Shin, Joongi; Hedderich, Michael A.; Rey, Bartlomiej; Lucero, Andrés; Oulasvirta, Antti

Understanding Human–AI Workflows for Generating Personas

Published in:
Proceedings of the 2024 ACM Designing Interactive Systems Conference, DIS 2024

DOI:
10.1145/3643834.3660729

Published: 01/07/2024

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Shin, J., Hedderich, M. A., Rey, B., Lucero, A., & Oulasvirta, A. (2024). Understanding Human–AI Workflows for Generating Personas. In A. Vallgarda, L. Jonsson, J. Fritsch, S. F. Alaoui, & C. A. Le Dantec (Eds.), *Proceedings of the 2024 ACM Designing Interactive Systems Conference, DIS 2024* (pp. 757-781). (Proceedings of the 2024 ACM Designing Interactive Systems Conference, DIS 2024). ACM. https://doi.org/10.1145/3643834.3660729

Understanding Human–AI Workflows for Generating Personas

Joongi Shin
Aalto University
Espoo, Finland
joongi.shin@aalto.fi

Michael A. Hedderich
LMU Munich & MCML
Munich, Germany
Michael.Hedderich@lmu.de

Bartłomiej Jakub Rey
Aalto University
Espoo, Finland
bartlomiej.rey@aalto.fi

Andrés Lucero
Aalto University
Espoo, Finland
lucero@acm.org

Antti Oulasvirta
Aalto University
Espoo, Finland
antti.oulasvirta@aalto.fi

ABSTRACT

One barrier to deeper adoption of user-research methods is the amount of labor required to create high-quality representations of collected data. Trained user researchers need to analyze datasets and produce informative summaries pertaining to the original data. While Large Language Models (LLMs) could assist in generating summaries, they are known to hallucinate and produce biased responses. In this paper, we study human–AI workflows that differently delegate subtasks in user research between human experts and LLMs. Studying persona generation as our case, we found that LLMs are not good at capturing key characteristics of user data on their own. Better results are achieved when we leverage human skill in grouping user data by their key characteristics and exploit LLMs for summarizing pre-grouped data into personas. Personas generated via this collaborative approach can be more representative and empathy-evoking than ones generated by human experts or LLMs alone. We also found that LLMs could mimic generated personas and enable interaction with personas, thereby helping user researchers empathize with them. We conclude that LLMs, by facilitating the analysis of user data, may promote widespread application of qualitative methods in user research.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; *Natural language interfaces*; *User studies.*

KEYWORDS

User research, persona generation, LLM

ACM Reference Format:
Joongi Shin, Michael A. Hedderich, Bartłomiej Jakub Rey, Andrés Lucero, and Antti Oulasvirta. 2024. Understanding Human–AI Workflows for Generating Personas. In *Designing Interactive Systems Conference (DIS '24), July 01–05, 2024, IT University of Copenhagen, Denmark.* ACM, New York, NY, USA, 25 pages. https://doi.org/10.1145/3643834.3660729

DIS '24, July 01–05, 2024, IT University of Copenhagen, Denmark

ACM ISBN 979-8-4007-0583-0/24/07
https://doi.org/10.1145/3643834.3660729

1 INTRODUCTION

User research is at the core of user-centered design [24, 38, 72]. The rich insight it produces about people can help in setting priorities, solving problems, and exposing new opportunities in design [4]. One of the most popular methods for summarizing user research is *personas* [11]. Personas are representations of archetypal or "median" user groups created from user data (e.g., interview, observation, survey, or log data). Apart from simply summarizing user data, personas should *personify user data* to encourage perspective-taking and evoke empathy toward user groups [11, 49, 71].

However, generating representative and empathy-evoking personas is labor-intensive. Trained user researchers must collect data, perform analysis, and narrate archetypal user behaviors that are important for their projects. Because of the high cost, corners are often cut. For an illustrative case, Pruitt and Grudin [62] reported issues encountered at Microsoft in the use of personas: The characters portrayed in personas were not believable, rather designed by committee, and "lacked a connection to user data". Also, the personas were poorly communicated, with long lists of characteristics that showed little effort to distill the main points. Among the undesired consequences are that the results can misdirect the design process and prove unconvincing to stakeholders [18]. We argue that these all are symptoms of an underlying cause: The high cost of producing high-quality representations.

To tackle this, we study human–AI workflows where Large Language Models (LLMs) are used for different subtasks of working with user data (Figure 1). LLMs could decrease human experts' efforts in summarizing collected data and identifying overlooked patterns. LLMs, such as GPT-4 [56], can extract keywords [41], cluster text by their semantic similarity [53], and summarize extensive text input [84, 86]. Yet LLMs also have technical limitations that could taint the quality of analysis. The models exhibit biases [15, 16, 31, 36, 40] and often synthesize non-factual information on account of training datasets' implicit biases and lack of domain knowledge [5, 50, 52]. Emerging evidence suggests that they fare poorly at reproducing the statistical distribution of input data [64]. Moreover, even if LLMs could analyze user data correctly, fully automating the process could impede human experts' understanding of the underlying data and its connection to personas. Therefore, we proceed on the assumption that human experts must have a role, but the question is which role. By studying human–AI workflows, we probed how LLMs' strengths can be best exploited to benefit human experts.

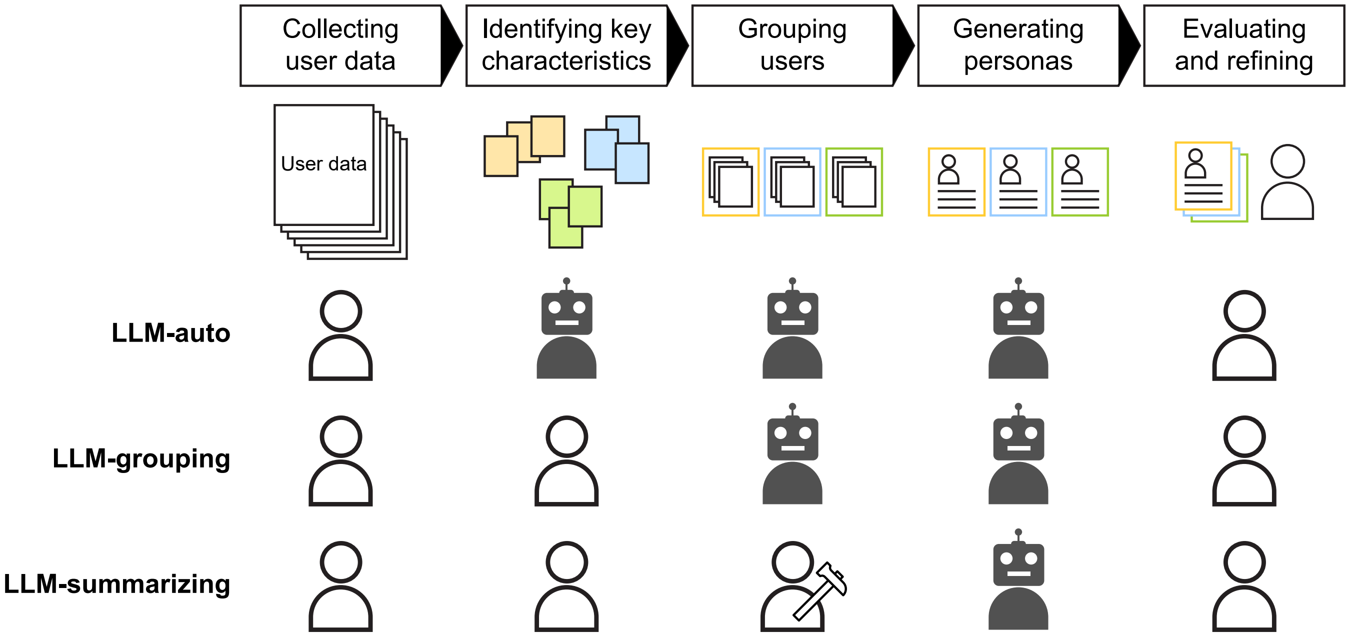

Figure 1: This paper studies human–AI workflows for supporting user research with LLMs. We study different subtask delegations between human experts and LLMs in persona generation. We show that the best outcomes are achieved with LLM-summarizing that leverages 1) human experts' ability to identify important user groups with clustering methods and 2) LLMs' ability to summarize text.

Our main research question is "which persona-generation subtasks should be delegated to user researchers vs. LLMs to produce representative and empathy-evoking personas?" Building on Cooper et al.'s process [11], we identified three subtasks that could be shared between user researchers and LLMs. Accordingly, we designed three workflows by varying the reliance on LLMs as shown in Figure 3. One workflow relies comprehensively on LLMs, from identifying users' characteristics to generating personas (LLM-auto). The second relies on human experts to identify characteristics (attitude, motivation, etc.) from user data, after which LLMs group the data by those characteristics and generate personas (LLM-grouping). The last workflow relies the most on human experts. Here, the experts identify users' characteristics and create a certain number of user groups supported by clustering methods [53]. LLMs' only task is to summarize the grouped data into personas (LLM-summarizing). To implement these workflows, we built on prompt engineering literature [12, 75, 79] and designed prompts that provide well-contextualized instructions to LLMs.

With our workflows, we also explored the exciting possibility of interacting with personas via LLMs (see Figure 2). Shanahan et al. [74] contend that LLM-based dialogue agents excel at answering users' questions by acting as helpful assistants. Likewise, we expect LLMs to act as generated personas and enable "role-play" [8, 60]. Conversing with personas would convey understanding more effectively than simply reading descriptions of them. We expect this interactivity to benefit user researchers working with their data, offestting some negative effects of delegating subtasks to LLM.

We evaluated the workflows' outputs (personas) in a series of four studies focused on statistical representativeness (Study 1), perceived quality (Study 2), preferred quality (Study 3), and interaction with the personas (Study 4). More precisely, in Study 1, we compared the personas by statistically measuring their similarity to the input user data. Study 2 involved assessing personas' various qualities with designers by taking designer-generated personas as a baseline. In Study 3, we refined the prompt of the best-performing workflow to narrate personas in accordance with designers' preferences observed in Study 2. Lastly, in Study 4, we conducted a formative study to explore how LLMs could converse with designers and assist their use of personas. Throughout the research, we used GPT-4 [1] and user data in the form of survey responses.

Our results demonstrate what both LLMs and human experts bring to the generation of high-quality personas. Though LLMs can generate personas with reasonable qualities 'out of the box', we found that the most statistically representative personas emerge when human experts have grouped user data for LLMs (LLM-summarizing). We found that this workflow can be augmented further by instructing human-preferred narration styles to LLMs (LLM-summarizing++). That resulted in the most empathy-evoking personas. To sum up, our work contributes to understanding optimal ways to divide the subtasks of data-intensive effort in user research between human experts and LLMs. In particular, we found that LLMs' utilities can be enhanced significantly by involving human experts in the loop at the correct persona-generation subtasks. With our findings, we make three contributions:

[1] Specifically, gpt-4-0314: https://platform.openai.com/docs/models/overview

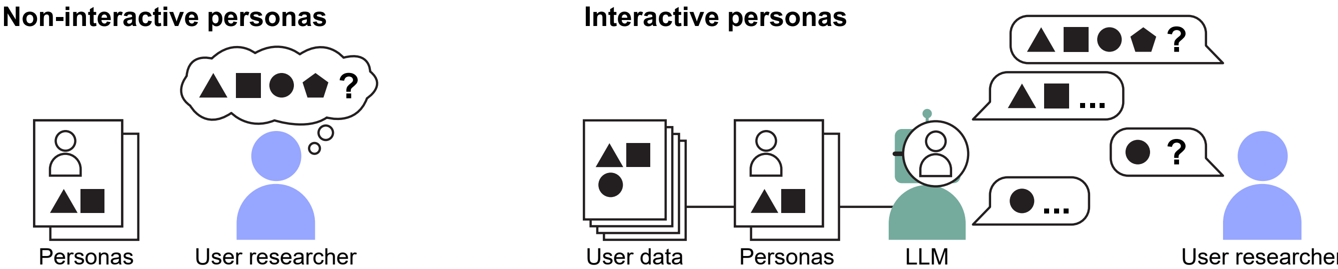

Figure 2: A depiction of the concepts of non-interactive personas (left) and interactive personas (right). The conventional approach to using personas has been to read them for understanding and empathizing with users. We find LLMs promising for enabling simulated interaction with personas, which could prove more effective for understanding users.

(1) We designed human–AI workflows wherein human experts and LLMs (GPT-4) can together create personas from qualitative user data. We offer guidelines for workflow implementations that support using these in practice.
(2) We found that the best workflow can aid user researchers without compromising outcome quality. The personas generated under the collaborative approach are at least as representative and empathy-evoking as those produced by human experts alone.
(3) We present evidence that human experts' preferences are crucial in augmenting LLMs' utility, which instruct LLMs to narrate personas to be more empathy-evoking.

2 RELATED WORK

Often, the conventional generation of personas is not cost-efficient [61]. User researchers typically depend on qualitative user data collected through interviews or direct observations. Since it demands significant time and effort, their analysis of the data often limits feasible dataset sizes or precludes more rigorous analyses [19, 77]. This qualitative approach has been criticized for resulting in personas that lack connection to the original data and larger population [51, 66]. Relying solely on human experts' interpretations has been questioned for objectivity reasons too [47]. With this paper, we aim to address such limitations by studying human–AI workflows that extend from analyzing user data to generating personas.

2.1 Data-driven persona generation

To address the shortcomings of qualitative persona generation, scholars have considered quantitative data-driven persona generation (DDPG) [51, 66, 70, 87], which Salminen et al. [67] define as "the use of algorithmic methods to create accurate, representative, and refreshable personas from numerical data". The goal behind DDPG is to enable efficient and accurate analysis of quantifiable user data. Zhang et al. [85] offer an example from studying two years of online customer behaviors. After collecting customers' click sequences in a user interface, they converted them into vectors and grouped them by means of a clustering algorithm. This resulted in six groups of users sharing similar interaction behaviors, which served as a foundation for creating distinctive personas.

Despite the advantages of DDPG, purely relying on quantitative data cannot furnish all of the contents for personas [68]. In the end, it is the qualitative data that holds the key to a deep understanding of users that can construct the attitude and motivations of personas [11, 61]. Recent advances in NLP technology attest to the potential of pairing qualitative data with DDPG [37]. For instance, Tan et al. [77] thereby generated personas based on more than 200,000 online comments about products. Having extracted keywords from each comment by means of NLP classifiers, they turned the keywords into word vectors and grouped them via a clustering algorithm. Still, the grouped data's interpretation and summarization as personas was done manually by user researchers alone: the connection between original user data and personas could still be tainted.

LLMs open a new horizon for DDPG. In collaborative activities, they could support the analysis of qualitative user data beyond keywords. Their competence in summarizing text could assist user researchers in creating contents for personas, with the resulting personas being better anchored in user data.

2.2 Natural language processing and large language models

NLP offers tools to analyze large quantities of text automatically. Functionality that extracts information has assisted sociologists [45], and it can likewise serve user research – by, for example, extracting key aspects of user stories [63] or coding qualitative interviews [46]. Besides extracting individual pieces of information, NLP methods afford summarizing entire texts in an abstractive way [2]. These techniques can speed up the process of otherwise manual analysis for the researcher or designer, thus opening the door to analyzing text in quantities that human reading alone cannot handle.

Recently, LLMs have revolutionized the field of NLP and made new applications possible [6]. Training on vast quantities of text and incorporating human feedback let state-of-the-art LLMs reach new heights of performance in many NLP tasks [55], among them text summarization [28], supporting people's creative activities [22, 35] and even following natural language instructions [57]. Many of LLMs' capabilities hold potential to inform persona generation. Creating personas based on interviews requires the designer to extract, group, summarize, and consolidate information from a set of texts. While LLMs could aid with one or more of these activities, studies have highlighted LLMs' tendency to hallucinate and exhibit intrinsic biases [5, 15, 16, 36, 40, 50, 52], both of which could influence representativeness of personas if LLMs on their own process

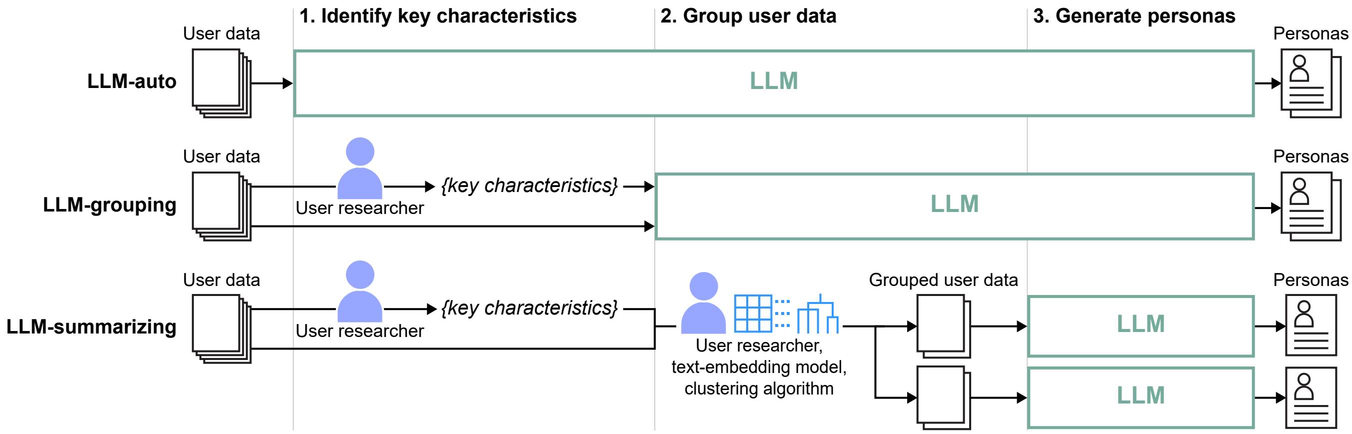

Figure 3: A conceptual diagram of our persona-generation workflows.

user data into personas. Hence, how best we can utilize LLMs for generating personas has remained unclear.

2.3 Human–AI workflows

The increasing competence of AI calls for rethinking which tasks should be done by machines and which by humans. This question has increased interest in *human–AI workflows*, especially how to assign subtasks between human experts and AI [20, 25, 26, 29, 43, 54]. In a well-known case, Kanarik et al. [34] designed cost-efficient workflows where experienced human engineers and AI build on each other's strengths in designing semiconductors. In a virtual lab, the authors tested how to find the optimal settings with fewer trials. They found that AI's advantages could not compensate for human experts' experiences/intuitions with its lower costs and fewer trials. Moreover, AI failed badly when working alone: it produced less optimal settings than the human experts in more than 95% of trials, costing more to find the optimal settings. Accordingly, the authors developed 'human first–compute last' workflows, wherein human experts set the initial parameters while AI performed the function of fine-tuning the parameters. This workflow dramatically increased the chances of finding the optimal settings (e.g., from 11% to 42% with the best-performing AI). Likewise, Ahn et al. [1] designed a workflow for humans and AI building on each other's diagnosis in visual classification tasks. Humans were better at detecting an association between visuals and concepts, while AI demonstrated better recall of existing ones. The authors' collaborative workflow could result in more accurate classification by correctly identifying misclassified instances. We envision user research benefiting similarly from human–AI workflow. Pursuing advantages in the context of persona generation, we investigated which subtasks should be performed by user researchers vs. LLMs such that more representative and empathy-evoking personas emerge.

3 HUMAN—AI WORKFLOWS FOR PERSONA GENERATION

Looking into the process of generating personas defined by Cooper et al. [11], we pinpointed three essential subtasks that could be delegated to LLMs. Accordingly, we designed three persona-generation workflows by varying the reliance between user researchers and LLMs (Figure 3). Considering LLMs' strength in extracting keywords [41], estimating semantic similarity [53], and summarizing input text [84, 86], we expect them to be able to (i) identify users' key characteristics, (ii) group user data, and (iii) generate personas from input user data. Below, we describe these workflows' fundamental distinctions. The procedure and technical aspects of each workflow are described in Appendix A.

(1) **LLM-auto:** Our first workflow relies on LLMs to perform all three tasks. In a single prompt, user researchers input all user data (e.g., a collection of user responses to interview or survey questions) and instruct LLMs to generate a minimum number of personas. This workflow does not require user researchers to state their interests in their target users; hence, LLMs could identify archetypal characteristics in user data unconstrained by human judgments.
(2) **LLM-grouping:** The second workflow depends on user researchers to identify key characteristics. Researchers instruct LLMs to group user data by the researcher-chosen characteristics and to generate one persona from each group. Often, user researchers rely on their "gut feeling" when encountering user data with vague distinctions [11]. LLMs could, instead, supply data-driven grouping by calculating the semantic similarity between user data and the description of characteristics.
(3) **LLM-summarizing:** Our last workflow relies mostly on user researchers, with LLMs active later on to summarize user data into personas. Researchers group user data by the key characteristics that they identify, supported by text-embedding models and clustering algorithms. Then, by inputting one group of user data per prompt, they prompt LLMs to generate one persona at a time. This workflow could generate personas aligned closely with the foci of user researchers' interest in their target users.

We presume that, generally, greater reliance on user researchers might well yield personas that match their interests more closely but require more effort. This effortful nature of working with user data might cause unassisted humans to create less representative

LLM-AUTO	LLM-GROUPING	LLM-SUMMARIZING
I want to have children someday because I believe that raising a family will bring additional joy and fulfillment to my life. ...	I want to have children someday, but not now because I am currently focused on growing in my career and enjoying the present moment. ...	I want to have children someday because it seems like an incredible experience to bring new life into this world and share in the growth process. ...
I want to have children because I am attracted to the new level of joy and fulfillment that raising a family might bring to my life. ...	I do not want to have children ever because I value my career and the pursuit of new discoveries. ...	I want to have children because I believe it will bring a new level of joy, fulfillment, and purpose to my life. ...
I do not want to have children ever. My career and the pursuit of new discoveries are what gives me the most fulfillment in life. ...	I do not want to have children ever because it does not align with my personal goals and priorities. ...	I do not want to have children ever, because I prioritize my career, personal interests, and rela tionships with my loved ones. ...

Table 1: Example sets of personas and partial narrations generated through our workflows. While they all generated three groups of personas from the same survey data (people's opinions on having children), they display nuanced differences in their creation of user groups. For instance, while LLM-AUTO and LLM-SUMMARIZING grouped users by their distinctive opinions ('I want to have children someday', 'I want to have children', and 'I do not want to have children'), LLM-GROUPING grouped them by the motivations behind those opinions, pinpointing three distinct reasons for not wanting to have children (blue highlights). Likewise, LLLM-AUTO and LLM-SUMMARIZING identified different reasons behind the same user groups (yellow highlights).

personas [48, 76]. In contrast, greater reliance on LLMs could reduce researchers' effort and potentially yield more statistically accurate analysis of user data [23, 78]. That said, the personas might be lower-quality, failing to highlight users' perspectives or key behaviors – LLMs might not capture such nuance. Table 1 illustrates differences revealed among the workflows.

To implement the workflows, we designed a prompt to delegate the subtasks to LLMs. Prior work attests that prompts' design significantly influences LLMs' output quality [12, 14, 75]. In addition, we observed that the most basic prompts fail to generate high-quality personas (see our exploration in Appendix B). Therefore, we adopted guidelines from prompt-engineering literature [12, 14, 75, 79] and iteratively experimented with GPT-4. This led to designing a prompt that is amendable to sharing across the workflows with only slight adjustments, making LLMs generate personas with reasonable qualities despite the workflows' differences. Our final prompt and resulting personas are shown in Appendix B and Appendix F, respectively.

4 STUDY 1: EVALUATION ON STATISTICAL REPRESENTATIVENESS

We assessed the representativeness of the personas each workflow creates (Figure 4). Measuring *How well a persona represents a user group* requires assessing statistics [64] and perceived quality [70] both. We can statistically measure the similarity between a persona and user data by computing their closeness for specific attributes (e.g., age and gender distribution) and text semantics (i.e., the similarity in meaning between two sentences or paragraphs). Whereas greater similarity can confirm that a persona captures more information from user data, it cannot reveal whether that persona highlights key characteristics or merely aggregates all user data (after all, the semantic-similarity score is highest when two sentences are identical). Therefore, an assessment by human experts is also needed to confirm that the personas indeed represent users' essential characteristics. We present our evaluation of personas' statistical representativeness below (Study 1). The next section reports on perceived representativeness with designers (Study 2).

While our workflows are, in principle, independent of any specific LLM, we used GPT-4 for all experiments in light of its superior performance at the time of writing. Multiple skill-based evaluations, such as comprehension and completeness metrics [82], attest to its effectiveness in our tasks.

4.1 Preparing user data for persona generation

We prepared 20 survey-response sets as input user data. To closely resemble the user data that designers collect, we created survey questions by adapting interview and questionnaire items found in persona-generation literature [30, 69, 77]. We included three types of questions, eliciting information on users' demographics, backgrounds, and design-related comments, all of which can contribute to the contents of personas [11].

Instead of collecting real user responses, we created synthetic responses, using GPT-3.5-turbo. This approach confers an advantage in establishing ground-truth user groups with specific demographic and attribute distributions, thereby enabling rigorous analysis of the user groups represented by personas. Also, recent studies attest to the competence of GPT models in reproducing subtleties of real user responses and creating human-like interview responses [16, 27, 73]. We chose a prominent design subject around which distinct user groups might cohere: designing civic services to address declining birth rates. For analysis, we formed four ground-truth user groups: "I want to have children", "I want to have children, but not now", "I do not want to have children", and "I do not want to have children ever". To simulate the responses of 20 users, we manually entered the age, gender, marital status, and child-plan details by following our ground-truth distribution. Then, we prompted GPT-3.5-turbo to generate responses to the rest of the survey questions on the basis of the pre-entered information. Appendix D presents all the questions and an example response.

4.2 Generating personas for analysis

We followed each persona-generation workflow 100 times and created a set of personas for each workflow (Figure 4, left). Whereas LLM-AUTO is designed to rely on LLMs to identify key characteristics in user data, LLM-GROUPING and LLM-SUMMARIZING require

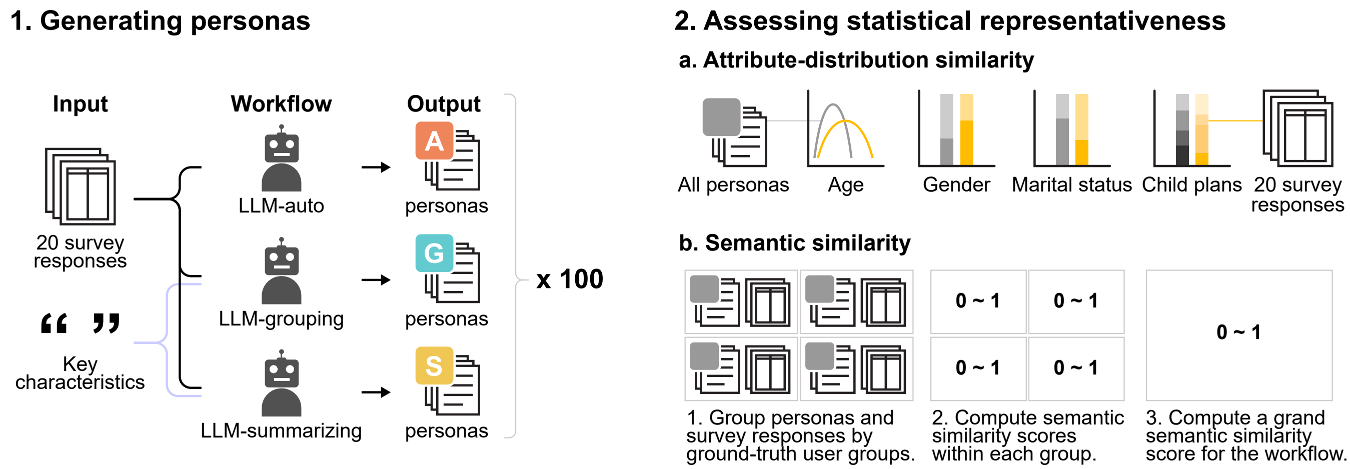

Figure 4: The structure of Study 1. We compared statistical similarity between input user data (20 survey responses) and personas generated by each workflow. From 100 runs with each workflow, we created 100 personas at minimum for it (left); then, we compared the attribute distribution (right, pane a) and semantic similarity between the personas and user data (right, pane b).

human experts to set the key characteristics for grouping user data. For this, we specified the characteristics as respondents' reasons for wanting or not wanting children to compare the personas with the ground-truth user groups. Accordingly, we set the persona contents to be name, age, gender, occupation, marital status, background, personality, plans for children, and motivation for using civic services. Despite working with the same settings, both workflows generate different personas in each trial as LLMs inherently introduce variances in their responses.

LLM-AUTO and LLM-GROUPING generated 2–3 personas at a time. Often, they yielded incomplete personas because of truncation of a few sentences on account of GPT-4's token limitation [2]. In such cases, we removed both complete and incomplete personas generated in the affected run, to avoid overrepresenting one of the user groups. Since LLM-SUMMARIZING outputs personas one at a time, it did not suffer from such token-related limitations. In the final material, LLM-AUTO created 268 personas, and LLM-GROUPING created 246. LLM-SUMMARIZING created three user groups; hence, 300 personas were created. We present example personas in Appendix F.

4.3 Assessing statistical similarity

We conducted two comparisons between persona sets and survey responses: attribute distribution and semantic similarity. We compared the age, gender, marital-status, and child-plan distributions, which can reveal how well LLMs reproduce statistical distribution in user data [64]. We used four common semantic-similarity metrics in Natural Language Generation (NLG) tasks [65]. Each measures separate aspects of semantic similarity between the source text (here, survey responses) and synthetic text (the personas):

- *ROUGE-L* measures the ratio for the longest word sequence in common between two sentences [42].
- *BERTScore* captures the distance between text embeddings of two sentences, generated with distilbert-based-uncased [83].
- *GPT-based-similarity* represents the distance between text embeddings of two paragraphs, generated with text-embedding-ada-002.
- *G-eval* evaluates the information validity of synthetic text vs. source text by asking GPT-4 to review both and assign a score between 0 (invalid) and 1 (valid) [21].

We conducted pairwise comparisons between the personas and survey responses corresponding to each ground-truth user group (e.g., we compared personas created to represent "I want to have children" to the survey responses with the same opinions). Appendix H provides further descriptions of the analysis.

4.4 Results

4.4.1 Attribute distributions. We used IBM SPSS Statistics for all analyses ($p < 0.05$ was considered significant). The age distribution of the survey responses (ground truth) and the personas are shown in Figure 5. We performed Levene's test with Bonferroni correction to compare the variance in age distribution between the survey response and each set of personas. The result showed that the age variances were statistically different across all comparisons (all $p < 0.05$). Consequently, we performed the Mann–Whitney U test (a non-parametric equivalent to independent-samples t-test) to compare the central tendency of the age distributions statistically. Across all comparisons, there were no statistically significant differences(all $p > 0.05$). This indicates that while all workflows did not reproduce the age variance, they captured the median age in the input data.

The distribution of gender, marital status, and child plans from the survey responses and the personas are shown in Figure 6. We found that all workflows captured every attribute present in the survey responses. LLM-SUMMARIZING is noteworthy for often creating personas with a new gender (e.g., "Non-binary" and "Male/Female") and new marital status (e.g., "In a committed relationship" and

[2]https://platform.openai.com/docs/models/gpt-4

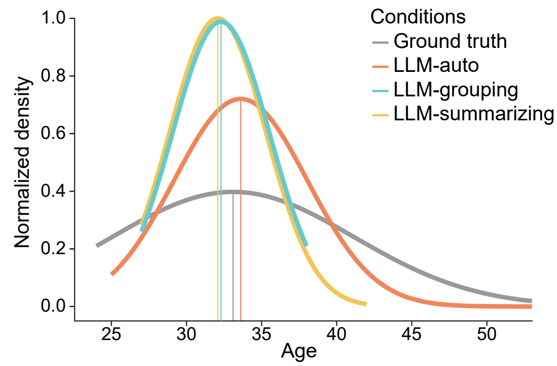

Figure 5: Distributions for gender, marital status, and child plans in Study 1's user data (ground truth) and generated personas. We found that all our workflows captured the median age in the user data, while they could not reproduce the age variance.

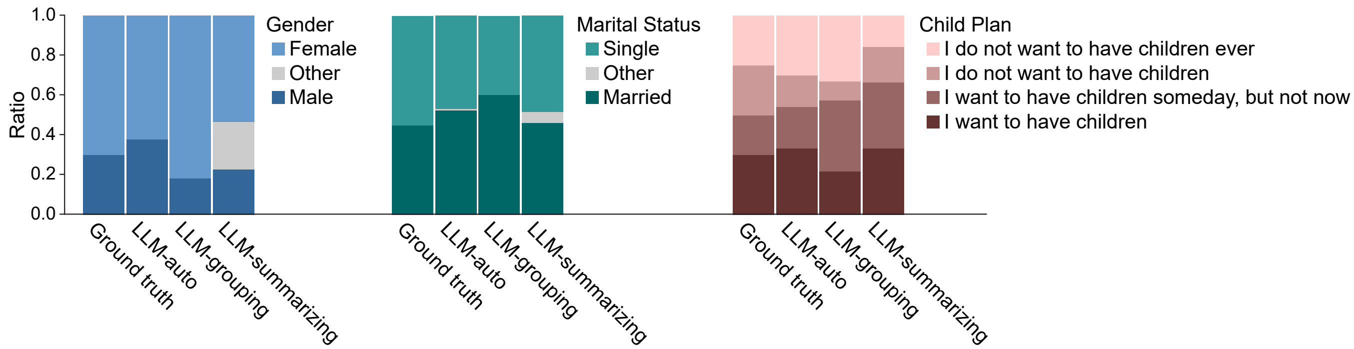

Figure 6: Distributions for gender, marital status, and child plans in Study 1's user data (ground truth) and generated personas. All workflows captured the attributes of the user data in these respects.

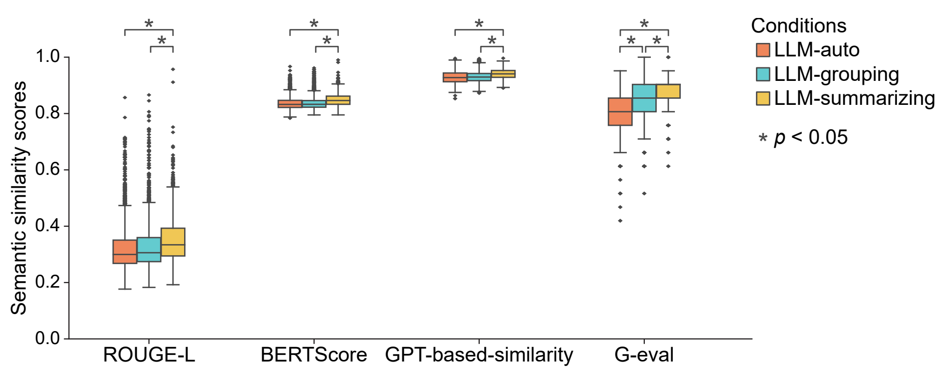

Figure 7: The semantic similarity found between the user data and generated personas in Study 1. The personas generated by LLM-SUMMARIZING proved most semantically similar to the input user data.

"Single/Married"). A few cases of this sort were observed with LLM-AUTO also. We statistically compared the attribute distributions between the survey responses and each set of personas using Fisher's exact test. Only LLM-AUTO showed no statistically significant differences from the survey responses (gender: $p = 0.16$, marital status: $p = 0.28$, and child plan: $p = 0.52$). From the results, we conclude that all three workflows can capture the attributes present in the input data, while LLM-AUTO can reproduce the attribute distribution also.

4.4.2 Semantic-similarity scores. The scores for semantic similarity between the survey responses and the personas are shown in Figure 7. We performed the Kruskal–Wallis test (a non-parametric equivalent one-way ANOVA) for statistically comparing each similarity score between the workflows. The result showed statistically significant differences in all similarity scores, with $p < 0.05$. We conducted the Mann–Whitney U test as a *post-hoc* analysis with Bonferroni correction to identify where the statistical differences lie.

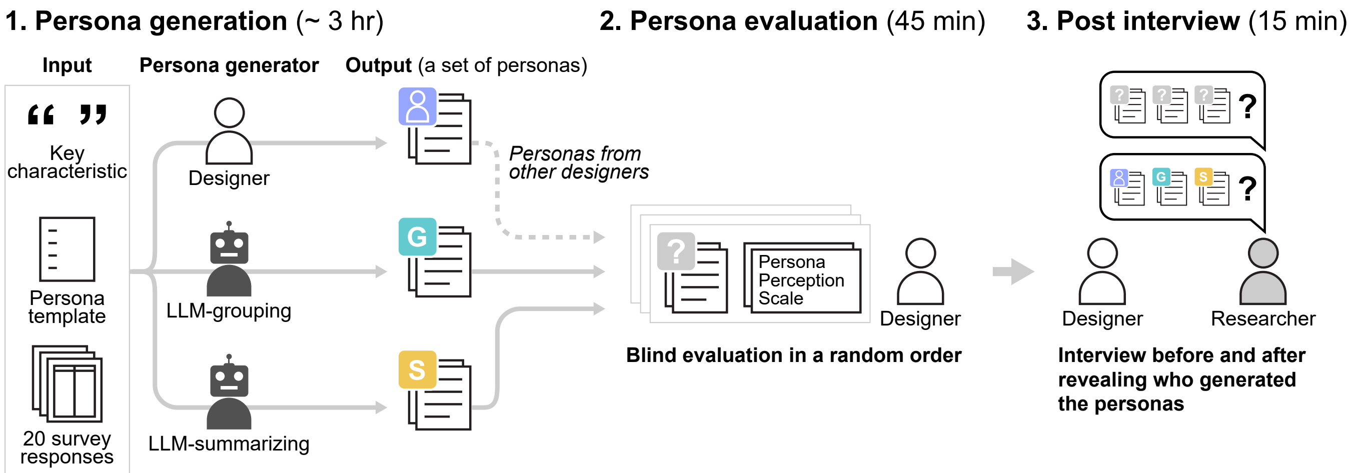

Figure 8: The procedure of evaluating personas with designers. We asked designers to review survey responses and generate personas (left). Then, we had them review these, one set of personas at a time, and assess each persona on the Persona Perception Scale (middle). Lastly, we conducted an interview to uncover designer-preferred qualities of personas (right).

By every similarity metric, LLM-SUMMARIZING reached statistically higher scores than the other workflows ($p < 0.01$). There were no statistically significant differences between LLM-AUTO and LLM-GROUPING, except for G-eval ($Z = -6.93$, $p < 0.01$). This indicates that LLM-SUMMARIZING generated the personas most semantically similar to the input data.

4.4.3 Summary. Proceeding from the results, we conclude that LLM-SUMMARIZING can generate the most statistically representative personas. Though unable to reproduce the exact distributions, it could capture all attributes present in user data and generate the personas that best match the data semantically. In practice, user researchers would not, for example, create one male and three female personas just to be faithful to their data's gender distribution. In other words, workflows do not need to match the data's distributions exactly as long as they capture the underlying attributes. Therefore, we looked further at the presence of existing attributes and also at semantic-similarity scores, continuing our evaluation with the two best-performing workflows (LLM-SUMMARIZING and LLM-GROUPING). Still, we believe that LLM-AUTO's strength in reproducing the attribute distribution could be beneficial in the early stages of user research, which we further discuss in Section 8.1.

5 STUDY 2: EVALUATION ON PERCEIVED QUALITY

We evaluated the perceived qualities of personas with user researchers (Figure 8). For this, we asked invited designers to create personas based on our user data. Then, we requested them to review the quality of personas generated by other designers (our baseline), LLM-SUMMARIZING (the most successful workflow), and LLM-GROUPING (the next-best workflow). With designers' assessments, we extended the statistical analysis from Study 1.

We collected 20 sets of survey responses from real people, via online crowdsourcing platform [3]. The survey was similar to that from Study 1, with the subject this time being designing AI courses for creative activities (See Appendix E for an example survey response). With the real user data collected for another subject, Study 2 not only tests persona generation in a more realistic setting but also demonstrates the generalizability of our workflows in working with a different set of user data.

To make a fair comparison between the conditions, each condition should generate the same number of personas with a shared goal. For this, we set "students' attitudes toward AI and what they want to learn by taking AI courses" for the key characteristics for grouping user data and specified the contents to be included in personas (name, age, gender, year of academic study, major, background, past experience of AI, attitude toward AI, and motivation for taking AI courses). From this setting, LLM-GROUPING created two personas at a time, depicting positive, negative, or mixed attitudes toward AI. Accordingly, we set the number of user groups to two for LLM-SUMMARIZING and designers. Appendix G presents an example persona from each condition.

5.1 Participants

We recruited 20 designers (mean age = 29, SD = 5.96), 10 of whom identified as female, nine as male, and one as "Other". Half of the participants had 1–5 years of experience in design, and the other half were above that experience threshold. All participants had experience in creating and using personas. They reported using these for various purposes, such as identifying the gaps between hypothetical and real users, communicating users' needs to collaborators, and defining customer journeys. All but one had used AI for their day-to-day tasks (e.g., utilizing ChatGPT for summarizing reports) and design activities (e.g., applying DALL-E [4] and Midjourney [5] to visualize concepts). For four hours of effort, we compensated each participant with a 40-euro voucher.

[3]https://www.prolific.co/

[4]https://openai.com/dall-e-2

[5]https://www.midjourney.com/

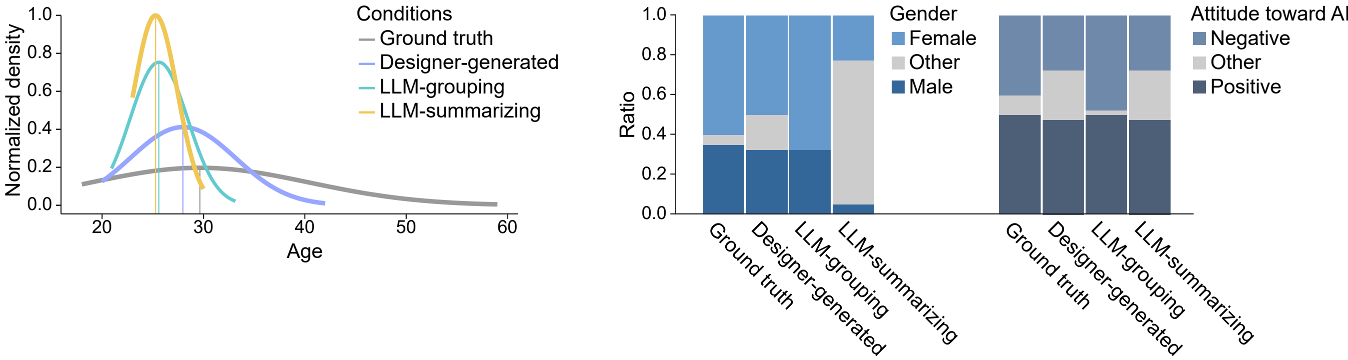

Figure 9: The distribution of age, gender, and attitude toward AI in the survey responses (ground truth) and generated personas in Study 2. Both designers and our workflows showed an ability to capture the median age and the attributes existing in the user data, but not all of them could reproduce the exact distribution of the attributes.

5.2 Tasks

Participants had two tasks: generating personas independently and evaluating personas, generated by other participants and our workflows. For the fair comparison, we instructed participants to create two personas based on the key characteristics, template, and survey responses used in our workflows. In all other respects, participants were free to decide how to analyze user data and create the personas' contents. Relying on their understanding of the survey responses, the participants evaluated personas from all three conditions (six personas in total), without knowing who generated them.

5.3 Measurement

We collected participants' perceptions of persona quality using the Persona Perception Scale (PPS) [70]. The PPS is designed to evaluate a persona's quality from designers' and stakeholders' perspectives. Following its developers' guidelines, we chose questions targeted at designers and excluded items for assessing the profile image (which personas sometimes include). The resulting instrument used 18 questions, measuring six qualities of personas: 1) *Consistency*, or how the various information in the persona is aligned; 2) *Completeness*, measuring how well the persona captures essential information about the users; 3) *Willingness to use (WTU)*, referring to how much practitioners wish to use the persona; 4) *Credibility*, denoting how realistic the persona appears to be; 5) *Clarity*, or how clearly the information in the persona is presented; and 6) *Empathy*, measuring how well the practitioners can empathize with the persona. Each question is answered on a seven-point Likert scale, from "strongly disagree" to "strongly agree".

5.4 Procedure

Our study had three parts: persona generation, persona evaluation, and post-evaluation interview. In persona generation, we met each participant online, introduced the purpose of our study, and obtained consent. Then, we gave participants the assignment of generating personas. Meanwhile, we created personas by following the LLM-grouping and LLM-summarizing workflows. For persona evaluation, we invited each participant to our lab and gave instructions to evaluate personas using the PPS. In a random order, we presented one set of personas at a time and asked participants to fill out the PPS form one persona at a time. We also supplied a printout of the survey responses in case they wanted to review them again. Lastly, with the interview, we asked about the participants' preferences and perceived differences between the personas created by designers and those from our workflows.

5.5 Results

5.5.1 Statistical representativeness. We performed statistical analyses similar to Study 1's. The distribution of age, gender, and attitude toward AI from each condition is shown in Figure 9. Our workflows demonstrated behaviors similar to those in Study 1. While they could capture the median age (all $p > 0.05$), they could not reproduce the age variance and the attribute distributions (all $p < 0.05$). Interestingly, the same was true for designer-generated personas. Nevertheless, both designers and our workflows captured all the attributes that were present in the data.

The semantic similar scores showed differences between our workflows and designers (Figure 10). For all scores, personas generated by designers were statistically less similar to the survey responses than those from the workflows (all $p < 0.01$). There were no statistically significant differences between LLM-summarizing and LLM-grouping (all $p > 0.05$). From the results, we conclude that the personas generated by the workflows are more statistically representative than the ones generated by unaided designers.

5.5.2 Perceived qualities of personas. The PPS results are shown in Figure 11. We performed the Friedman test (a non-parametric equivalent to repeated-measure ANOVA) with Bonferroni correction to compare the personas with regard to each quality. We found a statistically significant difference in how the participants perceived the completeness of the personas ($\chi^2(2) = 11.76$, $p = 0.02$). The Wilcoxon signed-Rank test with Bonferroni correction revealed that LLM-grouping produced personas less complete than designers' ($Z = -2.75$, $p = 0.02$) and LLM-summarizing ($Z = -2.95$, $p = 0.01$), which indicates that LLM-grouping's personas left out information that is vital for understanding the people represented. There were no statistically significant differences between the personas generated by designers and those from LLM-summarizing.

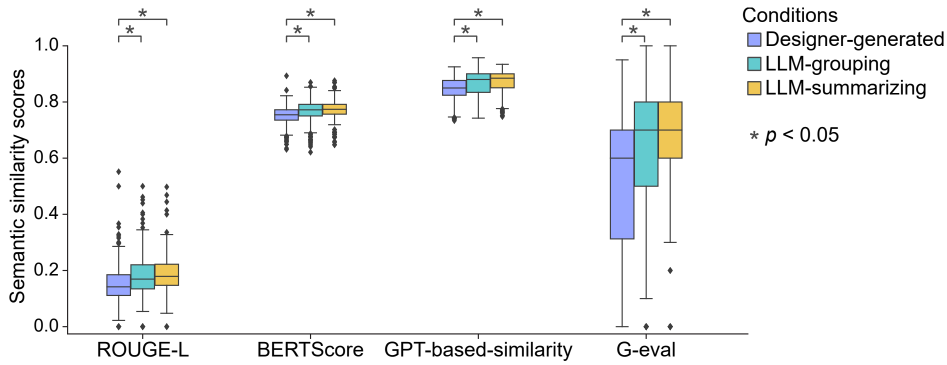

Figure 10: Scores for semantic similarity between user data and generated personas in Study 2. LLM-grouping and LLM-summarizing could generate personas that were more semantically similar to the user data than the designers' were.

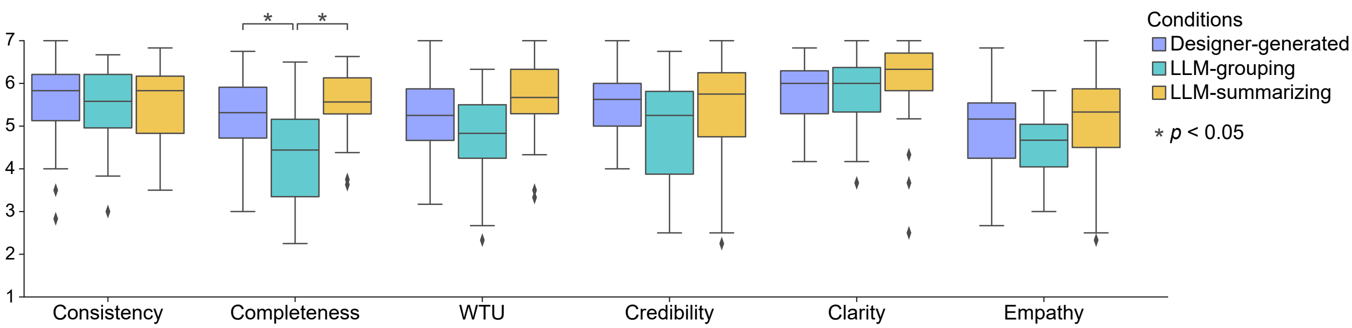

Figure 11: The perceived qualities of personas in Study 2, by PPS components. We found that the participants perceived the personas generated from LLM-summarizing to be similar to designer-generated ones.

The interview revealed that most participants could not correctly identify who had generated the personas. Ten of them thought that the LLM-summarizing personas were created by designers because they were more comprehensive than the other personas. Three others concluded that the LLM-grouping set had been created by designers since these highlighted the two most extreme attitudes toward AI. Our further queries, about the participants' preferred qualities in personas, identified three ares for improvement in LLM-generated personas:

- Firstly, the personas need to be more expressive (13 out of 20 participants). Participants reported favoring personas that use more expressive words and narration, which help them empathize on an emotional level. For instance, P11 commented, *"It's easier to connect with negative emotions when personas say they 'fear' something. Then I become empathetic."*
- Personas' backgrounds should provide motivation for the behaviors (11 out of 20 participants). This element often appeared lacking in both designer- and LLM-generated personas. For instance, P4 commented on a designer-generated persona that it *"talks about a person's interest, their likes, and dislikes, but it doesn't actually elaborate on their motivation."* Similarly, P13 commented on LLM-summarizing's output that *"I see a big gap in set C between personas' background, attitude, and motivation."*
- Lastly, personas should highlight only the most essential characteristics (10 out of 20 participants). Many participants noted that the personas that attempt to list all characteristics of a user group seemed too generic to represent archetypal users. For instance, P3 commented (on designer-generated and LLM-summarizing personas, respectively) that *"Set A is negative and positive about AI with a specific reasoning behind it, while C is just 'I am negative because of all these things'."*

5.5.3 Summary. We conclude that LLM-summarizing can generate personas that are statistically more representative than designer-generated ones while their perceived qualities do not differ. This implies that user researchers' effort is better spent on identifying key characteristics, rather than summarizing them. Accordingly, we believe that LLMs could offer a future where user researchers prioritize essential tasks of understanding users, with LLMs performing manual tasks. Potentially, this could make labor-intensive qualitative methods more efficient, increasing the adoption of rigorous user research rather than conducting them superficially.

6 STUDY 3: EVALUATION OF IMPROVED WORKFLOW

We conducted an evaluation study similar to Study 2 with an improved version of LLM-summarizing, which we refer to as LLM-summarizing++. In response to the three foci for improvement

Improvements	LLM-summarizing	LLM-summarizing++
Be more expressive	"I have a generally positive attitude toward AI. I believe it has the potential to..."	"Overall, I have a positive attitude towards AI. It makes me feel excited about the possibilities of..."
Recall previous contents as motivation	"I also don't want AI to replace human creativity... I'd like to know how AI could assist with tedious or time-consuming tasks, giving me more time to focus on the artistic aspects of my projects while also being aware of its limitations and implications."	"AI makes me a bit uneasy because I worry that it could take over the creative process, possibly replacing the originality and humane aspects in art... That's why I'm drawn to taking AI courses, to learn how to use this powerful tool while preserving the essence of human creativity in my work."
Emphasize the most important characteristics	"While studying graphic design, I also enjoy spending my free time doing a variety of activities like playing video games, hiking, painting, and playing instruments."	"As a graphic design major, I spend much of my free time engaged in various creative activities such as illustrating, playing video games, and painting."

Table 2: Examples of improvements from LLM-summarizing to LLM-summarizing++. We refined the prompt from LLM-summarizing to reflect the qualities that designers prefer to see in personas (the left column). We observed that LLM-summarizing++ narrates personas with emotional expressions (blue highlights), describes their behaviors aligned with their background (yellow highlights), and presents only the most important characteristics instead of listing all existing ones in the user data (green highlights).

found in Study 2, we added instructions to the prompt, which could get LLMs to (i) use more expressive words, (ii) recall previous contents for describing personas' motivation, and (iii) emphasize the characteristics most commonly expressed in the user data (Appendix C reproduced these instructions). With the updates from LLM-summarizing to LLM-summarizing++, we observed the anticipated improvements (Table 2). Using a crowdsourcing platform, we recruited user researchers to assess the personas generated by the two workflows. For this, we used the same instrument (PPS), user data, and personas as we had for Study 2. The only difference was that we did not have participants create personas, since we already obtained designer-generated personas.

6.1 Participants

We recruited 31 user researchers (mean age = 38, SD = 9.96), 12 self-identifying as females and 19 as males. The participants represented various fields that utilize personas, such as market research and product development. All participants had experience in creating and using personas, similar to those in Study 2. For their hour of effort, we compensated each participant with 10 euros.

6.2 Procedure

We used an online survey conducted via crowdsourcing site Prolific. We instructed participants to read through user data and evaluate personas. To make sure they read the data, we had them answer five multiple-choice questions testing their understanding of the users (e.g., "What are users' attitudes toward AI?"). After passing the test, participants completed one PPS survey for each of eight personas, presented in randomized order. We also included nonsensical questions between PPS tasks, to guarantee that participants were paying attention to the survey items. For this, we followed the attention-check policy of Prolific and guidelines for running experiments that employ crowdsourcing [58].

6.3 Result

6.3.1 Statistical representativeness. Comparing the attribute distribution and semantic similarity scores between LLM-summarizing and LLM-summarizing++ (Figure 12), we found that the updated prompt had little influence on the personas' statistical representativeness. The statistical analysis was the same as in Study 2. There were no statistically significant differences in either the age distribution and semantic-similarity scores between LLM-summarizing and LLM-summarizing++ (all $p > 0.05$). Also, while the age variances were statistically different between LLM-summarizing++ and the user data ($F(1, 58) = 31.63$, $p = 0.02$), their median ages were not statistically different ($Z = -1.87$, $p = 0.06$). Likewise, both workflows' distributions of gender and attitude toward AI were statistically different from the user data's (all $p < 0.01$). This indicates that LLM-summarizing++ retained most of the statistical representativeness of LLM-summarizing, and the two were similar in their divergences from the user data with regard to the gender and attitude distributions.

6.3.2 Perceived qualities of personas. We analyzed the PPS results by applying statistical analysis methods identical to those in Study 2, with adjusted Bonferroni correction to match the additional comparisons (Figure 13). Statistically significant differences were visible for consistency ($\chi^2(3) = 15.89$, $p < 0.01$), completeness ($\chi^2(3) = 13.61$, $p = 0.02$), WTU ($\chi^2(3) = 16.71$, $p < 0.01$), and empathy ($\chi^2(3) = 8.46$, $p = 0.04$). *Post-hoc* analysis showed that LLM-summarizing++ personas received statistically better scores than designer-generated ones for consistency ($Z = -2.95$, $p = 0.01$), WTU ($Z = -2.61$, $p = 0.01$), and empathy ($Z = -2.35$, $p = 0.02$) and had statistically higher scores than LLM-grouping personas for consistency ($Z = -2.63$, $p = 0.01$), completeness ($Z = -3.59$, $p < 0.01$), WTU ($Z = -3.36$, $p < 0.01$), and empathy ($Z = -3.11$, $p < 0.01$). While the mean score for each item was higher with LLM-summarizing++ relative to LLM-summarizing, these differences were not statistically significant (all $p > 0.05$). Echoing the result from Study 2, this study found no statistically significant difference between designer-generated and LLM-summarizing personas (all $p > 0.05$).

6.3.3 Summary. We conclude from Study 3 that LLM-summarizing++ can generate personas that are more consistent, useful, and empathy-evoking personas than the ones generated by designers alone. This is possible through prompting LLMs (GPT-4) with detailed descriptions of how they should narrate personas. This supports Kanarik et al.'s findings that human experts' experiences can augment AI's

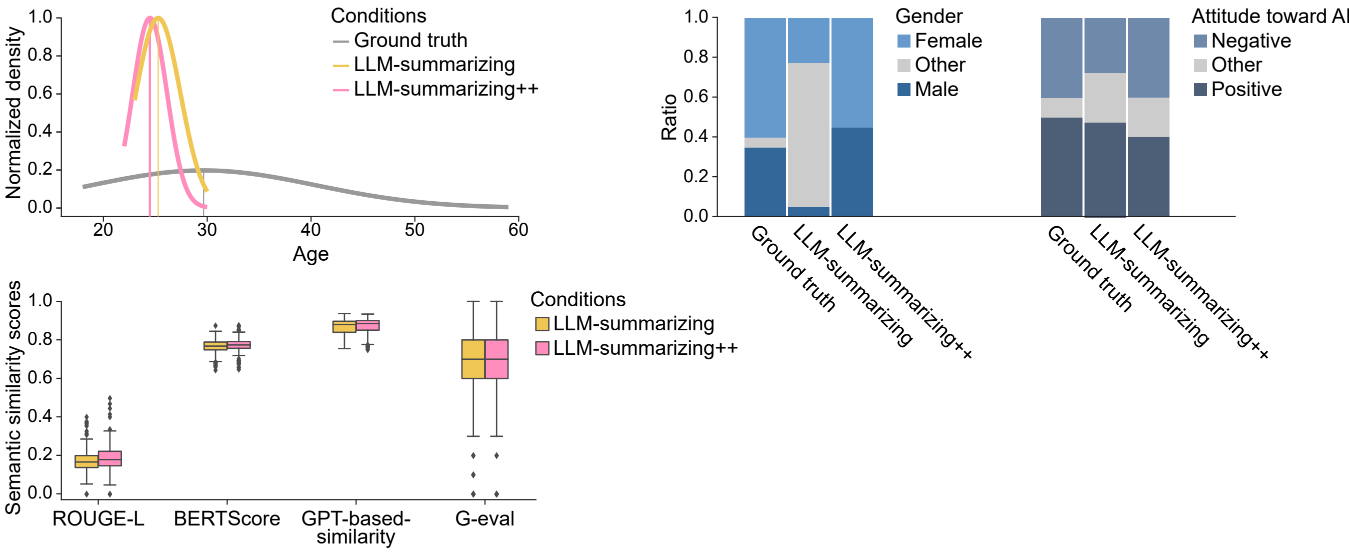

Figure 12: The attribute distributions and semantic similarity scores of LLM-summarizing and LLM-summarizing++ in Study 3. LLM-summarizing++ mostly retained the statistical representativeness of LLM-summarizing.

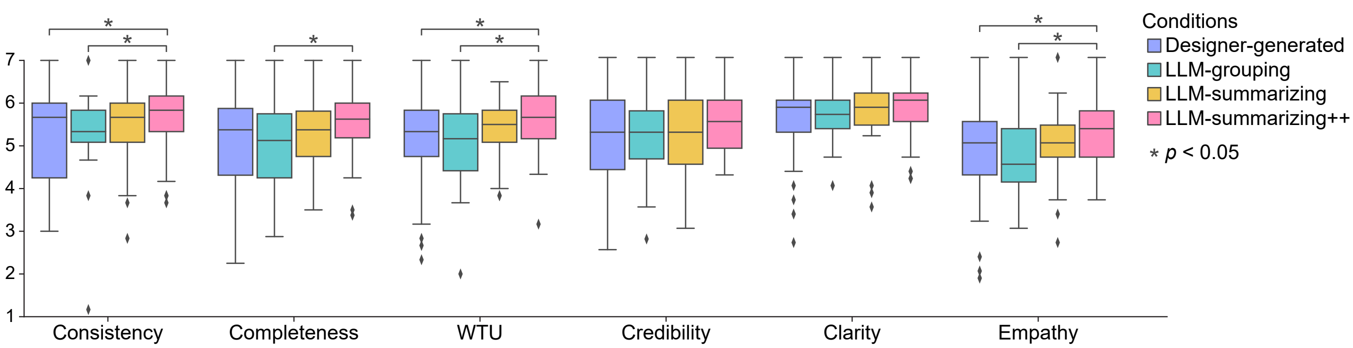

Figure 13: The perceived qualities of personas measured using PPS in Study 3. LLM-summarizing++ proved able to generate personas perceived as more consistent, useful (WTU), and empathy-evoking than those created by designers alone.

performance [34]. In the realm of persona generation, our results suggest that user researchers' abilities to recognize users' important traits and personas' qualities augment LLMs' summarization.

7 STUDY 4: EXPLORATION OF INTERACTIVE PERSONAS

The first three studies demonstrate that LLMs can help generate personas in human–AI workflows. To extend the workflows, we conducted a formative study exploring how LLMs can give user researchers additional assistance in using personas by enabling "interaction with personas" (Figure 2). To examine this, we implemented a system wherein users can generate personas by following our LLM-summarizing workflow (Figure 14). The system's chatbot assistant utilizes GPT-4 to guide the step-by-step process of persona generation. More importantly, the chatbot can converse with users by relying on both generated personas and input data (*Prompt: Here are survey responses: {data}. Here are personas based on the survey responses: {personas}. Respond to the following request based on the survey responses and the personas: {user input in the chat}*). We demonstrated our system to the 20 designers who had participated in Study 2 and asked them to interact freely with the chatbot. We recorded what participants asked the chatbot, after which interviews addressed the benefits and weaknesses of interacting with personas.

7.1 Results

Our formative study showcased how interaction with personas can enhance the use of them. Categorizing the interactions uncovered two main themes. The first of these is role-playing with personas. Most participants (16 out of 20) asked our chatbot to answer their questions while acting as one of the personas generated. This was handled similarly to interviewing a user. For instance, participants tried to ask what the personas would want or think about potential designs. To this end, P20 asked, *"How could an AI course on creativity address your fears, as Sara [the name of persona]?"* They also sought a simulation of user behavior that the user data had not captured. For

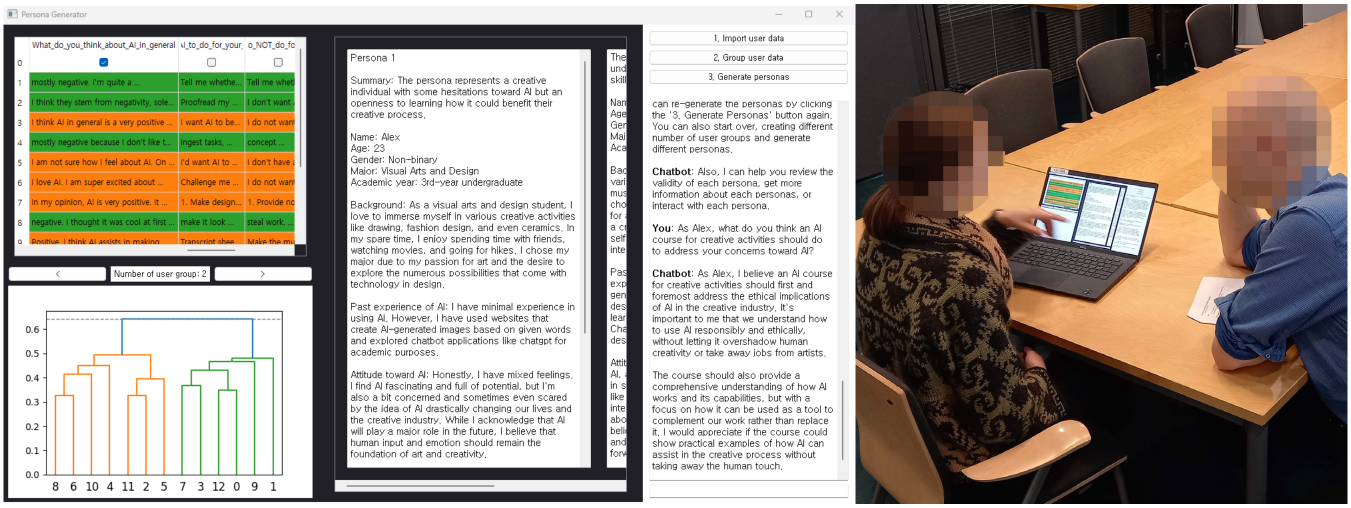

Figure 14: A screenshot showing our persona generator. It guides users in generating personas via the LLM-SUMMARIZING workflow and enables chatting with generated personas (left). We interviewed designers on interacting with personas (right) and found most of them are interested in interviewing personas to test their ideas or further inquire about users.

instance, P13 asked our chatbot to describe personas' daily activities, and P2 asked it to simulate interaction between generated personas. All such interactions were deemed to reflect interest in expanding the use of personas beyond the level of written narrations.

The second theme is inquiring about users through personas. Four participants expressed a desire to discover user characteristics that personas might not encapsulate. For example, P9 asked our chatbot whether there were any outliers whom the generated personas would not represent. Likewise, P16 asked, *"Can you tell me more personal things about persona 2 besides their hobbies?"* in hopes of delving further into the personas' contents. After the interaction, the participants commented that analyzing user data "through" personas could bring more efficiency to this task, which is often repetitive and challenging to carry out manually.

While the participants valued the interaction with personas, they did remark on possible limitations. With regard to the non-technical ones, all participants commented on LLMs' lack of proactivity. For instance, P20 stated, *"Data cannot capture everything that a human can feel. Those personas will only be as good as the data that they [LLMs] use."* Likewise, the participants pointed out that LLMs would not suggest collecting more user data or different interpretations unless directly asked to do so. Accordingly, the participants commented that human experts would still need to be aware of potential issues with the generation of personas.

8 DISCUSSION

User researchers, like anyone else, can be easily "lured into" using LLMs without deeply considering the most beneficial ways of using them. Our results demonstrate the importance of understanding what human experts and LLMs both bring to the table at subtask level, which enabled us to tailor human-AI workflows that significantly improve the outcome quality. We found specifically that the most representative and empathy-evoking personas can arise from human experts taking the lead role in creating key user groups and exploiting LLMs' summarization capabilities. In contrast, When LLMs take charge of the subtasks, lower-quality personas result, underrepresenting users' key characteristics that are fundamental for understanding user behaviors. Rather, we identified particular potential in LLMs "acting out" generated personas and thereby fleshing out user researchers' understanding of the personas and the users they represent. In summary, through the four studies, we obtained these key findings:

- LLM-SUMMARIZING generates the most statistically representative personas (Study 1). This is the workflow wherein user researchers identify key characteristics and group user data with clustering algorithms while LLMs summarize grouped user data into personas.
- LLM-SUMMARIZING produces personas that are more semantically similar to input data and perceivable as of a quality level similar to designer-generated personas' (Study 2).
- LLM-SUMMARIZING++ generates personas that may be perceived as more consistent, useful (as judged by willingness to use), and empathy-evoking than designer-generated personas (Study 3).
- LLMs offer the additional benefit of "role-plays" with staging wherein designers converse with a generated persona. When we let participants interact with LLMs, 16 out of 20 tried to ask about users' common opinions as if conducting interviews with the personas (Study 4).

8.1 Tradeoffs of persona-generation workflows

While LLM-SUMMARIZING++ performed best of all, distinctive tradeoffs attests to the potential of other workflows. We found a special strength of LLM-AUTO in reproducing the attribute distributions of user data. Study 1 showed this to be the only workflow that did not lead to statistically significant differences from the distributions in user data. We assume this outcome arose because the LLM was free to define user groups and contents without specific instructions, rather than being instructed in convergence of user data to form personas. Although reproducing the exact attribute distributions is not necessary for generating personas [11], this workflow could be useful when user researchers need to review the overall distributions of users' attributes. Especially at the outset, knowing the ground truth in user data could afford identifying key characteristics that personas must ultimately represent or pinpointing any minor but important groups [7, 33].

We found LLM-GROUPING beneficial for creating reasonable-quality personas while expending less effort than LLM-SUMMARIZING requires. In Study 2, LLM-GROUPING generated personas that were more semantically similar to the user data than designer-generated personas, while their perceived qualities were similar for the most part. The only difference was that LLM-GROUPING personas were perceptibly less complete, which means that the characteristics of the user data were accentuated less [70]. We assume that having LLMs group user data might make the model focus on distinctive characteristics and less on details. As Figure 9 shows, LLM-GROUPING indeed attended to two extreme cases (here, negative and positive attitudes toward AI), in marked contrast against the other workflows. Analogously to how we improved LLM-SUMMARIZING with more sophisticated prompt designs, researchers probably could take advantage of LLM-GROUPING likewise. Potentially, prompting LLMs to elaborate on users' key characteristics could make LLM-GROUPING another workflow that holds promise for generating high-quality personas with less effort.

8.2 LLMs and the changing landscape of user research

Beyond persona generation, our results open an exciting vista for using LLMs in user research. Firstly, LLMs allow applying algorithmic approaches to qualitative data, which can make qualitative user research more effective. Conventional approaches to analyzing qualitative user data have heavily relied on user researchers' manual work and interpretations. This has sparked criticisms of personas for being biased and unreliable as foundation for understanding target users [13, 62]. Whereas recent data-driven approaches have proven abilities to analyze user data statistically and mitigate user researchers' biases [47, 51, 66, 67], they have been employed almost exclusively with quantitative data, not to yield in-depth insights that can be synthesized from qualitative data only. In contrast, our results suggest a role for LLMs in helping experts identify interesting patterns in qualitative data. Further, we observed that auxiliary methods could help keep the distributional biases of LLMs. Accordingly, we find LLMs able to augment conventional approaches well, assisting user researchers to support their interpretations with algorithmic analysis.

Secondly, LLMs permit a completely new way to experience personas: role-playing. There is growing evidence that LLMs can produce consistent behaviors that adhere to given personalities [8, 10, 60, 80]. We also observed, in Study 4, that the LLM can act as generated personas and interact with designers. The central benefit of this could be a more effective route to understanding user data than simply reading about personas. For instance, user researchers can simulate interviews and further inquire about users' characteristics and opinions. Even more importantly, user researchers can test their ideas "with" personas, beyond written narrations in persona form. These interactions align well with the principle behind using personas, for greater empathy toward users and deep engagement with design problems from user perspectives [11]. Potentially, role-playing with LLMs can be applied to make other user-research methods more interactive. For instance, techniques such as using user scenarios and journey mapping could be augmented with LLMs that narrate the user experience from target users' perspectives.

Still, potential negative impacts of LLMs mimicking generated personas should not be overlooked. We saw that role-playing with LLMs could cause reliance on personas over real users, thereby leading to incorrect assumptions about them. Study 4 showed that the interactions involved might not encourage evaluation with real users – even when the LLMs might be generating user responses that are not grounded in the data. Accordingly, we conclude that LLMs can expand how user researchers work with user data. Yet, it remains an open question how role-playing affects possible biases in interpreting user data. The core challenge is how we can make the most of LLMs "without falling into the trap of anthropomorphism" [74] – i.e., considering them the same as real-world users.

8.3 Negative consequences of overreliance on LLMs

Excessively automating user research is naive and bound to fail: some human judgments cannot be replaced by state-of-the-art AI. We observed that user researchers bring unique human values to judging what is important for their projects. In particular, Study 2 showed that designers are good at identifying users' key characteristics and their relationships to design problems. Also, their ability to identify empathy-evoking qualities as another human being appeared irreplaceable. Therefore, removing the human from the loop would cause the outcomes' quality to deteriorate.

We believe that careless automation creates second-order consequences. Firstly, user researchers' ability to work with user data could suffer. For instance, their understanding of users grows through looking closely at user data as analysis progresses. If this task is delegated to LLMs in an attempt to reduce human effort, user researchers lose that opportunity to deepen their understanding of users. In addition, the researchers end up expending more effort in validating LLM-generated analysis. This is the "irony of automation" [3]: it might end up redirecting user researchers' efforts instead of reducing them. Secondly, people's level of willingness to delegate their tasks to AI must not be overlooked. Studies in human-centered AI indicate that effective collaboration depends not just on AI's performance but on human trust in AI also [32, 39]. As Lubars and Tan [44] showed, people prefer to take control of their tasks, with AI in an assisting role, even when AI displays competencies in completing the tasks alone. Accordingly, human–AI workflows that simply allocate more tasks to LLMs are bound not to endure in practice. Therefore, we recommend taking more time to examine LLMs' influence, so that we can (i) understand the consequences of human–AI workflow in user research and (ii) find the right balance for automation, one that benefits user researchers the most from the best aspects of LLMs.

8.4 Implementing persona-generation workflow in practice

Generating persona is a user study method that can be applied to any other user-research context that requires deep understanding of user data. Therefore, in principle, our workflows can be applied to diverse contexts, unlimited to designing civic services and course activities presented in this paper. In particular, we expect our workflows to be beneficial in contexts wherein practitioners need to cover a broad range of user behaviors or in which user behaviors frequently change over time (e.g., social media settings), swiftly generating multiple personas from high-volume user data.

Based on our workflow designs and insights, we present five steps for implementing the workflows in practice (we also make our code and prompts available to demonstrate the steps [6]). First, *organize user data in a question–answer format.* Grouping user responses by interview/survey item helps practitioners easily recognize and control what kinds of user data they enter in prompts. Also, question–answer formats align well with LLMs' input style, hence enabling better context recognition and summary generation. Second, *identify the group of question–answer data containing users' key characteristics.* Finding the most informative parts of dataset is an essential step in user research. To do this, practitioners could use a neutral prompt, as applied in LLM-AUTO, to summarize each group of data without introducing their biases. Practitioners can then exercise their judgments to select important data for their projects. Third, *create user groups with sufficient differences.* Each user group should be archetypal, distinctive from every other one. This could be achieved by means of the semantic-clustering method of LLM-SUMMARIZING (see Appendix A.3). By adjusting the number of clusters and prompting LLMs to summarize them, practitioners can ascertain the appropriate number of user groups and avoid focusing on often-trivial factors such as age or gender. Fourth, *create a template of personas.* The basic template could include the general contents such as name, age, gender, occupation, background, and motivation in line with previous work on persona generation [11, 49, 67]. In describing this template to LLMs, the relations between the contents should be articulated clearly, to support high-quality personas (see Section 6). Finally, *generate the personas.* Practitioners can achieve this by following our prompt design (see Table 4). we recommend, on the basis of Study 4, interacting with personas to assess their representation of user data and to deepen one's understanding of target users.

8.5 Limitations and future work

The first limitation deserving mention is that the workflows were tested with GPT-4 only. Since this is the most advanced LLM available [82], we selected it to explore the full potential of what LLMs currently offer. Testing human–AI workflows using other LLMs could be expected to offer benefits for generalizing workflow and prompt designs. Secondly, we evaluated personas with user researchers only. Alternatively, they can be examined by users, directly, in the course of evaluation processes. While we found the personas to be statistically representative and perceived as such by user researchers, users might focus on other qualities of personas [70]. To study this, future work could investigate human–AI workflows that draw from multiple domains of human expertise. Potential exists for adding other roles to the workflow (e.g., facilitators who delegate tasks between human experts and LLMs.

Considering the advancements in LLMs, we anticipate that future work could investigate multi-modal human–AI workflows. In practice, user researchers collect user data in various forms (e.g., pictures and videos). Recent LLM developments point to a shift toward multi-modal prompts that can accept content such as images as input [56]. Given our finding that the best results might unfold when human experts pre-group user data while LLMs summarize grouped user data, scholars could examine LLMs' ability to group or summarize such non-textual data. Another important direction is exploring more diverse prompt engineering techniques for delegating subtasks to LLMs. There are many unexplored prompting techniques that aim to divide complex tasks into simpler subtasks, such as prompt chaining [81] and automatic multi-step reasoning [59]. Investigating them could improve the delegability of human experts' tasks and expand the design space of human-AI workflows.

Future work could also investigate human–AI workflow in the other context of user research or design. For instance, the next step in persona generation could be to create use scenarios that can "vivify" personas in a specific sequence of actions. Investigating this could expand human–AI workflows in pursuit of actionable solutions or designs. In parallel with this, research could explore multiple variations of human–AI workflows. Our work focused on ones that exploit LLMs' primary competence in text summarization and human experts' strengths in identifying key characteristics. Potentially, subtasks could be further divided or performed collaboratively (e.g., with LLMs identifying initial characteristics to "jump-start" user researchers' analysis). Such solutions could identify additional synergy benefits from human experts working with LLMs. Lastly, we believe that ethical concerns surrounding AI-generated personas should be investigated thoroughly. For instance, whether people will accept AI that represents – and role-plays – them mertis longer-term study, especially as people and societies grow more exposed to AI technologies in their various forms.

9 CONCLUSION

Understanding which subtasks should be delegated to human experts versus AI is critical for designing interactive systems. Failing in this leads to arbitrary task delegation that risks replacing a unique advantage that the other agent cannot provide. In this paper, we studied human–AI workflows in the context of persona generation by differently delegating the subtasks between human experts and LLMs. Our findings suggest that experts' efforts are better spent defining important user groups and personas' qualities while LLMs summarize grouped user data accordingly. Our workflow that follows this approach generated the most representative and empathy-evoking personas compared to the other workflows and the experts working alone. Those personas highlight the most important characteristics of users using expressive narrations, thus helping user researchers understand target users' behaviors and empathize with them. Based on our findings, we discuss LLMs' positive as well as potentially negative influences on user research and inform the design directions of working with LLMs. Our work

[6] https://github.com/joongishin/persona-generation-workflow

extends the existing literature on AI-assisted user research by identifying human experts’ and LLMs’ values in persona generation.

ACKNOWLEDGMENTS

This research was supported by Human Automata, grant 328813; Subjective Functions, grant 357578; and IFI program of the German Academic Exchange Service (DAAD).

REFERENCES

[1] Yongsu Ahn, Yu-Ru Lin, Panpan Xu, and Zeng Dai. 2023. ESCAPE: Countering Systematic Errors from Machine’s Blind Spots via Interactive Visual Analysis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI ’23)*. Association for Computing Machinery, New York, NY, USA, Article 834, 16 pages. https://doi.org/10.1145/3544548.3581373

[2] Ayham Alomari, Norisma Idris, Aznul Qalid Md Sabri, and Izzat Alsmadi. 2022. Deep reinforcement and transfer learning for abstractive text summarization: A review. *Computer Speech& Language* 71 (2022), 101276. https://doi.org/10.1016/j.csl.2021.101276

[3] Gordon Baxter, John Rooksby, Yuanzhi Wang, and Ali Khajeh-Hosseini. 2012. The Ironies of Automation: Still Going Strong at 30?. In *Proceedings of the 30th European Conference on Cognitive Ergonomics* (Edinburgh, United Kingdom) *(ECCE ’12)*. Association for Computing Machinery, New York, NY, USA, 65–71. https://doi.org/10.1145/2448136.2448149

[4] David Benyon. 2019. *Designing user experience.* Pearson UK.

[5] Ali Borji. 2023. A Categorical Archive of ChatGPT Failures. arXiv:2302.03494 [cs.CL]

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[7] Peter Brusilovsky and Eva Millán. 2007. *User Models for Adaptive Hypermedia and Adaptive Educational Systems.* Springer Berlin Heidelberg, Berlin, Heidelberg, 3–53. https://doi.org/10.1007/978-3-540-72079-9_1

[8] Chris Callison-Burch, Gaurav Singh Tomar, Lara J. Martin, Daphne Ippolito, Suma Bailis, and David Reitter. 2022. Dungeons and Dragons as a Dialog Challenge for Artificial Intelligence. arXiv:2210.07109 [cs.CL]

[9] Rishabh Choudhary, Omar Alsayed, Simona Doboli, and Ali A. Minai. 2022. Building Semantic Cognitive Maps with Text Embedding and Clustering. In *2022 International Joint Conference on Neural Networks (IJCNN)*. 01–08. https://doi.org/10.1109/IJCNN55064.2022.9892429

[10] Yun-Shiuan Chuang, Siddharth Suresh, Nikunj Harlalka, Agam Goyal, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T. Rogers. 2023. Evaluating LLM Agent Group Dynamics against Human Group Dynamics: A Case Study on Wisdom of Partisan Crowds. arXiv:2311.09665 [cs.CL]

[11] Alan Cooper, Robert Reimann, David Cronin, and Christopher Noessel. 2014. *About face : the essentials of interaction design* (fourth edition. ed.). John Wiley W& Sons, Indianapolis, Indiana.

[12] Samuel Rhys Cox, Ashraf Abdul, and Wei Tsang Ooi. 2023. Prompting a Large Language Model to Generate Diverse Motivational Messages: A Comparison with Human-Written Messages. arXiv:2308.13479 [cs.CL]

[13] Nick De Voil. 2010. Personas considered harmful. *De Voil Consulting Limited* (2010).

[14] Adrian de Wynter, Xun Wang, Alex Sokolov, Qilong Gu, and Si-Qing Chen. 2023. An evaluation on large language model outputs: Discourse and memorization. *Natural Language Processing Journal* 4 (sep 2023), 100024. https://doi.org/10.1016/j.nlp.2023.100024

[15] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD: Dataset and metrics for measuring biases in open-ended language generation. In *ACM FAccT 2021*. https://www.amazon.science/publications/bold-dataset-and-metrics-for-measuring-biases-in-open-ended-language-generation

[16] Esin Durmus, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. Towards Measuring the Representation of Subjective Global Opinions in Language Models. arXiv:2306.16388 [cs.CL]

[17] Alex Endert, Seth Fox, Dipayan Maiti, Scotland Leman, and Chris North. 2012. The Semantics of Clustering: Analysis of User-Generated Spatializations of Text Documents. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (Capri Island, Italy) *(AVI ’12)*. Association for Computing Machinery, New York, NY, USA, 555–562. https://doi.org/10.1145/2254556.2254660

[18] Shamal Faily and Ivan Flechais. 2011. Persona Cases: A Technique for Grounding Personas. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI ’11)*. Association for Computing Machinery, New York, NY, USA, 2267–2270. https://doi.org/10.1145/1978942.1979274

[19] Bruna Moraes Ferreira, Simone D. J. Barbosa, and Tayana Conte. 2016. PATHY: Using Empathy with Personas to Design Applications that Meet the Users’ Needs. In *Human-Computer Interaction. Theory, Design, Development and Practice*, Masaaki Kurosu (Ed.). Springer International Publishing, Cham, 153–165.

[20] Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi. 2022. Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT ’22)*. Association for Computing Machinery, New York, NY, USA, 1362–1374. https://doi.org/10.1145/3531146.3533193

[21] Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like Summarization Evaluation with ChatGPT. arXiv:2304.02554 [cs.CL]

[22] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for Science Writing using Language Models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference* (<conf-loc>, <city>Virtual Event</city>, <country>Australia</country>, </conf-loc>) *(DIS ’22)*. Association for Computing Machinery, New York, NY, USA, 1002–1019. https://doi.org/10.1145/3532106.3533533

[23] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* 120, 30 (jul 2023). https://doi.org/10.1073/pnas.2305016120

[24] Elizabeth Goodman and Mike Kuniavsky. 2012. *Observing the user experience: A practitioner’s guide to user research.* Elsevier.

[25] Ben Green and Yiling Chen. 2019. The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 50 (nov 2019), 24 pages. https://doi.org/10.1145/3359152

[26] Hongyan Gu, Yuan Liang, Yifan Xu, Christopher Kazu Williams, Shino Magaki, Negar Khanlou, Harry Vinters, Zesheng Chen, Shuo Ni, Chunxu Yang, Wenzhong Yan, Xinhai Robert Zhang, Yang Li, Mohammad Haeri, and Xiang ‘Anthony’ Chen. 2023. Improving Workflow Integration with XPath: Design and Evaluation of a Human-AI Diagnosis System in Pathology. *ACM Trans. Comput.-Hum. Interact.* 30, 2, Article 28 (mar 2023), 37 pages. https://doi.org/10.1145/3577011

[27] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. 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* (Hamburg, Germany) *(CHI ’23)*. Association for Computing Machinery, New York, NY, USA, Article 433, 19 pages. https://doi.org/10.1145/3544548.3580688

[28] Pengcheng He, Baolin Peng, Song Wang, Yang Liu, Ruochen Xu, Hany Hassan, Yu Shi, Chenguang Zhu, Wayne Xiong, Michael Zeng, Jianfeng Gao, and Xuedong Huang. 2023. Z-Code++: A Pre-trained Language Model Optimized for Abstractive Summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 5095–5112. https://doi.org/10.18653/v1/2023.acl-long.279

[29] Elizabeth Y. Huang, Dario Paccagnan, Wenjun Mei, and Francesco Bullo. 2023. Assign and Appraise: Achieving Optimal Performance in Collaborative Teams. *IEEE Trans. Automat. Control* 68, 3 (2023), 1614–1627. https://doi.org/10.1109/TAC.2022.3156879

[30] Jina Huh, Bum Chul Kwon, Sung-Hee Kim, Sukwon Lee, Jaegul Choo, Jihoon Kim, Min-Je Choi, and Ji Soo Yi. 2016. Personas in online health communities. *Journal of Biomedical Informatics* 63 (2016), 212–225. https://doi.org/10.1016/j.jbi.2016.08.019

[31] Bernard J. Jansen, Soon gyo Jung, and Joni Salminen. 2023. Employing large language models in survey research. *Natural Language Processing Journal* 4 (2023), 100020. https://doi.org/10.1016/j.nlp.2023.100020

[32] Jialun Aaron Jiang, Kandrea Wade, Casey Fiesler, and Jed R. Brubaker. 2021. Supporting Serendipity: Opportunities and Challenges for Human-AI Collaboration in Qualitative Analysis. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 94 (apr 2021), 23 pages. https://doi.org/10.1145/3449168

[33] Plinio Thomaz Aquino Junior and Lucia Vilela Leite Filgueiras. 2005. User Modeling with Personas. In *Proceedings of the 2005 Latin American Conference on Human-Computer Interaction* (Cuernavaca, Mexico) *(CLIHC ’05)*. Association for Computing Machinery, New York, NY, USA, 277–282. https://doi.org/10.1145/1111360.1111388

[34] Keren J Kanarik, Wojciech T Osowiecki, Yu Lu, Dipongkar Talukder, Niklas Roschewsky, Sae Na Park, Mattan Kamon, David M Fried, and Richard A Gottscho. 2023. Human–machine collaboration for improving semiconductor process development. *Nature* 616, 7958 (2023), 707–711. https://doi.org/10.1038/s41586-023-05773-7

[35] Jeongyeon Kim, Sangho Suh, Lydia B Chilton, and Haijun Xia. 2023. Metaphorian: Leveraging Large Language Models to Support Extended Metaphor Creation

for Science Writing. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (<conf-loc>, <city>Pittsburgh</city>, <state>PA</state>, <country>USA</country>, </conf-loc>) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 115–135. https://doi.org/10.1145/3563657.3595996

[36] Ezgi Korkmaz. 2022. Revealing the Bias in Large Language Models via Reward Structured Questions. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*. https://openreview.net/forum?id=7uYsFvahSzx

[37] Dannie Korsgaard, Thomas Bjørner, Pernille Krog Sørensen, and Paolo Burelli. 2020. Creating user stereotypes for persona development from qualitative data through semi-automatic subspace clustering. *User modeling and user-adapted interaction* 30 (2020), 81–125. https://doi.org/10.1007/s11257-019-09252-5

[38] Mike Kuniavsky. 2003. *Observing the user experience: a practitioner's guide to user research*. Elsevier.

[39] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q. Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>New Orleans</city>, <state>LA</state>, <country>USA</country>, </conf-loc>) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 54, 18 pages. https://doi.org/10.1145/3491102.3501999

[40] Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Gunhee Kim, and Jung-Woo Ha. 2023. KoSBi: A Dataset for Mitigating Social Bias Risks Towards Safer Large Language Model Application. arXiv:2305.17701 [cs.CL]

[41] Wanhae Lee, Minki Chun, Hyeonhak Jeong, and Hyunggu Jung. 2023. Toward Keyword Generation through Large Language Models. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 37–40. https://doi.org/10.1145/3581754.3584126

[42] Chin-Yew LIN. 2004. ROUGE : a package for automatic evaluation of summaries. *Proceedings of the Workshop on Text Summarization Branches Out, 2004* (2004). https://cir.nii.ac.jp/crid/1571417125576321408

[43] Anthony Liu, Santiago Guerra, Isaac Fung, Gabriel Matute, Ece Kamar, and Walter Lasecki. 2020. Towards Hybrid Human-AI Workflows for Unknown Unknown Detection. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 2432–2442. https://doi.org/10.1145/3366423.3380306

[44] Brian Lubars and Chenhao Tan. 2019. Ask not what AI can do, but what AI should do: Towards a framework of task delegability. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/d67d8ab4f4c10bf22aa353e27879133c-Paper.pdf

[45] Ana Macanovic. 2022. Text mining for social science – The state and the future of computational text analysis in sociology. *Social Science Research* 108 (2022), 102784. https://doi.org/10.1016/j.ssresearch.2022.102784

[46] Megh Marathe and Kentaro Toyama. 2018. Semi-Automated Coding for Qualitative Research: A User-Centered Inquiry and Initial Prototypes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173922

[47] Nicola Marsden and Maren Haag. 2016. Stereotypes and Politics: Reflections on Personas. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4017–4031. https://doi.org/10.1145/2858036.2858151

[48] Nicola Marsden and Maren Haag. 2016. Stereotypes and Politics: Reflections on Personas. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4017–4031. https://doi.org/10.1145/2858036.2858151

[49] Tara Matthews, Tejinder Judge, and Steve Whittaker. 2012. How Do Designers and User Experience Professionals Actually Perceive and Use Personas?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1219–1228. https://doi.org/10.1145/2207676.2208573

[50] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. arXiv:2005.00661 [cs.CL]

[51] Jennifer (Jen) McGinn and Nalini Kotamraju. 2008. Data-Driven Persona Development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 1521–1524. https://doi.org/10.1145/1357054.1357292

[52] Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of Hallucination by Large Language Models on Inference Tasks. arXiv:2305.14552 [cs.CL]

[53] Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. 2022. Text and Code Embeddings by Contrastive Pre-Training. arXiv:2201.10005 [cs.CL]

[54] Timothy J. Nokes-Malach, Michelle L. Meade, and Daniel G. Morrow. 2012. The effect of expertise on collaborative problem solving. *Thinking & Reasoning* 18, 1 (2012), 32–58. https://doi.org/10.1080/13546783.2011.642206 arXiv:https://doi.org/10.1080/13546783.2011.642206

[55] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on Medical Challenge Problems. arXiv:2303.13375 [cs.CL]

[56] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[57] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[58] Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5, 5 (2010), 411–419. https://doi.org/10.1017/S1930297500002205

[59] Bhargavi Paranjape, Scott Lundberg, Sameer Singh, Hannaneh Hajishirzi, Luke Zettlemoyer, and Marco Tulio Ribeiro. 2023. ART: Automatic multi-step reasoning and tool-use for large language models. arXiv:2303.09014 [cs.CL]

[60] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442 [cs.HC]

[61] John Pruitt and Tamara Adlin. 2010. *The persona lifecycle: keeping people in mind throughout product design*. Elsevier.

[62] John Pruitt and Jonathan Grudin. 2003. Personas: practice and theory. In *Proceedings of the 2003 conference on Designing for user experiences*. 1–15.

[63] Indra Kharisma Raharjana, Daniel Siahaan, and Chastine Fatichah. 2021. User Stories and Natural Language Processing: A Systematic Literature Review. *IEEE Access* 9 (2021), 53811–53826. https://doi.org/10.1109/ACCESS.2021.3070606

[64] Alex Renda, Aspen Hopkins, and Michael Carbin. 2023. Can LLMs Generate Random Numbers? Evaluating LLM Sampling in Controlled Domains. In *ICML Workshop*.

[65] Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A Survey of Evaluation Metrics Used for NLG Systems. *ACM Comput. Surv.* 55, 2, Article 26 (jan 2022), 39 pages. https://doi.org/10.1145/3485766

[66] Joni Salminen, Kathleen Guan, Soon-Gyo Jung, Shammur A. Chowdhury, and Bernard J. Jansen. 2020. A Literature Review of Quantitative Persona Creation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376502

[67] Joni Salminen, Kathleen Guan, Soon-Gyo Jung, and Bernard J. Jansen. 2021. A Survey of 15 Years of Data-Driven Persona Development. *International Journal of Human–Computer Interaction* 37, 18 (2021), 1685–1708. https://doi.org/10.1080/10447318.2021.1908670 arXiv:https://doi.org/10.1080/10447318.2021.1908670

[68] Joni Salminen, Kathleen Guan, Lene Nielsen, Soon-gyo Jung, and Bernard J. Jansen. 2020. A Template for Data-Driven Personas: Analyzing 31 Quantitatively Oriented Persona Profiles. In *Human Interface and the Management of Information. Designing Information*, Sakae Yamamoto and Hirohiko Mori (Eds.). Springer International Publishing, Cham, 125–144.

[69] Joni Salminen, Bernard Jansen, and Soon-Gyo Jung. 2022. Survey2Persona: Rendering Survey Responses as Personas. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization* (Barcelona, Spain) *(UMAP '22 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 67–73. https://doi.org/10.1145/3511047.3536403

[70] Joni Salminen, Joao M. Santos, Haewoon Kwak, Jisun An, Soon gyo Jung, and Bernard J. Jansen. 2020. Persona Perception Scale: Development and Exploratory Validation of an Instrument for Evaluating Individuals' Perceptions of Personas. *International Journal of Human-Computer Studies* 141 (2020), 102437. https://doi.org/10.1016/j.ijhcs.2020.102437

[71] Joni Salminen, Kathleen Wenyun Guan, Soon-Gyo Jung, and Bernard Jansen. 2022. Use Cases for Design Personas: A Systematic Review and New Frontiers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>New Orleans</city>, <state>LA</state>, <country>USA</country>, </conf-loc>) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 543, 21 pages. https://doi.org/10.1145/3491102.3517589

[72] Jeff Sauro and James R Lewis. 2016. *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann.

[73] Neil Savage. 2023. Synthetic data could be better than real data. *Nature article* (2023). https://doi.org/doi.org/10.1038/d41586-023-01445-8

[74] Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature* 623, 7987 (01 Nov 2023), 493–498. https://doi.org/10.1038/s41586-023-06647-8

[75] Gabriel Simmons. 2023. Moral Mimicry: Large Language Models Produce Moral Rationalizations Tailored to Political Identity. arXiv:2209.12106 [cs.CL]

[76] Dimitris Spiliotopoulos, Dionisis Margaris, and Costas Vassilakis. 2020. Data-Assisted Persona Construction Using Social Media Data. *Big Data and Cognitive Computing* 4, 3 (2020). https://doi.org/10.3390/bdcc4030021

[77] Hao Tan, Shenglan Peng, Jia-Xin Liu, Chun-Peng Zhu, and Fan Zhou. 2022. Generating Personas for Products on Social Media: A Mixed Method to Analyze Online Users. *International Journal of Human–Computer Interaction* 38, 13 (2022), 1255–1266. https://doi.org/10.1080/10447318.2021.1990520 arXiv:https://doi.org/10.1080/10447318.2021.1990520

[78] Petter Törnberg. 2023. ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning. arXiv:2304.06588 [cs.CL]

[79] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. arXiv:2302.11382 [cs.SE]

[80] Ning Wu, Ming Gong, Linjun Shou, Shining Liang, and Daxin Jiang. 2023. Large Language Models are Diverse Role-Players for Summarization Evaluation. arXiv:2303.15078 [cs.CL]

[81] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>New Orleans</city>, <state>LA</state>, <country>USA</country>, </conf-loc>) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 385, 22 pages. https://doi.org/10.1145/3491102.3517582

[82] Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2023. FLASK: Fine-grained Language Model Evaluation based on Alignment Skill Sets. arXiv:2307.10928 [cs.CL]

[83] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. arXiv:1904.09675 [cs.CL]

[84] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking Large Language Models for News Summarization. arXiv:2301.13848 [cs.CL]

[85] Xiang Zhang, Hans-Frederick Brown, and Anil Shankar. 2016. Data-Driven Personas: Constructing Archetypal Users with Clickstreams and User Telemetry. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5350–5359. https://doi.org/10.1145/2858036.2858523

[86] Zoie Zhao, Sophie Song, Bridget Duah, Jamie Macbeth, Scott Carter, Monica P Van, Nayeli Suseth Bravo, Matthew Klenk, Kate Sick, and Alexandre L. S. Filipowicz. 2023. More Human than Human: LLM-Generated Narratives Outperform Human-LLM Interleaved Narratives. In *Proceedings of the 15th Conference on Creativity and Cognition* (Virtual Event, USA) *(C&C '23)*. Association for Computing Machinery, New York, NY, USA, 368–370. https://doi.org/10.1145/3591196.3596612

[87] Haining Zhu, Hongjian Wang, and John M. Carroll. 2019. Creating Persona Skeletons from Imbalanced Datasets - A Case Study using U.S. Older Adults' Health Data. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (San Diego, CA, USA) *(DIS '19)*. Association for Computing Machinery, New York, NY, USA, 61–70. https://doi.org/10.1145/3322276.3322285

APPENDIX

A PERSONA-GENERATION WORKFLOW DESIGNS

In this section, we describe how user researchers and LLMs perform the persona-generation subtasks in each workflow.

A.1 LLM-auto

In LLM-AUTO, user researchers simply prompt LLMs. In a single prompt, user researchers (i) input all user data; (ii) describe the persona-generation task (*"Generate a minimum number of personas to represent the user data"*); and (iii) specify persona contents such as name, age, attitude, and motivation (Appdendix B shows the design decisions behind the prompt). From the prompt, the LLM identifies key characteristics that can constitute content and generates a few distinctive personas.

A.2 LLM-grouping

LLM-GROUPING requires user researchers to identify key characteristics and prompt the LLM to group user data and generate personas. For instance, if researchers have interviewed users to understand people's motivations for having children, they could define the key characteristics as 'motivations behind child plans'. Then, in parallel to LLM-AUTO, the researchers input all user data in the prompt and give the instructions *"Group user data by {characteristics}"* and *"Generate a single persona for each group"* to the LLM. From the prompt, the model groups user data and generates one persona for each group. In principle, personas should highlight user characteristics that the researchers want to focus on. Pre-specified key characteristics would result in personas aligned with their interests, but LLMs might not create user groups with enough differences. For instance, while LLMs may distinguish 'having children for joy' from 'having children for personal growth', user researchers might find such differentiation trivial.

A.3 LLM-summarizing

LLM-SUMMARIZING requires user researchers to group qualitative user data by pre-identified key characteristics. This is laborious and potentially erroneous when reliant solely on human judgement [17]. Therefore, we employed text-embedding-based clustering (i.e., semantic clustering) to support the humans [9]. Semantic clustering is a computational method for grouping text inputs by their similar meanings. It converts the materials into numerical vectors (i.e., text embeddings) and groups them via clustering algorithms. Our workflow used text-embedding-ada-002 [7] to transform sentences into text embeddings and a hierarchical clustering algorithm to group them. This enables user researchers to adjust the similarity threshold for grouping user data.

Accordingly, in LLM-SUMMARIZING, user researchers first retrieve the parts of user data that are relevant to their key characteristics. For instance, if their key characteristic is 'motivations behind child plan', they retrieve the specific parts from data where users talk about why they want or do not want to have children (i.e., a collection of survey responses to a question "why do you want or do not want to have children."). Then, they create a certain number of user groups by clustering those data by means of the text-embedding model and clustering algorithm. In the process, they can explore alternative numbers of user groups to find an optimal number of user groups with meaningful differences (e.g., Two groups of 'I want to have children and I do not want to have children' versus Three groups of 'I want to have children, I want to have children but not now, and I do not want to have children'). For each user group, user researchers input all user data and give the instruction *"Generate a single persona to represent the user data"* to the LLM. Under this approach, the LLM can be tasked with summarizing user data.

B DESIGNING THE PROMPT FOR PERSONA GENERATION

We designed our prompt based on our experiments with GPT-4 and guidelines of prompt engineering [79]. We first tested the simplest prompt with GPT-4 — presenting all qualitative user data to LLMs and asking them to generate personas (Table 3) — and found limitations of this naive approach. First, the naive approach produces highly granular user groups that distinguish even slight variance between survey responses. It mostly creates one persona for each survey response, which prevents identifying behavioral patterns. Second, the naive approach prioritizes trivial characteristics in user data. Instead of emphasizing users' attitudes or motivations, it tends to highlight their occupations or hobbies. Lastly, the naive approach tends to generate a mere summary or list of keywords rather than narrating personas. As a result, the outcome does not provide a holistic understanding of individuals. In summary, simply prompting LLMs to create personas would mostly fail, generating impersonal summaries of user data.

Accordingly, we iteratively tested prompts on GPT-4 and found three approaches that could make LLMs generate improved personas. First, we found that specifying the number of personas can control the granularity of user groups. For instance, prompts such as 'Generate three personas' or 'Generate a minimum number of personas' made GPT-4 converge similar responses and generated personas that are distinctive from each other. Second, we found that specifying the template of personas can 'hint' GPT-4 about the important themes. Asking GPT-4 to include information such as background, attitude, or motivation made GPT-4 highlight relative information from user data. Lastly, we found that specifying the grammatical person can make GPT-4 generate comprehensive narrations instead of mere summaries or a list of keywords. Based on our findings, we designed a prompt that could provide clear instructions and rules for generating personas (Table 4). Our prompt had three main components:

[7] https://openai.com/blog/new-and-improved-embedding-model

- *Input user data* comes from material such as interview transcripts and survey responses. Each bracket contains a single user's data so that the LLM identifies them as such.
- *Persona-generation tasks* are the instructions for how the LLM should group user data and generate personas accordingly. They are modified to match the LLM's task in each workflow.
- *Persona-generation rules* are what the LLM should commonly follow in all workflows to produce high-quality personas. The rules are designed to prevent generating non-factual information and to instruct the LLM to include specific contents in personas, such as name, background, and motivation.

Prompt	Here are survey responses: *{10 survey responses}*. Generate personas from the survey responses.	
Outcome	1. The Cultured Explorer - Age: 38 - Gender: Male - Profession: Marketer - Relationship Status: Single - Hobbies: Attending cultural events - Values: Appreciating diversity, broadening perspectives, personal growth - Current Lifestyle Satisfaction: Mostly unsatisfied - Desire for Children: Wants children for a new level of joy and fulfillment ...	... 10. The Branding Buff - Age: 28 - Gender: Female - Profession: Marketing Coordinator - Relationship Status: Married - Hobbies: Attending workshops, reading marketing books, traveling - Values: Personal and professional growth, meaningful connections, balance - Current Lifestyle Satisfaction: Mostly unsatisfied - Desire for Children: Does not want children in order to focus on career and personal interests

Table 3: An example of simply prompting GPT-4 to generate personas from ten survey responses (an example response is shown in Figure 15). This resulted in generating one persona per response, which does not identify or represent archetypal user groups.

Prompt
Here are user data: [[Data from user 1], [Data from user 2], [Data from user 3], ... [Data from user *N*]]
{instructions for grouping user data}
{instructions for generating personas from the user data}
You must follow the rules below when generating the personas:
- Rule 1: Do not add any information that does not exist in the user data.
- Rule 2: You may combine, synthesize, or rephrase multiple user data into a single persona.
- Rule 3: The persona should have detailed descriptions of the following information: {a list of contents}
- Rule 4: Write {a list of contents} from the first person perspective.
After generating personas, compare the personas with the user data to validate Rule 1, 2, 3, and 4.
Make necessary updates such as updating information in personas, removing personas, or creating new personas.
Present only the final personas.

Table 4: A prompt commonly used in our persona-generation workflows. It comprises three main components: input user data (yellow-highlighted), persona-generation tasks (green-highlighted), and persona-generation rules (blue-highlighted). We adjusted the instructions for persona-generation tasks workflow-specifically and the list of contents in accordance with the user data.

C IMPROVING PROMPTS WITH HUMAN EXPERTS' PREFERENCES

Study 2 revealed the three points for improving the quality of personas. Accordingly, we added five rules in the original prompt by following the output customization patterns from prompt-engineering guidelines [79]. As shown below, Rule 1 instructs the LLM to use more expressive words when describing personas, while Rule 2 and 3 instruct it to recall the information from earlier contents (e.g., background) to describe later contents (e.g., motivation). Rule 4 and 5 indicate to emphasize the characteristics most commonly expressed in the user data.

- Rule 1: In 'Background', 'Past experience of AI', 'Attitude toward AI', and 'Motivation for taking AI courses', the persona MUST describe its emotion (e.g., It makes me feel... because...) and preference (e.g., I prefer... because...).

- Rule 2: In 'Attitude toward AI', the persona MUST describe WHY it has such attitudes considering 'Background' and 'Past experience of AI'.
- Rule 3: In 'Motivation for taking AI courses', the persona MUST describe WHY it has such motivations considering 'Background', 'Past experience of AI', and 'Attitude toward AI'.
- Rule 4: Exaggerate the survey respondents' most common characteristics.
- Rule 5: For 'Gender', 'Major', and 'Academic year', pick the most common ones from the survey responses.

D SYNTHETIC USER DATA

1. Age: 30 **2. Gender:** Male **3. Occupation:** Teacher **4. Marital status:** Single

5. What do you do during your free time?
I love spending my free time practicing yoga, reading books, and attending cultural events.

6. What do you value in your life?
I value overall wellness, knowledge, and experiencing new things in my life. Practicing yoga helps me maintain physical and mental well-being, while reading books and attending cultural events allows me to expand my understanding of the world and different perspectives. These things bring depth and meaning to my life, and help me feel fulfilled.

7. How much are you satisfied with current lifestyle?
Mostly unsatisfied with current lifestyle.

8. What is your child plan?
I want to have children.

9. Why do you want or do not want to have children?
I want to have children because I feel like it will add a new depth and dimension to my life. As much as I enjoy practicing yoga, reading books, and attending cultural events, I feel like having kids will bring a level of joy and fulfillment that I may not even understand yet. I value overall wellness and experiencing new things, and I believe that raising children will allow me to do both of those things in a new and exciting way. While I may not be satisfied with my current lifestyle at the moment, I believe that having children will bring a new sense of purpose and joy to my life.

10. What kinds of civic services would be helpful to you if you are a parent?
As a parent, I would greatly benefit from civic services that prioritize education and cultural enrichment. Accessible and affordable after-school programs, community events, and youth mentorship programs would help ensure that my children have opportunities to learn and grow beyond the classroom. Additionally, quality and affordable healthcare, including maternal and pediatric care, would provide peace of mind as a parent. Lastly, safe and well-maintained parks and playgrounds would allow my family to enjoy the outdoors and stay active together.

Figure 15: An example of synthetic survey responses generated with GPT-3.5-turbo. The survey collects information about users' demographics (1-4), backgrounds (5-7), and design-related comments (8-10).

E REAL USER DATA

1. Age: 27 **2. Gender:** Female **3. Major:** Music **4. Academic year:** 3rd year undergraduate

5. What do you do in your free time?
I listen to music, watch TV shows, socialize and do some writing.

6. Why did you choose your major?
I really enjoy music and wanted the opportunity to explore it further in studies, learning more and expanding my horizon and scope.

7. What are the most important aspects in your creative activities?
It is important to have fun, if it's not fun and enriching, what is the point?

8. How much experience do you have with AI?
Not a lot of experience, I have used ChatGPT a few times but I don't have a lot of experience.

9. What do you think about AI in general?
I am not sure how I feel about AI. On the one hand, it seems useful and convenient. It can write a resume for you, a song, a poem, book recommendations. And on the other hand, there are replacing jobs and there is the risk of identity theft and other nefarious outcomes.

10. What do you want AI to do for your creative activities?
I'd want AI to help with coming up with chord progressions or parameters to write a new song in. Or giving me ideas or music recommendations. I'd want AI to help make the lengthy decisions easier. I am very indecisive and I like rules and routine to help. AI would help with those things and I can be free to create without anything holding me back. It's like being given a framework or guide to work from without having to think of it yourself.

11. What would you like to achieve by taking AI courses?
I'd want to learn how AI works and what makes it possible. I'd like to learn how to use AI to work to my advantage. It would help inform my decisions and choices with AI and allow me to use it to its full capability so that I can focus on doing what I like and can do best.

12. What kinds of course activities would you like to do?
Write a song with AI, Business strategy with AI, Using AI to be more creative. It would be different and switch things up. It would supplement the teachers and lecturers.

13. What do you want the instructors to care about when they teach AI courses for you?
To care about how it helps us in the future. We all want to thrive in our fields so if there is something that can help, that would be very beneficial.

Figure 16: An example of survey responses from real people. The survey collects information about users' demographics (1-4), backgrounds (5-8), and design-related comments (9-13).

F PERSONAS CREATED FROM SYNTHETIC USER DATA

LLM-auto

Name: Andrew **Age:** 38 **Gender:** Male **Occupation:** Marketer **Marital status:** Single

Background:
I am a marketer who is passionate about attending cultural events. Experiencing different art forms, music genres, and literature works broadens my perspective and connects me with like-minded, creative individuals. Although I am mostly unsatisfied with my current lifestyle, I am eager to start a family someday.

Personality:
I am a culturally-driven, intellectually curious individual who appreciates diversity. I am passionate about the arts and enjoy surrounding myself with creative and like-minded people.

Plans for having children:
I want to have children someday because I believe it will bring new levels of joy and fulfillment to my life. I feel that having children will allow me to pass down my values and beliefs and provide me with a sense of purpose and legacy.

Motivation for using civic services:
I would appreciate access to quality art programs and cultural events for children, as well as affordable and reliable childcare services for working parents like myself.

LLM-grouping

Name: Elizabeth **Age:** 34 **Gender:** Female **Occupation:** Teacher **Marital status:** Single

Background:
I lead an active lifestyle, and I find joy in attending cultural events and exploring different art forms. I love to learn and grow through these experiences, which enriches my life and contributes to my personal growth.

Personality:
I am an open-minded and creative individual who values personal growth and the beauty in diversity. I also enjoy connecting with like-minded individuals who share a similar interest in the arts.

Plans for having children:
I want to have children because I feel like it will add a new depth and dimension to my life. I believe that having children will provide me with a sense of purpose and legacy that is important to me. While I may be mostly unsatisfied with my current lifestyle, having children will bring me the happiness and sense of accomplishment that I am looking for.

Motivation for using civic services:
As a future parent, I would greatly appreciate civic services that prioritize education and cultural enrichment for my children. Attending workshops, events, and mentorship programs would help them learn and grow beyond the classroom. Quality and affordable healthcare, as well as safe and well-maintained parks and playgrounds, would also benefit my family.

LLM-summarizing

Name: Alex **Age:** 36 **Gender:** Male **Occupation:** Creative Professional **Marital status:** Single

Background:
As a creative professional, I enjoy attending cultural events and exploring new ideas through reading and traveling. I have a strong passion for personal growth, learning, and staying current in my chosen field. I have also prioritized maintaining my physical and mental well-being by engaging in activities like yoga or going to the gym.

Personality:
I consider myself to be a well-rounded individual with a thirst for knowledge, adventure, and self-improvement. I value creativity and meaningful experiences in my life. I prioritize relationships with loved ones as they bring happiness and fulfillment to my life.

Plans for having children:
I want to have children because I believe that it will bring a new dimension and purpose to my life. While I enjoy the personal growth and experiences I gain from my chosen field and other activities, I feel that having children will bring a level of joy and fulfillment that cannot be replicated by any other experience. As a parent, I am excited to share my values and guiding principles with my future children and make a positive impact on their lives.

Motivation for using civic services:
As a future parent, I would greatly appreciate civic services that support the well-being and success of my future children. Accessible and affordable childcare options, educational and cultural programs for children, reliable healthcare services, and safe play areas in public spaces would greatly contribute to my family's quality of life. By investing in these civic services, I believe it would foster an enriching and supportive environment for both my children and the broader community.`

Figure 17: Examples of personas created from LLM-auto (top), LLM-grouping (middle), and LLM-summarizing (bottom).

G PERSONAS CREATED FROM REAL USER DATA

Designer-generated

Name: Tom **Gender:** Male **Age:** 26 **Major:** Film and TV prod **Academic year:** Final year MA

Background:
I have been into movies since I can remember. I took this major as I really want to get my career started in this field. In the past I took a couple of internships at a production agency and I really see myself working in such an environment.

Past experience of AI:
I remember using some AI tools before chat GPT was launched. I was impressed and also a bit worried about the future of artistic production. I used chat GPT many times, both for school and other things. I took screenplay classes, where it has been particularly useful for generating a decent storyline.

Attitude toward AI:
I am soon graduating, and I wonder how my degree and field is going to be affected by AI. I am not too worried that it will "steal" jobs, but I do see that it might impact things like screenplay, animation and editing of movies and TV. On the one hand this would save people a lot of time, but if it keeps evolving at this speed, I'm quite concerned that it will really end someone's career.

Motivation for taking AI courses:
The entertainment industry is really amazing. I wish to learn some more about AI tools and their impact on TV production. It would be nice to learn not just how to use them so that our work gets easier, but also how to better grasp the fact that AI needs to keep having human-generated inputs, and that we as people, do appreciate human-generated contents actually better than AI-generated ones.

LLM-grouping

Name: Alex **Gender:** Man **Age:** 25 **Major:** Fine Art **Academic year:** 3rd year undergraduate

Background:
I've been passionate about painting and creating art since I was young. In my free time, I love to immerse myself in different forms of expression, such as drawing and painting.

Past experience of AI:
I've had minimal experience with AI tools, mostly using them for fun rather than serious applications. I've experimented a bit with ChatGPT and some AI art generation software.

Attitude toward A:
I tend to view AI with a sense of apprehension and unease. While I acknowledge the potential benefits it could bring to the creative field, I worry about the risks of AI taking over the human touch in art and the ethical implications surrounding its use.

Motivation for taking AI courses:
Despite my reservations, I am curious about learning how AI works and what limitations it has. Gaining a better understanding of AI's capabilities and boundaries could help me make more informed decisions on when and how to incorporate it into my workflow, while still maintaining the human element and integrity of my art.

LLM-summarizing

Name: Alex **Gender:** Woman **Age:** 23 **Major:** Fine Arts and Design **Academic year:** 4th year undergraduate

Background:
I enjoy various creative activities including watching movies, playing games, and listening to music. I was drawn to my major because I love exploring different creative areas and wanted to study something that melds both practical and theoretical aspects of art. In my creative work, I value originality, self-expression, and pushing boundaries.

Past experience of AI:
My experiences with AI have mainly been for fun or limited use. I have dabbled in chatbots like GPT-3 and have seen friends using AI-generated art, but I haven't truly integrated AI into my own creative practice.

Attitude toward A:
I have mixed feelings about AI. On one hand, I recognize that it can be a helpful tool for some tasks, but on the other hand, I'm a bit wary and even scared of its potential implications. I don't like the idea of AI replacing human input in the creative process and believe that it may lead to a loss of authenticity in art.

Motivation for taking AI courses:
I'd like to learn more about AI, how it works, and how it can be used in a productive and ethical manner within the creative field. I'm interested in understanding its limitations and exploring its potential to enrich the creative process by aiding with tasks like research, idea generation, and technical problem-solving. Understanding AI better would help me feel more secure in my field and give me the tools to harness its potential advantages.

Figure 18: Examples of personas created by designers (top), LLM-GROUPING (middle), and LLM-SUMMARIZING (bottomw) from the same user data.

H COMPUTING SEMANTIC SIMILARITY SCORES

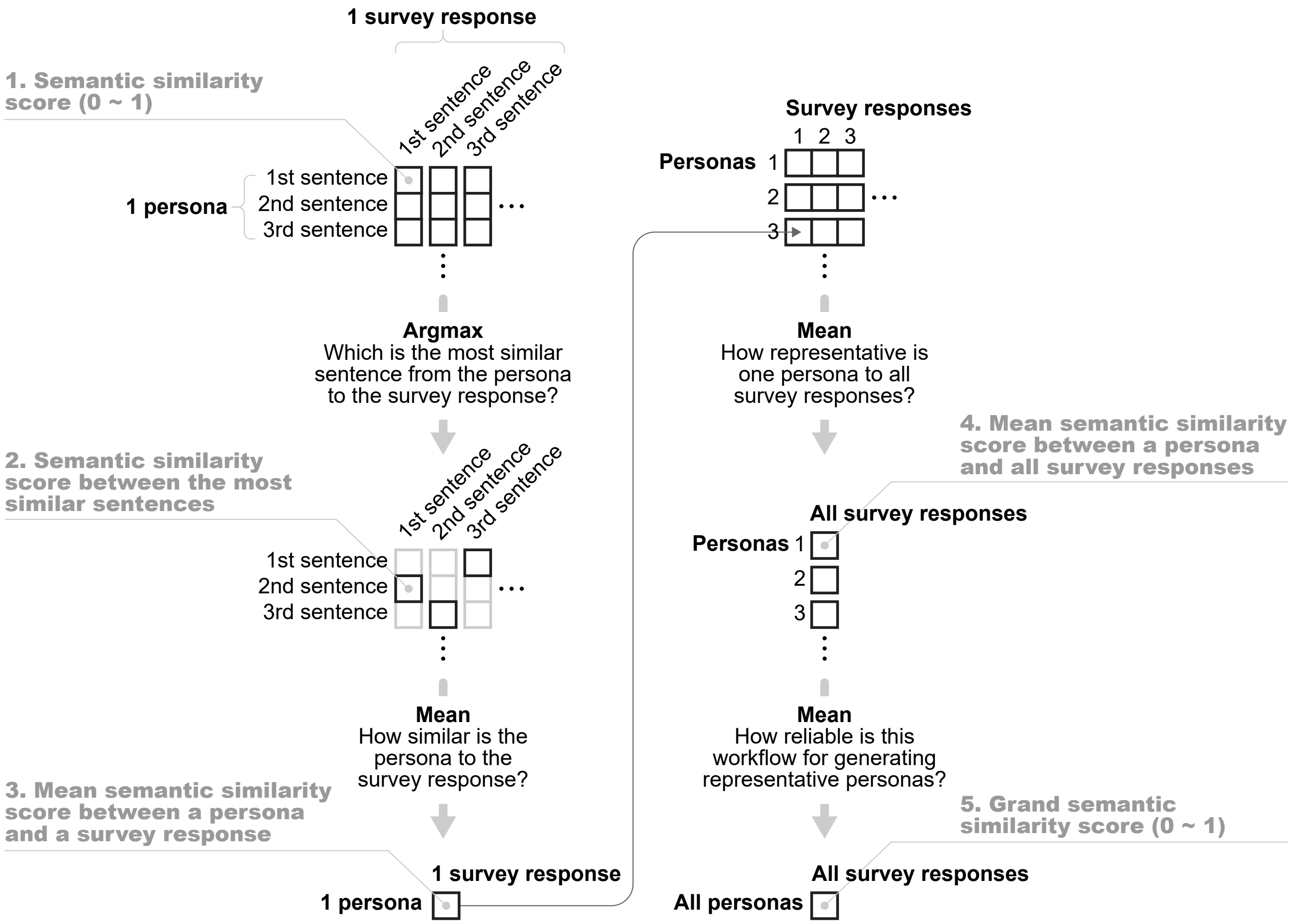

Figure 19: We computed a single grand semantic similarity score for each persona-generation workflow based on the semantic similarity scores between a set of personas and user data.