AI Personas for UX Research in Large Scale Retail

Karan Sharma[1,*], Abhijeet Phatak[1], Tatevik Petrosyan[1] and Chittaranjan Tripathy[1]

[1]*Walmart Global Tech, Sunnyvale, California, USA*

Abstract

Traditional moderated user testing is a cornerstone of User Experience (UX) research but is frequently constrained by high costs, long timelines, and difficulty in scaling. This paper introduces "Empathy-Based Simulation", a novel framework that utilizes multimodal Large Language Model (LLM)-based personas to mimic user behavior and provide qualitative feedback during the early stages of product development. By simulating responses to specific UI flows, the tool enables design teams to iterate rapidly before committing to expensive real-world studies. To validate the efficacy of this approach, we compare and evaluate LLM-generated insights with the traditional moderated UX testing results in the context of ecommerce and retail. Our findings demonstrate a high degree of alignment between AI personas and human participants regarding trust, clarity, and control, suggesting potential savings of over ninety percent in research vendor spend. Furthermore, we argue that Nielsen-Landauer's model extends naturally to AI personas in usability studies, where a small number of personas is sufficient to surface the majority of usability issues.

Keywords

AI personas, UX research, Large Language Models, Usability testing, Empathy-Based Simulation, Activation steering

1. Introduction

In the hyper-competitive landscape of global e-commerce, the ability to rapidly iterate on user interface (UI) designs is a critical survival advantage. However, traditional research methodologies often lag behind the development cycle due to recruitment and execution delays [1]. User interviews and moderated studies can require significant time for recruitment, execution, and analysis, creating bottlenecks in agile environments [1]. The convergence of multimodal Large Language Models (LLMs) and user persona development represents a natural evolution in research methodology [2]. LLMs are trained on diverse datasets containing varied linguistic and contextual information, while personas are traditionally developed as static user archetypes. This shared foundation allows LLMs to interpret and perceive UI elements in a human-like manner. This paper introduces a system tailored for large-scale retail environments, demonstrating that "artificial user" studies can provide actionable directional feedback that aligns with human behavior.

2. Motivation

Our work is motivated by the potential to bridge the gap between rapid product iteration and resource-intensive user testing by extending the Nielsen-Landauer [3] model of usability discovery to the domain of AI simulation. This foundational model is defined by the formula:

$$P = 1 - (1 - L)^n, \tag{1}$$

where n represents the number of participants, L denotes the proportion of usability problems uncovered by a single participant, and P represents the proportion of total problems discovered. This equation demonstrates that a sample size of five participants ($n = 5$) is typically sufficient to uncover

ACM CHI'26 Workshop on Responsible Use of AI Personas in Human-Centered Design and Research, April 13, 2026, Barcelona, Spain

*Corresponding author.

karan.sharma0@walmart.com (K. Sharma); abhijeet.phatak@walmart.com (A. Phatak); tatevik.petrosyan0@walmart.com (T. Petrosyan); ctripathy@walmart.com (C. Tripathy)

approximately 85% of an interface's core usability obstacles. We argue that this model extends naturally to AI-simulated personas. Just as a small but well-chosen cohort of human participants can surface the majority of usability problems, a lean set of AI personas can replicate the discovery dynamics of much larger human cohorts. This suggests that even modestly-sized pools of AI personas can achieve coverage comparable to that of traditional user studies.

3. Literature Review

3.1. The Evolution of Persona Generation

Traditionally, personas were static artifacts derived from multi-source data to represent specific target demographics. Recent advances have transitioned this practice toward dynamic, LLM-generated personas that human evaluators generally find informative, believable, positive, and relatable [4], although subject-matter experts rated them as slightly more stereotypical and less relatable compared to internal evaluators. By processing large-scale datasets from social media and consumer feedback, hybrid frameworks can generate richer, more contextually grounded personas [5], while autonomous LLM agents facilitate qualitative simulations that surface critical interaction patterns and usability defects early in the design cycle [1].

3.2. Methodologies in AI Persona Research

Current approaches to LLM-based persona generation follow three distinct trajectories that range from rapid creation to complex, multi-layered systems. Direct Prompt-Based Generation allows for the rapid development of personas using tailored prompts to achieve specific behavioral outcomes [4]. In contrast, more sophisticated systems employ Thematic Analysis Workflows, which involve initial coding and theme validation before persona generation to foster deeper human-AI collaboration [2]. Finally, Hybrid Persona Generation combines LLM-driven social media analysis with expert validation to bridge the gap between large-scale data and human intuition [5].

4. Methodology

4.1. System Design: Empathy-Based Simulation

The **Empathy-Based Simulation** tool utilizes a recursive qualitative pipeline. The process begins with defining a research objective and uploading the relevant UI screen or Figma prototype. In parallel, UX context (task flow, goals, pain points, success criteria) and research prompts (e.g., "What would you click next?") are prepared. AI persona profiles are then selected based on demographics, behaviors, or other signals, and the LLM is run to generate synthetic feedback including user-like quotes, heatmaps, and clarity/confusion scores. The results are compared against real user data where available — if the AI feedback aligns with real user insights, the process repeats for the next screen or variant; otherwise, the context and prompts are refined and the simulation is re-run. A schematic view of the workflow pipeline is shown in Figure 1.

4.2. Evaluation Methodology

To evaluate our methodology, we selected a real-world retail use case from Walmart. Following standard design protocols, Walmart UX designers typically conduct moderated usability studies using Figma prototypes prior to feature implementation. For this case study, we selected the "Quick Add with AI" feature for evaluation as shown in Figure 2. During the study, participants were presented with a series of qualitative probes, specifically: What do you think the "Quick Add with AI" button will do? What does the blue banner mean to you? Does it clearly explain the feature? What would encourage or discourage you from using this button? To benchmark these findings, we generated a diverse suite of AI personas and presented them with identical research questions. This dual-tracked approach allowed

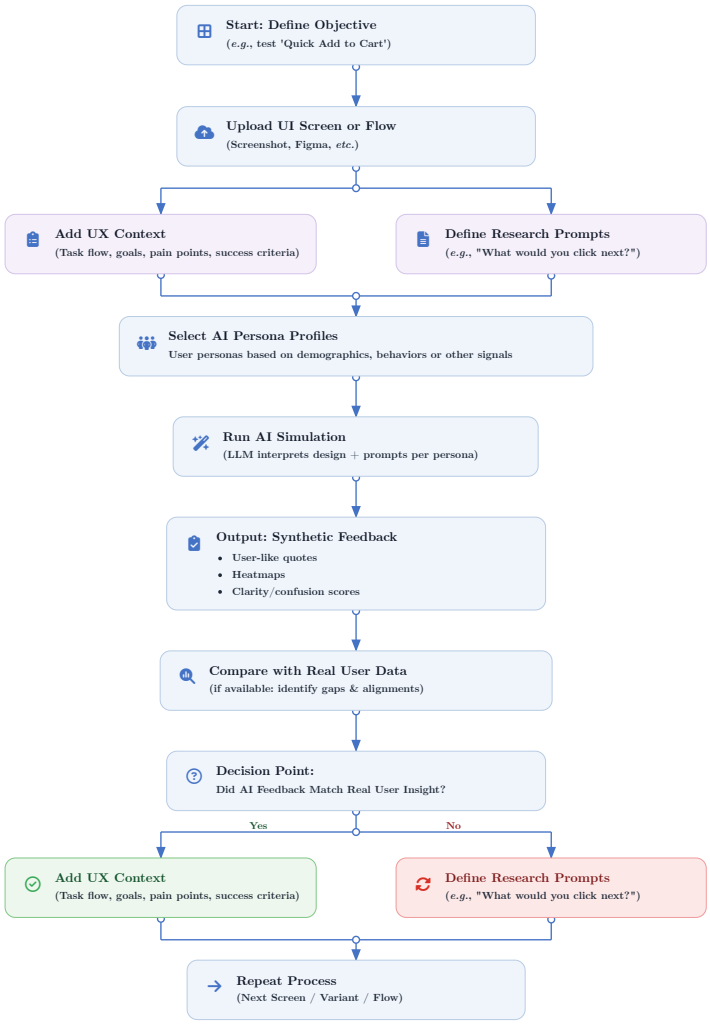


Figure 1: The Empathy-Based Simulation Workflow pipeline.

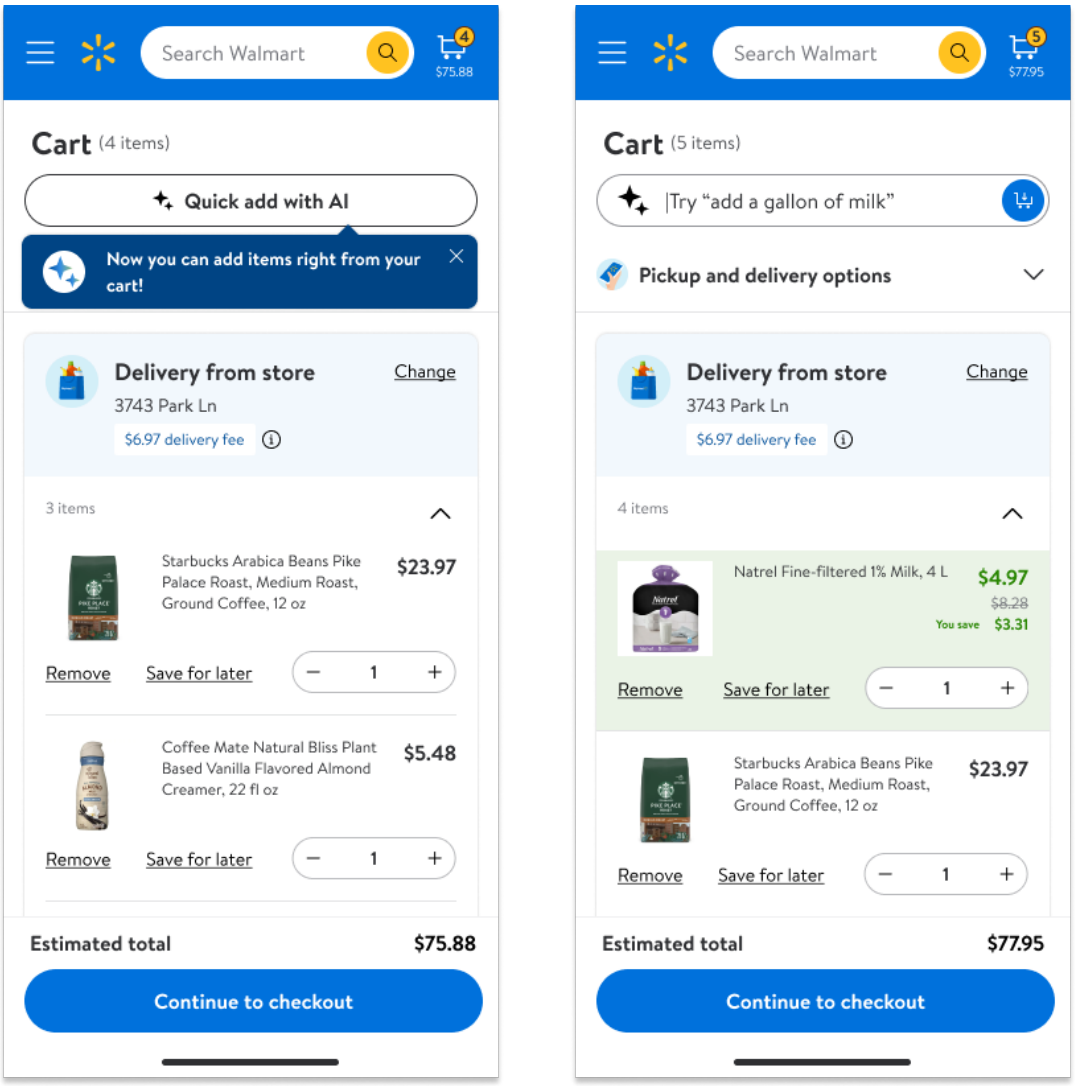
for a direct, systematic comparison between simulated insights from AI-based personas and traditional feedback from real users.

4.3. Persona Design and Demographics

The study deployed a deliberately lean persona set — 5 AI personas for Canada and 11 for Mexico — each engineered to represent a distinct user archetype across dimensions of digital literacy, financial background, and online shopping experience. Canadian profiles ranged from a tech-savvy parent in their mid-30s and a budget-conscious college student, to a hesitant 50-year-old digital adopter and a retiree encountering online shopping for the first time. These personas were constructed to mirror the demographic and behavioral characteristics observed in the 150-participant human benchmark dataset, enabling a direct comparison between simulated and real user insights.

5. Results

The comparative analysis revealed a remarkable degree of alignment between AI personas and real human feedback:



(a) Initial mobile cart state

(b) Updated cart after AI-assisted addition

Figure 2: UI stimulus: The 'Quick Add with AI' interface used for validation.

- **Trust:** Human participants expressed hesitation regarding brand accuracy ("What if it adds the wrong brand?"). AI personas mirrored this uncertainty exactly.
- **Control:** Both groups rejected automatic additions, preferring dropdown suggestions over automation.
- **Clarity:** Shared confusion was identified regarding the distinction between "Quick Add" and regular search.
- **Relevance:** Both groups agreed the feature's primary value was for mobile users to avoid navigation friction.

5.1. Failure Modes of AI Personas

While the AI personas demonstrated strong alignment with human participants on core usability themes, several divergence patterns were observed that highlight the current limits of synthetic simulation:

- **Emotional Hesitation:** Human participants frequently reacted emotionally (e.g., "This feels strange" or "I don't trust it"), whereas AI personas tended to provide more structured or rationalized explanations for their hesitation.

- **Assumed System Familiarity:** AI personas occasionally assumed an understanding of how predictive features worked that human participants did not possess, particularly regarding whether actions like "Quick Add" were automated or manual.
- **Behavioral Nuance:** Real participants showed physical hesitation patterns—such as pausing, re-reading UI elements, or exploring alternative options—which are inherently difficult to capture in text-based AI responses.
- **Context Sensitivity:** Some human participants reacted to specific contextual factors, such as immediate price sensitivity or deep-seated brand familiarity, that were less strongly represented in the synthetic persona responses.

5.2. Robustness of Simulation via Activation Engineering

The case study detailed above was conducted using AI personas created with prompts sent to a proprietary API. However, there are several shortcomings to this approach—notably relying on an external API, thus treating the model essentially as a blackbox. Latency and privacy issues can also render this approach non-pragmatic for enterprise applications.

To mitigate these shortcomings, we experimented with AI steering [6]. The objective is to create AI personas by directing the vectors extracted from intermediate layers in a direction that amplifies specific traits. **Activation Steering** solves this by mathematically enforcing the persona's perspective at the neuron level, rather than relying on the model to "remember" instructions from a system prompt:

1. **The "Direction" of Empathy:** We assume that concepts like "Budget Consciousness" or "Tech Illiteracy" are directions in the model's high-dimensional latent space.
2. **Contrast:** By taking the difference between a positive prompt (P_{pos}) and a negative prompt (P_{neg}), we isolate the vector V that represents that specific trait:

$$\mathbf{V} = \text{Activation}(P_{pos}) - \text{Activation}(P_{neg})$$

3. **Injection:** During the "Quick Add" test, we add this vector to the model's brain activity at every step controlled by a steering coefficient (α):

$$\text{NewActivation} = \text{OriginalActivation} + (\alpha \times \mathbf{V})$$

5.3. Results Based on Our Implementation

We implemented persona steering in BLIP-2 [7] (Salesforce/blip2-opt-2.7b), a state-of-the-art vision-language model that combines a vision encoder, Q-Former bridge module, and OPT-2.7B language model. Our approach targets only the language model component while preserving the vision processing pipeline intact. Using a Walmart shopping cart interface image, we tested tech enthusiast versus practical shopper personas by computing steering vectors from the difference between persona-specific and neutral prompt activations at layer 15 of the OPT model. The results demonstrate clear persona distinctions: the tech enthusiast persona (+1.5 coefficient) produces concise responses ("It adds the item to your cart") assuming technical familiarity, while the practical shopper persona (-1.5 coefficient) generates detailed explanations ("Quick add with AI is a feature that allows you to add items to your cart without having to go through the checkout process"). This approach successfully maintains BLIP-2's superior image understanding capabilities while enabling meaningful persona-driven response variations, proving that multimodal models can be effectively steered for personalized AI interactions without compromising visual accuracy.

6. Discussion

Despite the small persona count, AI-generated responses consistently surfaced the same core usability themes and critical defects identified by human participants — including hesitation around auto-add

features and concerns about transparency and user control. This alignment suggests that a carefully designed persona set can approximate the behavioral diversity of roughly 100 real-world users, pointing to a potential 90% improvement in early-stage research efficiency. This is aligned with Nielsen-Landauer model [3] of usability discovery.

While the AI personas demonstrated strong alignment with human participants on core usability themes, several divergence patterns were observed. AI-generated responses tended to exhibit more structured reasoning compared to the emotional and exploratory feedback provided by real users. Additionally, AI personas occasionally assumed familiarity with system behavior that human participants did not possess, particularly in early interactions with predictive features such as Quick Add. These differences highlight the importance of positioning AI simulation as a directional pre-research tool rather than a replacement for human validation.

The results confirm that LLM-based personas are viable for "directional feedback." While we acknowledge the "black-box" nature of LLMs as a limitation regarding interpretability, the strategic value in annual research savings is significant. AI should augment, rather than replace, human research to preserve the authenticity of human input [2]. Moving forward, we plan to implement a direct Figma Integration to allow designers to execute simulations within their primary prototyping environment for real-time iteration. We will also develop Visual Analytics, such as heatmap-style clustering, to pinpoint specific UI elements that trigger user confusion or friction. Finally, we aim to pursue Multimodal and Global Expansion by incorporating voice-enabled simulations and expanding our persona library to reflect diverse cultural nuances for international markets.

Acknowledgments

We thank Swati Kirti, Yog Domlur, and Siddharth Singh for their support. We also thank Jivesh Poddar and Ron Alunan for their help during the initial phases of this work.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] Y. Lu, B. Yao, H. Gu, J. Huang, Z. Wang, Y. Li, J. Gesi, Q. He, T. J.-J. Li, D. Wang, UXAgent: An LLM agent-based usability testing framework for web design, in: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), ACM, New York, NY, USA, 2025. doi:10.1145/3706599.3719729.

[2] S. De Paoli, Improved prompting and process for writing user personas with LLMs, arXiv:2310.06391 [cs.HC] (2023). doi:10.48550/arXiv.2310.06391.

[3] J. Nielsen, T. K. Landauer, A mathematical model of the finding of usability problems, in: Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems (CHI '93), CHI '93, Association for Computing Machinery, New York, NY, USA, 1993, pp. 206–213. doi:10.1145/169059.169083.

[4] J. Salminen, C. Liu, W. Pian, J. Chi, E. Häyhänen, B. J. Jansen, Deus Ex Machina and personas from large language models: Investigating the composition of AI-generated persona descriptions, in: Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), ACM, New York, NY, USA, 2024. doi:10.1145/3613904.3642036.

[5] M. Yin, H. Liu, B. Lian, C. Chai, Co-persona: Leveraging LLMs and expert collaboration to understand user personas through social media data analysis, arXiv:2506.18269 [cs.HC] (2025). doi:10.48550/arXiv.2506.18269.

[6] N. Rimsky, N. Peng, K. Damani, E. Hubinger, Steering Llama 2 via contrastive activation addition, arXiv:2312.06681 [cs.LG] (2023). doi:10.48550/arXiv.2312.06681.

[7] J. Li, D. Li, S. Savarese, S. Hoi, BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, in: Proceedings of the 40th International Conference on Machine Learning (ICML 2023), PMLR, 2023, pp. 19730–19742.