Nobody Talks About the Prompt Club: Evaluating Diverse Persona Generation Through System Prompts and Guidelines

Christopher Lazik*, Charlotte Kauter†, Isabella Graßl‡, Alina Pryma†, Christopher Katins†, Lars Grunske*, 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
‡Computer Science, Technische Universität Darmstadt, Darmstadt, Germany

***Abstract*—Personas are widely used to represent users' motivations, goals, and needs, yet the perspectives of underrepresented groups are rarely addressed in human-centered design. Large Language Models (LLMs) promise to generate personas from vast datasets, but achieving genuine diversity and accurate representation requires deliberate prompting. In this work, we investigate two strategies, namely system prompts and guidelines, to support users in crafting diverse personas and evaluate their effects in a controlled user study (N=66). Our results show that both strategies help users generate more diverse personas, with their combination achieving the highest coverage of diversity aspects. Consequently, this work aims to go beyond evaluating LLMs as tools for persona creation. We show and discuss strategies to support requirements engineers during their interaction with the LLM.**

***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 an empathetic approach to represent requirements in human-centered design [1]. In the form of fictional character descriptions, persona creators provide a text-based depiction of potential users of the system under development. While personas are more focused on describing an archetype than an actual person, persona designers elicit and group insightful information to create a representation of needs, motivations, and context that developers can relate to when creating the system. This progress in understanding human beings and in representing them as personas that feel realistic and believable takes time [2]. With the emergence of Large Language Models (LLMs), an opportunity arose to quickly generate well-formatted text and, therefore, a quicker approach to persona creation using models trained on large text corpora [3].

The literature highlights an ongoing diversity crisis in eliciting requirements from diverse stakeholder groups [4], [5]. The fact that many different diversity dimensions affect user requirements may seem obvious from a requirements engineering perspective, but the dimensions and resulting requirements are not obvious and need representation, since they need the awareness of them in order to be considered [6]. However, representing generally underrepresented groups is complicated per definition. Employing the broad knowledge embedded in LLMs could be a promising way to capture information that is otherwise difficult to obtain [3]. However, artificially generated personas were discussed and criticized [7], [8]. Especially if novices skip consulting an expert for persona creation, LLM-based personas can be dangerously convincing, establishing stereotypes and overly positive depictions that threaten software quality and, by extension, the potential of more inclusive requirements in software engineering [9]. To evaluate the potential of generated personas for diversity, we previously conducted a study showing that LLMs can help, especially novices, enrich diversity when prompted correctly [10]. On the other hand, our work notes that diverse personas do not necessarily contain meaningful information for software development. Consequently, generating personas using LLMs is a non-trivial process, requiring not only knowledge of the concept of personas but also of how to approach prompt engineering to achieve a meaningful, diverse set of personas.

As an initial step toward supporting LLM users in persona generation, we proposed a set of guidelines derived from qualitative findings [10]. While our work proposed such guidelines, it is mandatory to scientifically investigate their effect and how users can generally be supported when interacting with LLMs. Since LLMs such as OpenAI's GPT[1] support system prompts for contextual instructions, these guidelines can be embedded directly into the model's behavior. This creates a research gap in evaluating how guidelines and system prompts influence persona generation and human–LLM interaction in requirements engineering.

In this work, we evaluated the guidelines proposed by our previous work [10] through a mixed within–between subjects user study. Participants were assigned to one of three interfaces for persona creation, which differed in how they supported diversity: (1) Explicit guidelines presented to the user, (2) a pre-prompted LLM that followed the guidelines without the user's knowledge, and (3) a combination of both approaches. We analyzed both participant prompts and LLM outputs using qualitative methods. The identified

[1]https://openai.com (last accessed 2026-05-28)

diversity aspects were compared across conditions and against the findings from the prior study that introduced the guidelines. Our results show that both guidelines and system prompts help users generate more diverse personas, with the combined condition achieving the broadest coverage of diversity aspects. Overall, structured support can guide LLM-based persona creation toward greater diversity consideration, particularly for novice users. The findings have implications for the design of guideline-based support in chat-based LLM interfaces and for system prompt engineering, which, in our study, achieved greater diversity coverage than reported in the related work. We further discuss the influence of system prompts in human–LLM interaction, the responsibilities associated with their use, and the broader implications of relying on LLM-generated data to represent underrepresented groups.

The contribution of this work is threefold:

1) We present a user-centered evaluation of how guidelines and system prompts shape the interactive process of the creation of LLM-generated personas.
2) We show, through a mixed-methods user study, that these strategies foster more diverse persona generation, with their combination achieving the widest coverage but also revealing trade-offs between inclusiveness and meaningfulness.
3) Finally, we discuss the implications of relying on LLMs for persona generation and user representation.

Together, our work advances the understanding of guideline design, system prompt engineering, their support in human-LLM interaction, and the role of LLMs in representing underrepresented groups in human-centered software development.

II. Related Work

In the following, we summarize relevant work on persona creation, the diversity aspects of LLMs, and user support for LLMs.

A. Persona Creation – From Long Design Session to One Shot Prompting

Personas represent fictional yet realistic user characteristics, including the motivations, goals, and behavioral patterns of the target audience, in user-centered design [11].

Traditionally, persona development has been an intensive process requiring extensive user research, stakeholder interviews, and iterative refinement through collaborative design sessions [12]. This often demands significant time and resource investments that many development teams struggle to accommodate [13], [14]. The emergence of LLMs is transforming this process, introducing the possibility of quickly and easily generating personas through structured prompting techniques [3]. However, generated personas are also described as clear and consistent, yet stereotypical and overly positive [9].

Nevertheless, the transition introduces new challenges, particularly regarding the consistency and diversity of generated outputs. Current findings suggest that while LLMs can produce lexically diverse persona descriptions, the selection and configuration of specific models become critical decision factors in application success [15].

B. Diversity in Personas – Why Archetypes Are Not Equal to Stereotypes

Within computer science, the importance of diversity and inclusion has, for the past few years, gained recognition across various sub-disciplines, particularly in modeling and design [16], [17], where diversity considerations are increasingly recognized as essential for effective outcomes [18]. Within this context, *diversity* includes a set of characteristics, traditionally categorized into primary/internal dimensions including gender, age, and physical capabilities, and secondary/external dimensions including work experience, cultural background, and socioeconomic factors [19], [20].

While gender considerations dominated diversity discussions in computer science [21], recent research shows the need to adopt broader perspectives that account for cultural backgrounds, neurodiversity, and beyond binary gender definitions [5], [22], which significantly influence software experiences.

Despite this growing awareness, the integration of diversity considerations into requirements engineering practices remains underexplored [23], [24]. Still, traditional requirements gathering approaches often fail to capture the full spectrum of customer requirements and use cases [25]. A particular challenge arises when attempting to represent users with special needs or requirements, as it may not always be feasible to include representatives from all relevant groups directly in the requirements process [13], [14].

The emergence of LLMs has introduced new dimensions to the diversity challenge in computer science. Current research demonstrates that these LLMs exhibit biases related to gender, social norms, and professional expectations [26], [27]. It is reported that some LLMs tend to manifest traditional gender stereotypes by associating women with passive roles and highlighting physical appearance, while portraying men in contexts of authority and professional achievement [28], [29]. This pattern is particularly evident in professions related to computer science, for example, where software developers are predominantly portrayed as male [30]. Research defines stereotypes as widely held beliefs about the characteristics, attributes, and behaviors of particular groups [31]. While personas aim to represent archetypes of potential users, they should not reinforce stereotypes, preserving harmful biases against certain demographic backgrounds [32].

Research demonstrates that these biases can manifest across multiple dimensions, including age-based discrimination that appears through stereotypes, invisibility, exclusion, and differential treatment of various age cohorts [33]. Additionally, incomplete data representation for certain demographic groups, particularly older adults, can result in significantly inaccurate persona generation outcomes [34]. One of the most recent comprehensive analyses examining LLM-generated personas against human-written self-descriptions reveals that generation tends to emphasize racial markers and ‘culturally coded’ language, and create personas that appear sophisticated, but remain rather narrow [7].

However, while LLMs present challenges through their potential to project training data biases into educational and professional subjects, they also offer opportunities to actively address diversity concerns, when carefully designed [8], [11], [35] such as the framework Enactive Artificial Intelligence [36]. Using LLMs to

generate personas was shown to have potential to add diversity aspects that users did not consider in the first place [10]. However, this potential is limited to the correct prompting strategies, making the creation of diverse personas a prompt engineering problem as well. While our previous study proposed preliminary guidelines based on their findings, these recommendations remain largely unvalidated as concrete support strategies for LLM users. In particular, novice users, who may be more vulnerable to quality-related issues in tasks such as persona generation [9], are unlikely to be sufficiently supported by guidelines alone. To ensure reliable and effective use of LLMs in requirements engineering and human-centered design, it is essential to better understand how users can be actively supported when interacting with these systems.

C. Supporting Human-LLM Interaction for Persona-Generation

The effectiveness of LLM-generated personas depends highly on the quality of human interaction with the LLM, particularly through prompting. The idea of prompt engineering involves carefully crafting prompts to shape LLM output without altering the LLM itself [37]. Currently, researchers aim to describe approaches for formulating prompts to help LLM users achieve meaningful output [38]. While researchers even aim to develop automatic techniques for prompt engineering, the process of finding the right prompts remains a task of reasoning and complex understanding of the output and the model's capabilities [39]. General LLM users, especially novice users, however, might not be familiar with all the LLM's capabilities and therefore shouldn't be expected to prompt in the most efficient way.

To support LLM users and thus the field of Human-LLM interaction, it is necessary to investigate strategies to help LLM users during interaction [40]. While user-focused LLM interfaces can be enhanced with supportive features, it is necessary to understand how to empower users to properly prompt an LLM even without deeper knowledge regarding a specific use case. Researchers need to understand how to translate their insights into usable support strategies for general users. While it is possible to simply provide guidelines in the prompting interface, modern LLMs also allow system prompts that guide the LLM before the actual interaction. Thus, system prompts offer a direct way to support a user's interaction with the LLM from the perspective of an informed researcher, ensuring that the LLM is properly instructed to follow potential guidelines. However, to the best of our knowledge, the literature lacks user studies investigating the potential of those two support strategies. Especially the direct comparison between them, and the absence of any support at all for specific use cases such as persona creation, are research gaps we want to investigate.

Since LLMs are continuously evolving and consequently always in flux, good requirements engineering practice calls not only for investigating LLMs as a tool for RE but also for how requirements engineers should approach these tools in a meaningful way. While the related work suggests different potential guidelines or templates for prompting, the community needs evidence from user studies that conduct the actual effect of such support strategies on the interaction of the user with the LLM. Thus, focusing on a specific use case, this work aims to provide insights into strategies to support creators of LLM-based personas in crafting as diverse personas as possible.

III. Methodology

To investigate the research gap of the impact of prompting guidelines and system prompts on the process of diverse persona creation, we formulated the following research questions guiding our research, presented in this paper:

RQ1: To what extent do guidelines for users affect the process of persona generation?

RQ1.1: How do users change their prompting behavior for generating personas based on provided guidelines?

RQ1.2: To what extent does the diversity in generated personas change if users get explicit guidelines for enhancing diversity?

RQ2: To what extent does a system prompt affect the expressed and included diversity in the process of persona generation?

RQ2.1: To what extent does the prompting behavior of a user affect the output of an LLM interface with a system prompt for diverse persona generation?

RQ2.2: To what extent does the diversity in generated personas change if the used LLM gets a specific system prompt?

Our research questions investigated the interaction among users generating personas. We color coded the RQs based on Figure 2 to highlight the focus of each RQ related to the study design. The main distinction between RQ1 and RQ 2 is the focus on the guidelines and the system prompt as a support strategy. The sub-RQs aim to capture the human aspect of the process and the resulting output of the LLM Thus, we decided to employ a user study to collect chat protocols of such interactions. To understand the effect of system prompts and guidelines, we designed different interfaces to compare the resulting chat protocols with the baseline from our previous study [10]. We ensured that the study setup and prolific screening filters are the same as in our previous study that served as the baseline.

A. User Study

In this work, we designed a mixed within-between-subjects design, with the task of generating personas for a chat app as the within factor and the different interfaces for interacting with the LLM as the between factor. We collected demographic data via a LimeSurvey survey hosted by our university, which directed participants to the chat interfaces.

1) Participants: We collected the data of 66 participants (32 self-identified as female, 33 self-identified as male, and one self-identified as non-binary/non-conforming) with ages ranging from 18 to 74 ($\bar{x} = 39.23, s = 13.31$). The participants were recruited via the crowd-sourcing platform Prolific[2]. We adjusted the screening filters to an Prolific-internal approval rate of at least 98%, English speaking participants from either the US or the UK. This was the exact same screening filters as in our previous work [10]. Participants were randomly and exclusively assigned to the study groups until each group reached the target size of 22 participants, matching the size of the external baseline group. To ensure response quality, we additionally excluded insufficient responses in accordance with Prolific's guidelines

[2]https://www.prolific.com/ (last accessed 2026-05-28)

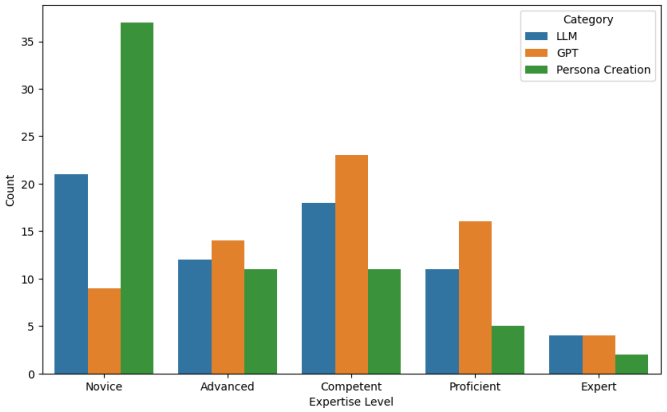

Fig. 1: Expertise levels of participants regarding LLMs, GPT, and Persona Creation

for participant exclusion[3], alongside the predefined screening filters. We explained the survey's purpose to the participants and informed them they could cancel it anytime. The participants self-reported their experience level with LLMs, GPT, and Persona Creation by choosing from "Novice", "Advanced Beginner", "Competent", "Proficient", and "Expert". As shown in Figure 1, the expertise regarding LLMs and ChatGPT was mixed, while the expertise regarding persona creation was mostly on the novice level, with some participants having more advanced expertise. This reflects a population of distributed LLM users who are mostly new to persona creation. A population that serves the purpose of evaluating the potential of guidelines and system prompts to empower non-expert users.

2) Task: While the task of generating personas for a chat app serves as the within-factor in our study, the interfaces used to solve the task varied. This makes the support strategies used the main independent variable in our study. All participants were introduced to the scenario first. We explained to the participants that they were asked to generate personas for a chat app called Chatter, which would be used by many different people. We defined what a persona is to ensure that both novices and others understand the concept. We show the study design with the separate groups in Figure 2. Since the support strategies are the independent variable in this study, we describe the three groups and the current support strategies for each group in the following.

a) Guidelines: In the first group, we introduced participants to the guidelines for generating personas outlined by our previous work [10]. The guidelines presented to the participants are listed in Table I. The interface used by this group listed the guidelines and asked participants to follow them. Guidelines were presetned using a guideline heading that explains the broader goal of the guideline, and an explanation on how to achieve the goal. Note that we designed the interface the same way as in the related study to avoid potential effects from other design differences beyond the guidelines.

b) System Prompt: We forwarded another separate group to an interface that included a hidden system prompt. We did not mention any guidelines. The system prompt itself was designed based on the related literature and the prompt cookbook provided by OpenAI[4]. The prompt aims to manipulate GPT to follow the guidelines itself and was formulated as shown in Listing 1.

c) Guidelines and System Prompt: In a last separate group, we combined the two treatments. The group was asked to use the guidelines when creating the personas, but also used a system-prompted interface to follow them.

3) Procedure: The study was conducted via Prolific, and the survey itself was carried out on LimeSurvey[5], which is hosted by our research institute. The surveys for all groups contained the same information and questions. The surveys differed only in whether guidelines were provided. All participants were first introduced to the methodology of personas and their purpose. Furthermore, it was clarified that ChatGPT, as a generative AI interface provided by us, would be used for the survey. This ensures that all participants use the same version and do not have to pay for their own license. For this purpose, the LLM 'gpt-4o-2024-11-20' was integrated into the interface via the API interface. Participants then gave their informed consent.

To help participants better understand the task, a scenario was provided. The scenario described the development of a new smartphone chat app called "Chatter" and asked participants to design potential target users as personas. To ensure participants were aware of the task, a validation question was included. Then, we presented the actual task. The participants were given the link and password for the interface. In the interface, they could then complete their respective tasks and download the conversation history as an HTML file, which was then uploaded to the survey. To capture participants' natural interaction behavior without introducing learning effects through the study design, no trial round was conducted beforehand. Participants decided themselves when the generated personas were finalized and submitted the downloaded chat protocol once they considered the task complete.

After completing the task, a demographic questionnaire was administered to collect data on age, occupation, highest level of education, self-identified gender, and unique Prolific ID. Following this, participants were requested to provide information about their experience with LLMs, ChatGPT, and persona creation. They were also asked how much they agreed with the statement "Diversity is important for persona creation" and how they define diversity for themselves.

4) Apparatus: One of the most important considerations when using LLMs in research is maintaining reproducibility and comparability. For this reason, ChatGPT from OpenAI was integrated into our generative AI interface via its API. This ensured that the same version was used by all participants. To maintain comparability with our baseline study [10], the latest version 'gpt-4o-2024-11-20' from gpt-4o [6] was used. This model has a knowledge base up to October 1, 2023. The interface was developed using the open-source Python framework 'Gradio' [7]. This framework was then hosted on Hugging Face to make it

[3]https://researcher-help.prolific.com/en/articles/445218-who-should-i-reject (last accessed 2026-05-28)

[4]https://cookbook.openai.com/examples/gpt4-1_prompting_guide (last accessed 2026-05-28)

[5]https://www.limesurvey.org/de (last accessed 2026-05-28)

[6]https://platform.openai.com/docs/models/gpt-4o (last accessed 2026-05-28)

[7]https://www.gradio.app (last accessed 2026-05-28)

TABLE I: Guidelines shown to participants for Creating Diverse Personas derived from our previous work [10].

Persona Creation Guideline Heading	Explanation
Broader Range of Diversity	Do not restrict to a limited set of diversity aspects. GPT can add more diverse aspects if there are no restrictions on certain aspects. Words such as "different" can lead to extra diversity.
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 in the personas.
Less Overly Positive Personas	Tell the LLM to include struggles, flaws, or pain points.
Less Stereotypical Personas	Provide context for the personas, and add information in the prompt that helps the LLM generate personas for your use case. Do not just ask for personas. The more specific the prompt, the less generic the personas are.
More Complex Diversity	Explicitly mention more complex diversity. GPT tends to generate only binary gender-diverse personas.
Personas with Rich Content	Tell the LLM to flesh out the personas' content. GPT tends to generate short persona descriptions that are not necessarily helpful for development. Additional attributes can help steer the focus of the information.

System Prompt: You are an expert in user-centered design and persona creation. Your task is to help create personas for a smartphone chat app called "Chatter". These personas should help to understand the different needs and backgrounds of potential users and make them as diverse as possible. Consider aspects such as gender, age, and ethnicity, as well as as many other relevant factors as possible. Diversity characteristics should be clearly described for each persona. Avoid overly positive or idealized profiles. Take into account personal problems and weaknesses in daily life, as well as limitations or pain points that could affect the potential user's use of the app. The personas should also include contextual details, such as their environment, goals, and communication behavior. Ensure that representations such as gender are not limited to binary categories. Consider more complex forms of diversity where applicable. Create detailed descriptions. Each persona should include categories such as background information, motivations, strategies, and relevant personal characteristics.

Listing 1: The system prompt we used to instruct the LLM to follow the guidelines.

available to participants. Hugging Face provides free hosting, requiring participants to provide only a link and eliminating the need for installation or registration. With this the apparatus is designed to work the same ways as the interface that was used in our baseline.

For each condition, a separate interface was designed for the study. All three include a login screen that requires a group-specific password. Participants then have access to the chat interface. This is very similar in structure to other chatbots such as ChatGPT. In addition, two of the three interfaces display the guidelines again, so participants have them readily available while solving the task. Furthermore, two of the interfaces have an integrated system prompt that is transmitted before the first prompt without the participants' knowledge. Another important function of the interface is downloading the chat history as an HTML file after entering the unique Prolific ID. The Prolific ID was used to assign the files to the participants afterward. This setup ensured that all participants interacted with the same ChatGPT version and received the same information in their groups' interfaces. Figure 3 shows the interface of group B that saw the guidelines and had a system prompt running in the back.

IV. Results

This study uses a mixed-method approach to evaluate the data. First, we employed a qualitative approach to label the data.

A. Qualitative Analysis

Following the exact same approach as used in our baseline study [10], we used thematic analysis [41]. Exactly as in the external baseline we labeled the prompting style, prompt content, diversity aspects in prompts, and the resulting diversity aspects in the personas using the same labels as in the previous study. While we started with the labels from our baseline study, we inductively introduced new labels when appropriate. Similar to the previous study three researchers in our research group separately analyzed the data and then met iteratively to review all personas and agree on the labels in regards of both whether the labels are given correctly and whether new labels are appropriate. Consequently the researchers had to revisit all protocols whenever a label was added to ensure the agreement on the given labels to each protocol. While this creates a total agreement between all researchers on each set of labels aiming to ensure that dimensions that result from different awareness perspectives are reflected as good as possible. Additionally, we collected the presence of common structural aspects in personas, such as age, gender, occupation, etc., and whether the structure of the personas differs between the personas in the set.

a) Prompting Strategy: To ensure comparability to our baseline study [10]the prompt strategies in the chat protocols were labeled in two dimensions. We described the approach to prompting frequency for a persona as either "oneshot" or "iterative", indicating whether a user used a single prompt or multiple prompts. The second dimension described how many personas were targeted by prompts. We coded for "all personas", "multiple personas", or "per persona" prompts. A prompt strategy for one protocol could follow a single strategy, but also rely on multiple strategies at the same time. A user could start with one-shot prompting for all personas, then refine some of the personas iteratively, one at a time. Additionally, we recognized that some users used an agent-based prompting style, in which they first told the LLM that it would serve as an agent for creating personas. Combining the two dimensions and the additional "agent" category, we collected the following strategies as shown in Table II. In our study, both groups that used a system-prompted interface for their interaction with the LLM had a high count for "Per Persona" prompting.

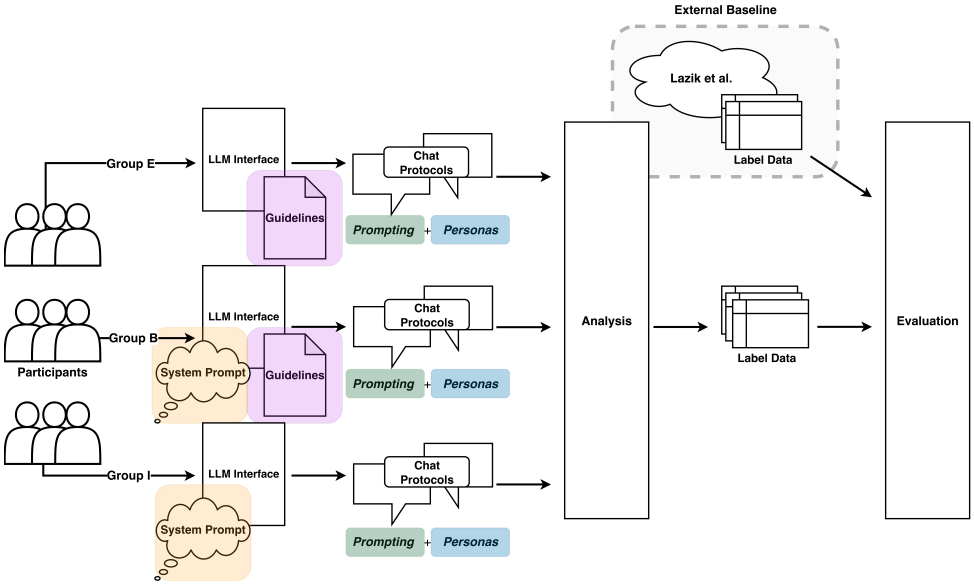

Fig. 2: The study design used in this study. We had three groups, each using a different interface to interact with the LLM. Group E used an interface that listed guidelines; Group I used an interface that was system prompted to follow the guidelines without the guidelines being shown to the users; and Group B used an interface that combined guidelines and a system prompt. We collected participants' chat protocols consisting of the participants prompts and the resulting personas and analyzed them. Our analysis produced labeled data that we evaluated and compared with the external baseline [10].

TABLE II: Prompting Strategies used by participants compared between the conditions: guidelines, system prompt, and both. Participants sometimes used multiple strategies for the same task.

Prompting Strategy	Guidelines	System Prompt	Guidelines + System Prompt
All Personas **one-shot**	4	1	4
All Personas **iterative**	7	7	4
Per Persona **one-shot**	8	14	13
Per Persona **iterative**	7	10	10
Multiple Personas **one-shot**	0	2	2
Multiple Personas **iterative**	0	0	3
Agent	1	0	2

b) Prompt Content: Regarding the prompt content, we used four labels that describe the kind of content prompts included for users. "Task", "Personal Details", "Structural Details", and "Diversity Aspects" aim to describe the prompts' focus, allowing analysis of the information a user provided to the LLM. The meaning of the codings is explained in Table III. The prompt content in combination with the prompting strategies provides inside on the users' prompting behavior. While the strategy aims to capture on what level participants tried to edit the persona set or its subset, the prompt content shows what the users actually included in the prompts.

c) Diversity Aspects in Prompts: Whenever a user's prompt contained information related to diversity, we described the aspects they mentioned using labels. This information can then provide insights into the context the LLM received and how it affected the resulting personas. We aimed to label both clear mentions of aspects such as age or gender, as well as more subtle ones, such as motivation, especially when the user did not directly name the aspect but rather described it as "a young person". Additionally, we coded the presence of information related to diversity itself, such as "diverse persons" or "personas that are different from each other". We describe the labels in Table IV. We are aware that we may not have fully identified all aspects of diversity in the prompts, since awareness of a diversity aspect is required to identify it. However, the three coding researchers collected all aspects to the best of their knowledge based on their own context and the related literature.

TABLE III: The Labels used to capture the Prompt Content by Participants and their explanation for applying them in the qualitative analysis.

Label	Description
Task	The participants' prompt contained information regarding the task of creating a chat app ("Chatter").
Personal Details	The participant defined specific information about personas (e.g. "generate a 20-year-old person").
Structural Details	The participant defined structural aspects of the personas (e.g. "Personas should contain their age").
Diversity Aspects	The participant's prompt contained diversity aspects or acknowledged that the personas should be diverse.

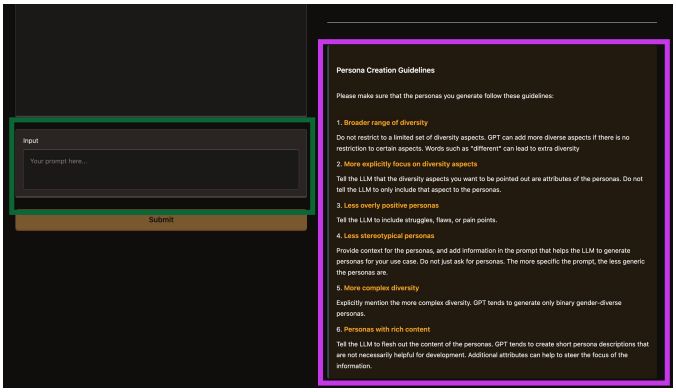

Fig. 3: Screenshot of the LLM interface with the guidelines shown in Table I. The screenshots shows the text field for the user prompts on the left and the guidelines on the right.

We open our data to the community to enable the identification of additional aspects to consider.

d) Diversity in Personas: To analyze the current diversity of the created personas, we labeled them according to the chat protocol. Some user protocols contained personas that were changed over several iterations, resulting in a final set of personas. The researchers in this work then considered the final set and evaluated whether it is diverse with respect to the specific aspect shown in Table IV. Here, a set of personas with a similar age group would not be considered age-diverse; in contrast, a set of personas with a broad range of young, old, and in-between individuals would be considered age-diverse. If a diversity aspect describes an attribute not commonly shared across personas, such as neurodiversity or different impairments, we considered the persona set diverse regarding that aspect if it was represented in the set. As well as the prompt diversity, the diversity in the persona could include aspects we might not be aware of during the coding process. However, the three coding researchers included as many aspects as possible, based on their best knowledge.

e) Followed Guidelines: We additionally coded whether participants seemed to follow the guidelines. This analysis aims to understand two major aspects. First, we wanted to check whether the participants followed the guidelines as intended. Second, we also wanted to understand how the participants interpreted the guidelines we gave to them, revealing potential issues of presenting guidelines to novice users in LLM interfaces.

f) Diversity Definition: The demographic section of the survey also asked for the definition of diversity. To understand emerging themes, we also analyzed the definitions. In our analysis, we identified three main themes that reflect the participants' definitions. "Abstract Definition" refers to definitions that describe diversity as people being different in general, without specific aspects. "Concrete Aspects" describes definitions that specify diversity based on specific aspects such as age, gender, ethnicity, etc.. "Representation" refers to definitions that explicitly mention representation of underrepresented groups or "everyone". "Other" describes such definitions that do not describe any of the aforementioned aspects. Table V shows the count of each of the present themes. Highlighting that most of the participants either chose to define diversity with concrete aspects or in an abstract way.

B. Qualitative Results

While analyzing the chat protocols, we noticed some details that are on a higher level of abstraction than labels or themes. In the following, we report these insights. The protocols are provided in the paper's supplementary material.

a) The Power of the System Prompt: Participants in the condition with the system prompt sometimes started the chat with the interface using just a simple phrase at the beginning, "hi" (PID: I50). The system prompt then clearly answered:

> "Hello! How can I assist you today? Would you like me to begin creating the personas for the "Chatter" app? Let me know how you'd like to proceed.", (PID: I50)

With a simple 'yes', the interface then created three personas leading the creation process. In the condition that had both the guidelines and the system, the interface even created five personas that follow the standard we aimed for, just by the prompt:

> "Fair enough. Worth a punt though", (PID: B35)

b) Misunderstood Guidelines: While coding whether participants in the conditions with guidelines followed the guidelines we provided, we noticed that all participants showed an intention to follow them. However, one assumed that the system was aware of those guidelines, even though they were in a state without any system prompt and thus without any description to the LLM of what the guidelines are.

> "Hi, I need you to help me create five user personas for a smartphone chat app called Chatter. Please follow these Persona Creation Guidelines", (PID E64)

Additionally, some participants misunderstood the guidelines from our perspective. While most participants understood the guidelines as intended by us, some did not prompt as we expected.

c) Characters instead of Personas: Sometimes participants were prompted with a rather than a classical persona. In the condition with guidelines, such prompts led the LLM to continue the story without generating a classical persona. While those characters also had many diverse aspects introduced by the participant "Focus more on the diversity aspects of Jane" (PID: E47), they do not serve the purpose of the persona to describe a potential user's needs, interests, or goals. However, similar prompting styles yielded a proper set of well-described personas when the system prompt was used. The prompting strategies presented in Table II show that many participants used "Per Persona" prompting. In combination with the character description-style prompts, the results suggest that the system prompt could have empowered users to work on a single persona at a time.

d) Cutting off Content: We noticed that participants prompted the LLM to shorten descriptions or remove specific attributes as structural features of the personas. By that, they created either short or one-paragraph personas.

> "We don't need as much detail. Why not break it down into a 70-word description and 2 key bullets to be aware of", (PID: B46)

While this ignores the guideline that calls for rich content, it shows that the participants were able to override the system prompt

TABLE IV: Labels for the diversity aspects present in the prompts and personas and their explanation for applying them in the qualitative analysis.

Label	Description
Just named Diversity	Acknowledgement of diversity without elaboration (e.g., diverse, different).
Age	The age range or specific age of the persona.
Gender	The gender identity of the persona.
Ethnicity	The ethnic background or heritage of the persona.
Neurodiversity	Cognitive differences such as autism, ADHD, dyslexia, or other neurological variations.
Personality	Character traits and tendencies (e.g., introverted, extroverted, open-minded, cautious).
Occupation	The persona's profession, role, or employment status.
Interests	Hobbies, passions, or areas of curiosity.
Motivation	The underlying goals, needs, or motivations to use the chat app.
Religion	Religious beliefs or practices.
Country of Origin	The country where the persona was born or primarily identifies with.
Sexual Orientation	The persona's sexual orientation (e.g., heterosexual, homosexual, bisexual, asexual).
Attitude to Technology	The persona's comfort level, openness, and approach to adopting and using technology.
Family Status / Household	Living situation, marital status, and family dynamics (e.g., single, married, parent, widowed).
Health / Impairments	Physical or mental health conditions, impairments, or disabilities affecting daily life.
Socio-Economic Background	Financial resources, and class background.
Location	The geographic place of residence.
Language	The primary language(s) spoken.
Culture	The persona's cultural background.
Appearance	The overall presentation of the persona, including their physical look, body shape, style, voice, and other traits that influence how they are perceived by others.

TABLE V: Number of aspects in diversity definitions.

Diversity Definition	Guidelines	System Prompt	Guidelines + System Prompt
Abstract Definition	11	13	10
Concrete Aspects	9	8	10
Representation	3	4	4
Other	1	1	1

that generated rich, structured personas before the participants told it to do otherwise.

e) Age-related Attitude to Technology: Looking at the generated personas, it was noticeable that a representation of a less tech-savvy individual was often also associated with an older age group. This frequent combination of traits was also sometimes intentionally prompted by participants.

> "Lets focus on an elderly man next. He is not so good with technology so needs things be simplistic. He has teenage grandchildren who are all about their mobiles. He loves to socialise and he is from an asian background. He is also retired", (PID: B37)

But also, we noticed a participant who explicitly criticized the LLM because it assumed that a less tech-savvy individual is from an older age group.

> "I think you have made unhelpful age-based assumptions. My experience is that generally people under 30 who didn't learn programming at college have a harder time learning about how to use apps and tech. Please revise the persona and make it less patronising about over 60s.", (PID: B29)

C. Quantitative Analysis

We analyzed the count of diversity aspects using descriptive analysis and a two-way Aligned Rank Transform (ART) ANOVA [42] with a mixed design: *Condition* (Guidelines + System Prompt, Guidelines, System Prompt, Baseline) as a between-subjects factor and *Source* (Prompts, Personas) as a within-subjects factor. Since we are working with count data, we decided to use nonparametric tests. Figure 4 shows an overview of the results. The baseline used in this study was conducted nine months before we sampled our data. While we ensured that the Prolific screening filters, Apparatus, and Setup was the same as in the baseline the time in between the two conducted study may still introduce confounding variables that we cannot control. To provide a sound analysis but show the potential impact of the supporting strategies we conducted the quantitative analysis with and without the baseline from our previous paper [10].

a) Descriptive Analysis of Diversity Counts: We begin by analyzing the descriptive diversity of user prompts. We measured the lowest number of diversity counts in the Baseline condition ($\bar{x}=3.31$, $s=2.81$), followed by the System Prompt condition ($\bar{x}=6.65$, $s=3.17$), the Guidelines + System Prompt condition ($\bar{x}=7.27$, $s=4.23$), and Guidelines condition ($\bar{x}=8.55$, $s=4.50$). Then we continued analyzing the diversity of the resulting personas. Here, we found the lowest number of diversity aspects in the Baseline condition ($\bar{x}=4.77$, $s=1.17$), followed by the System Prompt condition ($\bar{x}=8.75$, $s=1.68$), Guidelines condition ($\bar{x}=8.86$, $s=2.36$), and Guidelines + System Prompt condition ($\bar{x}=9.86$, $s=2.38$).

b) Number of Diversity Aspects in Prompts and Personas analyzed without Baseline: A two-way ART ANOVA revealed a significant main effect of *Source*, $F(1,61)=12.292$, $p<.001$, but no significant main effect of *Condition*, $F(2,61)=0.817$, $p=.447$. The *Condition* × *Source* interaction did not reach significance, $F(2,61)=2.802$, $p=.068$. All post-hoc p-values are Holm-adjusted. Within-condition contrasts compare Prompts minus Personas (negative estimates indicate Personas > Prompts on the aligned scale). Personas contained significantly more aspects than Prompts in the *Guidelines + System Prompt* condition (Estimate $=-19.2$, SE $=9.20$, $t(61)=-2.083$, $p_{\text{Holm}}=.041$) and the *System*

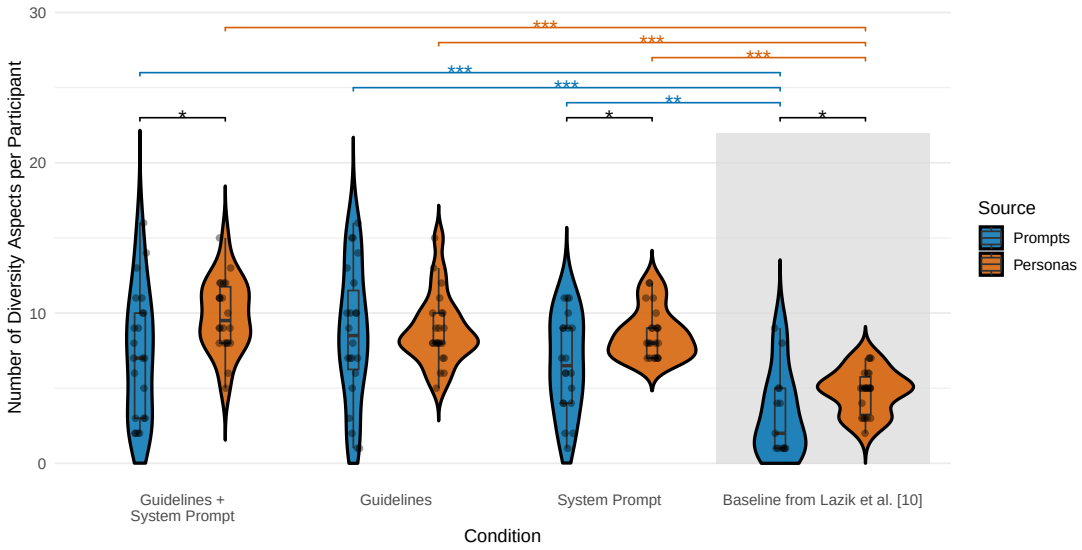

Fig. 4: The plot shows the number of diversity aspects in the user prompts and resulting persona. Using Guidelines + System Prompt resulted in the highest number of diversity aspects in the resulting personas, while the diversity aspects between the conditions remained similar, except for the Baseline condition. Consequently, the Baseline condition elicited the fewest diversity aspects in the prompts and the generated persona. The bar in the violin plots denotes the median. The baseline was conducted nine months before the other three conditions [10]. We marked it to underline the fact that the ART ANOVA might be affected by the time in between.

Prompt condition (Estimate $= -22.0$, SE $= 9.65$, $t(61) = -2.281$, $p_{\text{Holm}} = .026$); the difference in the *Guidelines* condition trended in the same direction but was not significant (Estimate $= -15.8$, SE $= 9.20$, $t(61) = -1.722$, $p_{\text{Holm}} = .090$). No significant between-condition differences were found within either Source.

c) Additional Analysis of the Number of Diversity Aspects in Prompts and Personas: As an addition, we included the baseline in the ANOVA to analyze the potential effect of the support strategies. Note that the baseline was conducted 9 months before our study. The two-way ART ANOVA revealed a significant main effect of *Condition*, $F(3, 73) = 12.761$, $p < .001$, and *Source*, $F(1, 73) = 18.482$, $p < .001$. However, the *Condition* $\times$ *Source* interaction was not significant, $F(3, 73) = 1.967$, $p = .126$. Because the interaction was not significant, we report pairwise post hoc tests for completeness but focus on the main effects of *Condition* and *Source* for interpretation. All post-hoc p-values are Holm-adjusted. Within-condition contrasts compare Prompts minus Personas (negative estimates indicate Personas $>$ Prompts on the aligned scale). Personas contained significantly more aspects than Prompts in *Guidelines + System Prompt* (Estimate $= -23.3$, SE $= 10.9$, $t(73) = -2.134$, $p_{\text{Holm}} = .0362$), *System Prompt* (Estimate $= -27.4$, SE $= 11.5$, $t(73) = -2.393$, $p_{\text{Holm}} = .0193$), and *Baseline* (Estimate $= -33.6$, SE $= 14.2$, $t(73) = -2.363$, $p_{\text{Holm}} = .0208$); the difference in *Guidelines* trended in the same direction but was not significant (Estimate $= -19.8$, SE $= 10.9$, $t(73) = -1.810$, $p_{\text{Holm}} = .0744$).

V. Threats to Validity

A. Threats to Internal Validity

Potential threats to internal validity stem from possible selection bias due to recruiting participants via the crowd-working platform Prolific. To ensure data quality, we included attention and comprehension checks to verify that participants understood the task. We also applied Prolific's rejection guidelines to exclude low-quality responses and restricted participation to workers with an approval rate of at least 98%. Additionally, pilot studies were conducted to ensure that the survey and interfaces were free of technical or procedural errors that could affect the results.

The authors are committed to advancing diversity, inclusion, and intersectionality in research and practice. While this commitment may introduce potential bias, the researchers approached the data analysis as objectively and neutrally as possible.

Our external baseline study [10] was conducted at another point of time as the study presented in this work. While we ensured that the setup, apparatus, and Prolific Screening Filters are exactly the same, the time of nine months might introduce confounding factors that we cannot control for. The time window could affect the general population of Prolific in such things as their perception of LLMs or how they approach this technology. While we acknowledge such and similar effects, we additionally analyzed the data in comparison with the baseline using the ART ANOVA to provide a deeper insight on the potential effect of the support strategies. However, we do not use the statistical significance of the additional ART ANOVA with the external baseline as to answers our research questions.

B. Threats to Construct Validity

Research on diversity is inherently limited by researchers' awareness of its many dimensions; unrecognized aspects cannot be analyzed, creating the risk of blind spots. To mitigate this risk and develop a coding scheme that reflects diverse perspectives, all researchers independently conducted an inductive analysis of the

protocols before jointly discussing their findings. This procedure ensured that each perspective informed the coding without being restricted by predefined categories. The resulting labels were developed to the best of our knowledge and are transparently documented in the supplementary material. Because the labeling of the protocols was refined through iterative consensus meetings to avoid overlooking relevant dimensions, we did not compute a traditional inter-rater agreement score, as the labels were not fixed prior to discussion, and, as part of our method fully agreed afterwards. All researchers fully agreed on the coding of each aspect and thus the inter-rator agreement is per desgin 100%. This follows the exact same approach as in our external baseline [10]. Since the question whether users seem to have followed the guidelines is especially highly subjective and defined by the perception of the researcher coding the protocol, the analysis of this construct should be treated as such. However, this work aims to contribute exploratory insights to guidelines for LLM usage and needs some kind of possibility to express their effect. Thus we provide our insights to the best of our knowledge.

C. Threats to External Validity

This study investigates the potential of prompting guidelines compared to system prompts in supporting users during LLM interactions. Participants were recruited via Prolific and largely consisted of novices in persona creation. Although this group does not represent domain experts, it enables us to assess how guidelines and system prompts support laypersons who may be unfamiliar with best practices described in the literature.

All participants rated the importance of diversity in persona creation as at least neutral on a 7-point Likert scale. Thus, the sample does not reflect individuals who explicitly reject the relevance of diversity. It also remains unclear whether participants who hold such views would openly report them in a survey setting.

System prompts can substantially influence LLM outputs, which introduces a potential threat to validity if prompts are designed to steer results toward expected outcomes. In this study, however, this influence was part of the research question and therefore intentional. The guideline-only condition demonstrated that effects were not solely driven by the system prompt, and our results further show that participants were able to override it. Nevertheless, both the guidelines and the system prompt were derived from the same metrics that we later used to evaluate output quality. This indicates that both support strategies were effective in promoting the targeted quality characteristics. However, our work does not demonstrate equivalent performance across all possible dimensions of diversity. As discussed earlier, we aimed to assess diversity as comprehensively and carefully as possible based on the current state of knowledge.

Finally, the evolving nature of LLMs poses a general threat to external validity, as model outputs may vary across versions. To support transparency and reproducibility, we report the exact model version used and provide the results generated in this study. We used this exact model to be comparable with the external baseline of the study. The discussion of our results aims to take away not only the state of the exact model under use but rather general insights on approaching human-LLM interaction.

VI. Discussion

To guide the discussion of this work, we structure the discussion section around our research questions. Since the research questions focus on the two support strategies for LLM users, we also discuss the overall findings afterward.

A. RQ1: To What Extent Do Guidelines Affect Persona Generation?

Our first research question aims to analyze the impact of the present guidelines on an LLM interface. Regarding the aspect of the users' prompts investigated by

a) RQ1.1: How do users change their prompting behavior for generating personas based on provided guidelines?: we showed an increase in the diversity aspects in prompts compared to the baseline [10]. Thus, the support strategies seem to help users prompt consideration of more diverse aspects.

However, our qualitative insights show that participants sometimes interpreted the guidelines differently than intended, leading to unexpected prompting behavior and, at times, unexpected output. Furthermore, participants assumed the system was aware of the guidelines even though it was not. We did not include active training sessions or instructional material in the user study. By intentionally refraining from demonstrating how the guidelines were meant to be applied, we were able to capture potential mismatches between our intended use of the guidelines and participants' actual interpretations and applications. While the results overall clearly show the effect of the guidelines we also identified their downside. This points out that such guidelines need meaningful explanations and a description of the LLM's capabilities regarding them. Authors of guidelines need to consider many potential perspectives to ensure guidelines that are understandable and less likely to be misinterpreted.

b) RQ1.2: To what extent does the diversity in generated personas change if users get explicit guidelines for enhancing diversity?: Regarding the diversity of the personas, we observed a difference between the guideline-assisted groups and the data reported in our baseline study [10]. Thus, participants who followed the guidelines achieved more diverse personas. The personas generated using the guidelines were rich in content and included many needs and requirements. Compared to the aforementioned study, personas (from our study) with many diverse aspects explained how these aspects lead to app needs. While our previous paper criticized the lack of such requirements [10], using our suggestion to prompt for rich persona content helped address this issue. We answer **RQ1** in the following:

> Answering **RQ1**, we show that participants with guidelines created more diverse personas through prompts containing more diversity aspects. The present guidelines guided their prompts. Sometimes the guidelines were misunderstood, leading to slightly different prompting behavior than expected.

B. RQ2: To What Extent Does a System Prompt Affect the Diversity in Persona Generation?

Our second research question aims to understand the effect of a system prompt as a supporting tool for LLM users. To understand this effect, we first discuss our question

a) RQ2.1: To what extent does the prompting behavior of a user affect the output of an LLM interface with a system prompt for diverse persona generation?: In our quantitative analysis, we showed a higher amount of diversity aspects in prompts with an interface that had a system prompt. Since the compared baseline only asked participants to generate personas, a chat interface that includes many diversity aspects seems to prompt participants to ask for more. This is shown because even the "system-prompt-only" group considered more diverse aspects than the baseline in the other study [10]. Nevertheless, participants tend to overwrite the system prompt, resulting in output that differs from expectations.

b) RQ2.2: To What Extent Does the Diversity in Generated Personas Change if the Used LLM Gets a Specific System Prompt?: In this study, the system prompt led to more diverse aspects compared to an interface without any guidelines or system prompts. Thus, the system prompt served its purpose. However, in case the participant just wrote unrelated prompts (e.g. "Hi", PID: I50) that would normally never lead to personas, the interface sometimes even created five diverse personas just by being prompted with "Fair enough. Worth a punt though" (PID: B35). Since we wrote the system prompt based on guidelines derived from a study analyzing prompt protocols for diversity aspects, the prompt was optimized to the measures used in this study. Without being visible, especially in conversations with participants, who right away focused on the persona creation, the system prompt could have produced better output without the user's actual impact. Because we showed that users could overwrite the prompts and that a group instructed only by guidelines produced similar effects, our findings demonstrate that we did not intentionally design a system prompt to generate false results. Nevertheless, GPT-4o's behavior in our study suggests a major threat to validity in LLM studies. Researchers who do not openly show their interface can manipulate results without a clear clue in the LLM's output.

Therefore, we can answer RQ2 with:

> Answering **RQ2**, our results show that participants with a system prompt created more diverse personas through prompts containing more diversity aspects. The output of the LLM seems to support the users by also leading to a chat protocol with more diversity aspects in prompts. Nevertheless, while some users were able to overwrite the system prompt, system prompts can be a potential threat to a study's validity if they are not openly communicated.

C. General Discussion

In this study, we investigated two potential support strategies for LLM users. While the results show potential for both strategies, the combination of the two appears to be the optimal solution. When participants do not fully understand the guidelines, the system prompt guides the conversation. Even though a system prompt is a powerful tool for shaping an LLM's output, participants who do not know how to interact with the LLM can accidentally overwrite it, leading to unexpected output. In such a case, guidelines can help users achieve a higher level of exposition, with a greater potential for better results.

The personas generated in this study were rich in content. Compared to the data presented in our baseline study [10], the personas in this newer study were not merely diverse demographic lists but had actual implications for requirements based on their aspects. However, GPT-4o often associates older adults as less tech savvy, which is a harmful stereotype [43] that contradicts current reality. Today's 60-plus population includes people who have used computers throughout their professional and personal life and adapted to decades of technological change [44], [45], yet this assumption persists in both society and computing [43]. When LLMs reproduce these biases, they transform social prejudices into seemingly objective design requirements, creating a critical cycle: biased personas lead to exclusionary designs, which reinforce stereotypes and validate the original biases [46]. The result is personas that appear diverse demographically but are homogenized by the ageism stereotype. This undermines a persona-based design mentality, as if the tools used to generate diverse user representations are systematically biased, and that inclusive design attempts are built on flawed foundations. Hence, meaningful diversity requires more than varying demographics; it requires a holistic approach to address the underlying assumptions. This essentially questions the use of LLM-based personas for diverse representation and reveals an ongoing gap in the need for tools that actually represent the underrepresented.

VII. Conclusion

In this study, we investigated two potential support strategies for Human-LLM interaction. This study specifically focuses on the use case of persona generation, aiming to produce diverse sets of personas. We compared the impact of directly prompting the user with guidelines to that of a system prompt that instructs the LLM. The results of our study show that both supporting strategies outperformed the baseline, indicating a positive impact. Especially the combination of providing guidelines and using a system prompt seems like the best solution to support an LLM user in prompting. Since this study focused on a specific use case, future work may investigate whether those strategies can also be used in other use cases. Nevertheless, we also showed that system prompts can strongly influence an LLM's output, posing a potential threat to validity if authors do not report their use. Additionally, the personas created in this study had many implications stemming from their diverse aspects, leading to specific needs. While the previous work criticized the absence of such implications [10], the guided, more engineered prompts in this study produced more meaningful results regarding needs. Thus. LLMs can help developers get inspired to include more aspects of diversity. However, the more enriched personas in our study also contained harmful stereotypes that threaten the overall idea of using LLM-based personas to represent underrepresented groups, asking for future work to create better approaches to ensure that user needs are reflected in a diverse but also correct representation that is not propagating stereotypes, harming both the regarding group and the developers who aim to understand their actual needs.

VIII. Data Availability Statement

The data of this work can be found via a DOI hosted on FigShare[8]. Together with the data provided by our previous study [10], the whole study from this paper can be reproduced.

[8]https://doi.org/10.6084/m9.figshare.31386934

References

[1] J. Pruitt and T. Adlin, *The persona lifecycle: keeping people in mind throughout product design*. Elsevier, 2010.

[2] 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, 2021.

[3] 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

[4] 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*. Lisbon Portugal: ACM, Apr. 2024, pp. 24–25.

[5] 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.

[6] 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.

[7] P. N. Venkit, J. Li, Y. Zhou, S. Rajtmajer, and S. Wilson, "A Tale of Two Identities: An Ethical Audit of Human and AI-Crafted Personas," May 2025.

[8] A. Li, H. Chen, H. Namkoong, and T. Peng, "LLM Generated Persona is a Promise with a Catch," Mar. 2025.

[9] C. K. 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?" in *Proceedings of the Mensch und Computer 2025*, ser. MuC '25. Association for Computing Machinery, pp. 434–451. [Online]. Available: https://dl.acm.org/doi/10.1145/3743049.3743057

[10] C. Lazik, C. Kauter, I. Nunes, A. Ziglowski, A. Pryma, C. Katins, L. Grunske, and T. Kosch, "The good, the bad, and the uncanny: Investigating diversity aspects of LLM-generated personas for requirements engineering," artwork Size: 846967 Bytes Pages: 846967 Bytes. [Online]. Available: https://figshare.com/articles/conference_contribution/The_Good_the_Bad_and_the_Uncanny_Investigating_Diversity_Aspects_of_LLM-Generated_Personas_for_Requirements_Engineering/28573244

[11] M. Prpa, G. Troiano, B. Yao, T. J.-J. Li, D. Wang, and H. Gu, "Challenges and opportunities of llm-based synthetic personae and data in hci," in *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, 2024, pp. 716–719.

[12] W. W. Sim and P. S. Brouse, "Empowering requirements engineering activities with personas," *Procedia Computer Science*, vol. 28, pp. 237–246, 2014.

[13] J. Ollerton, "IPAR, an inclusive disability research methodology with accessible analytical tools," *International Practice Development Journal*, vol. 2, no. 2, 2012.

[14] S. Fletcher-Watson, H. De Jaegher, J. Van Dijk, C. Frauenberger, M. Magnée, and J. Ye, "Diversity computing," *Interactions*, vol. 25, no. 5, pp. 28–33, Aug. 2018.

[15] S. Sethi, J. Salminen, D. Amin, and B. J. Jansen, ""When AI Writes Personas": Analyzing Lexical Diversity in LLM-Generated Persona Descriptions," in *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '25. New York, NY, USA: Association for Computing Machinery, Apr. 2025, pp. 1–8.

[16] R. Lukyanenko, D. Bork, V. C. Storey, J. Parsons, and O. Pastor, "Inclusive conceptual modeling: Diversity, equity, involvement, and belonging in conceptual modeling," 2023.

[17] G. Liebel, J. Klünder, R. Hebig, C. Lazik, I. Nunes, I. Graßl, J.-P. Steghöfer, J. Exelmans, J. Oertel, K. Marquardt, K. Juhnke, K. Schneider, L. Gren, L. Happe, M. Herrmann, M. Wyrich, M. Tichy, M. Goulão, R. Wohlrab, R. Kalantari, R. Heinrich, S. Greiner, S. A. Rukmono, S. Chakraborty, S. Abrahão, and V. Amaral, "Human factors in model-driven engineering: future research goals and initiatives for MDE," vol. 23, no. 4, pp. 801–819. [Online]. Available: https://doi.org/10.1007/s10270-024-01188-8

[18] 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. Cham: Springer International Publishing, 2021, vol. 12768, pp. 39–52.

[19] L. Gardenswartz and A. Rowe, *Diverse Teams at Work: Capitalizing on the Power of Diversity*. Society for Human Resource, 2003.

[20] 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*. Glasgow Scotland Uk: ACM, May 2019, pp. 1–16.

[21] I. Nunes, A. Moreira, and J. Araujo, "Gire: Gender-inclusive requirements engineering," *Data & Knowledge Engineering*, vol. 143, p. 102108, 2023.

[22] K. Gama and R. Santos, "It's not all about gender: A Multi-dimensional Course Perspective on Diversity and Inclusion in Software Engineering Education," in *Simpósio Brasileiro de Engenharia de Software (SBES)*. SBC, 2024, pp. 487–498.

[23] V. Gupta, "Requirement Engineering Challenges for Social Sector Software Development: Insights from Multiple Case Studies," *Digital Government: Research and Practice*, vol. 2, no. 4, pp. 1–13, Oct. 2021.

[24] J. Grundy, T. Kanij, J. McIntosh, H. Khalajzadeh, and I. Mueller, "Diverse End User Requirements," Oct. 2022.

[25] L. Okpara, C. Werner, A. Murray, and D. Damian, "A Case Study of Building Shared Understanding of Non-Functional Requirements in a Remote Software Organization," in *2022 IEEE 30th International Requirements Engineering Conference (RE)*. IEEE, 2022, pp. 1–13.

[26] H. Kotek, R. Dockum, and D. Sun, "Gender bias and stereotypes in Large Language Models," in *Proceedings of The ACM Collective Intelligence Conference*. Delft Netherlands: ACM, Nov. 2023, pp. 12–24.

[27] D. Zowghi and M. Bano, "AI for all: Diversity and Inclusion in AI," *AI and Ethics*, vol. 4, no. 4, pp. 873–876, Nov. 2024.

[28] L. Lucy and D. Bamman, "Gender and representation bias in GPT-3 generated stories," in *Proceedings of the Third Workshop on Narrative Understanding*, 2021, pp. 48–55.

[29] L. Sun, M. Wei, Y. Sun, Y. J. Suh, L. Shen, and S. Yang, "Smiling women pitching down: Auditing representational and presentational gender biases in image-generative AI," *Journal of Computer-Mediated Communication*, vol. 29, no. 1, p. zmad045, 2024.

[30] M. Bano, H. Gunatilake, and R. Hoda, "What Does a Software Engineer Look Like? Exploring Societal Stereotypes in LLMs," Jan. 2025.

[31] J. F. Dovidio, M. Hewstone, P. Glick, and V. M. Esses, "Prejudice, stereotyping and discrimination: Theoretical and empirical overview," *Prejudice, stereotyping and discrimination*, vol. 12, pp. 3–28, 2010.

[32] 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

[33] J. Stypinska, "AI ageism: A critical roadmap for studying age discrimination and exclusion in digitalized societies," *AI & SOCIETY*, vol. 38, no. 2, pp. 665–677, Apr. 2023.

[34] A. Rosales and M. Fernández-Ardèvol, "Ageism in the era of digital platforms," *Convergence: The International Journal of Research into New Media Technologies*, vol. 26, no. 5-6, pp. 1074–1087, Dec. 2020.

[35] J. Barambones, C. Moral, A. De Antonio, R. Imbert, L. Martínez-Normand, and E. Villalba-Mora, "ChatGPT for Learning HCI Techniques: A Case Study on Interviews for Personas," *IEEE Transactions on Learning Technologies*, vol. 17, pp. 1460–1475, 2024.

[36] I. Hipólito, K. Winkle, and M. Lie, "Enactive artificial intelligence: Subverting gender norms in human-robot interaction," *Frontiers in Neurorobotics*, vol. 17, Jun. 2023.

[37] G. Marvin, N. Hellen, D. Jjingo, and J. Nakatumba-Nabende, "Prompt engineering in large language models," in *Data Intelligence and Cognitive Informatics*, I. J. Jacob, S. Piramuthu, and P. Falkowski-Gilski, Eds. Springer Nature, pp. 387–402.

[38] L. Giray, "Prompt engineering with ChatGPT: A guide for academic writers," vol. 51, no. 12, pp. 2629–2633. [Online]. Available: https://doi.org/10.1007/s10439-023-03272-4

[39] Q. Ye, M. Axmed, R. Pryzant, and F. Khani, "Prompt engineering a prompt engineer." [Online]. Available: http://arxiv.org/abs/2311.05661

[40] B. Wang, J. Liu, J. Karimnazarov, and N. Thompson, "Task supportive and personalized human-large language model interaction: A user study," in *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*, ser. CHIIR '24. Association for Computing Machinery, pp. 370–375. [Online]. Available: https://dl.acm.org/doi/10.1145/3627508.3638344

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

[42] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins, "The aligned rank transform for nonparametric factorial analyses using only anova procedures," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '11. New York, NY, USA:

Association for Computing Machinery, 2011, p. 143–146. [Online]. Available: https://doi.org/10.1145/1978942.1978963
[43] J. L. Birkland, "How older adult information and communication technology users are impacted by aging stereotypes: A multigenerational perspective," vol. 26, no. 7, pp. 3967–3988, publisher: SAGE Publications. [Online]. Available: https://doi.org/10.1177/14614448221108959
[44] T. N. Friemel, "The digital divide has grown old: Determinants of a digital divide among seniors," *New media & society*, vol. 18, no. 2, pp. 313–331, 2016.
[45] S. Nash, "Older adults and technology: Moving beyond the stereotypes," *Stanford Center of Longevity*, 2019.
[46] A. Walling, "Ageism revisited," *American Family Physician*, vol. 110, no. 1, pp. 87–89, 2024.