Toward Building Behavioral Testbeds for Security and Privacy: LLM-Driven Personas as Crash Dummies

Amir Reza Asadi
University of Cincinnati

Joel Appiah
University of Cincinnati

Taiwo Peter Akinemi
University of Cincinnati

Hazem Said
University of Cincinnati

Abstract

The computing world increasingly focuses on data collection, and the integration of advanced IT technologies creates new privacy and security vulnerabilities. Traditional approaches to security and privacy testing lack the scale and diversity needed to anticipate the full range of potential vulnerabilities. This ongoing work proposes using large language models (LLMs) to identify these vulnerabilities by having LLMs role-play personas of diverse users, including threat actors, regular users, and security practitioners. We created a pool of 128 individual personas derived from security and privacy literature and developed a framework to evaluate how effectively LLMs can embody these personas across standardized security scenarios. We validate persona simulation using this framework.

1 Introduction

As information technology transformed our lives [18] and made it more entangled, from personal context to industrial scene, the balance of privacy and security became more critical, and technology development outpaced the development of regulatory frameworks [3, 4] and security measures [13].

LLM and diverse knowledge of web data [27, 30], have created opportunities for threat modeling [19, 29], and fine grained simulation of user behaviors [14], as LLM agent can simulate different user personas [27] which are imaginary people who represents real user segments [22]. LLMs and digital twin of human have been utilized in mirroring privacy risks in social media [7].

USENIX Symposium on Usable Privacy and Security (SOUPS) 2025.
August 10–12, 2025, Seattle, WA, United States.

We propose addressing the escalating threats to user privacy [2], and growing cybersecurity vulnerabilities [12] by creating a behavioral testbed [10] to simulate user personas. Bowles [8] describes personas as crash dummies for personified experimenting of privacy and security testing scenarios, and LLMs makes the concept of crash dummies interactive. The goal is to create a testbed that generates vulnerability reports for defined scenarios by testing on an extensive set of personas. The testbed focuses on vulnerabilities in system interaction flows and controls that are influenced by individual behavioral patterns, including social dynamics, trust relationships, and adversarial manipulation tactics [17].

By building on the structured approach of usability engineering through scenario-based design [9, 21], we developed a testing platform to simulate these personas. This testbed uses the GOMS framework [11] to structure persona-system interactions and generates critical decision points and user actions of defined systems, and scenarios.

Following the recent research which shows that LLMs as evaluators for persona role-playing achieve correlations of up to 80.7% with human evaluators [23], LLMs will be used for evaluating the behavior of personas in testing scenarios.

This effort aims to validate persona simulation in security and privacy testing to identify potential risks. The following research questions guide this investigation:

- **RQ.1** To what extent do LLMs adhere to persona characteristics in roleplaying security and privacy scenarios?
- **RQ.2** To what extent do different LLMs exhibit diverse behavioral patterns when simulating various personas?
- **RQ.3** To what degree does simulating personas in a scenario reveal new vulnerabilities?
- **RQ.4:** To what extent do LLM persona simulations match human analysts in vulnerability detection?

2 Methodology & Experiment Design

This section explains the experiment, and a schematic overview of the steps is presented in Figure 1.

Through a systematic literature review across Scopus, ACM Digital Library, and IEEE Xplore, 128 individual personas in security and privacy literature in 35 articles were identified [6]. We extended these personas by rounding them with quantified characteristics suitable for LLM simulation (see Appendix for examples). Personas were categorized into three categories of regular users, threat actors, and security practitioners. The testing platform (see Appendix A for the overview) will simulate these personas using three LLMs including GPT-4 [1], Claude 3.7 Sonnet [5], and Llama 3.1 70B [16]) to role-play their behaviors across security scenarios. The personas will perform actions in standardized security scenarios such as phishing email responses [20].

Each test scenario will undergo automatic assessment using three quantitative metrics. Additionally, LLM-generated and identified labels will be validated by human evaluators.

2.1 Evaluation Framework

We propose an evaluation framework that includes three quantitative metrics. The first, Persona Fidelity Index, measures to what extent the simulated behaviors adhere to defined persona attributes. The second, Behavioral Diversity Index, quantifies how different personas generate diverse actions within scenarios. The third, Vulnerability Detection Rate, measures how effectively each persona uncovers unique security vulnerabilities across testing scenarios.

2.1.1 Persona Fidelity Index (RQ1)

We create persona behavioral attributes as the vector $\vec{C}(p_i)$ for each persona with values from one to five. Each persona attribute is assigned a value of 1-5. For instance, for risk tolerance, a score of 1 means low risk-taking behavior while 5 indicates high risk tolerance. Next, the system generates $\vec{S}(p_i)$ by having an LLM grade the characteristics demonstrated in persona actions on a 1-5 scale. The Persona Fidelity Index measures similarity between these vectors using cosine similarity [15, 25], and score range range from 0 (complete disparity) to 1 (perfect match):

$$\text{PFI}(p_i) = \frac{\vec{C}(p_i) \cdot \vec{S}(p_i)}{|\vec{C}(p_i)| \cdot |\vec{S}(p_i)|} \tag{1}$$

2.1.2 Behavioral Diversity Index (RQ2)

To calculate Behavioral Diversity Index, we compare the actions taken by different personas. We create a persona-action matrix for each scenario where each row represents a persona and each column an observed action. For example, in such a matrix, persona p_1 might have performed actions 1 and 5 (represented by 1's in those positions and 0's elsewhere). These actions are identified automatically by LLMs.

To measure diversity, we calculate the entropy [24] of action usage:

$$\text{DI}(scenario) = -\sum_{j=1}^{m} p(a_j) \log p(a_j) \tag{2}$$

Where $p(a_j)$ is the proportion of personas that perform action a_j. Higher entropy indicates greater behavioral diversity across personas. To reduce variation among behaviors with overlapping meanings, behaviors will be transformed into vector embeddings, allowing semantic similarity to be measured.

2.1.3 Vulnerability Detection Rate (RQ3) and LLM Performance (RQ4)

Next, we measure how effectively different persona-LLM combinations discover security vulnerabilities by calculating Discovery Rate. The vulnerabilities in generated scenarios will be extracted with another prompting with an LLM. The experiment will compare vulnerability discovery between three expert security analysts and the three LLMs using identical scenarios.

$$\text{Discovery Score}(p_i, M_j) = \frac{\text{unique vulnerabilities found}}{\text{scenarios tested}} \tag{3}$$

2.2 Limitations

LLMs may hallucinate inconsistent persona behaviors [**?**]. We mitigate this through persona fidelity measurement, multi-model sampling, and sampling outputs for human evaluator grading. Our findings provide insights for expert review rather than definitive predictions.

3 Anticipated Contribution

This ongoing work will harness gulf of knowledge in LLMs for privacy and security plannings. It guides future LLM-based usability testing research, extends scenario-based design theory [9] to include persona simulation, and provides practical applications such as:

Proactive Treats & Vulnerability Discovery A large pool of personas can create a digital twin that brings opportunities for proactive vulnerability identification in underexplored situations by enabling testing across a broad spectrum of user behaviors.

Optimization Framework for Persona Simulation The evaluation framework enables optimization of persona simulations by quantifying LLM role-playing performance, which advances security engineering and privacy testings.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Saifuddin Ahmed and Sangwon Lee. The inhibition effect: Privacy concerns disrupt the positive effects of social media use on online political participation. *New Media & Society*, 27(1):203–224, 2025.

[3] Mohammed Fayyad Hassan Al-Jabouri. The evolution of privacy laws in the digital age: Balancing innovation and personal security. *Utu Journal of Legal Studies (UJLS)*, 1(1):39–45, 2024.

[4] Fnu Antara, Sarika Goel, and Pandi Kirupa Gopalakrishna Pandian. Network security measures in cloud infrastructure: A comprehensive study. *International Journal of Innovative Research in Technology (IJIRT)*, 9(3), August 2022. J-309, Pocket J, Sarita Vihar, Delhi, India; Research Supervisor, Mahgu, Uttarakhand; Sobha Emerald Phase 1, Jakkur, Bangalore.

[5] Anthropic. Claude 3.7 sonnet and claude code, 2025. Accessed.

[6] Amir Reza Asadi, Yuchong Zhang, and Hazem Said. What do personas say about privacy & security: a systematic literature review through human-ai collaboration. Under review, 2025.

[7] Frederik Simon Bäumer, Sergej Schultenkämper, Michaela Geierhos, and Yeong Su Lee. Mirroring privacy risks with digital twins: When pieces of personal data suddenly fit together. *SN Computer Science*, 5(8):1109, November 2024.

[8] John B. Bowles. The personification of reliability, safety, and security. In *2007 Annual Reliability and Maintainability Symposium*, pages 161–166, 2007.

[9] John M Carroll. Scenario-based design. In *Handbook of human-computer interaction*, pages 383–406. Elsevier, 1997.

[10] Kasthuri Jayarajah, Rajesh Krishna Balan, Meera Radhakrishnan, Archan Misra, and Youngki Lee. Livelabs: Building in-situ mobile sensing & behavioural experimentation testbeds. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '16, page 1–15, New York, NY, USA, 2016. Association for Computing Machinery.

[11] Bonnie E. John. Chapter 4 - information processing and skilled behavior. In John M. Carroll, editor, *HCI Models, Theories, and Frameworks*, Interactive Technologies, pages 55–101. Morgan Kaufmann, San Francisco, 2003.

[12] Antanas Kedys. Fast-changing cyber threat landscape and a new reality of cyber security. *Cyber Security: A Peer-Reviewed Journal*, 8(3):273–280, 2025. Full text not available for purchase; only accessible to subscribers.

[13] Qudus Lawal. Advancing cybersecurity: Strategies for mitigating threats in evolving digital and iot ecosystems. *International Research Journal of Modernization in Engineering Technology and Science*, 7(1), January 2025. Peer-Reviewed, Open Access, Fully Refereed International Journal. Impact Factor: 8.187.

[14] Kun Li, Chenwei Dai, Wei Zhou, and Songlin Hu. Fine-grained behavior simulation with role-playing large language model on social media. *arXiv preprint arXiv:2412.03148*, 2024.

[15] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, USA, 2008.

[16] Inc. Meta Platforms. Llama 3.1 70b. https://huggingface.co/meta-llama/Llama-3.1-70B, 2024. Accessed: 2025-05-13.

[17] Shutonu Mitra, Tomas Neguyen, Qi Zhang, Hyungmin Kim, Hossein Salemi, Chen-Wei Chang, Fengxiu Zhang, Michin Hong, Chang-Tien Lu, Hemant Purohit, and Jin-Hee Cho. Scvi: Bridging social and cyber dimensions for comprehensive vulnerability assessment, 2025.

[18] Sara Quach, Park Thaichon, Kelly D Martin, Scott Weaven, and Robert W Palmatier. Digital technologies: tensions in privacy and data. *Journal of the Academy of Marketing Science*, 50(6):1299–1323, November 2022.

[19] Maria Rigaki, Ondřej Lukáš, Carlos Catania, and Sebastian Garcia. Prompt. exploit. repeat: Automating network security testing with llms. In Ana Paula Rocha, Luc Steels, and Jaap van den Herik, editors, *Agents and Artificial Intelligence*, pages 15–36, Cham, 2025. Springer Nature Switzerland.

[20] Waldo Rocha Flores, Hannes Holm, Gustav Svensson, and Göran Ericsson. Using phishing experiments and scenario-based surveys to understand security behaviours in practice. *Information Management & Computer Security*, 22(4):393–406, 2014.

[21] Mary Beth Rosson and John M Carroll. *Usability engineering: scenario-based development of human-computer interaction*. Morgan Kaufmann, 2002.

[22] Joni Salminen, Kathleen Guan, Soon-Gyo Jung, and Bernard J. Jansen and. A survey of 15 years of data-driven persona development. *International Journal of Human–Computer Interaction*, 37(18):1685–1708, 2021.

[23] Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. Personagym: Evaluating persona agents and llms. *arXiv preprint arXiv:2407.18416*, 2024.

[24] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[25] Harald Steck, Chaitanya Ekanadham, and Nathan Kallus. Is cosine-similarity of embeddings really about similarity? In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 887–890, New York, NY, USA, 2024. Association for Computing Machinery.

[26] Muhammad Adnan Tariq, Joel Brynielsson, and Henrik Artman. Framing the attacker in organized cybercrime. In *2012 European Intelligence and Security Informatics Conference*, pages 30–37. IEEE, 2012.

[27] Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. User behavior simulation with large language model-based agents. *ACM Trans. Inf. Syst.*, 43(2), January 2025.

[28] xAI. Grok 3. https://grok.com, 2025. Artificial intelligence chatbot.

[29] Shuiqiao Yang, Tingmin Wu, Shigang Liu, David Nguyen, Seung Jang, and Alsharif Abuadbba. Threatmodeling-llm: Automating threat modeling using large language models for banking system. *arXiv preprint arXiv:2411.17058*, 2024.

[30] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2), 2023.

A Appendix A: Research Overview

A.1 Experiment Overview

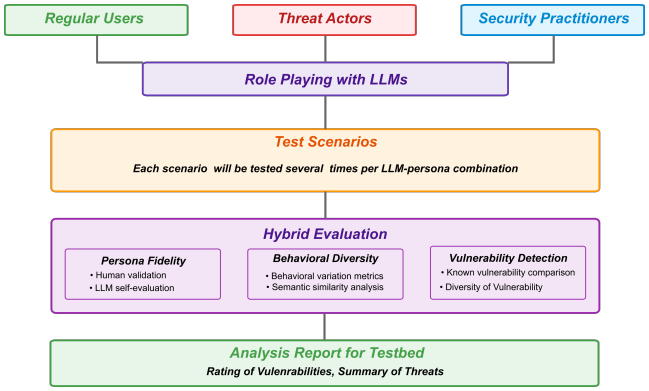

Figure 1: Experiment Overview

B Testbed Details

B.1 Testbed Design

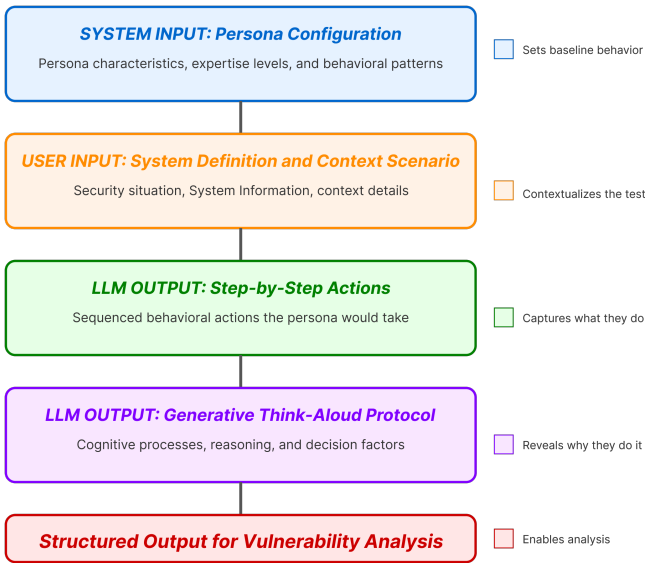

Figure 2: Testbed Platform

B.2 Working System

The ongoing implementation of our behavioral testbed is available at https://github.com/AmirrezaAsadi/behavorialtestbed and will be updated.

The system is structured around the GOMS (Goals, Operators, Methods, Selection) [9] framework, originally developed

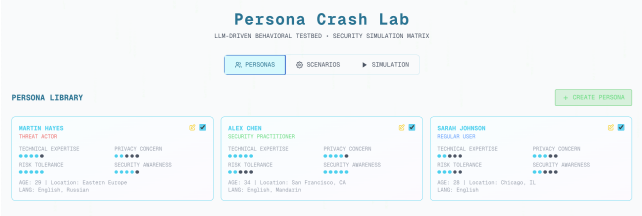

Figure 3: Persona Selection Page

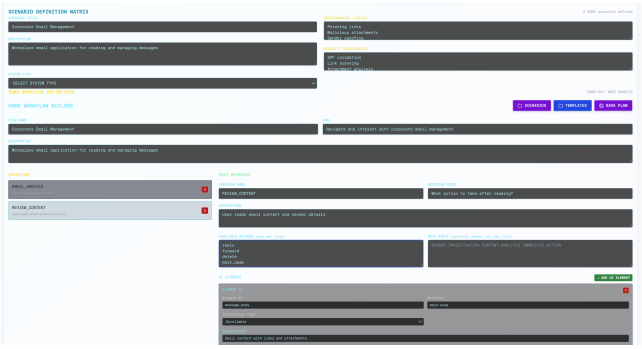

Figure 4: Scenario Builder Interface

for human-computer interaction modeling. In this testbed, Goals are derived from individual persona motivations and objectives, Operators represent system-defined workflow steps and interface elements, Methods are simulated through LLM-generated persona behaviors, and Selection rules emerge from persona-specific decision-making patterns based on their defined characteristics such as risk tolerance and security awareness. Figure 3 demonstrates the first tab of the application where users can select personas for simulation. In Figure 4 they define the system scenario in the second tab, and in the third tab 5 they view the simulation results and outcomes.

B.2.1 System and Scenario Configuration Interface

Figure 4 demonstrates the workflow builder interface used to define security testing scenarios. The left panel displays the operator hierarchy with two sequential operators: `EMAIL_ARRIVES` (initial trigger) and `REVIEW_CONTENT` (analysis phase).

The main configuration area includes detailed operator definition for `REVIEW_CONTENT`. It defines the decision point "What action to take after reading?" and available actions such as delete, mark spam, click links, and download attachments. The next steps field specifies subsequent operators: `SENDER_INVESTIGATION`, `CONTENT_ANALYSIS`, and `IMMEDIATE_ACTION`.

The UI Elements section defines the interface context with Element #1 configured as `message_body` positioned in the main-area with scrollable interaction type.

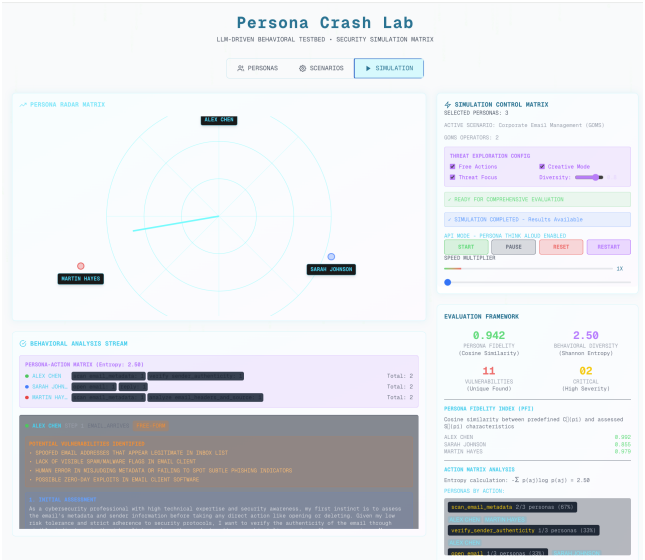

Figure 5: Simulation Progress

B.2.2 Evaluation Outcome

Figure 6 presents the comprehensive evaluation framework displaying real-time metrics during simulation execution. The results are based on simulation of only three personas using Grok 3 [28] LLM. It indicates high level of fidelity, and variety in actions taken by personas as behavorial diversity is above zero.

The Action Matrix Analysis section reveals behavioral distribution patterns, and provides insights on how many personas did which actions.

C Appendix B: Rounded Persona

Figure 6: Attacker Persona Example, Developed by extending [26]