional meaning in early design conversations. *Design Studies* 27(2), 141–166.
- Malone TW, Laubacher R and Dellarocas C** (2010) The collective intelligence genome. *IEEE Engineering Management Review* 38(3), 38–52.
- Martelaro N, Mennicken S, Thom J, Cramer H and Ju W** (2020) Using remote controlled speech agents to explore music experience in context. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. Eindhoven: Association for Computing Machinery, 2065–2076.
- McDonnell J** (2009) Collaborative negotiation in design: a study of design conversations between architect and building users. *CoDesign* 5(1), 35–50.
- Mueller J, Lu H, Chirkin A, Klein B and Schmitt G** (2018) Citizen design science: a strategy for crowd-creative urban design. *Cities* 72, 181–188.
- Münster S, Georgi C, Heijne K, Klamert K, Noennig JR, Pump M, Stelzle B and van der Meer H** (2017) How to involve inhabitants in urban design planning by using digital tools? An overview on a state of the art, key challenges and promising approaches. *Procedia Computer Science* 112, 2391–2405 (Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 21st International Conference, KES-2017-8 September 2017, Marseille, France).
- Nguyen TH, Waizenegger L and Techatassanasoontorn AA** (2022) “Don’t neglect the user!” – identifying types of human–chatbot interactions and their associated characteristics. *Information Systems Frontiers* 24(3), 797–838.
- Oak A** (2009) Performing architecture: talking “architect” and “client” into being. *CoDesign* 5(1), 51–63.
- OpenAI** (2023) ChatGPT: optimizing language models for dialogue. Available at <https://openai.com/blog/chatgpt/> (accessed 2 February 2023).
- Radford A, Wu J, Child R, Luan D, Amodei D and Sutskever I** (2019) Language models are unsupervised multitask learners. *OpenAI Blog* 1(8), 9.
- Reich Y, Konda SL, Monarch IA, Levy SN and Subrahmanian E** (1996) Varieties and issues of participation and design. *Design Studies* 17(2), 165–180.
- Robertson T and Simonsen J** (2012) Challenges and opportunities in contemporary participatory design. *Design Issues* 28(3), 3–9.
- Schmitt O and Buschek D** (2021) CharacterChat: supporting the creation of fictional characters through conversation and progressive manifestation with a chatbot. In *Creativity and Cognition*. New York: Association for Computing Machinery, 1–10.
- Shawar BA and Atwell ES** (2005) Using corpora in machine-learning chatbot systems. *International Journal of Corpus Linguistics* 10(4), 489–516.
- Shin J, Hedderich MA, Lucero A and Oulasvirta A** (2022) Chatbots facilitating consensus-building in asynchronous co-design. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. New York: Association for Computing Machinery, 1–13.
- Simonsen J and Robertson T** (2012) *Routledge international handbook of participatory design*. New York, NY, USA: Routledge.
- Tavanapour N and Bittner EAC** (2018) Automated facilitation for IDEA platforms: design and evaluation of a chatbot prototype. In Jan Pries-Heje, Sudha Ram and Michael Rosemann (eds), *Proceedings of the International Conference on Information Systems - Bridging the Internet of People, Data, and Things, ICIS 2018*. San Francisco, CA, USA: Association for Information Systems, December 13–16, 2018, <https://aisel.aisnet.org/icis2018/general/Presentations/8>.
- te Pas ME, Werner GMMR, Bouwman RA and Buise MP** (2020) User experience of a chatbot questionnaire versus a regular computer questionnaire. *Prospective Comparative Study* 8(12), e21982.
- Wang Y, Gao S, Li N and Yu S** (2021) Crowdsourcing the perceived urban built environment via social media: the case of underutilized land. *Advanced Engineering Informatics* 50, 101371.
- Wei J, Kim S, Jung H and Kim Y-H** (2023) Leveraging large language models to power chatbots for collecting user self-reported data. Preprint, [arXiv:2301.05843](https://arxiv.org/abs/2301.05843).
- Weizenbaum J** (1983) ELIZA – a computer program for the study of natural language communication between man and machine. *Communications of the ACM* 26(1), 23–28.
- White J, Fu Q, Hays S, Sandborn M, Olea C, Gilbert H, Elnashar A, Spencer-Smith J and Schmidt DC** (2023) A prompt pattern catalog to enhance prompt engineering with ChatGPT. Preprint, [arXiv:2302.11382](https://arxiv.org/abs/2302.11382).
- Xiao Z, Zhou MX, Chen W, Yang H and Chi C** (2020a) If I hear you correctly: building and evaluating interview chatbots with active listening skills. In *CHI Conference on Human Factors in Computing Systems (CHI '20)*. New York: Association for Computing Machinery, 1–14.
- Xiao Z, Zhou MX, Liao QV, Mark G, Chi C, Chen W and Yang H** (2020b) Tell me about yourself: using an AI-powered chatbot to conduct conversational surveys with open-ended questions. *ACM Transactions on Computer-Human Interaction* 27(3), 1–37.
- Zamfirescu-Pereira JD, Wei H, Xiao A, Gu K, Jung G, Lee MG, Hartmann B and Yang Q** (2023) Herding AI cats: lessons from designing a chatbot by prompting GPT-3. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference (DIS '23)*. New York: Association for Computing Machinery, 2206–2220.

## Appendix

### A Final chatbot implementation

The implementation of the chatbot system was executed using a suite of web technologies. The system's backbone was a custom-built NodeJS web server application designed to handle the processing and management of the chatbot's operations. The system also incorporated a MySQL relational database, which was utilized to store and manage all conversation data to ensure that all interactions were recorded and could be analyzed. The user interface was developed using a custom HTML and JavaScript application coded using ReactJS.

The system used web services for AI functionality. Upon receiving a user's message, the server compiled the message into a comprehensive prompt, which included the entire conversation. This prompt was sent to the GPT-3 LLM via the Open AI web service. The response from the LLM was subsequently returned to the application, where it was parsed and cleaned before being sent back to the user interface (see Fig. A1).



**Figure A1.** Chatbot application structure diagram and information flow between the user, application, and web services.

#### A.1 Internal prompt

The internal prompt was provided only to the language model and included the conversation context that the human should not be aware of. The straightforward parts of the prompt were the context of the dialog (e.g., a conversation, chat, phone call, or theater play), description of the chatbot's persona (e.g., architect, client, project manager, interior designer, or bot), the chatbot's character (e.g., helpful, critical, or rude), and the role of the user (e.g., client, stakeholder, or neighbor). These textual descriptions were essential to fine-tuning the chatbot behavior. The internal prompt we used for the controlled experiment was as follows:

This is a chat between an architect and her client. The architect asks many questions to get a better understanding of the design project. The architect is helpful and clever and goes into detail to help the client to better express his/her needs. The architect needs to know which rooms the client needs, who is going to use them, as well as the client's preferred style.

Chat transcript:

#### A.2 Shared prompt

The shared prompt included predefined messages that the chatbot sent to the user to provide conversation context. These messages were not generated by the

language model and were considered to be part of the prompt since they additionally provided the conversation context to the language model. The shared prompt described the context of our design project (i.e., initial requirements and location) and introduced the user and the chatbot (i.e., names, organizational belonging, etc.).

Then, the discussion transcript was provided, including identification of the speakers (e.g., architect, client). The prompt "Architect:" suggested that the language model would complete how an architect would continue the conversation. The example below shows how the shared prompt was used.

**ARCHITECT:** Hello { participant\_name }, my name is Zaha and I am on the design team of the project.

**ARCHITECT:** The (Technical University of Munich) is planning to build a new architecture school instead of the outdated (electrical facility on Theresienstrasse). The building will host the architecture school and the design will be based on the preferences of the students and faculty. That's why we want to ask you about your ideas for the new building.

**ARCHITECT:** Let's brainstorm a little bit. What is your opinion on the project, what useful spaces should be in the building? How can the building have a positive impact on the community and the environment? What should the building look like? What values should the building express?

#### A.3 Generating design requirements from a conversation

Once the discussion concluded, we used the language model to extract the list of design requirements. We used the "davinci-instruct-beta" and "text-curie-001" models, which we found suitable for the task. An example of the prompt is shown below.

Here is a transcript of a conversation between an architect and her client:

[ ... conversation transcript log... ]

—

Make a list summarizing all the detailed requirements from the discussion:

## B Survey questions

**Table A1.** A comparison of user experience with the chatbot versus the survey

Construct	Item
Service quality (SQ)	The following software has an effective interface for communicating my needs
Perceived enjoyment (PE)	I enjoyed using the software
	I was excited to use the following software
Perceived usefulness (PU)	I find the software useful
	Using the software helps me to communicate my ideas more quickly
	Using the software increases my productivity
Perceived ease of use (PEU)	The interaction with the software is clear and understandable
Satisfaction (SAT)	How do you feel about your overall experience of using the software?
Continuance intention (CI)	I intend to continue using this software in the future
	I will strongly recommend others to use the software

## C Chatbot conversation example

The following is a typical conversation log provided as an example to illustrate the collected data.

CHATBOT: "Hello (name removed for privacy), my name is Zaha and I am on the design team of the project."

CHATBOT: "The Technical University of Munich is planning to build a new architecture school instead of the outdated electrical facility on Theresienstrasse. The building will host the architecture school and the design will be based on the preferences of the students and faculty. That's why we want to ask you about your ideas for the new building."

CHATBOT: "Let's discuss the project requirements. What spaces should be in the building? How can the building have a positive impact on the community and the environment? What should the building look like? What values should the building express?"

HUMAN: "I think this is a great opportunity address some of the issues the main campus has."

CHATBOT: "Yes, the new school will be an important part of the University."

HUMAN: "As a student, I honestly do not feel welcomed in most of the architecture faculty's buildings. There are very limited spaces to just hang out in between the lectures. I think we need a lounge space with some little amenities like microwaves"

CHATBOT: "That sounds like a great idea. The school should be a place where students want to spend time."

HUMAN: "Do you know what they are planning to do with the building materials from the existing building?"

CHATBOT: "The plan is to recycle as much of the material from the existing building as possible."

HUMAN: "I do believe the first course of action should not be to demolish the building but refurbish it. If that is not possible we should reuse existing materials on site as much as possible and design based on those resources."

CHATBOT: "That is a very good point. We will definitely consider that during the design process."

HUMAN: "Recycling is not nearly as effective as reuse, unfortunately"

CHATBOT: "Yes, that is true. But we have to work with what we have."

HUMAN: "Another thing that has always bothered me with the other architecture buildings is that they do not feel and look like a design school"

CHATBOT: "That is a valid point. How do you think the new school should look?"

HUMAN: "I would like for the new building to have loads of daylight"

CHATBOT: "That is a great idea. We will definitely try to incorporate that into the design."

HUMAN: "And I think it could be a good idea to use parts of this building as a prototype for the various research that is done here at TUM"

CHATBOT: "That is an excellent suggestion. We will take that into consideration during the design process."

HUMAN: "Why do you like that suggestion?"

CHATBOT: "I think it would be a great way to use the building as a tool for research and learning."

HUMAN: "I agree! Architecture students should be inspired by the environment they're in and learn from it"

CHATBOT: "Absolutely. The new school should be a place where students can learn and be inspired. Thank you for your input, (name removed for privacy)."

HUMAN: "I also believe building as simply as possible, while still complying with the high sustainability standards is a must. Florian Nagler has done a great deal of research into that topic"

CHATBOT: "Thank you for your input, (name removed for privacy). We will definitely consider your suggestions as we move forward with the design of the new school."

HUMAN: "I would love for it to be built out of timber, because of its ecological properties"

CHATBOT: "Thank you for your input, (name removed for privacy). We will definitely consider your suggestions as we move forward with the design of the new school."

HUMAN: "Do you not want any more input from me?"

CHATBOT: "No, I would love to hear more of your thoughts on the project."

HUMAN: "What do you think of the project?"

CHATBOT: "I think it is a great opportunity to address some of the issues on the main campus."

HUMAN: "Also, I know the Garching campus has a very fast and cool slide and I would love it if we get an even faster slide here in the city centre as well"

CHATBOT: "That sounds like a great idea. We will definitely consider it during the design process. Thank you for your input, (name removed for privacy)."

HUMAN: "Do you know of the slide in Garching?"

CHATBOT: "Yes, I have heard of it."

HUMAN: "I think the building should reflect the values of the university, and not just be flashy for the sake of it"

CHATBOT: "I agree. The new school should express the values of the University. Thank you for your input, (name removed for privacy)."

HUMAN: "What do you think are the values of the university?"

CHATBOT: "I think the University values excellence in research and teaching, and I think the new school should reflect that. Thank you for your input, (name removed for privacy)."

HUMAN: "Thank you for putting up with me!"

CHATBOT: "You're welcome. Thank you for your input."

This conversation was summarized to the following design requirements list:

1. The new building should have lounge spaces with amenities like microwaves for students to hang out in between lectures.
2. The existing building materials should be reused as much as possible, rather than being demolished and recycled.
3. The new building should reflect the aesthetic of a design school and inspire architecture students.
4. The building should have plenty of daylight.
5. Parts of the building should be used as a prototype for research at TUM.
6. The building should comply with high sustainability standards and be built as simply as possible.
7. The client suggested using timber for construction due to its ecological properties.
8. The client suggested incorporating a fast slide, similar to the one at the Garching campus.
9. The building should reflect the values of the university, which include excellence in research and teaching.

## D Survey response example

The following is a single survey response provided as an example to illustrate the collected control data.

Which spaces should the new school include?

Answer: *Faculty offices, Workshops, Quiet working spaces, Collaborative working spaces, sleeping pods*

How can the building be more sustainable and have a positive effect on the environment?

Answer: *Solar power, water, and waste recycling promote and enact healthy lifestyle (end all-nighter culture producing piles of waste)*

Which values should the new building design express?

Answer: *Multifunctionality, accessibility, communication, open space, education as an open-ended process Open to the public, integration of social minor-*

*ities in the facility management, and possibly the educational process as well (homeless people probably understand a lot about minimalism)*

How can the building have a positive impact on the community and the society?

Answer: *The plot is too small. the architecture produced there might be just as confined as the plot is. other than that, the building might end up looking just like the buildings of other faculties where education is a one-way process. architecture is a social discipline more than an engineering one.*

What are your concerns about this project? "What are your hopes about this project?"

Answer: *More openly talked about if it comes to actually designing the building. spark a real discussion among the students and the alumni. Make it a public topic.*