AI Can See What You Can’t See: How LLM-Agents Complement Human-Based Gender-Inclusive Usability Testing

Joy Stefanie Geuenich, Frank Joublin, Antonello Ceravola, Stefan Brandenburg, Jonathan Dodge

2026

Preprint:

Copyright ACM 2026. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in Transactions on Interactive Intelligent Systems, https://doi.org/10.1145/3801978.

Check for updates

ACM DIGITAL LIBRARY | Association for Computing Machinery | acm open

Latest updates: https://dl.acm.org/doi/10.1145/3801978

RESEARCH-ARTICLE

AI Can See What You Can't See: How LLM-Agents Complement Human-Based Gender-Inclusive Usability Testing

JOY GEUENICH

FRANK JOUBLIN

ANTONELLO CERAVOLA

STEFAN BRANDENBURG

JONATHAN DODGE

PDF Download
3801978.pdf
16 March 2026
Total Citations: 0
Total Downloads: 0

Published: 13 March 2026
Accepted: 02 March 2026
Revised: 27 February 2026
Received: 25 August 2025

Citation in BibTeX format

AI Can See What You Can't See: How LLM-Agents Complement Human-Based Gender-Inclusive Usability Testing

JOY GEUENICH, Chemnitz University of Technology, Germany
FRANK JOUBLIN, Honda Research Institute Europe, Germany
ANTONELLO CERAVOLA, Honda Research Institute Europe, Germany
STEFAN BRANDENBURG, Chemnitz University of Technology, Germany
JONATHAN DODGE, Penn State University, USA

Inclusive usability testing, such as the GenderMag method, wants to identify gender-related usability problems in digital interfaces. Large Language Models (LLMs) have been used by usability engineers in usability evaluations but their contribution is still underexplored, especially regarding inclusive usability testing. Research has shown that GenderMag workshops can produce valuable insights but are resource intense and might show effects from the evaluators' ability to embody personas with a different cognitive style. Therefore, we need to assess if LLM-agent based testing can aid human-based evaluations. This study evaluates an LLM-agent system for GenderMag persona-based usability testing, and compares its performance to traditional human-led evaluations. The agent system integrates GenderMag persona facets into three LLM-agents, which analyze usability issues of four web interfaces, three generic and one intentionally flawed interface with gender-related usability issues. We quantitatively and qualitatively compare the types, severity, and relevance of usability problems LLM-agents identified to those produced in three GenderMag workshops involving nine participants. Findings show a broad overlap in detecting usability issues of humans and LLM-agents that are not gender specific via the generic interfaces. Agents thereby consistently assign significantly higher severity and relevance ratings. For the intentionally flawed interface, humans and LLM-agents assign similar ratings but the overlap between human and agent gender-related usability issues was low, with each missing issues the other caught. The agent system is an efficient tool for gender-related usability evaluation that humans may overlook, thereby expanding the coverage of evaluation.

CCS Concepts: • **Human-centered computing → Heuristic evaluations**; *Empirical studies in HCI*.

Additional Key Words and Phrases: Agent usability testing, GenderMag personas, Inclusive design, Human-computer interaction (HCI), Large language models (LLMs), Persona-based evaluation, Automated usability evaluation, Human-AI collaboration

1 Introduction

Usability problems reduce how effectively, efficiently, and satisfactorily people interact with technical systems, which lowers their acceptance and intention to use these systems. When designing these systems, usability engineers can employ user-related methods, to ensure that systems accurately reflect user requirements and needs. For user-related methods, engineers need to select participants for user studies from the estimated population(s)

Authors' Contact Information: Joy Geuenich, joy-stefanie.geuenich@psychologie.tu-chemnitz.de, Chemnitz University of Technology, Chemnitz, Saxony, Germany; Frank Joublin, Frank.Joublin@honda-ri.de, Honda Research Institute Europe, Offenbach/Main, Hesse, Germany; Antonello Ceravola, Antonello.Ceravola@honda-ri.de, Honda Research Institute Europe, Offenbach/Main, Hesse, Germany; Stefan Brandenburg, stefan.brandenburg@psychologie.tu-chemnitz.de, Chemnitz University of Technology, Chemnitz, Saxony, Germany; Jonathan Dodge, jxd6067@psu.edu, Penn State University, University Park, PA, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM 2160-6463/2026/3-ART
https://doi.org/10.1145/3801978

Just Accepted

of users, which is time consuming for diverse user groups and may be difficult in early stages of product development. Alternatively, usability engineers can approximate the fulfillment of universal usability criteria like understandability and error tolerance with expert-based methods. One such method is heuristic usability evaluation [43], which researchers have applied extensively in the past [22]. Researchers perform workshops that assess usability problems and rate their severity and relevance to analyze their actionability and prioritization depending on the extent to which the problem prevents the users from fulfilling their tasks [48]. However, having an expert perform a heuristic evaluation may make it difficult for the expert to adequately address the requirements and needs of diverse user groups. The present paper examines whether Large Language Model (LLM) agents may help to find usability problems and consider gender aspects at the same time by having them perform heuristic evaluations. More specifically, we want to examine mechanisms to integrate LLM-based agents in finding usability problems and gender-related issues in usability evaluations in early stages of human-centered product development according to DIN EN ISO 9241-210:2019 [16], which define principles and activities for this development.

Usability engineers still face challenges when applying heuristic evaluation, despite the extensive use of this method. For example, on the one hand, engineers have to apply usability heuristics with respect to universal criteria, applicable to all users of a technology. If usability engineers employ heuristics, they focus on universal usability issues, which may create tensions with accounting for issues that only exist for specific user perspectives. For example, Hofman et al. [29] show how accessibility features that work for some users, might introduce new barriers for others, if usability engineers aim to find one solution for all. To account for diverse perspectives, personas or similar methods can help engineers determine whether the system fulfills usability heuristics for users with different requirements and needs. In general, personas describe prototypical users with different characteristics such as gender, age, interests, special requirements concerning technology, etc. Research has led to specific forms of personas like the GenderMag [7] to focus on user requirements and needs that are associated with certain facets of their profile. Usability engineers following the GenderMag method seek to address gender bias in system design by incorporating gender-inclusive personas into the usability evaluation process. The GenderMag personas illuminate usability problems that disproportionately affect users with diverse problem-solving styles and attitudes linked to gender, e.g., focusing on male personas only [7].

Although GenderMag has proven successful for some applications (e.g., [1, 6]), at least two important challenges remain. First, applying the GenderMag method is intensive in labor, time, and expertise [31]. Second, GenderMag challenges usability engineers to cover a range of different personas. In doing so, usability engineers may introduce inconsistencies when applying diverse personas, as they may find it difficult to see things from the perspectives of different personas or unconsciously revert to implicit assumptions, stereotypes, or their own individual cognitive styles (e.g., exploring an interface by tinkering, rather than taking a systematic approach).

This work seeks to improve both issues by combining GenderMag personas into an LLM-based agent architecture. To simplify the process, we configured the LLM-agents so that they directly access the interface's HTML code. The LLM-agents will identify usability issues, whether universal in nature (i.e., from Nielsen's heuristics [42, 43]) or related to gender inclusiveness issues (i.e., from GenderMag [7]). To evaluate the quality of the results, we will compare the results of the LLM-agents with the findings from usability engineers conducting traditional GenderMag workshops based on direct interface interaction. The comparisons of LLM agent and human performance highlights the potential of LLM-agents for usability evaluations in general and gender sensitive usability evaluations in particular. To explore this potential, we investigate the following research questions (RQs):

- **RQ1:** What is the overlap between persona-specific usability problems of LLM-agents and usability experts?
- **RQ2:** How do the LLM-agent severity and relevance ratings of usability problems compare to usability experts?
- **RQ3:** How do the agent gender-related issues compare to the ones found by usability experts?

Just Accepted

In answering those RQs, we contribute the following to the research field on how to address usability testing in diverse user groups with LLM-based agents:

- An agent architecture that automates gender sensitive usability testing via Gender-Mag persona-based testing of HTML-based UIs
- Comparison of LLM-agent persona-based usability testing to human participants regarding severity, relevance and problem types
- Discussion on how a complementary perspective of usability engineers and agent architecture can contribute to a more complete picture of an interfaces strengths and weaknesses

2 Related Work

2.1 Persona-Based Usability Testing via GenderMag

Usability engineers define personas to represent important characteristics of user groups that interact with a given system in a given context [12]. Personas provide structured representations of users, helping designers to direct evaluations and design recommendations towards user characteristics, goals and behaviors. Personas have been an effective tool for usability testing, helping usability engineers understand different user perspectives. Still, Pruitt and Adlin [47] conclude that *"it is much easier for stakeholders to prioritize fewer personas"* (p. 51). Considering multiple personas at the same time may be confusing to stakeholders, which can lead usability engineers focus on some personas more than others. This, in turn, can lead to biased interfaces that may favor the requirements and needs of one user group more than the other. Also, usability engineers may focus too much on few user characteristics described in the personas, potentially overlooking important characteristics the personas do not cover or that engineers deem to be not very important, like gender.

The GenderMag persona method builds on data that illuminates gender-specific differences between users, with the goal of reducing bias via gender-inclusive personas. GenderMag is a cognitive walkthrough method to evaluate software using a set of personas with five facets each [7]. The five facets are structured around the five key cognitive dimensions in Table 1. The GenderMag method (and the current study) embodies these facets in the three personas Abi, Pat, and Tim. Each represents different combinations of the five facets, where Abi and Tim are at opposite ends of these dimensions and Pat is in the middle.

Field studies demonstrated the effectiveness of GenderMag in identifying usability issues that traditional heuristic evaluations without a focus on gender-related issues overlook, especially regarding non-dominant cognitive styles or motivations. Recent work further supports GenderMag's effectiveness in identifying inclusivity-related issues, including evidence that GenderMag-guided analyses can improve the detection of gender-related usability barriers [25]. For instance, some users are motivated by reaching their goal in the interface efficiently (see "goal-oriented" in Table 1), whereas others are motivated to learn and understand the interface in general (see "tech exploration" in Table 1). The first user group therefore values streamlined workflows whereas the other user group might like opportunities to explore further features (e.g., [1, 6, 7, 14, 41]). However, similar to

	Abi	**Pat**	**Tim**
Motivation	Goal-oriented	Balanced flexibility	Tech exploration
Information Processing	Step-by-step, careful	Adaptive (careful/skimming)	Skimming, fast experimentation
Risk Attitude	Risk-averse	Moderately open	Risk-tolerant
Self-Efficacy	Lower self-efficacy	Moderate self-efficacy	High self-efficacy
Learning Style	Prefers guided examples	Mix of learning styles	Trial-and-error experimentation

Table 1. Paraphrased Summary of GenderMag's Persona Facets [7]

Just Accepted

other persona methods, GenderMag's reliance on usability engineers' training and persona interpretation may lead to inconsistencies in usability problem identification. Usability engineers may unconsciously revert to their own cognitive styles or misinterpret persona facets. For example, the GenderMag field study [6] shows that an evaluator struggled to adopt Abis perspective: *"I am not like Abby... It was hard for me to step into those shoes..."* (p.10).

Also, applying GenderMag is labor intensive because it requires moderators to facilitate cognitive walkthrough sessions and needs active participation from stakeholders. Recent work has begun to explore automated approaches to identifying inclusivity and GenderMag-related issues. Chatterjee et al. [10, 11] propose automated inclusivity-bug detection techniques that operationalize GenderMag-inspired principles to flag potential barriers, primarily in educational software contexts. While these approaches demonstrate the feasibility of automation, they do not explicitly simulate the full set of GenderMag cognitive facets, leaving open questions about how automated systems compare to human evaluators across personas.

2.2 LLM-Agent Systems for Usability Testing

Usability engineers that integrate LLM-agent systems into usability evaluation can tackle long-standing challenges of scalability, inconsistency, efficiency and inclusiveness [23, 44]. LLM-agents may simulate the role of users [21, 39, 45, 55, 59, 60, 63] or usability engineers [30] in various test settings.

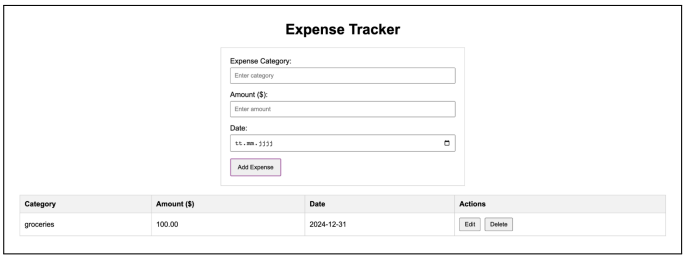
LLMs in heuristic and early-stage evaluations: LLMs can be be a part of usability evaluations to evaluate user interaction data (e.g., screenshots and interaction logs) [30]. Hsueh et al. [30] found that LLMs applied Nielsons [43] heuristics more consistently than human evaluators for the following heuristics: "visibility of system status", "match between system and real world", and "error prevention". However, they underperformed in detecting temporal issues (e.g., response delays) and subtle visual layout inconsistencies. Similarly, Duan et al. [19] used LLMs to generate heuristic feedback on Figma UI mockups. They found that LLM are effective for identifying poor visual design, especially when applying visualization-specific heuristics rather than the broader Nielson heuristics. However, the quality of LLM generated feedback declined as designers refined their mock-ups. Duan et al. [19] concluded that LLMs are helpful only in early stages of design, which is in line with other research [35, 51].

LLMs simulating personas/users: Recent research concerning the simulation of users [21, 39, 45, 55, 59, 60, 63], show the potential of LLMs to integrating persona data and applying this knowledge to solve complex usability analysis. For example, Lu et al. [39] used a persona generator to create diverse user profiles that completed different tasks using a simple UI, which usability engineers reported as insightful. Similarly, Xiang et al. [59] used LLM-agents and user characteristics to test a smartwatch interface successfully, as their system generated heuristic usability feedback that was 35.7% to 100% similar to their benchmark. Park et al. [45] simulated an online community via user agents to test social dynamics. While these studies show the technical feasibility of simulating personas at scale, few explicitly model underlying cognitive facets that represent differences in HCI. This gap limits understanding of how LLMs can reveal inclusive usability issues.

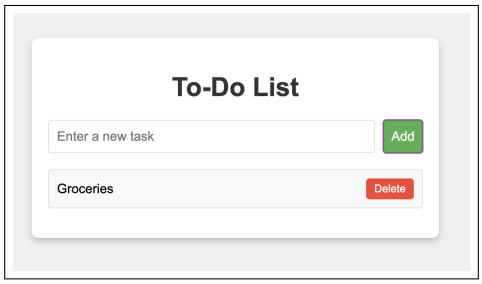
Challenges of LLM usage in usability evaluations: bias, hallucination, reproducibility: LLMs are a promising technology for simulating users because other than rule-based AIs, they are able to reason with a high degree of human-likeness [49].

However, the stochastic nature of LLMs arguably requires an analysis of the reliability and validity of their test results when integrating them into usability evaluation [38]. For example, Schuller et al. [50] found similar quality and acceptance of LLM generated personas to human generated. Other researchers stated that LLM outputs pose the problem of lacking reproducibility [36] and hallucinations [33]. Empirical tests of the consistency and appropriateness of LLM-generated outputs in usability testing are scarce, which are research gaps we also

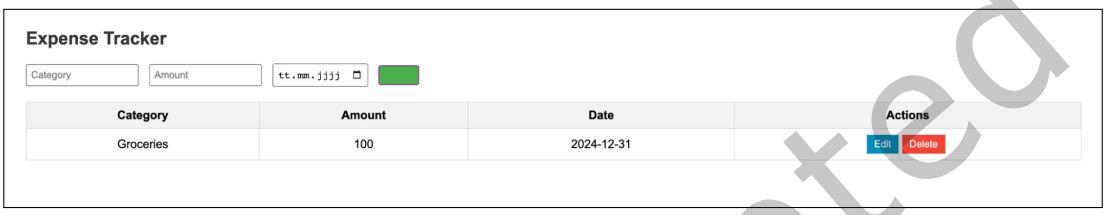
Just Accepted



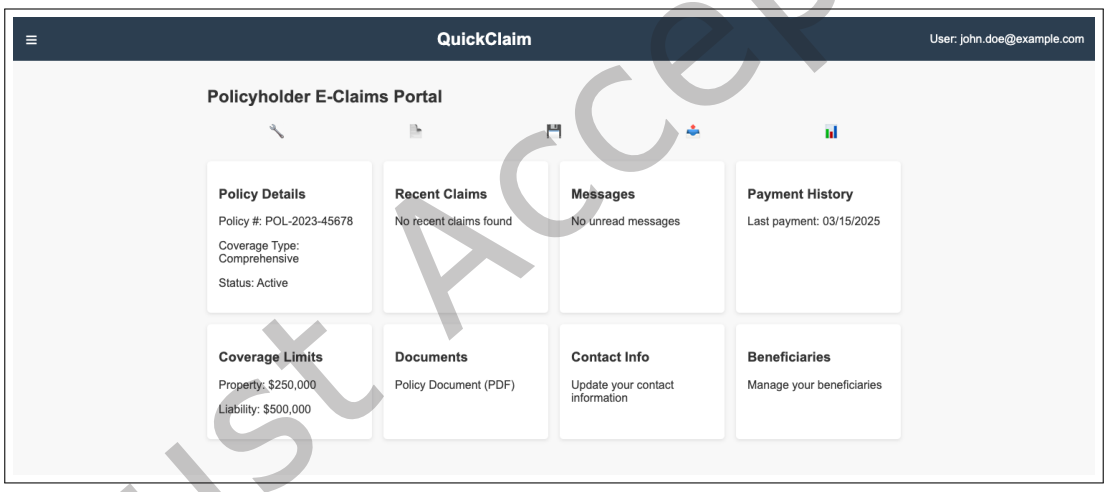
(a) Expense tracker screen I



(b) To-do list



(c) Expense tracker screen II



(d) One Screen of the Insurance Claim Interface

Fig. 1. The four interfaces: expense tracker screen I (a), a to-do list (b), expense tracker screen II (c), and an insurance claim screen (d).

Just Accepted

seek to address. This gap motivates our investigation into how usability engineers must carefully calibrate and benchmark LLM outputs against traditional human-powered methods.

3 Methodology

3.1 Creating a List of Usability Problems in Human GenderMag Workshops

We conducted three human-led GenderMag workshops to evaluate the usability and possible gender-related issues of four web interfaces. Three interfaces are generic, while the fourth is intentionally flawed as it has gender-related usability issues integrated (see Table 4). The generic interfaces are: Expense tracker screens one and two (Figure 1a and 1c), to-do list (Figure 1b), and the intentionally flawed interface for making an insurance claim (Figure 1d). The first workshops goal was to evaluate the generic interfaces (expense tracker and to-do list) to identify general usability and possibly gender-related issues. The second and third workshop evaluated the intentionally flawed interface (insurance claim), which differs from the other interfaces because we deliberately implemented usability issues that we hypothesized to trigger inclusive usability issues that the GenderMag literature outlined. This design allows us to especially address RQ3.

Six participants (5 female, mean age: 35, age range 28-40 years) took part in the initial workshop. Three of the original participants (2 female, mean age: 32, age range 28-36 years) and three independent participants (3 female, mean age: 39, age range 33–46 years) participated in two follow-up workshops. We chose this sample size based on established usability research demonstrating that the majority of usability issues are identified with three to five participants [42], and because our study focuses on qualitative analysis rather than statistical generalization. All participants were psychologists by training, familiar with usability evaluation, with each having a minimum of three years of experience in usability testing.

The workshop materials included the GenderMag personas [7, 15] (see Table 1) and usability problem reporting sheets [48] pre-formatted for all four interfaces. Tasks for the expense tracker included adding, editing, and deleting expenses. For the to-do list, tasks were adding and deleting an item. Finally, for the insurance claim interface, participants explored the interface freely and filed a claim.

We designed the intentionally flawed interface (Figure 1d) to trigger GenderMag-related issues (Table 4). We embedded 17 usability issues, which we hypothesized will trigger the personas differently due to their facets (see Table 1). Across these issues, raters achieved a mean inter-rater reliability (IRR) of 80% across personas. We drew from known problem areas related to cognitive styles, motivations, and strategies as the GenderMag literature describes [1, 6, 7] as well as design flaws that violate Nielsons heuristics [43]. For example, a gender-inclusive issue may be present when users need to provide details but the interface offers no explanation or context. Based on a field investigation of the GenderMag method [6], we expect Abi to identify this as a flaw (or at least assign it higher severity than Tim). This is because Abi's comprehensive information processing relies on having a distinct mental model before interacting. In contrast, Tim, as a tinkerer, is more likely to try to see what happens and therefore may perceive it as less problematic. To support that the implemented issues will affect the personas facets, we performed an inter-rater reliability check (see Table 2). Furthermore, Table 4 shows how each implemented issue is hypothesized to be detected by the human participants and the LLM-agents.

Naturally, humans access the interface by interacting with the UI, LLM-agents read the HTML code. Therefore, issue 11 (Table 2) should be undetectable by the LLM-agents. We included the undetectable issue to verify that humans would detect problems that the LLM-agents cannot access. For all other issues humans and LLM-agents can access the same problems because they are represented in the code, or can be inferred by the LLM and are visible in the UI during interaction.

The lead author of this paper moderated three 90-minute workshops, one in a laboratory setting with a laptop and a 58 inch screen (for the generic interfaces) and two online workshops using the browser-based screen sharing software Big Blue Button [3] (for the intentionally flawed interface). In all workshops, the workshop leader first instructed participants on the purpose and scope of the workshop, then on the GenderMag method (personas and their facets) and the heuristic evaluation method. Next, the workshop leader introduced and presented the web interfaces on the screen before the participants interacted with them. The participants had time to ask questions

Just Accepted

Implemented Issue	Abi	Pat	Tim
1. No clear call to action on landing page			
2. Technical jargon ("Policyholder E-Claims Portal")			
3. No step-by-step guide/onboarding			
4. Important info hidden in tooltips/accordions			
5. Buttons disabled with no error hint			
6. Unconventional navigation (sliding menu)			
7. Overloaded dashboard			
8. Only one fixed workflow (no going back)			
9. Hover-only feedback			
10. Doesn't explain why to upload an image			
11. Unrecoverable errors			
12. Error-prone design ("Cancel" near "Submit")			
13. No immediate feedback after "Submit"			
14. Strict linear form (no shortcuts)			
15. "Advanced" collapsed by default			
16. Ambiguous icon labels			
17. No summary/overview after submission			
IRR	94%	71%	76%

Table 2. Hypothesized probability of the personas identifying the implemented issue. Red is a high probability , yellow is a medium probability, green is a low probability . Top-left shade is the rating of rater 1, bottom-right of rater 2.

Just Accepted

before beginning the heuristic evaluation using the GenderMag personas. In the first workshop, they performed the tasks for each interface sequentially, completing each of the tasks for each interface per persona (i.e., two Abis, two Tims, two Pats). We randomized the order of personas and interfaces across participants to minimize order effects. In the second and third workshop, we followed the same procedure except that we only assigned one persona to each participant, which they applied to the intentionally flawed interface. At the end of the second workshop, participants had time for reflection and feedback to collect qualitative insights. Participants documented usability problems on reporting sheets, which we transcribed and qualitatively coded following the procedure of Reckin et al. [48]. We retrospectively surveyed workshop participants cognitive facets using the GenderMag Facet Survey [24] to control for the participants cognitive styles. The Supplementary Material includes LLM prompts, outputs, codebooks for qualitative analysis, and additional data.

3.2 Comparative Setup: Configuration of the LLM-Agent System using HyperGraphOS

In the present study, we used HyperGraphOS [8] to configure the LLM-agent. HyperGraphOS provides a graph-based platform to structure and execute AI workflows, enabling the execution of graph-based dataflow systems for evaluating web interfaces. Nodes of such graphs represent computational tasks, for example, data containers and JavaScript functions like calling OpenAI's API. Edges represent the flow of data/events between these tasks.

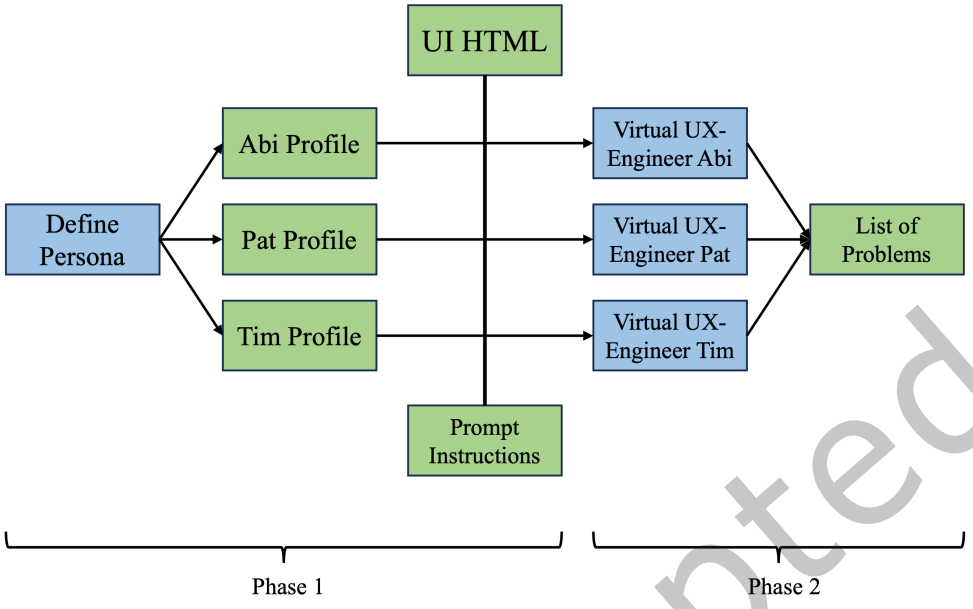


Fig. 2. Workflow of the Agent System in HyperGraphOS. In this high-level description of the system, computational tasks are shown as teal boxes and data containers as green boxes.

Just Accepted

The agent system emulated different usability testers configured with GenderMag personas Abi, Pat and Tim to evaluate self-contained HTML code of web interfaces through a two-phase process Figure 2 illustrates.

In **Phase 1**, we initialized each LLM agent (one per persona profile, see Figure 2) using GenderMag specifications, which define the characteristics and preferences of the personas they represent. These personas served as user models for assessing the usability of the web interface. We implemented every agent by calls to the OpenAI API using `GPT-4o-mini` (Version: October 2023). This model's maximum input is 128,000 tokens, which exceeded the size requirements for both the persona descriptions and the HTML interfaces. Therefore, its maximum output length of 16,384 tokens was sufficient for reporting bugs.

In **Phase 2**, we prompted the LLM agent with the persona definition (i.e., facets) and the HTML code of the web interface under evaluation. Those prompts asked LLM-agents to identify usability problems, including problem types, severity, and relevance ratings, as well as justify their ratings by referring to the appropriate facets. Each agent responded through the OpenAI API as a JSON structure that described usability problems. The parameters of the LLMs are the defaults, except temperature, which we set to 0.8 (range = 0.0 – 2.0) to allow variations in LLM responses. We refined prompts so that the model returned the expected JSON file schema we anticipated. In pilot runs, we checked the reported usability problems for semantic correctness and if the facet based justifications aligned with the persona definitions. In the preliminary tests, we did not observe inconsistent or unexpected outputs.

3.3 Data Analysis

To answer our research questions one and two (RQ1, RQ2), we compared human and LLM agent outputs using quantitative and qualitative methods. More specifically, we compared the severity and relevance ratings using Fisher's Exact Tests (RQ2). To evaluate the performance to detect usability issues of the generic interfaces, we

categorized the usability problems using a codebook (see Appendix B) and assessed, which problems are uniquely identifiable by evaluator type (LLM agent vs. human expert) across the three personas (RQ1, RQ3). To assess the overlap in the generic interfaces (RQ1), we defined a problem as unique if two raters determined that it differed from all other problems from the same interface across all personas. To assess the overlap in GenderMag specific usability issues in the intentionally flawed interface, we report precision, recall and F1 scores, which measure the degree to which either the human or agent can identify issues from a fixed set of 17 GenderMag usability issues we intentionally seeded (RQ3).

3.3.1 Controlling for Participant–Persona Alignment. The GenderMag method recommends to survey participants' facets to understand if a participant is more likely to identify with a persona. Prior work [1] used median splits to classify participants into either Tim-like or Abi-like for each of the five facets. The median splits allow researches to compare a participants facet to their "peer group", which are the other participants of the study. This methodological approach is feasible for participant groups that approximate a real distribution of users and where a median reflects a stable central tendency. In our study, we examined nine experts, leading to a non-representative sample, which is sensitive to outliers and distributional skew. Therefore, we compute facet scores by summing the participants' responses to the facet questionnaire items [24] and focus our analysis on extrema. We categorized participants as aligned or misaligned when we measured a strong tendency, which we operationalized as "disagree completely" (scores 1–3) or "agree completely" (scores 7–9) on the Likert scale. To avoid over-claims, we left participants neutral if they provided mid-range responses (scores 4–6). This approach provides estimates of the participants cognitive styles given that we conduct a usability engineering study.

3.3.2 Pre-Analysis Reliability Assessment of LLM Agent and Human Data. Before we tackle the research questions, we examined the reliability of our data. To analyze the reliability of LLM agent and expert data, we compared LLM- and human-generated usability problem descriptions and the problems' severity and relevance ratings. For the agent output, we generated the problem analysis results three times for each persona and interface combination (3x3x3=27). By coincidence, this process obtained 5 problems each time, leading to a total number of 135 problems. Then, we assessed the cosine similarity of the problem descriptions using a token-level semantic similarity analysis with the R package BERT [4] (see Appendix C). Further we confirmed that neither severity nor relevance ratings nor the distribution of problem types differed significantly across 25 trials (see Table 3 and Figure 3). This serves as a manipulation check of the LLMs' temperature setting, with all outputs achieving a dedicated cosine similarity of minimum 0.8, indicating high consistency. For the human-generated problem descriptions, two raters categorized the transcribed problems into problem types, severity, and relevance ratings. The inter-rater reliability [32] revealed agreement rates of 86% for problem type, 90% for problem severity and 84% for problem relevance, which is good to very good. To rule out confounding influences of familiarity, we qualitatively compared the outputs of the three participants who had been part of the first workshop with the three who had not engaged with the GenderMag personas before. We did not observe systematic differences or patterns in their outputs. To rule out gender-biases, we qualitatively examined both LLM- and human output for stereotype-laden content that go beyond the GenderMag facets (e.g., suggesting Abi would like a more feminine color palette). We did not find evidence of this, which is in line with the finding that the persona Abi is less susceptible to gender-based stereotyping than real individuals [28]. Furthermore, the persona configuration prompts did not use any gendered personal pronouns, merely the persona names might infer gender. Therefore, we compared the outputs of the Abi agent with those of an Abi agent configured with Tim's facet information, as well as with an unnamed agent that carried no persona label. We found no significant differences between these agents (see Table 19 and Table 20 in Appendix C).

Just Accepted

Interface	Severity (KW Test)		Relevance (KW Test)		Problem Type (Multinomial LR)	
	H	p-value	H	p-value	LR χ^2	p-value
To-Do List	12.8	0.97	8.65	0.10	94.50	1.00
Expense Tracker I	9.11	0.10	9.22	0.10	91.7	1.00
Expense Tracker II	10.7	0.99	15.0	0.92	83.3	1.00
Insurance Claim	10.5	0.99	16.5	0.87	84.4	1.00

Table 3. Stability analysis of LLM-generated usability issues across 25 trials. Severity and Relevance ratings were analyzed using Kruskal–Wallis tests. Problem Type distributions were evaluated using a multinomial logistic regression likelihood ratio test comparing a model with trial to a null model (LR χ^2)

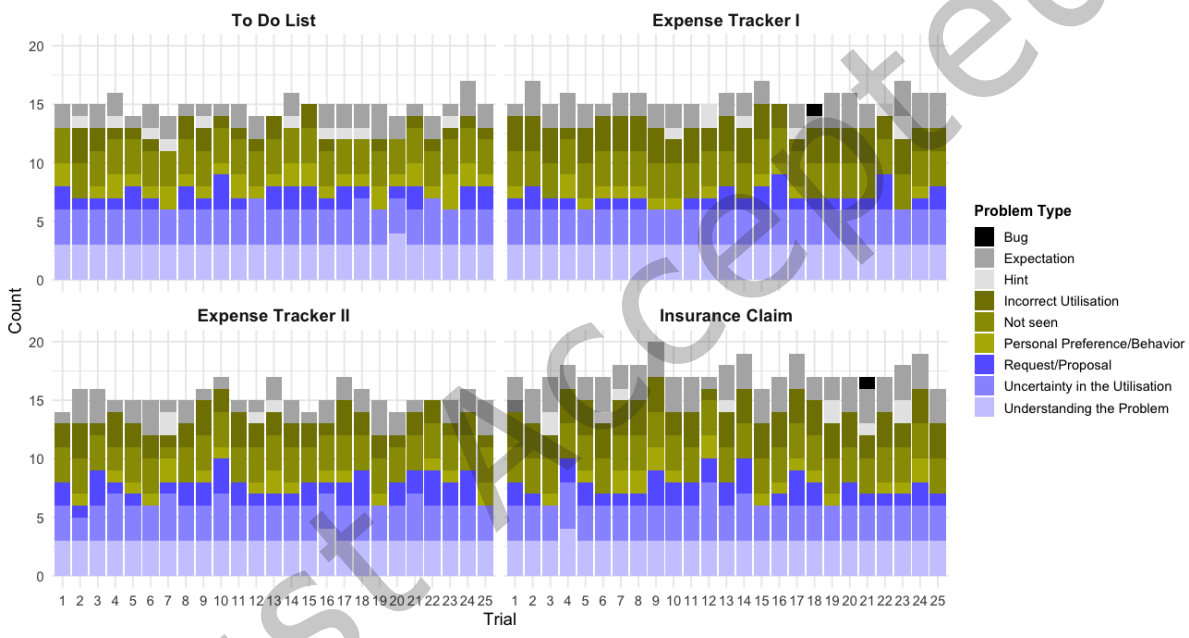


Fig. 3. Stacked bar charts showing the distribution of LLM-generated problem types across 25 trials for each interface.

4 Results

Table 5 provides an overview of the implemented usability issues, that we designed to trigger the GenderMag facets. By examining whether the agents find facet related problems, we can explicitly investigate RQ3. For example, the Abi persona is likely to flag the lack of step-by-step guidance (issue 3, in Table 5, because that persona is risk-averse and goal-orientated and therefore prefers a clear workflow [1]. Due to their lower self-efficacy they may not choose to proceed and flag it as an issue from their perspective. On the other side, users that like to tinker (e.g., Tim) might not find that issue 9 ("Hover-only feedback") or issue 1 ("No clear call to action") are problematic, because they like to explore the interface and eventually discover this information that Abis are more likely to overlook.

Just Accepted

Implemented Issue	**Access via...**
1. No clear call to action	*UI Interaction:* User may not identify primary action by perceiving interface. *Source Code:* HTML has no labels such as "Begin"
2. Technical jargon	*UI Interaction:* "Proprietary algorithmic adjudication methodology" difficult to understand and interpret. *Source Code:* HTML has "Proprietary algorithmic adjudication methodology" integrated, LLM needs to infer that the wording is difficult to understand and interpret.
3. No onboarding / step-by-step guidance	*UI Interaction:* User expects guidance information but finds none during interaction. *Source Code:* HTML contains no onboarding elements or instructions.
4. Important info hidden in tooltips/accordions	*UI Interaction:* User must find the hidden element (e.g., witness information) and understand that this is important information they would like to see straight away. *Source Code:* HTML can identify that witness information is structured in an accordion.
5. Disabled buttons with no error hint	*UI Interaction:* User notices that e.g. the icon resembling reports does not follow a function. *Source Code:* The disabled button has no event handler.
6. Unconventional navigation	*UI Interaction:* Layout irritates user. *Source Code:* HTML structure may appear unconventional, but LLM needs to infer that it is unconventional.
7. Overloaded dashboard / unclear flows	*UI Interaction:* User experiences difficulty identifying the correct path. *Source Code:* HTML represents many sections, LLM needs to infer that a point of "too much" has been reached.
8. Only one fixed workflow (no back navigation)	*UI Interaction:* User cannot return to previous steps. *Source Code:* HTML lacks back buttons or alternative paths.
9. Hover-only feedback	*UI Interaction:* User discovers text only by hovering (e.g., hovering over icon reveals that behind the icon is the site "reports"). *Source Code:* HTML represents tool-tip content.
10. No explanation of various steps	*UI Interaction:* User might wonder why image upload is necessary. *Source Code:* HTML contains no label explaining purpose, but LLM needs to infer that a user might expect this element.
11. Unrecoverable errors (form resets)	*UI Interaction:* User cannot access information they entered after clicking send and receives no information after submission. *Source Code:* HTML contains no reset mechanism.
12. Error-prone design	*UI Interaction:* User may notice that the buttons cancel and submit are placed closed to each other, or the system accepts impossible dates. *Source Code:* HTML shows placing of the buttons.
13. No immediate feedback after Submit	*UI Interaction:* Users may expect a confirmation message. *Source Code:* HTML includes no confirmation message.
14. Strict linear form (no shortcuts)	*UI Interaction:* User attempts to skip sections but cannot. *Source Code:* HTML contains no navigational shortcuts.
15. "Advanced" section collapsed by default	*UI Interaction:* User must discover hidden options. *Source Code:* HTML can identify that accordion is collapsed.
16. Ambiguous icon labels	*UI Interaction:* User doesn't understand icons. *Source Code:* HTML contains icon, but LLM needs to infer that is it ambiguous.
17. No summary/ overview after submission	*UI Interaction:* User expects confirmation or summary after completing task. *Source Code:* HTML has no summary or confirmation panel.

Table 4. Overview of implemented usability issues and how they can be detected through UI interaction versus HTML inspection. Bold issue indicates an issue detectable only by UI interaction.

Just Accepted

Usability Issue	Abi	Pat	Tim
1. No clear call to action on landing page		H	S
2. Technical jargon (e.g., "Policyholder E-Claims Portal")	S	S	
3. No step-by-step guide / onboarding	H	H	
4. Important info hidden in tooltips or accordions	S	S	S
5. Buttons are disabled with no error hint	H	H	H
6. Unconventional navigation	H		H
7. Overloaded dashboard w/ unclear task flows			H
8. Only one fixed workflow, not able to go back	H	S	
9. Hover-only feedback	H	H	H
10. No explanations			S
11. Unrecoverable errors (form resets without warning)			
12. Error-prone design	H		H
13. No immediate feedback after clicking "Submit"	S	S	S
14. Strict linear form (no shortcuts for experienced users)			
15. "Advanced" section collapsed by default w/ no hint	H		
16. Ambiguous icon labels	H	H	H
17. No summary/overview after submission	H	H	
N/A: Usability issues not implemented		H	

Table 5. Overview of usability issues that were flagged per GenderMag persona. magenta boxes (H) are issues found by humans, teal boxes (S) by the system and black boxes by human and system.

Inter-rater reliability (N = 2) for matching the generated problem descriptions to the implemented issues was 91%.

Just Accepted

4.1 RQ1: Usability Problem Detection for Generic Interfaces

Overall, humans identified most usability problems related to expectations, preference/behavior, understanding the problem, uncertainty in utilization and request/proposal (see Figure 4, left, see Appendix B for a description of problem types). The LLM-agent also showed similar high identification frequency for identified expectations, understanding the problem, uncertainty in utilization. However, the LLM-agent did not flag preference/behavior (6 human vs. 1 LLM), not seen (2 human vs. 9 LLM) and incorrect utilization (0 humans vs. 4 LLM) as often as the humans. For the unique problems (see Figure 4, right), we can see that the LLM-agent identifies unique problems across all problem categories. The human evaluators identify unique problems only for expectations (7), uncertainty in the utilization (3), request/ proposal (3) and preference/ behavior (4) (Appendix C presents all the problem descriptions of the unique problems identified by humans and LLM-agent.). Generally, the distribution of unique problem detection across the personas is similar except for Abi in the expense tracker 1 interface (Figure 1a), where the Abi agent did not identify a unique problem (see Appendix C, Table 14).

Our qualitative analysis of the unique problem descriptions of the generic interfaces reveals strengths of humans and LLMs that complement each other. Across all personas and interfaces the LLM-agents point out layout issues and overlooked elements (e.g., *"'Edit' and 'Delete' buttons are small and closely positioned, which might lead to accidental clicks or overlooking their visibility"* (Table 14, Appendix C), *"The 'Update Expense' and 'Cancel' buttons are initially hidden, which may lead to Abi not realizing they can edit or cancel expense entries*

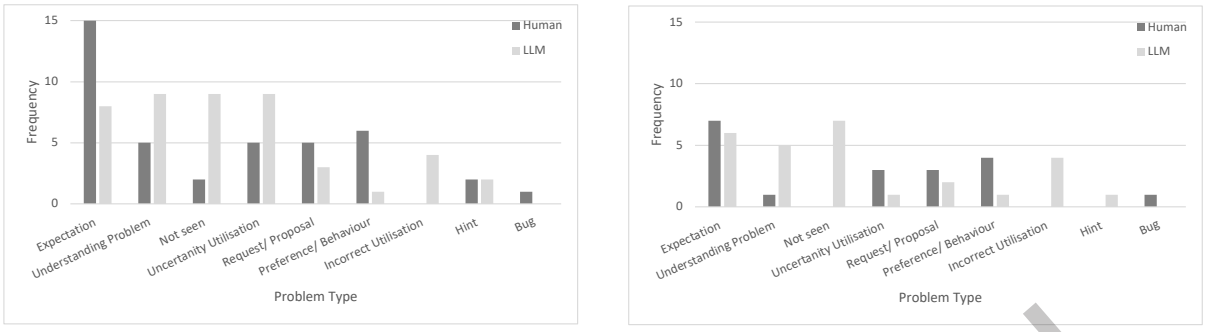


Fig. 4. Frequency of Problem Types Human vs. System: Total Number (**left** and unique **right**). For the total frequency of human-identified problems, we count some problems more than once because participants identified them with more than one persona.

without clear navigation cues." (Table 15, Appendix C), *"The update and cancel buttons are hidden during the initial use of the expense form"* (Table 15, Appendix C). Next, the LLM-agents identify insufficient system feedback e.g., *"If a user attempts to add an empty task, there is no error message or feedback explaining why the action didn't work."* (Table 16, Appendix C), *"There is no visual feedback confirming that a new task has been added successfully"*(Table 16, Appendix C). Also, the LLM-agents identify a lack of clarity in affordance, e.g., *"The 'Add' button does not clearly imply that it will add the typed task to the list."* (Table 16, Appendix C), *"When adding an issue, Pat noticed that the category does not provide a default"* (Table 15, Appendix C). Lastly, the LLM-agents address expectation mismatches such as *" Tim wants a dropdown in the category field"* (Table 16, Appendix C), *"Pat might expect an 'Are you sure you want to delete?'"* (Table 15, Appendix C). Taken together, the unique LLM-agents outputs reflect logical, structure-based usability issues.

In contrast, the unique usability problems reported by the human participants uncover experiential and affective insights. For example, they point out ambiguities such as button labeling *("Confused when editing an output because edit button is blue and not green or gray"* (Table 14, Appendix C), missing information *"the lack of information on the button and uncertainty about what "amount" means stops them from adding a new expense."* (Table 14, Appendix C). Also, they express emotional reactions such as "triggers a negative feeling" (Table 16, Appendix C), *"Abi is afraid that entries will be lost"* (Table 15, Appendix C). Lastly, they address learning and scenarios such as *"Tim would waste too much time trying out all the functions"* (Table 16, Appendix C).

4.2 RQ1, RQ3: Gender-Inclusive Usability Problem Detection for the Intentionally Flawed Interface

Table 6 reports F1 scores that reflect the overlap between human and LLM detections with the set of GenderMag related usability issues implemented in the intentionally flawed interface (see Figure 1d). Both human and LLM achieved moderate F1 scores across all personas. For Abi, humans performed better than the LLM (0.76 vs. 0.43) as they were more effective at identifying true issues without introducing many false positives. For Pat and Tim, LLM and human produced similar F1 scores (0.59 vs. 0.45 for Pat and 0.64 vs. 0.52 for Tim). Overall, humans and LLM have high precision scores concerning the detection of usability issues for all personas. Yet, Cohen's κ was negative for all personas, showing that the LLM misses issues the humans flag and the other way around.

4.3 RQ2: Severity and Relevance Ratings of the Generic and Intentionally Flawed Interfaces

For all interfaces the agent system identified a similar amount of problems as the humans (see sums in Table 7). Several comparisons achieved low statistical power (see Table 8). The only interface with adequate power was

Just Accepted

Metric		Abi	Pat	Tim
True Positives (TP)	Human	11	7	8
	LLM	5	5	6
False Positives (FP)	Human	1	1	0
	LLM	1	0	0
False Negatives (FN)	Human	6	9	9
	LLM	12	12	11
True Negatives (TN)	Human	0	0	0
	LLM	0	0	0
Precision	Human	0.92	0.88	1.00
	LLM	0.83	1.00	1.00
Recall	Human	0.65	0.44	0.47
	LLM	0.29	0.29	0.35
F1 Score	Human	0.76	0.59	0.64
	LLM	0.43	0.45	0.52
Cohen's κ	Human–LLM Agreement	-0.26	-0.27	-0.20

Table 6. Performance Comparison between Human and LLM Evaluators across Personas. True negatives are 0 across all conditions because the evaluation focused on identifying implemented usability problems rather than confirming the absence of problems. Usability issues that we considered "True" for F1 scores, are listed in Table 5.

Just Accepted

the To-Do List. For this generic interface, agents assigned significantly higher ratings for severity and relevance compared to the workshop participants after controlling the False Discovery Rate using the Benjamini-Hochberg Method [52] (see Table 7 and Table 8). However, for the intentionally flawed interface, the human and LLM-agents produced no significant differences in severity and relevance ratings (Table 7). Notably, this comparison was substantially underpowered (Table 8), increasing the risk of Type II error. Accordingly, this non-significant findings should not be interpreted as evidence of equivalence or true null effects. The effect sizes (see Table 8) are high for the generic interfaces (e.g., Cramér's V = 0.71 for severity effect of Expense Tracker II) and low to medium for the intentionally flawed interface (Cramér's V = 0.17 for severity and 0.22 for relevance).

4.4 RQ3: Persona Based Problems for the Generic Interfaces

Generally, the distribution of unique problems per persona is similar and the LLM-agent identifies more unique issues (Human Abi: 2, Human Pat: 1, Human Tim: 1, LLM Abi: 4, LLM Pat: 3, LLM Tim: 4, Appendix C, Table 16). On a facet level, the distribution of human vs. LLM-agent for unique problems is broadly similar (see Figure 5). The highest divergence between human and LLM agent evaluations were for the persona Tim (see Appendix C, Table 15). GenderMag characterizes Tim with an experimental and risk-tolerant style. However, the LLM flags that *"there is no clear message or help prompt explaining to users how to use the expense tracker effectively. Users will have to figure out how things work by trial and error."* (see Appendix C, Table 15). Arguably, this might not be an issue for Tim, as Tim likes to tinker. For Abi and Pat, human participants and LLM-agents semantically aligned more in simulating the tenancy for comprehensive information processing style and lower self-efficacy

Interface	Sev.	H	S	p	BH-p	Rel.	H	S	p	BH-p
To-Do List	S1	6	5	0.02*	0.04*	R1	6	3	0.01*	0.04*
	S2	0	8			R2	0	6		
	S3	0	1			R3	0	4		
						R4	0	1		
		6	14				6	14		
Expense Tracker I	S1	16	5	<0.01*	0.04*	R1	9	1	<0.01*	0.04*
	S2	1	10			R2	8	5		
	S3	0	0			R3	0	8		
						R4	0	1		
		17	15				17	15		
Expense Tracker II	S1	15	3	<0.01*	0.04*	R1	3	1	0.43	0.55
	S2	2	13			R2	11	8		
	S3	1	0			R3	4	7		
						R4	0	0		
		18	16				18	16		
Insurance Claim	S1	16	4	0.49	0.49	R1	1	2	0.41	0.55
	S2	18	10			R2	12	4		
	S3	5	3			R3	18	6		
						R4	8	5		
		39	17				39	17		

Table 7. Comparison of Fisher's Exact Test results for Severity and Relevance ratings by humans and system. We controlled the false discovery rate using the Benjamini–Hochberg method [52].

Interface	Severity V (95% CI)	Power	Relevance V (95% CI)	Power
To-Do List	.59 [.05, .61]	.66	.72 [.10, .68]	.86
Expense Tracker I	–	–	.71 [.08, .54]	.10
Expense Tracker II	.71 [.04, .47]	.10	–	–
Insurance Claim	.17 [.00, .41]	.19	.22 [.00, .46]	.26

Table 8. Cramér's V effect sizes with 95% confidence intervals and power estimates for Severity and Relevance ratings.

Just Accepted

(e.g., see Appendix C). Both human and LLM evaluators highlighted, for example, Abi's usability problem due to a missing confirmation message, reflecting the risk averse style (see Table 9).

4.5 RQ3: Persona Based Problems for the Intentionally Flawed Interface: Gender-Inclusive Issues

To assess the qualitative alignment of human evaluators and the LLM-agent for the intentionally flawed interface, we analyzed which usability problems humans and LLM-agents identified across the three personas. Performance metrics (see Table 6) showed that most issues that where flagged by humans and LLMs were part of the implemented set of issues (see precision metrics). However, recall metrics where lower (0.29–0.35 for the LLM and

#	Source	Problem Description	Severity	Relevance
1	Human	*When adding an output, they [Abi] do not notice if something is wrong.*	S1	R2
	System	*After clicking the 'Add' button, users receive no confirmation that their entry has been added successfully. This could lead to uncertainty about whether the action worked.*	S2	R3
2	Human	*Confused when adding an output because the green button is not labelled.*	S1	R2
	System	*The 'Add' button's text color matches its background, making it confusing and not immediately recognizable as a button for adding expenses.*	S2	R3
3	Human	*When editing an output, he [Tim] is confused that the output is edited directly [...]*	S1	R2
	System	*Users might expect the 'Edit' button to lead to a new interface or dialog for editing. Instead, the edit function happens inline without additional guidance.*	S2	R2

Table 9. Qualitative and Exemplary Descriptions of the same Issue Identified by Humans and the System with Different Severity and Relevance Ratings

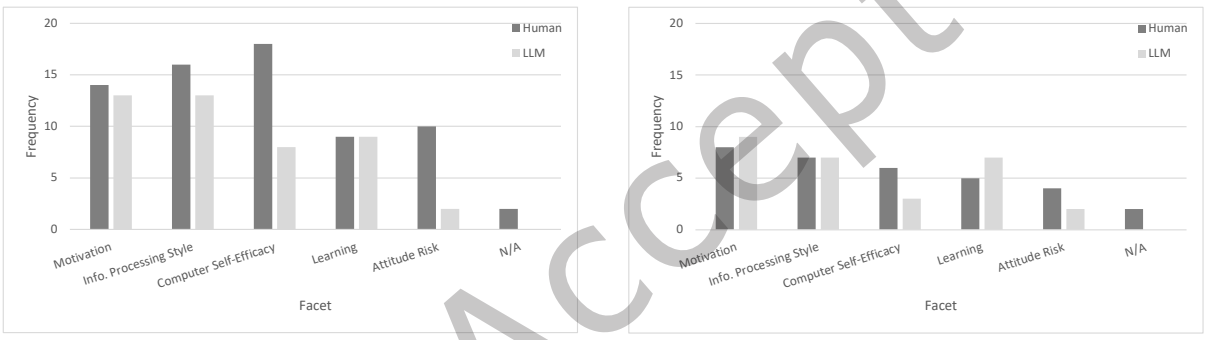


Fig. 5. Problem frequency per-facet total (**left**) and unique (**right**). Note that humans partly documented more than one facet per problem.

Just Accepted

0.44–0.65 for human), showing that each evaluator missed issues that were implemented. The F1 scores for the LLM (0.43–0.53) are lower compared to the human evaluators (0.59–0.76). Cohen's κ were negative across all personas (-0.26 to -0.20), showing that human and LLMs agree lower than chance. However, the Jaccard similarity between human and LLM-identified issues was medium to high (Abi: 0.39, Pat: 0.63 and Tim: 0.75), which shows that despite low agreement, the evaluators together flag a larger portion of the implemented issues (see Table 6 and the following example calculation of the Jaccard similarity).

Jaccard Similarity for Tim. For the Tim persona, the values from Table 6 are:

$$\text{Human: } TP = 8,\ FP = 0,\ FN = 9,$$
$$\text{LLM: } TP = 6,\ FP = 0,\ FN = 11.$$

From these numbers, we derive the number of issues flagged in terms of the human set H and the LLM set L (both are true positives plus false positives):

$$|H| = TP_{\text{Human}} + FP_{\text{Human}} = 8 + 0 = 8,$$

$$|L| = TP_{\text{LLM}} + FP_{\text{LLM}} = 6 + 0 = 6.$$

The definition for Jaccard similarity between two sets (human = H and LLM = L) is:

$$J(H, L) = \frac{|H \cap L|}{|H \cup L|} = \frac{|H \cap L|}{|H| + |L| - |H \cap L|}.$$

With $|H| = 8, |L| = 6, |H \cap L| = 6$, we observed a Jaccard index of 0.75.

For Abi, human and LLM evaluators identified only two issues (problem 1: Vague landing page and problem 10: No explanations), which represents the Cohen's κ (see Table 6). For Pat, there is the same overlap for problem 10, and for Tim, problem 8 (fixed workflow) overlapped.

One evaluator type systematically overlooked some usability problems. For instance, humans never identified problem 2 (technical jargon) but the LLM flagged it for Abi and Pat. In contrast, human evaluators recognized interaction-level problems 5 (buttons are disabled with no error hint) and 6 (unconventional navigation) across all personas, which the LLM-agents never detected.

Both human and LLM-agents did not identify problem 14 (No shortcuts for experienced users) and problem 11 (Unrecoverable errors), which would affect Tims high computer self-efficacy for problem 14. For the persona Tim, a workshop participant voiced that it was difficult for her to embody Tim, as she does not identify with him. We hypothesized that problem 11 would be difficult for the LLM-agents to identify because it required contextual knowledge that goes beyond the interface. Interestingly, the human participants also did not identify this problem.

4.6 Effect of Cognitive Facet Alignment of Participants and Personas

Participant	Risk	Self-Efficacy	Motivations	Learning	Info. Processing
A	Abi-like	Tim-like	neutral	Tim-like	Abi-like
B	Abi-like	Tim-like	Abi-like	Abi-like	neutral
C	Abi-like	Tim-like	Abi-like	Abi-like	neutral
D	neutral	Tim-like	neutral	neutral	Tim-like
E	Tim-like	Tim-like	neutral	neutral	neutral
F	Abi-like	Tim-like	neutral	neutral	Abi-like
G	neutral	Tim-like	neutral	neutral	Abi-like
H	neutral	Tim-like	neutral	neutral	Abi-like
I	neutral	Tim-like	Tim-like	Tim-like	neutral

Table 10. Persona-related cognitive facets for participants A–I.

Across Risk, Motivations, Learning and Information processing the participants represent either Abi-like, Tim-like or neutral facets. For self-efficacy all participants are Tim-like. Across workshop participants, most have Tim-like and Abi-like facets except for participant I, who has only Tim-like and neutral facets. Notably, Participant C voiced the difficulty embodying Tim and has an Abi-like facets in Risk, Motivations and Learning. To explore whether participants' cognitive styles influenced their interaction with the personas, we examined the alignment between each participant's facet profile and the GenderMag personas (see Table 10 and Table 18). We first tested whether participants used facets differently when they were aligned versus misaligned with the

Just Accepted

persona they enacted. A Wilcoxon signed-rank test shows no statistically significant difference between aligned and misaligned facet counts within participants ($W = 3$, $p = .46$, $r = .37$, $r_{\text{rb}} = -.40$, 95% CI $[-0.50, 0.50]$.) A Chi-square test of independence ($\chi^2(4) = 0.84$, $p = .93$, $V = .14$, 95% CI $[.13, .54]$) showed no evidence of an association between misalignment and which facet participants used.

5 Discussion

The present study examined whether, a) LLM-agents and humans identify the same persona-specific usability problems (RQ1), b) assign similar ratings concerning severity, relevance, and problem types (RQ2), and c) find gender-related issues (RQ3). To examine these research questions, we conducted a mixed-method study involving workshops with human usability experts and simulations with LLM-agents.

The results of our study show that LLM-agents identify different gender-related usability issues to human evaluators. Based on these results, **we claim that LLM-agents have the potential to complement humans performing inclusive usability evaluations**.

5.1 Overlap between Persona-Specific Usability Problems identified by LLM-agents and Humans (RQ1)

For the generic interfaces, we observed a broad agreement in the number and types of problems between LLM-agents and human evaluators (Figure 4). This balance between consistency and diversity in problem focus supports recent work by Ge et al. [21], and strengthens the case for integrating agent evaluations into iterative design workflows, especially when human capacity is limited or when broad coverage across user types is required [37, 46]. The use of LLM-agents is promising for early-stage evaluations, as we have tested simple interfaces in this study.

Although we have found a broad agreement, there are several unique issues identified either by human or LLM evaluators (Figure 4, right). We hypothesize that the most important cause for systematic difference in ratings is the nature of the input for the agent system (e.g., the HTML code) allows it to assess the complete functionality of the web interface and find bugs that are not necessarily visible to humans from interaction. While using HTML code as input can be especially valuable for web developers, as it allows the agent system to pinpoint specific lines of code that need fixing, this approach lacks dynamic insights into usability. As such, this might cause the agent system's high efficiency concerning technical usability problems and its relatively low performance in finding non-technical problems (e.g., understanding). Multi-modal approaches, such as integrating still images or videos, in addition to the code offer one step in this direction. Even more dynamic usability evaluations could incorporate the agent interacting with a simulation of the interface using reinforcement learning, which is an active area of research [20, 54].

5.2 Need for Calibration and Interpretation (RQ2)

The agent system assigns higher severity and relevance ratings in generic interfaces and we see some divergence of human vs. LLM-agent in found problem types (see Figure 4, right), which shows the need for calibration. Differences may occur, for example, because humans may integrate experience with developing and using web interfaces with their severity and relevance judgments. Also, difference in ratings could be from interpreting the definitions of ratings in the codebook. Humans may assign different levels of urgency or impact to descriptions like "delays users considerably" and "should be processed ASAP". This finding that evaluators integrate their experiences has been a long standing observation by Herzum et al., which they call the Evaluator Effect [27]. Calibrating the LLM-agent through few-shot examples or chain-of-thought prompting [57] may help align its judgments with human expectations. Few shot prompting could help calibration, but carries some risks since there is a competition for token count, and thus attention, from a variety of sources like personas, evaluation method,

Just Accepted

or interface source. Similarly, complicated prompting strategies (chain-of-thought [57], tree-of-thought [61], graph-of-thought [2]) tend to require large amounts of properly formatted data. However, stepping through the logic of identifying usability problems is a relatively data-poor environment, which also limits our ability to fine-tune.

5.3 LLM-agents' ability to identify gender-related usability issues (RQ3)

Concerning our results with respect to gender-specific issues, we found differences between LLM-agents and human evaluators. While humans and LLMs achieved moderate F1 scores against the predefined set of gender-related usability issues for the insurance claim sheet (Figure 1d), the agreement between human and LLM-agent in predefined problems identified was low (Cohen's Kappa ranged between -0.27 and -0.20). The agent system and human evaluators attended to different types of issues. The LLM agent finds structural and textual issues, likely due to analysis of HTML code, while humans find interaction-level problems. Prior work has found that agent systems provide breadth, consistency, and scalability [9], while human evaluators contribute depth, emotional nuance, and contextual judgment [23]. Such divergence emphasizes its role as a complementary tool.

The participants' expertise and alignment with the GenderMag personas can impact the validity of the results. For example, software engineering roles tend to align more closely with Tim's selective information-processing style [34]. Conversely, people in IT support are more likely to embody Abi's comprehensive processing approach [7].

Therefore, we posit that **researchers that conduct GenderMag workshops should assess the cognitive styles of workshop participants**. However, the existing literature only provides preliminary qualitative indications of alignment effects, which future work should explore.

These findings might inform how persona and evaluator facet alignment changes the validity of the GenderMag walkthroughs. This dependency of the results from the participant sample underlines the importance of contextual relevance: while Tim's style may lead to faster task performance, Abi's style may yield greater accuracy, which might be more or less important for certain interfaces and tasks. In our study we integrated facets of GenderMag personas by embedding cognitive styles to address gender-related usability issues, but changing its components could reveal other inclusiveness challenges like accessibility and cultural differences in cognitive styles [5, 58]. This approach of modeling a perspective can not only improve inclusiveness but the architecture is reusable in various domains that have personas. For example, we could integrate consumer configuration to model energy-use decision-making patterns [18].

5.4 Diverging Tasks of Human and LLM

Humans analyzed the UI by interacting with it directly while inhabiting the persona, interpreting its cognitive facets and making evaluation decisions based on experiential, emotional, and contextual reasoning. This is analogous to the software engineering concept of *dynamic analysis*, which is *"testing and evaluation of an application during runtime" [13]*. In contrast, the LLM-agents analyzed the structure of the HTML source code rather than assessing dynamic interactions. This is analogous to the software engineering concept of *static analysis*, which is *"testing and evaluation of an application by examining the code without executing the application"* [13]. Importantly, static and dynamic analyses surface different issues, since the former reveals *compile-time* errors, while the dynamic analysis reveals *runtime* errors. The advantage of dynamic analysis is that *"It reveals subtle defects or vulnerabilities whose cause is too complex to be discovered by static analysis"* [13]. Meanwhile, static analysis an advantage in that *"It examines all possible execution paths and variable values, not just those invoked during execution"* [13].

Just Accepted

Work in software engineering attempts to converge the static and dynamic perspectives, with the goal of gaining benefits from both [17]. In our context, the workflow could take several forms beyond the divide-and-conquer approach this paper adopts (LLMs do the static analysis; humans do the dynamic). First, GenderMag workshops could incorporate some form of static analysis, such as a code review. While code reviews may not be appropriate for all stakeholder groups, our results indicate that it might diversify the perspectives and thus the findings. Further, such an adaptation would be possible because GenderMag is a meta-method, requiring contextual adaptation to the current unit of analysis.

A second option to incorporate static analysis into GenderMag would be through adding an LLM-based analysis like the one found in this paper. This could occur by first running the LLM-based analysis and then using the outputs as an additional artifact in the human-based analysis. The human GenderMag session focuses on validating, contextualizing and extending these findings through dynamic interaction and experiential reasoning.

A third option to hybridize static and dynamic perspectives takes a different approach, giving the system the ability to perform dynamic analysis, which could take several forms. These forms range from providing textual program traces or video recordings of the software to live interaction via agentic AI controlling the interface. Each of these combination approaches bear different advantages and drawbacks, in terms of computational cost and requirements for user sophistication, which future work can explore.

Each of the three aforementioned recommendations would offer different benefits and drawbacks. For example, analyzing logs requires code to exist, be relatively bug free, and relatively feature complete. However, perhaps video recordings could come from recording a low-fidelity prototype. Similarly, pseudocode would suffice for some aspects of code review. As such, while the benefit of this type of analysis arises early in development, continued analysis post-deployment should yield continued benefits.

Similarly, the different task modalities of the human and LLM are understandable through the distinction between instrumental and hedonic product qualities [26]. The instrumental product qualities describe how effectively and efficiently a system supports users in achieving their goals. In our case, we intentionally hindered this quality (by e.g., clarity of terminology). Hedonic product qualities refer to users' experiential, emotional and value-laden perceptions for a system. For example, when comparing electric and gas-powered cars, instrumental qualities could concern refueling time, while hedonic qualities could relate to perceived environmental responsibility. Therefore, usability engineers can distinguish products not only by their functional performance by applying measures such as the meCUE model [40], where instrumental qualities map to pragmatic modules and hedonic qualities correspond to non-instrumental modules. In our study, the LLM-agents attended to instrumental qualities such as technical jargon or missing feedback mechanisms. Human evaluators also voice affective responses such as insecurity, confusion or negative emotional reactions. Notably, hedonic qualities are difficult to ground truth, as emotional user experience values *emerge* from situated interaction and subjective interpretation. The computational architecture presented in this paper reflects the GenderMag method via cognitive walkthrough but does not simulate interactional dynamics or affective responses that arise during interaction. Therefore, future work should explore how to integrate AI and human perspectives for additional meCUE modules, as well as whether and how interactive agent architectures are capable of engaging with UIs to approximate those modules.

5.5 Ground Truth in Persona-based Usability Testing

Again, humans identified emotional and interactive aspects, whereas the agents could identify guidance-related usability issues. Here, we would like to acknowledge our epistemic stance: the value of a complementary perspective. Due to subjective aspects of persona-based usability evaluation, it is difficult to pinpoint a ground truth. What one considers a true usability issue depends on interpretation, context and prioritization. Therefore, our general research goal is not to claim that either an agents or human evaluators perspective is correct or that we have a definitive reference for grounding. Instead, we interpret or reference list of usability issues as "rigorous,

Just Accepted

but noisy", so that we have a basis to compare and contrast the human and AI perspectives on the process of discovering such issues. Our goal is to leverage the Evaluator Effect [27] to expand the evaluators field of view, introducing new ideas and perspectives that contribute to a more complete picture of an interfaces strengths and weaknesses and a users' experience.

6 Threats to Validity

The current study has several limitations that readers need to keep in mind. In reviewing these threats, we follow the framework from Yin [62].

6.1 Internal Validity and Reliability

This study had 6 human participants per GenderMag workshop. Although this number should suffice according to Nielson [43], we should replicate study results with other samples. Post hoc power calculations based on the observed effect sizes suggest that reliably detecting the small effects observed for the intentionally flawed interface would require substantially larger samples. Given our small sample size, this increases the risk of missing real effects (Type II error). Therefore, non-significant results should not be taken as proof that no differences exist. Following guidance from the American Statistical Association [56], we interpret the statistical results together with effect sizes, confidence intervals, and descriptive patterns, and treat them as exploratory complements to the qualitative findings. Also, the three workshops have different structures, which limits the comparability between the generic (workshop 1) and intentionally flawed interfaces (workshop 2 and 3). The first workshop occurred in a laboratory setting and analyzed three interfaces that only had one screen. The second and third workshop were remote and analyzed a more complex interface. The participants of the second workshop were a subset of the participants from the first workshop. To mitigate potential learning effects concerning the persona facets, the third workshop included an independent sample. Their outputs mirrored the patterns observed in workshop 2, suggesting that any learning effects in the subset group were minimal. Nonetheless, we cannot fully exclude the possibility of residual familiarity influencing their reasoning.

In the first workshop each participant embodied each persona to control for effects regarding similarity of the personas' cognitive facets to their own. In the second and third workshop, each participant embodied only one persona, which means that the participant and persona pairing could affect the results more.

From the AI perspective, we sought to increase reliability by performing stability testing (see Table 3 and Figure 3).

6.2 External Validity

First, we tested interfaces via HTML code, which limits the generalizability of the results to more complex and dynamic interfaces (e.g., those with JavaScript interactions). Future studies should explore interfaces with dynamic artifacts and integrate multi-modal approaches. Next, due to clarity we solely focus on usability issues. To understand user experience (UX), which is a broader concept, requires integrating perception of non-instrumental interface qualities, such as visual aesthetics and commitment, as well as behavioral consequences, such as intention to use [53]. For this study we focused on the evaluation of LLM-agents in gender-inclusive usability issues, next steps should integrate more aspects of UX. Furthermore, we employed one prominent human centered design method: personas. In particular, we combined this with heuristic evaluation. Although these are prominent approaches, future research should also evaluate the alignment of LLM-agents and human evaluators for other expert-based methods, such as workflow analysis, and user-based methods, such as thinking aloud. Lastly, expanding the calibration of the LLM-agent by the scope of user types to include a broader range of emotional, cognitive, cultural, and accessibility considerations allows these agent systems to address a wider array of usability challenges.

Just Accepted

6.3 Construct Validity

A threat to construct validity is our operationalization of the GenderMag facets (Table 1) into implemented usability issues (Table 5). The 17 seeded issues were theoretically grounded in prior GenderMag literature and related findings on cognitive styles, but they were not empirically pre-validated through pilot testing with users whose facet distributions match the personas. Thus, the implemented issues represent theoretically derived hypotheses about gender-related usability challenges rather than empirically confirmed effects. Personas are a tool for usability engineers to understand and communicate about users to a variety of other stakeholders. In our study, we measured performance by comparing problem identification and triage from multiple sources, but this is only a proxy measure. Additionally, the proxy measure has some subjectivity, which is why we followed an IRR procedure with multiple raters, which had uneven agreement (e.g., Pat had only 71% agreement). This lower agreement compared to Abi (94%) and Tim (76%) suggests that the mapping of implemented issues to Pat's facets may be more ambiguous, reflecting variability in how certain cognitive characteristics are operationalized. The value added by the aforementioned understanding and communication is multifaceted.

7 Conclusion

The present study shows the potential of an agent system to address automatization and scalability challenges in persona-based inclusive usability testing. Integrating GenderMag facets into LLM-agents for usability evaluation, enabled the agent to identify usability issues that human-led GenderMag evaluations missed. Quantitatively, the agent system detected more functional issues, which human participants overlooked or did not report. In contrast, human evaluators captured emotional and interactive aspects of usability problems that the LLM-agent missed.

Based on these findings, we conclude that, especially for gender-inclusive usability issues, the LLM-agents do not capture the same aspects as human evaluators do when applying personas. Agent systems are less labor-intensive and provide the opportunity to scale across many personas. However, they lack the nuanced judgment and contextual sensitivity, and experience that informed usability engineers bring to the process. Also, we promote usability engineers to survey their workshop participants in order to control for cognitive styles when applying personas. A combination of humans and LLM-agents allows usability engineers to leverage the agent system's automatization and consistency while applying their expertise to interpret outputs and account for emotional, cultural, and contextual factors that automated methods may overlook. The findings suggest that agent systems are a valuable tool for human-led usability evaluations as it broadens the insights into the strengths and weaknesses of an interface.

Together, complementary human and agent approaches can pave the way towards a more inclusive HCI.

Just Accepted

Acknowledgments

We would like to acknowledge the participants of the workshop, whose feedback provided valuable insights; Jonas Scholz and Josephine Halama for rating the qualitative data and Julian Bornemeier for feedback on the analysis. We are thankful to Dr. Margaret Burnett for her feedback on a draft of this manuscript. We would also like to thank the editor and reviewers for their insightful comments which strengthened our papers rigor.

References

[1] Andrew Anderson, Jimena Noa Guevara, Fatima Moussaoui, Tianyi Li, Mihaela Vorvoreanu, and Margaret Burnett. 2024. Measuring User Experience Inclusivity in Human-AI Interaction via Five User Problem-Solving Styles. *ACM Transactions on Interactive Intelligent Systems* 14, 3 (Sept. 2024), 1–90. https://doi.org/10.1145/3663740

[2] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michał Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: solving elaborate problems with large language models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications*

of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'24/IAAI'24/EAAI'24). AAAI Press, Article 1972, 9 pages. https://doi.org/10.1609/aaai.v38i16.29720

[3] Inc. BigBlueButton. [n. d.]. BigBlueButton. https://bigbluebutton.org/. Accessed: 2025-08-22.

[4] Jonathan Bratt and Jon Harmon. 2019. RBERT: Implementation of BERT in R (Version 0.1.11) [R package]. https://github.com/jonathanbratt/RBERT. Retrieved from https://github.com/jonathanbratt/RBERT.

[5] Margaret Burnett. 2020. Doing Inclusive Design: From GenderMag in the Trenches to Inclusive Mag in the Research Lab. In *Proceedings of the International Conference on Advanced Visual Interfaces* (Salerno Italy, 2020-09-28). ACM, 1–6. https://doi.org/10.1145/3399715.3400871

[6] Margaret Burnett, Anicia Peters, Charles Hill, and Noha Elarief. 2016. Finding Gender-Inclusiveness Software Issues with GenderMag: A Field Investigation. 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, 2586–2598. https://doi.org/10.1145/2858036.2858274

[7] Margaret Burnett, Simone Stumpf, Jamie Macbeth, Stephann Makri, Laura Beckwith, Irwin Kwan, Anicia Peters, and William Jernigan. 2016. GenderMag: A Method for Evaluating Software's Gender Inclusiveness. *Interacting with Computers* 28, 6 (Nov. 2016), 760–787. https://doi.org/10.1093/iwc/iwv046

[8] Antonello Ceravola, Frank Joublin, Ahmed R. Sadik, Bram Bolder, and Juha-Pekka Tolvanen. 2024. HyperGraphOS: A Meta Operating System for Science and Engineering. https://doi.org/10.48550/arXiv.2412.04923

[9] C.S. Richard Chan, Charuta Pethe, and Steven Skiena. 2021. Natural language processing versus rule-based text analysis: Comparing BERT score and readability indices to predict crowdfunding outcomes. *Journal of Business Venturing Insights* 16 (2021), e00276. https://doi.org/10.1016/j.jbvi.2021.e00276

[10] Amreeta Chatterjee, Rudrajit Choudhuri, Mrinmoy Sarkar, Soumiki Chattopadhyay, Dylan Liu, Samarendra Hedaoo, Margaret Burnett, and Anita Sarma. 2024. Debugging for inclusivity in online cs courseware: Does it work?. In *Proceedings of the 2024 ACM Conference on International Computing Education Research-Volume 1*. 419–433.

[11] Amreeta Chatterjee, Mariam Guizani, Catherine Stevens, Jillian Emard, Mary Evelyn May, Margaret Burnett, and Iftekhar Ahmed. 2021. AID: An automated detector for gender-inclusivity bugs in OSS project pages. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. IEEE, 1423–1435.

[12] Alan Cooper. 1999. *The Inmates Are Running the Asylum: Why High-Tech Products Drive Us Crazy and How to Restore the Sanity*. Sams, Indianapolis, IN.

[13] Intel Corporation. 2025. Dynamic Analysis vs. Static Analysis. https://www.intel.com/content/www/us/en/docs/inspector/user-guide-windows/2024-2/dynamic-analysis-vs-static-analysis.html. Accessed: 2025-12-19.

[14] Sally Jo Cunningham, Annika Hinze, and David M. Nichols. 2016. Supporting Gender-Neutral Digital Library Creation: A Case Study Using the GenderMag Toolkit. In *Digital Libraries: Knowledge, Information, and Data in an Open Access Society*, Atsuyuki Morishima, Andreas Rauber, and Chern Li Liew (Eds.). Vol. 10075. Springer International Publishing, Cham, 45–50. https://doi.org/10.1007/978-3-319-49304-6_6 Series Title: Lecture Notes in Computer Science.

[15] Digitalzentrum Fokus Mensch. 2025. Fairness und Diversität. https://www.digitalzentrum-fokus-mensch.de/themen/fairness-und-diversitaet?standard=tru&nobots=1&listviewpagecpob_7876=3. Accessed: 2025-12-04.

[16] DIN. 2019. DIN EN ISO 9241-210:2019 Ergonomie der Mensch-System-Interaktion – Teil 210: Menschzentrierte Gestaltung interaktiver Systeme (ISO 9241-210:2019).

[17] Dino Distefano, Matteo Marescotti, Cons Ahs, Sopot Cela, Gabriela Cunha Sampaio, Radu Grigore, Akos Hajdu, Timotej Kapus, Ke Mao, and Thibault Suzanne. 2024. Enhancing Compositional Static Analysis with Dynamic Analysis. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering* (Sacramento, CA, USA) *(ASE '24)*. Association for Computing Machinery, New York, NY, USA, 2121–2129. https://doi.org/10.1145/3691620.3695599

[18] Jonathan Dodge, Michael Hilton, Ronald A. Metoyer, Josie Hunter, Karl Smeltzer, Catharina Vijay, and Andrew Atkinson. [n. d.]. Deriving Age Diverse Personas from a Participatory Design Study on Home Electricity Feedback. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Denver Colorado USA, 2017-05-06). ACM, 959–968. https://doi.org/10.1145/3027063.3053354

[19] Peitong Duan, Jeremy Warner, Yang Li, and Bjoern Hartmann. 2024. Generating Automatic Feedback on UI Mockups with Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 6, 20 pages. https://doi.org/10.1145/3613904.3642782

[20] Daniel Gaspar-Figueiredo, Marta Fernández-Diego, Ruben Nuredini, Silvia Abrahao, and Emilio Insfran. 2024. Reinforcement Learning-Based Framework for the Intelligent Adaptation of User Interfaces. In *Companion Proceedings of the 16th ACM SIGCHI Symposium on Engineering Interactive Computing Systems* (Cagliari, Italy) *(EICS '24 Companion)*. Association for Computing Machinery, New York, NY, USA, 40–48. https://doi.org/10.1145/3660515.3661329

[21] Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling Synthetic Data Creation with 1,000,000,000 Personas. https://doi.org/10.48550/ARXIV.2406.20094 Version Number: 2.

[22] Emily Gonzalez-Holland, Daphne Whitmer, Larry Moralez, and Mustapha Mouloua. 2017. Examination of the Use of Nielsen's 10 Usability Heuristics & Outlooks for the Future. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 61, 1 (Sept.

Just Accepted

2017), 1472–1475. https://doi.org/10.1177/1541931213601853
[23] Julián Grigera, Alejandra Garrido, José Matías Rivero, and Gustavo Rossi. 2017. Automatic detection of usability smells in web applications. *International Journal of Human-Computer Studies* 97 (Jan. 2017), 129–148. https://doi.org/10.1016/j.ijhcs.2016.09.009
[24] Md Montaser Hamid, Amreeta Chatterjee, Mariam Guizani, Andrew Anderson, Fatima Moussaoui, Sarah Yang, Isaac Escobar, Anita Sarma, and Margaret Burnett. 2024. How to measure diversity actionably in technology. In *Equity, diversity, and inclusion in software engineering: Best practices and insights*. Apress Berkeley, CA, 469–485.
[25] Md Montaser Hamid, Fatima A Moussaoui, Jimena Noa Guevara, Andrew Anderson, Puja Agarwal, Jonathan Dodge, and Margaret Burnett. 2026. Inclusive Design of AI's Explanations: Just for Those Previously Left Out? *ACM Transactions on Interactive Intelligent Systems* 16, 1 (2026), 1–44.
[26] Marc Hassenzahl. 2018. The thing and I: understanding the relationship between user and product. In *Funology 2: from usability to enjoyment*. Springer, 301–313.
[27] Morten Hertzum, Rolf Molich, and Niels Ebbe Jacobsen. 2014. What you get is what *you* see: revisiting the evaluator effect in usability tests. 33, 2 (2014), 144–162. https://doi.org/10.1080/0144929X.2013.783114
[28] Charles G. Hill, Maren Haag, Alannah Oleson, Chris Mendez, Nicola Marsden, Anita Sarma, and Margaret Burnett. 2017. Gender-Inclusiveness Personas vs. Stereotyping: Can We Have it Both Ways?. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6658–6671. https://doi.org/10.1145/3025453.3025609
[29] Megan Hofmann, Devva Kasnitz, Jennifer Mankoff, and Cynthia L Bennett. 2020. Living Disability Theory: Reflections on Access, Research, and Design. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event Greece, 2020-10-26). ACM, 1–13. https://doi.org/10.1145/3373625.3416996
[30] Nien-Lin Hsueh, Hsuen-Jen Lin, and Lien-Chi Lai. 2024. Applying Large Language Model to User Experience Testing. *Electronics* 13, 23 (Nov. 2024), 4633. https://doi.org/10.3390/electronics13234633
[31] Melody Y. Ivory and Marti A Hearst. 2001. The state of the art in automating usability evaluation of user interfaces. *Comput. Surveys* 33, 4 (Dec. 2001), 470–516. https://doi.org/10.1145/503112.503114
[32] Paul Jaccard. 1908. Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.* 44 (1908), 223–270.
[33] Ilkka Kaate, Joni Salminen, Soon-Gyo Jung, Trang Thi Thu Xuan, Essi Häyhänen, Jinan Y. Azem, and Bernard J. Jansen. 2025. "You Always Get an Answer": Analyzing Users' Interaction with AI-Generated Personas Given Unanswerable Questions and Risk of Hallucination. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*. ACM, Cagliari Italy, 1624–1638. https://doi.org/10.1145/3708359.3712160
[34] Tanjila Kanij, John Grundy, Jennifer McIntosh, Anita Sarma, and Gayatri Aniruddha. 2022. A new approach towards ensuring gender inclusive SE job advertisements. In *Proceedings of the 2022 ACM/IEEE 44th International Conference on Software Engineering: Software Engineering in Society*. ACM, Pittsburgh Pennsylvania, 1–11. https://doi.org/10.1145/3510458.3513016
[35] A. Baki Kocaballi. 2023. Conversational AI-Powered Design: ChatGPT as Designer, User, and Product. https://doi.org/10.48550/ARXIV.2302.07406 Version Number: 1.
[36] Thomas Kosch and Sebastian Feger. 2024. Risk or Chance? Large Language Models and Reproducibility in HCI Research. *Interactions* 31, 6 (Nov. 2024), 44–49. https://doi.org/10.1145/3695765 arXiv:2404.15782 [cs].
[37] María Inés Laitano, Roelvis Ortiz-Núñez, and Yudayly Stable-Rodríguez. 2024. User participation in digital accessibility evaluations: Reviewing methods and implications. *Education for Information* (Sept. 2024), 1–16. https://doi.org/10.3233/EFI-240014
[38] Vivian Liu and Lydia B Chilton. 2022. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–23.
[39] Yuxuan Lu, Bingsheng Yao, Hansu Gu, Jing Huang, Jessie Wang, Laurence Li, Jiri Gesi, Qi He, Toby Jia-Jun Li, and Dakuo Wang. 2025. UXAgent: An LLM Agent-Based Usability Testing Framework for Web Design. (2025). https://doi.org/10.48550/ARXIV.2502.12561 Publisher: arXiv Version Number: 2.
[40] Michael Minge, Manfred Thüring, Ingmar Wagner, and Carina V Kuhr. 2016. The meCUE questionnaire: a modular tool for measuring user experience. In *Advances in Ergonomics Modeling, Usability & Special Populations: Proceedings of the AHFE 2016 International Conference on Ergonomics Modeling, Usability & Special Populations, July 27-31, 2016, Walt Disney World®, Florida, USA*. Springer, 115–128.
[41] Emerson Murphy-Hill, Alberto Elizondo, Ambar Murillo, Marian Harbach, Bogdan Vasilescu, Delphine Carlson, and Florian Dessloch. 2024. GenderMag Improves Discoverability in the Field, Especially for Women: An Multi-Year Case Study of Suggest Edit, a Code Review Feature. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*. ACM, Lisbon Portugal, 1–12. https://doi.org/10.1145/3597503.3639097
[42] Jakob Nielsen and Robert L. Mack. 1994. *Usability Inspection Methods*. Wiley.
[43] Jakob Nielsen and Rolf Molich. 1990. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems Empowering people - CHI '90*. ACM Press, Seattle, Washington, United States, 249–256. https://doi.org/10.1145/97243.97281

Just Accepted

[44] Asil Oztekin, Dursun Delen, Ali Turkyilmaz, and Selim Zaim. 2013. A machine learning-based usability evaluation method for eLearning systems. *Decision Support Systems* 56 (Dec. 2013), 63–73. https://doi.org/10.1016/j.dss.2013.05.003

[45] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 74, 18 pages. https://doi.org/10.1145/3526113.3545616

[46] Helen Petrie, Fraser Hamilton, Neil King, and Pete Pavan. 2006. Remote usability evaluations With disabled people. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Montréal Québec Canada, 1133–1141. https://doi.org/10.1145/1124772.1124942

[47] John Pruitt and Tamara Adlin. 2005. *The Persona Lifecycle: Keeping People in Mind Throughout Product Design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[48] Ron Reckin and Vera Burckhardt. 2017. Instant Usability Report: Nutzertests mit Ergebnisdoku in 8 Stunden. In *Proceedings of Mensch und Computer 2017*. https://doi.org/10.18420/MUC2017-UP-0161 Publisher: Gesellschaft für Informatik e.V..

[49] Stuart J. Russell, Peter Norvig, and Ernest Davis. 2010. *Artificial intelligence: a modern approach* (3rd ed ed.). Prentice Hall, Upper Saddle River.

[50] Andreas Schuller, Doris Janssen, Julian Blumenröther, Theresa Maria Probst, Michael Schmidt, and Chandan Kumar. 2024. Generating personas using LLMs and assessing their viability. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–7. https://doi.org/10.1145/3613905.3650860

[51] Wilbert Tabone and Joost De Winter. 2023. Using ChatGPT for human–computer interaction research: a primer. *Royal Society Open Science* 10, 9 (Sept. 2023), 231053. https://doi.org/10.1098/rsos.231053

[52] David Thissen, Lynne Steinberg, and Daniel Kuang. 2002. Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of educational and behavioral statistics* 27, 1 (2002), 77–83.

[53] Manfred Thüring and Sascha Mahlke. 2007. Usability, aesthetics and emotions in human–technology interaction. 42, 4 (2007), 253–264. https://doi.org/10.1080/00207590701396674

[54] Kashyap Todi, Gilles Bailly, Luis Leiva, and Antti Oulasvirta. 2021. Adapting User Interfaces with Model-based Reinforcement Learning. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 573, 13 pages. https://doi.org/10.1145/3411764.3445497

[55] Dakuo Wang, Ting-Yao Hsu, Yuxuan Lu, Hansu Gu, Limeng Cui, Yaochen Xie, William Headean, Bingsheng Yao, Akash Veeragouni, Jiapeng Liu, Sreyashi Nag, and Jessie Wang. [n. d.]. AgentA/B: Automated and Scalable Web A/BTesting with Interactive LLM Agents. https://doi.org/10.48550/ARXIV.2504.09723 Version Number: 2.

[56] Ronald L Wasserstein and Nicole A Lazar. 2016. The ASA statement on p-values: context, process, and purpose. , 129–133 pages.

[57] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[58] Jacob O Wobbrock, Shaun K Kane, Krzysztof Z Gajos, Susumu Harada, and Jon Froehlich. 2011. Ability-based design: Concept, principles and examples. *ACM Transactions on Accessible Computing (TACCESS)* 3, 3 (2011), 1–27.

[59] Wei Xiang, Hanfei Zhu, Suqi Lou, Xinli Chen, Zhenghua Pan, Yuping Jin, Shi Chen, and Lingyun Sun. 2024. SimUser: Generating Usability Feedback by Simulating Various Users Interacting with Mobile Applications. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 9, 17 pages. https://doi.org/10.1145/3613904.3642481

[60] Jackie (Junrui) Yang, Yingtian Shi, Chris Gu, Zhang Zheng, Anisha Jain, Tianshi Li, Monica S. Lam, and James A. Landay. [n. d.]. GenieWizard: Multimodal App Feature Discovery with Large Language Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama Japan, 2025-04-26). ACM, 1–17. https://doi.org/10.1145/3706598.3714327

[61] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 11809–11822. https://proceedings.neurips.cc/paper_files/paper/2023/file/271db9922b8d1f4dd7aaef84ed5ac703-Paper-Conference.pdf

[62] Robert K. Yin. 2008. *Case Study Research: Design and Methods (Applied Social Research Methods)* (fourth edition. ed.). Sage Publications. http://www.amazon.de/Case-Study-Research-Methods-Applied/dp/1412960991%3FSubscriptionId%3D13CT5CVB80YFWJEPWS02%26tag%3Dws%26linkCode%3Dxm2%26camp%3D2025%26creative%3D165953%26creativeASIN%3D1412960991

[63] Hanning Zhao and Bilhanan Silverajan. 2024. Evaluating Cyber Security Dashboards for Smart Cities and Buildings: Enhancing User Modeling with LLMs. In *Proceedings of the 19th International Conference on Availability, Reliability and Security* (Vienna Austria, 2024-07-30). ACM, 1–10. https://doi.org/10.1145/3664476.3670943

Just Accepted

A Formats

A.1 Phase 1 Definition of Persona in JSON format

```
[Name]
[Age]
[Profession]
[Location]
[Role]
[Facet Description]
```

A.2 Phase 2 Prompt template for each Tester Agent

```
[Persona (taken from Phase 1)]
[html source code of web application]
[Usability Problem/ Suggestion Types (as found in Codebook)]
[Problem Severity (as defined in Codebook)]
[Problem Relevance (as defined in Codebook)]
[Task (Identify usability problems in the given web page using the DIN ISO 9241-
11 and produce a...
list of problems and suggestions in a JSON array using the following format) + Answer Format]
```

A.3 Output template in JSON format:

```
[Tester Name]
[ID]
[Problem/suggestion Type]
[Problem/suggestion Description]
[Problem/suggestion Severity]
[Problem/suggestion Relevance]
[Facets causing issues]
```

A.4 Sample Output

```
"Tester Name": "Abi",
"id": "Abi-01",
"Problem/suggestion Type": "Expectation",
"Problem/suggestion Description": "When I added an expense, I expected to see a confirmation...
...message indicating the expense was successfully added.",
"Problem/suggestion Severity": "S2",
"Problem/suggestion Relevance": "R2",
"Facets causing issues": "Abi's comprehensive information processing style means they like to have...
a clear understanding of task completion. The lack of feedback makes them uncertain about whether...
...their action was successful.",
"Solution": "After successfully adding an expense, display a confirmation message near the top of...
the form area, such as 'Expense added successfully!', which disappears after a few seconds."
```

B Codebooks

Just Accepted

Category	Description	*Example (Problems identified by humans)*
S1	Delays users briefly.	*Confused when deleting an output because no feedback*
S2	Delays users considerably, but ultimately allows them to fulfill the task.	*When editing an issue, Abi does not know where to edit the data and quickly becomes frustrated*
S3	Prevents users from fulfilling their tasks.	*When adding an expense, the lack of information on the button and uncertainty about what Amount means stops them from adding a new expense*

Table 11. Severity Codebook

Category	Description	*Example (Problems identified by humans)*
R1	Does not currently need to be edited.	*When adding a task, Pat is a little confused that the 'Delete' button is so present*
R2	Should be processed ASAP.	*Confused when adding an output because the green button is not labelled*
R3	Important: Should be processed ASAP.	*When editing an issue, Abi does not know where to edit the data and quickly becomes frustrated*
R4	Top priority: Absolutely *must* be processed - preferably first and as quickly as possible.	N/A for GenderMag Workshop output, but agent output used it.

Table 12. Relevance Codebook

Just Accepted

Category	Description	*Example*	Restriction on Use
Understanding the Problem	Users don't know what to do, don't recognise immediately that the displayed button leads to their intended action and overall goal. The possible causes of such issues are wordings in the interface are not meaningful, contents are not consistently named, and no explanations/information available.	*When adding an issue, Tim notices that the calendar is incorrect*	This generally causes restriction of use due to delays or failure.
Not seen	Users do not notice directly or overlook a button, object, or expect it elsewhere or users do not realize that the page continues below or users do not notice the feedback received after they performed an action. The possible causes for this are poor/unusual positioning of elements/objects/buttons, unfavourable design (e.g., contrast/ background), or lack of adherence with usability design standards.	*When editing an output, Tim would click on Edit several times and then realise that the text had already been transferred.*	This generally causes restriction of use due to delays or failure.
Uncertainty in the utilisation	Users don't understand what happened, don't understand if their action worked, don't understand where they navigated to, don't understand what to do next or if they have booked something when they clicked a button or object or if they can navigate back. The possible cause is that the user gets no feedback on user input (e.g., successful completion), no guidance on the following actions to reach their intention.	*When editing an issue, Abi does not know where to edit the data and quickly becomes frustrated*	This generally causes restriction of use due to delays or failure.
Personal Preference/ Behaviour	Relates to design or feature usage. Concerns include design/visualization, procedure, input of data, utilization of features.	*When adding a task, the red colour of the Delete button in Abi triggers a negative feeling*	This is an irritation or comment, but no restriction on use.
Expectation	The user expected a different display or interaction. This includes irritations or comments on display of content/symbols, links within the programme, interaction options, further steps. The user is surprised by the unexpected behaviour of the program (e.g., I would have expected a table.)	*When adding an expense, she is confused because she misses a currency for the expense*	This is an irritation or comment, but no restriction on use.
Request / Proposal	Suggestions for features or interface changes, types of interaction, information type, alternatives/ new ideas. e.g., *"I would like an overview here."*, *"Perhaps you could add another view here"*.	*When adding a task, a numbering of the entries or a different type of delimitation would have been desirable*	This is an irritation or comment, but no restriction on use.
Hint	Inconsistencies or ambiguities are pointed out. The user might be used to standards from other products. For example, *"A heading is missing here"*, *'I'm not quite sure what this is supposed to mean. A short description would be good"*.	*When adding a task, Abi asks herself where the list is saved*	This is an irritation or comment, but no restriction on use.
Bug	This is an error in the program, which can be functional (e.g., incorrect/missing links, incorrect interaction) or content issues (incorrect/missing content, copy-paste errors or typos).	*When adding an issue, Tim notices that the calendar is incorrect*	Might be a restriction depending on use

Table 13. Problem Type Codebook, adapted from Reckin and Burckhardt [48]

Just Accepted

C Data

Agreement per Persona	LLM-identified Diverging Problem	Human-identified Diverging Problem
As Abi, 13/14 (5 LLM, 9 Human)	N/A	• Confused when editing an output because edit button is blue and not green or grey
As Pat, 7/10 (5 LLM, 5 Human)	• The 'Edit' and 'Delete' buttons are small and closely positioned, which might lead to accidental clicks or overlooking their visibility. • Pat might expect a summary or overview display of their expenses after adding entries, which is currently absent in the UI.	• When adding an expense, the lack of information on the button and uncertainty about what "amount" means stops them from adding a new expense.
As Tim, 4/10 (6 LLM, 4 Human	• The user may overlook the feedback message upon pressing the 'Delete' button, as it does not provide immediate visual confirmation that the action is being processed. • The user may not realize that they can edit the expense entry immediately after adding it, as there is no instruction or indication that this is possible. • Tim may expect more robust features, such as an option to undo or cancel the delete action instantly instead of waiting for a timed removal. • A filter or search function for the expense table would help users like Tim manage entries more efficiently.	• When adding an output, Tim would waste too much time trying out all the functions • When editing an output, he is confused that the output is edited directly, but first goes back to the input and can be edited there

Table 14. Differences in Usability Problems Identified for Expense Tracker screen I

Just Accepted

Agreement per Persona	LLM-identified Diverging Problem	Human-identified Diverging Problem
As Abi, 2/9 (LLM 5, Human 4)	• The action buttons ('Add Expense', 'Update Expense', 'Cancel') may not clearly indicate their purpose for new users, particularly for Abi, who is unfamiliar with the interface. • The 'Update Expense' and 'Cancel' buttons are initially hidden, which may lead to Abi not realizing they can edit or cancel expense entries without clear navigation cues. • The 'Delete' action, implemented with a pending confirmation, may confuse Abi, especially since they are new to the system and might not understand how to cancel a pending deletion. • The arrangement and labeling of the input fields for expenses do not align with typical layouts seen in similar tools, which may lead to confusion for Abi, who has expectations based on prior experiences.	• When deleting an issue, Abi is afraid that entries will be lost • When editing outputs, the interface does not give any feedback, but only transfers the data • When deleting an issue, Abi is confused that the deletion process takes so long
As Pat, 4/11 (5 LLM, 6 Human)	• The table for displaying expenses may be overlooked if the user doesn't scroll down, which could discourage them from adding expenses. • The color scheme used for the pending delete action could be improved for better visibility; the alerting color may not be prominent enough to signal an important action. • Pat might expect an 'Are you sure you want to delete?' prompt when clicking delete, which is currently absent. Providing such a confirmation would match common user expectations. • When adding an issue, Pat noticed that the category does not provide a default and that confuses Pat	• When entering a task, the input format of the date pat gets lost because too many errors are possible • When editing an output, Pat is confused that the output appears at the top of the box, but quickly learns it • When deleting a task, Pat is briefly confused because the entry only turns red and wants a display with 'successfully deleted', but everything is still clear
As Tim, 5/12 (5 LLM, 7 Human)	• The labels on the buttons for adding and updating expenses ('Add Expense' and 'Update Expense') may not clearly signify that they will alter the table below. Users might wonder what will happen when they click these buttons, leading to confusion about their functionality. • The update and cancel buttons are hidden during the initial use of the expense form. Users might overlook these buttons after they are created, leading to confusion about how to edit or cancel an entry. • There is no clear message or help prompt explaining to users how to use the expense tracker effectively. Users will have to figure out how things work by trial and error. • When adding an issue, Tim wants a dropdown in the category field, but accepts the task easily • When adding an expense, Tim notices that the Euro field is not useful • When adding an issue, Tim notices that the calendar is incorrect	• When editing an output, Tim notices that the interface is not structured like familiar interfaces, but quickly gets to grips with the new interface

Table 15. Differences in Usability Problems Identified for Expense Tracker screen II

Just Accepted

Agreement per Persona	LLM-identified Diverging Problem	Human-identified Diverging Problem
As Abi, 2/8 (5 LLM, 3 Human)	• The 'Add' button does not provide a clear indication of its functionality. An unfamiliar user might not immediately recognize it as the action to add a task. • The input field for entering a new task may blend into the background, making it less likely that users will immediately notice it. • If a user attempts to add an empty task, there is no error message or feedback explaining why the action didn't work. This could lead to confusion and frustration. • The delete button could be mistaken for a secondary action rather than the primary action it represents. Expectations for button placement might differ from Abi's previous experiences.	• When adding a task, a numbering of the entries or a different type of delimitation would have been desirable • When adding a task, the red colour of the Delete button in Abi triggers a negative feeling
As Pat, 2/6 (4 LLM, 2 Human)	• The button to add a new task is labeled 'Add', which may not be immediately clear to all users that it is for submitting the new task input. A more descriptive label like 'Add Task' could help clarify its purpose. • The delete button is small and placed next to the task text. It may be overlooked, especially if users scan quickly through the list. It might be beneficial to make it more prominent. • Pat might expect the to-do items to display additional information, like timestamps or priority levels. The simplicity of the items may not meet their expectations for a task management tool.	• When adding a task, Pat is a little confused that the 'Delete' button is so present
As Tim, 0/6 (5 LLM, 1 Human)	• The 'Add' button does not clearly imply that it will add the typed task to the list. A more descriptive label such as 'Add Task' could improve clarity. • The input field and button could blend in with the background if not paying close attention, making the 'Add' action less prominent. • There is no visual feedback confirming that a new task has been added successfully; users might not realize their action was successful. • If Tim attempts to add an empty task (though it's already prevented), he might expect a message explaining why it didn't work. • Add a feature that allows users to edit tasks after they have been added to the list, which could enhance functionality.	• When adding a task, Tim is briefly confused that the entry appears directly below the input area because he is used to it from other applications

Table 16. Differences in Usability Problems Identified for To-do List

	Abi	Pat	Tim
To-Do List	0.89	0.88	0.88
Expense Tracker Screen I	0.90	0.87	0.88
Expense Tracker Screen II	0.88	0.93	0.88

Table 17. Human vs. System Token-Level Semantic Similarity Analysis of Usability Problem Descriptions Note. The scores consistently show high similarity [9] across all personas and interfaces.

Just Accepted

Participant	Interface	Facet	Count
F	Insurance Claim Interface	Self Efficacy	4
I	Insurance Claim Interface	Self Efficacy	3
I	Expense Tracker I	Self Efficacy	3
I	Insurance Claim Interface	Learning	2
I	Expense Tracker I	Motivations	2
C	Insurance Claim Interface	Risk	1
C	To Do List	Motivations	1
I	Expense Tracker I	Learning	1
I	To Do List	Self Efficacy	1
A	Expense Tracker I	Info Processing	1
E	Expense Tracker I	Risk	1
E	Expense Tracker II	Self Efficacy	1
D	Expense Tracker I	Self Efficacy	1
F	Insurance Claim Interface	Risk	0
C	Insurance Claim Interface	Self Efficacy	0
C	To Do List	Learning	0
C	To Do List	Self Efficacy	0
C	Expense Tracker II	Self Efficacy	0
C	Expense Tracker II	Motivations	0
C	Expense Tracker II	Learning	0
C	Expense Tracker II	Risk	0
I	To Do List	Motivations	0
B	To Do List	Self Efficacy	0
B	To Do List	Motivations	0
B	To Do List	Learning	0
B	To Do List	Risk	0
B	Expense Tracker II	Self Efficacy	0
B	Expense Tracker II	Motivations	0
B	Expense Tracker II	Learning	0
B	Expense Tracker II	Risk	0
A	Expense Tracker I	Self Efficacy	0
A	Expense Tracker I	Learning	0
A	Expense Tracker I	Risk	0
A	Expense Tracker II	Self Efficacy	0
A	Expense Tracker II	Learning	0
A	Expense Tracker II	Info Processing	0
A	Expense Tracker II	Risk	0
H	Insurance Claim Interface	Self Efficacy	0
E	Expense Tracker I	Self Efficacy	0
E	Expense Tracker II	Risk	0
D	Expense Tracker I	Info Processing	0
D	To Do List	Self Efficacy	0
D	To Do List	Info Processing	0

Table 18. Observed problem counts per participant, interface, and persona facet, sorted by count. Grey rows indicate misalignment between persona facet and observed problems.

Just Accepted

Comparison Group	Severity (KW Test)		Relevance (KW Test)		Problem Type (Multinomial LR)	
	H	p-value	H	p-value	χ^2	p-value
Abi vs Abi inverted vs Person A	0.26	0.88	5.43	0.07	8.09	0.88
Pat vs Person B	0.35	0.55	0.01	0.92	2.30	0.94
Tim vs Tim inverted vs Person C	1.09	0.58	2.29	0.32	8.35	0.87

Table 19. Bias Testing Results for Severity, Relevance, and Problem Type Across Persona Groups

Comparison	Severity ES	Relevance ES
Abi vs Abi inverted vs Person A	−0.12 [−0.14, 0.35]	0.25 [−0.05, 0.70]
Pat vs Person B	−0.15 [−0.58, 0.36]	0.03 [−0.55, 0.59]
Tim vs Tim inverted vs Person C	−0.06 [−0.13, 0.53]	0.02 [−0.12, 0.57]

Table 20. Effect sizes (η^2) for severity and relevance across tester groups. Confidence intervals in brackets.

Just Accepted