

## Article

# Exploring the Feasibility of Generative AI in Persona Research: A Comparative Analysis of Large Language Model-Generated and Human-Crafted Personas in Obesity Research

Urška Smrke <sup>1,\*</sup> , Ana Rehberger <sup>1</sup> , Nejc Plohl <sup>2</sup>  and Izidor Mlakar <sup>1</sup> 

<sup>1</sup> Faculty of Electrical Engineering and Computer Science, University of Maribor, 2000 Maribor, Slovenia; ana.rehberger@um.si (A.R.); izidor.mlakar@um.si (I.M.)

<sup>2</sup> Department of Psychology, Faculty of Arts, University of Maribor, 2000 Maribor, Slovenia; nejc.plohl1@um.si

\* Correspondence: urska.smrke@um.si

**Abstract:** This study investigates the perceptions of Persona descriptions generated using three different large language models (LLMs) and qualitatively developed Personas by an expert panel involved in obesity research. Six different Personas were defined, three from the clinical domain and three from the educational domain. The descriptions of Personas were generated using qualitative methods and the LLMs (i.e., Bard, Llama, and ChatGPT). The perception of the developed Personas was evaluated by experts in the respective fields. The results show that, in general, the perception of Personas did not significantly differ between those generated using LLMs and those qualitatively developed by human experts. This indicates that LLMs have the potential to generate a consistent and valid representation of human stakeholders. The LLM-generated Personas were perceived as believable, relatable, and informative. However, post-hoc comparisons revealed some differences, with descriptions generated using the Bard model being in several Persona descriptions that were evaluated most favorably in terms of empathy, likability, and clarity. This study contributes to the understanding of the potential and challenges of LLM-generated Personas. Although the study focuses on obesity research, it highlights the importance of considering the specific context and the potential issues that researchers should be aware of when using generative AI for generating Personas.

**Keywords:** user personas; obesity; large language models; value sensitive design; digital health interventions



Academic Editor: Jenhui Chen

Received: 2 December 2024

Revised: 5 February 2025

Accepted: 11 February 2025

Published: 13 February 2025

**Citation:** Smrke, U.; Rehberger, A.; Plohl, N.; Mlakar, I. Exploring the Feasibility of Generative AI in Persona Research: A Comparative Analysis of Large Language Model-Generated and Human-Crafted Personas in Obesity Research. *Appl. Sci.* **2025**, *15*, 1937. <https://doi.org/10.3390/app15041937>

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Digital interventions hold significant potential to facilitate behavior change by leveraging technology to deliver personalized, scalable, and accessible solutions [1,2]. They can provide tailored content and real-time feedback, continuously evolve to meet individual needs, and thereby increase motivation and reinforce positive behaviors. However, the main challenge of all digital interventions is low user engagement and adherence [3,4]. Engagement with digital health interventions (DHIs) is influenced by a complex interplay of contextual and dynamic factors [5,6]. Understanding these factors is crucial for developing effective interventions that can improve user engagement and, ultimately, outcomes. Persona-based research can contribute to a better understanding of these factors by capturing the diversity of user characteristics and through user archetypes, identifying how different user attributes may impact participation and motivation [7–9]. Moreover, Personas can inform the design of digital interventions by providing insights into user

preferences, needs, and goals. This knowledge can guide the development of engaging and effective digital interventions [10]. Finally, the Personas can efficiently capture and reveal the technological and environmental context in which users will interact with digital interventions, making digital interventions accessible, relevant, and acceptable to the target population [11,12].

While Personas are one of the most widely used methods of user research, their development, traditionally performed using qualitative methods, is facing the critical challenge of being extremely time-consuming and static [13]. Namely, the process involves multiple steps and requires the involvement of various experts and stakeholders to ensure that the Personas accurately represent the target audience [14]. As such, there is a great need for simpler, quicker, and more cost- and resource-effective ways to create them.

This gap may be addressed by artificial intelligence (AI), which is often described as “being able to think like humans” [15] or “being able to perform tasks that require human intelligence” [16]. Large language models (LLMs) can streamline the creation of high-quality Personas [14] by also automating qualitative data analysis, recognizing patterns, generating and refining Personas, personalizing content, and enhancing efficiency and scalability. Thus, employing LLMs in the Persona creation process can save time and resources, allowing for the development of high-quality Personas without extensive manual effort. However, there are several potential weaknesses and limitations that should be considered regarding LLMs, including their generalizability, data quality, interpretability, explainability, and prompt engineering. To understand the limitations further, this study analyzes and compares the quality of generated Personas using multiple LLMs and those generated by an expert human panel. To ensure replicability and the wider applicability of our research, we only analyze those models that can be accessed openly and for free, i.e., Google’s Bard (now Gemini), OpenAI’s ChatGPT 3.5, and Llama2 70B. Moreover, the prompts used (see Supplementary Materials) are straightforward to prevent bias from expert knowledge or exceptional skills and expertise. We are specifically interested in whether the different Personas will be perceived differently by human experts, and to what extent. Based on this aim, we developed our research question (RQ): “Which large language models are perceived to generate better or worse Personas in comparison with qualitatively developed Personas?”. To the best of our knowledge, no existing study has compared machine and qualitatively developed, human-crafted Personas targeting behavioral change in both the educational and healthcare domains. Moreover, by answering this question, we are, at the same time, reflecting the appeal from Salminen and colleagues [17] for a full comparison of AI-generated Personas with qualitatively developed Personas.

Consequently, the findings of such research can benefit the population of children and adolescents experiencing obesity. As it can provide evidence on the comparability of qualitatively developed and LLM-generated Personas, this can enable DHI technology developers in the future to utilize LLM models to construct relevant Personas for the field of obesity treatment and prevention more efficiently than with (only) traditional methods. This can accelerate the process of value-sensitive design incorporating Personas’ development, as discussed in the following section, and potentially shorten the time needed to develop technologies supporting the population experiencing obesity.

With the aim of answering our research question on the comparison of LLM-generated with qualitatively developed Personas, we structure the present paper as follows: in the next section, we provide the theoretical background relevant to the study, followed by the explanation of the context, sample characteristics, materials used, research procedure, and statistical analyses in the Materials and Methods. In the Results section, we present the results of our research, and in the Discussion section, we provide our interpretation of the study results, limitations, and future directions, and elaborate on the practical implications. Finally, the main findings are synthesized in the Conclusions section.

### *1.1. The Advantages of Value-Sensitive Design in Developing Digital Health Interventions*

The alignment of technology with human values, user needs, and an enhanced user experience is not just a benefit but a necessity in the development of mobile health applications, ensuring that they are effective, ethical, and embraced by users [18–20]. The value-sensitive design framework [21] has become increasingly popular in the development of technologies [18,19,22] as it ensures that the human values are incorporated from the start of the development process [23]. According to Schwartz’s theory of basic values [24], values possessed by an individual provide motivation for their behavior. In line with this, previous research shows that considering users’ values can be a solid entrance point into enhancing their motivation for technology use [18,25]. Value-sensitive design can increase trust and adoption, increase regulatory compliance, and promote the regular use of technology by enhancing positive user experiences and aligning with human values and needs [22,26].

Topics such as the motivation for the use, trust, and adoption of new technologies are important in the domain of DHIs, especially when target users are children. Their evolving ability to self-regulate, which also includes the regulation of behavior [27], and their own and their parents’ concerns about privacy in technology use [28,29] can result in lower technology adoption and consequently attenuate the effectiveness of DHIs. This can be particularly problematic in treatment, where a person’s rehabilitation depends on intervention success. Good user research can help to tailor the intervention to the user’s or patient’s needs and consequently improve engagement rates. For example, in a scoping review on digital technology use in children with chronic conditions [30], the authors reported on the importance of programs involving users in the design process to improve user engagement with the final product.

In obesity research, digital interventions are reported to have greater short-term effects than offline interventions and have the capacity to reach more people [31]. However, despite their popularity among researchers and clinicians, DHIs face high attrition rates among patients [31,32]. Stangeland Lie et al. [32] have found one of the reasons for discontinuing the eHealth intervention to be irrelevant content, and some participants have reported the content of the intervention as not being tailored to their needs and expectations. Similar conclusions were found for participants’ engagement rates, which are another vital precondition of intervention effectiveness; Yardley et al. [33], for example, state that engagement in digital behavioral change interventions is specifically promoted by interventions being tailored to individual needs. This is why the adoption of value-sensitive design is required to consider the needs and values of the target users.

### *1.2. Using Personas to Understand the Target Audience*

One of the methods that can be used to create representations of the users (i.e., their needs, requirements, and expectations) is the Persona development. Salminen et al. [34] define Personas as imaginary persons describing a user group, which are typically used in the exploration phases of product development, when developers are researching who their

users are. Personas serve the purpose of keeping the focus on the relevant user group that will potentially use the product [35]. The development of Personas as a user segmentation technique is used in various fields, such as design, marketing, software development, and health informatics [34]. In this research, we focus on Personas that are relevant to DHIs dedicated to obesity management and prevention. To design effective digital interventions for children and adolescent obesity management and prevention, the most relevant Personas are those representing stakeholders who have a direct interaction with adolescents and influence their health and lifestyle choices [36]. In particular, understanding the needs and preferences of adolescents who are already dealing with obesity or overweight issues, and those who are generally healthy but should be part of obesity prevention programs, can help in designing interventions that are more relatable and effective for them and accessible and engaging for a broader range of children and adolescents. The healthcare professionals, who are focused on helping adolescents with obesity or overweight issues (i.e., dietitians and pediatricians), are well-positioned to provide expert input on the development of digital interventions that can effectively support adolescents in managing their weight and health. Input from teachers can be valuable in designing interventions that are integrated into school settings and can effectively support students in maintaining healthy lifestyles. Finally, the administrative leaders (i.e., principals) can provide insight into how digital interventions can be effectively implemented and integrated into broader societal settings, such as school settings to support community-based obesity prevention.

Persona development is usually based on qualitative methods of data collection and analyzing techniques [37]. Data collection methods, for example, include interviews, surveys, and focus groups [35]. Despite the qualitative Persona approach being very beneficial regarding the depth and complexity of Persona descriptions [37], experts who are usually involved in the Persona development process [38] are facing several challenges. Qualitative Persona development is known to be time-consuming [13,37,38], subjective [13,37,38], questionably representative of the relevant user groups [13,37,39], and static [13,35], meaning that while users and their needs change over time, so should the respective Persona descriptions. In this fast-changing world, these changes might come quickly and often, which puts an extra burden on user research teams. In addition, for in-person activities, extra resources are spent on recruiting appropriate participants, handling the logistics of organization, and having enough space to accommodate participants for group activities (e.g., focus groups) [35]. In summary, qualitative Persona creation requires many experts and potential end users to be involved in the process [35] and still does not reliably ensure various socio-demographic and cultural aspects.

### *1.3. Large Language Models and Their Potential for Persona Development*

An LLM is a deep learning algorithm that can perform a variety of natural language processing (NLP) tasks [15]. It uses transformer models and is trained using massive datasets, enabling it to recognize, translate, predict, or generate text or other content [40]. With the advancement of AI, models such as ChatGPT, Bard, and Llama have surfaced. These models are especially relevant due to their advanced capabilities in NLP and their widespread adoption in various applications, including customer service, research, and education [41].

As LLMs can mimic human conversation, some authors have started testing to what extent the LLM-produced texts are similar to human-produced texts. For example, Hämäläinen et al. [42] compared text generated by LLM (GPT-3) [43] with real human responses on open-ended question responses related to a specific topic. Their study showed that participants could not distinguish between artificially generated responses and human responses and concluded that AI tools can create plausible responses that highly resemble real human responses. Another study that explored the human perception of Personas generated using LLMs (GPT-4) [43] reported that participants assessed artificially generated Personas to be of good quality [17].

One of the most important benefits, which is relevant to the context of the present article, is the efficiency that LLMs bring, which saves time and resources for development teams and researchers [15], as well as the ability to involve information from various fields and contexts, and thus potentially increasing the generalizability to a larger audience of target technology users than qualitatively developed Personas. Other benefits of using LLMs (specifically, ChatGPT) have been reported in a systematic review by Sallam [44]. Most benefits identified by relevant research are related to scientific writing, with some examples being versatility in writing, readability, improved language, and even qualitative data analysis [45] etc. LLMs have also been proven to be a valuable tool in creating culturally sensitive clinician vignettes, a process that is quite relevant to the research in this paper. Namely, the LLMs contributed to inclusive representation and the avoidance of the stereotypes and negative portrayals of specific groups by focusing on individual patient experiences rather than generalizations about entire communities [46]. In line with this, several authors have concluded that LLMs have remarkable potential in the healthcare domain [47]. However, a quantitative, meaning objective, comparison of Personas generated using LLMs and qualitatively developed Personas, has not been performed in the context of childhood obesity or even in the broader context of health technologies.

Despite the potential of LLMs to create believable content, the advantages come with certain limitations. Perhaps most importantly, Hämäläinen [42] showed that LLMs made several factual errors, while Salminen et al. [17] report the potential of stereotypes and biases emerging in LLM-generated Persona descriptions. One of the possible reasons for the inaccuracies in generated information can be linked to security concerns regarding the attacks that manipulate the model and result in inaccurate or undesirable output content [15]. Another reason could be the training dataset size; if the dataset is too small, the model may not be able to generalize to various situations [47]. As stated by Hämäläinen et al. [42], AI tools can create content that highly resembles real human-generated text, but generated data is still of a lower quality than the real data. Thus, the trade-off between data quality and cost needs to be considered [42].

## 2. Materials and Methods

The study is part of the value-sensitive design process of the EU-funded project BIO-STREAMS (<https://www.bio-streams.eu/>). Project BIO-STREAMS is a multi-pillar framework for children's anti-obesity behavior, building on an EU Biobank, Micro-Moments, and Mobile Recommendation Systems. The digital solutions focus on clinical (i.e., management) and educational (i.e., community-based prevention) contexts. While the general methodology was the same in both contexts, they differed in the Personas evaluated and the participants who were eligible to participate in each part of the study. In the clinical setting, the Personas were designed to enable the co-design and co-development process of effective digital interventions in obesity management. In the educational setting, the Personas were selected to best represent the most relevant stakeholders for the design of effective digital interventions in the community-based prevention of obesity.

### 2.1. Participants

In the study, two samples of participants were employed, i.e., one of participants with a clinical professional background and another with an educational professional background.

In the clinical context, a total of 93 participants clicked on the survey, and 85 (91.40%) began filling in the survey. After the exclusion of participants who did not complete the whole survey ( $n = 4$ , 4.71%) and those who failed more than one attention check ( $n = 2$ , 2.35%) integrated in the survey [48], the final sample consisted of 63 participants. Of those, 55.56% were female, 42.86% were male, and 1.59% self-described their gender as 'other'. Participants had, on average, 9.24 years of professional work experience ( $SD = 9.06$ ). In total, 25.40% of participants were employed as nurses and 20.63% as doctors (e.g., general practitioners, dentists, orthopedists), followed by public health prevention managers, psychologists, students, and scientific researchers/consultants (4.76% each), and other occupations with two or fewer participants per occupation (altogether 34.92% of the sample). Overall, 49.21% of participants were employed in primary health care, 23.81% in secondary health care, 9.52% in tertiary health care, 4.76% were at university, 3.17% were students, and 9.52% were employed in other types of organizations. Regarding educational background, 39.68% of participants studied medicine, 23.81% studied nursing, 6.35% each studied psychology and social work, 4.76% studied pharmacy, and 19.05% of participants responded other. About half of the participants (41.27%) obtained a bachelor's degree or equivalent, followed by master's degree or equivalent (36.51%), higher secondary education (12.70%), a doctoral degree or equivalent (6.35%), and primary education or less (3.17%). All the participants resided in the European region at the time of the study, about a quarter of them in the United Kingdom (26.98%), followed by Portugal (19.05%), Poland (17.46%), Italy (9.52%), The Netherlands and Spain (4.76% each), and other countries represented by two or fewer participants.

In the educational context, a total of 82 participants clicked on the survey, and 78 (95.12%) began filling in the survey. After the exclusion of participants who did not complete the whole survey ( $n = 4$ , 5.13%) and those who failed more than one attention check ( $n = 1$ , 1.28%), the final sample consisted of 64 participants. Of those, 60.94% were female, 37.50% were male, and 1.56% preferred not to answer the question regarding gender. Participants had an average of 11.09 years of professional work experience ( $SD = 10.31$ ). Half of the participants (50.00%) were employed as teachers, followed by psychologists, researchers, and teaching assistants (each 4.69%), administrative workers, assistants, and students (each 3.13%), and other occupations (29.69%). In total, 40.63% of participants were employed in higher education, 23.44% in primary schools, 18.75% in secondary or high school, in a language school, or were self-employed (each 3.13%), and 10.94% employed in other types of organizations. Educational background was, for most participants, education in teaching a specific subject (e.g., history; 39.06%): primary education (15.63%), psychology (9.38%), social work (4.69%), and economics, engineering, and business management (3.13% each), while 21.88% had other educational backgrounds. About half of the participants (43.75%) obtained a master's degree or equivalent, followed by a bachelor's degree or equivalent (40.63%), higher secondary education (3.13%), a doctoral degree or equivalent (2.13%), and lower secondary education (1.56%). All the participants resided in the European region at the time of the study, most of them again in the United Kingdom (25.94%), followed by Italy (14.06%), The Netherlands (10.94%), Portugal and Spain (9.38% each), and other countries represented by two participants or less.

## 2.2. Materials

Participants were presented with three Personas in each context, namely Doctor (Pediatrician), Dietitian, and Child patient in the clinical context, and Teacher, Principal, and Student in the educational context. Four versions of each Persona were presented, one developed by experts and three developed by LLMs, specifically Bard, ChatGPT 3.5, and Llama2 70B.

Qualitative Personas were developed in combination with a quantitative approach. In the first step, experts answered open-ended questions, asking about the motivations, needs of potential users and their goals, the expected benefits, and the digital solutions they seek in relation to the respective platform. Then, the key characteristics of each Persona were further refined in a mapping exercise with experts. Following this, we sought additional input from experts in several sessions to refine the Persona list and create the Persona descriptions. The second step was the quantitative evaluation of Personas with the Persona Perception Scale (PPS) [34], which was similar to the method described below.

The prompts were constructed to seek comprehensive, narrative-driven answers to key questions about the Persona's background, goals, needs, motivations, benefits from the platform, and sought-after digital solutions. At the same time, the prompts were quite generic to ensure generalizability of the approach, without a specific expert knowledge about LLMs required by the researchers. Namely, each prompt begins with a clear directive to create a structured representation of a Persona, with a specific role (e.g., Dietitian, Pediatrician, Child, Patient, Teacher, Principal). Limited additional context, such as workplace (e.g., primary care, clinic, school) and target group (e.g., adolescents suffering from obesity), is also defined. The prompts also list specific questions to structure the Persona profile, i.e., background description of the Persona, the Persona's goals related to obesity, the Persona's needs, the Persona's motivations that drive the engagement, the digital solutions sought, and the expected benefits from the digital solutions/platform. In this way, we ensure that the Personas are detailed, relatable, and actionable for designing effective digital interventions for obesity prevention and treatment. Specific prompts used for generating the Personas via LLMs and all the Personas' descriptions are available in the Supplementary Materials.

Participants completed a questionnaire on their perception of different Personas' descriptions and questions on their socio-demographic characteristics.

The perception of Personas' descriptions was assessed via the PPS [34]. The PPS consists of eight subscales; however, in this study, the following six were used: Credibility (three items, e.g., The Persona seems like a real person), Consistency (four items, e.g., The Persona information seems consistent), Completeness (four items, e.g., The Persona profile is not missing vital information), Clarity (three items, e.g., The information in the persona profile is easy to understand), Empathy (three items, e.g., I feel like I understand this persona), and Likability (four items, e.g., I could be friends with this Persona). Most subscales showed acceptable to great internal consistency (Credibility:  $\alpha = 0.72$ – $0.89$ , Completeness:  $\alpha = 0.76$ – $0.91$ , Empathy:  $\alpha = 0.72$ – $0.87$ , Likability:  $\alpha = 0.81$ – $0.93$ ). We modified one item of the Completeness subscale (i.e., The Persona profile is detailed enough to make decisions about the consumers it describes) by replacing the term 'consumers' with 'people' to align better with the Personas used in this study, and omitted one item from the Consistency subscale due to its reference to parts of the Persona description that we did not include in the Personas used in this study. The scale, therefore, consisted of 20 items answered on a 7-point scale (1—"Strongly disagree", 7—"Strongly agree"), with the additional "Not applicable" option, and it was completed for each of the Persona presented. The Consistency and Credibility subscales each contained one item that refers to a picture of the Persona, which was not applicable in the Persona descriptions

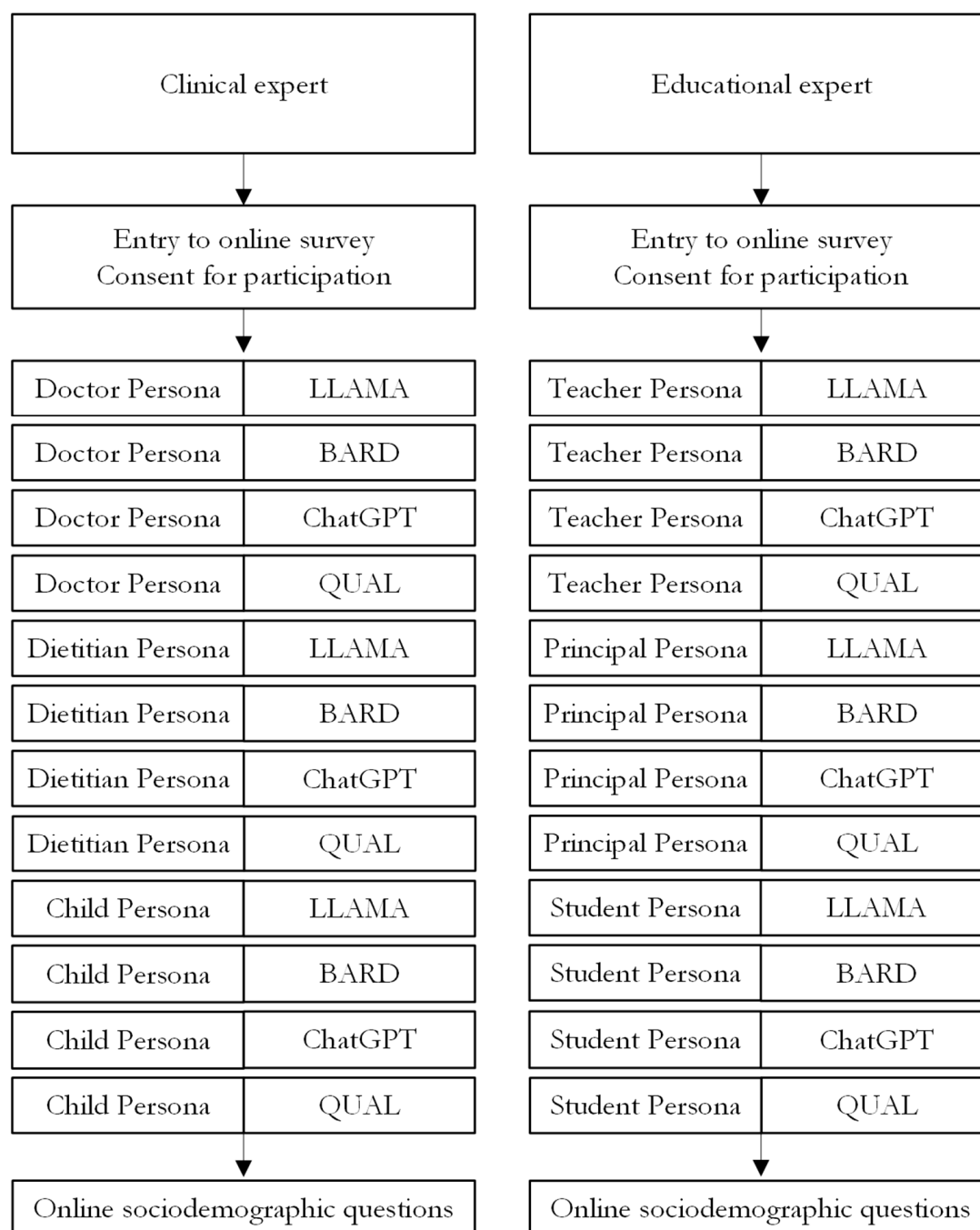
developed using LLMs in the present study (i.e., a picture of the Persona was included only in the qualitatively developed Persona descriptions); therefore, these two items were excluded from the analyses. This was also reflected in some of the lower internal consistency coefficients (Consistency:  $\alpha = 0.38$ – $0.84$ , Credibility:  $\alpha = 0.15$ – $0.73$ ) for the LLM-generated Persona descriptions, which was not the case in the qualitatively developed Persona descriptions where no items were excluded (Consistency:  $\alpha = 0.61$ – $0.86$ , Credibility:  $\alpha = 0.51$ – $0.77$ ). All alpha coefficients are reported in the Supplementary Materials (Tables S1–S6).

Participants were also asked to respond to four attention checks that were spread throughout the questionnaires, all in the form of directed questions (e.g., please mark as “Not applicable”) [48].

### 2.3. Procedure

Participants were recruited using the Prolific panel, a data collection platform that recruits participants within its own pool and has been exhibited to provide high-quality data (i.e., respondents are likely to pass attention checks, follow instructions, and provide meaningful answers) [49], in February 2024 to participate in an online study. Participants had to be aged 18 years or above, reside in Europe, have a good knowledge of English, and be employed in the medicine sector or education and training sector to participate in the clinical or educational context of this study, respectively. Required sample sizes were estimated a priori via G\*Power 3.1 [50], by selecting the MANOVA global effects (4 groups, 6 response variables), entering the medium effect size ( $f^2 = 0.15$ ), and a conventional significance threshold and power ( $\alpha = 0.05$ ,  $1 - \beta = 0.80$ ), which suggested that at least 52 participants must be recruited in each study context.

Participants were first informed about the study, the anonymous and voluntary participation, and the right to terminate the study at any time. After consenting to participate, they were presented with three Personas that they are likely to encounter in their professional context (i.e., clinical or educational), and each Persona was presented in four variations as described above, i.e., three developed via a specific large language model, and a Persona developed by experts. Each participant, therefore, evaluated 12 descriptions of Personas presented in a random order to eliminate any potential order effects (Figure 1). Moreover, participants were not informed of how each of the presented descriptions was developed. After completing all the evaluations, participants were asked to answer the socio-demographic questions. The study procedure took approximately 33 min to complete. Participants were rewarded with 6.5£ for their participation. Following the institutional and national guidelines, ethical approval was not required for this study.



**Figure 1.** Study procedure diagram. For each context (clinical and educational) different Personas generated using a different method (Llama, Bard, Chat-GPT, QUAL) appeared in a random order.

#### 2.4. Statistical Analyses

Statistical analyses were conducted using R 4.3.1 and IBM SPSS Statistics 29.0. Participants who dropped out before the end of the study and those who failed more than one attention check were excluded from the analyses. Next, we analyzed the pattern of missing data in the final samples. As the results showed that the data are likely missing at random (Little's Missing Completely at Random test,  $\chi^2(2983) = 0.000$ ,  $p = 1.000$  for the clinical context sample, and  $\chi^2(2405) = 0.000$ ,  $p = 1.000$  for the educational context sample), we replaced missing values using the expectation–maximization algorithm. Then, we calculated skewness and kurtosis, and normality distribution tests. We proceeded with calculating the descriptive statistics and Pearson correlation coefficients and examined the

inter-rater reliability via intraclass correlation coefficients. Values below 0.5 were deemed as pointing to poor reliability, values between 0.5 and 0.75 to moderate, between 0.75 and 0.9 to good, and values above 0.9 as pointing to excellent reliability [51].

To evaluate potential differences between the Persona descriptions, we performed multivariate analysis of variance (MANOVA) tests for each type of Persona separately. As for the Teacher, Principal, Student, and Patient Personas, the assumption of homogeneity of the covariance matrices was violated, along with slight deviations from normality in several variables for all Personas, and for the moderate sample sizes, we employed the bootstrap procedure on 1000 samples [52]. Pillai's trace was used for the multivariate analyses due to its robustness [53]. The significance threshold was adjusted with the Bonferroni correction due to multiple comparisons, resulting in a new threshold of 0.017 for the multivariate analyses. Significant multivariate comparisons were followed up by analyses of variance (ANOVAs), which were further explored with post-hoc tests in cases with significant results, using the Games–Howell correction due to violations of the homogeneity of variance assumption. Due to multiple comparisons, the significance threshold was adjusted using the Bonferroni correction and was set to 0.001.

### 3. Results

We present the means and standard deviations of the Persona perception dimensions, i.e., dimensions of Clarity, Completeness, Consistency, Credibility, Empathy, and Likability (separately for each Persona) in Tables 1 and 2, while the correlations between these dimensions are presented in the Supplementary Materials (Tables S1–S6). All the correlations between the observed dimensions were positive, and most were significant, exhibiting intermediately strong relationships. Overall, the inter-rater reliability was excellent in a clinical,  $ICC(2, 63) = 0.92$ , 95% CI [0.90, 0.93] and educational context,  $ICC(2, 64) = 0.92$ , 95% CI [0.91, 0.94], pointing towards a high agreement of participants in their evaluations of the Personas.

Next, we performed multivariate analyses of variance (MANOVA) for each Persona separately using Pillai's trace to test for the differences in the perception of dimensions among the variants of Persona created with different methodologies. The results showed a significant effect of different Persona variants on perception dimensions in the case of the Patient Persona,  $V = 0.125$ ,  $F(18, 756735) = 1822.161$ ,  $p < 0.001$ ,  $\eta_p^2 = 0.042$ ; the Principal Persona,  $V = 0.251$ ,  $F(18, 768747) = 3903.271$ ,  $p < 0.001$ ,  $\eta_p^2 = 0.084$ ; the Student Persona,  $V = 0.209$ ,  $F(18, 747) = 3.103$ ,  $p < 0.001$ ,  $\eta_p^2 = 0.070$ ; and the Teacher Persona,  $V = 0.212$ ,  $F(18, 747) = 3.149$ ,  $p < 0.001$ ,  $\eta_p^2 = 0.071$ . On the other hand, the effect of different Persona variants on the perception dimensions was not significant for the Dietitian Persona,  $V = 0.065$ ,  $F(18, 735) = 0.906$ ,  $p = 0.571$ ,  $\eta_p^2 = 0.022$ ; and the Pediatrician Persona,  $V = 0.088$ ,  $F(18, 735) = 1.230$ ,  $p = 0.230$ ,  $\eta_p^2 = 0.029$ .

The significant MANOVA results for the Patient, Principal, Student, and Teacher Personas were followed up with ANOVA tests to test for specific differences in Persona variants per each perception dimension, and where they revealed significant differences in the Persona perception dimensions, this was followed up with pairwise comparisons.

In the clinical context, as the MANOVA results revealed significant differences in the Persona perception dimensions only for the Patient Persona, these were followed up with ANOVA tests (Table 2). These tests revealed significant differences in the perception of the Patient Persona, where the effects of the Persona description variants were significant for all the Persona perception dimensions.

Significant ANOVA tests for the Patient Persona were followed up with pairwise comparisons, reported in the Supplementary Materials (Table S7). All the comparisons, i.e., between each combination of two Persona description variants within each of the Persona

perception dimensions, were significant, except for the Bard and Llama variants in the Completeness dimension, and the ChatGPT and QUAL (qualitatively developed) variants in the Credibility dimension.

These post-hoc pairwise comparisons revealed some consistent results among the Persona perception dimensions for the Patient Persona (Figure 2). The Bard variant of this Persona was most favorably perceived on all the Persona perception dimensions, except for the Credibility dimensions, where there was no significant difference between the Bard and Llama variants, and the Completeness dimension, where the Llama variant of the Patient Persona was not significantly differently perceived. The second-most favorable perceived variant of the Patient Persona dimension was the one produced by Llama, except for the Completeness dimension, where the Bard variant was equally well perceived. It is interesting to note that the qualitatively developed variant of the Patient Persona was significantly least favorably perceived on all the Persona perception dimensions.

**Table 1.** Descriptive statistics and post-hoc differences among the perception of Personas' dimensions in the educational context.

|                              | Bard<br>M (SD) |        | ChatGPT<br>M (SD) |        | Llama<br>M (SD) |        | QUAL<br>M (SD) |        | ANOVA<br>F   | $\eta_p^2$ |
|------------------------------|----------------|--------|-------------------|--------|-----------------|--------|----------------|--------|--------------|------------|
| <b>Principal<sup>a</sup></b> |                |        |                   |        |                 |        |                |        |              |            |
| Clarity                      | 5.90           | (0.82) | 5.77              | (0.84) | 5.69            | (1.14) | 5.02           | (1.65) | 7397.049 *** | 0.080      |
| Completeness                 | 5.11           | (1.20) | 5.22              | (1.03) | 5.40            | (0.92) | 5.34           | (1.14) | 897.126 ***  | 0.010      |
| Consistency                  | 5.61           | (0.91) | 5.59              | (0.95) | 5.51            | (1.02) | 5.75           | (0.81) | 756.364 ***  | 0.009      |
| Credibility                  | 5.50           | (0.92) | 5.28              | (0.89) | 5.34            | (1.03) | 5.41           | (0.97) | 625.840 ***  | 0.007      |
| Empathy                      | 4.96           | (1.08) | 4.57              | (1.16) | 4.91            | (1.18) | 4.42           | (1.32) | 3123.551 *** | 0.035      |
| Likability                   | 5.12           | (1.07) | 4.66              | (1.26) | 4.93            | (1.20) | 4.55           | (1.38) | 2876.472 *** | 0.033      |
| <b>Student<sup>b</sup></b>   |                |        |                   |        |                 |        |                |        |              |            |
| Clarity                      | 5.76           | (0.95) | 5.83              | (0.90) | 5.96            | (0.92) | 5.67           | (1.07) | 1.019        | 0.012      |
| Completeness                 | 4.74           | (1.25) | 5.35              | (0.94) | 5.29            | (1.12) | 5.29           | (1.03) | 4.417 **     | 0.050      |
| Consistency                  | 5.06           | (1.42) | 5.59              | (0.95) | 5.60            | (1.09) | 5.42           | (1.21) | 2.896 *      | 0.033      |
| Credibility                  | 4.77           | (1.47) | 5.56              | (1.07) | 5.57            | (1.13) | 4.88           | (1.35) | 7.384 ***    | 0.081      |
| Empathy                      | 4.33           | (1.38) | 4.76              | (1.19) | 4.81            | (1.28) | 4.18           | (1.43) | 3.583 *      | 0.041      |
| Likability                   | 4.36           | (1.28) | 4.79              | (1.21) | 5.00            | (1.22) | 4.17           | (1.38) | 5.821 ***    | 0.065      |
| <b>Teacher<sup>c</sup></b>   |                |        |                   |        |                 |        |                |        |              |            |
| Clarity                      | 5.99           | (0.93) | 5.71              | (0.86) | 5.71            | (0.96) | 5.74           | (0.90) | 1.409        | 0.017      |
| Completeness                 | 5.35           | (1.11) | 4.93              | (1.13) | 5.05            | (1.09) | 5.46           | (1.10) | 3.243 *      | 0.037      |
| Consistency                  | 5.75           | (0.81) | 5.60              | (0.94) | 5.43            | (1.08) | 5.77           | (0.82) | 1.869        | 0.022      |
| Credibility                  | 5.74           | (0.86) | 5.17              | (1.01) | 5.38            | (0.91) | 5.58           | (0.94) | 4.601 **     | 0.052      |
| Empathy                      | 5.17           | (1.10) | 4.72              | (1.12) | 4.98            | (1.11) | 4.70           | (1.22) | 2.505        | 0.029      |
| Likability                   | 5.52           | (1.07) | 4.76              | (1.15) | 5.09            | (1.19) | 4.88           | (1.21) | 5.302 **     | 0.059      |

Notes. QUAL = qualitatively developed Persona description. <sup>a</sup> For all ANOVA comparisons within the Principal Persona,  $F(3, 256252)$ . <sup>b</sup> For all ANOVA comparisons within the Student Persona,  $F(3, 252)$ . <sup>c</sup> For all ANOVA comparisons within the Teacher Persona,  $F(3, 252)$ . \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ .

In the educational context, the MANOVA results revealed significant differences in the overall perception of all three evaluated Personas, i.e., Principal, Student, and Teacher Personas. Therefore, these differences were followed up with ANOVA tests (Table 1).

For the Principal Persona, significant effects of Persona description variants were found for all the Persona perception dimensions. For the Student Persona, all but the Clarity dimension showed significant effects of the Persona description variants, and for the Teacher Persona, only the dimensions of Completeness, Credibility, and Likability showed significant effects.

Significant ANOVA tests were, similar to the procedure in the clinical context, followed up with pairwise comparisons, reported in the Supplementary Materials (Tables S8–S10). For the Principal Persona, all the pairwise comparisons showed significant differences between the Persona description variants in all the Persona perception dimensions. The Clarity, Credibility, Empathy, and Likability dimensions were most favorably perceived for the Bard variant of the Persona description, while for the Completeness dimension,

the most favorably perceived was the Llama variant, and for the Consistency dimension, the qualitatively developed Persona variant was the most favorably perceived. On the other hand, the qualitatively developed Persona variants were least favorably evaluated in the Clarity, Empathy, and Likability dimensions, while in the remaining dimensions, i.e., Completeness, Consistency, and Credibility, the variants produced by Bard, Llama, and ChatGPT were the least favorably perceived, respectively. All comparisons are presented in Figure 2.

For the Teacher Persona, only the comparison in the Likability dimension between the Bard and ChatGPT variants was significant, where the Bard variant was more favorably perceived compared with the ChatGPT variant.

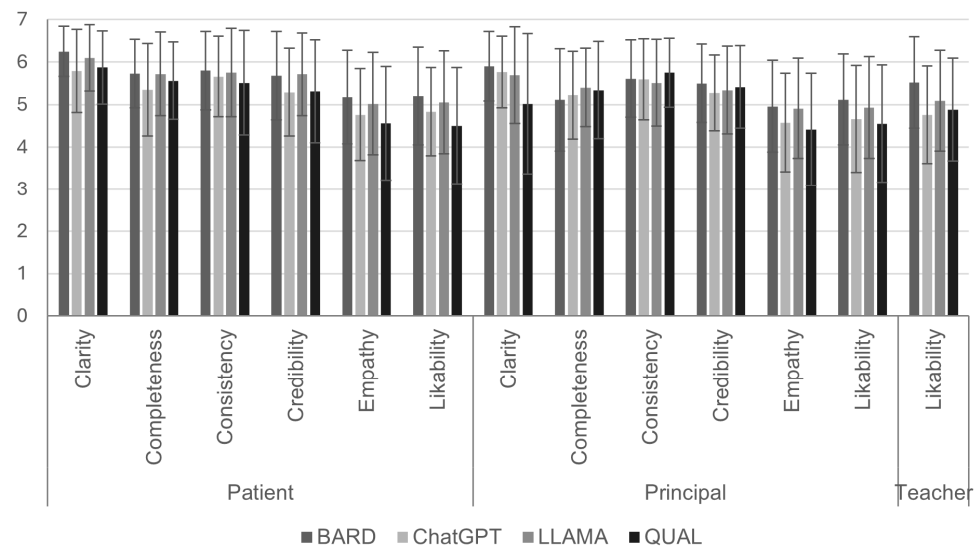
For the Student Persona, none of the pairwise comparisons revealed significant differences between the Persona description variants in any of the perception dimensions.

**Table 2.** Descriptive statistics and post-hoc differences among the perception of Personas' dimensions in the clinical context.

|                            | Bard<br>M (SD) |        | ChatGPT<br>M (SD) |        | Llama<br>M (SD) |        | QUAL<br>M (SD) |        | ANOVA<br>F   | $\eta_p^2$ |
|----------------------------|----------------|--------|-------------------|--------|-----------------|--------|----------------|--------|--------------|------------|
| <b>Dietitian</b>           |                |        |                   |        |                 |        |                |        |              |            |
| Clarity                    | 5.86           | (0.82) | 5.81              | (0.89) | 6.20            | (0.76) | 5.84           | (0.91) |              |            |
| Completeness               | 5.39           | (1.15) | 5.30              | (1.17) | 5.75            | (0.91) | 5.36           | (1.10) |              |            |
| Consistency                | 5.68           | (0.96) | 5.57              | (1.10) | 5.89            | (0.82) | 5.57           | (0.91) |              |            |
| Credibility                | 5.14           | (1.05) | 5.12              | (1.20) | 5.48            | (1.08) | 4.90           | (1.25) |              |            |
| Empathy                    | 4.60           | (1.28) | 4.67              | (1.28) | 4.97            | (1.19) | 4.51           | (1.33) |              |            |
| Likability                 | 5.09           | (1.23) | 5.09              | (1.30) | 5.52            | (1.05) | 4.90           | (1.18) |              |            |
| <b>Patient<sup>a</sup></b> |                |        |                   |        |                 |        |                |        |              |            |
| Clarity                    | 6.25           | (0.59) | 5.79              | (0.98) | 6.10            | (0.78) | 5.87           | (0.86) | 4283.722 *** | 0.048      |
| Completeness               | 5.73           | (0.80) | 5.35              | (1.09) | 5.72            | (0.98) | 5.56           | (0.91) | 2200.721 *** | 0.026      |
| Consistency                | 5.80           | (0.92) | 5.66              | (0.95) | 5.75            | (1.04) | 5.51           | (1.23) | 960.804 ***  | 0.011      |
| Credibility                | 5.68           | (1.04) | 5.29              | (1.03) | 5.71            | (0.97) | 5.31           | (1.21) | 2926.731 *** | 0.034      |
| Empathy                    | 5.18           | (1.10) | 4.76              | (1.08) | 5.02            | (1.20) | 4.56           | (1.34) | 3491.944 *** | 0.040      |
| Likability                 | 5.20           | (1.15) | 4.83              | (1.04) | 5.05            | (1.21) | 4.50           | (1.37) | 4142.605 *** | 0.047      |
| <b>Pediatrician</b>        |                |        |                   |        |                 |        |                |        |              |            |
| Clarity                    | 5.94           | (0.97) | 5.85              | (0.94) | 5.99            | (0.76) | 5.96           | (1.02) |              |            |
| Completeness               | 5.58           | (0.94) | 5.58              | (0.99) | 5.46            | (0.98) | 5.80           | (0.80) |              |            |
| Consistency                | 5.72           | (0.94) | 5.68              | (0.94) | 5.72            | (0.82) | 5.87           | (0.84) |              |            |
| Credibility                | 5.27           | (1.04) | 5.33              | (1.06) | 5.43            | (0.93) | 5.66           | (0.84) |              |            |
| Empathy                    | 4.93           | (1.25) | 4.68              | (1.29) | 4.86            | (1.25) | 5.03           | (1.16) |              |            |
| Likability                 | 5.37           | (1.04) | 5.11              | (1.19) | 5.23            | (1.12) | 5.52           | (1.02) |              |            |

Notes. QUAL = qualitatively developed Persona description. <sup>a</sup> For all ANOVA comparisons within the Patient Persona,  $F(3, 252248)$ . \*\*\*  $p < 0.001$ . ANOVA results are not reported for Dietitian and Pediatrician Personas, as ANOVA results did not reveal significant effects.

Where post-hoc pairwise comparisons were significant, they revealed some differences among the Persona description variants that were consistent across Personas (Figure 2). The Empathy and Likability dimensions were most favorably perceived in the Bard variant, followed by Llama and ChatGPT, while the qualitatively developed variant was perceived least favorably, that is for the Principal and Patient Personas. Similarly, the Likability dimension in the Teacher Persona was most favorably perceived in the Bard variant, and least favorably in the ChatGPT variant. Some consistency in the pairwise comparisons was also found for the Consistency dimension, where the Bard variant was perceived as most or second-most favorable for the Patient and Principal Personas, respectively, and for the Credibility dimension, where, again, the Bard variant was perceived as most or second-most favorable for Principal and Patient Personas, respectively. Other pairwise comparisons, although significant for the Principal and Patient Personas, did not reveal similar patterns.



**Figure 2.** Means and standard deviations of the Personas' perception dimensions with significant differences in pairwise comparisons.

#### 4. Discussion

Digital interventions offer several advantages that can enhance the effectiveness and sustainability of behavior change [1]. However, effective digital interventions should be holistic, integrating various components to address the physical, emotional, and social aspects of health. Moreover, personalization, engagement, and accessibility are critical to ensuring these interventions are impactful and sustainable for children and adolescents. To achieve this goal, it is essential to incorporate user perspectives during the development process [5]. Personas are instrumental to support such a user-centered development process leading to long-term sustainability digital interventions [8,9]. By focusing on relevant user groups with their needs and values, such approaches can ensure that interventions are well-targeted, user-friendly, and effective, thereby enhancing the user experience and increasing engagement rates [22,26,30]. However, the most prominent method of Persona development, the qualitative approach, which creates the most representative Personas, presents some challenges, such as being time-consuming and subjective, potentially misrepresenting relevant user groups, and being slow to adapt to changing user needs [14]. With advancements in AI, particularly LLMs, more efficient methods for generating Personas have been proposed and explored [8,42,54–56].

The current study aimed to investigate how experts in the field perceive and evaluate Persona descriptions generated by three different LLMs and qualitatively developed Persona descriptions. The study focused on this exploration in two contexts, i.e., clinical and educational, and more broadly in the context of obesity research. In each context, three Personas were defined, and their descriptions were generated using both traditional, qualitative methods and three LLMs (i.e., Bard, Llama, and ChatGPT), resulting in a total of 12 Persona descriptions per context. These descriptions were then evaluated using the Persona Perception Scale by experts in the respective fields.

Previous studies have found that LLMs can generate responses that are highly similar to human responses [42,57] and that LLM-generated Personas have been assessed as being of good quality [17]. The current study extends this work by comparing the descriptions generated by various LLMs among themselves and with qualitatively developed Persona descriptions. This approach builds upon and adds to the findings of related studies, such as those by Alessa and Al-Khalifa [54] and Hong et al. [56], which did not evaluate the perceptions of LLM-generated Personas; a study by Cheng et al. [55], which did not include

expert perspectives on the Personas; and the work of Salminen et al. [17], which is most closely related to the present study but explored Personas generated by one LLM.

The results of this study showed that in many comparisons, perceptions of Persona descriptions did not significantly differ among LLM-generated Personas and qualitatively developed Personas, supporting the finding of Salminen et al. [17] that “LLMs can generate consistent Personas that are perceived as believable, relatable, and informative for design/.../as perceived by the human evaluators”. However, post-hoc comparisons of significant ANOVA analyses revealed some differences that are worth exploring. In general, where these comparisons were significant, in most cases the descriptions generated using Bard model were evaluated most favorably, which is especially evident in evaluations of the Empathy, Likability, and Clarity of Personas’ descriptions (i.e., in Patient and Principal Personas, and the Likability dimension also in the Teacher Persona). In these cases, Bard was evaluated significantly better than ChatGPT- and Llama-generated descriptions, as well as qualitatively developed Personas that were in many cases evaluated with the lowest or second-to-lowest scores. Qualitatively developed Personas were formed by experts in the BIO-STREAMS project consortium that consists of partners from various EU countries, thus lacking native English speakers, which could potentially impact its development, which was reflected especially in less favorable evaluations of dimensions such as Clarity and Likability compared with Persona descriptions developed by LLMs. As Salminen et al. [17] point out, LLMs can have, in generating Personas, “a property of fluency that can explain why human evaluators give such high ratings to LLM-generated Personas”.

Taken together, our results show that LLMs can be efficiently employed in the process of Persona development. By comparing the expert evaluations of Persona descriptions generated by LLMs with qualitative methods, this study provides valuable insights into the potential of AI-driven Persona development. The findings contribute to the growing body of research on the application of LLMs in user experience design and highlight the need for further exploration of this promising approach.

#### *4.1. Limitations and Future Directions*

While our study expands the current knowledge by exploring evaluations of different LLMs and comparing them with qualitatively developed Personas in two contexts and in a relatively large sample of experts, some limitations need to be mentioned. While the Persona Perception Scale [34] is an instrument with generally good psychometric properties, the consistency and credibility subscales exhibited a low internal consistency for evaluating LLM-generated Personas, most likely due to the exclusion of items that referred to the picture of the Persona that our LLM-generated descriptions did not entail. Therefore, some caution is advised in interpreting the results regarding these two dimensions of Persona perception.

In addition, while LLMs are imposed by heavy safety restrictions that aim to prevent misuse, the generation of specific Personas might be limited, particularly those exhibiting harmful or unethical behaviors. Namely, the limited ability to study negative Personas may, for instance, hamper research on detecting and mitigating harmful AI behaviors. Moreover, the inability to generate certain Personas could create blind spots in safety testing and the evaluation of AI systems’ responses to problematic interactions. Future work should consider methods to ethically study these edge cases while maintaining appropriate safeguards.

Furthermore, based on the results of this study, we corroborate further research implications for LLM-generated Personas put forward by Salminen et al. [17]. These include verifying the LLM-generated Persona descriptions using subject-matter experts, adjusting the prompts if observing challenges, and comparing LLM-generated Personas

with additional sources when possible, such as with Personas developed via the qualitative way, as performed in the present study. Additionally, we recommend exploring the use of various LLMs for developing the Persona description in question, as they can provide qualitatively different Persona descriptions.

#### 4.2. Practical Implications

Since this study delves into a highly interdisciplinary area that involves both academia and business, some practical implications of the current study can be addressed. Firstly, Personas developed within the scope of the study can be found in the Supplementary Materials (Section S2) and further used in obesity practice or research, if they are found to be relevant. Secondly, following the need to create new Personas, the prompts developed for this study can be used as an example, which are also available in the Supplementary Materials (Section S1). Thirdly, our findings have shown that LLM-generated Personas were in general evaluated similarly to the qualitatively developed Personas on all the perception dimensions. Moreover, in some instances, LLM-generated Personas were even evaluated more favorably than the qualitatively developed Personas. This shows that LLMs can be used equally effectively for Persona creation as a qualitative method, which is more time-consuming and static [35,39]. Fourthly, these findings (and Personas or prompts used as a template) are applicable in any field that values the presentations of real-life persons, such as design, marketing, software development, and health informatics [13].

### 5. Conclusions

In this article, the research question “Which large language models are perceived to generate better or worse Personas in comparison with qualitatively developed Personas?” has been investigated by generating Personas via three different LLMs and comparing the perception of them to that of qualitatively developed Persona descriptions in two contexts: clinical and educational. The results show that LLM-generated and qualitatively developed Persona descriptions were mostly perceived similarly by expert evaluators on several dimensions, i.e., significant differences in the perception of different variants of Personas were found only in a subset of Persona perception dimensions for a part of the evaluated Persona variants.

Where significant differences occurred, i.e., in the evaluations of Patient, Principal, and Teacher Personas, some consistencies can be observed through several dimensions, Personas, and both the clinical and educational context. The most consistent results were found regarding the Persona descriptions using the Bard model, which were in most cases evaluated most or second-most favorably, especially in the evaluations of the empathy, Likability, Consistency, and Clarity dimensions of Personas’ descriptions. In these cases, Bard was evaluated significantly better than ChatGPT and Llama-generated descriptions, as well as qualitatively developed Personas that were in many cases evaluated with the lowest or second-to-lowest scores.

Although the differences found in the Persona perception dimensions for some of the Personas were significant and interesting, especially in their consistency, an even more interesting finding of this study is in that for the majority of Persona variants, there were no differences to be found in any of the Persona perception dimensions.

Therefore, this study provides support for the ability of LLMs to create comparable descriptions to the traditional qualitative method. As such, the study represents a valuable contribution to the current body of research on the topic. However, caution in their use must be applied. As previous studies highlight some important limitations of LLMs in Persona research (e.g., factual errors [42], potential stereotypes, and bias generation [17,58]), employing LLMs cannot be seen as a substitute to the qualitative development, but rather

as a useful and efficient tool that might set the user perspective research in technology development on a more agile path.

**Supplementary Materials:** The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/app15041937/s1>, Prompts used in generating Personas with the use of LLM models, Persona descriptions used in the study, Tables S1–S6: Correlations and alpha coefficients of dimensions of Persona perceptions, Tables S7–S10: Post-hoc pairwise comparisons regarding the perception of Personas.

**Author Contributions:** Conceptualization, U.S., A.R. and I.M.; methodology, U.S. and I.M.; software, I.M.; formal analysis, U.S.; investigation, U.S., A.R., N.P. and I.M.; data curation, U.S.; writing—original draft preparation, U.S., A.R., N.P. and I.M.; writing—review and editing, U.S., A.R., N.P. and I.M.; funding acquisition, I.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work has been funded by the European Union Horizon Europe Research and Innovation Programme project Bio-Streams (grant number 101080718) and the Slovenian Research and Innovation Agency, Advanced methods of interaction in telecommunication research programme (grant number P2-0069). The content of this works does not reflect the official position of the European Union or any other institution. The information and views expressed are the sole responsibility of the authors.

**Institutional Review Board Statement:** The study was conducted in accordance with the Declaration of Helsinki. Ethical review and approval were not required for this study following the institutional and national guidelines.

**Informed Consent Statement:** Informed consent was obtained from all the subjects involved in the study.

**Data Availability Statement:** The raw data supporting the conclusions of this article will be made available by the authors on request.

**Acknowledgments:** The authors would like to thank the BIO-STREAMS project consortium members for their collaboration in developing Personas.

**Conflicts of Interest:** The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of the data; in the writing of the manuscript; or in the decision to publish the results.

## References

1. Chatterjee, A.; Prinz, A.; Gerdes, M.; Martinez, S. Digital Interventions on Healthy Lifestyle Management: Systematic Review. *J. Med. Internet Res.* **2021**, *23*, e26931. [\[CrossRef\]](#) [\[PubMed\]](#)
2. Bucher, A.; Blazek, E.S.; Symons, C.T. How Are Machine Learning and Artificial Intelligence Used in Digital Behavior Change Interventions? A Scoping Review. *Mayo Clin. Proc. Digit. Health* **2024**, *2*, 375–404. [\[CrossRef\]](#)
3. Iyamu, I.; Gómez-Ramírez, O.; Xu, A.X.; Chang, H.-J.; Watt, S.; Mckee, G.; Gilbert, M. Challenges in the Development of Digital Public Health Interventions and Mapped Solutions: Findings from a Scoping Review. *Digit. Health* **2022**, *8*, 205520762211022. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Kassiopoulou, O.; Vasiliou, V.; Karekla, M. Overcoming Challenges in Adherence and Engagement Digital Interventions: The Development of the ALGEApp for Chronic Pain Management. *Internet Interv.* **2023**, *32*, 100611. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Sharpe, E.E.; Karasouli, E.; Meyer, C. Examining Factors of Engagement with Digital Interventions for Weight Management: Rapid Review. *JMIR Res. Protoc.* **2017**, *6*, e205. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Bijkerk, L.E.; Spigt, M.; Oenema, A.; Geschwind, N. Engagement with Mental Health and Health Behavior Change Interventions: An Integrative Review of Key Concepts. *J. Context. Behav. Sci.* **2024**, *32*, 100748. [\[CrossRef\]](#)
7. Idoughi, D.; Seffah, A.; Kolski, C. Adding User Experience into the Interactive Service Design Loop: A Persona-Based Approach. *Behav. Inf. Technol.* **2012**, *31*, 287–303. [\[CrossRef\]](#)
8. Salminen, J.; Wenyun Guan, K.; Jung, S.-G.; Jansen, B. Use Cases for Design Personas: A Systematic Review and New Frontiers. In Proceedings of the CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April–5 May 2022; pp. 1–21.

9. Bartels, S.L.; Taygar, A.S.; Johnsson, S.I.; Petersson, S.; Flink, I.; Boersma, K.; McCracken, L.M.; Wicksell, R.K. Using Personas in the Development of EHealth Interventions for Chronic Pain: A Scoping Review and Narrative Synthesis. *Internet Interv.* **2023**, *32*, 100619. [\[CrossRef\]](#)
10. Yardley, L.; Morrison, L.; Bradbury, K.; Muller, I. The Person-Based Approach to Intervention Development: Application to Digital Health-Related Behavior Change Interventions. *J. Med. Internet Res.* **2015**, *17*, e30. [\[CrossRef\]](#)
11. LeRouge, C.; Ma, J.; Sneha, S.; Tolle, K. User Profiles and Personas in the Design and Development of Consumer Health Technologies. *Int. J. Med. Inf.* **2013**, *82*, e251–e268. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Lee, J.H.; Kim, Y.M.; Rhiu, I.; Yun, M.H. A Persona-Based Approach for Identifying Accessibility Issues in Elderly and Disabled Users' Interaction with Home Appliances. *Appl. Sci.* **2021**, *11*, 368. [\[CrossRef\]](#)
13. Salminen, J.; Guan, K.; Jung, S.G.; Chowdhury, S.A.; Jansen, B.J. A Literature Review of Quantitative Persona Creation. In Proceedings of the Conference on Human Factors in Computing Systems—Proceedings, Honolulu, HI, USA, 25–30 April 2020; Association for Computing Machinery: New York, NY, USA, 2020.
14. Salminen, J.; Jung, S.; Almerikhi, H.; Cambria, E.; Jansen, B. How Can Natural Language Processing and Generative AI Address Grand Challenges of Quantitative User Personas? In Proceedings of the 25th International Conference on Human-Computer Interaction, Copenhagen, Denmark, 23–28 July 2023; AC Bella Sky Hotel and Bella Center: København, Denmark, 2023; pp. 211–231.
15. Deng, J.; Lin, Y. The Benefits and Challenges of ChatGPT: An Overview. *Front. Comput. Intell. Syst.* **2023**, *2*, 81–83. [\[CrossRef\]](#)
16. Sarker, I.H. AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems. *SN Comput. Sci.* **2022**, *3*, 158. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Salminen, J.; Santos, J.M.; Jung, S.; Jansen, B.J. Picturing the Fictitious Person: An Exploratory Study on the Effect of Images on User Perceptions of AI-Generated Personas. *Comput. Hum. Behav. Artif. Hum.* **2024**, *2*, 100052. [\[CrossRef\]](#)
18. Cruz-Martínez, R.R.; Wentzel, J.; Bente, B.E.; Sanderman, R.; van Gemert-Pijnen, J.E. Toward the Value Sensitive Design of EHealth Technologies to Support Self-Management of Cardiovascular Diseases: Content Analysis. *JMIR Cardio* **2021**, *5*, e31985. [\[CrossRef\]](#)
19. Jongsma, K.R.; Jongepier, F. Value-Sensitive Design and Global Digital Health. *Bull. World Health Organ.* **2020**, *98*, 579–580. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Jull, J.E.; Davidson, L.; Dungan, R.; Nguyen, T.; Woodward, K.P.; Graham, I.D. A Review and Synthesis of Frameworks for Engagement in Health Research to Identify Concepts of Knowledge User Engagement. *BMC Med. Res. Methodol.* **2019**, *19*, 211. [\[CrossRef\]](#)
21. Friedman, B.; Kahn, P.H.; Borning, A.; Hultgren, A. Value Sensitive Design and Information Systems. In *Early Engagement and new Technologies: Opening Up the Laboratory*; Doorn, N., Schuurbiers, D., van de Poel, I., Gorman, M.E., Eds.; Springer: Dordrecht, The Netherlands, 2013; Volume 14, pp. 55–95.
22. Winkler, T.; Spiekermann, S. Twenty Years of Value Sensitive Design: A Review of Methodological Practices in VSD Projects. *Ethics Inf. Technol.* **2021**, *23*, 17–21. [\[CrossRef\]](#)
23. Friedman, B.; Kahn, P.H.; Borning, A.; Hultgren, A. Value Sensitive Design and Information Systems. In *Human-Computer Interaction in Management Information Systems: Foundations*; Zhang, P., Galleta, D., Eds.; M.E. Sharpe: New York, NY, USA, 2006; pp. 348–372.
24. Schwartz, S.H. Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries. *Adv. Exp. Soc. Psychol.* **1992**, *25*, 1–65.
25. Oreg, S.; Nov, O. Exploring Motivations for Contributing to Open Source Initiatives: The Roles of Contribution Context and Personal Values. *Comput. Hum. Behav.* **2008**, *24*, 2055–2073. [\[CrossRef\]](#)
26. Simon, J. Value-Sensitive Design and Responsible Research and Innovation. In *The Ethics of Technology—Methods and Approaches*; Hansson, S.-O., Ed.; Rowman & Littlefield International: London, UK, 2016; pp. 219–236.
27. Hautakangas, M.; Kumpulainen, K.; Uusitalo, L. Children Developing Self-Regulation Skills in a Kids' Skills Intervention Programme in Finnish Early Childhood Education and Care. *Early Child Dev. Care* **2022**, *192*, 1626–1642. [\[CrossRef\]](#)
28. Blower, S.; Swallow, V.; Maturana, C.; Stones, S.; Phillips, R.; Dimitri, P.; Marshman, Z.; Knapp, P.; Dean, A.; Higgins, S.; et al. Children and Young People's Concerns and Needs Relating to Their Use of Health Technology to Self-Manage Long-Term Conditions: A Scoping Review. *Arch. Dis. Child.* **2020**, *105*, 1093–1104. [\[CrossRef\]](#)
29. Walker, K.L.; Bodendorf, K.; Kiesler, T.; de Mattos, G.; Rostom, M.; Elkordy, A. Compulsory Technology Adoption and Adaptation in Education: A Looming Student Privacy Problem. *J. Consum. Aff.* **2023**, *57*, 445–478. [\[CrossRef\]](#)
30. Edwards, J.; Waite-Jones, J.; Schwarz, T.; Swallow, V. Digital Technologies for Children and Parents Sharing Self-Management in Childhood Chronic or Long-Term Conditions: A Scoping Review. *Children* **2021**, *8*, 1203. [\[CrossRef\]](#)
31. Beleigoli, A.M.; Andrade, A.Q.; Cançado, A.G.; Paulo, M.N.L.; Diniz, M.D.F.H.; Ribeiro, A.L. Web-Based Digital Health Interventions for Weight Loss and Lifestyle Habit Changes in Overweight and Obese Adults: Systematic Review and Meta-Analysis. *J. Med. Internet Res.* **2019**, *21*, e9609. [\[CrossRef\]](#) [\[PubMed\]](#)

32. Stangeland Lie, S.; Karlsen, B.; Oord, E.R.; Graue, M.; Oftedal, B. Dropout from an Ehealth Intervention for Adults with Type 2 Diabetes: A Qualitative Study. *J. Med. Internet Res.* **2017**, *19*, e187.
33. Yardley, L.; Spring, B.J.; Riper, H.; Morrison, L.G.; Crane, D.H.; Curtis, K.; Merchant, G.C.; Naughton, F.; Blandford, A. Understanding and Promoting Effective Engagement With Digital Behavior Change Interventions. *Am. J. Prev. Med.* **2016**, *51*, 833–842. [\[CrossRef\]](#)
34. Salminen, J.; Santos, J.M.; Kwak, H.; An, J.; Jung, S.; Jansen, B.J. Persona Perception Scale: Development and Exploratory Validation of an Instrument for Evaluating Individuals' Perceptions of Personas. *Int. J. Hum. Comput. Stud.* **2020**, *141*, 102437. [\[CrossRef\]](#)
35. Baxter, K.; Courage, C.; Caine, K. *Understanding Your Users: A Practical Guide to User Research Methods*, 2nd ed.; Elsevier Inc.: Amsterdam, The Netherlands, 2015.
36. Partridge, S.R.; Redfern, J. Strategies to Engage Adolescents in Digital Health Interventions for Obesity Prevention and Management. *Healthcare* **2018**, *6*, 70. [\[CrossRef\]](#)
37. Jansen, B.J.; Jung, S.G.; Nielsen, L.; Guan, K.W.; Salminen, J. How to Create Personas: Three Persona Creation Methodologies with Implications for Practical Employment. *Pac. Asia J. Assoc. Inf. Syst.* **2022**, *14*, 1. [\[CrossRef\]](#)
38. Brickey, J.; Walczak, S.; Burgess, T. Comparing Semi-Automated Clustering Methods for Persona Development. *IEEE Trans. Softw. Eng.* **2012**, *38*, 537–546. [\[CrossRef\]](#)
39. Chapman, C.N.; Milham, R.P. The Personas' New Clothes: Methodological and Practical Arguments Against A Popular Method. In Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting, San Francisco, CA, USA, 16–20 October 2006; pp. 634–636.
40. Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 1–45. [\[CrossRef\]](#)
41. Alipour, H.; Pendar, N.; Roy, K. ChatGPT Alternative Solutions: Large Language Models Survey. *arXiv* **2024**, arXiv:2403.14469.
42. Hämäläinen, P.; Tavast, M.; Kunnari, A. 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, New York, NY, USA, 19 April 2023; ACM: New York, NY, USA, 2023; pp. 1–19.
43. OpenAI. ChatGPT. [Large Language Model]. 2023. Available online: <https://openai.com/chatgpt/> (accessed on 20 December 2023).
44. Sallam, M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare* **2023**, *11*, 887. [\[CrossRef\]](#) [\[PubMed\]](#)
45. De Paoli, S. Performing an Inductive Thematic Analysis of Semi-Structured Interviews With a Large Language Model: An Exploration and Provocation on the Limits of the Approach. *Soc. Sci. Comput. Rev.* **2024**, *42*, 997–1019. [\[CrossRef\]](#)
46. Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S.S.; Wei, J.; Chung, H.W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. Large Language Models Encode Clinical Knowledge. *Nature* **2023**, *620*, 172–180. [\[CrossRef\]](#)
47. Hadi, M.U.; Qasem, A.T.; Qureshi, R.; Shah, A.; Muneer, A.; Irfan, M.; Zafar, A.; Shaikh, M.B.; Akhtar, N.; Wu, J.; et al. A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage. *TechRxiv* **2020**, 1–29. [\[CrossRef\]](#)
48. Maniaci, M.R.; Rogge, R.D. Caring about Carelessness: Participant Inattention and Its Effects on Research. *J. Res. Pers.* **2014**, *48*, 61–83. [\[CrossRef\]](#)
49. Douglas, B.D.; Ewell, P.J.; Brauer, M. Data Quality in Online Human-Subjects Research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLoS ONE* **2023**, *18*, e0279720. [\[CrossRef\]](#) [\[PubMed\]](#)
50. Faul, F.; Erdfelder, E.; Lang, A.-G.; Buchner, A. G\*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences. *Behav. Res. Methods* **2007**, *39*, 175–191. [\[CrossRef\]](#)
51. Koo, T.K.; Li, M.Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* **2016**, *15*, 155–163. [\[CrossRef\]](#)
52. Konietzschke, F.; Bathke, A.C.; Harrar, S.W.; Pauly, M. Parametric and Nonparametric Bootstrap Methods for General MANOVA. *J. Multivar. Anal.* **2015**, *140*, 291–301. [\[CrossRef\]](#)
53. Field, A. *Discovering Statistics Using SPSS*, 3rd ed.; SAGE Publications Ltd.: London, UK, 2009.
54. Alessa, A.; Al-Khalifa, H. Towards Designing a ChatGPT Conversational Companion for Elderly People. In Proceedings of the 16th International Conference on Pervasive Technologies Related to Assistive Environments, New York, NY, USA, 5 July 2023; ACM: New York, NY, USA, 2023; pp. 667–674.
55. Cheng, M.; Durmus, E.; Jurafsky, D. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. *arXiv* **2023**, arXiv:2305.18189.
56. Hong, M.K.; Hakimi, S.; Chen, Y.-Y.; Toyoda, H.; Wu, C.; Klenk, M. Generative AI for Product Design: Getting the Right Design and the Design Right. *arXiv* **2023**, arXiv:2306.01217.

57. Zhou, Z.; Li, J.; Zhang, Z.; Yu, J.; Duh, H. Examining How the Large Language Models Impact the Conceptual Design with Human Designers: A Comparative Case Study. *Int. J. Hum. Comput. Interact.* **2024**, 1–17. [[CrossRef](#)]
58. Sison, A.J.G.; Daza, M.T.; Gozalo-Brizuela, R.; Garrido-Merchán, E.C. ChatGPT: More Than a “Weapon of Mass Deception” Ethical Challenges and Responses from the Human-Centered Artificial Intelligence (HCAI) Perspective. *Int. J. Hum. Comput. Interact.* **2024**, *40*, 4853–4872. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.