Active Personas for Synthetic User Feedback: A Design Science Study

Mario Simaremare and Henry Edison

Blekinge Institute of Technology, Karlskrona, Sweden
mario.simaremare@bth.se, henry.edison@bth.se

Abstract. *Context*: Securing consistent user feedback is a critical yet resource-intensive challenge in new product development (NPD). Active Personas (APs), user archetypes powered by Generative AI, offer a novel approach to generate realistic user feedback on demand, enabling internal product experimentation. *Objective*: This study evaluates the effectiveness of AP in generating "user" feedback by evaluating the alignment between AP-generated feedback and actual (human) users and investigates how persona and Large Language Model (LLM) choice influence the feedback. *Method*: Adopting a Design Science Research methodology, we designed and developed AP as a novel artifact. We demonstrated and evaluated the artifact in a case of a mobile transport app, creating eight AP instances from two personas and four LLMs. The AP-generated feedback was compared against human feedback from interviews and Google Play reviews using a triangulated analysis based on Nielsen's usability heuristics. *Results*: Our findings reveal a strong alignment between AP and human feedback, with APs effectively identifying various usability and accessibility issues, confirming what the actual users also found. The persona definition significantly dictated the evaluation's criticality, and the choice of LLM further influenced the evaluative stance. *Conclusion*: APs have the potential to augment early-stage usability evaluations and are a viable option for human users for internal experimentation, providing rapid, low-cost feedback. While they supplement, not replace, direct human interaction, this work validates the transformation of personas from static artifacts into dynamic, generative agents for NPD.

Key words: active personas, user personas, generative ai, user feedback, experimentation, new product development

1 Introduction

User feedback plays a critical role in new product development (NPD) because end users offer invaluable expertise about their lived experiences [17]. In software engineering (SE), development teams integrate user feedback into every phase of the NPD process, from idea validation and requirements engineering to design, testing, and support activities [11, 30]. User feedback enables teams to identify real-world requirements, constraints, clarify assumptions, and make strategic design decisions [27, 30]. To structure user feedback, teams generate

requirements engineering artifacts such as user scenarios, user stories, concept mind maps, and user personas [25]. Among these artifacts, personas vividly capture user characteristics, foster team empathy, and help the team focus design efforts on authentic user profiles [15, 29]. Personas also guide participant selection in usability tests, clarify assumptions, and inspire interface and experience design decisions [8, 16]. Despite its importance, securing consistent user engagement and substantive feedback at NPD stages remains challenging [16]. Relying on a small, self-selecting group restricts the diversity of viewpoints and risks overlooking critical requirements.

The persona technique has been promoted as a strong tool to understand the needs of potential users. However, this technique is not widely used in the software development industry [3]. Recently, with the emergence of Large Language Models (LLMs) and Generative AI (GenAI), an advanced AI technology capable of creating novel human-generated content [10], several attempts have been made to include personalities into conversational agents in various domains, such as mental health [1] and games [2]. Moreover, several studies have been conducted to develop personas using LLMs [12, 20]. A study by Salewski et al. [28] showed that GenAI provides better feedback when it pretends to be a domain expert or persona while performing simple tasks.

To address this gap, Simaremare et al. [32] proposed the concept of Active Personas (APs), dynamic persona instances that generate realistic user feedback. Unlike traditional personas, which are static design documents [8], APs transform these descriptions into dynamic agents that can generate feedback on demand, shifting from passive design references to active evaluation participants.

APs leverage techniques like retrieval-augmented generation (RAG) and fine-tuning [14] to feed persona definitions into multi-modal LLMs. APs can simulate user feedback on demand, empowering teams to rapidly *build, measure, and learn* (BML) through internal experimentation without the logistical overhead of recruiting actual users [27]. This capability would make APs particularly well-suited for internal usability evaluations [4].

This study is built upon the AP concept [32] by answering the following research question: *How effective is Active Personas in generating "user" feedback?* We define effectiveness as the alignment between AP-generated feedback and human user feedback, measured using Nielsen's heuristics. We used Skånetrafiken[1], a transportation-related application running on the Android platform, as a case. We used four multimodal LLMs to power our APs.

This study makes contributions to research and practice. For research, our study provides an empirical evaluation of APs as a practical method for accelerating internal BML cycle. For practice, by transforming personas from static documents into dynamic, generative agents, this research offers a novel paradigm for integrating simulated user feedback into the NPD lifecycle.

[1] `https://play.google.com/store/apps/details?id=se.skanetrafiken.washington`

The remainder of this paper is organized as follows. Section 2 elaborates on the study's research method, followed by the results in Section 3, and discussion in Section 4. The conclusion and future work are laid out in Section 5.

2 Research Methodology

We applied the six-phase process model as described in the Design Science Research (DSR) [26] methodology to answer our research question. DSR is a problem-solving paradigm focused on creating and evaluating innovative artifacts to solve real-world problems while contributing new knowledge.

Phase 1 & 2: Problem Identification and Solution Objectives

The problem motivating this research, as discussed earlier in Section 1, is the significant challenge development teams face in securing consistent and diverse user feedback throughout the NPD process. This issue is particularly acute for resource-constrained teams, such as software startups [25]. The primary objective was to design a tool capable of generating realistic, contextualized, and actionable user feedback on demand, thereby reducing the reliance on direct, continuous engagement with actual human users. The solution should be efficient, scalable, and capable of representing diverse user perspectives, such as those with specific needs.

Phase 3: Artifact Design and Development

Persona Definition: We adopted and adapted two personas for this study. The first, "Ingrid Henke," a data-driven persona, was adopted from a prior work by Mueller et al. [22] and adapted to emphasize the accessibility needs of a user with a vision impairment. We used the UK Government's guidelines for personas with access needs[2]. The second persona, "Claudio Eriksson," represents an average user: a young adult with moderate tech skills without accessibility needs.

Model Configuration: We developed the APs using four state-of-the-art multimodal LLMs: GPT-5, Claude 4, Gemini 2.5 Pro, and Llama 4 routed through OpenRouter[3] APIs. We used a temperature of 1.0 for each model to balance between creativity and accuracy.

Active Persona Creation: We created eight AP instances, each is a unique combination of a persona (Ingrid or Claudio) and an LLM. We instantiated each persona within its respective model by using a system prompt that applied a speaker-specific role approach [35]. In the prompt, we directed the LLM to scope its knowledge and reasoning proportionally to the persona and to respond in a speaking-like tone.

[2] `https://alphagov.github.io/accessibility-personas/`
[3] https://openrouter.ai/

Phase 4 & 5: Demonstration and Evaluation

We evaluated the effectiveness of AP in generating contextual feedback by investigating its alignment with feedback collected from human users in a specific case using a consistent evaluation framework.

Case: We selected the Skånetrafiken mobile app on the Android platform as our case. Skånetrafiken is the public transport authority for Skåne County, Sweden, managing a comprehensive network of buses, trains, and other services[4]. The app, with over 1 million downloads and a 3.1 overall rating from more than 4,000 reviews, provides functionalities for journey planning, real-time travel information, and ticket purchasing. This case was chosen due to the app's widespread use and the availability of a large body of existing user feedback. The company is committed to being fully compliant with Swedish law on the accessibility of digital public services.

Usability Evaluation: Usability refers to the extent to which a system can be used by specified users to achieve goals with effectiveness, efficiency, and satisfaction [13]. Usability can be decomposed into five dimensions: learnability, efficiency, memorability, error prevention, and satisfaction [23].

A widely used method for evaluating usability is Nielsen's heuristics evaluation framework, where a small group of usability experts assesses a user interface against a set of ten established criteria (i.e., heuristics) [23]. These heuristics include: visibility of system status (N01), match between the system and the real world (N02), user control and freedom (N03), consistency and standards (N04), error prevention (N05), recognition rather than recall (N06), flexibility and efficiency of use (N07), aesthetic and minimalist design (N08), help users recognize, diagnose, and recover from errors (N09), and help and documentation (N10). Nielsen's heuristics is a fast, inexpensive, and practical alternative to resource-intensive usability testing, with the goal of identifying and addressing usability issues early in the design process.

While Nielsen's heuristics evaluation was originally intended for usability experts, an empirical study shows that end users are capable of understanding and applying these heuristics [18], suggesting that early user involvement in internal evaluations can be highly effective. Studies by Branco et al. [5] and Matos et al. [21] are two examples of end-user involvement in usability evaluation.

The use of APs aligns with Nielsen's heuristics idea of internal experimentation, offering a method for getting simulated user feedback without needing actual user participants (i.e., external experimentation [32]). Therefore, our work investigates the viability of APs to augment or even replace human evaluators in a heuristic evaluation setting, combining the efficiency of the heuristic method with the realism of persona-driven synthetic feedback generation.

Evaluation Design: Our evaluation was based on Nielsen's heuristics for two reasons: first, it accommodates both usability experts and users as the evaluators [18], allowing us to integrate both human and AP feedback. Second, the heuristics are well-suited for evaluating static artifacts, such as the screenshots we used.

[4] `https://www.skanetrafiken.se/english/`

To achieve a comprehensive and triangulated evaluation [19], we consolidated data from three sources: (human) user interviews, Google Play review analysis, and AP feedback generation. Fig. 1 depicts an overview of the evaluation phase.

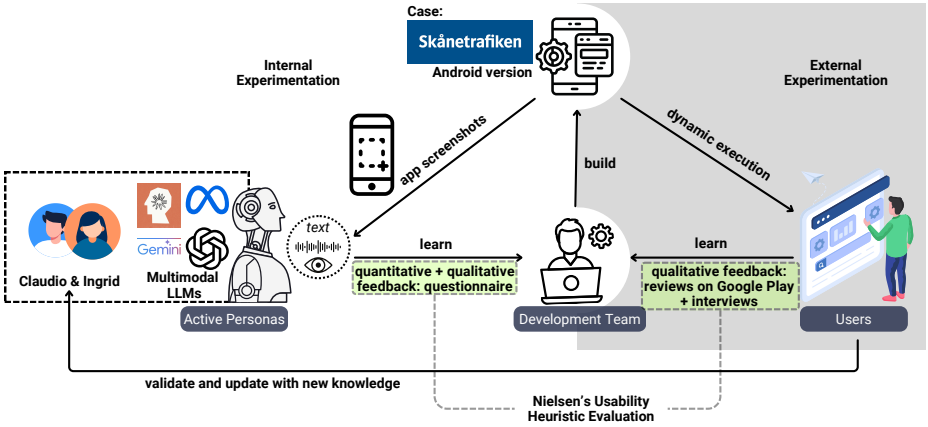

Fig. 1. Evaluation: an overview

Semi-structured Interviews: We applied online interviews [19] to gather qualitative data for in-depth evaluation. We applied a semi-structured approach, allowing us to explore the unexpected discussion [31]. We conducted four online interview sessions in August 2025, each lasting between 30 and 45 minutes and was recorded and transcribed. We transformed Nielsen's heuristic evaluations workbook[5] into 20 interview questions (Q01-20). Each of the heuristics was translated into two questions to capture different aspects of each heuristic. For example, the first two questions (Q01 and Q02) were aimed at collecting data related to the first Nielsen's heuristic, visibility of system status (N01).

Participants: We used convenience sampling to recruit four participants from our contacts who frequently used the app at least once a week (see Table 1). Participants were systematically selected to align with our predefined personas: P1 (vision impairment) matched Ingrid's accessibility needs; P2 and P3 (software professionals) aligned with Claudio's tech-savvy profile; and P4 represented another variant of Claudio. This purposive sampling ensured meaningful comparison between human and AP feedback.

Table 1. Interview participants

ID	Age	Background	Experience	App language
P1	30-35	Master student in social science	2 years	Swedish
P2	30-35	Professional in a software company	5 years	English
P3	25-30	Master student in software engineering	1.5 years	English
P4	30-35	Professional in logistics	2 years	English

[5] https://www.nngroup.com/articles/ten-usability-heuristics/

Synthesis: We employed thematic analysis [6, 9] with a deductive approach to identify patterns and insights from the interview data, using Nielsen's heuristics as our initial coding.

Google Play Review Analysis: We scraped over 500 textual reviews from the app's Google Play on July 30, 2025. To ensure relevance, we filtered the reviews to include only reviews posted from 2024 onwards, resulting in 46 reviews. Non-English reviews were translated using Google Translate[6] API to ensure consistent analysis. Each review was given a unique identifier (e.g., f001) for traceability. We then applied thematic analysis to map the review to Nielsen's heuristics, mirroring the synthesis approach used for the interview data.

Active Persona Feedback Generation: To generate both quantitative and qualitative synthetic feedback, we provided each of the eight AP instances with a questionnaire derived from our earlier interview protocol. This ensured direct comparability between the human and AP-generated feedback. The questionnaire consisted of 20 statements, which the AP instances evaluated on a 5-level Likert scale (1: *"strongly disagree"*, 5: *"strongly agree"*). For each statement, the AP was instructed to provide a brief rationale to support its choice.

The AP feedback was generated through a zero-shot prompting approach [7]. This method leverages the pre-trained knowledge of the LLMs without requiring any in-context examples. The prompt contains the questionnaire accompanied by nine screenshots of the app's user journey (see Table 2), and a request for a structured JSON-formatted response. The screenshots, two of which were in Swedish to evaluate language appropriateness and consistency, covered key user flows from searching for a trip to purchasing a ticket. For each AP instance, we ran ten independent evaluations, generating a total of 80 results (for eight AP instances). Each evaluation used identical inputs: the same screenshots, questionnaire, and system prompts. Nothing was changed between runs to get reliable responses. Samples of the app's screenshots are shown in Fig. 2.

Table 2. Description of the screenshots used in this study

No	Screenshot content	Language
1	the app's homepage, accessible through the "Search journey" tab.	English
2	the search journey screen	English
3	the journey list screen	English
4	the journey filter screen	English
5	the journey detail screen	English
6	the ticket selection screen	English
7	the ticket selection screen	Swedish
8	the ticket detail screen	English
9	the ticket detail screen	Swedish

Execution costs, in terms of input tokens, output tokens, and execution time, varied across persona-model combinations. Llama-based personas were the

[6] https://translate.google.com/about/

fastest (< 30s) but consumed the most input tokens ($\approx 13,000$) while generating the fewest output tokens ($\approx 1,400$). Conversely, Gemini-based personas were the most token-efficient, making efficient use of input tokens ($\approx 3,700$) while consuming a large number of output tokens ($\approx 6,000$) at a moderate time cost.

Synthesis: We mapped the qualitative feedback from the questionnaire to Nielsen's heuristic, consistent with our approach for the other two earlier sources. We then performed a comparative analysis of the feedback from all three sources.

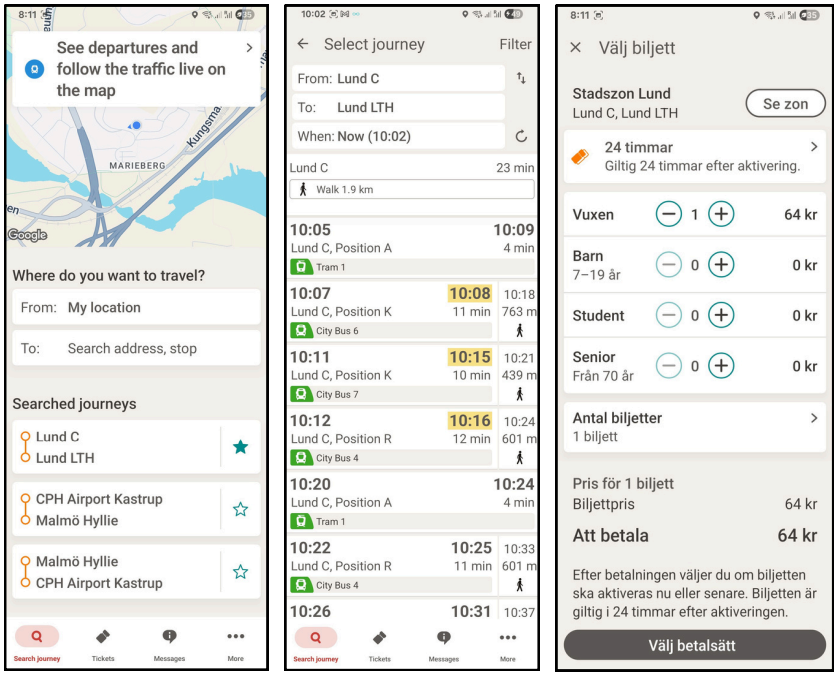

Fig. 2. Screenshots of the Skånetrafiken app

Moreover, we used descriptive statistics [34] of the ratings for each statement in the questionnaire to evaluate how the persona-model combination affects the feedback. We calculated the mean of ratings [24] given by AP instances per persona-model. We further investigated how these combinations affect the ratings per statement in the questionnaire (i.e., Q01-20). Following this, we conducted a Kruskal-Wallis test [34] to determine if the observed differences among the eight AP configurations were statistically significant. This non-parametric test is well-suited for our analysis because it is designed to compare more than two independent groups using ordinal data, such as the Likert scale ratings from our questionnaire, and does not require the data to follow a normal distribution.

Appendix: Detailed information on the used user personas (Ingrid and Claudio), the interview and questionnaire protocols, the system and evaluation prompts, the complete app's screenshots, the cost evaluation, the complete sta-

tistical analysis, reviews from Google Play, AP-generated feedback, and source code we used in this study can be found in [33].

3 Results

The effectiveness of AP in generating user feedback is measured through the alignment between AP-generated feedback and human user feedback by mapping them to Nielsen's heuristics. We measured the alignment through: (1) thematic mapping of both human and AP feedback, (2) comparing specific usability issues identified by each source, (3) analyzing rating patterns, and (4) triangulating findings across interviews, Google Play reviews, and AP feedback. We then investigate the influence of the persona-model combination on generated feedback using statistical analysis.

3.1 Alignment Between APs and Human User Feedback

Visibility of system status (N01): The app generally communicates its system status effectively. Interview participants and AP instances noted clear information like departure times and track numbers. For example, an instance of Claudio shared, *"I can clearly see 'From: Lund C' and 'To: Lund LTH' with the current time. The journey detail screen ... shows the complete route with clear departure and arrival times, and the ticket selection screens ... clearly display the zone and journey (Q02, nielsen_4, claudio_claude)."* However, an instance of Ingrid highlighted an accessibility issue, stating, *"The journey page shows Changes and Travel time plus 'Today Monday' (s5). For my eyes, some small, pale text makes it a bit harder to read at a glance (Q02, nielsen_71, ingrid_openai)."* Google Play reviews offered a real-world counterpoint, with one user (f16) noting, *"Not showing correct departure time. When trains are delayed, the information is shown too late, and you have no chance to take an earlier train."* This feedback highlighted a potential gap between the app's static interface and the dynamic needs of a user in real-time.

Match between the system and the real world (N02): All interview participants found that the app's terminology reflected real-world terms. For instance, P1 stated, *"the language is consistent with its website, which is in Swedish."* Comments from AP instances aligned with the interview participants'. An instance of Claudio shared, *"It uses all the normal words I'd use when talking about taking the bus or train ... and the Swedish versions ... are exactly what you'd expect. There's no confusing jargon, which makes it quick to use (Q3, nielsen_17, claudio_gemini)."* An instance of Ingrid added a nuance related to icons, stating that while they are generally recognizable, *"some icons might not be immediately clear to all users, especially those with visual impairments (Q04, nielsen_65, ingrid_llama)."*

User control and freedom (N03): All interview participants reported sufficient freedom to navigate the app, though some preferred Android's native

sliding feature over the in-app back button, as noted by P2: *"I typically navigate by sliding. I rarely use the back button."* AP instances provided similar feedback, recognizing the app's consistent in-app back button but suggesting that some screens still require a more explicit exit button. For example, an instance of Claudio noted, *"Every screen has a consistent back arrow in the top left ... and the ticket selection screens have a clear 'X' button for closing ... This consistency makes it easy for me to know exactly how to exit any screen without getting lost ... (Q05, nielsen_ 7, claudio_ claude)."* Similarly, an instance of Ingrid pointed out, *"the back buttons are always in the top-left corner where you'd expect them. The main menu is always at the bottom, which is good. The button for 'Filter' on screen 3 is just a small symbol and could be a bit bigger for my eyes, but it's in a normal spot, so I would find it (Q06, nielsen_ 58, ingrid_ gemini)."* A Google Play review from a user (f030) further emphasized a navigation problem, stating, *"There is no bottom navigation bar or hamburger menu ..."*

Consistency and standards (N04): The app maintains its consistency in styles, typography, and layouts across its screens, a finding corroborated by both human and AP feedback. The app also applies layouts similar to other transportation apps. All interview participants agreed on this point. For instance, P3 noted: *"When I look at it from the color perspective, I can really see that there is consistency, for example, in the background and foreground color. I get the same colors everywhere."* The AP instances provided similar observations. An instance of Claudio shared, *"Typography, pill buttons, and card lists are consistent from journey search to tickets. It all feels part of one system (Q07, nielsen_ 32, claudio_ openai)."* An instance of Ingrid added, *"Standard patterns: bottom nav, back arrow, toggles, plus/minus steppers, and card lists ... It behaves like other Swedish travel apps I use (Q08, nielsen_ 79, ingrid_ openai)."*

Error prevention (N05): Interview participants agreed that the app helps them avoid errors through clear layouts and confirmation dialogs. For example, P3 highlighted a helpful modal that prevents accidental ticket activation. However, participants also identified potential issues. P3 found icons confusing and noted a problem with button color: *The app has this red color here (a button) as the highlighted or selected color. But then, when you click on this, maybe you would expect that little bit of red tint to be highlighted on this as well, because this is the highlighted color to show you an active color."* P1 found it difficult to read train names due to small text size. AP instances reported similar feedback on the potential cause of errors. An instance of Claudio warned of a possible tap error: *"On the journey detail, Monitor sits next to Select ticket and has similar weight, so I could tap the wrong one when rushing (Q09, nielsen_ 39, claudio_ openai)."* An instance of Ingrid's feedback aligned with the button color issue, stating, *"The 'Select ticket' button is a small grey oval that almost blends in with the background ... these grey buttons are not very prominent and I might have to search for them (Q09, nielsen_ 54, ingrid_ gemini)."*

Recognition rather than recall (N06): The app effectively leverages recognition over recall through a "recent search" feature, which was identified by most interview participants. However, P4's comment that she *"never knew*

that feature existed" highlights a potential discoverability issue. The AP instances also recognized the feature's value. An instance of Ingrid noted how helpful it is for users with routines, stating, *"On the very first screen, it shows my 'Searched journeys'. I take the bus to the market every Tuesday and Saturday, the same route. Having it right there without typing is perfect. And I see a little star, so I can save it (Q11, nielsen_ 60, ingrid_ gemini)."*

Flexibility and efficiency of use (N07): Both interview participants and AP instances identified features that enhance efficiency, though they did not highlight the same ones. All interview participants identified two key shortcuts: the "favorite journey" and the "buy again" ticket option. P3 illustrated the value of the latter, stating, *"When you previously purchased a ticket, you have the 'buy again' option."* In contrast, AP instances identified two shortcuts: the "favorite journey" and the "my location" feature, which helps select the nearest stop.

Aesthetic and minimalist design (N08): The app's design was widely seen as uncluttered, minimalist, and aesthetically pleasing by both interview participants and APs. P3 noted that while the journey information was generally sufficient, a complex journey could make the screen difficult to read due to information overload. The APs echoed this sentiment. For instance, an instance of Claudio shared, *"Most screens are clean and focused. Screenshot 2 with the destination list is well-organized, and the journey details in screenshot 5 show just what I need. However, screenshot 3 with multiple journey options feels a bit dense with all the times and details - it could benefit from slightly more spacing or clearer visual separation between options (Q15, nielsen_ 9, claudio_ claude)."*

Help users recognize, diagnose, and recover from errors (N09): Only one interview participant, P3, had a direct experience with an error. P3 shared that while rushing to buy a ticket, a network issue occurred, but the error message was "quite clear (P3)." In contrast, none of the AP instances recognized application errors, which was expected given that no error screens were included in the provided screenshots. For example, an instance of Claudio stated, *"No errors are shown in these screenshots, so I cannot judge. Neutral based on what I can see (Q17, nielsen_ 38, claudio_ openai)."* However, a Google Play review from a user (f031) underscored a real-world issue with the app's error handling, noting a generic and unhelpful message: *"When I tried to register my number, I got the following error message: 'Oops, something went wrong! Contact customer service for more information'."*

Help and documentation (N10): No interview participants reported using the in-app help or documentation features. For example, P4 shared, *"I never check the help in the app. I prefer asking friends or ChatGPT."* Similarly, the AP instances found little in-app help, tooltips, or visible guides based on the provided screenshots. An instance of Claudio noted, *"Some screens offer additional information, such as the 'Read about our tickets' option (screenshots 6 and 7). However, more visible help options or tooltips could be beneficial, especially for less familiar features (Q19, nielsen_ 23, claudio_ llama)."*

3.2 Relationship Between Personas, LLMs, and APs Feedback

The quantitative feedback, measured on a 5-level Likert scale (1: *"strongly disagree"*, 5: *"strongly agree"*), reveals a clear interplay between the persona's definition and the LLM in shaping the final evaluation. Two distinct trends were evident. First, the persona's specific traits consistently guided the tone of the feedback. The Ingrid persona, designed with the accessibility needs of a visually impaired user, was invariably more critical than the Claudio persona. This pattern was consistent across all four LLMs, with Ingrid's agreement level (3.78 - 4.12) consistently lower than Claudio's (3.88 - 4.66).

Second, the choice of LLM acted as a distinct lens, influencing the overall level of criticality. The Llama model emerged as the most critical evaluator, producing the lowest agreement level for both personas (3.88 for Claudio and 3.78 for Ingrid). Conversely, Gemini was the most generous, yielding the highest agreement level (4.66 for Claudio and 3.99 for Ingrid). GPT and Claude performed in the median range. These results suggest that AP-generated feedback is not determined by one factor alone, but by the combination of a persona's defined perspective and the unique evaluative tendencies of the LLM.

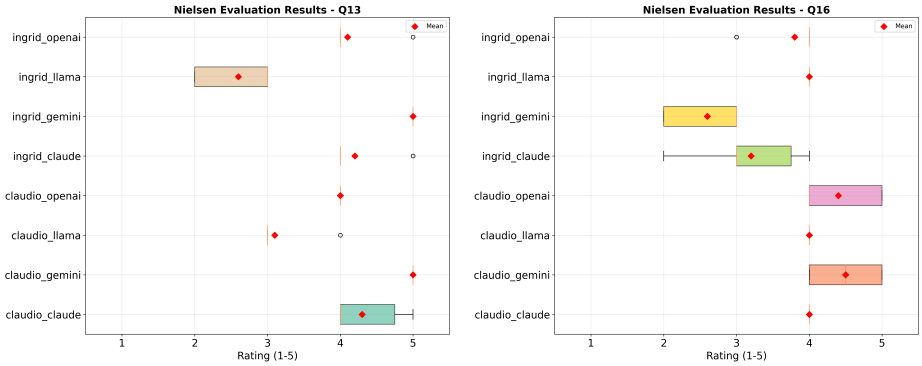

Fig. 3. Agreement levels on Q13 and Q16

A per-statement analysis provides deeper insight into these patterns. The divergence between the two personas was most pronounced in statements directly related to visual clarity and accessibility. For instance, in statements Q16 (*"Important information immediately grabs your attention"*) as depicted in Fig. 3 (the right side), Ingrid's instances gave significantly lower agreement levels, reflecting her defined visual impairment. Our analysis also highlighted specific LLM tendencies, such as Llama's consistent criticality on matters of user efficiency. For example, Q13 (*"The app visibly offers shortcuts (e.g., favorite routes) for common actions."*) as depicted in the left side of Fig. 3. Despite these variations, the AP instances reached a consensus on certain issues. For example, uniformly neutral scores for Q17 *("If visible, error messages clearly explain what went wrong,"*) as no error screens were provided. Our Kruskal-Wallis tests further confirm that

the persona-model combination significantly impacts the agreement level with $pvalue < 0.05$ in all statements, except Q17.

4 Discussion

Our research question sought to understand the effectiveness of AP in generating user feedback measured through the alignment between AP-generated feedback and human user feedback. Our qualitative analysis revealed a substantial alignment, with APs successfully identifying many of the same usability issues and strengths as human participants and Google Play reviewers across Nielsen's heuristics. This suggests that APs, when properly configured, can effectively simulate a user's perspective in a heuristic evaluation context. For instance, the APs' ability to detect issues related to consistency (N04), recognition over recall (N06), and minimalist design (N08) demonstrates their capacity to process visual information and apply usability principles in a human-like manner.

The Ingrid persona, designed with a vision impairment, consistently highlighted accessibility issues such as small text and low-contrast elements. This finding shows that APs can be tailored to represent specific user needs, aligning with the core purpose of personas: to foster empathy and focus design efforts on diverse user profiles [15, 29].

The APs' inability to comment on dynamic, real-world issues (e.g., incorrect train departure times) or application errors not depicted in the screenshots highlights a key limitation of our static evaluation method. Human feedback is shaped by lived experiences and interactions with a live system, a context that static images cannot fully replicate.

Our statistical analysis clearly showed that the Ingrid instances were consistently more critical than the Claudio instances, regardless of the underlying model. This demonstrates that the persona definition successfully guided the LLM's output, causing it to adopt the specified critical perspective. Furthermore, the variation in criticality among the LLMs themselves (with Llama being the most critical) suggests that factors like training data, architectural differences, and built-in safety protocols influence the model's evaluative stance. This implies that the selection of both the persona and the LLM is a critical configuration that can be tuned to achieve different evaluation goals.

Our results also indicate that all of the underlying LLMs are capable of avoiding hallucinating responses by showing their neutrality on insufficient context, for example, in Q17, where no screenshot indicating errors was given.

We recognized two limitations of the study. Firstly, the use of static screenshots prevented the APs from evaluating dynamic interactions and real-time context. Secondly, our human feedback was gathered from a small interview participant sample and reviews on Google Play.

Implications for Practice and Research

For practitioners, APs offer a practical, low-cost method for rapid internal evaluations, allowing teams to address resource-intensive issues of user involvement in early stages of NPD [25]. We also recommend practitioners not to oversimplify user experiences by relying too heavily on APs' feedback since the underlying LLMs have the potential of inherent biases accumulated from the training data. Gaps between APs' feedback and actual user feedback should be expected. We suggest continuously informing APs with new information gathered from external experimentations to make APs a better reflection of the underlying user group. However, practitioners must also consider APs' computational costs. As our findings on execution performance show, these costs are driven by factors, such as API calls, token consumption, and execution time.

For researchers, this work validates the novel GenAI application, transforming static personas into dynamic evaluation tools and opens new avenues for research into further APs' involvement in various NPD stages, such as requirement engineering and behavioral testing.

5 Conclusion

This study investigated AI-powered Active Personas (APs) to address the challenge of securing consistent user feedback in new product development. Through Design Science Research, we transformed traditional static personas into dynamic, generative agents capable of providing realistic user feedback on demand.

Our empirical evaluation confirms that APs generate high-fidelity feedback that closely aligns with human users across Nielsen's usability heuristics. A key finding is the significant influence of both persona definition and underlying LLM choice on the generated feedback.

This study validates the viability of APs in generating contextual user feedback. APs offer a practical solution for securing diverse user perspectives while maintaining focus on user empathy. Future work is to enable APs to interact with the dynamic context of applications, validate APs in other domains, and establish ethical guidelines for responsible deployment.

Acknowledgements

This work has been supported by ELLIIT, the Swedish Strategic Research Area in IT and Mobile Communications.

References

1. Ahmad, R., Siemon, D., Gnewuch, U., Robra-Bissantz, S.: The benefits and caveats of personality adaptive conversational agents in mental health care. In: Americas Conference on Information Systems (2021)

2. Barthet, M., Khalifa, A., Liapis, A., Yannakakis, G.: Generative personas that behave and experience like humans. In: International Conference on the Foundations of Digital Games (2022)
3. Billestrup, J., Stage, J., Nielsen, L., Hansen, K.S.: Persona usage in software development: Advantages and obstacles. In: International Conferenc on Advances in Computer-Human Interactions (2014)
4. Bjarnason, E., Lang, F., Mjöberg, A.: An empirically based model of software prototyping: a mapping study and a multi-case study. Empirical Software Engineering **28**(5), 115 (2023)
5. Branco, A.C., Sacramento, E., Oliveira, E., Tymoshchuk, O., Antunes, M., Almeida, M., Pedro, L., Ramos, F., Carvalho, D.: Usability evaluation of a community-led innovation mobile app. In: International Conference on Computer-Human Interaction Research and Applications. p. 81 – 88 (2022)
6. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qualitative Research in Psychology **3**(2), 77–101 (2006)
7. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: International Conference on Neural Information Processing Systems (2020)
8. Cooper, A., Saffo, P.: The Inmates Are Running the Asylum. Macmillan (1999)
9. Cruzes, D.S., Dyba, T.: Recommended steps for thematic synthesis in software engineering. In: International Symposium on Empirical Software Engineering and Measurement. pp. 275–284 (2011)
10. Ebert, C., Louridas, P.: Generative ai for software practitioners. IEEE Software **40**(4), 30–38 (2023)
11. Fabijan, A., Olsson, H.H., Bosch, J.: Customer feedback and data collection techniques in software r&d: A literature review. In: International Conference of Software Business. pp. 139–153 (2015)
12. Huang, Y., Kanij, T., Madugalla, A., Mahajan, S., Arora, C., Grundy, J.: Unlocking adaptive user experience with generative ai. In: International Conference on Evaluation of Novel Approaches to Software Engineering (2024)
13. ISO: Iso 9241-11 ergonomics of human-system interaction - part 11: Usability: Definitions and concepts (2019)
14. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM Computing Surveys **55**(12) (2023)
15. Karolita, D., Grundy, J.C., Kanij, T., McIntosh, J., Obie, H.O.: Lessons learned from persona usage in requirements engineering practice. In: IEEE International Requirements Engineering Conference. pp. 116–128 (2024)
16. Karolita, D., McIntosh, J., Kanij, T., Grundy, J., Obie, H.O.: Use of personas in requirements engineering: A systematic mapping study. Information and Software Technology **162**, 107264 (2023)
17. Kujala, S., Kauppinen, M., Lehtola, L., Kojo, T.: The role of user involvement in requirements quality and project success. In: IEEE International Conference on Requirements Engineering. pp. 75–84 (2005)
18. Lechner, B., Fruhling, A., Petter, S., Siy, H.: The chicken and the pig: User involvement in developing usability heuristics. In: Americas Conference on Information Systems. vol. 5, pp. 3263–3270 (01 2013)

19. Lethbridge, T.C., Sim, S.E., Singer, J.: Studying software engineers: Data collection techniques for software field studies. Empirical Software Engineering **10**(3), 311–341 (2005)
20. Leão, R., Ayach, F., Lameirão, V., Fontão, A.: A prompt engineering-based process to build proto-personas during lean inception. In: Proceedings of the Brazilian Symposium on Software Engineering (2024)
21. Matos, D.P., Torres, M.D., da Silva, L.S.R., dos Santos, C.A.A.S., de Oliveira, F.J.F., de Araújo, M.F.M., de Oliveira Serra, M.A.A.: Hansenapp: Development of a mobile application to assist primary healthcare providers to control leprosy. Tropical Medicine & International Health **27**(8), 719–726 (2022)
22. Mueller, A., Beyer, S., Kopp, G., Deisser, O.: User-centered development of a public transportation vehicle operated in a demand-responsive environment. In: Stanton, N. (ed.) Advances in Human Factors of Transportation. pp. 545–555. Springer International Publishing, Cham (2020)
23. Nielsen, J.: Usability Engineering. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1994)
24. Norman, G.: Likert scales, levels of measurement and the "laws" of statistics. Advances in Health Sciences Education **15**(5), 625–632 (2010)
25. Paternoster, N., Giardino, C., Unterkalmsteiner, M., Gorschek, T., Abrahamsson, P.: Software development in startup companies: A systematic mapping study. Information and Software Technology **56**(10), 1200–1218 (2014)
26. Peffers, K., Tuunanen, T., Rothenberger, M.A., Chatterjee, S.: A design science research methodology for information systems research. Journal of Management Information Systems **24**(3), 45–77 (2007)
27. Ries, E.: Lean Startup: How Today's Entrepreneurs Use Continous Innovation to Create Radically Successful Businesses. Crown Business (2011)
28. Salewski, L., Alaniz, S., Rio-Tirto, I., Schulz, E., Akata, Z.: In-context impersonation reveals large language models' strengths and biases. In: International Conference on Neural Information Processing Systems (2023)
29. Salminen, J., Jansen, B.J., An, J., Kwak, H., Jung, S.g.: Are personas done? evaluating their usefulness in the age of digital analytics. Persona Studies **4**(2), 47–65 (2018)
30. Sauvola, T., Lwakatare, L.E., Karvonen, T., Kuvaja, P., Olsson, H.H., Bosch, J., Oivo, M.: Towards customer-centric software development: A multiple-case study. In: Euromicro Conference on SEAA. pp. 9–17 (2015)
31. Seaman, C.B.: Qualitative methods in empirical studies of software engineering. IEEE Transactions on Software Engineering **25**(4), 557–572 (1999)
32. Simaremare, M., Edison, H.: Accelerating new product development: A vision on active personas. In: Software Business. pp. 461–466. Springer Nature Switzerland, Cham (2025)
33. Simaremare, M., Edison, H.: Active personas for synthetic user feedback - appendix (2025). https://doi.org/10.6084/m9.figshare.29925158
34. Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A.: Experiments, pp. 73–83. Springer Berlin Heidelberg (2024)
35. Zheng, M., Pei, J., Logeswaran, L., Lee, M., Jurgens, D.: When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Findings of the Association for Computational Linguistics. pp. 15126–15154. Association for Computational Linguistics (2024)