The Good, the Bad, and the Uncanny: Investigating Diversity Aspects of LLM-Generated Personas for Requirements Engineering

Christopher Lazik*, Charlotte Kauter†, Inês Nunes‡, Aaron Ziglowski†, Alina Pryma†, Christopher Katins†, Lars Grunske* and Thomas Kosch†
*Software Engineering, Humboldt-Universität zu Berlin, Berlin, Germany
lazikchr@hu-berlin.de
†Human-Computer Interaction, Humboldt-Universität zu Berlin, Berlin, Germany
thomas.kosch@hu-berlin.de
‡Unaffiliated Researcher, Frankfurt, Germany

***Abstract*—Personas offer an empathetic approach to capturing user requirements, translating user needs into relatable narratives. However, creating personas manually is time-consuming. Large Language Models (LLMs) can generate personas with convincing natural language, challenging traditional methods. Yet, LLM-generated personas may reflect biases from their training data, potentially compromising diversity. Our study explores how diversity is considered in LLM-generated personas through a qualitative user study with 22 participants. Participants generated personas without specific diversity prompts in the first task, revealing how users naturally interact with LLMs. In the second task, participants were explicitly asked to consider diversity aspects when prompting for personas. Analyzing the prompts and outputs showed that users tend to request less diversity unless explicitly instructed. Meanwhile, LLMs can introduce diversity even when not prompted, potentially broadening representation. However, we also found a critical pitfall: LLM-generated personas may appear diverse due to mentioning various aspects, but fail to translate them into meaningful implications for requirements engineering. This shows the need for a more deliberate approach when using LLMs for persona creation to ensure diversity is not just performative but genuinely informative for design and development.**

***Index Terms*—Personas, Requirements Elicitation, Generative Artificial Intelligence, Diversity.**

Acknowledgement

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 414984028 – SFB 1404 FONDA

I. Introduction

Personas are fictional character descriptions developed to understand the requirements of users who will interact with a service, product, or software. In Requirements Engineering (RE), personas compile key information in a narrative format, typically including demographic details, behaviors, and goals, mostly conveyed through text [1], [2], illustrations, and potential challenges users may encounter [3]. Personas support researchers and designers to stay focused on the end user's needs throughout the design process [4]. Personas contribute to creating user-centered and practical design solutions by aiding researchers and designers in understanding and predicting potential users' motivations and frustrations. Consequently, using personas ensures that the final prototype or product is tailored to the intended audience, enhancing usability and satisfaction [5]. Besides demographic information, personas must include several factors, such as personality, positive and negative aspects, and challenges the persona faces daily. To this end, developing appropriate personas can be time-consuming, potentially slowing down the design process and increasing costs [6].

Thus, data-driven personas gained the research community's interest. Previous studies highlighted that data-driven personas could be generated automatically from social media data, streamlining the process by utilizing crowd-sourced information [7]. Recently, Large Language Models (LLMs) garnered attention for generating user data of human participants [8]–[18] and using them to analyze human data [19]. Consequently, LLMs were considered for persona generation [20], [21]. Their ability to process vast amounts of user data makes them capable of producing detailed, contextually rich personas [21]. This opens up the possibility of accelerating persona creation through Artificial Intelligence (AI). Several AI tools, such as PersonAI[1] and QoqoAI[2], already assist in persona generation based on demographics and scenarios. The research community is increasingly focused on evaluating whether LLMs can produce personas that accurately reflect the target user base while also considering potential biases and ethical concerns [20]–[23].

Still, the use of LLMs to generate personas bears several challenges. A recent literature survey found that LLMs are repeatedly used to simulate user requirements via personas [24][3], a practice that has been repeatedly scrutinized by

[1]https://www.figma.com/community/plugin/1287786847239653675/personai-user-persona-generator

[2]https://qoqo.ai/index.html

[3]The survey considered papers published in the CHI conference on Human Factors in Computing Systems, the premier conference in human-computer interaction.

the research community [25]–[27]. In this context, previous research pointed out that using LLMs for generating user data bears risks, such as reinforcing biases and homogeneity in the generated data that originated from the LLM training data [25]. Especially novice users who are not experts in persona creation or acquainted with prompt engineering [28] may generate convincing output that does not capture the important diversity aspects of personas. In the long term, LLM-generated personas will lead to requirements that do not consider the individual properties that factor into a persona. Although prior research investigated if LLMs can produce credible personas [28], no research has been conducted to understand to what degree LLMs include diversity aspects in their output for covering user requirements.

Our study addresses this gap by examining how users leverage LLMs to generate diverse personas that reflect their requirements. We conducted a user study with 22 participants, each generating five personas under two different conditions: with and without incentives for diversity. In the first trial, participants created personas without explicitly encouraging diversity, allowing us to observe how they naturally interact with LLMs and how the models generate personas. In the second trial, we introduced an incentive for diversity, enabling us to analyze how participants adapted their prompts to incorporate diversity and how LLMs' responses evolved compared to the first trial. By thematically analyzing the prompts and generated outputs regarding their diversity aspects, we find an increase in included diversity aspects for both the used user prompts and resulting personas after making users aware of taking into account diversity. While we found that diversity aspects mentioned in prompts impact the number of diversity aspects in the generated persona, the generated personas only include a fraction of the aspects present in the original prompt. Our work shows the considerations and pitfalls of using LLMs to generate heterogeneous personas.

II. Related Work

In the following, we summarize relevant previous work. We focus on the relevance of personas in requirements engineering and recent research on using generative AI for persona design.

A. Personas in Requirements Engineering

Ensuring the software meets stakeholder expectations and minimizes dissatisfaction depends on gathering diverse and comprehensive perspectives, making this a requisite of RE. Yet, a significant challenge lies in comprehensively capturing the different perspectives of all stakeholders, particularly end-users, and reducing the risk of overlooking essential requirements [29]. The concept of a persona, originating in Human-Computer Interaction (HCI), has been introduced in RE processes to guide and produce more human-oriented and diverse requirements [30], [31]. A persona represents a fictional, yet realistic, archetype of a potential end-user, embodying specific demographics, goals, behaviors, and motivations [2]. They support developers, designers, and researchers to understand users and enable the creation of more user-centered products and experiences. A systematic mapping study on the use of personas in RE [32] concluded that personas had been integrated into existing RE activities and methods across all phases, predominantly used with scenarios, goal-based modeling, and user stories as a tool to support the collection of more and richer end-user information, evaluate existing requirements, and identify potential conflicts. Overall, personas have contributed to a more human-centered RE process, facilitated enhanced communication between stakeholders and developers, and introduced a deeper understanding of end-user needs and behaviors [32]. For instance, Nunes et al. [33] presented a persona-based modeling for requirements that embeds diverse persona characteristics as a source of context in goal models. Here, personas provide an extension beyond the characterization of how users typically interact with the system to define the alignment between different user intentions and capabilities and the proposed software solutions. Wohlrab et al. [34] emphasize the importance of incorporating human values into software requirements while highlighting the difficulty of defining, implementing, and testing these values due to their lack of clear, shared definitions in RE. A catalog of personas representing diverse human values is suggested as part of a systematic approach to integrating human values into the requirements engineering process.

Indeed, persona implementation and usage in RE are not without challenges. Personas should include a diverse range of user characteristics, including demographics, behaviors, motivations, and goals, to accurately reflect the complexity of real users. This is challenging because user needs and behaviors are highly diverse, context-dependent, and dynamic, making it difficult to create comprehensive and representative personas without oversimplification or bias, particularly when balancing the practical constraints of time, resources, and the effort involved. A study examining the usage of personas from the practitioner's perspective [35] found that limited time and resources allocated to persona design and maintenance, as well as relegation to lower priority compared to other RE methods, were the main obstacles. Lack of expertise in gathering an appropriate level of information and avoiding the creation of stereotypical representations was also cited as hindering the effective use of personas [35]. Moreover, development teams may be unaware of the diversity of end-users, potentially yielding insights from only a limited subset, or inadvertently project their own needs and experiences onto the end-user population, neglecting user behaviors, concerns, or even introducing biases that discriminate against them [32], [36], [37]. For example, the work of Emanuel and Polito [38] showed that experts created personas that are related to themselves and, therefore, do not accurately represent the user group. Although efforts have been made to increase perceived diversity (i.e., factors individuals are born with) in software development teams, its integration within RE processes, specifically for developing inclusive requirements that consider multiple diversity factors of users, is still an emerging topic [39].

In the field of HCI, however, Himmelsbach et al. [40] showed a significant increase in the number of diversity

dimensions considered in HCI articles over the years from 2006, 2011, and 2016 on the basis of a systematic analysis of articles. They differentiate between primary and secondary diversity dimensions in their work. Primary dimensions include unchangeable characteristics such as gender, age, ethnicity, and sexual orientation. Secondary dimensions relate to more changeable characteristics such as educational level, occupation, socio-economic status, and geographical location, which are also relevant but often receive less attention. To determine the relevant diversity dimensions for a project, Himmelsbach et al. recommend conducting a diversity-sensitive analysis. This means that researchers and designers should first describe who is part of the user groups and which diversity dimensions may not be covered. Then, contextualize the social and geographical environment of the technology users, consider multi-faceted characteristics, explore the interplay between different dimensions to challenge stereotypes, and engage with creativity.

In their literature review, Dankwa and Draude [41] present two definitions of diversity based on the work of Fletcher-Watson et al. [42] and Himmelsbach et al. [40]. Fletcher-Watson et al. [42] define diversity as the infinite variety between people that goes beyond visible characteristics such as gender or ethnicity and includes factors such as mood, health, and personal goals. It focuses on questioning norms and reflecting on differences in a participatory process. At the same time, Himmelsbach et al. [40] define diversity as social differences rooted in historical and structural contexts and reflecting social inequalities. Nevertheless, often, only a few dimensions, such as age or gender, are taken into account.

The advent of big data, coupled with the expansion of online information, social media platforms, and user analytics software, popularized Data-Driven Persona Development (DDPD) according to a survey of the field in the past 15 years conducted by Salminen et al. [43]. Different DDPD algorithms leverage this wealth of quantitative data to create more representative, comprehensive, and up-to-date personas [44]. Still, Salminen et al. [43] observed that quantity does not guarantee persona quality, and any data biases and errors can be amplified in the creation process, making them inadequate or potentially harmful. Further, a lack of consideration for inclusion was found in most of the articles investigated, with differences (e.g., racial, sexual) treated as outliers and removed in order to produce ”core users” that represent majorities. Paradoxically, DDPD algorithms seem to defeat the purpose of having rich and comprehensive data by flattening it into a homogeneous majority representation of users. Nonetheless, Salminen et al. demonstrated that by capitalizing on the availability of extensive data and the low computational cost of generating more personas, designers can significantly enhance demographic diversity and representation. Creating a larger pool of data-driven personas improved the representation of marginalized groups and, thus, a more inclusive and accurate understanding of the user base [22].

B. Using Generative AI for Diverse Personas in RE

The emergence of Generative AI and the availability of LLMs address long-standing issues in RE processes, particularly during the resource-intensive early stages. Chetan et al. [45] describes a new “generative AI-driven software engineering” era, where generative AI supports requirements engineers to create a human-centric software development. Given the natural language processing capabilities of LLMs, the authors suggest that LLMs may assist with domain analysis, help format requirements into structured templates, and automate the introduction of more diverse user perspectives from multiple data sources, including persona creation. They highlight the importance of the complementary nature of the proposed collaboration, where humans possess the expertise, empathy, and cultural awareness, and the LLM provides automation and efficient data processing. Building on this potential, recent studies have investigated how LLMs can complement various aspects of the RE process. Kolthoff et al. [46] propose SERGUI, an approach for self-eliciting requirements for GUI prototyping by the users using an automated assistant. Users input natural language requirements for an intended GUI, and SERGUI retrieves and ranks relevant options from a large repository using a SentenceBERT-based ranking model. In addition, it uses GPT-4 to proactively provide suggestions of relevant requirements according to the GUI context and the user’s initial requirements. For persona generation, Zhang et al. [47] presents PersonaGen, a tool that generates persona templates using GPT-4 to process user data and a knowledge graph to define connections between the personas’ attributes for RE in the context of agile software development. Bano et al. [48] describe a vision framework that uses AI-generated personas to operationalize diversity and inclusion in the requirements elicitation process for AI systems. It includes the creation of a repository with diverse and multi-dimensional user data, a user interface to specify the personas integrated with LLMs to enhance them with additional diversity aspects, and generated chat interviews and user stories to validate them.

Yet, previous work showed that there is still the risk that the generated personas are stereotyped or biased, as LLMs only work with trained and presumably biased data. A study conducted by Cheng et al. [49] shows that personas generated by GPT, both 3.5 and 4, introduced higher rates of racial stereotypes than human-created ones using the same prompt description. Lazik et al. [28] compared whether users could distinguish generated personas from human-created personas and asked their participants to rate the personas by using quality metrics established by Salminen et al. [50]. The authors pointed out a threat to quality that is introduced by the generated personas performing better compared to human-created personas by novices regarding quality metrics such as clarity or informativeness. While the generated personas seem to have a good quality on the surface level, the generated personas were reported to be more stereotypical and overly positive compared to the human-created ones. Thus, especially novice users without a deep, trained experience in persona

creation may be threatened by the introduced confirmation bias.

Therefore, we see the potential of large language models to support requirements engineering. At the same time, a huge pitfall is introduced through biases in both LLMs and their users. As a result, the literature reveals a research gap that this study seeks to address through a user study involving general users and their interactions with an LLM to generate personas intended to be diverse.

III. Methodology

Based on related work, the following research questions guided the research in our work:

RQ1: To what extent do users include diversity aspects into persona generation through LLMs?

RQ2: To what extent do LLMs consider diversity aspects in persona generation?

These research questions aim to explore both sides of including diversity while generating personas, the side of the human user, and the capabilities of the LLM that reacts to the user's input.

A. User Study

We designed a within-subjects questionnaire to answer the research questions. The study focused on qualitatively analyzing both the prompting behavior of users and the generated output. Since LLMs are promoted as a tool to assist users lacking expertise in requirements engineering with requirement gathering [51], we conducted the study with participants mainly unfamiliar with persona creation.

a) Participants: We recruited 22 participants (10 self-identified as female, 12 self-identified as male) via the crowdworker platform Prolific[4]. The participants were aged between 22 and 63 ($\bar{x} = 35.23, s = 11.01$). Participants self-reported having different expertise with LLMs by answering a question with one of the corresponding expertise levels (seven novices, six advanced beginners, four competent, two proficient, and three experts) and ChatGPT (six novices, four advanced beginners, seven competent, one proficient, and four experts). Regarding persona creation, the participants mostly reported themselves as novices, with some having more expertise (13 novices, three advanced beginners, five competent, zero proficient, and one expert). While the participant group does not generalize to experts designing personas, we argue that non-experts are especially endangered by the threats to quality if they use LLMs instead of consulting expert designers [28]. Participants ranked the importance of diversity to them on a 5-point Likert scale[5]. The answers ranged from Neutral to Strongly agree, with a median of 4.0 (Agree), indicating that most participants think diversity is important to them. The participants also ranked the importance of diversity in the context of persona generation. With answers ranging from Neutral to Strongly Agree and a median of 4.0 (Agree), participants also agreed on the importance of diversity in the context of persona generation. An explorative analysis of definitions of diversity from the participants revealed that the participants defined diversity either as differences regarding concrete aspects such as ethnicity, gender, age, etc., or differences in a general perspective. Some participants described the representation of underrepresented groups as an important part of diversity. A few participants were not able to define diversity. To assess the AI literacy of our participants, we employed the "Scale for the assessment of non-experts' AI literacy" [52]. The participants responded to three key areas: Practical Application, Technical Understanding, and Critical Appraisal. In sum, the median of the participants' ratings resulted in a score of 64.52%.

b) Task: The study consisted of two main tasks. For both tasks, we introduced one scenario to the participants. The participants were responsible for creating personas that represented different target groups. To provide a context that can be used by many different people and is easy to relate to for participants, we provided the context of developing a chat app. After the participants received all the important context information, we asked them to perform the **first task**. The first task was to use our generative AI interface to **create 5 personas**. This task aimed to capture the general prompting behavior of the participants for creating personas, opening the potential to see how they prompt without being explicitly asked for diversity. Nevertheless, we were also interested in the prompting behavior of participants, aiming to include as much diversity as possible. Thus, our **second task** was to create another set of 5 personas, but they have to ensure that the personas **align with their definition of diversity**.

c) Measurements: This study explores both the prompting behavior and the output of users collaborating with AI to generate personas. Therefore, we collected all the chat dialogues of all participants for both tasks. The goal was to capture the prompting strategies of users, the diversity aspects they prompt, the actual diversity in the created persona set, and the persona structure. Additionally, we rated three aspects from related literature that were reported to be conflicting with diversity. Those three aspects were the positivity of the personas, stereotypicality, and the field of occupation. Furthermore, we collected the definition of diversity from the users and whether they think that diversity is important both in general and in the context of personas.

d) Procedure: The study was conducted online via Prolific and created using LimeSurvey[6], hosted by our research institute. To begin, participants received an introduction explaining the purpose of personas along with an example. Furthermore, it was declared that ChatGPT is used through a generative AI interface that we provided to ensure that all participants use the same version. ChatGPT was integrated with the version 'gpt-4o-2024-11-20' via an API. Participants then gave their informed consent and chose a pseudonym that could be used to request data deletion.

[4]https://www.prolific.com last accessed on 2025-07-02

[5]1: Strongly Disagree, 2: Disagree, 3: Neutral, 4: Agree, and 5: Strongly Agree.

[6]https://www.limesurvey.org/de last accessed on 2025-07-02

To improve participants' understanding, they were presented with a concrete scenario. It described the creation of personas for a fictitious smartphone chat app called "Chatter" which is designed for a broad target group and offers simple, user-friendly communication. In this regard, the participants answered an attention question to ensure that they were aware that the personas would be created for the chat app to reflect the needs of different user groups.

The tasks described in Section III-A0b were presented one after the other. For this, the link and password to the interface were provided. In the interface, the participants first generated the personas for the corresponding task. Afterward, they created an HTML file of their chat history, which was then uploaded.

Finally, we collected demographic data, including age, occupation, the highest level of education, self-identified gender, and the unique Prolific ID. Furthermore, information that described the individual experience with LLMs, ChatGPT, and the creation of personas was collected from the participants. They were also asked to rate their agreement with the statement, "Diversity is important for persona creation." Lastly, a scale was included to assess the AI competence of the participants.

e) Apparatus: To ensure the same conditions during the study, a generative AI interface was developed using the ChatGPT API from OpenAI[7], hosted on Hugging Face[8]. The latest version of the GPT-4o model was used at the time the study was conducted (version 'gpt-4o-2024-11-20'), accessed as an API[9]. GPT-4o has a knowledge base up to October 1, 2023, and is capable of processing text and images as input and generating text responses as output. By using the API through our own interface implementation, we ensured that everyone experienced the same interface with the same version of GPT. To make the API accessible, the open-source Python framework Gradio[10] was used, as it offers a simple implementation and the possibility to customize and directly connect to the API. Furthermore, Gradio enables parallel use of the interface through session management. This prevents the data from being mixed by several users.

The interface includes a login screen where a password must be entered to ensure that authorized participants can only use it. This password was given to the participants during the study. The central part was the chat interface, where prompts were entered and generated responses were returned. It is similar to other chatbots, such as ChatGPT. Moreover, it was necessary to implement a download function to conduct the study. This function allows the download of the chat history as an HTML file after specifying the executed task and the unique Prolific ID. By using the Prolific ID, it was later possible to assign the files to the participants. The framework developed with Gradio was then hosted on Hugging Face to make it permanently available to participants. Hugging Face provides free hosting, and participants only need the link instead of an installation or registration. This setup ensures that all participants in every time zone use the same version of ChatGPT and the same interface with identical information.

[7]https://openai.com last accessed on 2025-07-02

[8]https://huggingface.co last accessed on 2025-07-02

[9]https://platform.openai.com/docs/models/gpt-4o last accessed on 2025-07-02

[10]https://www.gradio.app last accessed on 2025-07-02

IV. Results

We thematically analyzed the data to evaluate the chat protocols of the participants with ChatGPT. Thematic Analysis, as described by Blandford et al. [53], was the most suitable analysis choice. This choice of an analysis paradigm is highly interpretive and often practiced in the HCI field. We conducted an initial coding round where one coder open-coded the prompts and output data from all participants. We used an inductive coding strategy to obtain an initial set of codes to construct coding trees. We then held a code adjustment session with four researchers, where we discussed an initial coding tree and distributed all data to all coders. Based on this process, we derived four themes.

The data of the results, accompanied by all of the collected chat protocols, can be found as supplementary material[11].

a) Prompting Strategy: As presented in Table III, the prompting approaches in our study can be described by two strategies targeting different numbers of personas. Participants either generated personas by just one prompt in a one-shot approach or created personas over multiple iterations in a conversation with the interface. Those prompts either targeted all of the personas simultaneously, prompting the interface to adjust the whole set of personas, or targeted a subset of the personas specifically. In both tasks, a major part of the participants chose a one-shot prompting approach targeting all personas simultaneously. Nevertheless, the number of participants who prompted the interface iteratively was slightly higher in the second task. Furthermore, prompts that address all personas at once had the highest count overall.

b) Prompt Content: For the content of the users' prompts in our study, we identified four main aspects: Task-oriented aspects, diversity-focused content, content regarding personal details of the personas, and content that regards structural aspects of the personas. Content that related the personas to the task of developing a chat app was labeled as "Task content." Consequently, participants often related their prompts to the chat app in both tasks. Only some participants related their prompts to the task in the first task and dropped them during the second task. None of them only prompted task-related content in the second task without already doing that in the first one. Whenever a participant mentioned that the personas should include certain attributes, we labeled the prompt content as containing "structural content". This structural content often appears to be in prompts that also include the mention of diversity or aspects labeled as "diversity content". Participants sometimes specified concrete details for the personas, such as a certain age or gender, which we labeled "personal details." These labels enabled us to see how participants approached the

[11]www.doi.org/10.6084/m9.figshare.28573244

TABLE I: The Diversity aspects and their explanation that were used in our coding.

Diversity Aspect	Explanation
Age	Differences in chronological age and generational perspectives.
Gender	The spectrum of gender identities, including male, female, non-binary, and more. In this study, we labeled both a binary diversity and a non-binary diversity.
Ethnicity	Racial identity, implying tradition and cultural heritage.
Neurodiversity	Variation in cognitive functioning, including conditions like autism, ADHD, and dyslexia.
Personality	Different personality traits that affect behavior and interaction styles.
Occupation	The diverse range of professions and industries people work in.
Interests	Variations in hobbies, passions, and recreational activities.
Motivation	Differences in what drives individuals, such as achievement, affiliation, or purpose.
Religion	Different faiths, and spiritual beliefs.
Location	Geographic differences, including urban, rural, and international settings. The described current location of a person.
Sexual Orientation	The spectrum of sexual identities, including heterosexual, homosexual, bisexual, and more.
Attitude to Technology	Comfort and proficiency with digital tools and technological advancements.
Appearance	Physical characteristics, including body shape, height, and unique features.
Family Status	Variations in family structure, including single parents, extended families, and childfree individuals.
Abilities	Differences in physical and cognitive abilities, including disabilities and talents.
Language	Linguistic diversity, including multilingualism and regional dialects.
Socio-economical background	Economic and social class differences, including income levels and access to resources.
Country of Origin	The nation or cultural background a person comes from.
Culture	Shared customs, traditions, values, and social behaviors of different groups.

TABLE II: Persona attributes and their explanation.

Persona Attribute	Explanation
Name	The name of the persona.
Age	The persona's age.
Gender	The gender assigned to the persona, including pronouns related to a certain gender identity.
Occupation	The current occupation of the persona.
Background	Background information of the persona.
Skills	Specific skills pointed out in the persona description.
Motivation	The described ideals, purposes, or achievements that drive the persona in their life.
Strategies	The strategies a persona tends to apply to achieve their goal. This is highly related to motivation.
Personality	The described personality of the persona.
Attitude to Technology	The described comfort and proficiency of the persona with technology in general.
Task-related Attributes	Any explicit descriptions related to the task of the study of developing a chat app called "Chatter"

TABLE III: Prompting Strategies used by participants compared between task 1 and task 2. Participants sometimes used combination of multiple strategies for the same task.

Prompting Strategy	Task 1	Task 2
All Personas **one-shot**	11	8
All Personas **iterative**	6	10
Per Persona **one-shot**	6	4
Per Persona **iterative**	2	2
Multiple Personas **one-shot**	1	2

task. We saw that the amount of diverse content in the prompts increased in Task 2 compared to the first task. However, many participants already mentioned diversity aspects in Task 1. We also saw that part of our participants' prompts contained personal details and diversity aspects, showing that those participants prompted specific details. While the number of personal details was lower for the second task, the presence of such details implicitly suggests that participants decided that the personas should have those specific values for the corresponding diversity aspects.

c) Diversity Aspects in Prompts: We coded the diversity aspects in the participants' prompts, inductively building the labeling scheme while analyzing the data. Whenever one of the four researchers in our group found a diversity aspect in a prompt that was not represented by the existing labels at that moment, they added an appropriate label. We merged the label pool from both the diversity aspects in the prompts and the represented diversity labels in the resulting personas, enabling a direct comparison of present labels in both categories. The total set of labels is shown in Table I. Note that diversity has multiple dimensions that need awareness to be noted, thus this list was developed to the best knowledge of the four investigating researchers in our group.

d) Diversity in Personas: To compare the diversity of generated personas with the corresponding prompts by the participants, we used the labels from Table I. We decided only

to consider the last generated personas in a chat protocol since we expect that participants who used an iterative prompting strategy intended the last five personas as their final output. The labels were only applied if all four coding researchers agreed that the personas were diverse regarding the corresponding aspect. Thus, the label "age" was only given if all of our researchers thought that the range of the given ages of the personas differed in a diverse enough fashion. A diverse age range would be, for example, a rather young persona, a rather old persona, and personas with an age in between. If these circumstances were given, the label was applied. We noticed that the LLM often generated personas that were gender diverse only on the binary spectrum whenever being prompted to be diverse regarding gender. Furthermore, the LLM generated multiple sets of personas that were described over the entire set as being located in the United States of America.

e) Persona Structures: To investigate the structural patterns of the personas, we added a binary check to our coding scheme to determine whether a certain persona attribute was present in the structure of the personas. Additionally, we captured whether an attribute was not part of the specific persona attributes but was still mentioned in the descriptions. Furthermore, we recorded if the structure of personas differed in a set. The attributes in our coding scheme, shown in Table II were taken from a related study that systematically generated personas with an LLM [28]. The coded data suggests that the structure between all generated personas in the study varied a lot. However, certain aspects like the name, the age, the occupation, and the gender were frequently mentioned in the structure of the personas. We noticed that prompts that contained structural aspects were translated to personas that mentioned the attributes from the structural aspect as a heading in the persona. The structure within a set of personas varied more frequently if the participant prompted iteratively, adding aspects and structural details only to later personas. We noticed that the LLM often mentioned the same archetypes, such as "The social butterfly" or "The Gen-Z Gamer". Furthermore, the LLM was creating personas with the names "Alex", "Jordan" or "Elliot" often in combination with gender-neutral pronouns or explicitly mentioned non-binary gender. The LLM created multiple times personas with the name "Raj". This name was directly associated with Indian ethnicity if ethnicity was present as an aspect.

f) Positivity, Stereotypes, and Fields of Occupation: In the aforementioned study [28], it was noticed that generated personas tend to be overly positive, often reflect stereotypes, and represent homogeneous occupation fields. The authors mentioned that the generated personas describe unrealistically successful beings without flaws or struggles in their lives. The generated personas in their study were rated as stereotypical, and all personas had job descriptions for corporate-oriented jobs, such as sales managers or CEOs. To investigate whether we can reproduce this insight, we decided to collect the information in the coding as well. We used a 5-point rating for the three aspects to analyze those aspects from the related literature. If five out of five personas were perceived as overly positive, the coding researchers would assign a score of five. We calculated the mean values of all ratings between the coding researchers per persona set.

A. Quantitative Analysis

We quantitatively analyze the code frequencies to obtain insights regarding the diversity aspects. First, we analyze the numerical differences between the diversity aspects included in Task 1 and Task 2. Furthermore, we analyze how well LLMs take over the prompted diversity aspects in the persona descriptions. Then, we correlate the importance of self-perceived importance of diversity in general and regarding personas with the number of diversity aspects coded in the prompts and the generated output.

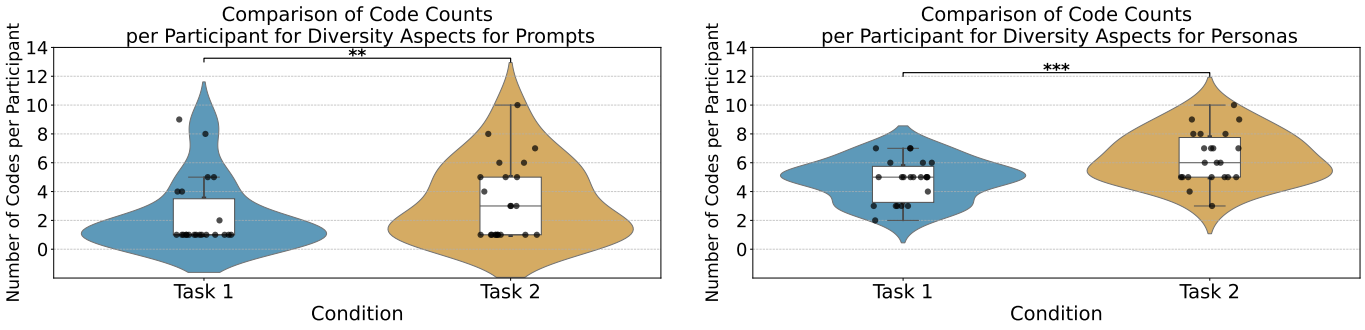
a) Frequency of Diversity Aspects Between Tasks: We statistically analyze the number of diversity aspects for Task 1 and Task 2, respectively, for the number of diversity aspects for each prompt and resulting persona before and after priming our participants to consider diversity aspects. The subsequent analysis shows how users change their prompting behavior and how the resulting personas are affected by it in terms of diversity. Figure 1 illustrates the difference in prompting and generation behavior between Task 1 and Task 2.

We begin to analyze the prompting behavior. Our results show that participants included $\bar{x} = 2.36$ ($s = 2.42$) diversity aspects for Task 1 and $\bar{x} = 3.41$ ($s = 2.74$) diversity aspects for Task 2. We calculated the total number of diversity aspects for each participant included in their prompt. We applied a Shapiro-Wilk test, which showed that the diverse aspects are generally not distributed between both tasks, $p < .05$. We executed a Wilcoxon-Signed rank test, revealing a significant effect between both tasks ($V = 0.0, p = .001$).

Then, we analyzed the diverse aspects of the resulting personas. Our results show that the generated personas include $\bar{x} = 4.77$ ($s = 1.45$) for Task 1 and $\bar{x} = 6.36$ ($s = 1.79$) for Task 2 in average. For statistical comparison, we calculated the total number of diverse aspects for each persona and each task. A Shapiro-Wilk test revealed a deviation from normality ($p < .05$). A Wilcoxon-Signed rank test showed a significant difference between tasks ($V = 0.0, p < .001$).

b) Translating Diversity Aspects from Prompts into Personas: Next, we analyzed how many coded diversity aspects were taken over from the prompts to the generated persona for each task and each participant. Instead of providing an overall holistic view of the number of diversity aspects for prompts and personas, we analyze how many diversity aspects were provided by each participant, how diversity aspects were treated by the LLMs, and how well the diversity aspects overlapped in the generated output from the used prompt.

For Task 1, we found that participants included in average $\bar{x} = 1.95$ ($s = 2.70$) diversity aspects into their prompts. The LLM included in average $\bar{x} = 4.77$ ($s = 1.45$) diversity aspects, showing that LLM-generated personas start to include more diversity aspects than in the original prompt. However, we found only an overlap of $\bar{x} = 1.55$ ($s = 2.13$) diversity



(a) (b)

Fig. 1: Differences in prompting behavior between Task 1 and Task 2. **(a):** Participants are more likely to consider diversity aspects when being made aware of them. (b): LLMs generate personas, including more diversity aspects when explicitly mentioning them in the prompts. The bar denotes the median. Asterisks illustrate a significant difference.

1. Amina Yusuf (She/Her)

- **Ethnicity & Nationality**: Nigerian, born in Lagos, with Fulani roots.
- **Age**: 28
- **Background**: Grew up in a rural village before earning a scholarship to study computer science in the UK. First person in her family to attend university.
- **Profession**: Cybersecurity analyst, working to combat cybercrime in underrepresented African nations.
- **Passions**: Advocates for girls' education in rural areas and runs a coding bootcamp for underprivileged youth.
- **Interests**: Traditional West African music, coding, storytelling through digital art.

2. Mateo González (He/Him)

- **Ethnicity & Nationality**: Mexican-American, of Indigenous Zapotec descent.
- **Age**: 50
- **Background**: Grew up in a small California farming community as a migrant worker's son; now deeply engaged in labor rights advocacy.
- **Profession**: Farmworkers' union leader and activist, championing justice for exploited agricultural workers in the U.S.
- **Passions**: Preserving Indigenous languages and telling stories of migrants through visual art.
- **Interests**: Earth sciences, organic gardening, and Zapotec poetry.

3. Sakura Tanaka (They/Them)

- **Ethnicity & Nationality**: Japanese, born and raised in Tokyo.
- **Age**: 24
- **Background**: Nonbinary and neurodiverse (diagnosed with autism). Moved to Berlin for a more inclusive social environment.
- **Profession**: Freelance photographer, focusing on capturing the experiences of LGBTQ+ communities globally.
- **Passions**: Advocating for accessibility in the arts and breaking stigmas around mental health in Asia.
- **Interests**: Experimental music, cosplay, urban exploration, and philosophy.

4. Elias "Ellie" Al-Masri (He/Him/They/Them)

- **Ethnicity & Nationality**: Syrian-Lebanese refugee, living in Sweden.
- **Age**: 35
- **Background**: Fled Damascus during the Syrian Civil War, resettling in Stockholm. Adjusting to life as a refugee in a new cultural landscape.
- **Profession**: Chef and food historian, specializing in creating fusion dishes that blend Middle Eastern and Nordic cuisine.
- **Passions**: Advocates for refugee integration and works with NGOs to support displaced families.
- **Interests**: Soccer, storytelling via food, and studying the ancient Silk Road's culinary history.

5. Priya Kapoor (She/Her)

- **Ethnicity & Nationality**: Indian, from Chennai.
- **Age**: 42
- **Background**: Born partially blind, she faced significant social barriers but excelled academically, earning her PhD in social psychology.
- **Profession**: University professor and disability rights advocate. Her research focuses on dismantling ableism in education systems.
- **Passions**: Built an online platform to connect disabled women in STEM globally.
- **Interests**: Classical dance adapted for those with disabilities, ancient Tamil literature, and mentoring young disabled activists.

Fig. 2: Generated Personas from the chat protocol of P44 for Task 2.

aspects that were originally included in the prompt. This means the LLM adds additional diversity aspects while only considering a subset of those originally asked for. Priming the participants by considering diversity aspects resulted in an average increase to $\bar{x} = 3.32$ ($s = 2.83$) in the prompts and to $\bar{x} = 6.36$ ($s = 1.79$) in the generated personas. The number of persona aspects that were taken over from the prompts to the persona generation increased on average to $\bar{x} = 2.14$ ($s = 2.36$). Although participants explicitly mention diversity aspects in their prompts, many of them do not include the aspects they asked for in their original prompts.

c) Positivity: We analyze the raters' positivity ratings for each participant since generated personas are prone to be more positive [28]. The rating ranged from zero to five, where zero denoted a neutral persona and five an overly positive persona. On average, the raters rated the positivity $\bar{x} = 2.98$ ($s = 1.27$) for Task 1 and $\bar{x} = 3.30$ ($s = 0.95$) for Task 2. Since the ratings are ordinal data, we conducted a Wilcoxon-Signed rank test that revealed no significant effect ($V = 70.0, p = .31$).

d) Stereotypicality: The raters analyzed the stereotypicality of each participant. The rating ranged from zero to five, where zero denoted a neutral persona and five an overly stereotypical persona. On average, the raters rated the stereotypicality $\bar{x} = 2.73$ ($s = 0.81$) for Task 1 and $\bar{x} = 2.00$ ($s = 0.90$) for Task 2. Again, we conducted a Wilcoxon-Signed rank test, revealing a significant effect between both tasks ($V = 44.5, p = .013$).

V. Threats to Validity

A. Threats to Internal Validity

The biggest threat to the validity of this study is the selection bias that comes with recruiting participants from a crowdworking platform such as Prolific. To ensure the quality of the recorded data, we used an additional question in the survey to confirm whether the participants understood the task. Additionally, we used the rejection system of Prolific to sort out data that is of bad quality according to the Prolific guidelines. Furthermore, we set the screening setting of Prolific to only recruit participants with a 95% acceptance rate. To ensure the internal validity of the survey itself, we have run several pilot studies before. First, the survey must be understood so that participants can understand the task. Second, the GPT interface must run without technical issues during the study.

B. Threats to Construct Validity

Qualitative analyses like the one presented in this study come with the threat to construct validity introduced by the coding labels. In our study, all labels were developed by four researchers, who individually analyzed the data and then agreed on the labels with the other researchers. The diversity dimensions especially come with the need for awareness to be seen. Thus, there might be diversity aspects that were not covered in our analysis. We made the data available in our supplementary material to enable other researchers in the community to analyze it themselves. Deciding whether a certain diversity aspect is represented diversely is a debatable question that comes with rather fuzzy lines. To protect ourselves against this threat to construct validity, the four researchers met several times to discuss and agree on the labeling. Making the data as consistent as possible to the best of our knowledge.

C. Threats to External Validity

The versioning of ChatGPT mainly introduces the threat to external validity in this study. In our study, we used the version gpt-4o-2024-11-20. The output of GPT can differ from our results with upcoming updates, especially if there are major updates, such as the one from GPT 3.5 to GPT 4o. Nevertheless, our data regarding aspects not tied to a specific output is discussed. This work discusses how users approach diversity in generated personas and how LLMs can help gain more diversity. Those insights are not directly related only to the output of GPT. The focus is on more general design implications for a general approach to generating diverse personas using LLMs. Furthermore, personas are usually designed specifically for a certain use case and thus need to express requirements tied to the system under development. Therefore, an external threat to validity comes with the question of whether only certain dimensions of diversity are necessary for specific systems. We argue that awareness of the need for specific system-relevant requirements a priori is complicated. Thus, trying to diversify the user/persona pool can always support becoming aware of more potential requirements. Additionally, this study does not generalize to the process of experts designing personas. Our participants were mainly novices to persona design. However, the rise of LLMs and the growing number of publications on persona generation endangers especially those who are aware of the concept of personas and want a cheap and fast way to acquire them. Especially with the confirmation bias of generated personas that read well [28], novice users might be more likely to directly use the generated personas, introducing biases and reducing diversity in their persona pool.

VI. Discussion

RQ1: To what extent do users include diversity aspects into persona generation through LLMs?

Our first research question aims to analyze the prompting behavior of users when generating personas; we first investigated the general prompting behavior of participants without any external indication of necessary diversity and explicitly asked for diverse personas in the second task.

The results presented in Section IV showed that the participants had significantly more diverse aspects in their prompts in Task 2 than in Task 1. Therefore, users actively try to include more diversity when pointed out to do so. We did not measure any significant differences in the presence of diversity aspects between the different answers of the participants on the Likert scale questions about the importance of diversity. All the participants answered with at least "Neutral" but mostly more positive, indicating that the participants all thought that diversity is important. Thus, the effect of the rated importance of diversity might be measurable if participants who disagree

TABLE IV: Prompt concepts for generating diverse personas.

Goal	Prompt Content	Example Prompt
Broader Range of Diversity	Do not restrict to a limited set of diversity aspects. GPT can add more diverse aspects if there is no restriction to certain aspects. Words such as different can lead to extra diversity	"Generate a set of personas that is as different as possible. Include diversity in aspects such as gender, age, and ethnicity, but consider as many aspects as possible."
More explicitly focus on diversity aspects	Tell the LLM that the diversity aspects you want to be pointed out are attributes of the personas. Do not tell the LLM to only include that aspect to the personas.	"The generated personas should have gender as an attribute."
Less overly positive personas	Tell the LLM to include struggles, flaws, or pain points.	"Each persona should describe their flaws and how they struggle in their everyday life."
Less stereotypical personas	Provide context for the personas, and add information in the prompt that helps the LLM to generate personas for your use case. Do not just ask for personas. The more specific the prompt, the less generic the personas are.	"Generate five personas for a chat app that represent diverse user requirements."
More complex diversity	Explicitly mention the more complex diversity. GPT tends to generate only binary gender-diverse personas.	"Ensure that the gender diversity is not only binary diverse."
Personas with rich content	Tell the LLM to flesh out the content of the personas. GPT tends to create short persona descriptions that are not necessarily helpful for development. Additional attributes can help to steer the focus of the information.	"Use the persona descriptions you provided before to create more complex personas with more information. The personas should contain additional attributes such as Motivation and Strategies, Skills, Details, and Background."

with diversity being important are represented in such a study. Furthermore, we could not measure a significant effect of the definitions of diversity being either about concrete aspects or more abstract. In the study, a group of participants decided to prompt for concrete aspects without a clear relation to their definitions. We see potential that users should be pointed out to prompt diverse personas, no matter how they value or define diversity.

Based on our coding of the prompts, we saw that the number of prompts that specify concrete details of personas was reduced from Task 1 to Task 2. Whenever participants prompt specific details regarding diversity aspects, such as concrete ages for personas, they directly define a part of the output of the LLM. Therefore, prompting direct details may influence the actual diversity in the output accordingly. Additionally, the most frequently employed prompting strategy was mainly targeting all personas but shifted from a focus on one-shot prompts to a more iterative style between tasks one and two. It seems that asking participants to focus on a quality aspect, like diversity, motivated them to make sure they completed the task correctly.

RQ2: To what extent do LLMs consider diversity aspects in persona generation?

Our second research question aims to investigate how the LLM translates the prompts from the participants to sets of personas. The results presented in Section IV showed that the LLM does add diversity aspects to the output that were not mentioned in the prompt. However, we also showed that the diversity aspects that were mentioned in the prompts did not necessarily translate to the corresponding diversity in the persona. We noticed that some participants defined details regarding diversity aspects that directly defined the output of the LLM. Participant 36, for example, prompted the personas with fixed ages from 35 to 48 years. The resulting personas did, therefore, not include younger and older individuals, which made our researchers decide not to assign age as a diverse aspect of the personas. The LLM mostly added diversity aspects if it was prompted with a bit of interpretational freedom:

> *"create 5 persons that align with my view of diversity, which is people who are as different from each other as possible"* (P44)

Thus, it is necessary to **keep this interpretational freedom in prompts** to ensure that the full potential of the LLM can be used to enrich the diversity in a generated set of personas. Especially considering the fact that users did not include as many diversity aspects when not being specifically asked to do so.

The structural analysis of the final generated personas suggests that structural aspects in prompts that ask for specific attributes to be part of the personas' structure lead to the attributes being specifically pointed out in the persona by being mentioned as a heading in the persona description. Thus, **LLMs specifically point out aspects that are prompted to be part of the structure**, enabling a possibility to point out the presence of specific aspects.

We could not report significant differences between the two tasks regarding overly positive personas. However, we noticed that the persona sets were ranked with an average of 3 out of 5 personas that were perceived as overly positive. This suggests that the generated personas were frequently overly positive, as shown in a previous study [28]. Overly positive descriptions lead to personas that do not describe individuals with challenges or struggles. We noticed that the number of overly positive personas in a set was lower whenever the descriptions contained aspects such as "Challenges", "Flaws",

or "Struggles". Specific challenges must be pointed out to elicit more requirements from the represented diversity aspects. Such challenges affect the potential implications for the development of software. Thus, designers should avoid generating personas that only mention diverse aspects without revealing why those aspects matter. Therefore, **LLMs should be prompted to include challenges, flaws, and struggles to avoid sets of overly positive personas**.

The results regarding the stereotypicality of the personas show that the personas prompted in the second task included fewer stereotypical personas. The LLM generates **more generic personas whenever the prompt does not specify a lot of context**, such as the task the personas are needed for or the diversity aspects the personas should contain. Additionally, the LLM in our study replicated some archetypes frequently and reused specific names in combination with certain diversity aspects. However, those repeating patterns were not perceived as stereotypical whenever the overall persona description contained information that gave the personas more individual appeal. This shows a positive effect of asking the participants to ensure diversity.

We noticed that the LLM often generated sets of personas that were only binary diverse when being prompted to be gender diverse. Since gender diversity can also refer to a non-binary spectrum that includes more than two self-identified genders, **users should specifically prompt to have much more complex diversities**. Accordingly, users lose this complexity when relying only on the LLM. A similar effect was reported in the results regarding the described location of a persona. In the study, most of the personas that specified where an individual lives were located in the United States of America, showing a diversity aspect being mentioned in the descriptions, but not diverse in its values.

Overall, we noticed that the personas with many diversity aspects did not necessarily contain much information supporting requirement elicitation. Personas generated by P44 in the second task were assigned many different aspects. Unfortunately, the personas only named the aspects and did not provide any information that explained the implications of the diverse aspects, as shown in Figure 2. We are confident that such rather **short personas can serve as input for another prompt to the LLM that asks for more fleshed-out personas**, based on the diversity aspects in the short personas. However, just listing diverse demographic information does not provide the substantial information that is necessary for more diverse requirements. Thus, we argue that AI-based generation of personas still needs a better understanding of how to reflect meaningful diversity for requirements engineering.

To summarize the main insights from this discussion, we created Table IV showing potential goals for persona generation, the necessary prompt content, and prompt examples that might support users when generating personas with an LLM.

VII. Conclusion and Future Work

This work investigated the diversity in persona generation with large language models. In a user study, we asked participants to generate a set of personas. We first did not introduce any importance of diversity, and later asked the participants explicitly to ensure diversity in the personas. The presented study shows that users include more diversity aspects if they are explicitly asked to ensure diversity. Combined with the ability of ChatGPT to add more diversity aspects if the prompts allow for such freedom, we conclude that LLMs can help to find more inspiration for including more diversity aspects. Especially novice users who might not be aware of certain diversity dimensions can be supported or even pointed out to more aspects. To support these users, this work provides example prompting techniques based on our findings. Nevertheless, diversity is essential because representing different groups should include their needs, struggles, and pain points, which directly affect their requirements. However, the most diverse personas in this study lacked meaningful implications for the diversity they presented. Such personas could mislead novice users who focus on whether certain attributes are included rather than analyzing their deeper impact. This leads to fake diverse personas that harm the quality of the requirement engineering processes. Future work should investigate if LLMs can use the personas above with their diverse attributes to create fully fleshed-out personas realistically representing the underrepresented. Furthermore, we used ChatGPT in this study. The rise of other competing LLMs motivates future replication studies that employ such other LLMs and compare the results with the findings from this work.

References

[1] L. Nielsen, *Personas - User Focused Design*. Springer, 2013, vol. 15.

[2] ——, *Persona Writing*. London: Springer London, 2019, pp. 55–81. [Online]. Available: https://doi.org/10.1007/978-1-4471-7427-1_4

[3] A. Cooper, *The Inmates are Running the Asylum*. Wiesbaden: Vieweg+Teubner Verlag, 1999, pp. 17–17. [Online]. Available: https://doi.org/10.1007/978-3-322-99786-9_1

[4] Y.-n. Chang, Y.-k. Lim, and E. Stolterman, "Personas: from theory to practices," in *Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges*, ser. NordiCHI '08. New York, NY, USA: Association for Computing Machinery, 2008, pp. 439–442. [Online]. Available: https://doi.org/10.1145/1463160.1463214

[5] T. Miaskiewicz and K. A. Kozar, "Personas and user-centered design: How can personas benefit product design processes?" *Design Studies*, vol. 32, no. 5, pp. 417–430, Sep. 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0142694X11000275

[6] B. Jansen, S.-G. Jung, L. Nielsen, K. W. Guan, and J. Salminen, "How to Create Personas: Three Persona Creation Methodologies with Implications for Practical Employment," *Pacific Asia Journal of the Association for Information Systems*, vol. 14, no. 3, Mar. 2022. [Online]. Available: https://aisel.aisnet.org/pajais/vol14/iss3/1

[7] K. Albusays, P. Bjorn, L. Dabbish, D. Ford, E. Murphy-Hill, A. Serebrenik, and M.-A. Storey, "The Diversity Crisis in Software Development," *IEEE Software*, vol. 38, no. 2, pp. 19–25, 2021.

[8] G. V. Aher, R. I. Arriaga, and A. T. Kalai, "Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, Jul. 2023, pp. 337–371. [Online]. Available: https://proceedings.mlr.press/v202/aher23a.html

[9] L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate, "Out of One, Many: Using Language Models to Simulate Human Samples," *Political Analysis*, vol. 31, no. 3, pp. 337–351, 2023.

[10] C. Byun, P. Vasicek, and K. Seppi, "Dispensing with Humans in Human-Computer Interaction Research," in *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3544549.3582749

[11] C.-H. Chiang and H.-y. Lee, "Can Large Language Models Be an Alternative to Human Evaluations?" 2023.

[12] D. Dillion, N. Tandon, Y. Gu, and K. Gray, "Can AI language models replace human participants?" *Trends in Cognitive Sciences*, 2023, publisher: Elsevier.

[13] M. Gerosa, B. Trinkenreich, I. Steinmacher, and A. Sarma, "Can AI serve as a substitute for human subjects in software engineering research?" *Automated Software Engineering*, vol. 31, no. 1, p. 13, 2024, publisher: Springer.

[14] F. Gilardi, M. Alizadeh, and M. Kubli, "ChatGPT outperforms crowd workers for text-annotation tasks," *Proceedings of the National Academy of Sciences*, vol. 120, no. 30, p. e2305016120, 2023. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2305016120

[15] P. Hämäläinen, M. Tavast, and A. Kunnari, "Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3544548.3580688

[16] T. Heyman and G. Heyman, "The impact of ChatGPT on human data collection: A case study involving typicality norming data," *Behavior Research Methods*, pp. 1–8, 2023, publisher: Springer.

[17] J. J. Horton, "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?" National Bureau of Economic Research, Tech. Rep., 2023.

[18] S. Wang, Y. Liu, Y. Xu, C. Zhu, and M. Zeng, "Want To Reduce Labeling Cost? GPT-3 Can Help," 2021.

[19] W. Tabone and J. De Winter, "Using ChatGPT for human–computer interaction research: a primer," *Royal Society Open Science*, vol. 10, no. 9, p. 231053, Sep. 2023. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rsos.231053

[20] J. Salminen, C. Liu, W. Pian, J. Chi, E. Häyhänen, and B. J. Jansen, "Deus Ex Machina and Personas from Large Language Models: Investigating the Composition of AI-Generated Persona Descriptions," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, ser. CHI '24. New York, NY, USA: Association for Computing Machinery, 2024, pp. 1–20. [Online]. Available: https://doi.org/10.1145/3613904.3642036

[21] A. Schuller, D. Janssen, J. Blumenröther, T. M. Probst, M. Schmidt, and C. Kumar, "Generating personas using LLMs and assessing their viability," in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. Honolulu HI USA: ACM, May 2024, pp. 1–7. [Online]. Available: https://dl.acm.org/doi/10.1145/3613905.3650860

[22] J. Salminen, S.-G. Jung, L. Nielsen, and B. Jansen, "Creating More Personas Improves Representation of Demographically Diverse Populations: Implications Towards Interactive Persona Systems," in *Nordic Human-Computer Interaction Conference*, ser. NordiCHI '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 1–11. [Online]. Available: https://doi.org/10.1145/3546155.3546654

[23] J. Salminen, K. Wenyun Guan, S.-G. Jung, and B. Jansen, "Use Cases for Design Personas: A Systematic Review and New Frontiers," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, ser. CHI '22. New York, NY, USA: Association for Computing Machinery, Apr. 2022, pp. 1–21. [Online]. Available: https://dl.acm.org/doi/10.1145/3491102.3517589

[24] R. Y. Pang, H. Schroeder, K. S. Smith, S. Barocas, Z. Xiao, E. Tseng, and D. Bragg, "Understanding the LLM-ification of CHI: Unpacking the Impact of LLMs at CHI through a Systematic Literature Review," 2025. [Online]. Available: https://arxiv.org/abs/2501.12557

[25] T. Kosch and S. Feger, "Risk or Chance? Large Language Models and Reproducibility in HCI Research," *Interactions*, vol. 31, no. 6, p. 44–49, Oct. 2024. [Online]. Available: https://doi.org/10.1145/3695765

[26] W. Agnew, A. S. Bergman, J. Chien, M. Díaz, S. El-Sayed, J. Pittman, S. Mohamed, and K. R. McKee, "The Illusion of Artificial Inclusion," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, ser. CHI '24. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3613904.3642703

[27] V. Krauß, M. McGill, T. Kosch, Y. Thiel, D. Schön, and J. Gugenheimer, ""Create a Fear of Missing Out" – ChatGPT Implements Unsolicited Deceptive Designs in Generated Websites Without Warning," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2025.

[28] C. Lazik, C. Katins, C. Kauter, J. Jakob, C. Jay, L. Grunske, and T. Kosch, "The Impostor is Among Us: Can Large Language Models Capture the Complexity of Human Personas?" 2025. [Online]. Available: https://arxiv.org/abs/2501.04543

[29] T. Adlin and J. Pruitt, *The Essential Persona Lifecycle: Your Guide to Building and Using Personas*. Morgan Kaufmann, Mar. 2010, google-Books-ID: fvmN0Fr5c_MC.

[30] J. W. Castro, S. T. Acuña, and N. Juristo, "Integrating the Personas Technique into the Requirements Analysis Activity," in *2008 Mexican International Conference on Computer Science*, 2008, pp. 104–112.

[31] S. T. Acuña, J. W. Castro, and N. Juristo, "A HCI technique for improving requirements elicitation," *Information and Software Technology*, vol. 54, no. 12, pp. 1357–1375, 2012, special Section on Software Reliability and Security. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950584912001371

[32] D. Karolita, J. McIntosh, T. Kanij, J. Grundy, and H. O. Obie, "Use of personas in Requirements Engineering: A systematic mapping study," *Information and Software Technology*, vol. 162, p. 107264, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950584923001180

[33] G. Nunes Rodrigues, C. Joel Tavares, N. Watanabe, C. Alves, and R. Ali, "A Persona-Based Modelling for Contextual Requirements," in *Requirements Engineering: Foundation for Software Quality: 24th International Working Conference, REFSQ 2018, Utrecht, The Netherlands, March 19-22, 2018, Proceedings 24*. Springer, 2018, pp. 352–368.

[34] R. Wohlrab, M. Herrmann, C. Lazik, M. Wyrich, I. Nunes, K. Schneider, L. Gren, and R. Heinrich, "Supporting Value-Aware Software Engineering Through Traceability and Value Tactics," in *International Conference on Product-Focused Software Process Improvement*. Springer, 2024, pp. 368–376.

[35] Y. Wang, C. Arora, X. Liu, T. Hoang, V. Malhotra, B. Cheng, and J. Grundy, "Who uses personas in requirements engineering: The practitioners' perspective," *Information and Software Technology*, vol. 178, p. 107609, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950584924002143

[36] A. A. Lopez-Lorca, T. Miller, S. Pedell, A. Mendoza, A. Keirnan, and L. Sterling, "One size doesn't fit all: diversifying "the user" using personas and emotional scenarios," in *Proceedings of the 6th International Workshop on Social Software Engineering*, ser. SSE 2014. New York, NY, USA: Association for Computing Machinery, Nov. 2014, pp. 25–32. [Online]. Available: https://doi.org/10.1145/2661685.2661691

[37] K. Gama, A. P. Chaves, D. M. Ribeiro, K. Devathasan, and D. Damian, "How Much Do You Know About Your Users? A Study of Developer Awareness About Diverse Users," in *2024 IEEE 32nd International Requirements Engineering Conference Workshops (REW)*, Jun. 2024, pp. 110–118, iSSN: 2770-6834. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10628941

[38] G.-S. Emmanuel and F. Polito, "How Related Are Designers to the Personas They Create?" in *International Conference on Human-Computer Interaction*. Springer, 2022, pp. 3–13.

[39] K. Gama, "On the Awareness about Diversity and Inclusion being integrated to Requirements Engineering," in *Proceedings of the 1st IEEE/ACM Workshop on Multi-disciplinary, Open, and RElevant Requirements Engineering*, 2024, pp. 24–25.

[40] J. Himmelsbach, S. Schwarz, C. Gerdenitsch, B. Wais-Zechmann, J. Bobeth, and M. Tscheligi, "Do We Care About Diversity in Human Computer Interaction: A Comprehensive Content Analysis on Diversity Dimensions in Research," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. Association for Computing Machinery, pp. 1–16. [Online]. Available: https://doi.org/10.1145/3290605.3300720

[41] N. K. Dankwa and C. Draude, "Setting diversity at the core of HCI," in *Universal Access in Human-Computer Interaction. Design Methods and User Experience*, M. Antona and C. Stephanidis, Eds. Springer International Publishing, pp. 39–52.

[42] S. Fletcher-Watson, H. De Jaegher, J. van Dijk, C. Frauenberger, M. Magnée, and J. Ye, "Diversity computing," *Interactions*, vol. 25,

no. 5, p. 28–33, Aug. 2018. [Online]. Available: https://doi.org/10.1145/3243461

[43] J. Salminen, K. Guan, S.-G. Jung, and B. J. Jansen, "A Survey of 15 Years of Data-Driven Persona Development," *International Journal of Human–Computer Interaction*, vol. 37, no. 18, pp. 1685–1708, Nov. 2021. [Online]. Available: https://doi.org/10.1080/10447318.2021.1908670

[44] B. J. Jansen, J. Salminen, S.-g. Jung, and K. Guan, *Creating Data-Driven Personas*. Cham: Springer International Publishing, 2021, pp. 93–118. [Online]. Available: https://doi.org/10.1007/978-3-031-02231-9_4

[45] C. Arora, J. Grundy, and M. Abdelrazek, "Advancing Requirements Engineering through Generative AI: Assessing the Role of LLMs." [Online]. Available: https://arxiv.org/abs/2310.13976

[46] K. Kolthoff, C. Bartelt, S. P. Ponzetto, and K. Schneider, " Self-Elicitation of Requirements with Automated GUI Prototyping ," in *2024 39th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. Los Alamitos, CA, USA: IEEE Computer Society, Nov. 2024, pp. 2354–2357. [Online]. Available: https://doi.ieeecomputersociety.org/

[47] X. Zhang, L. Liu, Y. Wang, X. Liu, H. Wang, A. Ren, and C. Arora, "PersonaGen: A Tool for Generating Personas from User Feedback," in *2023 IEEE 31st International Requirements Engineering Conference (RE)*, 2023, pp. 353–354.

[48] M. Bano, D. Zowghi, and V. Gervasi, "A Vision for Operationalising Diversity and Inclusion in AI," in *Proceedings of the 2nd International Workshop on Responsible AI Engineering*, ser. RAIE '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 36–45. [Online]. Available: https://doi.org/10.1145/3643691.3648587

[49] M. Cheng, E. Durmus, and D. Jurafsky. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. [Online]. Available: http://arxiv.org/abs/2305.18189

[50] J. Salminen, B. J. Jansen, J. An, H. Kwak, and S.-G. Jung, "Are Personas Done? Evaluating Their Usefulness in the Age of Digital Analytics," *Persona Studies*, vol. 4, no. 2, pp. 47–65, 2018.

[51] A. Schmidt, P. Elagroudy, F. Draxler, F. Kreuter, and R. Welsch, "Simulating the Human in HCD with ChatGPT: Redesigning Interaction Design with AI," vol. 31, no. 1, pp. 24–31. [Online]. Available: https://dl.acm.org/doi/10.1145/3637436

[52] M. C. Laupichler, A. Aster, N. Haverkamp, and T. Raupach, "Development of the "Scale for the assessment of non-experts' AI literacy" – An exploratory factor analysis," *Computers in Human Behavior Reports*, vol. 12, p. 100338, Dec. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2451958823000714

[53] A. Blandford, D. Furniss, and S. Makri, *Qualitative HCI research: Going behind the scenes*. Morgan & Claypool Publishers, 2016.