# ChatGPT for Learning HCI Techniques: A Case Study on Interviews for Personas

Jose Barambones, Cristian Moral, Angélica de Antonio, Ricardo Imbert, Loïc Martínez- Normand, and Elena Villalba-Mora

***Abstract*—Before interacting with real users, developers must be proficient in human–computer interaction (HCI) so as not to exhaust user patience and availability. For that, substantial training and practice are required, but it is costly to create a variety of high-quality HCI training materials. In this context, chat generative pretrained transformer (ChatGPT) and other chatbots based on large language models (LLMs) offer an opportunity to generate training materials of acceptable quality without foregoing specific human characteristics present in real-world scenarios. Personas is a user-centered design method that encompasses fictitious but believable user archetypes to help designers understand and empathize with their target audience during product design. We conducted an exploratory study on the Personas technique, addressing the validity and believability of interviews designed by HCI trainers and answered by ChatGPT-simulated users, which can be used as training material for persona creation. Specifically, we employed ChatGPT to respond to interviews designed by user experience (UX) experts. Two groups, HCI professors and professionals, then evaluated the validity of the generated materials considering quality, usefulness, UX, and ethics. The results show that both groups rated the interviews as believable and helpful for Personas training. However, some concerns about response repetition and low response variability suggested the need for further research on improved prompt design in order to generate more diverse and well-developed responses. The findings of this study provide insight into how HCI trainers can use ChatGPT to help their students master persona creation skills before working with real users in real-world scenarios for the first time.**

***Index Terms*—Chatbots, computer science education, human–computer interaction (HCI), large language model (LLM), training, user-centered design.**

Manuscript received 29 August 2023; revised 16 January 2024 and 22 March 2024; accepted 1 April 2024. Date of publication 8 April 2024; date of current version 25 April 2024. This work was supported in part by the Universidad Politécnica de Madrid, through the project INNOVATION IN HUMAN COMPUTER INTERACTION under Grant RP2210470068 and in part by the CIBER—Consorcio Centro de Investigación Biomédica en Red, Instituto de Salud Carlos III, Ministerio de Ciencia e Innovación under Grant CB06/01/0051. *(All authors contributed equally to this work.) (Corresponding author: Cristian Moral.)*

Jose Barambones, Cristian Moral, Angélica de Antonio, Ricardo Imbert, and Loïc Martínez- Normand are with the Madrid HCI Laboratory, Department of Computer Languages and Systems and Software Engineering, Escuela Técnica Superior de Ingenieros Informáticos, Universidad Politécnica de Madrid, Campus de Montegancedo, 28040 Madrid, Spain (e-mail: j.barambones@upm.es; cristian.moral@upm.es; angelica.deantonio@upm.es; ricardo.imbert@upm.es; loic.mnormand@upm.es).

Elena Villalba-Mora is with the Madrid HCI Laboratory, Department of Computer Languages and Systems and Software Engineering, Escuela Técnica Superior de Ingenieros Informáticos, Universidad Politécnica de Madrid, Campus de Montegancedo, 28040 Madrid, Spain, and also with the CIBER de Bioingeniería, Biomateriales y Nanomedicina, Instituto de Salud Carlos III, Avda. Monforte de Lemos, 28029 Madrid, Spain (e-mail: elena.villalba@upm.es).

Digital Object Identifier 10.1109/TLT.2024.3386095

## I. Introduction

EXPERTS in the field of HCI require thorough training in a number of techniques used to analyze potential users and the context of the use of a computer system before it is designed. HCI education frequently prioritizes the creation of tangible project deliverables, such as wireframes, prototypes, and design materials, as well as focusing on a set of thorough communication and analysis skills for the user research that is required for this purpose [1].

Conducting user interviews can be challenging for different reasons, including confirmation and cognitive biases. Recruiting the right participants and crafting effective questions for eliciting helpful information are preconditions for obtaining valid data to be further interpreted. Incomplete and/or poor quality information gathered in interviews hampers a meaningful analysis of users, tasks, and the future system environment [2], [3]. User archetypes, such as personas, can only be adequately extracted through a correct analysis of the context of use, allowing HCI professionals to design, prototype, and evaluate products according to a human-centered design approach [4].

On the other hand, the emergence of generative AI and the rapid expansion of its use through chatbots trained by large language models (LLMs) has offered an emerging opportunity for learning performance reinforcement and improvement, with Chat generative pretrained transformer (ChatGPT) currently leading this trend due to its effectiveness and availability. In this context, recent case studies have examined the use of chatbots pretrained by these models with respect to the acquisition of pedagogical skills for trainee teachers [5], the effects on students' intrinsic motivation, cognitive load, and learning performance [6], and the use of prompting techniques to obtain quality information and improve learning in flipped classrooms [7]. To the best of the authors' knowledge, however, the potential of these tools for human-centered design learning has not yet been explored.

### A. Personas Technique

In the context of HCI, the Personas technique is frequently used to represent current and potential users of software applications, services, or other products or outputs developed by an organization. As opposed to a user archetype, a persona is a fictional, yet realistic representation of a potential user, exemplifying the key context-associated attitudes of users with respect to a specific product. The definition of a persona varies depending on the specific approach employed, which is driven by the strategy devised by UX experts to identify key user perspectives, including a selection of goals, roles, behaviors, concerns, skills, and user knowledge [8], [9].

Personas are used a great deal in software engineering processes that include UX activities. In their meta-study, Losana et al. [9] evidenced that personas are well-integrated into agile processes for analyzing target users and can be refined according to new iterative findings. This technique is not the result of a unified method and is applied by UX researchers according to different approaches as documented in the literature.

1) *Goal-oriented personas* are the most basic and oldest approach, where archetypes are differentiated by identifying specific and shared goals and objectives that characterize a set of users [10]. This was the approach used to conduct our research, since it is the most straightforward and instructive technique for teaching human–computer interaction (HCI).
2) *The role-based* approach focuses on the personas' lifecycle, which includes and relates to user goals and involved behaviors [11], [12].
3) *The ad-hoc personas*, conceived by Norman [13] as a fast approach that does not require a large body of user research, are based more on designer intuition and experience.
4) *Engaging persona* is a perspective defined by Nielsen [8] and rooted in the idea that stories can produce engagement, insight, and a better understanding of users through vivid and realistic descriptions of characters and narratives.

A persona includes a description of attitudes and knowledge regarding a specific context of use and a selection of key personality traits evidenced when gathering user data, typically through observation, surveys, and interviews [14]. These traits provide helpful clues for designers seeking user engagement and good usability. Anvari et al. [15] studied the influence of this technique on conceptual design. They suggested that personas with different personalities result in a tailored design and help to identify suitable features. Recently, Holzinger et al. [16] proposed an adaptation of the original process to create personas for artificial intelligence (AI), where they were shown to be useful for designing human–AI interaction. However, personas open a space for subjective engagement with data and may lead to stereotypical descriptions that fail to investigate all the user dimensions, merely focusing on behaviors and omitting other aspects, such as fears, motivations, and beliefs [8], [17]. Far from being exclusive to the data source, subjectivity also moves the designer interpreting the data in decision making on which archetypes to focus on. This is even more difficult for inexperienced professionals or learners [18], [19].

Extracting personas from the information gathered during user research (mostly responses to interviews and records from observation) is challenging, requiring designers to create abstractions that condense information and highlight the key issues to be considered in the design. Some research has been performed on simplifying (by partially automating) the analysis of the information collected to create the personas. Nevertheless, the information required to extract a good set of personas has to be of high quality [20]. Therefore, it is undeniable that good training material, such as a quality collection of interviews, is essential to maximize learning opportunities. If the starting material is potentially unsatisfactory, for instance, sourced from interviews conducted by HCI students with their users, there is a risk that students may get sidetracked struggling with deficient information instead of focusing on learning how to create personas.

The rest of this article is organized as follows. Section II reports the preliminary investigation on which this research is based. It highlights the key findings that motivate this work, which laid the groundwork for the subsequent methods. Section III describes the methods followed in the research. It covers the formulated research questions, the research design, and the process enacted during the different steps of our method. Section IV reports the results of the research. It comprehensively analyzes the collected data and showcases the outcomes with respect to the defined hypotheses. Section V discusses the results in depth, exploring, first, the implications and significance of the findings to discuss their strengths and limitations and, second, the design recommendations for leveraging ChatGPT to teach UX techniques, drawn from the insights gained during this study. Finally, Section VI presents the main conclusions and proposes future work.

## II. Motivation

Universidad Politécnica de Madrid (UPM) HCI courses provide bachelor of computer engineering students with their first opportunity to interact with real users through a team project. This firsthand experience exposes students to the challenges and complexities of designing interactive systems catering to user needs, preferences, and behaviors.

The early project work consists of planning in detail how the user research activities are to be performed, defining the recruitment (who, how many, how, when, and where), which data collection methods (question- and observation-based) are to be used, what is to be observed or asked (defining the questions, the tasks to be observed) and preparing the materials needed to collect the data (templates, audio/video recorders, etc.). After students have completed the planned activities, they are asked to analyze and specify, based on the collected information, the three components of the context of use, i.e., users, tasks, and environment. In the specific case of users, students are asked to define user profiles or personas.

As HCI professors, we have observed over the years that students gathering user information in a real scenario find it difficult to acquire the appropriate user-centered design knowledge. Factors such as limited access to real users, time constraints and lack of interview practice and skills are obstacles to the collection of useful data about the context of use from real users.

Besides, the ethics of gaining access to specific user groups or wasting professionals' time for mere training purposes is questionable, at least in the early training stages. All of this highlights that a different approach is needed.

We conducted a preliminary survey of HCI course students across two undergraduate computer engineering programs at UPM to gain a better understanding of their perception of the problems that they face when performing activities related to the analysis of real users and their context of use. The results are reported below.

### A. Preliminary Study Details

A survey was sent to 456 students enrolled in HCI courses as part of the Bachelor of Computer Engineering and Bachelor of Mathematics and Informatics degree programs at UPM. The survey included closed-ended Likert scale questions and open-ended questions concerning their experiences and aimed to identify problems or difficulties encountered during the user research-related activities carried out as part of the course. In particular, questions were designed to gather information on interview design, execution and completeness, recruitment activities, categorization of interviewed users, and activities related to knowledge extraction and the definition of potential users. The survey responses were anonymous, although the survey respondents were known to us. This was required to encourage participation, as students who completed the survey received a 0.5-point (out of 10) bonus to be added to their final course grade. The study was conducted before students were given their final grade to prevent a course pass/fail from influencing the survey results.

### B. Findings

A total of 255 students, representing 56% of the target population, completed the survey. The quantitative and qualitative analysis of their responses provided the following findings.

*1) Recruitment Difficulties:* HCI students highlighted several problems.

1) Scheduling conflicts between team members or potential users, especially when they have busy schedules or conflicting commitments. This can lead to delays in conducting interviews and other user-related activities, impacting the overall project timeline.
2) Complex user profiles: students often face the challenge of finding users who fit the specific criteria required for their projects (i.e., they had to recruit experienced soccer trainers for a project that aimed to develop an app to help trainers track a soccer team's performance). Thus, it can be challenging to identify and recruit users who can actually provide valuable insights and feedback relevant to the project's objectives.
3) Issues with user participation and engagement: students reported difficulties engaging users and encouraging their active involvement. Some users may be hesitant or reluctant to invest their time and provide input for the project.

These problems demotivate student teams. In fact, 71% of respondents stated that they opted to recruit people from their close circle (family and friends) to avoid such situations. This leads to a strong bias in the resulting data and, above all, distorts what the students learn as the situation is not realistic.

*2) Lack of User Personality Variability:* Quantitative analysis must account for personality variability in order to obtain the richest possible data set. Even though 74% of the students identified different personality traits from interviewees, they did not perceive any bias or recognize the importance of practicing with different personality profiles. Most of these students only identified two personalities—extroverted or shy—and did not consider any other possible personality traits that an interviewer may face.

From the learning perspective, students should recognize that practicing with different personality profiles enhances their ability to interview and collect data from a realistic and diverse user community. A more comprehensive and representative understanding of user experiences (UXs) can be achieved by including individuals with varying personality traits.

*3) Interview Quality Shortcomings:* With respect to interview question design, the evidence shows that students need help defining and asking effective questions because of their inexperience (14% of the students), poor definition of questions in substance and form (24%), and unawareness of which questions are important (52%). Regarding interview responses, 58% of the students reported that they encountered difficulties related to the poor quality and quantity of such responses.

From a learning perspective, students must understand that successful data collection goes beyond simply engaging with users, as the design of human-centered solutions requires not only gathering but also effectively analyzing meaningful and high-quality data from users. However, this contradicts the false perception held by many students that the activities that they had conducted with users were sufficient, even though they needed help to successfully complete the milestones. Specifically, 72% considered that they had gathered enough or all of the information required to identify profiles/personas in stark contrast to the above-mentioned shortcomings.

We concluded that students are only able to get enough responses of good quality if they have recruited participants correctly, formulated their questions appropriately, and participants are open to sharing key information.

### C. Exploratory Proposal

Conversational AI applications, commonly known as chatbots, can provide human-like responses to user queries and assist users to reduce their workload, improve task performance, or provide advice in several ways. In particular, ChatGPT is a chatbot application that uses GPT-3.5 and GPT-4, which are transformer-based LLMs developed by OpenAI. In 2022, these models achieved state-of-the-art performance for natural language processing tasks, such as question answering, text generation, and knowledge extraction from documents. Chat apps based on GPT are conversational interfaces where models expect input formatted in a specific chat-like transcript format and return a completion representing a message written by the model within the chat [21]. GPT models have become hugely popular because they can be integrated as AI tools for building chatbots for various purposes, such as customer service, virtual assistants, tutoring activities, document assessment, consulting, research, and other conversational interfaces.

Students have already taken advantage of interactive chatbots in solutions, such as educational collaborative games, social apps, and intelligent tutoring systems [22], [23], [24]. In our context, we hypothesized that ChatGPT could be used to roleplay a set of fictional users intended to perform a wide variety of interviews designed by trainers, which can be used as input material to train students in the Personas technique. On the one hand, this approach allows trainers to generate abundant, meaningful materials without too much effort. On the other hand, such materials serve as input for trainees to learn skills and get insights about the use of interviews to achieve helpful information without "wasting" real users' time. Note that this proposal is not intended to replace fieldwork with real users and environments. Instead, it is proposed as preliminary or intensive training where learners are introduced to and gain knowledge of how to gather data from the user interviews that they analyze in order to create personas before starting any fieldwork.

As a collateral benefit, learners with little user research experience will also learn from a set of well-designed interviews, observing the kind of questions and information that are key to building meaningful personas.

Our proposal is that, by adequately configuring ChatGPT through system prompts and model parameterization, bots could simulate unique users with specific behaviors and experiences, which could be easily interviewed by the trainers to produce a large amount of handy material for students. The research described in this article involves investigating how to create a set of bots that roleplay users according to a real-world use case. Our objective is to assess whether the resulting interviews can be used as the basis for creating personas and how effective these interviews are as training material for the Personas technique.

To sum up, we aimed to study and validate if, properly configured and parametrized, the current version of the LLM-based chatbot ChatGPT can generate fictitious but good-quality and realistic information that can be used as HCI training materials. We proposed ChatGPT due to its fast adoption, extensive use, and impact on the scientific and educational communities.

## III. Methods

Based on this motivation, the central question of this article posits whether an LLM-based chatbot can serve as an effective educational tool by facilitating training in HCI activities and particularly for the analysis of the context of use. Specifically, we hypothesized that the ChatGPT bot, trained solely on the data of the ongoing version, can enhance the educational experience by simulating conversational interactions and, consequently, help the students learn about user archetyping through personas.

### A. Research Questions

Based on this hypothesis, the following research questions were formulated.

*RQ1:* Can ChatGPT represent a fictional but believable user from a proper configuration and a one-shot prompt description?

*RQ2:* Is it possible to employ these fictional users to generate a set of good-quality interviews that can be used as valid input to extract realistic personas?

If the answers to the above-mentioned two questions were positive, we aimed to establish some criteria/patterns/guidance on configuring ChatGPT as a conversational chatbot for user research training.

We intend to make the results of this exploratory research addressing these issues available to the broader HCI research community, encompassing learners, experts, and instructors. By proposing an approach that requires little setup, we aim to provide a valuable contribution that individuals with varying levels of expertise and access to ChatGPT and other LLM-based chatbots can follow.

### B. Research Design

Interviews are one of the best techniques for collecting information to profile system users [25]. There are several types of interviews, which vary depending on how flexible the question definitions are (i.e., structured, semi-structured, and unstructured interviews), the number of users who are interviewed in each session (individual interview versus focus group) or whether the interview is carried out in person, by telephone or via video call [26]. The choice of interview type must be made by identifying the study requirements and analyzing the benefits and disadvantages of each one. In this research, one of the main requirements was that the results of the different interviews had to be comparable. As a result, the questions always had to be the same, and therefore we used a structured interview. Obviously, this choice meant that we no longer had the ability to dynamically adapt to user responses, as we would have been able to do with a semi-structured interview. However, it ensured that the results of the experiment were not influenced by any such variability and could then be replicated.

To answer our research questions, it is necessary to evaluate whether we have succeeded in simulating believable users and if the set of generated responses to the interviews designed by HCI trainers helped analyze the context of use. One of the main challenges in the design of this research was the operationalization of the evaluation of the concepts of "believable user" and "quality set of interviews."

As a playground for experimentation, we used the context and data of a real project developed by the UPM's Vice-Rectorate for Strategy and Digital Transformation, aimed at redesigning the university's intranet. This project is being developed according to a user-centered design process. At the time of research, the context of use analysis activity had already been completed. For this purpose, 122 university-related users (potential or real users of the current intranet) had participated in user research (focus groups, guerrilla tests, and 1-to-1 interviews). Analyzing the collected information, the university IT Service's UX professionals extracted a total of eight personas following Cooper et al. [10], [14] goal-oriented approach. They kindly allowed us to use their user research information for our study (see top block in Fig. 1).

Our research involved the following teams, where individuals were not allowed to be members of more than one group.

1) *Research supervisor:* One author coordinated the teams and prepared and collected data during the experiment.
2) *UX experts:* Two authors transformed the eight goal-oriented personas from the selected domain into a ChatGPT configuration. Their work covered the prompt and

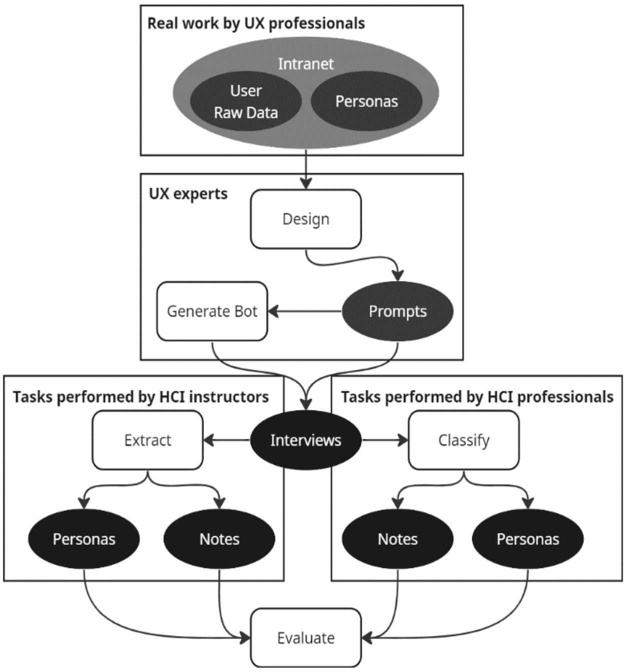



Fig. 1. Diagram of the research design, including teams, their tasks (squares with rounded corners), and outputs (ellipses).

property configuration, and the interview design process. They also generated the interview responses for further evaluation (see central block in Fig. 1).

3) *UX professionals:* Based on the generated responses to the interview designed by the HCI experts, the four members of the university IT Service UX team evaluated the capability of ChatGPT to simulate users consistent with the eight reference personas (see bottom right block in Fig. 1). They are all members of the university's intranet project and have, therefore, prior knowledge of the domain, the real users and the eight personas.
4) *HCI professors:* The remaining three authors had no previous knowledge of the intranet project and extracted personas from the generated interview responses (see bottom left block in Fig. 1). They later evaluated the quality of personas as useful learning material, as well as the capability of ChatGPT to represent valid interviewees for training.

Fig. 1 illustrates the whole process. First, the UX experts interviewed the team of UX professionals and gathered all the information needed to design the prompts for the bot that would be used to generate the interviews. Then, the goal of UX experts was to generate, for each of the eight reference personas, five fictitious users to interview. In other words, 40 fictional users and their corresponding interviews were created.

We asked the UX professionals to classify the answers generated by ChatGPT to the interviews designed by the HCI trainers within the eight reference personas. We considered these fictional users to be "believable" if the same professionals who analyzed the context of use and extracted the eight personas could match each interview with its corresponding persona. Thus, we measured "believability" as the degree of success in classifying the 40 interviews (RQ1).

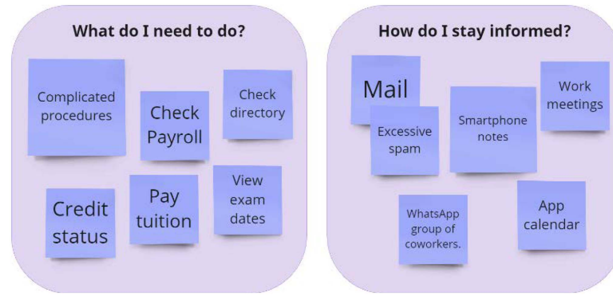



Fig. 2. Examples of descriptive sections of a persona to be filled with information from interviews.

The quality of the material (RQ2) (the ChatGPT-generated answers to the interviews) was measured in two ways.

1) *UX professionals:* Through an assessment of the quality of the material received from the technical point of view, evaluating how closely it represented the domain in which they had worked.
2) *HCI professors:* Through an assessment of the quality of the material received from the educational point of view, including its usefulness for applying the Personas technique.

To evaluate the quality of the set of interviews as a valid input for the persona extraction process from an educational point of view, the team of HCI professors individually extracted their own set of personas from the interviews provided. To rule out bias, they were blinded to the number and characteristics of the initial personas and did not have access to the real intranet redesign project documentation. For ease of comparison, they were only given the interview package and exactly the same template for defining the personas as the model used in the original project.

Finally, a postevaluation was carried out by HCI professors and UX professionals to quantitatively and qualitatively compare the results (explained in Section III-E). We compared the personas generated by HCI professors and the eight reference personas developed by UX professionals. They were not expected to match 100% since there is always some degree of subjectivity in the application of the Personas technique, especially if done individually (personas are usually the result of a consensus among the experts involved in analyzing the collected information) [8], [17], [18], [19]. By analyzing the alignment of personas created by both teams, plus the soundness of the generated interviews, we can gauge the effectiveness of ChatGPT for generating valuable content for HCI technique training, as well as for simulating user characteristics and capturing the essence of real users.

### C. Materials—Persona Template

The persona template used consisted of a schema with a bio and description of the persona. The bio includes a photograph, a persona name, the persona's role/profile, and a quote summarizing the persona's attitude, concerns, and goals. The persona description consists of sections that should be completed with information from the interviews, as illustrated in Fig. 2. The sections to be completed are as follows.

1) *Who am I?*—Persona's demographics.

2) *What do I do?* —Persona's role and occupation.
3) *What are my daily actions?* —Day-by-day interactions with the current intranet.
4) *What do I need to do?* —Tasks the persona needs to accomplish.
5) *Why do I need to do it?* —Motivation behind the persona's actions.
6) *How do I stay informed?* —Sources of information for daily work.
7) *What do I value positively?* —What is working fine with the current intranet?
8) *What complications do I have?* —What is not working correctly in the current intranet?
9) *How do I solve it today?* —Strategies used to overcome current difficulties.

This template, which the UX professionals originally used during the creation of the eight reference personas, was given to the HCI professors as a basis on which to build their personas as a constraint for proper comparison.

### D. Procedure for Interview Creation

We defined a procedure whereby ChatGPT would be used to generate a collection of interview subjects (users) based on each reference persona. The users should be different enough from each other to give the impression that the interviews have been conducted with different people and similar enough to respond to the particularities of the same persona. This implied ensuring that ChatGPT had sufficient levels of both variability and consistency.

First, a thorough investigation was conducted of the configuration options offered by ChatGPT. As a result, we concluded that there were two mechanisms on which we could act: the prompt and the model configuration variables.

A prompt in ChatGPT refers to the initial text input or command given to the model to start generating a response. Different types of prompts can be used.

1) *Text input for conversational behaviors:* A distinction can be made between *single-turn prompts,* in which only the user's message is provided, and *multiturn prompts,* which include both the user's message and the model's previous response.
2) *Prompt types for conversational behaviors:* To configure interactions, ChatGPT prompts can be divided into *user prompts,* conveying messages or queries from the user, or *system prompts,* which provide instructions and additional information to drive the model's behavior.
3) *Prompt approach for conversational behaviors: One-shot prompts* provide a single thorough instruction as input, allowing the model to generate a coherent response for the new context. In contrast, *multiturn prompts* entail a more continuous interaction, providing a sequence of exchanges between the user and the model.

However, there are limitations when using only prompts to configure a conversation and control the model's behavior. While ChatGPT now boasts a larger context window (the recent conversation history that the LLM maps to generate relevant and coherent responses), it does not have a built-in memory of previous interactions. Therefore, it can sometimes lose the context of a conversation. In addition, bias in the training data influences the model's responses and may generate inaccurate or inappropriate information. Careful monitoring and fine-tuning are sometimes necessary to ensure that the model's behavior aligns with the desired outcomes. To complement prompts, ChatGPT can also be configured using several variables (called hyperparameters) to customize its behavior and optimize the conversation experience by controlling the diversity and creativity of responses. Some of the key variables for the HCI community, which were used in this research, are as follows.

1) *TEMPERATURE:* This hyperparameter controls the randomness of the model's predictions in its responses. A value close to 1 introduces more variability and creativity in the generated text, while a value close to 0 produces more focused and deterministic responses.
2) *PRESENCE_PENALTY:* This hyperparameter rewards/penalizes word repetition. When the value is positive, the model pays a "penalty" each time it repeats a word or phrase, which means that it can be more diverse and talk about new topics.
3) *FREQUENCY_PENALTY:* This hyperparameter rewards/penalizes word frequency. When it is positive, the model is discouraged from using high-frequency words, which prevents repetition of phrases (line verbatim) and can lead to more diverse or creative texts.

By tweaking the model configuration variables, we can control the diversity of the operating mode of ChatGPT as a language model. At the same time, prompt design must ensure consistency with the reference persona, while allowing the simulation of differentiated users.

The general outline of a ChatGPT interview conforms to the following model.

1) Initial simulation setup prompt.
2) The sequence of pairs (prompt-response) corresponding to the interview questions and ChatGPT responses.

Once the available mechanisms had been identified, a preliminary study was carried out to test ChatGPT operation using different configurations of the variables and the initial prompt in order to analyze the impact of such manipulations on the resulting interviews. Fig. 3 summarizes the iterative process used to configure and validate interview generation.

Although we had a ChatGPT 4 license, there was a limit of 25 interactions every 3 h at the time of the research. Therefore, we decided to run this initial exploration using ChatGPT 3.5 Playground and its web API.

We randomly selected the data of one of the personas and proposed an initial prompt instructing the bot to act like a university intranet user (user definition). Then, all the specific persona data were compiled in different lines (user description). Interviewing the bot with this prompt, we discovered some interesting tips for avoiding poor or robotic behaviors and enhancing helpful and believable answers.

After versioning our simulation setup prompt four times, we were satisfied with the quality of the interview, extracting some lessons learned.

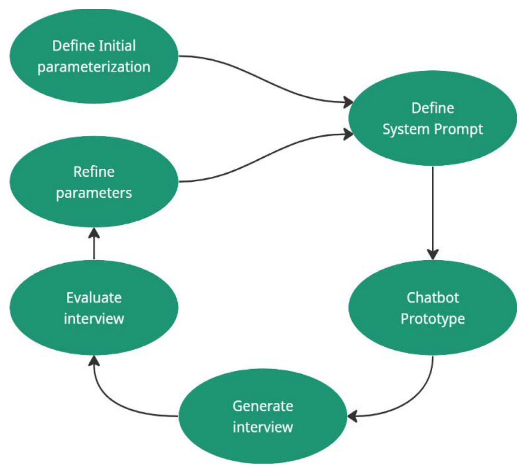



Fig. 3. Interview generation process.

1) It is better to provide only essential information without going into too much detail over specific data. When ChatGPT is flooded with information, it tries to incorporate it all into its responses, resulting in clone-like interviews with little space for improvisation or elaboration.
2) It is crucial to curate the tone of the interviewee. The mere specification of age and role is insufficient to ensure, for instance, the colloquial style of a 22-year-old student. The reference to university members leads ChatGPT to adopt a rather formal manner, which is not believable in most cases. The introduction of the intended tone of the fictional user into the user definition improves the vocabulary and sentence structure (i.e., "as a user of the intranet, act like a young person in higher education, using polite but colloquial language").
3) ChatGPT sometimes forgets some of the initial instructions during the interview and gives responses that ignore the definition. We discovered that it is generally sufficient to instruct ChatGPT to maintain the definition throughout the interview to ensure consistency.
4) ChatGPT overuses lists and enumerations, resulting in unnatural responses to questions compared to how a person would spontaneously answer interview questions. It has to be explicitly instructed to avoid such enumerations.
5) ChatGPT can be given several alternatives and asked to make a choice. This is especially useful when the user description contains a limited number of acceptable answers on a particular aspect. It is also essential to remind ChatGPT that, during the interview, it should not reveal that it is a virtual assistant since it has a strong tendency to use statements like "as a language model…" in its responses.

Once the prompt strategy had been established, we generated combinations of values for the afore-mentioned three configuration variables. We chose three possible values for each variable: 0 (no effect), 0.7 (default value), and 1 (high value). Then, we also tried values above 1 for each one.

As a result of this initial exploration, we drew some conclusions about the best way to configure ChatGPT hyperparameters for our purposes. Thus, it was verified that the following configuration of model variables provided the best results.

1) *TEMPERATURE (range [−2, 2])*: 1.0. Lower values offer fewer creative dialogs, affecting interview believability. A temperature above 1.0 causes poorly constructed sentences with grammatical and spelling mistakes.
2) *FREQUENCY_PENALTY (range [−2, 2])*: 1.0. Lower values produce repetitive sentences, which would sound unnatural for a human being. Again, values above 1.0 cause poorly formulated and incorrect sentences.
3) *PRESENCE_PENALTY (range [−2, 2])*: 1.0. Lower values result in information being repeated in different responses, which is unnatural. And above 1.0, there were again problems with sentence structure.

After the different tests run on ChatGPT 3.5 and later on ChatGPT 4, we arrived at a prompt design with the following structure.

1) Tell ChatGPT who the target person to be simulated is (using a general definition like a university intranet user or a football team coach).
2) Ask ChatGPT to answer a series of questions for a study, applying a specific configuration of the model (in ChatGPT 4, this is specified using brackets).
3) Ask ChatGPT to act as a person with specific characteristics (name, age, gender, educational level, and type of language spoken).
4) Ask ChatGPT not to reveal that it is a virtual assistant during the conversation.
5) Ask ChatGPT not to use enumerations in its responses.
6) Ask ChatGPT to respond to questions according to a specification that will be delimited by “”".
7) Include the specification dependent on the context of use (delimited by “”").
8) Ask ChatGPT to acknowledge that it has understood the instructions.

As for the more general characteristics of the user to be simulated (name, age, gender, educational level, and type of language spoken), we decided to ask ChatGPT to invent a name, age, and gender in order not to have to make specific changes to the prompt for each user to be simulated. It proved to be able to do so. In some cases, however, inconsistencies were observed throughout the interview (e.g., it first answered saying that it was a woman and then went on to respond using the male gender). This problem may have been exacerbated by the fact that the interviews were conducted in Spanish, given that the university intranet project is being developed in Spanish and the UX professionals mostly used this language during their user research. Gender is much less explicit in English, and ChatGPT would have behaved better in this regard if we had used this language. To address this issue, we added an instruction that the same name, age, and gender should be used throughout the interview.

We found it necessary to include a final request at the end of the initial prompt asking ChatGPT 4 (2023 May 3 version) to answer "OK" if it had understood the instructions. This avoids the undesirable behavior of responding to this first configuration prompt by revealing information about itself that should only be provided if it is asked a question in this regard. This is not necessary in ChatGPT 3.5 since the initial prompt text is entered separately as a system prompt, and there is no chatbot response. In ChatGPT 4, however, the system treats the first prompt like any other prompt and always reacts by giving an answer. OpenAI had not published the API for entering ChatGPT 3.5-style system prompts in ChatGPT 4 at the time of the research.

After the preliminary study had been concluded, the UX experts wrote eight initial prompts (one for each of the eight reference personas, see Appendix A for an example of these prompts). They interviewed ChatGPT five times using each prompt, generating 40 interviews that were copied into separate files, eliminating the initial prompt and any trace that might lead a reader to believe that the interview had been generated using ChatGPT.

The interview was an abridged version of the real interview used in the intranet redesign project and was composed of eight questions. See Appendix B for a transcript of one of the 40 answers to the interview design by the HCI trainers generated by ChatGPT.

### *E. Procedure for Interview Analysis*

As explained above, these 40 responses were assessed according to a dual research procedure. Each UX professional analyzed the generated conversations, identified critical information, and classified the interviews within the initial personas. Each HCI professor assessed the educational value and learning potential of these answers. This involved considering the extent to which the interviews effectively represented and simulated real user interactions, the quality of the insights and resulting information, and the possibility of "re-personification." In our context, repersonification stands for creating new personas from chatbot interactions to evaluate whether interviews with ChatGPT-simulated users are valid inputs for the Personas technique. This process aims to replicate the workflow of a real-world UX expert in extracting and synthesizing user data.

Interviews were shared with HCI professors and UX professionals separately according to their purpose: professors had to repersonificate (create new personas based on the responses given by ChatGPT during the interviews), whereas professionals had to match the interviews to the eight reference personas used to generate ChatGPT bots. Note, as discussed above, that we did not expect the professors to generate exactly the same personas as the UX professionals since persona creation is a subjective task and the input material differed. Neither is the quality of the original work of the professionals under evaluation. To minimize possible biases on repersonification, each HCI professor was assigned an unnamed set of 22 out of the 40 interviews, corresponding to 5 or 6 personas out of the 8 original personas, and different numbers of interviews per persona were included in each set (from 2 to 5). We tried to minimize the possibility of inferring a global data pattern from the structure of already identified personas. We ensured that each of the original eight personas was assigned to two different HCI professors, although they received a different (albeit overlapping) input set of interviews. Then, they were each asked to design their personas individually using the same template as the professional team. Similarly, each professional member separately tried to match interviews to the existing personas.

At the end of the procedure, the members of both teams were asked to complete a questionnaire about their experience in order to measure the quality of the interviews. Questionnaires consisted of two subsets of Likert and open-ended questions adapted to each team for different measures. The metrics were inspired by Han et al. [27], which addressed the design of effective interviews with chatbots. Both questionnaires (see Table VI in Appendix) aimed to assess the following aspects.

1) *Interview design:* Questions related to measuring the quality of the interview design, which is the basis of the experiment for data extraction from the bot as interviewee. This included the order, number, and definition of questions.
2) *Informativeness:* Questions related to how relevant the information provided by bot answers were for extracting knowledge enabling HCI designers to create meaningful user archetypes in the shape of personas. This included the difficulty of gathering data regarding HCI/UX expert skills. This section contained four questions for professors and two extra questions for professionals to compare insights and soundness with their previous personas.
3) *Soundness:* Questions related to measuring the coherence of answers about their self-experience as HCI designers with real users. This included identifying useless information, contradictions, unelaborated responses, ease of explanation, etc.
4) *Learnability/educational utility:* For HCI professors only, these questions measured the potential of the responses generated to the designed interviews being used as learning content for personas training. This subset of questions was omitted for professionals, as this issue is beyond their field of expertise.
5) *UX:* These questions aimed to measure how closely bots emulated real interviewees with a view to effective user interviewing practice. This included measuring the capability of detecting behaviors, attitudes, patterns, simulated background, and possible emotions and their alignment to produce a bot that is consistent as a potential user. This section contained eight questions for professors and two extra questions for professionals to measure the proximity of the responses to the domain on which this case study is based.
6) *Ethics:* We included questions to evaluate the interview ethics where private and sensitive information, fictional or otherwise, could have been exposed.

## IV. Results

### *A. Repersonification by HCI Professors*

Each HCI professor received 22 out of the 40 answers generated by ChatGPT and the template to create the personas. Of the eight original personas used to configure the ChatGPT bot, seven were identified by the HCI professors. Only the *new (administrative) staff* persona was not recognized.

Table I shows the results of the repersonification performed individually and separately by the three HCI professors, detailing the match that they each achieved. Overall, professors correctly

TABLE I
RESULTS OF THE REPERSONIFICATION PERFORMED BY HCI PROFESSORS

| | HCI professor | | |
|---|---|---|---|
| | P1 | P2 | P3 |
| First-year student | - | 5 / 5 | 3 / 3 |
| Last-year student | - | 3 / 3 | 5 / 5 |
| Exchange student | - | 5 / 5 | 3 / 3 |
| Ph.D. student and new professor | 3 / 3 | 5 / 5 | - |
| Experienced professor and research group director | 5 / 5 | 4 / 4 | - |
| Administrative staff who regularly use the intranet | 9 / 5 | - | 6 / 4 |
| Administrative staff who occasionally use the intranet | 5 / 5 | - | 5 / 5 |
| New staff | - / 4 | - | - / 2 |
| Degree of coincidence | 81,82% | 100,00% | 90,91% |

The first value represents the number of interviewees assigned to a given persona by the professor. The second value stands for the persona to which the interviewee belongs.

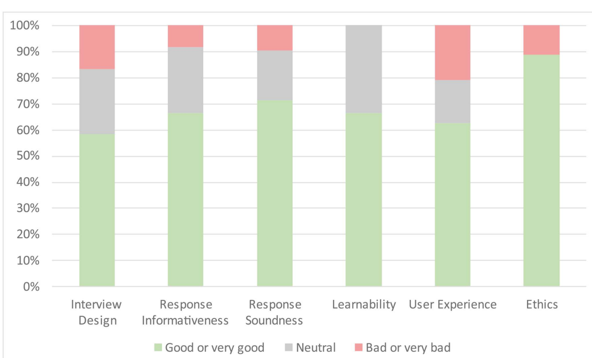



Fig. 4. Aggregated results of the quantitative questionnaire for HCI professors.

assigned the interviewees to the original personas in 90.91% of the cases.

### B. Match With Original Personas by UX Professionals

In this case, all UX professionals received all the interviews and assigned them to the existing personas. All the professionals were able to match all the interviewees to their respective persona, achieving a 100% match.

### C. Questionnaire Responses

Table II shows the responses provided by the HCI professors to each question, whereas Fig. 4 synthesizes the results grouped by categories. Finally, Table III summarizes the responses to the open-ended questionnaire questions.

In addition, Table IV shows the responses provided by the UX professionals to each question, whereas Fig. 5 synthesizes the

TABLE II
RESPONSES TO THE LIKERT-BASED QUESTIONS BY HCI PROFESSORS

| | | Strongly disagree | Somewhat disagree | Neutral | Somewhat agree | Strongly agree |
|---|---|---|---|---|---|---|
| Interview Design | 1 | 0 | 0 | 0 | 3 | 0 |
| | 2 | 0 | 0 | 0 | 1 | 2 |
| | 3 | 0 | 1 | 2 | 0 | 0 |
| | 4 | 0 | 1 | 1 | 1 | 0 |
| Answers Informativeness | 5 | 0 | 0 | 2 | 1 | 0 |
| | 6 | 0 | 1 | 1 | 0 | 1 |
| | 7 | 2 | 1 | 0 | 0 | 0 |
| | 8 | 0 | 0 | 0 | 2 | 1 |
| Answers Soundness | 9 | 0 | 2 | 0 | 0 | 1 |
| | 10 | 0 | 1 | 1 | 0 | 1 |
| | 11 | 0 | 0 | 0 | 0 | 3 |
| | 12 | 2 | 0 | 1 | 0 | 0 |
| | 13 | 0 | 0 | 1 | 1 | 1 |
| | 14 | 2 | 1 | 0 | 0 | 0 |
| | 15 | 0 | 0 | 1 | 1 | 1 |
| Learnability | 16 | 3 | 0 | 0 | 0 | 0 |
| | 17 | 0 | 0 | 2 | 1 | 0 |
| UX | 18 | 0 | 0 | 1 | 1 | 1 |
| | 19 | 0 | 0 | 0 | 3 | 0 |
| | 20 | 0 | 0 | 1 | 2 | 0 |
| | 21 | 0 | 1 | 1 | 1 | 0 |
| | 22 | 0 | 0 | 0 | 3 | 0 |
| | 23 | 0 | 0 | 0 | 3 | 0 |
| | 24 | 0 | 0 | 0 | 1 | 2 |
| | 25 | 0 | 1 | 1 | 0 | 1 |
| Ethics | 26 | 3 | 0 | 0 | 0 | 0 |
| | 27 | 3 | 0 | 0 | 0 | 0 |
| | 28 | 1 | 0 | 0 | 0 | 2 |

TABLE III
RESPONSES TO OPEN-ENDED QUESTIONS BY HCI PROFESSORS

| |
|---|
| **If it is the case, what repetitive patterns have you identified in the responses?** |
| All HCI professors consider that several answers were exactly the same, or with very slight differences. |
| **What tones have you identified in the interviewees (neutral, polite, aggressive, direct, sarcastic, etc...)?** |
| Two of the HCI professors consider that most of the interviewees had a neutral and polite tone. The third HCI professor, on the contrary, states that there were quite diverse tones (neutral, polite, direct and colloquial). |
| **What ambiguities have you detected in the answers?** |
| Two of the HCI professors state that the answers lacked detail and did not provide enough information. The third one considers that there was an inconsistency, as foreign students indicated that they did not speak Spanish but, on the contrary, they answered the questionnaire (in Spanish) perfectly. |
| **What information has been the easiest to obtain during data extraction? Why?** |
| Two of the HCI professors consider that the difficulties have been the easiest information to extract, whereas the third one states that all the information was easy to extract. |
| **What information has been the most difficult to obtain during data extraction? Why?** |
| All HCI professors have found difficulties extracting how problems and difficulties are overcome by interviewees. One of the HCI professors also found problems filling the information related to the dedication of the interviewee, his/her daily actions and for what he/she needs/wants it, as the questions are very open and abstract and then do not allow to obtain such in-depth information. |

TABLE IV
RESPONSES TO THE LIKERT-BASED QUESTIONS BY UX PROFESSIONALS

| | | Strongly disagree | Somewhat disagree | Neutral | Somewhat agree | Strongly agree |
|---|---|---|---|---|---|---|
| Interview Design | 1 | 0 | 0 | 1 | 2 | 1 |
| | 2 | 0 | 0 | 0 | 1 | 3 |
| | 3 | 0 | 0 | 1 | 2 | 1 |
| | 4 | 0 | 0 | 0 | 3 | 1 |
| Response Informativeness | 5 | 0 | 0 | 1 | 2 | 1 |
| | 6 | 0 | 0 | 0 | 3 | 1 |
| | 7 | 0 | 0 | 0 | 2 | 2 |
| | 8 | 3 | 1 | 0 | 0 | 0 |
| | 9 | 0 | 1 | 0 | 0 | 3 |
| | 10 | 0 | 0 | 0 | 2 | 2 |
| Response Soundness | 11 | 0 | 4 | 0 | 0 | 0 |
| | 12 | 0 | 1 | 2 | 1 | 0 |
| | 13 | 0 | 0 | 0 | 0 | 4 |
| | 14 | 1 | 3 | 0 | 0 | 0 |
| | 15 | 0 | 0 | 0 | 2 | 2 |
| | 16 | 1 | 3 | 0 | 0 | 0 |
| | 17 | 0 | 0 | 0 | 1 | 3 |
| UX | 18 | 0 | 0 | 0 | 2 | 2 |
| | 19 | 0 | 0 | 1 | 2 | 1 |
| | 20 | 1 | 1 | 1 | 0 | 1 |
| | 21 | 0 | 1 | 0 | 2 | 1 |
| | 22 | 0 | 0 | 0 | 3 | 1 |
| | 23 | 0 | 0 | 0 | 2 | 2 |
| | 24 | 0 | 0 | 1 | 1 | 2 |
| | 25 | 0 | 0 | 1 | 2 | 1 |
| | 26 | 0 | 0 | 0 | 0 | 4 |
| | 27 | 0 | 1 | 0 | 1 | 2 |
| Ethics | 28 | 4 | 0 | 0 | 0 | 0 |
| | 29 | 4 | 0 | 0 | 0 | 0 |
| | 30 | 0 | 0 | 0 | 0 | 4 |

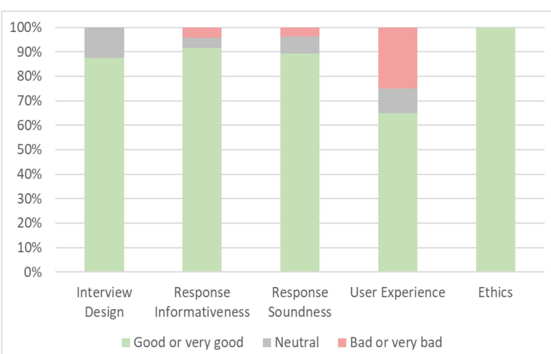



Fig. 5. Aggregated results of the quantitative questionnaire for UX professionals.

results grouped by categories. Finally, Table V summarizes the responses to the open-ended questions of this questionnaire.

TABLE V
UX PROFESSIONAL RESPONSES TO OPEN-ENDED QUESTIONS

| |
|---|
| **If this is the case, what repetitive patterns have you identified in the responses?** |
| All UX professionals consider that several answers were very similar. |
| **What tones have you identified in the interviewees (neutral, polite, aggressive, direct, sarcastic, etc...)?** |
| All UX professionals consider that there are different tones that match with the specific personas' background, providing trustability to the results. |
| **What ambiguities have you detected in the answers?** |
| None of the UX professionals have detected ambiguities or inconsistencies. |
| **What problems did you find matching the interviewees with your existing personas?** |
| All UX professionals agree that it has been very easy to match the interviewees with the original Personas, especially with the answer to the introductory question about the interviewee background. |

## D. Research Question 1

The first research question is whether ChatGPT can represent a fictional but believable user from a proper configuration and a one-shot *prompt* description.

The quality of the answers provided by ChatGPT to the interview designed by the HCI trainers was evaluated based on two questionnaires: one questionnaire for HCI professors and a slightly different one for UX professionals (see Fig. 4 and 5). In general, the interview items were better rated by professionals than by professors, which is probably because the UX professionals were already acquainted with the project personas and they performed a simpler task (classification) than the HCI professors (creation of personas). Below is a summary of the main findings by questionnaire section.

1) *Interview design:* UX professionals gave very positive feedback (almost 90% of good or very good, and no negative responses), whereas HCI professor feedback was less positive (almost 60% of good or very good, and some negative responses). The main issues identified by professors were missing questions or missing information required to create personas (especially the section on how users currently solve the problems that they encounter). Note that the number of interview questions was reduced on purpose to make the research more manageable and deal with the use-time constraints imposed by ChatGPT. Moreover, some information was missing or very concise in the original UPM intranet project personas, and, therefore, we were unable to provide ChatGPT with enough information in the initial prompt without inventing details. As a result, ChatGPT could not provide this information in its responses. This aspect could be easily improved in the future by extending the amount of information included in the prompt. This could be done by designing longer interviews and by improvising new questions throughout the interview in order to further explore issues on which the information initially provided by ChatGPT is considered insufficient.
2) *Response informativeness:* Again, the feedback from UX professionals is much better than from HCI professors (90% versus 65% of good or very good responses). Professors found it easy to create the personas based on the interviews, even though some information was missing. The additional two questions answered by the professionals did not provide any further findings.
3) *Response soundness:* UX professionals again provided more positive feedback than HCI professors (almost 90% compared to 70%) regarding the consistency of the responses related to their own experience with real users.

The most interesting finding is that there is disagreement on whether or not the responses were too short. This highlights some differences in the personal preferences of professionals/professors. All the other points were well rated.

4) *Learnability:* Only HCI professors completed this section only, with around 65% giving ratings of good/very good. They agreed that the interviews were simple enough, but two professors were noncommittal with respect to whether they were suitable for learning how to create personas. This issue was probably influenced by the deficiencies noted in the interview design (missing questions) and content (missing information).
5) *UX:* There is more agreement between experts and professors on this issue, with good/very good accounting for around 65% of responses. It is also the issue that attracted the most negative responses. Expectedly, the ChatGPT-generated answers to the interview designed by the HCI trainers were regarded as only partially natural and believable. A significant issue is interview repetitiveness. We found that, when given a list of alternatives, ChatGPT preferred some options over others, with an uneven distribution that had not been anticipated during the prompt creation process. Further research needs to investigate how to improve response randomization.
6) *Ethics:* This is the issue with the best results for both groups. Only one of the professors found that the responses were not fully ethically accountable because of the uneven gender distribution of the synthetic users that were created. ChatGPT had been asked in the prompt to select a random name and gender (in this order) for each new interviewee. We discovered that the proportion was balanced by interchanging this order.

As far as the responses to the qualitative questions are concerned, the main findings were as follows.

1) The interviews sounded repetitive, and the responses to some questions were, in some cases, the same. This is because the prompts and settings were designed to rule out the possibility of ChatGPT inventing responses for this controlled experiment.
2) The responses were superficial and did not provide enough information. ChatGPT was mostly unable to expand on the information given in the prompt.
3) Language inconsistencies were detected. In one case, the "user" was an international student claiming not to speak Spanish (the interview language) who produced responses in excellent Spanish. Second, there were some, but not too many, gender inconsistencies. We believe that this is partly due to the use of Spanish for the interviews, as ChatGPT is more limited to non-English languages.
4) The interviews did not contain all of the information that was required to complete the personas template. This is due to missing information in the original personas.
5) Finally, the classification of the interviews into personas was straightforward given response repetition.

In summary, even though there are some issues with the generated answers to the interviews, mostly associated with the constraints imposed by the study, the answer to RQ1 is, according to the results of the evaluation, positive. When each interview is considered separately, the UX professionals and HCI professors found them to be believable users in the context of the system.

Further research is needed to improve the prompts and settings so that the system can provide better and more diverse interview responses, bearing in mind that the goal is to create a set of different interviews based on a single prompt.

### E. Research Question 2

The second research question is whether it is possible to employ these fictional users to generate a quality set of interviews that can be used as valid input to extract realistic personas in a training situation.

This has been tested in the experiment in two ways. First, the UX professionals unequivocally matched all the interviewees with their respective persona. This means that the process used to create the prompts from the personas enables ChatGPT to provide the correct information in response to interview questions.

Second, and more importantly, the HCI professors were able, except in one case—the persona representing a new administrative staff member—to analyze the set of interviews and create new personas that matched the original ones.

A detailed analysis of the interviews for that specific persona revealed that it represented staff from different services, and the professors tended to classify staff based on their service rather than on the fact that they were all new to the university.

In summary, the response to the second research question is also positive, despite the abovementioned limitations found in the set of interviews. The major limitation of the collection of interviews is the repetitiveness of their content, which was mainly due to: 1) the prompts limiting the possible responses that ChatGPT could provide in order to maximize consistency and 2) the mismanagement by ChatGPT of the possible alternatives, with a nonuniform distribution of the chosen option.

## V. Discussion

### A. Strengths, Flaws, and Possible Biases

This exploratory research has demonstrated the viability of using ChatGPT to impersonate users and answer interviews provided that the user characteristics, needs, and expectations are described in an appropriate prompt. It has also demonstrated the feasibility of generating several different interviews based on a single prompt, albeit with some issues with regard to reduced response diversity.

This is, in fact, the major flaw of the approach applied in this research. The prompts created for each persona tried to force ChatGPT to provide correct answers by listing several valid alternatives for each user feature, behavior, or need. The current versions of ChatGPT have shown two limitations in this respect. First, ChatGPT did not elaborate on these responses and merely copied the answers verbatim in its dialog. Second, ChatGPT did not apply a good random selection of the alternatives, choosing some options much more frequently than others. This led to excessive repetition among the interviews generated for the same persona.

In the future, we will explore options for ChatGPT to rephrase the concepts expressed in the prompt in several ways, as well as the possibility of using the new ChatGPT Code Interpreter plug-in, which can interpret Python code, to program a randomizer algorithm to select alternatives.

Regarding ChatGPT behavior, gender selection was wholly biased towards males. We asked ChatGPT to select the gender of the interviewee, resulting in 39 males out of 40 interviewees. We need to explore whether this may have been influenced by the use of gender in the prompt, given that Spanish is a strongly gendered language, which may have induced ChatGPT to consider itself male. Another possibility is that the gender bias directly derives from how the GPT has been trained.

Finally, we believe that the research that was conducted, and especially the repersonification, may be subject to some bias. The HCI professors were not informed about the initial personas and had to identify them based on the information included in the answers provided by ChatGPT. However, the project topic (the university intranet) is closely connected to the daily work of the professors, who are familiar with and use the intranet tools and who are also acquainted with the major types of user roles (students, administrative/services staff, and professors).

### B. Design Recommendations

One of the major challenges in developing a robust model for interviews and reliably configuring our synthetic characters was the restriction imposed by OpenAI, which limited the number of prompts to 25 every 3 h when using ChatGPT 4. This limitation significantly impacted the efficiency of our trial-and-error process. Consequently, we chose to refine our experiments using the unrestricted sandbox offered by ChatGPT 3.5, with the expectation of smoothly transferring our results to the latest version later.

While most of our assumptions held true across both versions, transitioning our configuration to ChatGPT 4 was not seamless. Interviewees behaved differently, which meant that much of the fine-tuning work had to be repeated in the new platform. Surprisingly, changes to the configuration parameter ranges (extended to include negative values in ChatGPT 4) did not affect the behavior of our fictional characters, and the previously selected values continued to work satisfactorily. However, we had to run another trial-and-error process to retune the manner in which instructions were given in prompts, as interviewee behavior changed subtly, impacting the believability of the results. For example, bot instructions to use “colloquial language” for a young student resulted in the expected language register in ChatGTP 3.5. In ChatGPT 4, however, the same instruction provoked the use of overly casual language in the student interview. Thus, the prompt had to be refactored to request “colloquial but polite language.” A key recommendation is not to assume that both versions will behave similarly, emphasizing the importance of using the same version for both experiment design and execution.

Concerning interview design, our initial version drew inspiration from the interviews UX professionals crafted for their user research in the intranet project. It comprised 11 questions aimed at extracting interesting information and delving deeper into valuable details with each successive question. However, we found that ChatGPT attempted to offer all related information about the subject in response to each new question. This led to peculiar interviews where nearly all information was revealed in the early questions, and subsequent questions redundantly provided the same information. To address this robotic, repetitive behavior, we simplified the interview to eight questions, including an introductory item. Each question focused on a specific topic, and subsequent questions were designed not to gather further information on the same topic. Our observation indicates that ChatGPT does not consider the option of rationing its information for later questions and tends to share all its knowledge as soon as possible. A potential solution for more natural results may be to provide the entire list of questions and ask ChatGPT to generate the complete interview. Alternatively, concise interviews appear to yield better results.

We also found that, unless clear instructions are provided, there is no guarantee of diversity if the choice of the variability in responses of individuals with similar profiles is left to the model. The system tends to use all the information to which it has access in each interview, resulting in responses that are overly homogeneous. To address this, we employed two strategies. When variability depended on a possible attribute value, we instructed the model to select it randomly, either freely (e.g., “make up a name”) or from a range of accepted values (e.g., “choose an age between 21 and 26 years”). When the variation depended on selecting from a set number of options, we directly asked the model to choose the desired number of items from a list (e.g., “choose only two tools that you usually access from the intranet from the list delimited by parenthesis and separated by a semicolon: [ …]”). In this case, our unconfirmed impression is that the system attaches more importance to the first items on the list. Further research is needed to establish a better strategy for randomization.

## VI. Conclusion and Future Work

The research described in this article is a first step in the use of ChatGPT to enhance learning about user research as part of HCI courses. When given adequate instructions through prompts and settings, this chatbot can impersonate users providing information on who they are, their needs, and what issues they have. In addition, the same one-shot prompt can be used to generate a set of different interviews, although more research is needed to improve the diversity of the system responses.

In this process, we have learned some valuable lessons.

1) It is crucial to have a good understanding of how a given LLM-based chatbot works and the settings and prompts that can be used to give instructions. The quality of the responses provided by the system impersonating a user depends on the prompt contents and selected parameters. Different versions of the same chatbot can even lead to significant differences in the results.

2) When designing the prompts, it is essential to strike the right balance between constraining system responses to exclusively correct answers and giving it enough flexibility to avoid repeated responses. So far, trial-and-error has been the only way to achieve such a tradeoff. However, we expect new system capabilities (like code interpreters) to facilitate this task.

The results of this research are encouraging, and we plan to conduct further research in the future on the application of ChatGPT and other chatbots as tools to enhance user research learning. Two initial activities are planned. First, as a continuation of the reported research, an LLM-based chatbot will be used to enable students to practice the creation of personas from a set of interviews. Given a specific subject, the professor will use the bot to generate a set of interviews based on some pre-defined personas. Then, students will analyze the resulting interviews to identify and populate personas descriptions. Second, a LLM-based chatbot will be used to help students practice interview skills. The professor may set up an interview session with a one-shot prompt and then let the students freely interview the user impersonated by the system to gather as much helpful information as possible. We expect this exercise to raise student awareness of the importance of the quality of the questions that they ask and to improve their interviewing proficiency. Our study marks a step towards integrating these AI technologies into HCI training by using ChatGPT to generate training materials. In no case does it intend to replace real users, which are the foundational basis of human-centered design. The use of ChatGPT not only showcases the practical applications of LLMs but also gives a glimpse of a future where they become an integral part of educational frameworks. As we continue to explore, interact, and refine these technologies, the possibilities for enhancing and transforming educational practices in HCI are promising.

Indeed, through this approach, we aspire to contribute to the growth and development of the HCI learning community.

## Appendix

### A. System prompt for the persona of Lucas (reduced version)

You are a user of a **university intranet**.

As an *intranet user*, act as a *young* person who uses *colloquial but polite language*, making up a name, a gender, and an age within the range of *21* and *26* years that you will maintain throughout the interview.

Respond to my questions for a context of use study. Do not reveal that you are a virtual assistant during the conversation. Answer, without enumeration, with the following description *delimited by triple quotes*:

“””

you are studying *engineering at the university*,

you are finishing *your degree*,

you use *the intranet very sporadically*,

you use *your computer and smartphone* to access the intranet,

*choose only one* of the options *delimited by parentheses and separated by a semicolon* as *activity in the university*:

(

1. you are working as an intern;
2. you are doing curricular internships at a research group at a school of the university;
3. you are doing curricular internships at a company;

),

*choose only one* of the options *delimited by parentheses and separated by a semicolon* as *something you don’t like about the intranet*:

(

1. you find it difficult to find the information on the university’s intranet;
2. you do not think the intranet is well adapted for consultation from a smartphone;
3. you find it difficult to access information and regulations regarding scholarships on the university’s website;
4. you think that WhatsApp groups and conversations with friends are more useful for keeping up to date;
5. you find that the institutional language that the university uses on the intranet is overly formal and uninviting;

),

…

“””

### B. Example of interview in GPT-4 (translated from Spanish)

**-***Interviewer*: Hello, I am collecting information for a study of the context of use of the university’s intranet. Can I interview you?

**-***GPT as Lucas*: Hello! Sure, I would be delighted to help you with your study. Ask whatever you need.

**-** *I*: To begin with, please introduce yourself: What is your name, how old are you and what are you doing at the university?

**-** *G*: Hello! My name is Lucas, I am 24 years old, and I study engineering at the university. In addition, I am doing curricular internships in a research group at the university’s school.

**-** *I*: How often do you use the university’s intranet?

**-** *G*: The truth is that I only use the university’s intranet very sporadically. I don’t have much time or get the chance to check it out as often as I would like.

**-** *I*: What device do you use most often to access the intranet and why that device?

- *G*: I don’t have a device that I use more frequently, the truth is that both my mobile phone and my computer are my allies in this. When I am out or on the move, my mobile phone is the quickest and most practical option. But if I am at home or in the library, I prefer the computer because I can see everything on a larger screen.

- *I*: What corporate applications do you access through the intranet?

- *G*: Well, most of the time I use the intranet to access Moodle and Webmail. I also use the intranet occasionally to request to borrow books from the library and to know my situation of approved credits for the degree and the ones I have yet to pass.

- *I*: What information do you get from the intranet for your daily work?

TABLE VI
LIKERT QUESTIONS USED FOR EVALUATING INTERVIEWS FROM BOTH TEAMS

| HCI Professors question no. | UX Experts question no. | Likert statement |
|---|---|---|
| | | **Quality: Interview Design** |
| 1 | 1 | I think that interviews are well-defined and questions are adecuate for data extraction |
| 2 | 2 | I think that question ordering is adecuate for data extraction |
| 3 | 3 | I think that the amount of questions is adecuate for data extraction |
| 4 | 4 | I think that questions were relevant enough for creating meaningful Personas |
| | | **Quality: Answers Informativeness** |
| 5 | 5 | I found that interviews are complete to extract meaningful knowledge from users. |
| 6 | 6 | I think that I was capable to create meaningful Personas from users. |
| | 7 | I found that I discovered useful findings from users. |
| 7 | 8 | I consider that the extraction of information has been difficult with regard to my skills. |
| 8 | 9 | I think that I was capable to condense the information extracted from interviews and conduct an executive summary of users. |
| | 10 | In general, I think that my developed Personas are sound |
| | | **Quality: Answers Soundness** |
| 9 | 11 | I think that answers are simple (lack of elaboration) |
| 10 | 12 | I think that answers are short |
| 11 | 13 | I think that answers are easy to understand |
| 12 | 14 | I think that answers provide useless information |
| 13 | 15 | I think that answers are direct/straight to the point |
| 14 | 16 | I think that answers possess contradictory information. |
| 15 | 17 | In general, I think that answers are coherent with my previous self-experience |
| | | **Quality: Learnability** |
| 16 | | I found the interviews unnecessarily complex for practicing Personas |
| 17 | | In general, I think the interviews are adecuate for HCI/UX non-experts for learning how to create Personas. |
| | | **User Experience** |
| 18 | 18 | The interviewee has expressed his/her concerns and preserved his/her priorities. |
| 19 | 19 | The jargon has been aligned with the attitude and background of the interviewee. |
| | 20 | I think that the diversity of responses are consistent within the interview application domain |
| | 21 | I think that answers to questions are close to reality. |
| 20 | 22 | In general, I think the responses of the interviewees are believable. |
| 21 | 23 | I believe that the interviewees exhibit some emotion in their responses. |
| 22 | 24 | I think I could understand the frustrations of the interviewees refering his/her experiencience with the context. |
| 23 | 25 | I think I could understand the strengts of the interviewees refering his/her experiencience with the context. |
| 24 | 26 | I found that different interviews sounded very repetitive. |
| 25 | 27 | I found that during the same interview the answers to different questions were very repetitive. |
| | | **Ethics** |
| 26 | 28 | I found discriminatory speech in the interventions of the interviewees. |
| 27 | 29 | I found a harmful use of language by the interviewees (insults, slurs, taunts, provocations, etc...). |
| 28 | 30 | In general, I think that the interviewees to have been ethically accountable. |

Each number matches the result in tablesII and IV, respectively

- *G*: Mainly, I use the intranet to check my institutional email through Webmail and see if I have messages from my professors or from the administration. I also use Moodle to see updates on my courses, submit assignments, review grades, and sometimes participate in discussion forums. In addition, it also helps me keep track of my academic credits, to know what I have left to take and how my progress in the degree is going.

- *I*: What do you like most about the intranet?

- *G*: I like that the intranet allows me to have all the resources I need in one place. For example, being able to access Moodle and my university email from the same place is very useful. In addition, the possibility of seeing information about my credits and the progress of my degree is really convenient.

- *I*: What difficulties do you encounter in using the intranet?

- *G*: Although the intranet is quite useful, I find it complicated to find information in the university's intranet. Sometimes I feel that the information is scattered and it's not always easy to get to what I need quickly and easily. This can be a bit frustrating, especially when I have little time.

## ACKNOWLEDGMENT

The authors would like to thank the members of the UX team at the Universidad Politécnica de Madrid's Vice-Rectorate of Strategy and Digital Transformation—M. Velaochaga, I. Sánchez-Rivero, D. Acebedo and J. Martínez—for their contribution to this research.

## REFERENCES

[1] C. M. MacDonald, E. J. Rose, and C. Putnam, "How, why, and with whom do user experience (UX) practitioners communicate? Implications for HCI education," *Int. J. Hum.–Comput. Interaction*, vol. 38, no. 15, pp. 1422–1439, Dec. 2021, doi: 10.1080/10447318.2021.2002050.

[2] J. Aberg, "Challenges with teaching HCI early to computer students," in *Proc. 15th Annu. Conf. Innov. Technol. Comput. Sci. Educ.*, 2010, pp. 5–7.

[3] S. Caffiau, J. C. Campos, C. Martinie, L. Nigay, P. Palanque, and L. D. Spano, "Teaching HCI Engineering: Four case studies," in *Proc. IFIP Conf. Hum.-Comput. Interaction*, 2021, pp. 295–210.

[4] *Ergonomics of human-system interaction, Part 210: Human-centred design for interactive systems, ISO Standard 9241-210:2019*, 2019.

[5] M. Wang, M. Wang, X. Xu, L. Yang, D. Cai, and M. Yin, "Unleashing ChatGPT's power: A case study on optimizing information retrieval in flipped classrooms via prompt engineering," *IEEE Trans. Learn. Technol.*, vol. 17, pp. 629–641, 2023, doi: 10.1109/TLT.2023.3324714.

[6] J. Lu, R. Zheng, Z. Gong, and H. Xu, "Supporting teachers' Professional development with generative AI: The effects on higher order thinking and self-efficacy," *IEEE Trans. Learn. Technol.*, vol. 17, pp. 1279–1289, 2024, doi: 10.1109/TLT.2024.3369690.

[7] J. Yin, T.-T. Goh, B. Yang, and Y. Hu, "Using a chatbot to provide formative feedback: A longitudinal study of intrinsic motivation, cognitive load, and learning performance," *IEEE Trans. Learn. Technol.*, to be published, doi: 10.1109/TLT.2024.3364015.

[8] L. Nielsen, "Personas," in *The Encyclopedia of Human-Computer Interaction*, 2nd ed. Aarhus, Denmark: IxDF, 2014, ch. 30, pp. 2039–2092.

[9] P. Losana, J. W. Castro, X. Ferre, E. Villalba-Mora, and S. T. Acuña, "A systematic mapping study on integration proposals of the Personas technique in agile methodologies," *Sensors*, vol. 21, no. 18, Sep. 2021, Art. no. 6298, doi: 10.3390/s21186298.

[10] A. Cooper, *The Inmates are Running the Asylum*, INpolis, IN, USA, Sams Publication, 1998.

[11] T. Adlin and J. Pruitt, *Persona Lifecycle: Keeping People in Mind Throughout Product Design*, Burlington, MA, USA: Morgan Kaufmann, 2006.

[12] J. Grudin and J. Pruitt, "Personas: Practice and theory," in *Proc. Conf. Designing User Experiences*, 2003, pp. 1–15.

[13] D. Norman, "Ad-hoc personas & empathetic focus," 2004, revisited May 2023. [Online]. Available: https://jnd.org/ad-hoc-personas-empathetic-focus

[14] A. Cooper, R. Reimann, D. Cronin, and C. Noessel, *About Face: The Essentials of Interaction Design*. Hoboken, NJ, USA: Wiley, 2014.

[15] F. Anvari, D. Richards, M. Hitchens, M. A. Babar, H. M. T. Tran, and P. Busch, "An empirical investigation of the influence of persona with personality traits on conceptual design," *J. Syst. Softw.*, vol. 134, pp. 324–339, Dec. 2017, doi: 10.1016/j.jss.2017.09.020.

[16] A. Holzinger, M. Kargl, B. Kipperer, P. Regitnig, M. Plass, and H. Müller, "Personas for Artificial Intelligence (AI) an open source toolbox," *IEEE Access*, vol. 10, pp. 23732–23747, 2022, doi: 10.1109/ACCESS.2022.3154776.

[17] C. Bradley, L. Oliveira, S. Birrell, and R. Cain, "A new perspective on personas and customer journey maps: Proposing systemic UX," *Int. J. Hum.-Comput. Stud.*, vol. 148, pp. 1–17, Apr. 2021, doi: 10.1016/j.ijhcs.2021.102583.

[18] R. Sinha, "Persona development for information-rich domains," in *Proc. Extended Abstr. Hum. Factors Comput. Syst.*, 2003, pp. 830–831, doi: 10.1145/765891.766017.

[19] J. Salminen, K. Guan, S.-G. Jung, and B. J. Jansen, "A survey of 15 years of data-driven persona development," *Int. J. Hum.–Comput. Interaction*, vol. 37, no. 18, pp. 1685–1708, Apr. 2021, doi: 10.1080/10447318.2021.1908670.

[20] D. Korsgaard, T. Bjørner, P. Sorensen, and P. Burelli, "Creating user stereotypes for persona development from qualitative data through semi-automatic subspace clustering," *User Model User-Adapt Interact*, vol. 30, pp. 81–125, Jan. 2020, doi: 10.1007/s11257-019-09252-5.

[21] OpenAI, "GPT-4 technical report," Mar. 2023. *ArXiv*, doi: abs/2303.08774.

[22] T. Terzidou, T. Tsiatsos, C. Miliou, and A. Sourvinou, "Agent supported serious game environment," *IEEE Trans. Learn. Technol..*, vol. 9, no. 3, pp. 217–230, Jul. 2016, doi: 10.1109/TLT.2016.2521649.

[23] P. Smutny and P. Schreiberova, "Chatbots for learning: A review of educational chatbots for the Facebook Messenger," *Comput. Educ.*, vol. 151, pp. 1–11, Jul. 2020, doi: 10.1016/j.compedu.2020.103862.

[24] A. Kerly, P. Hall, and S. Bull, "Bringing chatbots into education: Towards natural language negotiation of open learner models," *KBS*, vol. 20, pp. 177–185, Mar. 2007, doi: 10.1016/j.knosys.2006.11.014.

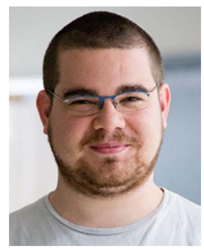
[25] S. Portigal, *Interviewing Users: How to Uncover Compelling Insights*. New York, NY, USA: Rosenfeld Media, 2013.

[26] W. Chauncey, *Interview Techniques for Ux Practitioners: A User-Centered Design Method*. Burlington, MA, USA: Morgan Kaufmann, 2014.

[27] X. Han, M. Zhou, M. J. Turner, and T. Yeh, "Designing effective interview chatbots: Automatic chatbot profiling and design suggestion generation for chatbot debugging," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2021, pp. 1–15.


**Jose Barambones** received the Ph.D. (international mention and cum laude) degree in computer science from the Universidad Politécnica de Madrid (UPM), Madrid, Spain, in 2023.

He is a Senior Researcher with the Madrid HCI Laboratory and the Ageing Laboratory based at UPM. He is also an Assistant Professor of HCI, agent-based software development, and interactive systems courses at UPM in Madrid. He was a Guest Research Intern at the National Institute of Informatics in Tokyo, Japan, in 2016 and 2019, and was a Senior Computer Engineer in Distributed Systems at the European Organisation for Nuclear Research (CERN) in 2016–2017. His research focuses on distributed artificial intelligence and user experience in games on intelligent multiagent systems and human–agent interaction.



**Cristian Moral** received the M.Sc. and Ph.D. degrees in computer science from the Universidad Politécnica de Madrid, Madrid, Spain, in 2011 and 2016, respectively.

He is an Associate Professor with the UPM's Escuela Técnica Superior de Ingenieros Informáticos (ETSIINF) where he has been teaching courses on human–computer interaction and virtual reality since 2014. He has been a Researcher with the Madrid HCI Laboratory based at UPM since 2011 and at the UPM's Center for Biomedical Technology since 2017. In 2014, he was a visiting Ph.D. scholar at the Center for Human–Computer Interaction, Virginia Tech. His main teaching and research interests are HCI in ICT for healthy aging, information retrieval and visualization, virtual and augmented reality, and eHealth. His current research interests include the use of sensor-based platforms to monitor and improve the quality of life of the aging population. Since 2022, he also been Vice Dean for Quality and Accreditations at ETSIINF-UPM.

**Angélica de Antonio** received the M.Sc. degree in computer science and the Ph.D. degree in artificial intelligence from the Universidad Politécnica de Madrid (UPM), Madrid, Spain, in 1990 and 1994, respectively, and the M.Sc. degree in psychology from the Universidad Complutense de Madrid, Madrid, in 2016.

She was a resident affiliate at the Software Engineering Institute, Carnegie Mellon University, USA, in 1995. Since 1997, she has been an Associate Professor with the UPM. She is the Director of the Research Group on Human-Computer Interaction and Advanced Interactive Systems, Madrid HCI Laboratory, UPM and Coordinator of the European Master in Software Engineering. Since 2020, she is also been the UPM's Associate Vice-Rector for Quality. Her main research interests include user modeling and adaptive systems, 3-D virtual environments and intelligent virtual agents, and technology acceptance and adoption.

**Ricardo Imbert** received the M.Sc. degree in computer science, the M.Sc. degree in software engineering, and the Ph.D. degree in computer science from the Universidad Politécnica de Madrid (UPM), Madrid, Spain, in 1997, 2000, and 2005, respectively.

He has been a Researcher with the Madrid HCI Laboratory, UPM, since 1996 and with the Center for Biomedical Technology, UPM, since 2021. Since 2000, he has been an Associate Professor with the UPM's Escuela Técnica Superior de Ingenieros Informáticos (ETSIINF). He coordinated the European Master in software engineering at UPM, from 2008 to 2017. He was a Vice Dean for Student Affairs at ETSIINF-UPM from 2012 to 2016, and he has been the UPM's Associate Vice-Rector for Information Systems since 2016. His research interests include address human–computer interaction, interactive systems, user experience, cognitive architectures, and agent-based software engineering.

**Loïc Martínez- Normand** received the M.Sc. degree in computer science and the Ph.D. degree in artificial intelligence from the Universidad Politécnica de Madrid (UPM), Madrid, Spain, in 2003 and 1993, respectively.

He has been an Associate Professor with UPM, since 1996, where he teaches courses on HCI and ICT accessibility. He has been Vice Dean for Academic Affairs of the Escuela Técnica Superior de Ingenieros Informáticos (ETSIINF) since 2016. Since 1995, he has been a member of the UPM's Official Research Group on Computing and Information Technologies, where he leads and develops projects on ICT accessibility for organizations, such as Spanish National Association for the Blind (ONCE), Spanish National Confederation for the Deaf (CNSE), and Institute for Migrations and Social Affairs (IMSERSO). He actively participates in national and international standardization activities in the field of ICT accessibility.

**Elena Villalba-Mora** received the M.Sc. degree in telecommunications engineering (cum laude) and the Ph.D. (Doctor Europeus) degree in biomedical engineering from the Universidad Politécnica de Madrid (UPM), Madrid, Spain, in 2004 and 2008, respectively.

She is currently an Associate Professor in computer science, human–computer interaction, and eHealth with the Universidad Politécnica de Madrid (UPM), Madrid, Spain, where she is also a Principal Investigator with the Ageing Laboratory, Centre for Biomedical Technology. She has participated in more than 20 international and national research projects, mostly as a Principal Investigator, and research contracts. She has authored more than 50 papers in high-impact journals, international conference proceedings, and scientific reports. Her research interests include modeling, design, and validation of health information technologies for long-term care of older adults to keep them independent and autonomous.